Supplement

Robust Inference for High-Dimensional Linear Models via Residual Randomization

In the supplement, we give proofs for statements in the main manuscript, propose an alternative procedure for selecting $M^\star$, and give additional simulation details.

1 Proofs

Recall the conditions required in the main text.

Condition 1 (Covariates). *Suppose that $X_{i,:} \in \mathbb{R}^p$ are generated i.i.d. with mean 0 and covariance Σ. Let $\lambda_{\max}$ and $\lambda_{\min}$ denote the largest and smallest eigenvalues of Σ. Suppose each element of $X_{i,:}$ is sub-Weibull(α) and the de-correlated covariates $\tilde{X}_{i,:}^\top = \Sigma^{-1/2} X_{i,:}^\top$ are jointly sub-Weibull(α) with*

$$\max\left(\|\tilde{X}_{i,:}^\top\|_{J,\Psi_\alpha}, \max_v \|X_{i,v}\|_{\Psi_\alpha}\right) \leq \kappa. \tag{1}$$

Moreover,

$$\Gamma = \max\left(\max_{v\in[p]} \mathbb{E}\left(\left[a^\top \Sigma^{-1} X_{i,:}^\top X_{i,v}\right]^2\right), \max_{v\in[p]} \mathbb{E}\left(\left[a^\top \Sigma^{-1} X_{i,:}^\top X_{j,v}\right]^2\right), \max_{u,v\in[p]^2} \mathbb{E}\left(\left[X_{i,u} X_{i,v}\right]^2\right)\right) \tag{2}$$

Condition 2 (Sample Size). *Suppose $\kappa^\star = \kappa^2 \max\left(|a|_2 \frac{\sqrt{\lambda_{\max}}}{\sqrt{\lambda_{\min}}}, 1\right)$, and*

$$n > \max\left\{\frac{4C_\alpha^2(\kappa^\star)^2 \left[\log(2n)\right]^{4/\alpha} \left[3\log(pn)\right]^{4/\alpha-1}}{\Gamma}, 64\Gamma(\log(pn) + 2\log(p))\right\} \tag{3}$$

for some constant C_α which only depends on α.

Condition 3 (Exchangeability). *Let $\mathcal{G} \subset \mathcal{G}_p$ where $\mathcal{G}_p$ is the set of all matrices corresponding to a permutation g of $[n]$ such that (i) $[n] = N_1 \cup N_2$ for some N_1 and N_2 equal-sized disjoint sets, and (ii) for all $j \in N_1$, $g(j) \in N_2$ and for all $j \in N_2$, $g(j) \in N_1$.*

Condition 4 (Sign Symmetry). *Let $\mathcal{G} \subset \mathcal{G}_s$ where $\mathcal{G}_s$ is the set of all diagonal matrices containing only ± 1 such that there is an equal number of positive and negative 1's.*

Condition 5 (Cluster Exchangeability). *Suppose there exist n_c disjoint sets L_k with $[n] = \bigcup_k^{n_c} L_k$ and $|L_k| = n/n_c = J$ such that $\{\varepsilon_i\}_{i\in L_k}$ are exchangeable, but may otherwise be dependent. That is, $\mathcal{G} \subset \mathcal{G}_c$, where $\mathcal{G}_c$ is the set of all block diagonal matrices where the G_{L_k,L_k} block is a permutation matrix satisfying Condition 3.*

Condition 6 (Lasso with sub-Weibull errors). *Suppose ε_i is sub-Weibull(α) with $\|\varepsilon_i\|_{\Psi_\alpha} \leq \kappa$. Suppose that*

$$\lambda_{\min} \geq 54 \min_{1\leq h\leq p}\left\{\Xi_{n,h} + \frac{32k\Xi_{n,h}}{h}\right\} \tag{4}$$

where

$$\begin{aligned}\Xi_{n,h} &= 14\sqrt{2}\sqrt{\frac{\Upsilon_{n,h} h\log(36np/h)}{n}} \\ &\quad + \frac{C_\alpha \kappa^2 h(\log(2n))^{2/\alpha}(h\log(36np/h))^{2/\alpha}}{n} \\ \Theta_h &= \{\theta \in \mathbb{R}^p : |\theta|_0 \le h, |\theta|_2 \le 1\} \\ \Upsilon_{n,h} &= \sup_{\theta\in\Theta_h} \operatorname{var}\left[\left(X_{i,:}^\top \theta\right)^2\right]\end{aligned} \tag{5}$$

Furthermore, suppose that the Lasso penalty term λ_1 is set such that

$$\lambda_1 = 14\sqrt{2}\sigma\sqrt{\frac{\log(np)}{n}} + \frac{C_{\alpha/2}\kappa^2(\log(2n))^{2/\alpha}(2\log(np))^{2/\alpha}}{n}, \tag{6}$$

and in addition to Condition 2

$$n > \frac{C_{\alpha/2}^2 \kappa^4 (\log(pn))^{8/\alpha - 1}}{\sigma^2}, \tag{7}$$

where $\sigma = \max_{v\in[p]} \operatorname{var}(X_{i,v}\varepsilon_v)$ and $C_{\alpha/2}$ is a constant only depending on α.

Lemma 1

Lemma 1. *For any $M \in \mathbb{R}^{p\times p}$, let $d_1\left(F_t(X,\varepsilon), F_{\hat{t}}(X,\varepsilon)\right)$ denote the Wasserstein-1 distance between the oracle randomization distribution and attainable randomization distributions. Then,*

$$\begin{aligned}d_1\left(F_t(X,\varepsilon), F_{\hat{t}}(X,\varepsilon)\right) &\le \left|\hat{\beta}^l - \beta\right|_1 \times \\ &\left[\left|\sqrt{n}a^\top(I - MS)\right|_\infty + \left|a^\top M\right|_1 \mathbb{E}_Q\left(\left|X^\top G X/\sqrt{n}\right|_\infty\right)\right].\end{aligned} \tag{8}$$

where Q is the uniform distribution over G in $\mathcal{G}$.

Proof. The debiased Lasso $\hat{\beta}^{d,M}$ is defined as

$$\hat{\beta}^{d,M} = \hat{\beta}^l + \frac{1}{n} M X^\top (Y - X\hat{\beta}^l) \tag{9}$$

so that

$$\begin{aligned}\hat{\beta}^{d,M} - \beta &= \hat{\beta}^l - \beta + \frac{1}{n}MX^\top(Y - X\hat{\beta}^l) + \frac{1}{n}MX^\top(Y - X\beta) - \frac{1}{n}MX^\top(Y - X\beta) \\ &= \hat{\beta}^l - \beta + \frac{1}{n}MX^\top X(\beta - \hat{\beta}^l) + \frac{1}{n}MX^\top\varepsilon \\ &= \left(I - \frac{1}{n}MX^\top X\right)(\hat{\beta}^l - \beta) + \frac{1}{n}MX^\top\varepsilon.\end{aligned} \tag{10}$$

So that for any M, under the null hypothesis that $a^\top\beta = a_0$, we have

$$\begin{aligned}T_n &= \sqrt{n}(a^\top\hat{\beta}^{d,M} - a_0) \\ &= \sqrt{n}a^\top(\hat{\beta}^{d,M} - \beta) \\ &= \sqrt{n}a^\top(I - MS)(\hat{\beta}^l - \beta) + \frac{1}{\sqrt{n}}a^\top MX^\top\varepsilon.\end{aligned} \tag{11}$$

Thus, the oracle randomization distribution which has access to the realization of ε would be

$$t(G\varepsilon) = \sqrt{n}a^\top\left[\left(I - \frac{1}{n}MX^\top X\right)(\hat{\beta}^l - \beta) + \frac{1}{n}MX^\top G\varepsilon\right] \tag{12}$$

where G is drawn uniformly from $\mathcal{G}$. The attainable randomization distribution which we actually use is

$$\begin{aligned}t(G\hat{\varepsilon}) = \frac{1}{\sqrt{n}}a^\top MX^\top G\hat{\varepsilon} &= \frac{1}{\sqrt{n}}a^\top MX^\top G(\varepsilon + X(\beta - \hat{\beta}^l)) \\ &= \frac{1}{\sqrt{n}}a^\top MX^\top G\varepsilon + \frac{1}{\sqrt{n}}a^\top MX^\top GX(\beta - \hat{\beta}^l)).\end{aligned} \tag{13}$$

For any Q which is a joint distribution over (G_1, G_2) where, marginally, G_1 and G_2, are uniform from $\mathcal{G}$, we have

$$d_1\left(F_t(X,\varepsilon), F_{\hat{t}}(X,\varepsilon)\right) \leq \mathbb{E}_Q\left(|t(G_1\varepsilon) - t(G_2\hat{\varepsilon})|\right) \tag{14}$$

Setting Q to the distribution where $G_1 = G_2$ are drawn uniformly from $\mathcal{G}$, and using (11) and (13), we have:

$$\begin{aligned}
d_1\left(F_t(X,\varepsilon), F_{\hat{t}}(X,\varepsilon)\right) &\leq \mathbb{E}_Q\left(\left|\frac{a^\top MX^\top G\varepsilon}{\sqrt{n}} + \sqrt{n}a^\top(I-MS)(\hat{\beta}^l - \beta)\right.\right. \\
&\qquad \left.\left. - \frac{a^\top MX^\top G\varepsilon}{\sqrt{n}} - \frac{a^\top MX^\top GX(\beta - \hat{\beta}^l)}{\sqrt{n}}\right|_1\right) \\
&\leq \mathbb{E}_Q\left(\left|\sqrt{n}a^\top(I-MS)(\hat{\beta}^l - \beta) + \frac{a^\top MX^\top GX(\hat{\beta}^l - \beta)}{\sqrt{n}}\right|_1\right) \\
&= \mathbb{E}_Q\left(\left|\sqrt{n}a^\top(I-MS) + a^\top MX^\top GX/\sqrt{n}\right|_\infty\right)\left|\hat{\beta}^l - \beta\right|_1 \\
&= \left[\frac{1}{|\mathcal{G}|}\sum_{G\in\mathcal{G}}\left(\left|\sqrt{n}a^\top(I-MS) + a^\top MX^\top GX/\sqrt{n}\right|_\infty\right)\right]\left[|\beta^l - \beta|_1\right] \\
&\leq \left[|\beta^l - \beta|_1\right]\left(\sqrt{n}\left|a^\top(I-MS)\right|_\infty + \frac{1}{|\mathcal{G}|}\sum_{G\in\mathcal{G}}\left|\frac{1}{n}a^\top MX^\top GX\right|_\infty\right)
\end{aligned} \tag{15}$$

□

Lemma 2

Lemma 2. *Under Conditions 1 and 2 and either Condition 3, 4, or 5, we have*

$$P\left(\frac{1}{|\mathcal{G}|}\sum_{G\in\mathcal{G}}\left(X^\top GX|_\infty \geq 8\sqrt{\frac{2\Gamma(\log(pn) + 2\log(p))}{n}}\right)\right) \leq 6|\mathcal{G}|(np)^{-1}. \tag{16}$$

Proof. We bound $|X^\top GX|_\infty$ for each $G \in \mathcal{G}$, and then the final result follows from a union bound.

Exchangeability For some fixed $G \in \mathcal{G}$, letting $g(i) = \{j : G_{ij} \neq 0\}$ and $\gamma_i = \mathrm{vec}(X_{i,:}^\top X_{g(i),:})$ where the vec operator vectorizes the $p \times p$ matrix so that $\gamma_i \in \mathbb{R}^{p^2}$. Note that $\mathbb{E}(\gamma_i) = \mathbb{E}\left(\mathrm{vec}(X_{i,:}^\top X_{g(i),:})\right) = 0$ since $i \neq g(i)$. Furthermore,

$$\|X_{i,u}X_{g(i),v}\|_{\psi_{\alpha/2}} \leq \|X_{i,u}\|_{\Psi_\alpha}\|X_{g(i),v}\|_{\Psi_\alpha} \leq \kappa^2 \leq \kappa^\star. \tag{17}$$

Thus, each element of γ_i is sub-Weibull($\alpha/2$) with Orlicz-norm bounded by $\kappa^\star$. Now,

$$\left|\frac{1}{n}X^\top GX\right|_\infty = \left|\frac{1}{n}\sum_i X_{i,:}^\top X_{g(i),:}\right|_\infty = \left|\frac{1}{n}\sum_i \gamma_i\right|_\infty, \tag{18}$$

but the γ_i's are not independent of each other because $X_{i,:}$ appears both in γ_i and $\gamma_{g^{-1}(i)}$. However, by construction, each $X_{i,:}$ only appears in one term of $\{\gamma_i\}_{i\in N_1}$ and one term in $\{\gamma_i\}_{i\in N_2}$. Thus, we can decompose the entire sum with possibly dependent terms into two separate sums of independent terms.

$$\begin{aligned}
\left|\frac{1}{n}\sum_i \gamma_i\right|_\infty &= \left|\frac{1}{n}\sum_{i\in N_1}\gamma_i + \frac{1}{n}\sum_{i\in N_2}\gamma_i\right|_\infty \\
&\leq \max\left(\left|\frac{2}{n}\sum_{i\in N_1}\gamma_i\right|_\infty, \left|\frac{2}{n}\sum_{i\in N_2}\gamma_i\right|_\infty\right)
\end{aligned} \tag{19}$$

We apply Kuchibhotla & Chakrabortty (2018, Theorem 3.4) to each term with $t > 0$ so that

$$
\begin{aligned}
P & \left(\left| \frac{1}{n} \sum_i \gamma_i \right|_\infty \geq 7\sqrt{\frac{2\Gamma(t + 2\log(p))}{n}} + \frac{2C_\alpha \kappa^\star \log(n)^{2/\alpha}(t + 2\log(p))^{2/\alpha}}{n} \right) \\
&\leq P\left(\max\left(\left| \frac{2}{n} \sum_{i \in N_1} \gamma_i \right|_\infty, \left| \frac{2}{n} \sum_{i \in N_2} \gamma_i \right|_\infty \right) \geq 7\sqrt{\frac{2\Gamma(t + 2\log(p))}{n}} + \frac{2C_\alpha \kappa^\star \log(n)^{2/\alpha}(t + 2\log(p))^{2/\alpha}}{n} \right) \\
&\leq 2P\left(\left| \frac{2}{n} \sum_{i \in N_1} \gamma_i \right|_\infty \geq 7\sqrt{\frac{2\Gamma(t + 2\log(p))}{n}} + \frac{2C_\alpha \kappa^\star \log(n)^{2/\alpha}(t + 2\log(p))^{2/\alpha}}{n} \right) \\
&\leq 6\exp(-t).
\end{aligned} \tag{20}
$$

Note that when applying the concentration inequality to each term, the sample size is $n/2$ rather than n. Letting $t = \log(pn)$ and using Condition 2 implies that the

$$
7\sqrt{\frac{2\Gamma(\log(pn) + 2\log(p))}{n}} + \frac{2C_\alpha \kappa^\star \log(n)^{2/\alpha}(\log(pn) + 2\log(p))^{2/\alpha}}{n} \leq 8\sqrt{\frac{2\Gamma(\log(pn) + 2\log(p))}{n}} \tag{21}
$$

and using a union bound over all $G \in \mathcal{G}$ completes the proof.

Cluster Exchangeability Because $\mathcal{G}_c \subset \mathcal{G}_p$, then the proof for exchangeability directly implies that the statement holds for cluster exchangeability as well.

Symmetry We repeat the same arguments with a slight modification due to sign-flipping. For some fixed $G \in \mathcal{G}$, let $\gamma_i = \text{vec}(G_{ii} X_{i,:} X_{i,:}^\top)$ so that $\gamma_i \in \mathbb{R}^{p^2}$. Note that $\mathbb{E}(\gamma_i) = G_{ii}\mathbb{E}\left(X_{i,:} X_i^\top\right) \neq 0$, so we instead pair together each $i \in N_1 = \{i : G_{ii} = 1\}$ with some $j \in N_2 = \{i : G_{ii} = -1\}$. Specifically, assume that N_1 and N_2 are ordered and let $N_1(i)$ and $N_2(i)$ denote the ith element of N_1 and N_2 respectively. We then define

$$
\tilde{\gamma}_i = \frac{1}{2}\left(\gamma_{N_1(i)} - \gamma_{N_2(i)}\right) \tag{22}
$$

so that $\mathbb{E}\left(\tilde{\gamma}_i\right) = 0$. Each element of $\tilde{\gamma}_i$ is sub-Weibull$(\alpha/2)$ with Orlicz-norm bounded by $\kappa^\star$. Now,

$$
\begin{aligned}
\left| X^\top G X/n \right|_\infty &= \left| \frac{1}{n} \sum_i G_{ii} X_{i,:} X_i^\top \right|_\infty \\
&= \left| \frac{2}{n} \sum_{i \in [n/2]} \tilde{\gamma}_i \right|_\infty .
\end{aligned} \tag{23}
$$

Now we again apply Kuchibhotla & Chakrabortty (2018, Theorem 3.4) to each term with $t > 0$ so that

$$
\begin{aligned}
P & \left(\left| \frac{2}{n} \sum_{i \in [n/2]} \tilde{\gamma}_i \right|_\infty \geq 7\sqrt{\frac{2\Gamma(t + 2\log(p))}{n}} + \frac{2C_\alpha \kappa^\star \log(n)^{2/\alpha}(t + 2\log(p))^{2/\alpha}}{n} \right) \\
&\leq 3\exp(-t).
\end{aligned} \tag{24}
$$

Again, letting $t = \log(pn)$ and using Condition 2 implies that the

$$
7\sqrt{\frac{2\Gamma_s(\log(pn) + 2\log(p))}{n}} + \frac{2C_\alpha \kappa^\star \log(n)^{2/\alpha}(\log(pn) + 2\log(p))^{2/\alpha}}{n} \leq 8\sqrt{\frac{2\Gamma(\log(pn) + 2\log(p))}{n}}. \tag{25}
$$

Taking a union bound over all $G \in \mathcal{G}$ completes the proof. Note that for sign-flips, we actually get a tighter upper bound on the probability, but we use looser leading term of 6 from the permutation setting for simplicity. □

Lemma 3

Suppose we select $M^\star = M^\star_\lambda$ by solving

$$\lambda^\star = \arg\min_{\lambda\in[0,1)} \delta|a^\top(I - M_\lambda S)|_\infty + \frac{|a^\top M_\lambda|_1}{|\mathcal{G}|}\left(\sum_G \left|\frac{X^\top G X}{n}\right|_\infty\right), \tag{26}$$

where

$$\begin{aligned} M_\lambda = \arg\min_M\ & |a^\top M|_1 \\ \text{s.t. } & \left|a^\top(I - MS)\right|_\infty \le \lambda. \end{aligned} \tag{27}$$

Lemma 3. *Under the Conditions 1 and 2, we have*

$$P\quad |a^\top(I - \Sigma^{-1}S|_\infty \ge 8\sqrt{\frac{\Gamma(\log(pn) + 2\log(p))}{n}}\right) \le 3(np)^{-1}. \tag{28}$$

Thus, with probability at least $1 - 3(np)^{-1}$ *the feasible set of* (27) *is non-empty with* $\lambda = 8\sqrt{\frac{\Gamma(\log(pn)+2\log(p))}{n}}$ *and*

$$|a^\top M_\lambda|_1 \le |a^\top \Sigma^{-1}|_1. \tag{29}$$

Proof. We show that (28) holds which then trivially implies that Σ^{-1} is in the feasible set for $\lambda = 8\sqrt{\frac{\Gamma(\log(pn)+2\log(p))}{n}}$ and that $|a^\top M_\lambda|_1 < |a^\top\Sigma^{-1}|_1$ by the optimality of M_λ.

Let $\gamma_i = a^\top(I - \Sigma^{-1}X_{i,:}^\top X_{i,:})$ such that $\gamma_i \in \mathbb{R}^p$. Note that $\mathbb{E}(\gamma_i) = a^\top\mathbb{E}(I - \Sigma^{-1}X_{i,:}^\top X_{i,:}) = 0$. Furthermore,

$$\begin{aligned} \|\gamma_{i,v}\|_{\psi_{\alpha/2}} &= \|(a^\top\Sigma^{-1}X_{i,:}^\top X_{i,:})_v\|_{\psi_{\alpha/2}} \\ &= \|a^\top\Sigma^{-1/2}\tilde{X}_{i,:}^\top\left(\tilde{X}_{i,:}\Sigma^{1/2}\right)_v\|_{\psi_{\alpha/2}} \\ &\le \|a^\top\Sigma^{-1/2}\tilde{X}_{i,:}^\top\|_{\Psi_\alpha}\|\left(\tilde{X}_{i,:}\Sigma^{1/2}\right)_v\|_{\Psi_\alpha} \le |a|_2\frac{\sqrt{\lambda_{\max}}}{\sqrt{\lambda_{\min}}}\kappa^2 \le \kappa^\star. \end{aligned} \tag{30}$$

Thus, each γ_i is sub-Weibull($\alpha/2$) with Orlicz-norm bounded by $\kappa^\star$. Now, $a^\top(I - \Sigma^{-1}S) = \frac{1}{n}\sum_i\gamma_i$ so we again apply Kuchibhotla & Chakrabortty (2018, Theorem 3.4) which implies that for any $t \ge 0$,

$$P\quad \left|\frac{1}{n}\sum_i\gamma_i\right|_\infty \ge 7\sqrt{\frac{\Gamma(t + 2\log(p))}{n}} + \frac{C_\alpha\kappa^\star\log(2n)^{2/\alpha}(t + 2\log(p))^{2/\alpha}}{n}\right) \le 3\exp(-t). \tag{31}$$

Letting $t = \log(pn)$ and assuming Condition 2, we have

$$\begin{aligned} \frac{C_\alpha\kappa^\star\log(2n)^{2/\alpha}(\log(pn) + 2\log(p))^{2/\alpha}}{n} &= \sqrt{\frac{\Gamma(\log(pn) + 2\log(p))}{n}}\frac{C_\alpha\kappa^\star\log(2n)^{2/\alpha}(\log(pn) + 2\log(p))^{2/\alpha - 1/2}}{\sqrt{n\Gamma}} \\ &\le \sqrt{\frac{\Gamma(\log(pn) + 2\log(p))}{n}}. \end{aligned} \tag{32}$$

Thus, the first term in the lower bound of (31) dominates. Again, since $a^\top(I - \Sigma^{-1}\frac{1}{n}X^\top X) = \frac{1}{n}\sum_i\gamma_i$, we then have

$$P\quad \left|a^\top\left(I - \Sigma^{-1}\frac{1}{n}X^\top X\right)\right|_\infty \ge 8\sqrt{\frac{\Gamma(\log(pn) + 2\log(p))}{n}}\right) \le 3(np)^{-1}. \tag{33}$$

□

Corollary 1. *Assume the conditions of Lemma 3 and Lemma 2. Then with probability greater than* $1 - 3(np)^{-1} - 6|\mathcal{G}|(np)^{-1}$ *using* $M^\star$ *selected from* (27) *and* (26) *yields*

$$d_1\left(F_t(X,\varepsilon), F_{\hat{t}}(X,\varepsilon)\right) \leq \left|\hat{\beta}^l - \beta\right|_1 \times \left[8\left(\delta + \left|a^\top \Sigma^{-1}\right|_1\right)\sqrt{2\Gamma(\log(pn) + 2\log(p))}\right] \tag{34}$$

Proof. Let $b = 8\sqrt{\frac{\Gamma(\log(pn)+2\log(p))}{n}}$ and suppose that Σ^{-1} is in the feasible set for $\lambda = b$. Note, that by Condition 2, $b < 1$. By the optimality of $\lambda^\star$, $M^\star$, and M_b we have

$$\begin{aligned}\delta|a^\top(I - M^\star S)|_\infty + |a^\top M^\star|_1 \mathbb{E}_Q\left(\left|X^\top GX/n\right|_\infty\right) &\leq \delta|a^\top(I - M_b S|_\infty + |a^\top M_b|_1 \mathbb{E}_Q\left(\left|X^\top GX/n\right|_\infty\right) \\ &\leq \delta b + |a^\top \Sigma^{-1}|_1 \mathbb{E}_Q\left(\left|X^\top GX/n\right|_\infty\right).\end{aligned} \tag{35}$$

Lemma 2 and 3 imply that with probability greater than $1 - 3(np)^{-1} - 6|\mathcal{G}|(np)^{-1}$ that Σ^{-1} is feasible for $\lambda = b$ and that $\mathbb{E}_Q\left(\left|X^\top GX/n\right|_\infty\right) \leq \sqrt{2}b$. Applying Lemma 1 then implies that

$$d_1\left(F_t(X,\varepsilon), F_{\hat{t}}(X,\varepsilon)\right) \leq \left|\hat{\beta}^l - \beta\right|_1 \times \left[8\left(\delta + \left|a^\top \Sigma^{-1}\right|_1\right)\sqrt{2\Gamma(\log(pn) + 2\log(p))}\right] \tag{36}$$

□

Theorem 1 (Sub-Weibull Errors and Covariates)**.** *Suppose Conditions 1, 2, and 6 hold. Under either Condition 3 or 4, with probability no less than* $1 - \frac{6|\mathcal{G}|+3}{np} + \frac{3}{np} + \frac{3}{n}$,

$$d_1\left(F_t(X,\varepsilon), F_{\hat{t}}(X,\varepsilon)\right) \leq \frac{10752 s\sqrt{3\Gamma}\sigma}{\lambda_{\min}}(\delta + |a^T\Sigma^{-1}|_1)\frac{\log(np)}{\sqrt{n}} \tag{37}$$

Proof. We combine Corollary 1 with results from Kuchibhotla & Chakrabortty (2018). Specifically, to bound $|\beta^l - \beta|_1$, we apply Kuchibhotla & Chakrabortty (2018, Theorem 4.5), which states that with probability at least $1 - 3(np)^{-1} - 3n^{-1}$, letting the Lasso penalty parameter be:

$$\lambda_1 = 14\sqrt{2}\sigma\sqrt{\frac{\log(np)}{n}} + \frac{C_\gamma \kappa^2 (\log(2n))^{2/\alpha}(2\log(np))^{2/\alpha}}{n} \tag{38}$$

yields $\hat{\beta}^l$ such that

$$|\hat{\beta}^l - \beta|_2 \leq \frac{84\sqrt{2}}{\lambda_{\min}}\left[\sigma\sqrt{\frac{s\log(np)}{n}} + \frac{C_{\alpha/2}\kappa^2\sqrt{s}(\log(np))^{4/\alpha}}{n}\right] \tag{39}$$

We require the corresponding bound on $|\hat{\beta}^l - \beta|_1$. As part of proving Theorem 4.5 (Appendix E.4), Kuchibolta and Chakroborty show that with the probability stated above,

$$\lambda_n \geq 2\left|\frac{X^\top \varepsilon}{n}\right|_\infty. \tag{40}$$

This allows us to apply Hastie et al. (2015, Lemma 11.1) which states that when (40) holds, the estimation error belongs to the cone set:

$$\hat{\nu} = \hat{\beta}^l - \beta \in C(\mathcal{S}, 3) = \left\{\nu : |\nu_{\mathcal{S}^C}|_1 \leq 3|\nu_{\mathcal{S}}|_1\right\}, \tag{41}$$

where $\mathcal{S} = \{j : \beta_j \neq 0\}$ and $\mathcal{S}^C$ is its complement. Thus, we have

$$\begin{aligned}|\hat{\beta} - \beta|_1 &\leq |(\hat{\beta} - \beta)_{\mathcal{S}}|_1 + |(\hat{\beta}^l)_{\mathcal{S}^C}|_1 \\ &\leq 4|(\hat{\beta} - \beta)_{\mathcal{S}}|_1 \leq 4\sqrt{s}|(\hat{\beta} - \beta)_S|_2 \leq 4\sqrt{s}|\hat{\beta} - \beta|_2.\end{aligned} \tag{42}$$

Thus, we have under the same conditions and stated probability that

$$
\begin{aligned}
|\hat{\beta} - \beta|_1 &\leq \frac{336s\sqrt{2}}{\lambda_{\min}} \left[\sigma\sqrt{\frac{\log(np)}{n}} + \frac{C_{\alpha/2}\kappa^2(\log(np))^{4/\alpha}}{n}\right]\Bigg(\\
&\leq \frac{672s\sqrt{2}}{\lambda_{\min}} \left[\sigma\sqrt{\frac{\log(np)}{n}}\right]\Bigg(
\end{aligned}
\tag{43}
$$

Combining with Corollary 1, we then have with probability no less than $1 - \frac{6|\mathcal{G}|+3}{np} - \frac{3}{np} - \frac{3}{n}$,

$$
\begin{aligned}
d_1\left(F_T(X,\varepsilon), F_t(X,\varepsilon)\right) &\leq \sqrt{n}\frac{672s\sqrt{2}}{\lambda_{\min}} \left[\sigma\sqrt{\frac{\log(np)}{n}}\right]\Bigg(\times 8(\delta + |a^T\Sigma^{-1}|_1)\sqrt{\frac{6\Gamma\log(np)}{n}}\Bigg(\\
&= \frac{10752s\sqrt{3\Gamma}\sigma}{\lambda_{\min}}(\delta + |a^T\Sigma^{-1}|_1)\frac{\log(np)}{\sqrt{n}}
\end{aligned}
\tag{44}
$$

□

2 Assumptions of Belloni et al. (2016) for Cluster Dependence

The proof of Theorem 2 follow directly from Corollary 1 and Theorem 1 of Belloni et al. (2016).

Using our notation, we restate the relevant portion of Theorem 1 of Belloni et al. (2016) as well as the conditions required. Recall that we assume that clusters are indexed by $i = 1, \ldots, n_c$ and observations within each cluster are indexed by $j = 1, \ldots, J$ so that $n = n_c J$. To accommodate this notation, we let $X_{ij} \in \mathbb{R}^p$ be the covariates of the jth observation from the ith cluster. Furthermore, let $X_{ijv} \in \mathbb{R}$ denote the vth covariate of the jth observation from the ith cluster. Similarly, let Y_{ij} denote the jth outcome from the ith cluster.

Belloni et al. (2016) begin with a more general additive fixed effects model where:

$$Y_{ij} = f(w_{ij}) + e_i + \varepsilon_{ij} \qquad \text{where} \qquad \mathbb{E}(\varepsilon_{ij} \mid w_{i1}, \ldots w_{iJ}) = 0. \tag{45}$$

However, the Approximately Sparse Model condition stated below requires that f is well approximated by a linear model so that $f(w_{ij}) = X_{ij}^\top \beta + r(w_{ij})$ for some sparse β and $r(w_{ij})$ term which vanishes as p increases. We require the stronger assumption of a linear model; i.e., $r(w_{ij}) = 0$.

Belloni et al. (2016) define the "demeaned observations"

$$\ddot{X}_{ij} = X_{ij} - \frac{1}{J}\sum_{j=1}^{J} X_{ij}, \qquad \ddot{Y}_{ij} = Y_{ij} - \frac{1}{J}\sum_{j=1}^{J} Y_{ij}, \qquad \text{and} \qquad \ddot{\varepsilon}_{ij} = \varepsilon_{ij} - \frac{1}{J}\sum_{j=1}^{J} \varepsilon_{ij}. \tag{46}$$

The Cluster-Lasso estimate is then defined as

$$\hat{\beta} \in \arg\min_b \frac{1}{n_c J}\sum_{i=1}^{n_c}\sum_{j=1}^{J}(\ddot{Y}_{ij} - \ddot{X}_{ij}^\top b)^2 + \frac{\lambda_1}{n_c J}\sum_{v=1}^{p}\hat{\phi}_v |b_v|, \tag{47}$$

where

$$\lambda_1 = 2c\sqrt{n_c J}\,\Phi^{-1}(1 - \gamma/2p) \tag{48}$$

with $c > 1$ being is a constant slack parameter, $\gamma = o(1)$, and Φ is the CDF of the standard Gaussian. Furthermore, $\hat{\phi}_v^2$ are estimates of

$$\phi_v^2 = \frac{1}{n_c J}\sum_{i=1}^{n_c}\left(\sum_{j=1}^{J}\ddot{X}_{ijv}\ddot{\varepsilon}_{ij}\right)^2. \tag{49}$$

Since we do not have access to $\ddot{\varepsilon}_{ij}$, we instead use

$$\hat{\phi}_v^2 = \frac{1}{n_c J}\sum_{i=1}^{n_c}\left(\sum_{j=1}^{J}\ddot{X}_{ijv}\hat{\varepsilon}_{ij}\right)^2. \tag{50}$$

where $\hat{\varepsilon}_{ij}$ are preliminary estimates of $\ddot{\varepsilon}_{ij}$. Belloni et al. (2016) give a procedure for calculating $\hat{\phi}_v$, but ultimately only require with probability $1 - o(1)$ for all $v \in [p]$ that

$$l\phi_v \leq \hat{\phi}_v \leq u\phi_v \tag{51}$$

for some $l \to 1$ and $u \leq C < \infty$.

The Sparse Eigenvalues condition concerns the empirical Gram matrix of the re-centered data

$$\ddot{M} = \{M_{jk}\}_{u,v\in[p]^2}, \qquad M_{u,v} = \frac{1}{n_c J}\sum_{i=1}^{n_c}\sum_{j=1}^{J}\ddot{X}_{ijv}\ddot{X}_{iju}, \tag{52}$$

and requires its minimum and maximum m-sparse eigenvalues to be bounded. Specifically, they require conditions on the quantities

$$\varphi_{\min}(m)(\ddot{M}) = \min_{\delta\in\Delta(m)} \delta^\top \ddot{M}\delta \qquad \text{and} \qquad \varphi_{\max}(m)(\ddot{M}) = \max_{\delta\in\Delta(m)} \delta^\top \ddot{M}\delta \tag{53}$$

where $\Delta(M) = \{\delta \in \mathbb{R}^p : |\delta|_0 \leq m, |\delta|_2 = 1\}$.

Finally, the regularity conditions require two additional quantities. The first, $\bar{\omega}_v$, involves the third moment of the vth covariate and error:

$$\bar{\omega}_v = \left(\mathbb{E}\left[\left| \frac{1}{\sqrt{J}} \sum_{j=1}^{J} \ddot{X}_{ijv} \ddot{\varepsilon}_{ij} \right|^3 \right] \right)^{1/3}. \tag{54}$$

They additionally require a measure of dependence within cluster, $\imath_J$:

$$\imath_J = J \min_{1 \leq v \leq p} \frac{\mathbb{E}\left(\frac{1}{J} \sum_{j=1}^{J} \ddot{X}_{ijv}^2 \ddot{\varepsilon}_{ij}^2 \right)}{\mathbb{E}\left(\frac{1}{J} \left[\sum_{j=1}^{J} \ddot{X}_{ijv}^2 \ddot{\varepsilon}_{ij}^2 \right]^2 \right)}. \tag{55}$$

With no intra-cluster dependence, $\imath_J = J$, but in the worst case, $\imath_J = 1$.

Theorem 1 of Belloni et al. (2016)

Let $\{P_{n,J}\}$ be a sequence of probability laws, such that $\{(Y_{ij}, w_{ij}, X_{ij})\}_{j=1}^{J} \sim P_{n,J}$, i.i.d. across i for which $n_c, J \to \infty$ jointly or $n_c \to \infty, J$ fixed. Suppose that Conditions ASM, SE, and R hold for probability measure $P = P_{P_{n,J}}$ induced by $P_{n,J}$. Consider a feasible Cluster-Lasso estimator with penalty level set by (48) and penalty loadings obeying (51). Then

$$|\hat{\beta} - \beta|_1 = O_p\left(\sqrt{\frac{s^2 \log(p \vee n)}{n_c \imath_J}} \right) \tag{56}$$

Condition ASM (Approximately Sparse Model) The function $f(w_{ij})$ is well approximated by a linear combination of a dictionary of transformations, $X_{ij} = X_{n_c J}(w_{ij})$ where X_{ij} is a $p \times 1$ vector with $p \gg n$ allowed, and $X_{n_c J}$ is a measureable map. That is, for each i and j,

$$f(w_{ij}) = X_{ij}^{\top} \beta + r(w_{ij}), \tag{57}$$

where the coefficient β and the remainder term $r(w_{ij})$ satisfy

$$|\beta|_0 \leq s = o(n_c \imath_J) \qquad \text{and} \qquad \left[\frac{1}{n_c J} \sum_{i=1}^{n_c} \sum_{j=1}^{J} r(w_{ij})^2 \right]^{1/2} \leq A_s = O_p(\sqrt{s/n_c \imath_J}). \tag{58}$$

Condition SE (Sparse Eigenvalues). For any $C > 0$, there exists constants $0 < \kappa' < \kappa'' < \infty$, which do not depend on n but may depend on C, such that with probability approaching one, as $n \to \infty$ $\kappa' \leq \varphi_{\min}(Cs)(\ddot{M}) \leq \varphi_{\max}(Cs)(\ddot{M}) \leq \kappa''$.

Condition R (Regularity Conditions). Assume that for data $\{y_{ij}, w_{ij}\}$ that are i.i.d. across i, the following conditions hold with X_{ij} defined as in Condition ASM with probability $1 - o(1)$:

1. $\frac{1}{J} \sum_{j=1}^{J} \mathbb{E}(\ddot{X}_{ijv}^2 \ddot{\varepsilon}_{ij}^2) + \left[\frac{1}{J} \sum_{j=1}^{J} \mathbb{E}(\ddot{X}_{ijv}^2 \ddot{\varepsilon}_{ij}^2) \right]^{-1} = O(1)$
2. $1 \leq \max_{v \in [p]} \phi_v / \min_{v \in [p]} \phi_v = O(1)$
3. $1 \leq \max_{v \in [p]} \bar{\omega}_v / \sqrt{\mathbb{E}(\phi_v^2)} = O(1)$
4. $\log^3(p) = o(n_c J)$ and $s \log(p \vee n_c J) = o(n_c \imath_J)$
5. $\max_{v \in [p]} |\phi_v - \sqrt{\mathbb{E}(\phi_v^2)}| / \sqrt{\mathbb{E}}(\phi_v^2) = o(1)$.

3 Alternative Procedure for Selecting M

Recall that

$$\begin{aligned} M_\lambda = \arg\min_M \ & |a^\top M|_1 \\ \text{s.t. } & \left|a^\top (I - MS)\right|_\infty \leq \lambda. \end{aligned} \tag{59}$$

Define

$$\begin{aligned} d(\lambda) &= |a^\top (I - M_\lambda S)|_\infty + \frac{1}{|\mathcal{G}|} \sum_G \left| \frac{a^\top M_\lambda X^\top G X}{n} \right|_\infty \\ d'(\lambda) &= |a^\top (I - M_\lambda S)|_\infty + |a^\top M_\lambda|_1 \frac{1}{|\mathcal{G}|} \sum_G \left| \frac{X^\top G X}{n} \right|_\infty \end{aligned} \tag{60}$$

such that $d(\lambda) \leq d'(\lambda)$. When Γ (or some reasonable upper bound) is known, select δ_1 so that $\delta_1 \geq 8\sqrt{\Gamma}$ and $1 > \delta_1 \sqrt{(\log(pn) + 2\log(p))/n}$. Condition 2 ensures that such a δ_1 exists. Then, an alternative way to select $M^\star$ is

$$\begin{aligned} \lambda^\star = \min_{\lambda \in [0,1)} \ & |a^\top (I - M_\lambda S)|_\infty \\ \text{s.t. } & (59) \text{ has non-empty feasible set for } \lambda \text{ and } d(\lambda) \leq d'(\delta_1 \sqrt{(\log(pn) + 2\log(p))/n}). \end{aligned} \tag{61}$$

Similar to the procedure described in the main text, (61) selects a $\lambda^\star$ which minimizes $|a^\top (I - M_\lambda S)|_\infty$.

This procedure, which we refer to as `RR Tuning Free`, may be preferable to the one (`RR`) presented in the main manuscript since it involves selecting a tuning parameter δ_1 which is tied to a population quantity, Γ, rather than picking δ which may be hard to interpret. However, when δ_1 is not large enough to satisfy $\delta_1 \geq 8\sqrt{\Gamma}$, the procedure may not be asymptotically valid. This is in contrast to the original procedure which is asymptotically valid for any δ, though the empirical performance may be affected by selecting δ too small.

We show that this alternative selection procedure is also valid by slightly modifying the proof of Corollary 1.

Corollary 2. *Assume the conditions of Lemma 3 and Lemma 2. Suppose in* (61) *that* $\delta_1 \geq 8\sqrt{\Gamma}$ *and* $1 > \delta_1 \sqrt{(\log(pn) + 2\log(p))/n}$. *Let* $\delta_2 = \delta_1/(8\sqrt{\Gamma})$. *Then with probability greater than* $1 - 3(np)^{-1} - 6|\mathcal{G}|(np)^{-1}$ *using* $M^\star$ *selected from* (61) *yields*

$$\begin{aligned} d_1\left(F_t(X,\varepsilon), F_{\hat{t}}(X,\varepsilon)\right) \leq & \left|\hat{\beta}^l - \beta\right|_1 \times \\ & \left[8\left(\delta_2 + \left|a^\top \Sigma^{-1}\right|_1\right) \sqrt{2\Gamma(\log(pn) + 2\log(p))}\right] \end{aligned} \tag{62}$$

Proof. Let $b = \delta_1 \sqrt{(\log(pn) + 2\log(p))/n}$. Suppose that Σ^{-1} is in the feasible set for $\lambda = b$. By the optimality of $\lambda^\star$, $M^\star$, and M_b we have

$$\begin{aligned} |a^\top (I - M^\star S)|_\infty + \mathbb{E}_Q\left(\left|a^\top M^\star X^\top G X/n\right|_\infty\right) &\leq |a^\top (I - M_b S|_\infty + |a^\top M_b|_1 \mathbb{E}_Q\left(\left|X^\top G X/n\right|_\infty\right) \\ &\leq b + |a^\top \Sigma^{-1}|_1 \mathbb{E}_Q\left(\left|X^\top G X/n\right|_\infty\right). \end{aligned} \tag{63}$$

Lemma 2 and 3 imply that with probability greater than $1 - 3(np)^{-1} - 6|\mathcal{G}|(np)^{-1}$ that Σ^{-1} is feasible for $\lambda = b$ and that $\mathbb{E}_Q\left(\left|X^\top G X/n\right|_\infty\right) \leq \sqrt{2}b$. Applying Lemma 1 then implies that

$$\begin{aligned} d_1\left(F_t(X,\varepsilon), F_{\hat{t}}(X,\varepsilon)\right) \leq & \left|\hat{\beta}^l - \beta\right|_1 \times \\ & \left[8\left(\delta_2 + \left|a^\top \Sigma^{-1}\right|_1\right) \sqrt{2\Gamma(\log(pn) + 2\log(p))}\right] \end{aligned} \tag{64}$$

□

4 Experiment Details

We compare the empirical coverage of 95% confidence intervals produced by `BLPR` (Liu et al., 2017), `HDI` (Dezeure et al., 2015), `SSLASSO` (Javanmard & Montanari, 2014)[1], `SILM` (Zhang et al., 2019) (non-studentized confidence intervals) and residual randomization (`RR`). For each setting, we replicate the experiment 1000 times for for $(n = 50, p = 100)$ and again for $(n = 100, p = 300)$.

In each setting, we sample random $X \in \mathbb{R}^{n \times p}$ with rows drawn i.i.d. from either

- **N1**: $X_{i,:} \sim N(0, I)$
- **G1**: $X_{iv} \sim \text{Gamma}(1, 1) - 1$; i.e., a centered gamma with shape $= 1$ and rate $= 1$
- **N2**: $X_{iv} \sim N(\mu, 1)$ with $P(\mu = -2) = P(\mu = 2) = 0.5$
- **NT**: $X_{i,:} \sim (0, \Sigma)$ for $\Sigma_{ij} = .8^{|i-j|}$
- **GT**: $X_{i,:} \sim \text{Gamma}(\Sigma) - 1$ for $\Sigma_{ij} = .8^{|i-j|}$; i.e., each X_{iv} is marginally a centered gamma with shape $= 1$ and rate $= 1$, but the covariance is Topelitz.
- **WB**: $X_{iv} \sim \text{Weibull}(1, 0.5) - \Gamma(2)$; i.e., a centered Weibull with scale $= 1$ and shape $= 1/2$.

We sample the errors $\varepsilon \in \mathbb{R}^n$ from

- **N1**: $\varepsilon_i \sim N(0, 1)$
- **N2**: $\varepsilon_i \sim N(\mu, 1)$ with $P(\mu = -2) = P(\mu = 2) = 0.5$;
- **HN**: $\varepsilon_i \sim N(0, 2\|X_{i,:}\|_2^2/p)$; i.e., the errors are **h**eteroskedastic and drawn from a **n**ormal distribution.
- **HM**: $\varepsilon_i \sim N(\mu, 2\|X_{i,:}\|_2^2/p)$ with $P(\mu = -2) = P(\mu = 2) = 0.5$; i.e., the errors are **h**eteroskedastic and drawn from a **m**ixture of normal distributions.
- **WB** $\varepsilon_i \sim \text{Weibull}(1, 0.5) - \Gamma(2)$; i.e., a centered Weibull with scale $= 1$ and shape $= 1/2$.

For each setting, we draw $\beta \in \mathbb{R}^p$ with $s = 4$ or 15 active (i.e., non-zero) coordinates drawn from the Rademacher distribution and set the remaining $p - s$ inactive coordinates to 0. We arrange entries in β in such a way that there is one active entry between two inactive entries (isolated) so that $\beta_j = 1$ and $\beta_{j-1} = \beta_{j+1} = 0$, one active between an active entry and an inactive entry (adjacent) so that $\beta_j = \beta_{j-1} = 1$ and $\beta_{j+1} = 0$, and one active entry between two other active entries (sandwiched) $\beta_j = \beta_{j-1} = \beta_{j+1} = 1$. We also use the same scheme for the inactive variables. We then set $Y = X\beta + \varepsilon$.

Since in practice we do not know the appropriate tuning parameter λ_1 a priori, for the residual randomization procedure we employ the Square-Root Lasso (Belloni et al., 2011) implemented in `RPtests` (Shah & Buhlmann, 2017) to obtain estimates for $\hat{\beta}^l$. We follow Zhang & Cheng (2017) and rescale $\hat{\varepsilon}$ by $\sqrt{n/(n - |\hat{\beta}^l|_0)}$ as a finite-sample correction.

Empirically, a larger value of δ generally results in `RR` producing better coverage at the expense of confidence interval length. We set $\delta = 10000$ for all settings; broadly speaking though, we see that for $\delta \geq 1000$, the performance of the proposed procedure is fairly insensitive to the value of δ.

Given Corollary 2 requires $\lambda_2 = \alpha\sqrt{(\log(pn) + \log(p))/n}$ for some $\alpha \geq 8\sqrt{\Gamma}$. In practice, we may not know the value of Γ, but can provide a reasonable upper bound. Since we assumed that $8\sqrt{\Gamma(\log(pn) + \log(p))/n} < 1$ and require that $\lambda^\star < 1$, in the implementation of `RR Tuning Free` used for the simulations, we set $\lambda = .99$. For added interpretability, we parameterize λ with $\alpha\sqrt{\log(p))/n}$ and use `R`'s `optimize` function to find the smallest $\alpha \in [0.001, 0.99/\sqrt{\log(p))/n}]$ whose $d(\lambda^\star) < d(0.99)$.

Throughout our simulations, we use $1,000$ draws for the bootstrap-based methods, and $1,000$ group actions for our method.

[1] https://web.stanford.edu/ montanar/sslasso/code.html

5 Additional Experiments

5.1 Inactive variables: $(n = 50, p = 100)$

In Figures 1 and 2, we show empirical coverage and confidence interval length for the *inactive variables* over 1000 trials when the errors and covariates are all sub-exponential with $(n = 50, p = 100)$ assuming exchangeable and sign symmetric errors. The same plots for the *active variables* are shown in the main document.

Generally, all `DLASSO`, `SILM`, and `RR` achieve (or exceed) nominal coverage. `HDI` performs well under exchangeable errors, but generally undercovers in the symmetric setting. `BLPR` generally performs poorly in all settings.

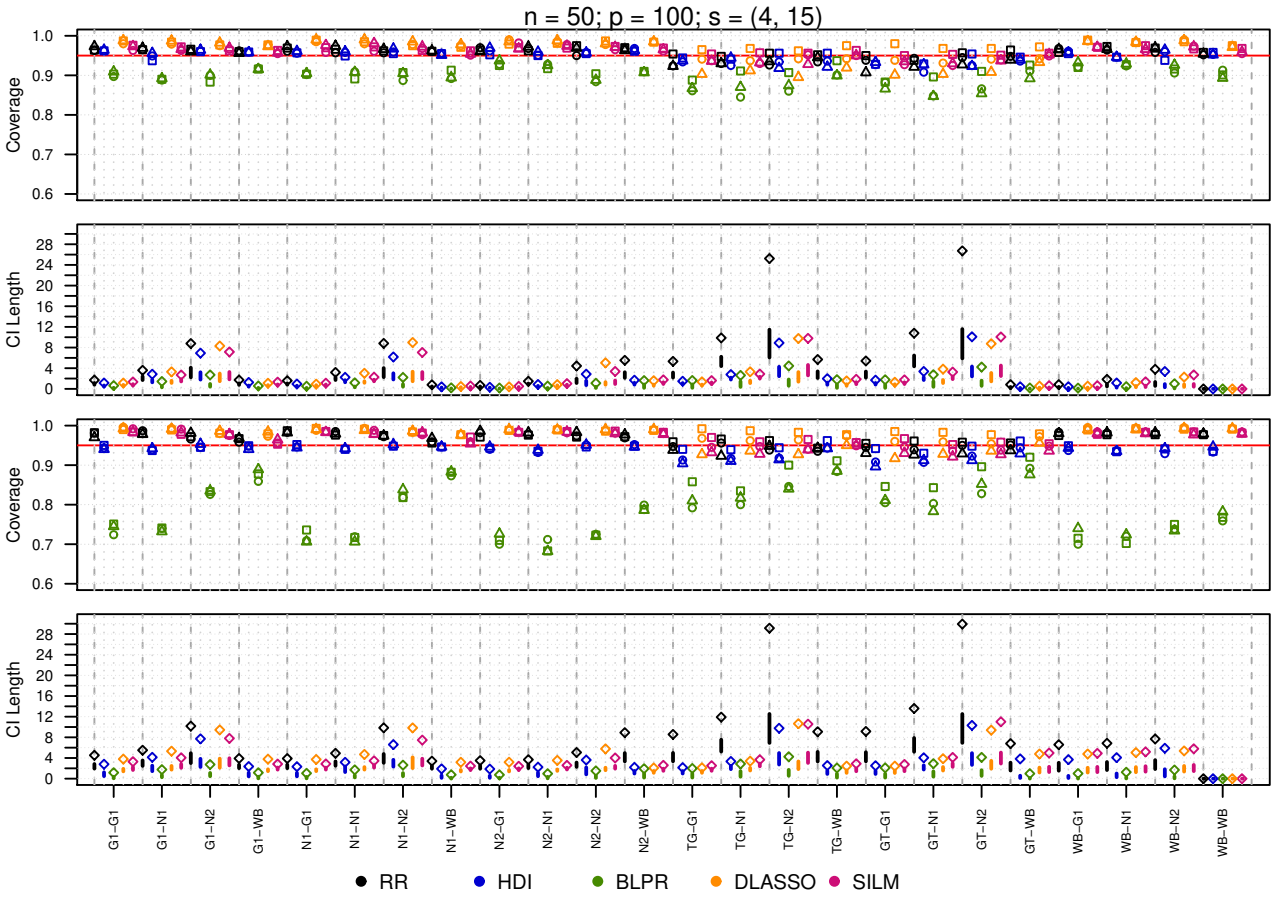

Figure 1: Empirical coverage and confidence interval length for the *inactive* variables with $n = 50$, $p = 100$, 1000 replications and exchangeable errors. The top two panels are for $s = 4$ and the bottom two are for $s = 15$. The first and third panels show empirical coverage rates for each procedure; the sandwich coordinate is denoted by Δ, isolated is $\square$, and adjacent is $\circ$. In the bottom panel, the line segment spans the .25 quantile and .75 quantile of the confidence interval lengths and the single point indicates the .99 quantile. Instead of showing the quantiles for each coordinate, we instead plot the maximum .25 (or .75, .99) quantile across the sandwich, isolated, and adjacent coordinates. The labels on the horizontal axis indicate a different simulation setting and are coded as "Covariate - Errors" where the different covariate and error settings are detailed in the main text. For some settings and procedures, the empirical coverage drops below .6 and is not shown.

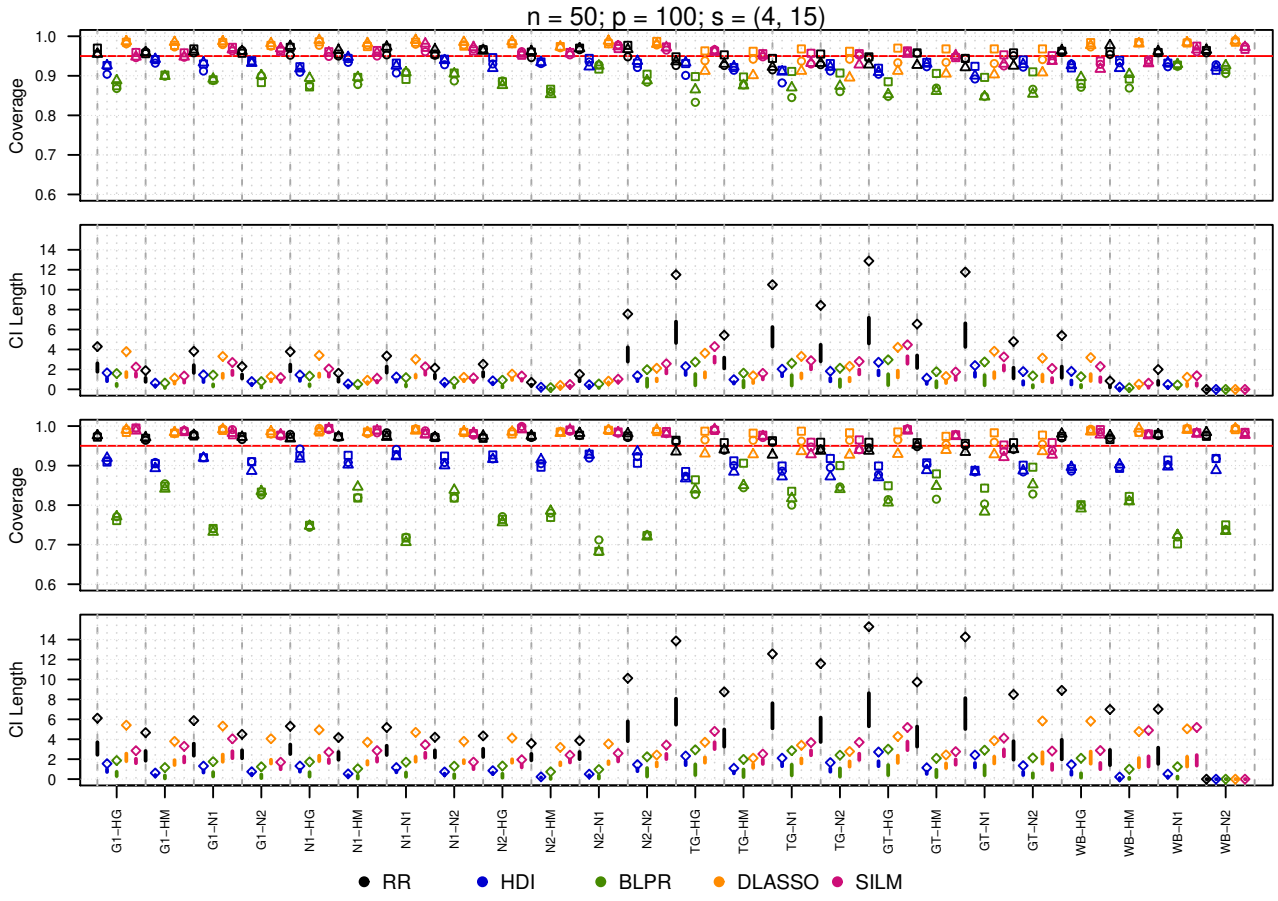

Figure 2: Empirical coverage and confidence interval length for *inactive variables* when $n = 50$ and $p = 100$ for *sign symmetric errors*. All other elements remain the same as Figure 1.

5.2 All variables: $(n = 100, p = 300)$

In Figures 3 and 4, we show empirical coverage and confidence interval length for the active variables over 1000 trials when the errors and covariates are sub-exponential with $(n = 100, p = 300)$ assuming exchangeable and sign symmetric errors. Figures 5 and 6 show the analogous plots for inactive variables.

The conclusions are qualitatively similar to the $(n = 50, p = 100)$ experiments for both active and inactive variables. However, we note that in this case the confidence intervals produced by the residual randomization procedure have lengths comparable to the competing methods in most settings.

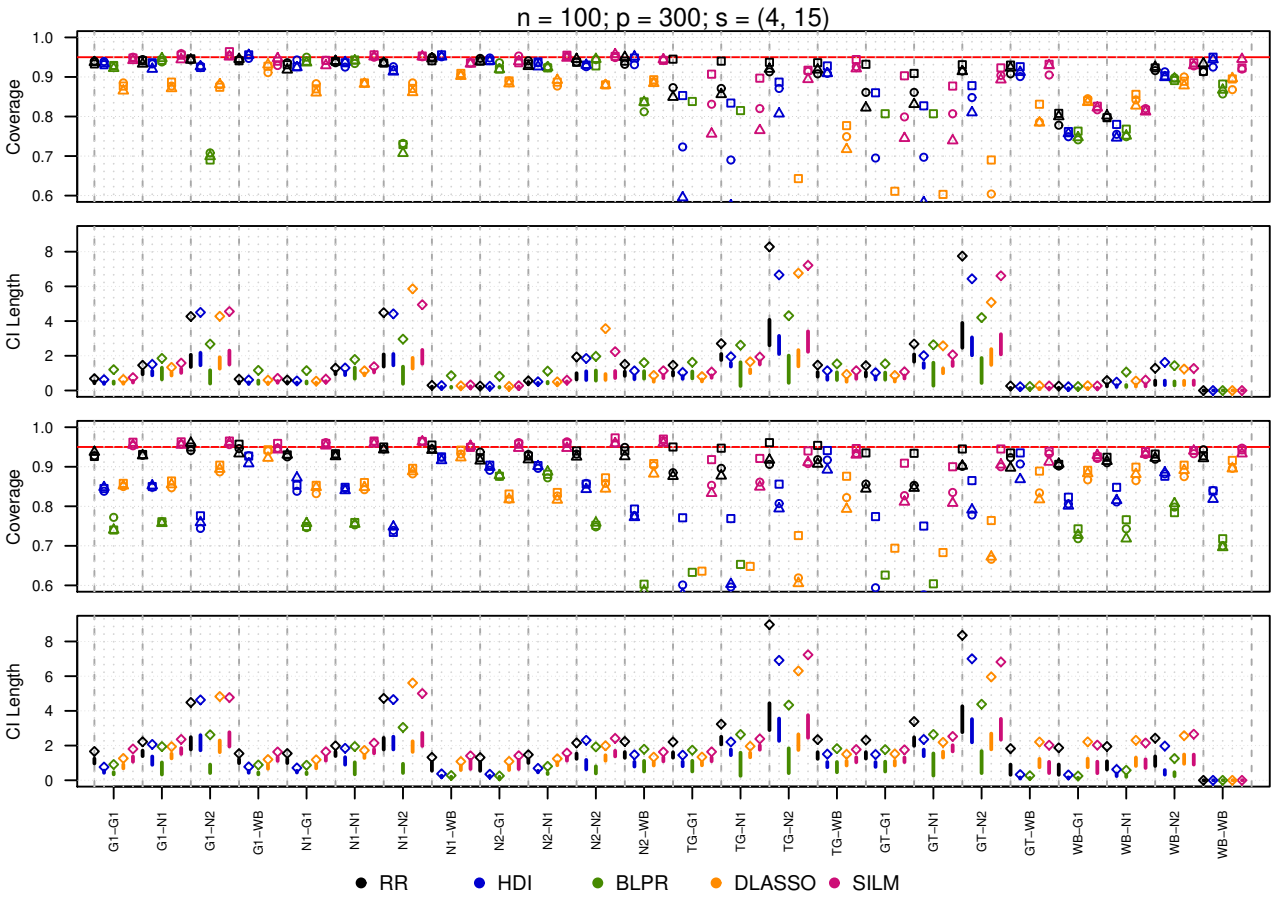

Figure 3: Empirical coverage and confidence interval length for the active variables with $n = 100$, $p = 300$, 1000 replications and exchangeable errors. The top two panels are for $s = 4$ and the bottom two are for $s = 15$. The first and third panels show empirical coverage rates for each procedure; the sandwich coordinate is denoted by Δ, isolated is $\square$, and adjacent is $\circ$. In the bottom panel, the line segment spans the .25 quantile and .75 quantile of the confidence interval lengths and the single point indicates the .99 quantile. Instead of showing the quantiles for each coordinate, we instead plot the maximum .25 (or .75, .99) quantile across the sandwich, isolated, and adjacent coordinates. The labels on the horizontal axis indicate a different simulation setting and are coded as "Covariate - Errors" where the different covariate and error settings are detailed in the main text. For some settings and procedures, the empirical coverage drops below .6 and is not shown.

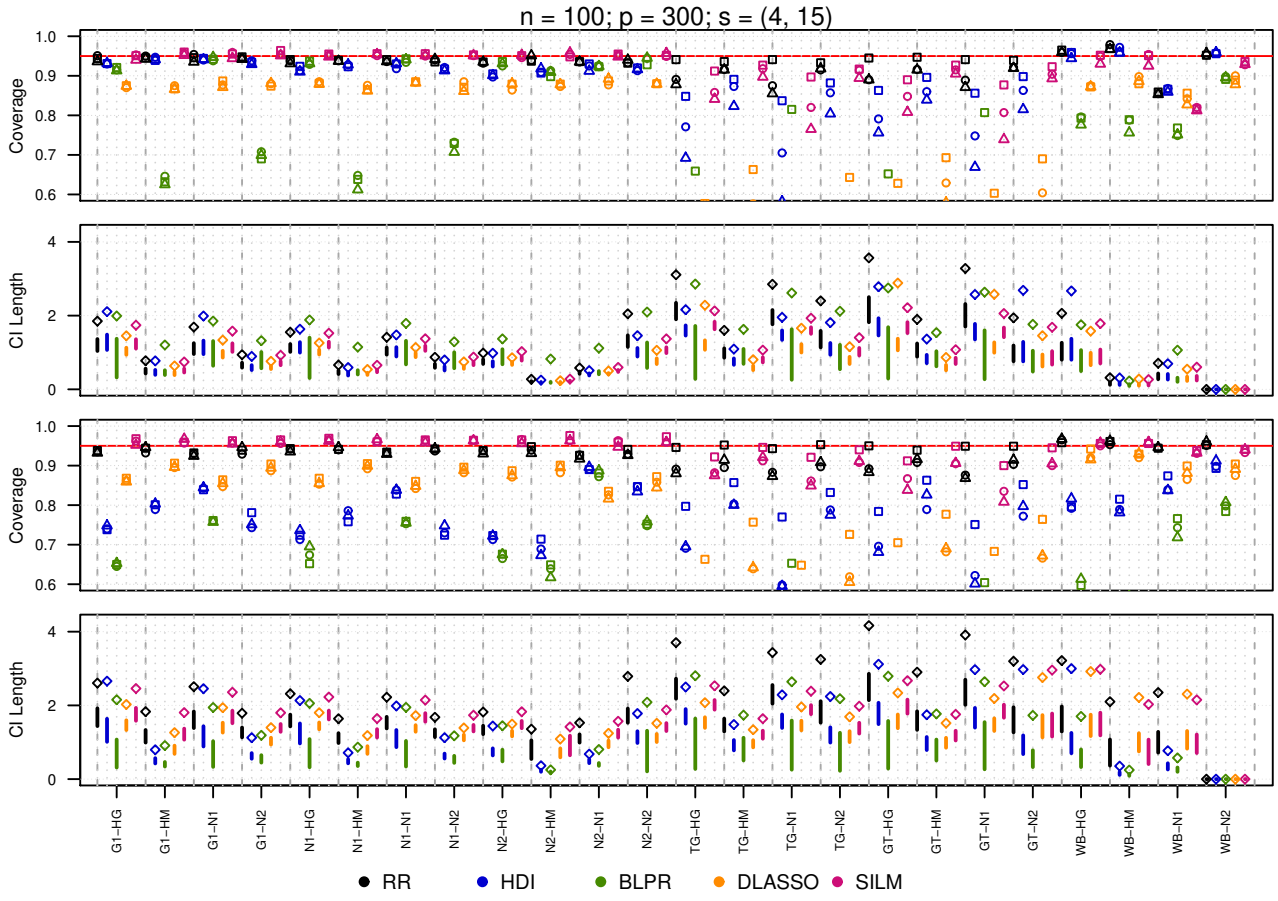

Figure 4: Empirical coverage and confidence interval length for *active variables* when $n = 100$ and $p = 300$ for *sign symmetric errors*. All other elements remain the same as Figure 3.

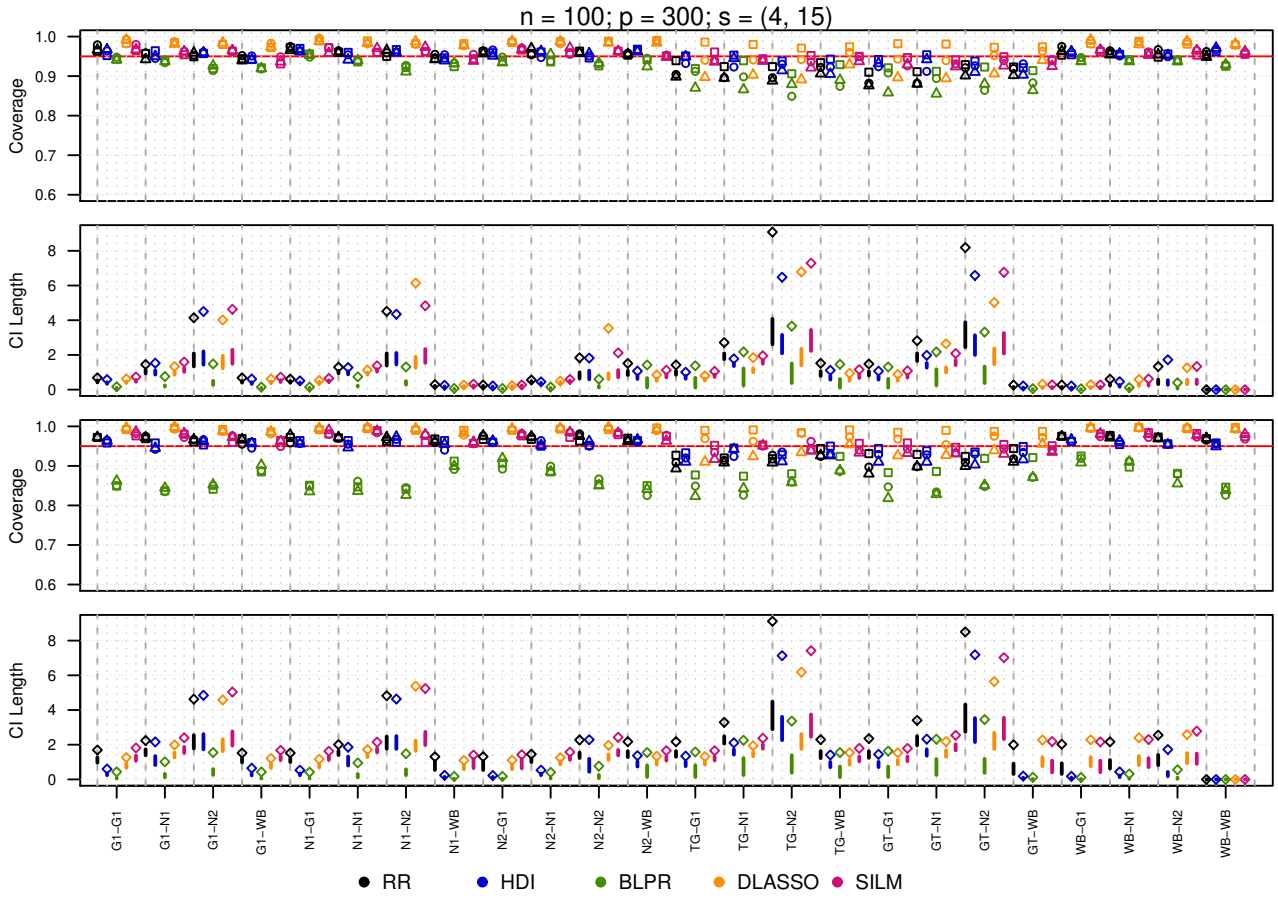

Figure 5: Empirical coverage and confidence interval length for *inactive variables* when $n = 100$ and $p = 300$ for *exchangeable errors*. In the top two and bottom two panels, the true support of β are 4 and 15 respectively. The first and third panels show empirical coverage rates for each procedure. In the bottom panels, the line segment indicates the .25 and .75 quantiles of the confidence interval lengths (averaged across all inactive variables for each run) and the single point indicates the .99 quantile. The labels on the horizontal axis indicate a different simulation setting and are coded as "Covariate - Errors" where the different covariate and error settings are detailed in the main text.

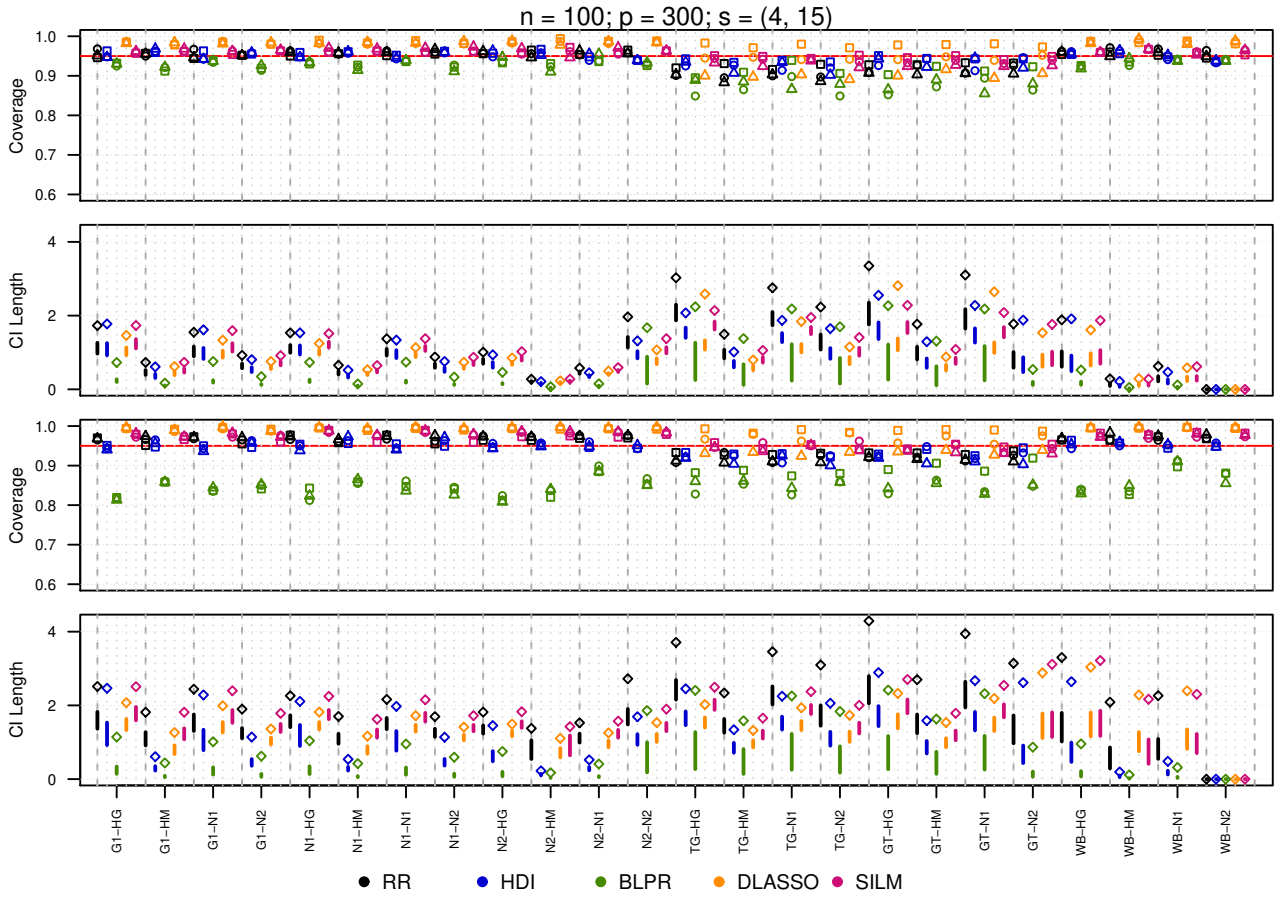

Figure 6: Empirical coverage and confidence interval length for *inactive variables* when $n = 100$ and $p = 300$ for *sign symmetric errors*. All other elements remain the same as Figure 5.

5.3 Comparison of `RR` to `RR Tuning Free`

In the following set of figures, we compare the performance of `RR` to `RR Tuning Free`. We note that when $s = 4$, `RR Tuning Free` generally performs slightly worse compared to `RR`. However, when $s = 15$, `RR Tuning Free` performs comparably. In both cases, `RR Tuning Free` yields shorter CI lengths compared to `RR`, especially with covariates with Toeplitz covariances. With $\delta = 10000$, we would expect the solution from the selection procedure of the original method to be very close to what is obtained via that of `RR Tuning Free` albeit with a less precise grid search. The two main sources of discrepancies comes from 1) the second term in $d(\lambda)$ (60) being a tighter upper bound compared to that in eq. 12 in the main text and 2) `fastclime` symmetrizing M.

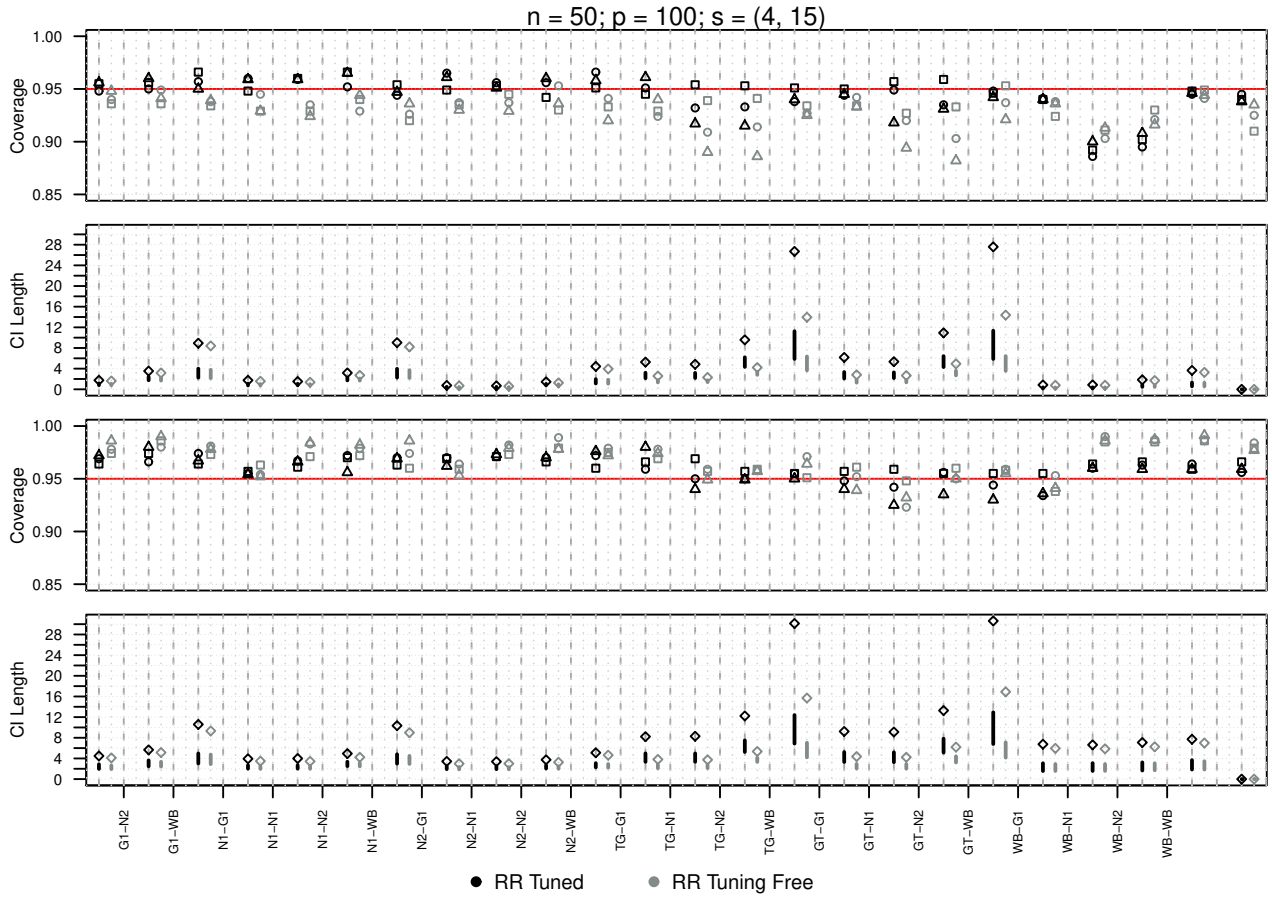

Figure 7: Empirical coverage and confidence interval length for *active variables* when $n = 50$ and $p = 100$ for *exchangeable errors*. In the top two and bottom two panels, the true support of β are 4 and 15 respectively. The first and third panels show empirical coverage rates for each procedure. In the bottom panels, the line segment indicates the .25 and .75 quantiles of the confidence interval lengths (averaged across all inactive variables for each run) and the single point indicates the .99 quantile. The labels on the horizontal axis indicate a different simulation setting and are coded as "Covariate - Errors" where the different covariate and error settings are detailed in the main text.

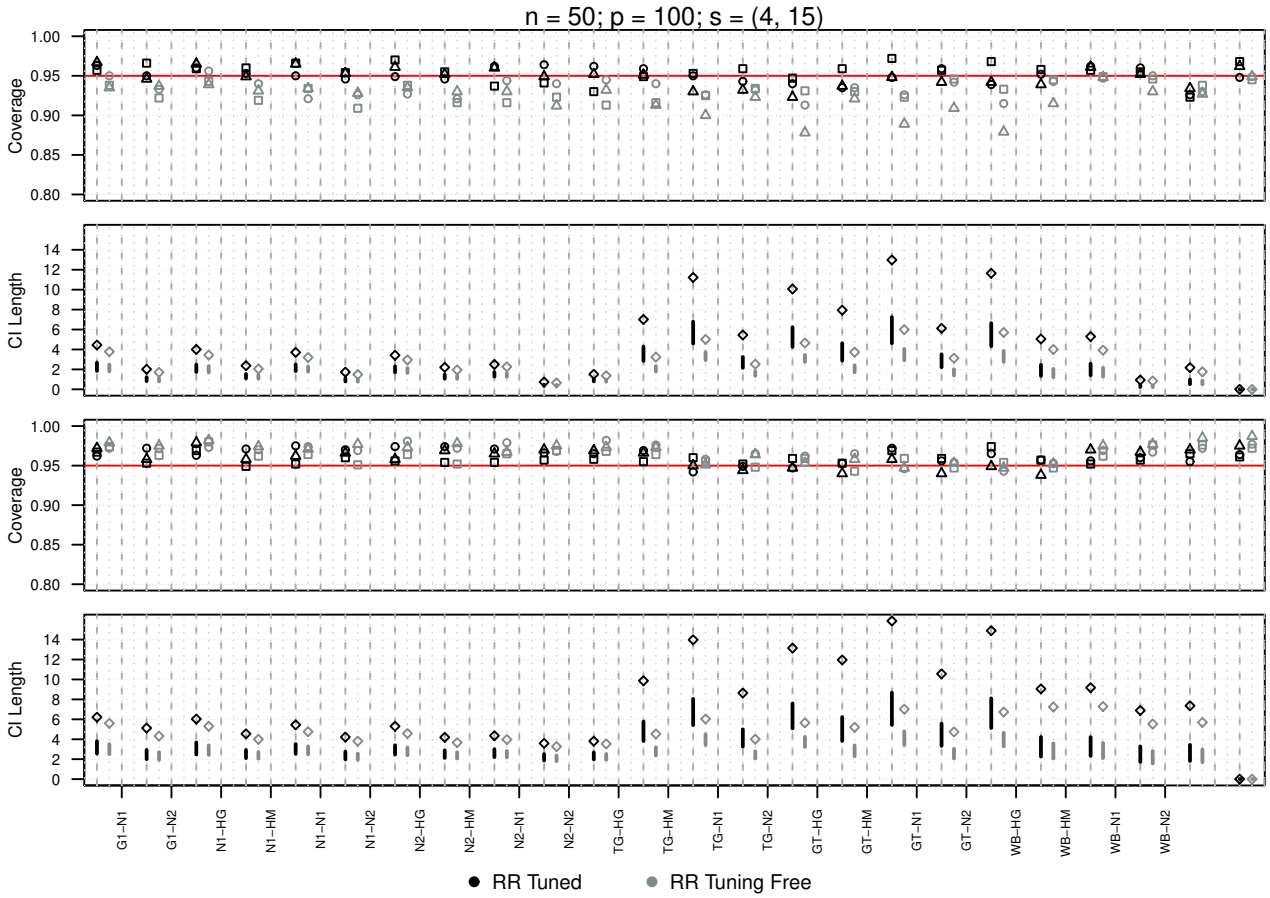

Figure 8: Empirical coverage and confidence interval length for *active variables* when $n = 50$ and $p = 100$ for *sign symmetric errors*. All other elements remain the same as Figure 7.

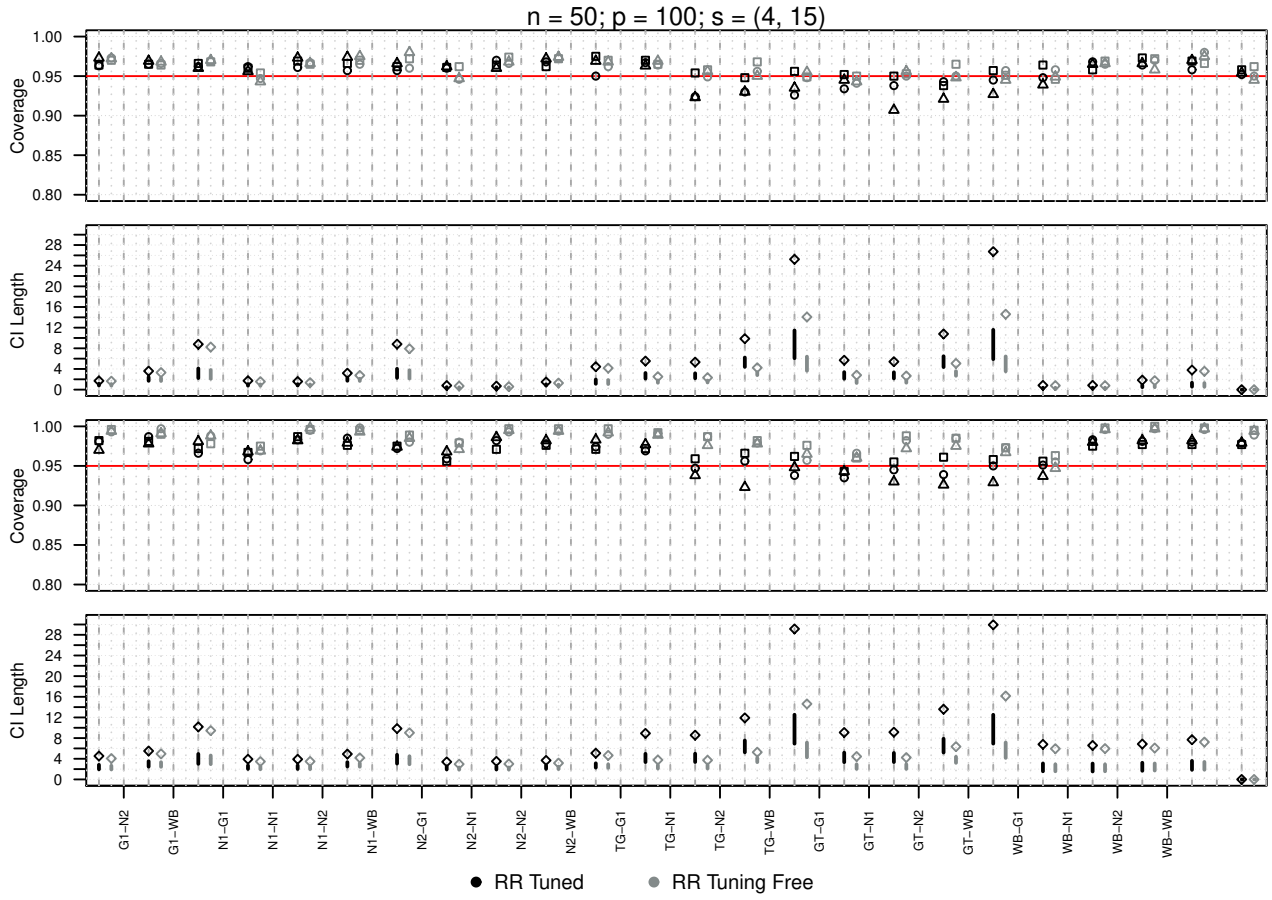

Figure 9: Empirical coverage and confidence interval length for *inactive variables* when $n = 50$ and $p = 100$ for *exchangeable errors*. In the top two and bottom two panels, the true support of β are 4 and 15 respectively. The first and third panels show empirical coverage rates for each procedure. In the bottom panels, the line segment indicates the .25 and .75 quantiles of the confidence interval lengths (averaged across all inactive variables for each run) and the single point indicates the .99 quantile. The labels on the horizontal axis indicate a different simulation setting and are coded as "Covariate - Errors" where the different covariate and error settings are detailed in the main text.

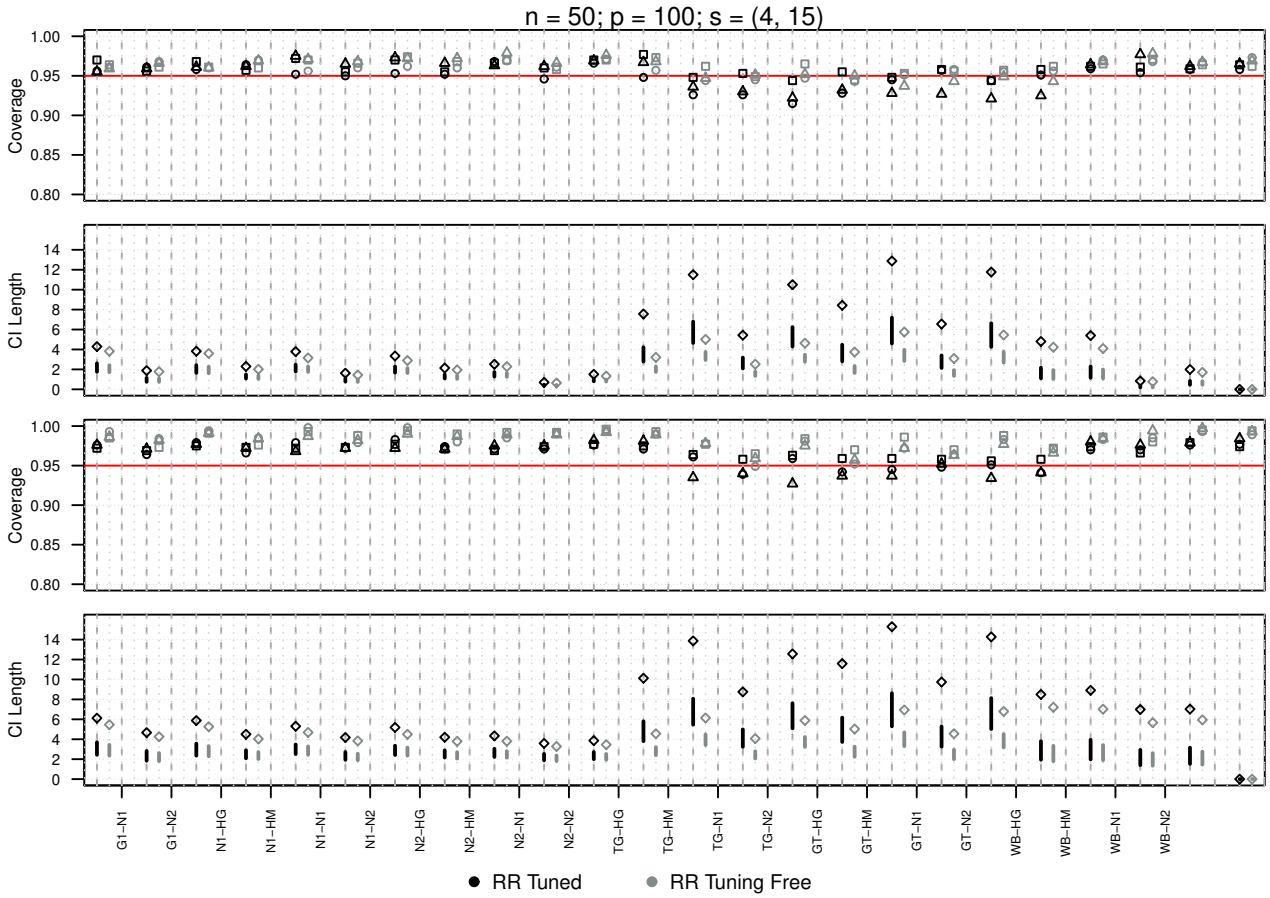

Figure 10: Empirical coverage and confidence interval length for *inactive variables* when $n = 50$ and $p = 100$ for *sign symmetric errors*. All other elements remain the same as Figure 9.

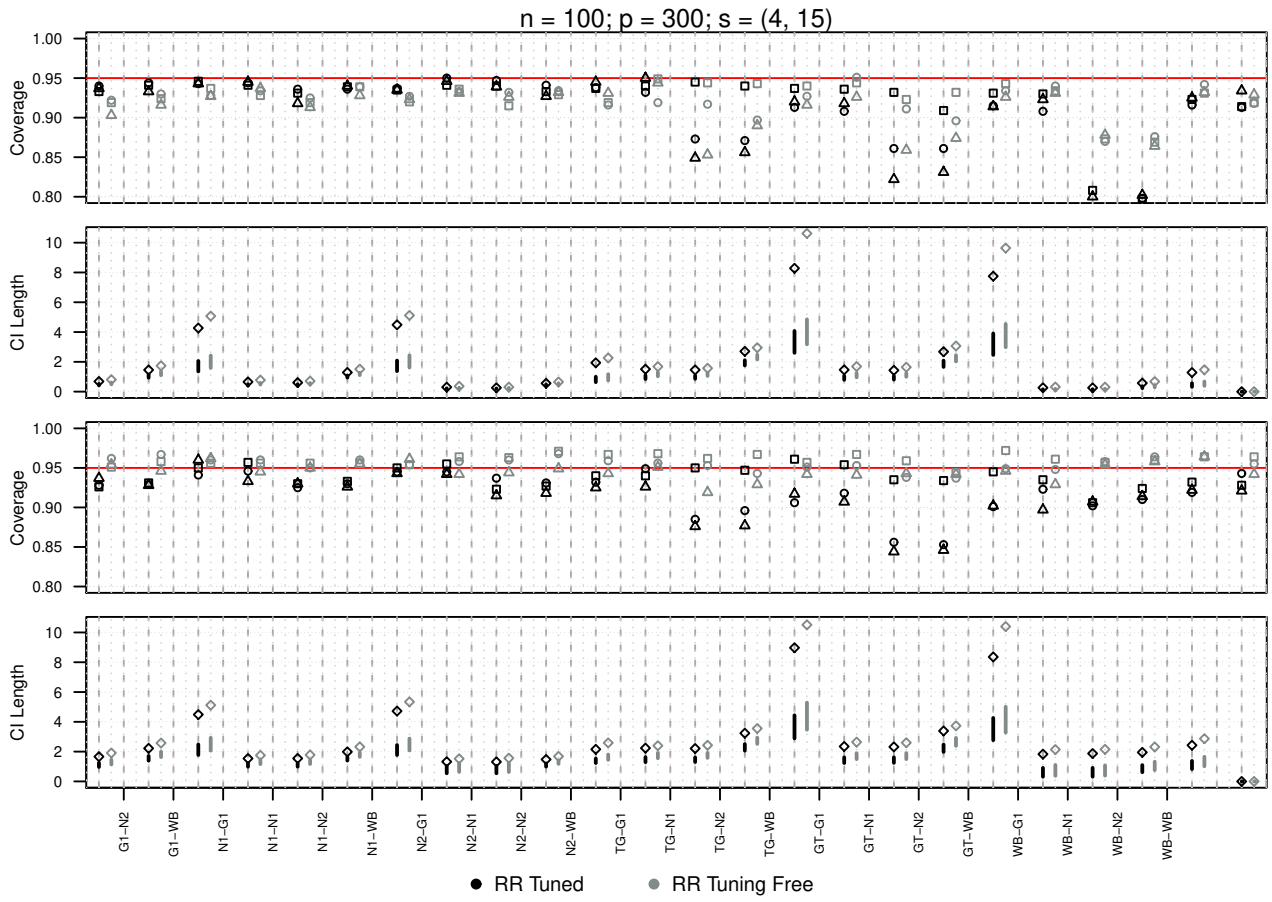

Figure 11: Empirical coverage and confidence interval length for *active variables* when $n = 100$ and $p = 300$ for *exchangeable errors*. In the top two and bottom two panels, the true support of β are 3 and 10 respectively. The first and third panels show empirical coverage rates for each procedure. In the bottom panels, the line segment indicates the .25 and .75 quantiles of the confidence interval lengths (averaged across all inactive variables for each run) and the single point indicates the .99 quantile. The labels on the horizontal axis indicate a different simulation setting and are coded as "Covariate - Errors" where the different covariate and error settings are detailed in the main text.

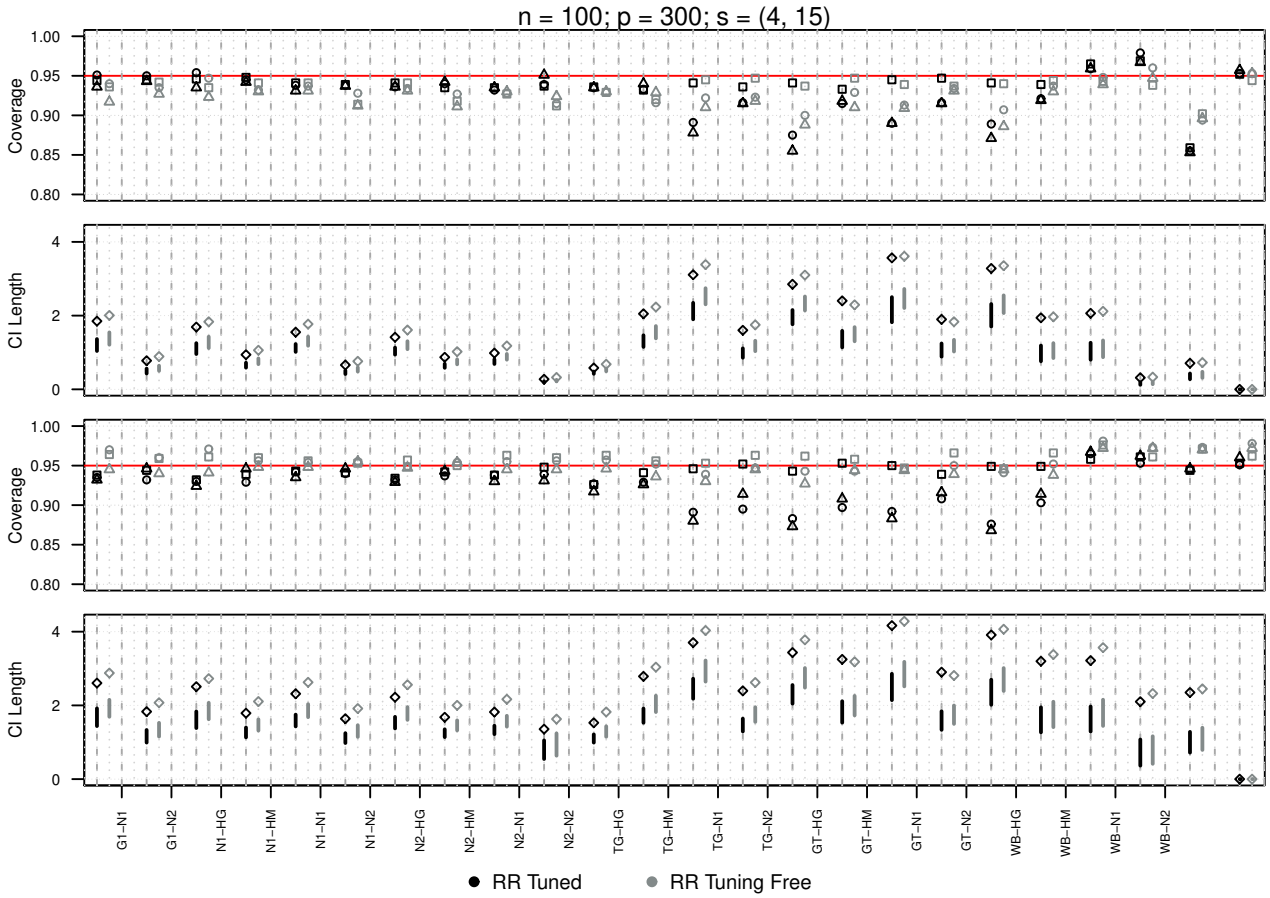

Figure 12: Empirical coverage and confidence interval length for *active variables* when $n = 100$ and $p = 300$ for *sign symmetric errors errors*. All other elements remain the same as Figure 11.

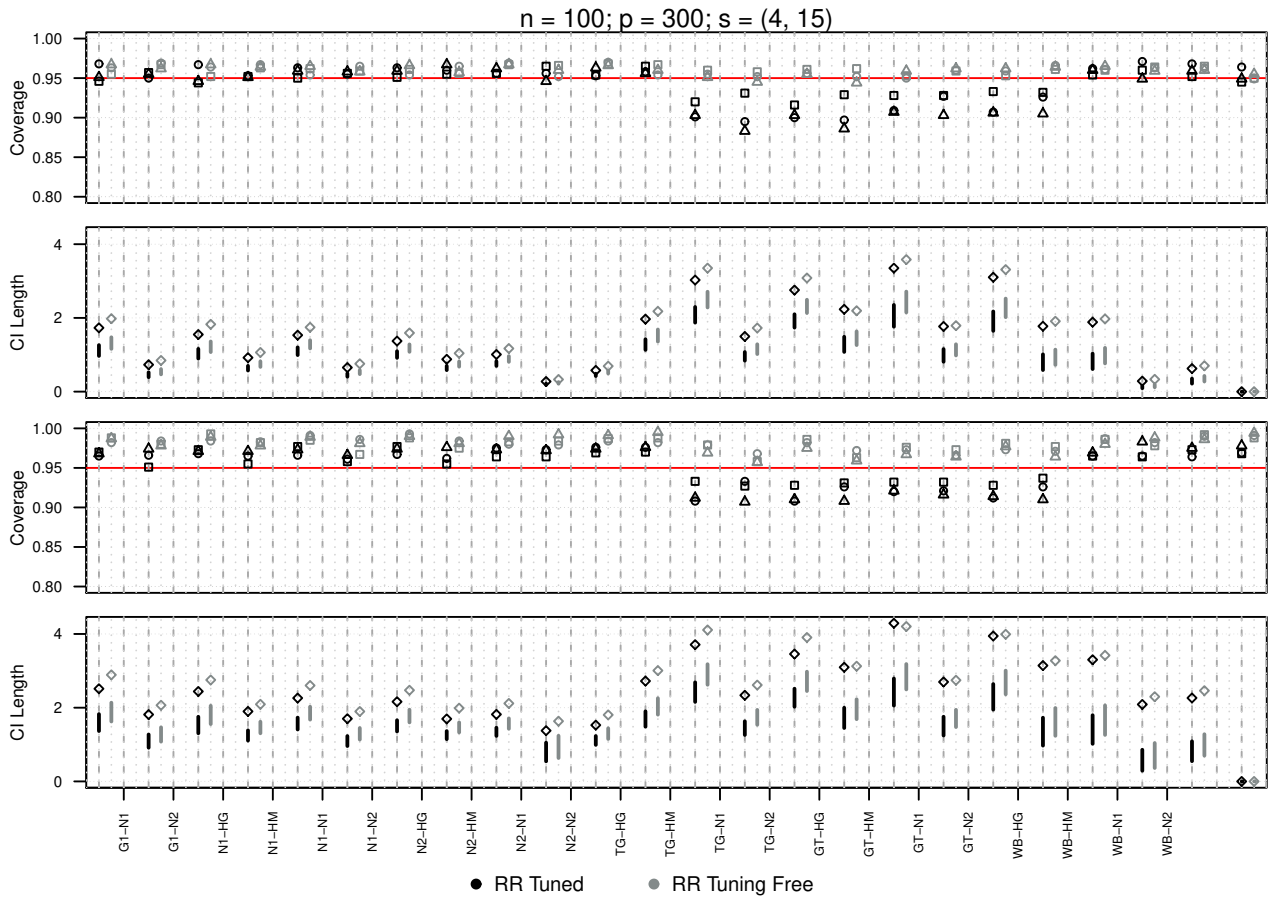

Figure 13: Empirical coverage and confidence interval length for *inactive variables* when $n = 100$ and $p = 300$ for *exchangeable errors*. In the top two and bottom two panels, the true support of β are 4 and 15 respectively. The first and third panels show empirical coverage rates for each procedure. In the bottom panel, the line segment indicates the .25 and .75 quantiles of the confidence interval lengths (averaged across all inactive variables for each run) and the single point indicates the .99 quantile. The labels on the horizontal axis indicate a different simulation setting and are coded as "Covariate - Errors" where the different covariate and error settings are detailed in the main text.

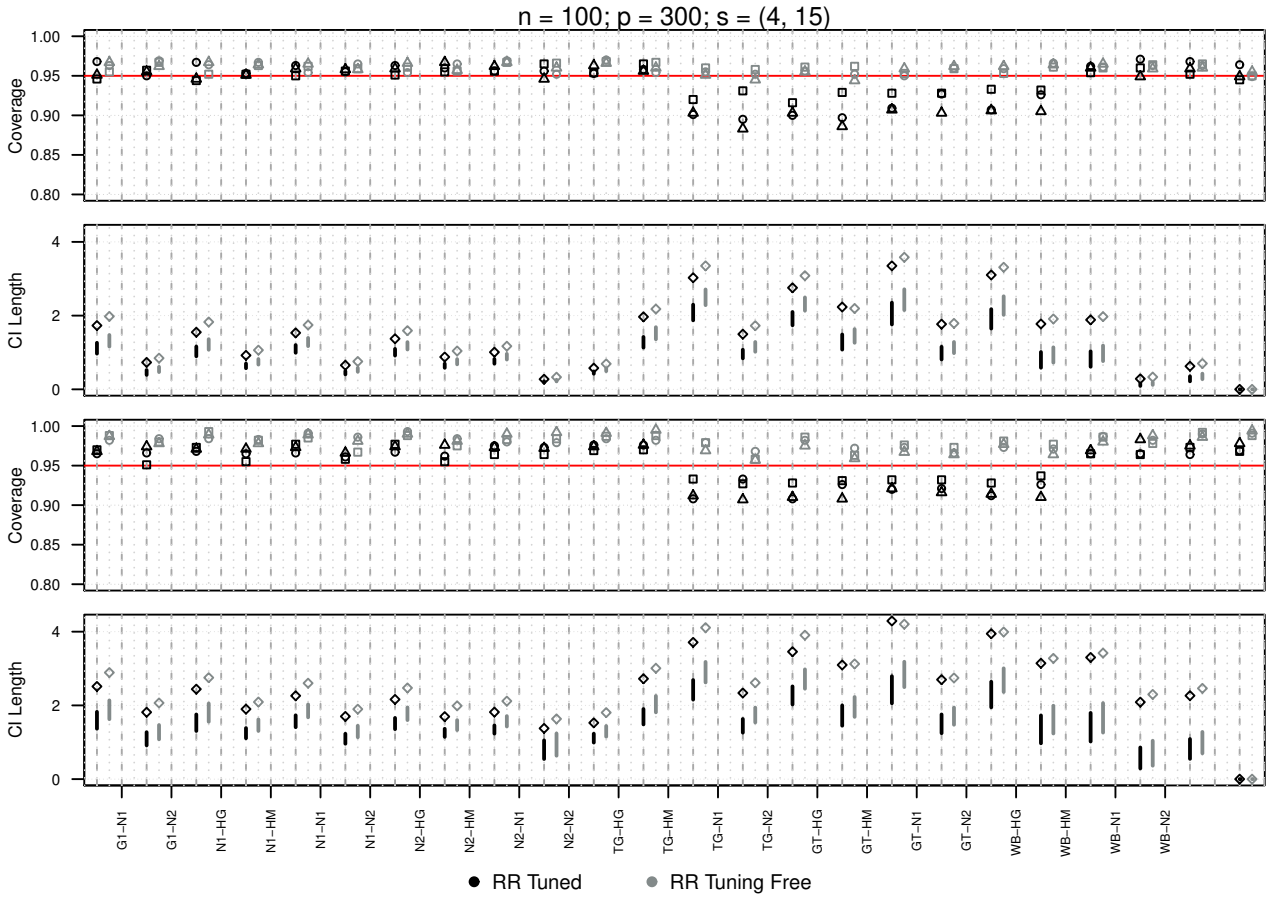

Figure 14: Empirical coverage and confidence interval length for *inactive variables* when $n = 100$ and $p = 300$ for *sign symmetric errors*. All other elements remain the same as Figure 13.

References

Belloni, A., Chernozhukov, V., and Wang, L. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011. ISSN 0006-3444. doi: 10.1093/biomet/asr043. URL `https://doi.org/10.1093/biomet/asr043`.

Belloni, A., Chernozhukov, V., Hansen, C., and Kozbur, D. Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics*, 34(4):590–605, 2016.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. High-dimensional inference: Confidence intervals, p-values and R-software hdi. *Statistical Science*, 30(4):533–558, 2015.

Hastie, T., Tibshirani, R., and Wainwright, M. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

Javanmard, A. and Montanari, A. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Kuchibhotla, A. K. and Chakrabortty, A. Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*, 2018.

Liu, H., Xu, X., and Li, J. J. *HDCI: High Dimensional Confidence Interval Based on Lasso and Bootstrap*, 2017. URL `https://CRAN.R-project.org/package=HDCI`. R package version 1.0-2.

Shah, R. and Buhlmann, P. *RPtests: Goodness of Fit Tests for High-Dimensional Linear Regression Models*, 2017. URL `https://CRAN.R-project.org/package=RPtests`. R package version 0.1.4.

Zhang, X. and Cheng, G. Simultaneous inference for high-dimensional linear models. *J. Amer. Statist. Assoc.*, 112(518):757–768, 2017. ISSN 0162-1459. doi: 10.1080/01621459.2016.1166114. URL `https://doi.org/10.1080/01621459.2016.1166114`.

Zhang, X., Cheng, G., and Bai, J. *SILM: Simultaneous Inference for Linear Models*, 2019. URL `https://CRAN.R-project.org/package=SILM`. R package version 1.0.0.