A Proofs

A.1 Proof of Theorem 4.1

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

$$\widehat{\beta}_i = \begin{cases} 0, & \text{if } |Y_i| \leq \Delta. \\ [|Y_i| - \sigma^2 \lambda^*(\widehat{\beta}_i)]_+ \text{sign}(Y_i), & \text{otherwise.} \end{cases} \tag{1}$$

After re-weighting, our objective function becomes

$$\begin{aligned} \widetilde{\boldsymbol{\beta}} &= \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ -\sum_{i=1}^{n} \frac{w_i}{2\sigma^2}(Y_i - \beta_i)^2 + \sum_{j=1}^{p} \log \pi(\beta_j - \mu_j \mid \theta) \right\} \\ &= \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ -\sum_{i=1}^{n} \frac{1}{2\sigma^2}(Y_i^* - \beta_i^*)^2 + \sum_{j=1}^{p} \log \pi^*(\beta_j^* \mid \theta) \right\} \end{aligned}$$

where $Y_i^* = \sqrt{w_i}(Y_i - \mu_i)$, $\beta_i^* = \sqrt{w_i}(\beta_i - \mu_i)$, and we are imposing prior π^* with $\lambda_0' = \lambda_0/\sqrt{w_i}$ and $\lambda_1' = \lambda_1/\sqrt{w_i}$. So from (1), the mode estimator $\widehat{\beta_i^*}$ for β_i^* satisfies

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

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

$$\widetilde{\beta}_i = \begin{cases} \mu_i, & \text{if } |\sqrt{w_i}(Y_i - \mu_i)| \leq \Delta. \\ \mu_i + [|Y_i - \mu_i| - \frac{\sigma^2}{w_i}\lambda^*(\tilde{\beta}_i - \mu_i)]_+ \text{sign}(\sqrt{w_i}(Y_i - \mu_i)), & \text{otherwise.} \end{cases} \tag{2}$$

For active coordinates,

$$\begin{aligned} &\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(\tilde{\beta}_i - \beta_i^0)^2 \mid Y] = \mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(\tilde{\beta}_i - Y_i + Y_i - \beta_i^0)^2 \mid Y] \\ &\leq 2\sigma^2 + 2\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(\tilde{\beta}_i - Y_i)^2 \mid Y] \\ &= 2\sigma^2 + 2\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(\tilde{\beta}_i - Y_i)^2 \mid Y]\mathbb{I}(\sqrt{w_i}|Y_i - \mu_i| \leq \Delta) + 2\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(\tilde{\beta}_i - Y_i)^2 \mid Y]\mathbb{I}(\sqrt{w_i}(Y_i - \mu_i) > \Delta) \end{aligned} \tag{3}$$

where

$$\begin{aligned} &\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(\tilde{\beta}_i - Y_i)^2 \mid Y]\mathbb{I}(\sqrt{w_i}|Y_i - \mu_i| \leq \Delta) \\ &= \mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(\mu_i - Y_i)^2 \mid Y]\mathbb{I}(|Y_i - \mu_i| \leq \frac{\Delta}{\sqrt{w_i}}) \leq \Delta^2 \mathbb{E}_{w_i}\frac{1}{w_i} \end{aligned} \tag{4}$$

and

$$\begin{aligned} &\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(\tilde{\beta}_i - Y_i)^2 \mid Y]\mathbb{I}(\sqrt{w_i}(Y_i - \mu_i) > \Delta) \\ &= \mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(\frac{1}{w_i}\lambda^*(\tilde{\beta}_i - \mu_i))^2 \mid Y]\mathbb{I}(\sqrt{w_i}|Y_i - \mu_i| > \Delta) \leq 4\mathbb{E}_{w_i}\frac{1}{w_i^2} \end{aligned} \tag{5}$$

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., $p^*(\tilde{\beta}_i - \mu_i) < c_+$ when $|\sqrt{w_i}(Y_i - \mu_i)| > \Delta$ (notice p_i^* uses λ_0 and λ_1). Thus, since $c_+(1-c_+) = \frac{1}{(\lambda_0-\lambda_1)^2}$, we have $\lambda^*(\tilde{\beta}_i - \mu_i) < c_+(\lambda_1 - \lambda_0) + \lambda_0 = (1-c_+)(\lambda_0 - \lambda_1) + \lambda_1 < \frac{2}{\lambda_0 - \lambda_1} + 1 < 2$ when $|\sqrt{w_i}(Y_i - \mu_i)| > \Delta$.

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

$$\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(\tilde{\beta}_i - \beta_i^0)^2 \mid Y] \le 2\sigma^2 + 2\Delta^2\mathbb{E}_{w_i}\frac{1}{w_i} + 8\mathbb{E}_{w_i}\frac{1}{w_i^2} \le C_3\Delta^2$$

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

$$\begin{aligned}
&\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(\tilde{\beta}_i - \beta_i^0)^2 \mid Y] = \mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(\tilde{\beta}_i)^2 \mid Y] \\
=&\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(\tilde{\beta}_i - \mu_i + \mu_i)^2 \mid Y]\mathbb{I}(\sqrt{w_i}|Y_i - \mu_i| \ge \Delta) \\
\le&\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(|Y_i - \mu_i| + |\mu_i|)^2 \mid Y]\mathbb{I}(\sqrt{w_i}|Y_i - \mu_i| \ge \Delta) \\
\le&2\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(|Y_i - \mu_i|^2 + |\mu_i|^2) \mid Y]\mathbb{I}(\sqrt{w_i}|Y_i - \mu_i| \ge \Delta)\mathbb{I}(w_i \le t)\mathbb{I}(|\mu_i| \le \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}}) \\
&+ 2\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(|Y_i - \mu_i|^2 + |\mu_i|^2) \mid Y]\mathbb{I}(\sqrt{w_i}|Y_i - \mu_i| \ge \Delta)\mathbb{I}(w_i \le t)\mathbb{I}(|\mu_i| > \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}}) \\
&+ 2\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(|Y_i - \mu_i|^2 + |\mu_i|^2) \mid Y]\mathbb{I}(\sqrt{w_i}|Y_i - \mu_i| \ge \Delta)\mathbb{I}(w_i > t) \\
\doteq&U_1 + U_2 + U_3
\end{aligned} \tag{6}$$

For the first term U_1 in (6),

$$\begin{aligned}
&U_1 = 2\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}\left(|Y_i - \mu_i|^2 + |\mu_i|^2 \mid Y\right)\mathbb{I}\left(\sqrt{w_i}|Y_i - \mu_i| \ge \Delta\right)\mathbb{I}(w_i \le t)\mathbb{I}\left(|\mu_i| \le \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}}\right) \\
\le&2\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i}\left(|Y_i - \mu_i|^2 \mid Y\right)\mathbb{I}\left(|Y_i - \mu_i| \ge \frac{\Delta}{\sqrt{t}}\right)\mathbb{I}\left(|\mu_i| \le \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}}\right) + 2\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}\mu_i^2\mathbb{I}\left(|Y_i| \ge \frac{\Delta}{\sqrt{t}} - \frac{1}{\lambda_0} - \frac{1}{\sqrt{\lambda_0}}\right) \\
\overset{(a)}{=}&2\sigma^2\mathbb{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{\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)\left(1 - \Phi\left(\frac{\mu_i}{\sigma} + \frac{\Delta}{\sigma\sqrt{t}}\right) + \Phi(\frac{\mu_i}{\sigma} - \frac{\Delta}{\sigma\sqrt{t}})\right)\right] \\
&\times\mathbb{I}\left(|\mu_i| \le \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}}\right) + \frac{4}{\lambda_0^2}2\Phi\left(-\frac{\Delta}{\sigma\sqrt{t}} + \frac{1}{\sigma\lambda_0} - \frac{1}{\sigma\sqrt{\lambda_0}}\right) \\
\le&4\sigma^2\mathbb{E}_{\mu_i}\left[\left(\frac{\Delta}{\sigma\sqrt{t}} + \frac{|\mu_i|}{\sigma}\right)\phi\left(\frac{|\mu_i|}{\sigma} - \frac{\Delta}{\sqrt{t}}\right) + (1+\mu_i^2)\Phi\left(\frac{|\mu_i|}{\sigma} - \frac{\Delta}{\sqrt{t}}\right)\right]\mathbb{I}\left(|\mu_i| \le \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}}\right) + \frac{16}{\lambda_0^2}\frac{\phi\left(-\frac{1}{\sigma\lambda_0} + \frac{1}{\sigma\sqrt{\lambda_0}} + \frac{\Delta}{\sigma\sqrt{t}}\right)}{-\frac{1}{\sigma\lambda_0} + \frac{1}{\sigma\sqrt{\lambda_0}} + \frac{\Delta}{\sigma\sqrt{t}}} \\
\le&4\sigma^2\mathbb{E}_{\mu_i}\left[\left(\frac{\Delta}{\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]\mathbb{I}\left(|\mu_i| \le \frac{1}{\lambda_0} + \frac{1}{\sqrt{\lambda_0}}\right) \\
&+ 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) \\
\overset{(c)}{\le}&5\sigma\frac{\Delta}{\sqrt{t}}\phi(\frac{\Delta}{\sigma\sqrt{t}} - \frac{1}{\sigma\lambda_0} + \frac{1}{\sigma\sqrt{\lambda_0}})
\end{aligned} \tag{7}$$

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

$$
\begin{aligned}
&\mathbb{E}_{Y_i}\left(|Y_i-\mu_i|^2 \mid Y\right)\mathbb{I}\left(|Y_i-\mu_i| \geq \frac{\Delta}{\sqrt{t}}\right) \\
\overset{(b)}{=}&\sigma^2\int_{Y_i-\mu_i/\sigma>\frac{\Delta}{\sigma\sqrt{t}}}(Y_i-\mu_i/\sigma)^2\phi(Y_i)dY_i+\sigma^2\int_{Y_i-\mu_i/\sigma<-\frac{\Delta}{\sigma\sqrt{t}}}(Y_i-\mu_i/\sigma)^2\phi(Y_i)dY_i \\
=&\int_{Y_i-\mu_i/\sigma>\frac{\Delta}{\sigma\sqrt{t}}}Y_i^2\phi(Y_i)dY_i-2\mu_i\sigma\int_{Y_i-\mu_i/\sigma>\frac{\Delta}{\sigma\sqrt{t}}}Y_i\phi(Y_i)dY_i+\mu_i^2\int_{Y_i-\mu_i/\sigma>\frac{\Delta}{\sigma\sqrt{t}}}\phi(Y_i)dY_i \\
&+\sigma^2\int_{Y_i-\mu_i/\sigma<-\frac{\Delta}{\sigma\sqrt{t}}}Y_i^2\phi(Y_i)dY_i-2\mu_i\sigma\int_{Y_i-\mu_i/\sigma<-\frac{\Delta}{\sigma\sqrt{t}}}Y_i\phi(Y_i)dY_i+\mu_i^2\int_{Y_i-\mu_i/\sigma<-\frac{\Delta}{\sigma\sqrt{t}}}\phi(Y_i)dY_i \\
=&-\sigma^2\int_{Y_i-\mu_i/\sigma>\frac{\Delta}{\sigma\sqrt{t}}}Y_i\phi'(Y_i)dY_i+2\mu_i\sigma\int_{Y_i-\mu_i/\sigma>\frac{\Delta}{\sigma\sqrt{t}}}\phi'(Y_i)dY_i+\mu_i^2\left[1-\Phi(\mu_i/\sigma+\frac{\Delta}{\sigma\sqrt{t}})\right] \\
&-\sigma^2\int_{Y_i-\mu_i/\sigma<-\frac{\Delta}{\sigma\sqrt{t}}}Y_i\phi'(Y_i)dY_i+2\mu_i\sigma\int_{Y_i-\mu_i/\sigma<-\frac{\Delta}{\sigma\sqrt{t}}}\phi'(Y_i)dY_i+\mu_i^2\Phi(\mu_i/\sigma-\frac{\Delta}{\sigma\sqrt{t}}) \\
=&-\sigma^2\left[Y_i\phi(Y_i)\mid_{Y_i+\frac{\Delta}{\sigma\sqrt{t}}}^{+\infty}-\int_{Y_i>\mu_i/\sigma+\frac{\Delta}{\sqrt{t}}}\phi(Y_i)dY_i\right]+2\mu_i\sigma\phi(Y_i)\mid_{\mu_i+\frac{\Delta}{\sigma\sqrt{t}}}^{+\infty}+\mu_i^2\left[1-\Phi(\mu_i/\sigma+\frac{\Delta}{\sigma\sqrt{t}})\right] \\
&-\sigma^2\left[Y_i\phi(Y_i)\mid_{-\infty}^{\mu_i-\frac{\Delta}{\sigma\sqrt{t}}}-\int_{Y_i<\mu_i/\sigma-\frac{\Delta}{\sigma\sqrt{t}}}\phi(Y_i)dY_i\right]+2\mu_i\sigma\phi(Y_i)\mid_{-\infty}^{\mu_i/\sigma-\frac{\Delta}{\sigma\sqrt{t}}}+\mu_i^2\left[1-\Phi(\mu_i/\sigma-\frac{\Delta}{\sigma\sqrt{t}})\right] \\
\overset{(d)}{=}&\sigma^2\left(\frac{\Delta}{\sigma\sqrt{t}}-\mu_i/\sigma\right)\phi\left(\mu_i/\sigma+\frac{\Delta}{\sigma\sqrt{t}}\right)+\left(\frac{\Delta}{\sigma\sqrt{t}}+\mu_i/\sigma\right)\phi\left(\mu_i/\sigma-\frac{\Delta}{\sigma\sqrt{t}}\right) \\
&+(1+\mu_i^2)\left(1-\Phi\left(\mu_i/\sigma+\frac{\Delta}{\sigma\sqrt{t}}\right)+\Phi(\mu_i/\sigma-\frac{\Delta}{\sigma\sqrt{t}})\right)
\end{aligned}
$$

where (b) use change of variable $Y_i \leftarrow Y_i/\sigma$. Equality (d) and (c) in (7) uses Mills ratio ($\frac{1-\Phi(x)}{\phi(x)} \to \frac{1}{x}$ as $x \to \infty$).

For the second term in (6),

$$
\begin{aligned}
U_2 &= 2\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}\left[(|Y_i-\mu_i|^2+|\mu_i|^2)\mid Y\right]\mathbb{I}(\sqrt{w_i}|Y_i-\mu_i|\geq\Delta)\mathbb{I}(w_i\leq t)\mathbb{I}\left(|\mu_i|>\frac{1}{\lambda_0}+\frac{1}{\sqrt{\lambda_0}}\right) \\
&\leq 2\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(|Y_i-\mu_i|^2+|\mu_i|^2)\mid Y]\mathbb{I}\left(|\mu_i|>\frac{1}{\lambda_0}+\frac{1}{\sqrt{\lambda_0}}\right) \\
&\leq 2\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(2Y_i^2+3\mu_i^2)\mid Y]\mathbb{I}\left(|\mu_i|>\frac{1}{\lambda_0}+\frac{1}{\sqrt{\lambda_0}}\right) \\
&= 2\mathbb{E}_{\mu_i}(2\sigma^2+3\mu_i^2)\mathbb{I}\left(|\mu_i|>\frac{1}{\lambda_0}+\frac{1}{\sqrt{\lambda_0}}\right) \\
&= 4\sigma^2\mathbb{P}(|\mu_i|>\frac{1}{\lambda_0}+\frac{1}{\sqrt{\lambda_0}})+6\mathbb{E}_{\mu_i}\mu_i^2\mathbb{I}\left(|\mu_i|>\frac{1}{\lambda_0}+\frac{1}{\sqrt{\lambda_0}}\right) \\
&\overset{(e)}{=} 4\sigma^2e^{-1-\sqrt{\lambda_0}}+6\left[\left(\frac{1}{\lambda_0}+\frac{1}{\sqrt{\lambda_0}}\right)^2+\frac{2}{\lambda_0}\left(\frac{2}{\lambda_0}+\frac{1}{\sqrt{\lambda_0}}\right)\right]e^{-1-\sqrt{\lambda_0}}<\sigma\frac{\Delta}{\sqrt{t}}\phi\left(\frac{\Delta}{\sigma\sqrt{t}}-\frac{1}{\sigma\lambda_0}+\frac{1}{\sigma\sqrt{\lambda_0}}\right)
\end{aligned}
$$

where (e) follows from integration by parts.

For the third term in (6), utilizing condition (iii), we have $\mathbb{P}(w_i > t) \leq \widetilde{C}_3 \frac{\Delta}{\sqrt{t}} \phi\left(\frac{\Delta}{\sqrt{t}}\right)$, thus

$$
\begin{aligned}
U_3 =& 2\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(|Y_i-\mu_i|^2+|\mu_i|^2)\,|\,Y]\mathbb{I}(\sqrt{w_i}|Y_i-\mu_i|\geq\Delta)\mathbb{I}(w_i>t) \\
\leq& 2\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(2|Y_i|^2+3|\mu_i|^2)\,|\,Y]\mathbb{I}(w_i>t) = \left(4\sigma^2+\frac{12}{\lambda_0^2}\right)\mathbb{E}_{w_i}\mathbb{I}(w_i>t) \leq C_2\sigma\frac{\Delta}{\sqrt{t}}\phi\left(\frac{\Delta}{\sigma\sqrt{t}}\right)
\end{aligned}
$$

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

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

and the risk for all coordinates satisfies

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{Y}}\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{w}}[|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0|^2\,|\,Y] &\leq qC_1\Delta^2+(n-q)\sigma(6+C_2)\frac{\Delta}{\sqrt{t}}\phi\left(\frac{\Delta}{\sigma\sqrt{t}}-\frac{1}{\sigma\lambda_0}+\frac{1}{\sigma\sqrt{\lambda_0}}\right) \\
&= qC_1\Delta^2+\widetilde{C_2}(n-q)\sigma\frac{\Delta}{\sqrt{t}}e^{-\frac{1}{2}\left(\frac{\Delta}{\sigma\sqrt{t}}-\frac{1}{\sigma\lambda_0}+\frac{1}{\sigma\sqrt{\lambda_0}}\right)^2} \leq qC_1\Delta^2+\widetilde{C_2'}(n-q)\frac{\Delta^U}{\sqrt{t}}e^{-\frac{1}{2}\left(\frac{\Delta^L}{\sigma\sqrt{t}}\right)^2} \\
&\leq qC_1\Delta^2+\widetilde{C_2'}(n-q)\frac{\Delta^U}{\sqrt{t}}e^{-\frac{1}{2}\frac{2\log[1/p^*(0)]-2}{t}} \leq qC_1\Delta^2+\widehat{C_2}(n-q)\Delta^U\left(\frac{q}{n}\right)^{\frac{\eta+\gamma}{t}}
\end{aligned}
$$

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

Set $t = \eta+\gamma$, $\mathbb{E}_{\boldsymbol{Y}}\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{w}}[|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0|^2\,|\,Y] \leq qC_3\Delta^2$. Then from Markov Inequality, for any $M_n \to \infty$,

$$
\mathbb{E}_{\boldsymbol{Y}}\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{w}}\left(\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|_2^2 > M_nq\log\left(\frac{n}{q}\right)\,|\,Y\right) \leq \mathbb{E}_{\boldsymbol{Y}}\frac{\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{w}}\left[\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|^2\,|\,Y\right]}{M_nq\log\frac{n}{q}} \leq \frac{qC_3\Delta^2}{M_nq\log\frac{n}{q}}
$$

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

$$
\mathbb{E}_{\boldsymbol{Y}}\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{w}}\left(\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|_2^2 > M_nq\log\left(\frac{n}{q}\right)\,|\,Y\right) \to 0
$$

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 $\boldsymbol{w} \sim n\text{Dir}(\alpha,\cdots,\alpha)$, $w_i \sim n \times \text{Beta}(\alpha,(n-1)\alpha)$. Both of them satisfy condition (ii). So we only need to check

condition (iii). Write $B(\cdot)$ as the Beta function. When $\boldsymbol{w} \sim n\text{Dir}(\alpha, \cdots, \alpha)$, then for any $t \geq \frac{\alpha+1}{\alpha}$, the following equation holds

$$
\begin{aligned}
&\mathbb{P}_{w_i}(w_i > t) = \mathbb{P}(Beta(\alpha, n\alpha - \alpha) > \frac{t}{n}) = \frac{\int_{t/n}^{1} v_i^{\alpha-1}(1-v_i)^{n\alpha-\alpha-1}dv_i}{B(\alpha, n\alpha-\alpha)} \\
&\overset{z=\alpha n v_i}{=} \frac{1}{B(\alpha, n\alpha-\alpha)} \int_{\alpha t}^{\alpha n} (\frac{z}{\alpha n})^{\alpha-1}(1-\frac{z}{\alpha n})^{n\alpha-\alpha-1} d\frac{z}{\alpha n} \\
&= \frac{1}{B(\alpha, n\alpha-\alpha)(\alpha n)^{\alpha}} \int_{\alpha t}^{\alpha n} z^{\alpha-1}[(1-\frac{1}{\alpha n/z})^{-\alpha n/z+1}]^{\frac{n\alpha-\alpha-1}{1-\alpha n/z}} dz \\
&\overset{(a)}{\leq} \frac{1}{B(\alpha, n\alpha-\alpha)(\alpha n)^{\alpha}} \int_{\alpha t}^{\alpha n} z^{\alpha-1} e^{\frac{n\alpha-\alpha-1}{z-\alpha n}z} dz \\
&\leq \frac{1}{B(\alpha, n\alpha-\alpha)(\alpha n)^{\alpha}} \int_{\alpha t}^{\alpha n} z^{\alpha-1} e^{\frac{n\alpha-\alpha-1}{\alpha t-\alpha n}z} dz \\
&\overset{(b)}{\leq} \frac{1}{B(\alpha, n\alpha-\alpha)(\alpha n)^{\alpha}} \int_{\alpha t}^{\alpha n} z^{\alpha-1} e^{-z} dz = \frac{\Gamma(n\alpha)}{\Gamma(n\alpha-\alpha)(\alpha n)^{\alpha}\Gamma(\alpha)} \int_{\alpha t}^{\alpha n} z^{\alpha-1} e^{-z} dz \\
&\leq \frac{\Gamma(n\alpha)}{\Gamma(n\alpha-\alpha)(\alpha n)^{\alpha}} \mathbb{P}(Gamma(\alpha, 1) > \alpha t) \leq \mathbb{P}(Gamma(\alpha, 1) > \alpha t) \\
&\overset{(c)}{\leq} \frac{\mathbb{E}_{v\sim Gamma(\alpha,1)} e^{xv}}{e^{\alpha t x}} = \frac{(1-x)^{-\alpha}}{e^{\alpha t x}} = e^{\alpha(-log(1-x)-tx)}
\end{aligned}
$$

where (a) uses the fact $(1-\frac{1}{x})^{-x+1} \leq e$ 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

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

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

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

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

Now $\boldsymbol{w} = (w_1, w_2, \cdots, w_n) \sim nDir(\alpha, \alpha, \cdots, \alpha)$. One important property we will use

below is that $nBeta(\alpha, n\alpha-\alpha) \xrightarrow{\mathrm{d}} \frac{1}{\alpha}Gamma(\alpha,1)$ for any fixed α.

$$\begin{aligned}
\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(\tilde{\beta}_i - Y_i)^2 \mid Y]\mathbb{I}(\sqrt{w_i}(Y_i-\mu_i) > \Delta) &= \mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(\frac{1}{w_i}\lambda^*(\tilde{\beta}_i-\mu_i))^2 \mid Y]\mathbb{I}(\sqrt{w_i}|Y_i-\mu_i| > \Delta) \\
&\geq \lambda_1^2 \mathbb{E}_{Y_i,\mu_i}\mathbb{E}_{w_i}\frac{1}{w_i^2}\mathbb{I}(w_i > \frac{\Delta^2}{(Y_i-\mu_i)^2}) \\
&\overset{(a)}{=} \lambda_1^2 \mathbb{E}_{Y_i,\mu_i}\mathbb{E}_{v_i\sim Gamma(\alpha,1)}\frac{\alpha^2}{v_i^2}\mathbb{I}(v_i > \frac{\alpha\Delta^2}{(Y_i-\mu_i)^2}) + o(1) \\
&= \lambda_1^2\alpha^2 \mathbb{E}_{Y_i,\mu_i}\mathbb{E}_{z_i\sim Inv-Gamma(\alpha,1)} z_i^2 \mathbb{I}(z_i < \frac{(Y_i-\mu_i)^2}{\alpha\Delta^2}) + o(1)
\end{aligned}$$

where (a) follows from the following derivations:

$w_i \xrightarrow{\mathrm{d}} \frac{1}{\alpha}Gamma(\alpha,1)$, 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 $f(w_i) = \frac{1}{w_i^2}\mathbb{I}(w_i > \frac{\Delta^2}{(Y_i-\mu_i)^2})$ is X-continuous and bounded by $\frac{(Y_i-\mu_i)^2}{\Delta^4}$, so $\mathbb{E}f(w_i) \to \mathbb{E}f(v_i/\alpha)$. It means that if we let $f_n(Y_i,\mu_i) = \mathbb{E}_{w_i\sim n\mathrm{Beta}(\alpha,n\alpha-\alpha)}\left[\frac{1}{w_i^2}\mathbb{I}(w_i > \frac{\Delta^2}{(Y_i-\mu_i)^2})\right]$ and $f(Y_i,\mu_i) = \mathbb{E}_{v_i\sim Gamma(\alpha,1)}\left[\frac{\alpha^2}{v_i^2}\mathbb{I}(v_i > \frac{\alpha\Delta^2}{(Y_i-\mu_i)^2})\right]$, then, $f_n(Y_i,\mu_i) \to f(Y_i,\mu_i)$ for any Y_i and μ_i. Also notice that $|f_n(Y_i,\mu_i)| \leq \frac{(Y_i-\mu_i)^4}{\Delta^4}$ and $\frac{(Y_i-\mu_i)^4}{\Delta^4}$ is integrable. So from Dominated Convergence Theorem, $\mathbb{E}_{Y_i,\mu_i}f_n(Y_i,\mu_i) \to \mathbb{E}_{Y_i,\mu_i}f(Y_i,\mu_i)$, i.e, $\mathbb{E}_{Y_i,\mu_i}\mathbb{E}_{w_i}\frac{1}{w_i^2}\mathbb{I}(w_i > \frac{\Delta^2}{(Y_i-\mu_i)^2}) \to \mathbb{E}_{Y_i,\mu_i}\mathbb{E}_{v_i\sim Gamma(\alpha,1)}\frac{\alpha^2}{v_i^2}\mathbb{I}(v_i > \frac{\alpha\Delta^2}{(Y_i-\mu_i)^2})$.

Since

$$\begin{aligned}
&\mathbb{E}_{z\sim Inv-Gamma(\alpha,1)} z^2\mathbb{I}(z<M) = \int_0^M z^2\frac{1}{z^{\alpha+1}}e^{-1/z}dz = \int_0^M z^{1-\alpha}e^{-1/z}dz \\
&= \frac{1}{2-\alpha}\int_0^M e^{-1/z}dz^{2-\alpha} = \frac{1}{2-\alpha}(e^{-1/z}z^{2-\alpha} \mid_0^M - \int_o^M z^{2-\alpha}de^{-1/z}) \\
&= \frac{1}{2-\alpha}(e^{-1/M}M^{2-\alpha} + \int_0^M \frac{1}{z^\alpha}e^{-1/z}dz) > \frac{1}{2-\alpha}e^{-1/M}M^{2-\alpha}
\end{aligned}$$

So

$$\begin{aligned}
&\mathbb{E}_{Y_i}\mathbb{E}_{\mu_i,w_i}[(\tilde{\beta}_i - Y_i)^2 \mid Y]\mathbb{I}(\sqrt{w_i}(Y_i-\mu_i) > \Delta) \geq \lambda_1^2\alpha^2 \mathbb{E}_{Y_i,\mu_i}\mathbb{E}_{z_i\sim Inv-Gamma(\alpha,1)} z_i^2 \mathbb{I}(z_i < \frac{(Y_i-\mu_i)^2}{\alpha\Delta^2}) + o(1) \\
&\geq \lambda_1^2\alpha^2 \mathbb{E}_{Y_i,\mu_i}\frac{1}{2-\alpha}e^{-\alpha\Delta^2/(Y_i-\mu_i)^2}(\frac{(Y_i-\mu_i)^2}{\alpha\Delta^2})^{2-\alpha} + o(1)
\end{aligned}$$

When $\alpha < 2$, if we want $\lambda_1^2\mathbb{E}_{Y_i,\mu_i}\mathbb{E}_{w_i}\frac{1}{w_i^2}\mathbb{I}(w_i > \frac{\Delta^2}{(Y_i-\mu_i)^2}) \preceq \Delta^2$ for any β_i^0, we need

$$\mathbb{E}_{Y_i,\mu_i}e^{-\alpha\Delta^2/(Y_i-\mu_i)^2}(\frac{(Y_i-\mu_i)^2}{\alpha\Delta^2})^{2-\alpha} \preceq \Delta^2$$

to hold for any β_i^0. But if we set $\beta_i^0 = 2\Delta^t$ where $t = \frac{\widetilde{C}}{2-\alpha} + 1$, then

$$
\begin{aligned}
&\mathbb{E}_{Y_i,\mu_i} e^{-\alpha\Delta^2/(Y_i-\mu_i)^2} \big(\frac{(Y_i-\mu_i)^2}{\alpha\Delta^2}\big)^{2-\alpha} \\
&\geq \mathbb{E}_{Y_i,\mu_i} e^{-\alpha\Delta^2/(Y_i-\mu_i)^2} \big[\frac{(Y_i-\mu_i)^2}{\alpha\Delta^2}\big]^{2-\alpha} \mathbb{I}(|Y_i| \geq \frac{3}{2}\Delta^t)\mathbb{I}(|\mu_i| < 1) \\
&\geq \mathbb{E}_{Y_i,\mu_i} e^{-\frac{\alpha\Delta^2}{(\frac{3}{2}\Delta^t-1)^2}} \big[\frac{(\frac{3}{2}\Delta^t-1)^2}{\alpha\Delta^2}\big]^{2-\alpha} \mathbb{I}(|Y_i| \geq \frac{3}{2}\Delta^t)\mathbb{I}(|\mu_i| < 1) \\
&\geq e^{-\frac{\alpha\Delta^2}{(\frac{3}{2}\Delta^t-1)^2}} \big[\frac{(\frac{3}{2}\Delta^t-1)^2}{\alpha\Delta^2}\big]^{2-\alpha} \mathbb{E}_{Y_i}\mathbb{I}(|Y_i| \geq \frac{3}{2}\Delta^t)\mathbb{E}_{\mu_i}\mathbb{I}(|\mu_i| < 1) \\
&\overset{(b)}{\geq} e^{-\frac{\alpha\Delta^2}{(\Delta^t)^2}} \big[\frac{(\Delta^t)^2}{\alpha\Delta^2}\big]^{2-\alpha} C = Ce^{-\alpha\Delta^{2-2t}} \frac{1}{\alpha^{2-\alpha}} \Delta^{(2t-2)(2-\alpha)} \\
&\geq \widehat{C}\Delta^{(2t-2)(2-\alpha)}\alpha^{\alpha-2} \geq C_5\Delta^{2\widetilde{C}}
\end{aligned}
$$

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 $\widetilde{C}$ can be arbitrarily large, the risk (depending on the truth β_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 $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. We denote $pen(\boldsymbol{\beta} \mid \theta) = log[\frac{\pi(\boldsymbol{\beta} \mid \theta)}{\pi(\mathbf{0}_p \mid \theta)}]$. We write $\boldsymbol{W} = \text{diag}(\sqrt{w_1}, \cdots, \sqrt{w_n})$. We denote by $Q(\boldsymbol{\beta}) = -\frac{1}{2}||\boldsymbol{W}\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{X}\boldsymbol{\mu} - \boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta}||^2 + pen(\boldsymbol{\beta} \mid \theta)$ and $\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} Q(\boldsymbol{\beta})$. Notice that BB-SSL solution $\widetilde{\boldsymbol{\beta}}$ is $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} + \boldsymbol{\mu}$. The matrix norm $\|\cdot\|_a$ is defined as $\|\boldsymbol{X}\|_a = \sup_{\boldsymbol{\beta}} \frac{\|\boldsymbol{X}\boldsymbol{\beta}\|_a}{\|\boldsymbol{\beta}\|_a}$ where $\|\cdot\|_a$ is the vector a-norm. Write $\boldsymbol{\Theta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$. We use $|\boldsymbol{\beta}|$ and $\|\boldsymbol{\beta}\|_1$ 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 $\boldsymbol{X}$ with penalty $pen(\boldsymbol{\beta})$ satisfies the $\tilde{\eta}$-null consistency ($\tilde{\eta}$-NC) condition if*

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

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

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

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

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

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

where $\eta^* = \max\left\{\widetilde{\eta} + C_n\frac{\|\boldsymbol{X}\|}{\lambda_1}, \frac{\widetilde{\eta}}{m}\right\}$ *and* $\widetilde{C}_n$ *is any sequence that satisfies* $C_n \to \infty$.

Lemma A.3. *If* $\arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p}\{-\frac{1}{2}||\boldsymbol{W}(\boldsymbol{\epsilon}-\boldsymbol{X}\boldsymbol{\mu})/\eta^* - \boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta}||^2 + pen(\boldsymbol{\beta}\,|\,\theta)\} = \mathbf{0}_p$ *where* $\eta^* \in (0,1]$*, then BB-SSL estimate* $\widetilde{\boldsymbol{\beta}}$ *lies in the cone* $C(\eta^*;\boldsymbol{\beta}) = \{\boldsymbol{\Theta}\in\mathbb{R}^p : (\eta^*+1)pen(\boldsymbol{\Theta}_S\,|\,\theta) \le (1-\eta^*)pen(\boldsymbol{\Theta}_{S^C}\,|\,\theta)\}$ *with high probablity, where* S *is the active set of* β_j*'s.*

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

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

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

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

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

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 $\text{pen}'(\boldsymbol{\beta}) = \text{pen}(\boldsymbol{\beta})/n$ and thus, $\rho'(t) = \rho(t)/n$. Notice that ρ' satisfies $\rho'(t) \ge \lambda_1(|t|)/n$. From condition (5), $\lambda/n \ge (1+\xi_0)\frac{\sigma}{\tilde{\eta}}\sqrt{n}(1+\sqrt{2\log(2p/\delta)})$ 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

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

and thus

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

A.6 Proof of Lemma A.2

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

$$
\begin{aligned}
&Var_{\boldsymbol{\mu},\boldsymbol{w}}((\boldsymbol{\epsilon}-\boldsymbol{X}\boldsymbol{\mu})^T\boldsymbol{W}^2\boldsymbol{X}\boldsymbol{\beta}\,|\,\boldsymbol{\epsilon}) = Var_{\boldsymbol{\mu},\boldsymbol{w}}\left(\sum_i(\epsilon_i-\boldsymbol{x}_i^T\boldsymbol{\mu})w_i\boldsymbol{x}_i^T\boldsymbol{\beta}\,|\,\boldsymbol{\epsilon}\right)\\
=&\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{w}}\left[\sum_i(w_i-1)\epsilon_i\boldsymbol{x}_i^T\boldsymbol{\beta}-\sum_i w_i\boldsymbol{x}_i^T\boldsymbol{\mu}\boldsymbol{x}_i^T\boldsymbol{\beta}\right]^2\\
=&\mathbb{E}_{\boldsymbol{w}}\left[\sum_i(w_i-1)\epsilon_i\boldsymbol{x}_i^T\boldsymbol{\beta}\right]^2+\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{w}}\left[\sum_i w_i\boldsymbol{x}_i^T\boldsymbol{\mu}\boldsymbol{x}_i^T\boldsymbol{\beta}\right]^2\\
=&\sum_i Var(w_i)(\epsilon_i\boldsymbol{x}_i^T\boldsymbol{\beta})^2+\sum_{i\neq j}Cov(w_i,w_j)\epsilon_i\boldsymbol{x}_i^T\boldsymbol{\beta}\epsilon_j\boldsymbol{x}_j^T\boldsymbol{\beta}+\sum_{i,j}\boldsymbol{x}_i^T\boldsymbol{\beta}\boldsymbol{x}_j^T\boldsymbol{\beta}\boldsymbol{x}_i^T\mathbb{E}(\boldsymbol{\mu}\boldsymbol{\mu}^T)\boldsymbol{x}_j\mathbb{E}w_jw_j\\
\overset{(a)}{\leq}&\frac{C_1}{\log n}\sum_i(\epsilon_i\boldsymbol{x}_i^T\boldsymbol{\beta})^2+\frac{C_2}{n\log n}\sum_{i\neq j}\frac{1}{2}\left[(\epsilon_i\boldsymbol{x}_i^T\boldsymbol{\beta})^2+(\epsilon_j\boldsymbol{x}_j^T\boldsymbol{\beta})^2\right]+\mathbb{E}w_iw_j\frac{2}{\lambda_0^2}\sum_{i,j}\boldsymbol{x}_i^T\boldsymbol{\beta}\boldsymbol{x}_j^T\boldsymbol{\beta}(\boldsymbol{x}_i^T\boldsymbol{x}_j)\\
\overset{(b)}{\leq}&\frac{C_1}{\log n}\sum_i(\epsilon_i\boldsymbol{x}_i^T\boldsymbol{\beta})^2+\frac{C_2}{n\log n}(n-1)\sum_i(\epsilon_i\boldsymbol{x}_i^T\boldsymbol{\beta})^2+\mathbb{E}w_iw_j\frac{2n}{\lambda_0^2}(\sum_i\boldsymbol{x}_i^T\boldsymbol{\beta})^2\\
\overset{(c)}{\leq}&\frac{\widetilde{C}_1}{\log n}\|\boldsymbol{\epsilon}\|_\infty^2\|\boldsymbol{X}\boldsymbol{\beta}\|^2+C_3\frac{2n^2}{\lambda_0^2}\sum_i(\boldsymbol{x}_i^T\boldsymbol{\beta})^2\\
\overset{(d)}{\leq}&\widetilde{C}_3\|\boldsymbol{X}\boldsymbol{\beta}\|^2
\end{aligned}
$$

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

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

$$
\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{w}}\left(|(\boldsymbol{\epsilon}-\boldsymbol{X}\boldsymbol{\mu})^T\boldsymbol{W}^2\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{\epsilon}^T\boldsymbol{X}\boldsymbol{\beta}|>t\,|\,\boldsymbol{X},\boldsymbol{\epsilon}\right)\leq\frac{Var_{\boldsymbol{\mu},\boldsymbol{w}}((\boldsymbol{\epsilon}-\boldsymbol{X}\boldsymbol{\mu})^T\boldsymbol{W}^2\boldsymbol{X}\boldsymbol{\beta})}{t^2}\leq\frac{\widetilde{C}_3\|\boldsymbol{X}\boldsymbol{\beta}\|^2}{t^2}
$$

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

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

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

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

Notice that

$$
\|\boldsymbol{X}\boldsymbol{\beta}\|+\frac{\|\boldsymbol{X}\|}{\lambda_1}\mathrm{pen}(\boldsymbol{\beta}\,|\,\theta)\overset{(e)}{\leq}\|\boldsymbol{X}\boldsymbol{\beta}\|+\frac{\|\boldsymbol{X}\|}{\lambda_1}(-\lambda_1\boldsymbol{\beta})\leq\|\boldsymbol{X}\|\times|\boldsymbol{\beta}|-\|\boldsymbol{X}\|\times|\boldsymbol{\beta}|=0
$$

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

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

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

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

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

$$\begin{aligned} & -\frac{\eta^*}{2}\|\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta}\|^2 + (\boldsymbol{\epsilon} - \boldsymbol{X}\boldsymbol{\mu})^T \boldsymbol{W}^2 \boldsymbol{X}\boldsymbol{\beta} + \eta^* \text{pen}(\boldsymbol{\beta} \mid \theta) \\ \overset{(f)}{\leq} & -\frac{\eta^* m}{2}\|\boldsymbol{X}\boldsymbol{\beta}\|^2 + |\boldsymbol{\epsilon}^T \boldsymbol{X}\boldsymbol{\beta}| + (\eta^* - C_n \frac{\|\boldsymbol{X}\|}{\lambda_1})\text{pen}(\boldsymbol{\beta} \mid \theta) \\ \overset{(g)}{\leq} & -\frac{\tilde{\eta}}{2}\|\boldsymbol{X}\boldsymbol{\zeta}\|^2 + \boldsymbol{\epsilon}^T \boldsymbol{X}\boldsymbol{\zeta} + \tilde{\eta}\text{pen}(\boldsymbol{\zeta} \mid \theta) \leq 0 \end{aligned} \tag{13}$$

where (f) follows from assumption (3) in Theorem 4.2 and equation (11). Inequality (g) follows from the definition of η^* and the fact that $\text{pen}(\boldsymbol{\beta} \mid \theta) \leq 0$ for any $\boldsymbol{\beta}$. We set $\boldsymbol{\zeta} = \boldsymbol{\beta}$ if $\boldsymbol{\epsilon}^T \boldsymbol{X}\boldsymbol{\beta} \geq 0$ and $\boldsymbol{\zeta} = -\boldsymbol{\beta}$ if $\boldsymbol{\epsilon}^T \boldsymbol{X}\boldsymbol{\beta} < 0$. The last inequality directly follows from (12).

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

$$\begin{aligned} & \mathbb{P}_{\boldsymbol{\mu},\boldsymbol{w}}\left(|(\boldsymbol{\epsilon} - \boldsymbol{X}\boldsymbol{\mu})^T \boldsymbol{W}^2 \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\epsilon}^T \boldsymbol{X}\boldsymbol{\beta}| \leq C_n \|\boldsymbol{X}\boldsymbol{\beta}\| \mid \boldsymbol{X}, \boldsymbol{\epsilon}\right) \\ \leq & \mathbb{P}_{\boldsymbol{\mu},\boldsymbol{w}}\left(-\frac{\eta^*}{2}\|\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta}\|^2 + (\boldsymbol{\epsilon} - \boldsymbol{X}\boldsymbol{\mu})^T \boldsymbol{W}^2 \boldsymbol{X}\boldsymbol{\beta} + \eta^* \text{pen}(\boldsymbol{\beta} \mid \theta) \leq 0 \mid \boldsymbol{X}, \boldsymbol{\epsilon}\right) \end{aligned}$$

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

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

A.7 Proof of Lemma A.3

Starting from basic inequality $Q(\widehat{\boldsymbol{\beta}}) \geq Q(\boldsymbol{\beta}_0)$, we get

$$||\boldsymbol{W}\boldsymbol{X}\boldsymbol{\Theta}||^2 - 2(\boldsymbol{W}\boldsymbol{\epsilon} - \boldsymbol{W}\boldsymbol{X}\boldsymbol{\mu})^T \boldsymbol{W}\boldsymbol{X}\boldsymbol{\Theta} + 2\text{pen}(\boldsymbol{\beta}_0 \mid \theta) - 2\text{pen}(\widetilde{\boldsymbol{\beta}} \mid \theta) \leq 0 \tag{14}$$

From $\arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p}\{-\frac{1}{2}||\boldsymbol{W}(\boldsymbol{\epsilon} - \boldsymbol{X}\boldsymbol{\mu})/\eta^* - \boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta}||^2 + pen(\boldsymbol{\beta} \mid \theta)\} = \mathbf{0}_p$, we have

$$-2\boldsymbol{\Theta}^T \boldsymbol{X}^T \boldsymbol{W}^2(\boldsymbol{\epsilon} - \boldsymbol{X}\boldsymbol{\mu}) \geq -\eta^*||\boldsymbol{W}\boldsymbol{X}\boldsymbol{\Theta}||^2 + 2\eta^*\text{pen}(\boldsymbol{\Theta} \mid \theta) \tag{15}$$

Plug (15) into (14), using the fact that $\text{pen}(\cdot \mid \theta)$ is super-additive, we get

$$
\begin{aligned}
&(1-\eta^*)||\boldsymbol{W}\boldsymbol{X}\boldsymbol{\theta}||^2 \leq -2\eta^* pen(\boldsymbol{\Theta} \mid \theta) - 2\text{pen}(\boldsymbol{\beta} \mid \theta) + 2\text{pen}(\boldsymbol{\Theta}+\boldsymbol{\beta} \mid \theta) \\
&= -2\eta^*\text{pen}(\boldsymbol{\Theta}_S \mid \theta) - 2\eta^*\text{pen}(\boldsymbol{\Theta}_{S^C} \mid \theta) + (-2\text{pen}(\boldsymbol{\beta}_S \mid \theta) + 2\text{pen}(\boldsymbol{\Theta}_S + \boldsymbol{\beta}_S \mid \theta)) + 2\text{pen}(\boldsymbol{\Theta}_{S^C} \mid \theta) \\
&\leq -2\eta^*\text{pen}(\boldsymbol{\Theta}_S \mid \theta) - 2\eta^*\text{pen}(\boldsymbol{\Theta}_{S^C} \mid \theta) - 2\text{pen}(\boldsymbol{\Theta}_S \mid \Theta) + 2\text{pen}(\boldsymbol{\Theta}_{S^C} \mid \theta) \\
&= -2(\eta^*+1)\text{pen}(\boldsymbol{\Theta}_S \mid \theta) - 2(\eta^*-1)\text{pen}(\boldsymbol{\Theta}_{S^C} \mid \theta)
\end{aligned}
$$

Since $(1-\eta^*)||\boldsymbol{W}\boldsymbol{X}\boldsymbol{\theta}||^2 \geq 0$, 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 $Q(\widehat{\boldsymbol{\beta}}) \geq Q(\widehat{\boldsymbol{\beta}} + t\mathbf{1}_j)$ where $\mathbf{1}_j$ 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 $\text{pen}(\boldsymbol{\beta}) = \sum_j \rho(\beta_j)$, we have

$$
\begin{aligned}
t\left(\boldsymbol{W}\boldsymbol{X}_j\right)^T\left(\boldsymbol{W}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\mu}) - \boldsymbol{W}\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) &\leq \frac{t^2}{2}\|\boldsymbol{W}\boldsymbol{X}_j\|^2 + \text{pen}\left(\widehat{\boldsymbol{\beta}}\right) - \text{pen}\left(\widehat{\boldsymbol{\beta}} + t\mathbf{1}_j\right) \\
&\leq \frac{t^2 M}{2}\|\boldsymbol{X}_j\|^2 + \rho\left(t\right) = \frac{t^2 M n}{2} + M\rho\left(t\right)
\end{aligned}
$$

Thus, for any t,

$$
\boldsymbol{X}_j^T\boldsymbol{W}^2\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\mu}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) \leq M\left(\frac{1}{2}nt + \rho(t)/t\right)
$$

From the defintion of Δ in Equation (5), we have

$$
\|\boldsymbol{X}^T\boldsymbol{W}^2\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\mu}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)\|_\infty \leq M\Delta
$$

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

$$
\|\boldsymbol{X}^T\boldsymbol{W}^2\left(\boldsymbol{\epsilon}-\boldsymbol{X}\boldsymbol{\mu}\right)\|_\infty \leq \eta^* M\Delta
$$

A.9 Proof of Theorem 4.2

The proof follows from proof of Theorem 7 in Ročková and George (2018). We denote $c_+^w = 0.5(1+\sqrt{1-\frac{4||\boldsymbol{W}\boldsymbol{X}_j||^2}{(\lambda_0-\lambda_1)^2}})$ and $\delta_{c_+^w} = \frac{1}{\lambda_0-\lambda_1}\log(\frac{1-\theta}{\theta}\frac{\lambda_0}{\lambda_1}\frac{c_+^w}{1-c_+^w})$, 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 $0 \geq Q(\boldsymbol{\beta}_0) - Q(\widehat{\boldsymbol{\beta}})$,

we get

$$
\begin{aligned}
0 &\geq ||\boldsymbol{W}\boldsymbol{X}\boldsymbol{\Theta}||^2 - 2(\boldsymbol{W}\boldsymbol{\epsilon} - \boldsymbol{W}\boldsymbol{X}\boldsymbol{\mu})^T\boldsymbol{W}\boldsymbol{X}\boldsymbol{\Theta} + 2\log\frac{\pi(\boldsymbol{\beta}_0 \mid \theta)}{\pi(\tilde{\boldsymbol{\beta}} \mid \theta)} \\
&\geq ||\boldsymbol{W}\boldsymbol{X}\boldsymbol{\Theta}||^2 - 2(\boldsymbol{W}\boldsymbol{\epsilon} - \boldsymbol{W}\boldsymbol{X}\boldsymbol{\mu})^T\boldsymbol{W}\boldsymbol{X}\boldsymbol{\Theta} + 2\left[-\lambda_1|\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}| + \sum_{j=1}^{p}\log\frac{p^*_\theta(\tilde{\boldsymbol{\beta}}_j)}{p^*_\theta(0)} + \sum_{j=1}^{p}\log\frac{p^*_\theta(0)}{p^*_\theta(\boldsymbol{\beta}_{0j})}\right] \\
&\geq ||\boldsymbol{W}\boldsymbol{X}\boldsymbol{\Theta}||^2 - 2(\boldsymbol{W}\boldsymbol{\epsilon} - \boldsymbol{W}\boldsymbol{X}\boldsymbol{\mu})^T\boldsymbol{W}\boldsymbol{X}\boldsymbol{\Theta} + 2\left[-\lambda_1|\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}| + \tilde{q}b^w + (\tilde{q}-q)\log\frac{1}{p^*_\theta(0)}\right] \\
&\geq ||\boldsymbol{W}\boldsymbol{X}\boldsymbol{\Theta}||^2 - 2||(\boldsymbol{W}\boldsymbol{\epsilon} - \boldsymbol{W}\boldsymbol{X}\boldsymbol{\mu})^T\boldsymbol{W}\boldsymbol{X}||_\infty \times ||\boldsymbol{\Theta}||_1 - 2\lambda_1||\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}||_1 + 2\tilde{q}b^w + 2(\tilde{q}-q)\log\frac{1}{p^*_\theta(0)}
\end{aligned}
\tag{16}
$$

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

$$\|\boldsymbol{X}^T\boldsymbol{W}^2(\boldsymbol{\epsilon} - \boldsymbol{W}\boldsymbol{\mu})\|_\infty \leq M\eta^*\Delta$$

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

$$
\begin{aligned}
0 &\geq ||\boldsymbol{W}\boldsymbol{X}\boldsymbol{\Theta}||^2 - 2(M\eta^*\Delta + \lambda_1)|\boldsymbol{\Theta}| + 2\tilde{q}b^w + 2(\tilde{q}-q)\log\frac{1}{p^*_\theta(0)} \\
&\geq mc^2||\boldsymbol{\Theta}||^2||\boldsymbol{X}||^2 - 2(M\eta^*\Delta + \lambda_1)|\boldsymbol{\Theta}| + 2\tilde{q}b^w + 2(\tilde{q}-q)\log\frac{1}{p^*_\theta(0)} \\
&\geq mc^2||\boldsymbol{\Theta}||^2||\boldsymbol{X}||^2 - 2(M\eta^*\Delta + \lambda_1)||\boldsymbol{\Theta}||_2||\boldsymbol{\Theta}||_0^{1/2} + 2\tilde{q}b^w + 2(\tilde{q}-q)\log\frac{1}{p^*_\theta(0)}
\end{aligned}
$$

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

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

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

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

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 $\frac{1-\theta}{\theta} = C_1p^\eta, \lambda_0 = C_2p^\gamma$ with $C_1C_2 > 4$. Then $B = \log(1 + \frac{1-\theta}{\theta}\frac{\lambda_0}{\lambda_1}) >$

$\log(C_1C_2/4) + (\eta+\gamma-1)\log p$, so $\lambda_1 < 4\sqrt{n\log p} < 4\sqrt{nB/(\eta+\gamma-1)}$ and $r = \frac{A}{B} < (\frac{M\eta^*\Delta}{\sqrt{2nmBc}} + \frac{\lambda_1}{\sqrt{2mnBc}})^2 < (\frac{M\eta^*}{\sqrt{mc}}\sqrt{\frac{\eta+\gamma}{\eta+\gamma-1}} + \frac{2\sqrt{2}}{c\sqrt{m(\eta+\gamma-1)}})^2 = D$.

Notice that

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

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

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

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

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

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

We have

$$\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\beta}}\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{w}}\left(\tilde{q} \le q(1+\frac{2r}{1-r}) \,|\, \boldsymbol{Y}^{(n)}\right) = 1$$

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 $\|\boldsymbol{\Theta}\|$. Since $\log\frac{\pi(\boldsymbol{\beta}_0\,|\,\theta)}{\pi(\boldsymbol{\beta}\,|\,\theta)} \ge -\lambda_1|\boldsymbol{\Theta}| + q\log p^*_\theta(0)$, from $Q(\widehat{\boldsymbol{\beta}}) \ge Q(\boldsymbol{\beta}_0)$ and Lemma A.4, we know if (i) $\arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p}\{-\frac{1}{2}||\boldsymbol{W}(\boldsymbol{\epsilon}-\boldsymbol{X}\boldsymbol{\mu})/\eta^* - \boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta}||^2 + pen(\boldsymbol{\beta}\,|\,\theta)\} = \mathbf{0}_p$; (ii) $\max w_i \le M$, then

$$\begin{aligned} 0 &\ge ||\boldsymbol{W}\boldsymbol{X}\boldsymbol{\Theta}||^2 - 2(\boldsymbol{W}\boldsymbol{\epsilon} - \boldsymbol{W}\boldsymbol{X}\boldsymbol{\mu})^T\boldsymbol{W}\boldsymbol{X}\boldsymbol{\Theta} + 2\log\frac{\pi(\boldsymbol{\beta}_0\,|\,\theta)}{\pi(\widehat{\boldsymbol{\beta}}\,|\,\theta)} \\ &\ge ||\boldsymbol{W}\boldsymbol{X}\boldsymbol{\Theta}||^2 - 2(M\eta^*\Delta + \lambda_1)|\boldsymbol{\Theta}| + 2q\log p^*_\theta(0) \end{aligned} \tag{17}$$

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

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

Plug into (17), we know (i), (ii), plus (iii) $\min w_i \geq m$ and (iv) $\|\boldsymbol{\epsilon}\|_\infty \lesssim \sqrt{\log n}$ implies the following:

$$\begin{aligned}0 &\geq ||\boldsymbol{W}\boldsymbol{X}\boldsymbol{\Theta}||^2 - \frac{m||\boldsymbol{X}\boldsymbol{\Theta}||^2}{2} - \frac{5(K+1)q(M\eta^*\Delta+\lambda_1)^2}{m||\boldsymbol{X}||^2\phi^2} + (M\eta^*\Delta+\lambda_1)|\boldsymbol{\Theta}| + 2q\log(p^*_\theta) \\ &\geq \frac{m}{2}||\boldsymbol{X}\boldsymbol{\Theta}||^2 - \frac{5(K+1)q(M\eta^*\Delta+\lambda_1)^2}{m||\boldsymbol{X}||^2\phi^2} + (M\eta^*\Delta+\lambda_1)|\boldsymbol{\Theta}| + 2q\log p^*_\theta(0)\end{aligned}$$

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

$$\begin{aligned}\frac{m}{2}||\boldsymbol{X}\boldsymbol{\Theta}||^2 + (M\eta^*\Delta+\lambda_1)|\boldsymbol{\Theta}| &\leq \frac{5(K+1)q(C_3M\eta^*\sqrt{n\log p})^2}{mn\phi^2} + 2qC_4\log p \\ &< \frac{C_5^2M^2(\eta^*)^2}{m\phi^2}q(1+K)\log p\end{aligned}$$

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

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

It follows from definition of c that

$$\lim_{n\to\infty}\mathbb{P}_{\boldsymbol{w},\boldsymbol{\mu},\boldsymbol{\beta}_0}\left(||\boldsymbol{\Theta}|| \leq \frac{C_5\eta^*}{\sqrt{m}\,\phi\,c}\sqrt{q(1+K)\frac{\log p}{n}}\right) \geq \lim_{n\to\infty}\mathbb{P}\left(\text{condition (i)-(iv) holds}\right) = 1$$

Notice that the difference between $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$ only depends on $\boldsymbol{\mu}$ and satisfies

$$\mathbb{P}_{\boldsymbol{\mu}}\left(||\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}||^2 > u \,|\, \boldsymbol{Y}^{(n)}\right) \leq \frac{\mathbb{E}\sum_{j=1}^p(\hat{\beta}_j-\tilde{\beta}_j)^2}{u} = \frac{1}{u}\sum_{j=1}^p(\frac{1}{\lambda_0^2}+\frac{2}{\lambda_0^2}) = \frac{1}{u}\frac{3p}{\lambda_0^2}$$

Set $u = \frac{C_5^2(\eta^*)^2M^2}{m\,\phi^2\,c^2}q(1+K)\frac{\log p}{n}$, then $\frac{1}{u}\frac{3p}{\lambda_0^2} \to 0$ when $n,p\to\infty$, thus, from the tower law and triangle inequality,

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\beta}_0}\mathbb{P}_{\boldsymbol{w},\boldsymbol{\mu}}\left(||\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0|| > \frac{C_5\eta^*M}{\sqrt{m}\,\phi\,c}\sqrt{q(1+K)\frac{\log p}{n}} \,|\, \boldsymbol{Y}^{(n)}\right) \\ &\leq \mathbb{E}_{\boldsymbol{\beta}_0}\mathbb{P}_{\boldsymbol{w},\boldsymbol{\mu}}\left(||\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}||^2 > \frac{C_5\eta^*M}{2\sqrt{m}\,\phi\,c}\sqrt{q(1+K)\frac{\log p}{n}} \,|\, \boldsymbol{Y}^{(n)}\right) + \mathbb{E}_{\boldsymbol{\beta}_0}\mathbb{P}_{\boldsymbol{w},\boldsymbol{\mu}}\left(||\boldsymbol{\Theta}|| > \frac{C_5\eta^*M}{2\sqrt{m}\,\phi\,c}\sqrt{q(1+K)\frac{\log p}{n}} \,|\, \boldsymbol{Y}^{(n)}\right) \\ &= \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{\mu}}\mathbb{P}_{\boldsymbol{w}}\left(||\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}||^2 > \frac{C_5\eta^*M}{2\sqrt{m}\,\phi\,c}\sqrt{q(1+K)\frac{\log p}{n}} \,|\, \boldsymbol{Y}^{(n)},\boldsymbol{\mu}\right) + \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{\mu}}\mathbb{P}_{\boldsymbol{w}}\left(||\boldsymbol{\Theta}|| > \frac{C_5\eta^*M}{2\sqrt{m}\,\phi\,c}\sqrt{q(1+K)\frac{\log p}{n}} \,|\, \boldsymbol{Y}^{(n)},\boldsymbol{\mu}\right) \\ &= \mathbb{E}_{\boldsymbol{\beta}_0,\boldsymbol{\mu}}\mathbb{I}\left(||\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}||^2 > \frac{C_5\eta^*M}{2\sqrt{m}\,\phi\,c}\sqrt{q(1+K)\frac{\log p}{n}} \,|\, \boldsymbol{Y}^{(n)},\boldsymbol{\mu}\right) + \mathbb{E}_{\boldsymbol{\beta}_0}\mathbb{P}_{\boldsymbol{w}}\left(||\boldsymbol{\Theta}|| > \frac{C_5\eta^*M}{2\sqrt{m}\,\phi\,c}\sqrt{q(1+K)\frac{\log p}{n}} \,|\, \boldsymbol{Y}^{(n)}\right) \\ &\leq \frac{1}{u}\frac{3p}{\lambda_0^2} + \mathbb{P}_{\boldsymbol{\beta}_0,\boldsymbol{\mu},\boldsymbol{w}}\left(||\boldsymbol{\Theta}|| > \frac{C_5\eta^*M}{2\sqrt{m}\,\phi\,c}\sqrt{q(1+K)\frac{\log p}{n}} \,|\, \boldsymbol{Y}^{(n)}\right)\end{aligned}$$

where the right hand side

$$\frac{1}{u}\frac{3p}{\lambda_0^2} + \mathbb{P}_{\boldsymbol{\beta}_0,\boldsymbol{\mu},\boldsymbol{w}}\left(||\boldsymbol{\Theta}|| > \frac{C_5\eta^* M}{2\sqrt{m}\,\phi\,c}\sqrt{q(1+K)\frac{\log p}{n}} \mid \boldsymbol{Y}^{(n)}\right) \to 0$$

Thus, we have

$$\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\beta}}\mathbb{P}_{\boldsymbol{w},\boldsymbol{\mu}}\left(||\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| > \frac{C_5\eta^* M}{\sqrt{m}\,\phi\,c}\sqrt{q(1+K)\frac{\log p}{n}} \mid \boldsymbol{Y}^{(n)}\right) = 0$$

A.11 Proof of Corollary 4.1

We notice that if $w_i \sim \frac{1}{\alpha}Gamma(\alpha,1)$, we have $\mathbb{E}w_i = 1$, $Var(w_i) = \frac{1}{\alpha} \lesssim \frac{1}{\log n}$, $Cov(w_i, w_j) = 0$. 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,

$$\begin{aligned}
&\mathbb{P}(\min w_i < \frac{1}{e}) \le n\mathbb{P}(w_i < \frac{1}{e}) = n\mathbb{P}_{v\sim Gamma(\alpha,1)}(v < \alpha/e) = ne^{-\alpha/e}\sum_{i=\alpha}^{\infty}\frac{(\alpha/e)^i}{i!} \\
&\le ne^{-\alpha/e}\sum_{i=\alpha}^{\infty}\frac{(\alpha/e)^i}{\sqrt{2\pi}i^{i+1/2}e^{-i}} = n\frac{e^{-\alpha/e}}{\sqrt{2\pi}}\sum_{i=\alpha}^{\infty}(\frac{\alpha}{i})^i \le n\frac{e^{-\alpha/e}}{\sqrt{2\pi}}[\sum_{i=\alpha+1}^{\infty}(\frac{\alpha}{i})^i + 1] \\
&\le n\frac{e^{-\alpha/e}}{\sqrt{2\pi}}[\sum_{i=\alpha+1}^{\infty}(\frac{\alpha}{\alpha+1})^i + 1] = n\frac{e^{-\alpha/e}}{\sqrt{2\pi}}[(\frac{\alpha}{\alpha+1})^{\alpha+1}(\alpha+1) + 1] \le C_0 n\frac{e^{-\alpha/e}}{\sqrt{2\pi}}[\frac{\alpha+1}{e} + 1] \\
&\le C_1(\log p)\, n\, p^{-\frac{2(\eta+\gamma)}{e}}
\end{aligned}$$

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

$$\mathbb{P}(\max w_i > t) = ne^{\alpha(log(t)+1-t)+1} \le Cn\,(p)^{-\frac{\eta+\gamma}{t}}$$

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

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

$$\mathbb{P}(\min w_i < \frac{1}{e}) \le \widetilde{C}_1(\log p)\, n\, p^{-\frac{2(\eta+\gamma)}{e}}$$

and

$$\mathbb{P}(\max w_i > t) \le \widetilde{C}n\,(p)^{-\frac{\eta+\gamma}{t}}$$

Thus, condition (1)-(4) hold for both $\boldsymbol{w} \sim nDir(\alpha,\cdots,\alpha)$ 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

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

Next, define

$$\phi_0^{(-)}(x) = \phi(x; y_i - \lambda_0/n, 1/n), \qquad \phi_0^{(+)}(x) = \phi(x; y_i + \lambda_0/n, 1/n),$$

where $\phi(x;\mu,\sigma^2)$ is the Gaussian density with mean μ and variance σ^2. The quantities $c_1^{(-)}$, $c_1^{(+)}$, $\phi_1^{(-)}(x)$, $\phi_1^{(+)}(x)$ are defined in a similar way in 3.1. Δ_j is defined as in (5). Define $c_+ = 0.5\left(1+\sqrt{1-4/(\lambda_0-\lambda_1)^2}\right)$ and $\delta_{c_+} = 1/(\lambda_0-\lambda_1)\log\left[\frac{1-\theta}{\theta}\frac{\lambda_0}{\lambda_1}\frac{c_+}{1-c_+}\right]$. 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 $||\boldsymbol{X}_j|| = \sqrt{n}$. So here we have $\Delta^L < \Delta_j = \inf_{t>0}[nt/2 - \rho(t\,|\,\theta)/t] \doteq \Delta < \Delta^U$ where $\Delta^L = \sqrt{2n\log 1/p^*(0) - d} + \lambda_1$ and $\Delta^U = \sqrt{2n\log 1/p^*(0)} + \lambda_1$, $d = -[\lambda^*(\delta_{c_+}) - \lambda_1]^2 - 2n\log p^*(\delta_{c_+})$. Throughout this section we assume that the parameter λ_0, λ_1 satisfies: $(1-\theta)/\theta \asymp n^a$, $\lambda_0 \asymp n^d$ where $a, d \geq 2$ and $1/\sqrt{n} < \lambda_1 \leq c_0$, c_0 is a constant.

A.12.2 Active Coordinates

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

Proof. Without loss of generality, we assume $y_i > \beta_i/2 > 0$.

$$\begin{aligned} w_1 = \pi(\gamma_i = 1\,|\,y_i) &= \frac{\pi(y_i\,|\,\gamma_i=1)\pi(\gamma_i=1)}{\pi(y_i\,|\,\gamma_i=1)\pi(\gamma_i=1)+\pi(y_i\,|\,\gamma_i=0)\pi(\gamma_i=0)} \\ &= \frac{\int_{\beta_i}\pi(y_i\,|\,\beta_i)\pi(\beta_i\,|\,\gamma_i=1)\pi(\gamma_i=1)d\beta_i}{\int_{\beta_i}\pi(y_i\,|\,\beta_i)\pi(\beta_i\,|\,\gamma_i=1)\pi(\gamma_i=1)d\beta_i + \int_{\beta_i}\pi(y_i\,|\,\beta_i)\pi(\beta_i\,|\,\gamma_i=0)\pi(\gamma_i=0)d\beta_i} \\ &= \frac{\int_0^\infty c_1^{(-)}\phi_1^{(-)}(\beta_i)d\beta_i + \int_{-\infty}^0 c_1^{(+)}\phi_1^{(+)}(\beta_i)d\beta_i}{\int_0^\infty c_1^{(-)}\phi_1^{(-)}(\beta_i)d\beta_i + \int_0^\infty c_0^{(-)}\phi_0^{(-)}(\beta_i)d\beta_i + \int_{-\infty}^0 c_1^{(+)}\phi_1^{(+)}(\beta_i)d\beta_i + \int_{-\infty}^0 c_0^{(+)}\phi_0^{(+)}(\beta_i)d\beta_i} \end{aligned} \tag{18}$$

We consider the four terms in denominator separately. It is helpful to divide each of them by $\theta\lambda_1$. For the first term in denominator:

$$\frac{1}{\theta\lambda_1}\int_0^\infty c_1^{(-)}\phi_1^{(-)}(\beta_i)d\beta_i = e^{-y_i\lambda_1+\lambda_1^2/2n}\left(1-\Phi\left(-\sqrt{n}(y_i-\frac{\lambda_1}{n})\right)\right) \to e^{-y_i\lambda_1} \tag{19}$$

For the second term in denominator, from Mills ratio, we have:

$$\frac{1}{\theta\lambda_1}\int_0^\infty c_0^{(-)}\phi_0^{(-)}(\beta_i)d\beta_i = \frac{1}{\theta\lambda_1}c_0^{(-)}\Phi(-\sqrt{n}(y_i-\frac{\lambda_0}{n}))$$
$$\leq C\frac{(1-\theta)\lambda_0}{\theta\lambda_1}e^{-y_i\lambda_0+\lambda_0^2/2n}\frac{\phi\left(\sqrt{n}(y_i-\frac{\lambda_0}{n}\right)}{\sqrt{n}(y_i-\frac{\lambda_0}{n})} = C\frac{(1-\theta)\lambda_0}{\sqrt{n}(y_i-\frac{\lambda_0}{n})\theta\lambda_1}\frac{1}{\sqrt{2\pi}}e^{-\frac{n}{2}y_i^2}$$

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

$$\frac{1}{\theta\lambda_1}\int_0^\infty c_0^{(-)}\phi_0^{(-)}(\beta_i)d\beta_i \to 0 \tag{20}$$

For the third term in denominator:

$$\frac{1}{\theta\lambda_1}\int_{-\infty}^0 c_1^{(+)}\phi_1^{(+)}(\beta_i)d\beta_i = e^{y_i\lambda_1+\lambda_1^2/2n}\Phi\left(-\sqrt{n}(y_i+\frac{\lambda_1}{n})\right) \to 0 \tag{21}$$

For the fourth term in denominator:

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

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

$$\frac{1}{\theta\lambda_1}\int_{-\infty}^0 c_0^{(+)}\phi_0^{(+)}(\beta_i)d\beta_i \to 0 \tag{22}$$

From (19), (20), (21) and (22), we know that $w_0 \to \frac{e^{-y_i\lambda_1}}{e^{-y_i\lambda_1}} = 1$. Thus, $w_1 = 1-w_0 \to 0$. □

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

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

$$\pi\left(u_i\,|\,y_i,\gamma_i=1\right) = \frac{1}{\sqrt{n}}\frac{\mathbb{I}(u_i\geq-\sqrt{n}y_i)c_1^{(-)}\phi_1^{(-)}(u_i/\sqrt{n}+y_i)+\mathbb{I}(u_i<-\sqrt{n}y_i)c_1^{(+)}\phi_1^{(+)}(u_i/\sqrt{n}+y_i)}{\int_0^\infty c_1^{(-)}\phi_1^{(-)}(\beta_i)d\beta_i+\int_{-\infty}^0 c_1^{(+)}\phi_1^{(+)}(\beta_i)d\beta_i} \tag{23}$$

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 $\pi\left(u_i\,|\,y_i,\gamma_i=1\right) \to \phi(u_i;0,1)$. □

Proposition 3. *For fixed WBB estimate $\widehat{\beta}_i$, conditioning on the event that $|y_i| > \frac{|\beta_i^0|}{2} > 0$, when $\widehat{\beta}_i \neq 0$, $n\,(\widehat{\beta}_i - y_i) \mid y_i \to -\frac{1}{w_i}\lambda_1$.*

Proof. The objective function for each fixed WBB sample is

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

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

$$\begin{aligned}
&(n\,(\widehat{\beta}_i - y_i) \mid Y_i = y_i, w_i) = -\frac{1}{w_i}\lambda_1 - \frac{1}{w_i}(\lambda_0 - \lambda_1)(1 - p^*(\widehat{\beta}_i)) \\
&= -\frac{1}{w_i}\lambda_1 - \frac{1}{w_i}\frac{\frac{(1-\theta)\lambda_0}{\theta\lambda_1}(\lambda_0 - \lambda_1)e^{-|\,\widehat{\beta}_i\,|\,(\lambda_0 - \lambda_1)}}{1 + \frac{(1-\theta)\lambda_0}{\theta\lambda_1}e^{-|\,\widehat{\beta}_i\,|\,(\lambda_0 - \lambda_1)}} \to -\frac{1}{w_i}\lambda_1
\end{aligned}$$

□

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

Proof. From Markov Inequality,

$$\begin{aligned}
\mathbb{P}_{w_i}(\widehat{\beta}_i = 0 \mid y_i) &\leq \mathbb{P}\left((\widehat{\beta}_i - y_i)^2 > y_i^2\right) \leq \frac{\mathbb{E}\,(\widehat{\beta}_i - y_i)^2}{y_i^2} \\
&= \frac{1}{y_i^2}\left\{ \mathbb{E}\left[y_i^2 \mathbf{1}\left(|\sqrt{w_i} y_i| \leq \frac{\Delta}{n} \right) \right] + \mathbb{E}\left[\left(\frac{1}{w_i n}\lambda^*(\widehat{\beta}_i) \right)^2 \mathbb{I}\left(|\sqrt{w_i} y_i| > \frac{\Delta}{n} \right) \right] \right\} \\
&\leq \mathbb{P}(w_i \leq \frac{\Delta^2}{n^2 y_i^2}) + \frac{1}{y_i^2}\mathbb{E}\left[\left(\frac{c_+(\lambda_1 - \lambda_0) + \lambda_0}{\sqrt{n w_i}} \right)^2 \mathbb{I}\left(\sqrt{w_i} > \frac{\Delta}{n y_i} \right) \right] \\
&\leq \mathbb{P}(\frac{1}{w_i} \geq \frac{n^2 y_i^2}{\Delta^2}) + \frac{1}{y_i^2}\frac{4}{n}\mathbb{E}\frac{1}{w_i}\mathbb{I}\left(\sqrt{w_i} > \frac{\Delta}{n y_i} \right) \\
&\leq \frac{\mathbb{E}\frac{1}{w_i}}{\frac{n^2 y_i^2}{\Delta^2}} + \frac{1}{y_i^2}\frac{4}{n}\mathbb{E}\frac{1}{w_i}
\end{aligned} \tag{24}$$

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

$$\lambda^*(\widehat{\beta}_i) = \sqrt{n w_i}\lambda^*(\sqrt{n w_i}\widehat{\beta}_i) < \sqrt{n w_i}(c_+(\lambda_1 - \lambda_0) + \lambda_0) \text{ when } |\sqrt{w_i} y_i| > \frac{\Delta}{n}$$

and the fourth row of (24) follows directly from the definition of c_+, Δ and the conditions on λ_0, λ_1 as described at the beginning of this section. The right hand size of (24) satisfies

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

Thus $\mathbb{P}_{w_i}(\widehat{\beta}_i = 0 \mid y_i) \to 0$ holds. □

A.12.3 Inactive coordinates

Proposition 5. *For w_0 and w_1 in (9), when conditioning on $y_i \asymp \frac{1}{\sqrt{n}}$, 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 $\theta\lambda_1$. For the first term:

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

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

$$\begin{aligned}\frac{1}{\theta\lambda_1}\int_0^\infty c_0^{(-)}\phi_0^{(-)}(\beta_i)d\beta_i &= \frac{1}{\theta\lambda_1}\int_0^\infty (1-\theta)\lambda_0\sqrt{\frac{n}{2\pi}}e^{-\frac{n}{2}(\beta_i-y_i)^2-\beta_i\lambda_0}d\beta_i \\ &\geq \frac{1}{\theta\lambda_1}\int_0^{\epsilon y_i} (1-\theta)\lambda_0\sqrt{\frac{n}{2\pi}}e^{-\frac{n(\epsilon-1)^2}{2}y_i^2-\beta_i\lambda_0}d\beta_i \\ &= \frac{1}{\theta\lambda_1}(1-\theta)\sqrt{\frac{n}{2\pi}}e^{-\frac{n(\epsilon-1)^2}{2}y_i^2}\left(1-e^{-\epsilon y_i\lambda_0}\right)\end{aligned}$$

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

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

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,

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

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

$$\begin{aligned}\frac{1}{\theta\lambda_1}\int_{-\infty}^0 c_0^{(+)}\phi_0^{(+)}(\beta_i)d\beta_i &= \frac{1}{\theta\lambda_1}\int_{-\infty}^0 (1-\theta)\lambda_0\sqrt{\frac{n}{2\pi}}e^{-\frac{n}{2}(\beta_i-y_i)^2+\beta_i\lambda_0}d\beta_i \\ &\geq \frac{1}{\theta\lambda_1}\int_{-\epsilon y_i}^0 (1-\theta)\lambda_0\sqrt{\frac{n}{2\pi}}e^{-\frac{n(1+\epsilon)^2}{2}y_i^2+\beta_i\lambda_0}d\beta_i \\ &= \frac{1}{\theta\lambda_1}(1-\theta)\sqrt{\frac{n}{2\pi}}e^{-\frac{n(1+\epsilon)^2}{2}y_i^2}\left(1-e^{-\epsilon y_i\lambda_0}\right)\end{aligned}$$

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

$$\frac{1}{\theta\lambda_1}(1-\theta)\sqrt{\frac{n}{2\pi}}e^{-\frac{n(1+\epsilon)^2}{2}y_i^2}\left(1-e^{-\epsilon y_i\lambda_0}\right) \to \infty$$

we know

$$\frac{1}{\theta\lambda_1}\int_{-\infty}^{0} c_0^{(+)}\phi_0^{(+)}(\beta_i)d\beta_i \to \infty \tag{28}$$

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 $|y_i| \asymp O\left(\frac{1}{\sqrt{n}}\right)$, we have $\pi\left(\lambda_0\beta_i \,|\, y_i, \gamma_i = 0\right) \to \frac{1}{2}e^{-|\lambda_0\beta_i|}$.*

Proof. Notice that

$$\pi\left(\beta_i \,|\, y_i, \gamma_i = 0\right) \propto \pi\left(y_i \,|\, \beta_i, \gamma_i = 0\right)\pi\left(\beta_i \,|\, \gamma_i = 0\right) \propto e^{-\frac{n}{2}(\beta_i - y_i)^2}e^{-|\beta_i|\lambda_0}$$

Thus, let $\beta_i' = \lambda_0\beta_i$, since $\lambda_0 \to \infty$, for any fixed β_i', we have

$$\pi\left(\beta_i' \,|\, y_i, \gamma_i = 0\right) \propto e^{-\frac{n}{2}(\beta_i'/\lambda_0 - y_i)^2}e^{-|\beta_i'|} \to e^{-\frac{n}{2}(y_i)^2}e^{-|\beta_i'|}$$

which implies that

$$\pi\left(\beta_i' \,|\, y_i, \gamma_i = 0\right) \to e^{-|\beta_i'|}$$

From Scheffé (1947), we know that β_i' converges in total variation to $e^{-|\beta_i'|}$. □

Proposition 7. *For fixed WBB estimate $\widehat{\beta}_i$, conditioning on $|y_i| \asymp O\left(\frac{1}{\sqrt{n}}\right)$, if $\exists\, t > 0$ s.t. $Ee^{tw_i} \le C$, $\forall i$, we have $\mathbb{P}_{w_i}\left(\widehat{\beta}_i = 0 \,|\, y_i\right) \to 1$.*

Proof. The definition of the fixed WBB sample is in $\widehat{\beta}_i$ (12). From Chernoff bound,

$$\mathbb{P}_{w_i}\left(\sqrt{w_i} > \frac{\Delta}{n|y_i|} \,|\, y_i\right) \le \frac{\mathbb{E}e^{tw_i}}{e^{t\Delta^2/(n^2y_i^2)}} \le \frac{C}{e^{t\Delta_L^2/(n^2y_i^2)}} = \frac{C}{e^{t(2\log 1/p^*(0) - d/n)/(ny_i^2)}} = \frac{\widetilde{C}}{n^{2(\eta+\gamma)t/(ny_i^2)}}$$

where $\frac{\widetilde{C}}{n^{2(\eta+\gamma)t/(ny_i^2)}} \to 0$. Thus, $\mathbb{P}_{w_i}\left(\widehat{\beta}_i = 0 \,|\, y_i\right) \to 1$. □

Remark A.1. *Random WBB is equivalent to fixed WBB by setting weights to be $\boldsymbol{w}/w_p$ 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. $\mathbb{E}_{w_i,w_p}e^{tw_i/w_p} \le C$, $\forall i$, we have $\mathbb{P}_{w_i}\left(\beta_i^{random} = 0 \,|\, y_i\right) \to 1$.*

B Details of Connection to NPL in Section 4.2

We define the loss function $l(\cdot)$ as

$$l(\boldsymbol{x}_i, y_i, \boldsymbol{\beta}) = -\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \frac{1}{n}\log[\int_{\theta}\prod_{j=1}^{p}\pi(\tilde{\beta}_j \,|\, \theta)d\pi(\theta)] \tag{29}$$

Motivation for the Prior For paired data $(\boldsymbol{x}_i, y_i)$, Fong et al. (2019) uses independent prior which assumes that y_i does not depend on $\boldsymbol{x}_i$:

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

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

One possible solution is to use $y = \boldsymbol{x}^T\widehat{\boldsymbol{\beta}} + \epsilon$ where $\widehat{\boldsymbol{\beta}}$ is the MAP of $\boldsymbol{\beta}$ 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 $\boldsymbol{\mu}$ on $\widehat{\boldsymbol{\beta}}$. We want $\boldsymbol{\mu}$ 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_π becomes

$$\widetilde{\boldsymbol{x}}_k \sim \hat{F}_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\delta(\boldsymbol{x}_i), \quad \widetilde{y}_k \,|\, \widetilde{\boldsymbol{x}}_k = \boldsymbol{x}_k^T(\widehat{\boldsymbol{\beta}} + \boldsymbol{\mu}) + \epsilon$$

where $\boldsymbol{\mu} \sim$ Spike and $\epsilon \sim N(0, \sigma^2)$. If $\widehat{\boldsymbol{\beta}}$ is close enough to truth, we have $\widetilde{\boldsymbol{x}}_k^T\widehat{\boldsymbol{\beta}} \approx \widetilde{\boldsymbol{x}}_k^T\boldsymbol{\beta}_0$. Since $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta}_0 + \epsilon_i$ and $\epsilon_i \overset{d}{=} \epsilon$, we can set $\widetilde{y}_k \,|\, \widetilde{\boldsymbol{x}}_k = y_i + \boldsymbol{x}_i^T\boldsymbol{\mu}$ where i satisfies $\widetilde{\boldsymbol{x}}_k = \boldsymbol{x}_i$. Then the above prior becomes

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

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

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}}^t &= \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{-\frac{1}{2}\sum_{i=1}^{n} w_i(Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 - \frac{1}{2}\sum_{i=1}^{n}\tilde{w}_i(Y_i + \boldsymbol{x}_i'\boldsymbol{\mu} - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \frac{1}{n}\log[\int_\theta\prod_{j=1}^{p}\pi(\beta_j \,|\, \theta)d\pi(\theta)]\right\} \\
&= \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{-\frac{1}{2}\sum_{i=1}^{n}(w_i + \tilde{w}_i)(Y_i + \frac{\tilde{w}_i}{w_i + \tilde{w}_i}\boldsymbol{x}_i^T\boldsymbol{\mu} - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \frac{1}{n}\log[\int_\theta\prod_{j=1}^{p}\pi(\beta_j \,|\, \theta)d\pi(\theta)]\right\} \\
&\approx \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{-\frac{1}{2}\sum_{i=1}^{n}(w_i + \tilde{w}_i)(Y_i + \frac{c}{c+n}\boldsymbol{x}_i^T\boldsymbol{\mu} - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \frac{1}{n}\log[\int_\theta\prod_{j=1}^{p}\pi(\beta_j \,|\, \theta)d\pi(\theta)]\right\} \\
&\overset{\tilde{\boldsymbol{\beta}}=\boldsymbol{\beta}-c/(c+n)\boldsymbol{\mu}}{=} \arg\max_{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^p}\left\{-\frac{1}{2}\sum_{i=1}^{n} w_i^*(Y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})^2 + \log[\int_\theta\prod_{j=1}^{p}\pi(\tilde{\beta}_j + \frac{c}{c+n}\mu_j \,|\, \theta)d\pi(\theta)]\right\} + \frac{c}{c+n}\boldsymbol{\mu}
\end{aligned}$$

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

$$\tilde{\boldsymbol{\beta}}^t \overset{\text{D}}{=} \arg\max_{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^p}\left\{-\frac{1}{2}\sum_{i=1}^{n} w_i^*(Y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})^2 + \log[\int_\theta\prod_{j=1}^{p}\pi(\tilde{\beta}_j - \frac{c}{c+n}\mu_j^* \,|\, \theta)d\pi(\theta)]\right\} - \frac{c}{c+n}\boldsymbol{\mu}^*$$

Setting	$\mathbf{p=1000, \rho=0.6}$	$\mathbf{p=1000, \rho=0.9}$	$\mathbf{p=2000, \rho=0.6}$	$\mathbf{p=2000, \rho=0.9}$	$\mathbf{p=5000, \rho=0.6}$	$\mathbf{p=5000, \rho=0.9}$
SSVS1	33.77±0.34	33.67±0.19	194.45±0.38	193.95±0.18	2526.67±0.72	2526.87±1.16
SSVS2	2.50±0.04	2.50±0.03	4.99±0.06	4.98±0.04	12.77±0.09	12.76±0.10
Skinny Gibbs	6.59±0.76	6.57±0.61	15.82±1.37	16.02±0.98	57.17±3.01	56.50±2.45
BB-SSL (single λ_0)	**0.36±0.04**	**0.37±0.04**	**0.67±0.07**	**0.77±0.10**	**1.79±0.32**	**1.80±0.31**
BB-SSL (sequence of λ_0's)	20.67±0.26	21.22±0.37	47.69±0.60	49.74±0.37	171.49±2.72	173.36±3.61

(a) Fixed $n = 100$ and increasing p.

Setting	$\mathbf{n=100}$	$\mathbf{n=500}$	$\mathbf{n=1000}$	$\mathbf{n=2000}$	$\mathbf{n=5000}$
SSVS	0.47±0.03	0.50 ± 0.01	0.66 ± 0.01	1.19 ± 0.09	**2.11 ± 0.02**
Bhattacharya	0.39 ± 0.02	1.37 ± 0.01	5.18 ± 0.04	37.47 ± 1.96	388.491 ± 4.59
Skinny Gibbs	0.90 ± 0.04	1.05 ± 0.14	1.22 ± 0.11	1.99 ± 0.44	2.32 ± 0.10
WLB	0.20 ± 0.002	1.06 ± 0.01	3.31 ± 0.004	16.22 ± 0.86	92.00 ± 0.18
BB-SSL (single λ_0)	**0.07 ± 0.005**	**0.17 ± 0.02**	**0.29 ± 0.03**	**0.70 ± 0.19**	2.81 ± 0.93
BB-SSL (sequence of λ_0's)	1.61 ± 0.03	5.34 ± 0.21	6.73 ± 0.33	8.72 ± 0.41	15.67 ± 0.78

(b) Fixed $p = 100, \rho = 0.6$ and increasing n.

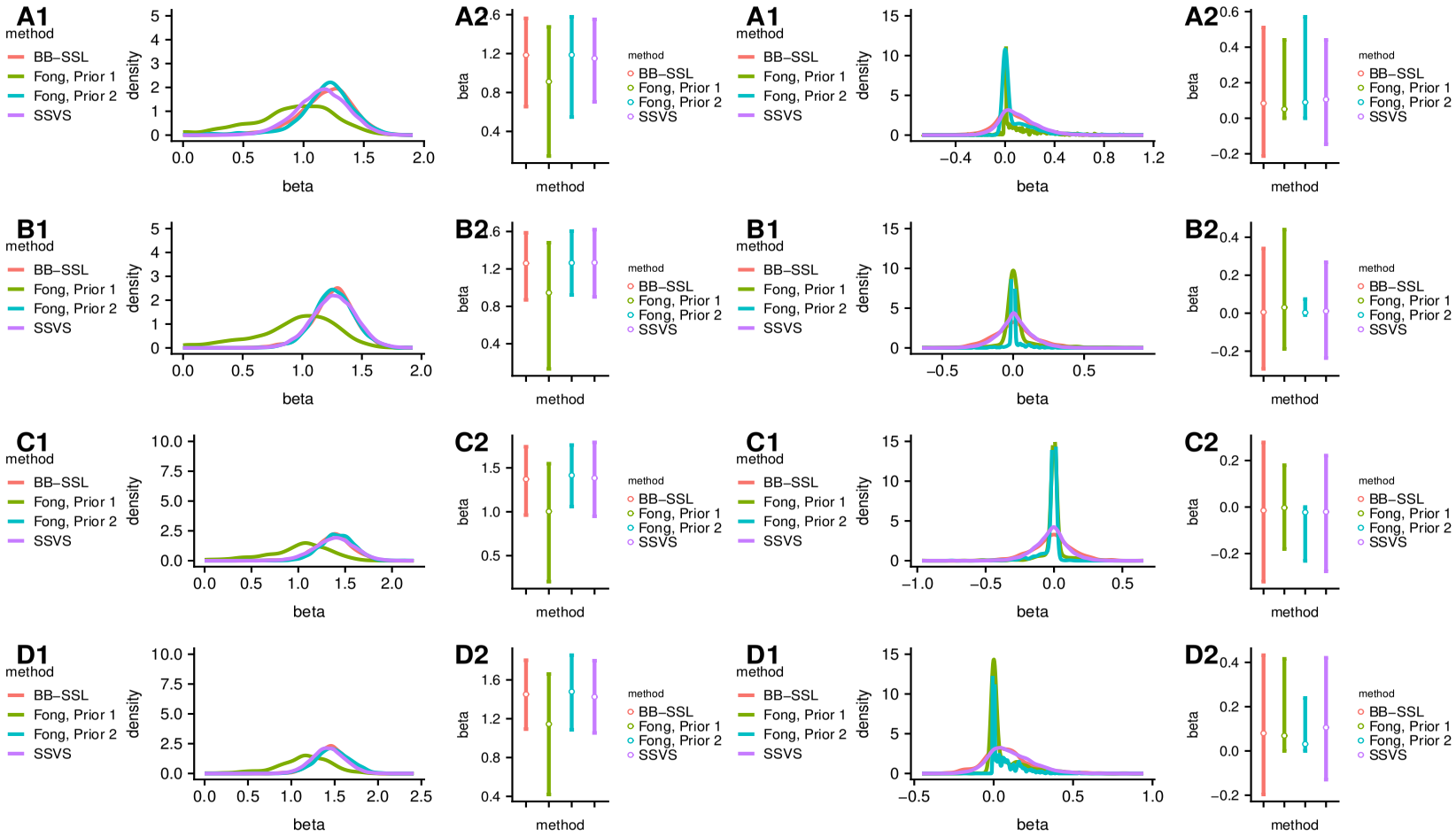
Table 1: Running time (in seconds, based on average of 10 independent runs) of each algorithm (per 100 iterations). The number after '±' is the standard deviation of 10 independent runs. BB-SSL are either fitted using a sequence of λ_0's: an equal difference series of length $\lceil \lambda_0 \rceil$ 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.

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

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 $\boldsymbol{\beta}$ 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 $\boldsymbol{X}_j$ – this takes $O(n)$; another way is to pre-compute the Gram matrix and for each coordinate, we calculate the inner product between $\boldsymbol{X}^T\boldsymbol{X}_j$ and $\hat{\boldsymbol{\beta}}$ – this takes $O(p)$. For both ways, we update θ every c iterations where each update is $O(p)$. So for a single value of λ_0, SSLASSO is $O\left(\min\left(\text{maxiter} \times p(n + \frac{p}{c_1}), (n + \text{maxiter}) \times p^2\right)\right)$. For a sequence of λ_0's, complexity is $O\left(L \times \min\left(\text{maxiter} \times p(n + \frac{p}{c_1}), (n + \text{maxiter}) \times p^2\right)\right)$ where L is the length of λ_0's. Usually if the biggest λ_0 is large enough, we would expect that the larger λ_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 $n = 50, p = 12$, $\boldsymbol{\beta} = (1.3, 0, 0, 1.3, 0, 0, 1.3, 0, 0, 1.3, 0, 0)^T$, predictors are grouped into 4 blockes with correlation coefficient $\rho = 0.6$, $\alpha = 2$, $m = 50$, $c = 10$, $\lambda_0 = 7, \lambda_1 = 0.15$.

convergence quickly using the estimated β from previous λ_0's, so usually it takes less than $O\left(L \times \min\left(\text{maxiter} \times p(n + \frac{p}{c_1}),\ (n + \text{maxiter}) \times p^2\right)\right)$.

Empirically, when λ_0 is small and does not provide enough shrinkage, SSL takes more iterations to converge. When choosing λ_0, we suggest plotting the regularization path from SSL and choose a large enough λ_0 after which further increasing λ_0 does not affect the chosen model. This is also reasonable from the goal of variable selection. When λ_0 is extremely small, a single value fitted SSL might have difficulty converging so we have to use a sequence of λ_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 λ_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 $(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{D}_\tau^{-1})^{-1} = \boldsymbol{D}_\tau - \boldsymbol{D}_\tau\boldsymbol{X}^T(\boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{D}_\tau\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{D}_\tau$, whose complexity is $O(p^2n)$.

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

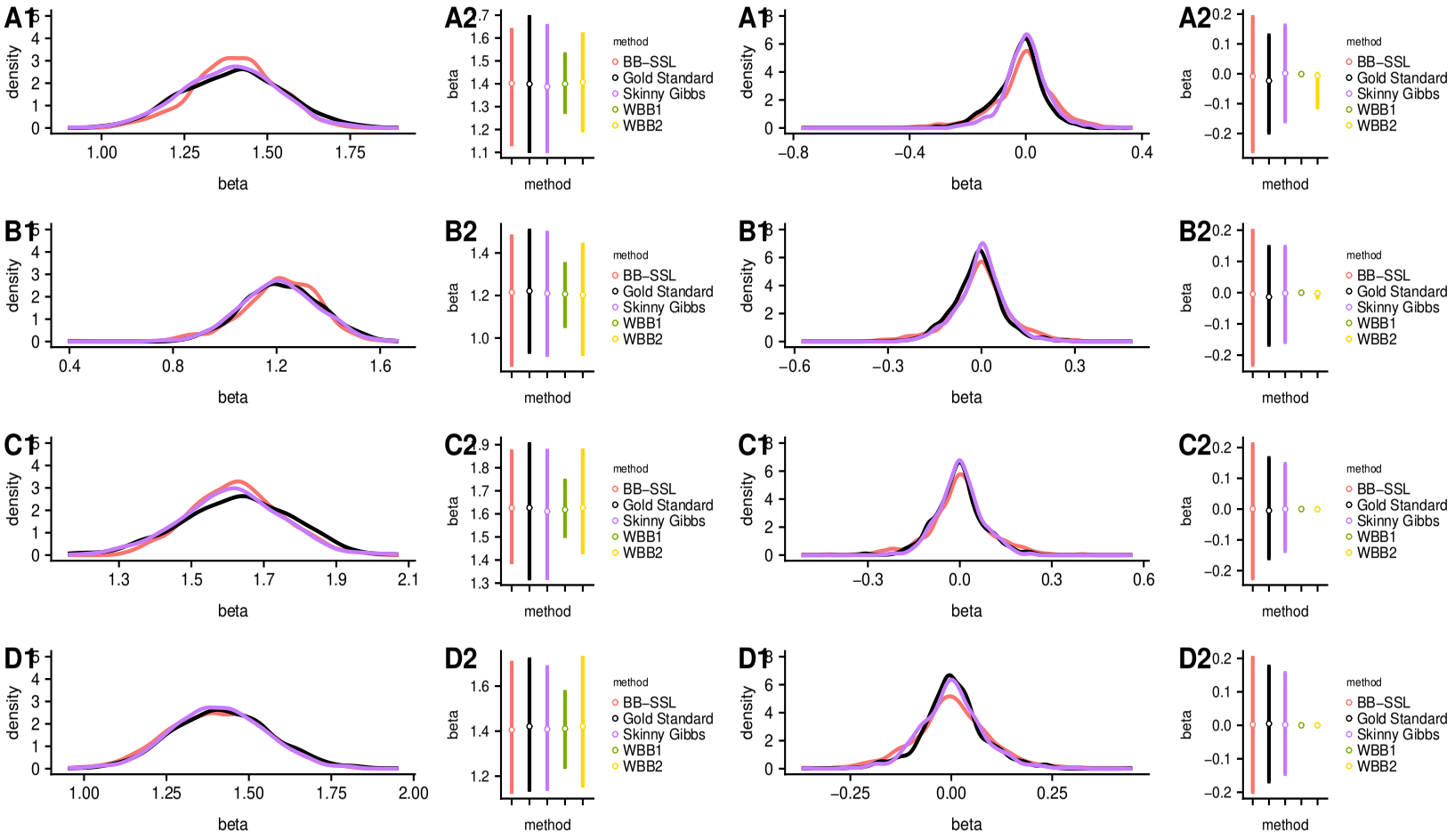
Skinny Gibbs We modified Skinny Gibbs to sample from the posterior using SSL prior. Theoretically it is of complexity $O(np)$. However, sometimes when n is relatively small, we observe that the running time of Skinny Gibbs is slower than Bhattacharya's method. This is because Skinny Gibbs involves an $O(n)$ matrix product for each coordinate and the update for each coordinate is implemented via for-loop, whereas in Bhattacharya's method the $O(n^2p)$ 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 $O(p^2n)$ 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 β_i^0, as shown in figure 2. We find that all methods perform well for active β_i's. WBB1 and WBB2 are doing poorly for inactive β_i's. For the marginal mean of γ_i's, as shown in figure 3, all methods perform pretty well. All methods can detect over 95% of models.



(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 2: Density and credible interval of β_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 λ_0. Since WBB1 and WBB2 produce a point mass and do not fit into the y-axis, we exclude it in the density plot.

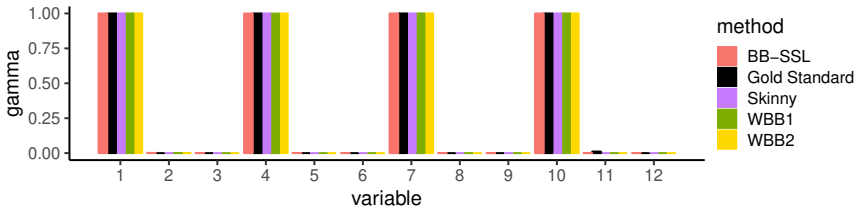


Figure 3: Mean of $\gamma_i, i = 1, 2, \cdots, 12$ in low-dimensional, independent case. $n = 50, p = 12, \beta_{active} = (1.3, 1.3, 1.3, 1.3)$ and predictors mutually independent. Each method is (thinned to) 1,000 sample points. $\lambda_0 = 13, \lambda_1 = 0.05$. SSLASSO is fitted using a single λ_0.

Metric			Skinny Gibbs	WBB1	WBB2	BB-SSL
β_i's	KL divergence	active	3.02	3.66	1.80	**0.67**
		inactive	0.04	3.09	3.09	**0.01**
	MSE	active	0.62	1.15	0.44	**0.20**
		inactive	0.009	**0.004**	0.005	0.007
γ_i's	MSE	active	2.36	0.25	0.12	**0.10**
		inactive	0.006	**0.001**	**0.001**	**0.001**
Selected Model	Hamming distance	all	5	3	2	3

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 β_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.

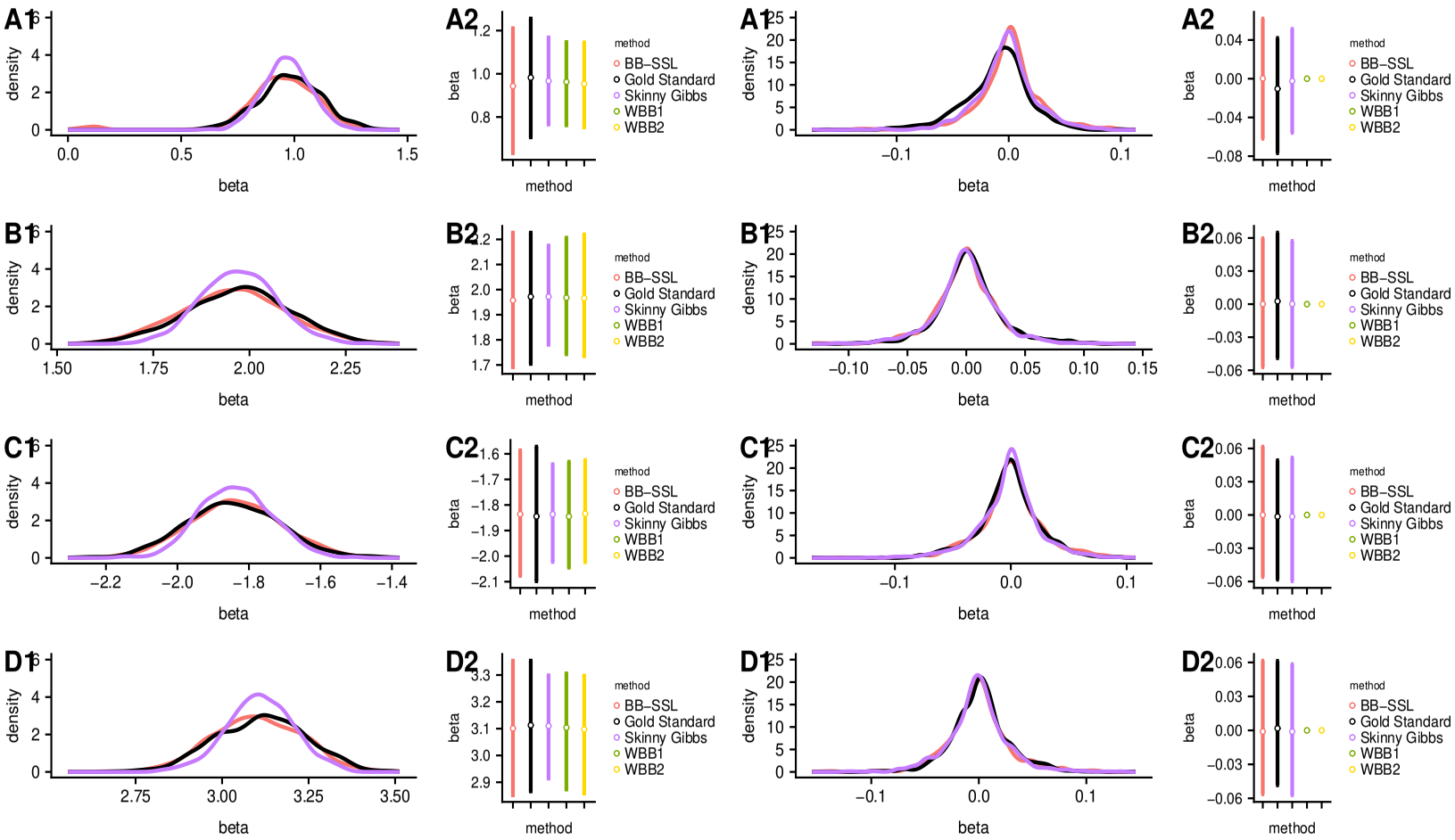
D.3 High Dimensional, $\rho = 0.99$

Here we have $n = 100, p = 1\,000, \rho = 0.99$. 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 $\boldsymbol{\beta}$ 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 β_2 and β_{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 α On The Posterior

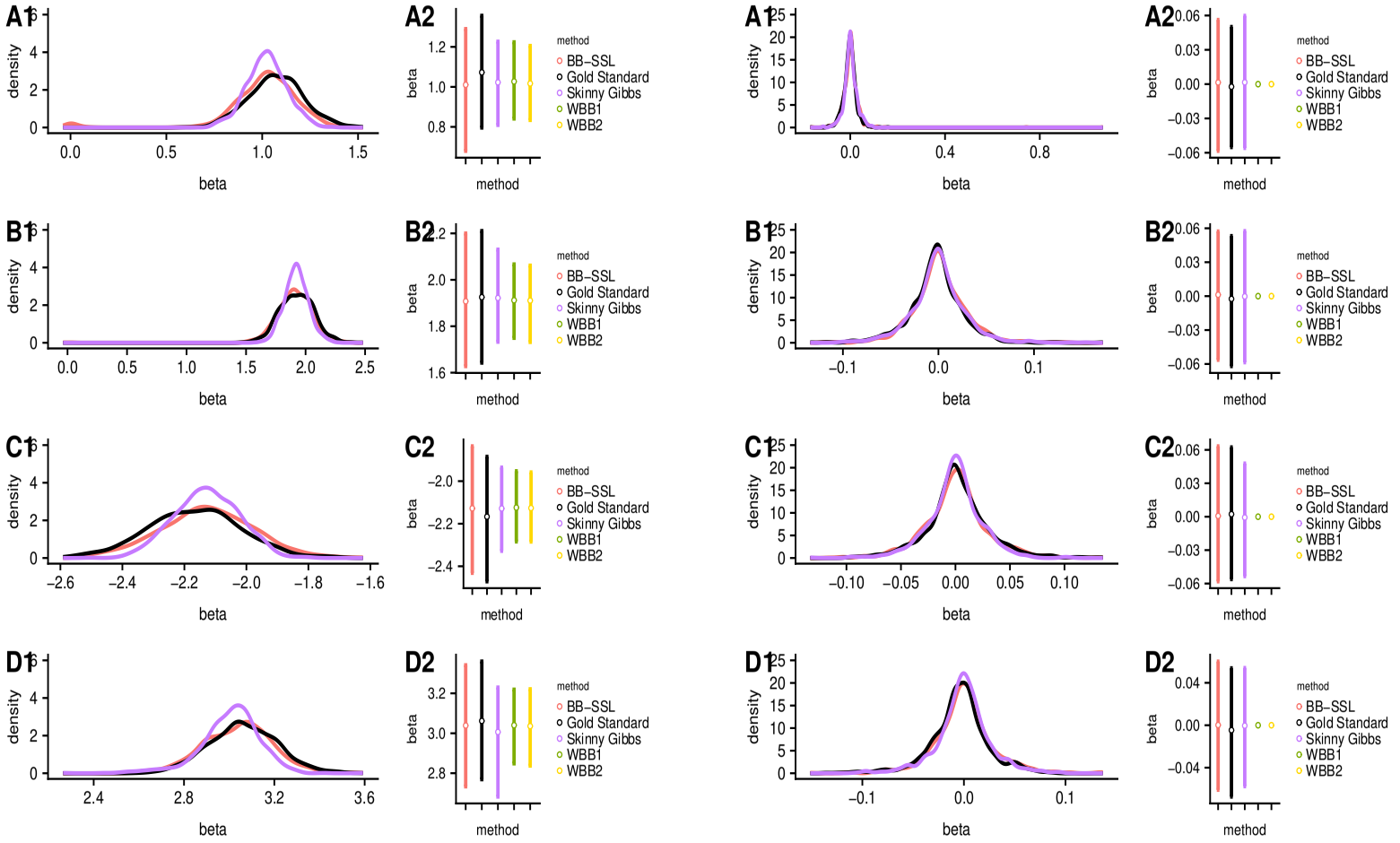
In this section we investigate the influence of α on BB-SSL posterior under (1) high-dimensional, moderately correlated ($\rho = 0.6$) setting, as shown in Figure 8, and (2) Durable Goods Marketing Data Set, as shown in Figure 9.

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



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

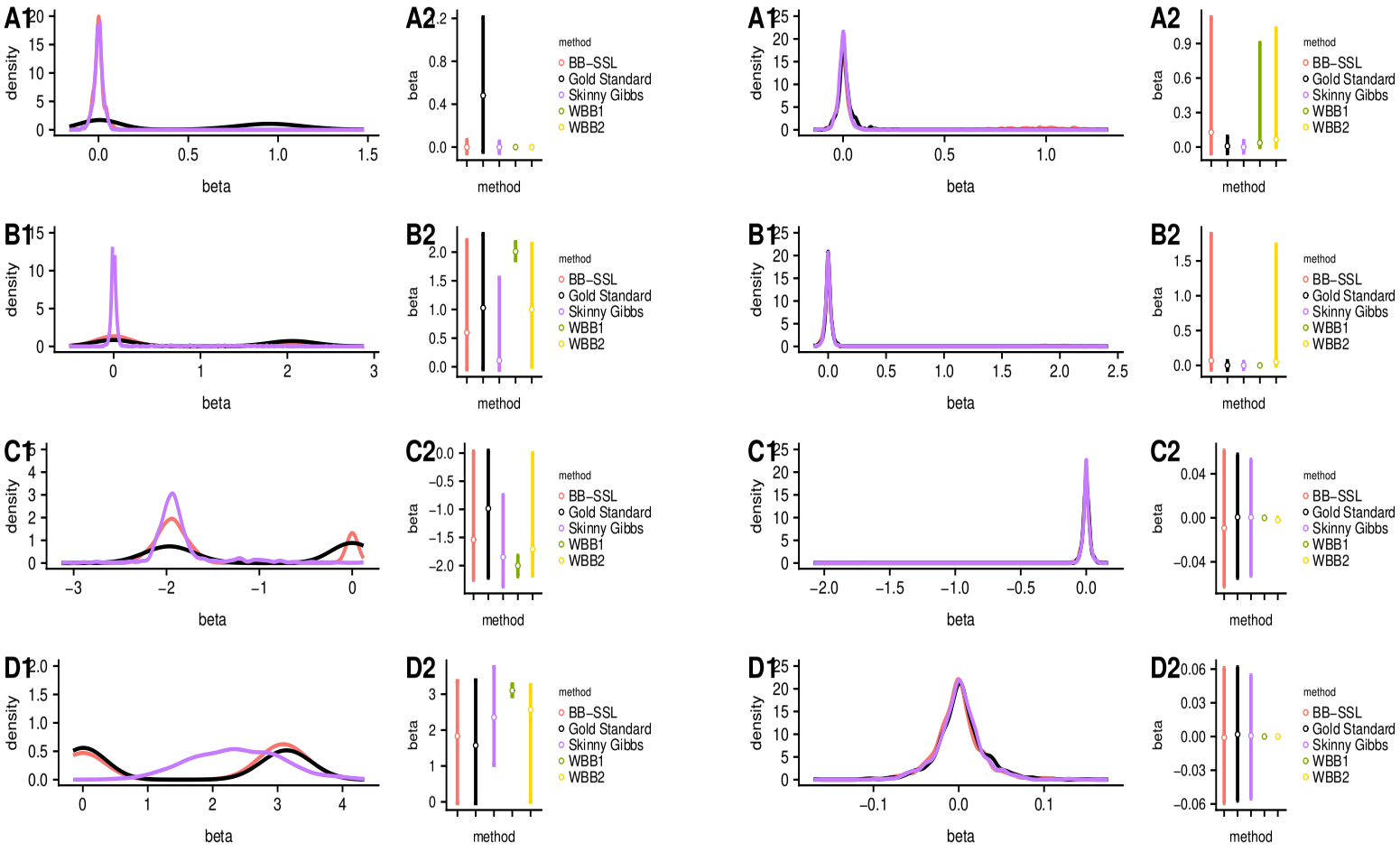
Figure 4: Density and credible interval of active β_i's in high-dimensional, independent case. $n = 100, p = 1000, \beta_{active} = (1, 2, -2, 3), \lambda_0 = 50, \lambda_1 = 0.05$ and predictors are mutually independent. Each method (is thinned to) 1,000 sample points. SSLASSO is fitted using a single λ_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: $\beta_1, \beta_{11}, \beta_{21}, \beta_{31}$

(b) Inactive predictors, from top to bottom: $\beta_2, \beta_{12}, \beta_{22}, \beta_{32}$

Figure 5: Density and credible interval of β_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 λ_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.



(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 6: Density and credible interval of β_i's in high-dimensional, extremely correlated case. $n = 100, p = 1000, \beta_{active} = (1, 2, -2, 3), \lambda_0 = 50, \lambda_1 = 0.05$ 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 λ_0's.

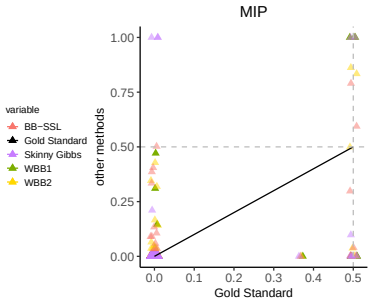
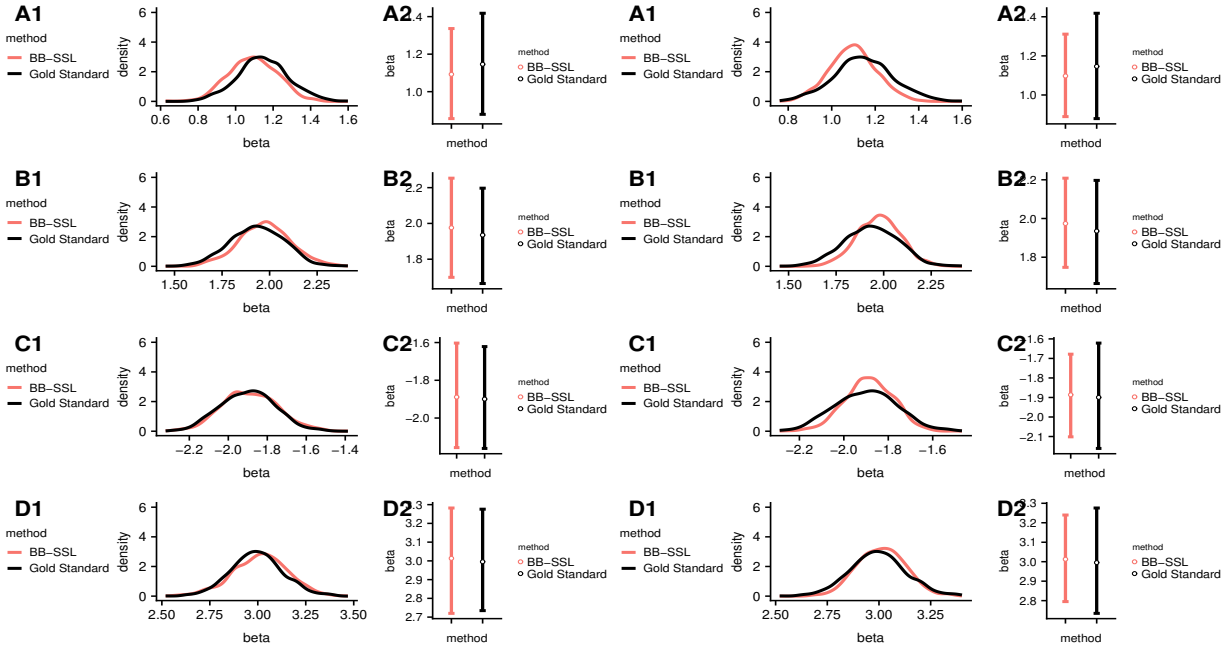
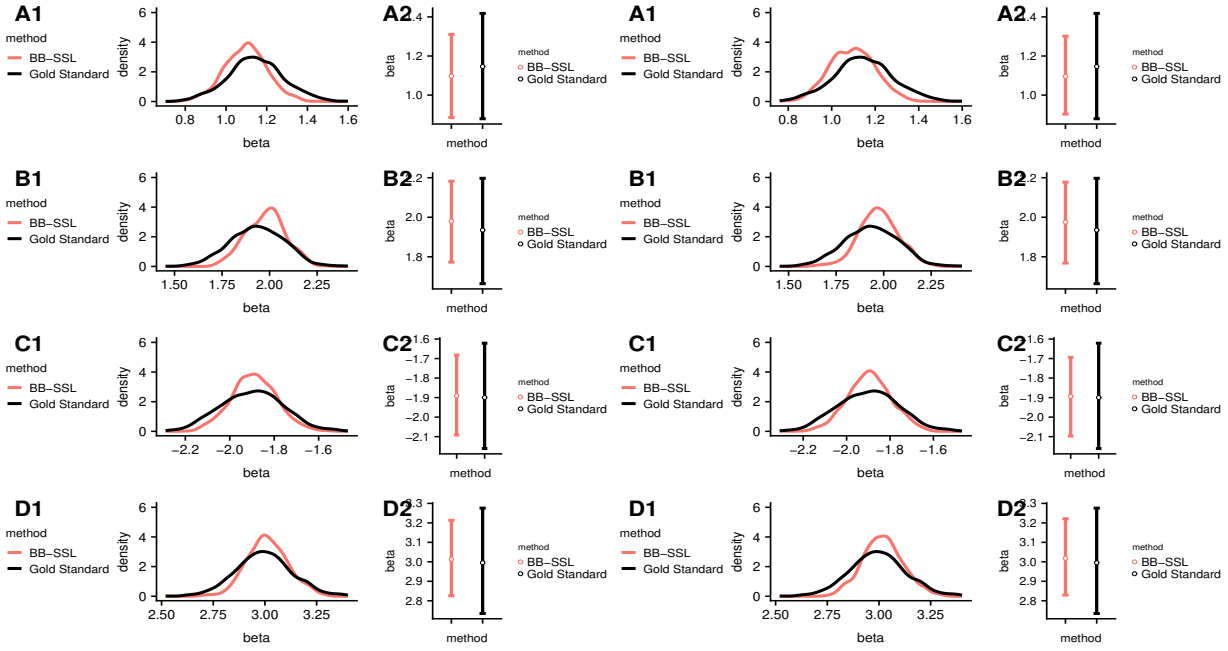


Figure 7: Mean of $\gamma_i, i = 1, 2, \cdots, 1000$ in high-dimensional, correlated case. $n = 100, p = 1000, \beta_{active} = (1, 2, -2, 3), \lambda_0 = 50, \lambda_1 = 0.05$ and predictors are grouped into blocks of size 10 with $\rho = 0.99$. SSL is fitted using a sequence of λ_0's.



(a) $\alpha = 2$.

(b) $\alpha = 10$.



(c) $\alpha = 100$.

(d) $\alpha = 1000$.

Figure 8: Comparison of posterior density for active β_i's when choosing different α in high-dimensional, moderately correlated setting ($\rho = 0.6$). SSL is fitted with λ_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.

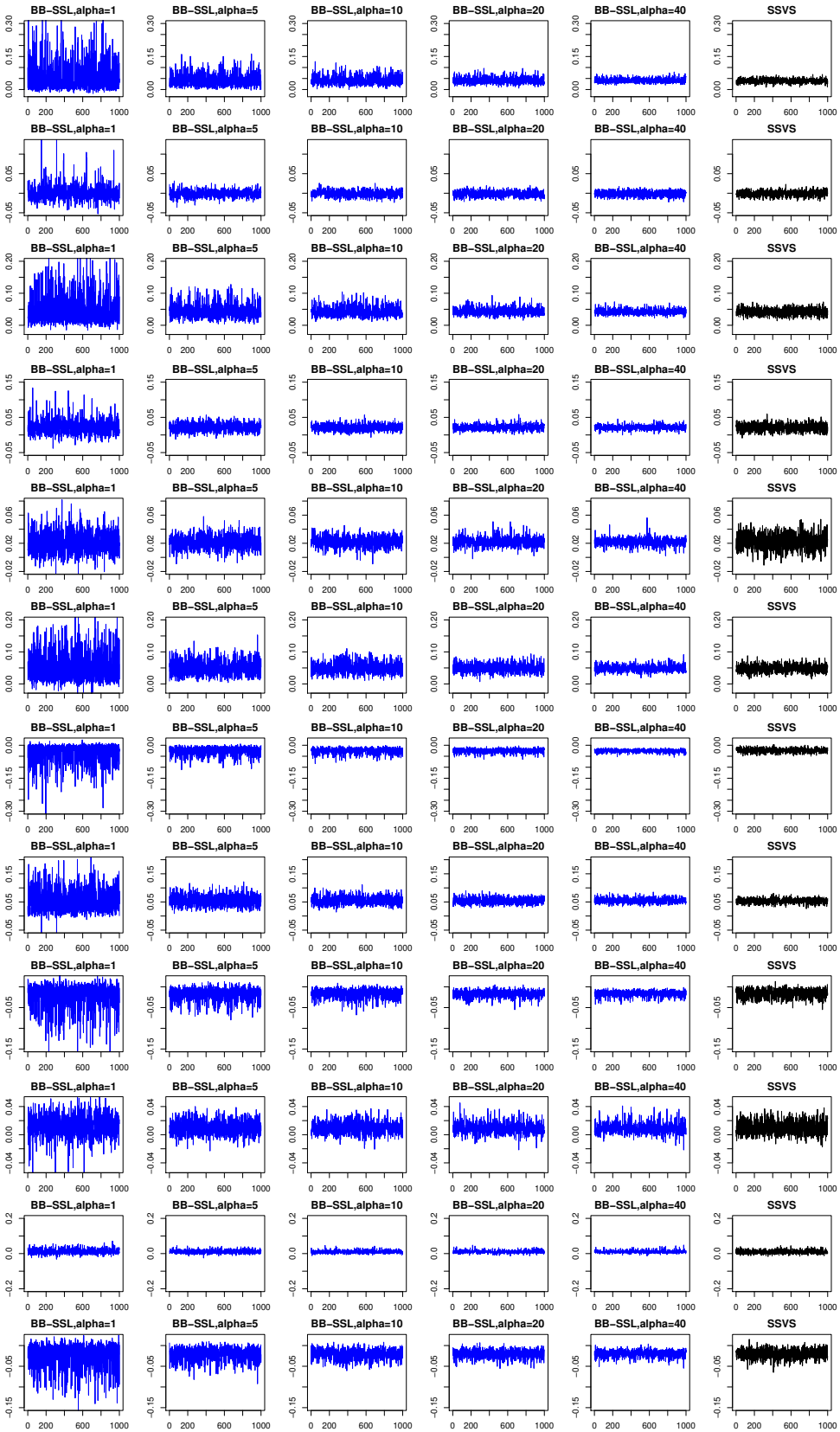


Figure 9: Trace plots of $\beta_i, i, 1, 2, \cdots, 12$ under varying α's in model (20).

References

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, page asw042.

Fong, E., Lyddon, S., and Holmes, C. (2019). Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap. *arXiv:1902.03175*.

Martin, R. and Walker, S. G. (2014). Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electronic Journal of Statistics*, 8(2):2188–2206.

Ročková, V. (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *The Annals of Statistics*, 46(1):401–437.

Ročková, V. and George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444.

Rocková, V. and Moran, G. (2017). SSLASSO: The Spike-and-Slab LASSO. *URL https://cran. r-project. org/package= SSLASSO*, 1:25.

Scheffé, H. (1947). A useful convergence theorem for probability distributions. *The Annals of Mathematical Statistics*, 18(3):434–438.

Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593.