Large-scale inference of multivariate regression for heavy-tailed and asymmetric data

Youngseok Song[♭], Wen Zhou[§], and Wen-Xin Zhou[♯]

[♭] *Ecole Polytechnique Fédérale de Lausanne,* [§] *Colorado State University*

[♯] *University of California San Diego*

Supplementary Material

This online Supplementary Material contains technical details and additional numerical results. Section S1 provides proofs of the main theorems. The technical lemmas are detailed in Section S2. Section S3 considers our method with $q = 1$, *i.e.*, the linear combination of regression coefficients. We further discuss our method under fixed designs in Section S4. Section S5 reports additional simulation studies and numerical results. More details about the real data application on the Project Gutenberg are given in Section S6.

S1 Proofs of Main Theorems

S1.1 Proof of Proposition 1

For ease of exposition, we use z^2 in place of z in this proof. To prove Proposition 1, we will prove a stronger result that

$$p_0^{-1}V^\circ(z^2) = \mathbb{P}(\chi_q^2 > z^2) + O_{\mathbb{P}}(q^{1/2}p^{-\kappa_1} + n^{-1/2}q^{7/4} + q[n^{-1}\{\log(np) + d\}]^{\delta/(2+\delta)}) \tag{A.1}$$

uniformly over $z \geqslant 0$ as $n, p \to \infty$. Denote $\mathbf{A}_j := \mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}} \in \mathbb{R}^{q\times q}$ the true covariance matrix of $n^{1/2}\mathbf{C}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)$, and $\mathbf{A}_j^{1/2}$ the square root of $\mathbf{A}_j$. The proof consists of two steps: first, we sandwich the number of false discoveries using the Bahadur representation of

$$\mathbf{T}_j^{\circ} = n^{1/2}\mathbf{A}_j^{-1/2}\mathbf{C}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j), \tag{A.2}$$

where $V_j^{\circ} = \|\mathbf{T}_j^{\circ}\|^2$; then we will show that the bounds converge to $\mathbb{P}(\chi_q^2 > z^2)$ as $n, p \to \infty$.

First, we will show that $\mathbf{T}_j^{\circ}$ can be approximated by a q-dimensional multivariate normal distribution, so that V_j° can be approximated by the χ_q^2 distribution due to Lemma B.6. Let

$$\begin{aligned}
\mathbf{S}_j &= n^{-1/2}(\mathbf{C}\boldsymbol{\Sigma}_Z^{-1}\mathbf{C}^{\mathrm{T}})^{-1/2}\mathbf{C}\boldsymbol{\Sigma}_Z^{-1/2}\sum_{i=1}^{n}\boldsymbol{\Sigma}_Z^{-1/2}\left[\ell'_{\tau_j}(\epsilon_{ij})\mathbf{Z}_i - \mathbb{E}\{\ell'_{\tau_j}(\epsilon_{ij})\mathbf{Z}_i\}\right], \\
\mathbf{R}_j &= n^{-1/2}(\mathbf{C}\boldsymbol{\Sigma}_Z^{-1}\mathbf{C}^{\mathrm{T}})^{-1/2}\mathbf{C}\boldsymbol{\Sigma}_Z^{-1/2}\sum_{i=1}^{n}\boldsymbol{\Sigma}_Z^{-1/2}\,\mathbb{E}\{\ell'_{\tau_j}(\epsilon_{ij})\mathbf{Z}_i\}.
\end{aligned} \tag{A.3}$$

Note that $\mathbf{R}_j$ is negligible by Conditions 1 and Proposition B.1. By Corollary B.2,

$$\|\mathbf{T}_j^{\circ} - \sigma_{\epsilon,jj}^{-1/2}(\mathbf{S}_j + \mathbf{R}_j)\| \leqslant C_2\tau_{0j}\frac{d+t}{(n\sigma_{\epsilon,jj})^{1/2}} \tag{A.4}$$

with probability $1 - 3\exp(-t)$ as long as $n \geqslant C_3(d+t)$. For $j = 1, \ldots, p$, let $E_{1j}(t)$ be the event on which (A.4) holds. Set $E_1(t) = \bigcap_{j=1}^{p} E_{1j}(t)$, on which

$$\begin{aligned}
&\sum_{j\in H_0}\mathbb{I}\{\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{S}_j\| \geqslant z + C_2\tau_{0j}(n\sigma_{\epsilon,jj})^{-1/2}(d+t)\} \\
\leqslant\ & V^{\circ}(z^2) \\
\leqslant\ & \sum_{j\in H_0}\mathbb{I}\{\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{S}_j\| \geqslant z - C_2\tau_{0j}(n\sigma_{\epsilon,jj})^{-1/2}(d+t)\}.
\end{aligned} \tag{A.5}$$

with probability $1-3p\exp(-t)$ as long as $n \geqslant C_3(d+t)$.

Define $D_j = \mathbb{I}(\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{S}_j\| \geqslant z)$ and $\mathscr{P}_j = \mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{S}_j\| \geqslant z)$ for $j = 1,\ldots,p$ and $z \geqslant 0$. Under Condition 1, $D_1,\ldots,D_p$ are weakly correlated. Recall that $\mathscr{H}_0 = \{j : 1 \leqslant j \leqslant p, H_{0j} \text{ is true}\}$, it holds

$$\begin{aligned}\operatorname{var}\left(p_0{}^{-1}\sum_{j\in\mathscr{H}_0} D_j\right) &= \frac{1}{p_0^2}\sum_{j\in\mathscr{H}_0}\operatorname{var}(D_j) + \frac{1}{p_0^2}\sum_{j,k\in\mathscr{H}_0;j\neq k}\operatorname{cov}(D_j,D_k)\\ &\leqslant \frac{1}{4p_0} + \frac{1}{p_0^2}\sum_{j,k\in\mathscr{H}_0;j\neq k}\{\mathbb{E}(D_jD_k) - \mathscr{P}_j\mathscr{P}_k\}.\end{aligned} \tag{A.6}$$

We first study $\mathscr{P}_j$. Note that $\mathbf{S}_j$ is a sum of independent random vectors with $\mathbb{E}(\mathbf{S}_j) = \mathbf{0}$ and $\operatorname{cov}(\mathbf{S}_j) = s_j^2\mathbf{I}$ where $s_j^2 = \mathbb{E}[\{\ell'_{\tau_j}(\epsilon_{ij})\}^2]$. Let $\mathbf{G} \sim N(\mathbf{0},\mathbf{I})$ be a standard normal random vector. Lemmas B.2 and B.3 imply that

$$\max_{1\leqslant j\leqslant p}|\mathscr{P}_j - \mathbb{P}(\|\mathbf{G}\| \geqslant z)| \lesssim n^{-1/2}q^{7/4} + q^{1/2}\frac{2}{\delta\sigma_{\epsilon,jj}}\frac{v_{j,\delta}^{2+\delta}}{\tau_{0j}^{\delta}}\left(\frac{d+t}{n}\right)^{\delta/(2+\delta)} \tag{A.7}$$

holds uniformly over $z \geqslant 0$.

Following that, we consider $\mathbb{E}(D_jD_k)$ for each pair (j,k) with $1 \leqslant j \neq k \leqslant p$. Set $\mathbf{S} = (s_j^{-1}\mathbf{S}_j, s_k^{-1}\mathbf{S}_k)^{\mathrm{T}}$. Let $\mathbf{G} = (\mathbf{G}_1,\mathbf{G}_2) = (G_{11},\ldots,G_{1q},G_{21},\ldots,G_{2q})^{\mathrm{T}} \in \mathbb{R}^{2q}$ be a Gaussian vector with $\mathbb{E}(\mathbf{G}) = \mathbf{0}$ and $\operatorname{cov}(\mathbf{G}) = \operatorname{cov}(\mathbf{S})$. The block-structured matrix $\operatorname{cov}(\mathbf{S})$ has unit diagonal entries with $\operatorname{cov}(s_j^{-1}\mathbf{S}_j) = \mathbf{I}$. Also, $\operatorname{cov}(s_j^{-1}\mathbf{S}_j, s_k^{-1}\mathbf{S}_k) = (ns_js_k)^{-1}\sum_{i=1}^n \operatorname{cov}(\ell'_{\tau_j}(\epsilon_{ij}),\ell'_{\tau_k}(\epsilon_{ik}))\mathbf{I}$ and

$$\left|(ns_js_k)^{-1}\sum_{i=1}^n \operatorname{cov}(\ell'_{\tau_j}(\epsilon_{ij}),\ell'_{\tau_k}(\epsilon_{ik})) - r_{\epsilon,jk}\right| \lesssim \frac{v_{jk}}{\tau_0^{\delta}}\left(\frac{d+t}{n}\right)^{\delta/(2+\delta)} \tag{A.8}$$

by Corollary B.1 and Proposition B.2, where $\tau_0 = \min(\tau_{0j},\tau_{0k})$ and $v_{jk} = \max\{\mathbb{E}(|\epsilon_j|^{2+\delta}),\mathbb{E}(|\epsilon_k|^{2+\delta})\} < \infty$ for some $\delta \in (0,2]$ and $1 \leqslant j \neq k \leqslant p$. Putting together (A.8), Condition 1 (iv),

Corollary B.1, and Lemmas B.2 and B.5 with our choice on τ_j yield

$$
\begin{aligned}
&|\,\mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2}s_j\mathbf{G}_1\| \geqslant x, \|\sigma_{\epsilon,kk}^{-1/2}s_k\mathbf{G}_2\| \geqslant x) - \mathbb{P}(\|\mathbf{Z}_1\| \geqslant x)\,\mathbb{P}(\|\mathbf{Z}_2\| \geqslant x)| \\
\leqslant &|\,\mathbb{P}(\|\mathbf{G}_1\| \geqslant \sigma_{\epsilon,jj}^{1/2}s_j^{-1}x, \|\mathbf{G}_2\| \geqslant \sigma_{\epsilon,kk}^{1/2}s_k^{-1}x) - \mathbb{P}(\|\mathbf{Z}_1\| \geqslant \sigma_{\epsilon,jj}^{1/2}s_j^{-1}x)\,\mathbb{P}(\|\mathbf{Z}_2\| \geqslant \sigma_{\epsilon,kk}^{1/2}s_k^{-1}x)| \\
&+ |\,\mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2}s_j\mathbf{Z}_1\| \geqslant x)\,\mathbb{P}(\|\sigma_{\epsilon,kk}^{-1/2}s_k\mathbf{Z}_2\| \geqslant x) - \mathbb{P}(\|\mathbf{Z}_1\| \geqslant x)\,\mathbb{P}(\|\mathbf{Z}_2\| \geqslant x)| \\
\lesssim &\, q^{1/2}|r_{\epsilon,jk}| + q\{n^{-1}(d+t)\}^{\delta/(2+\delta)}.
\end{aligned}
$$

It follows that

$$
|\,\mathbb{P}(\|\mathbf{G}_1\| \geqslant s_j^{-1}\sigma_{\epsilon,jj}^{1/2}z, \|\mathbf{G}_2\| \geqslant s_k^{-1}\sigma_{\epsilon,kk}^{1/2}z) - \{\mathbb{P}(\|\mathbf{Z}\| \geqslant z)\}^2| \lesssim q^{1/2}|r_{\epsilon,jk}| + q\left(\frac{d+t}{n}\right)^{\delta/(2+\delta)} \tag{A.9}
$$

for $\mathbf{Z} \sim N_q(\mathbf{0}, \mathbf{I})$. In addition, Lemma B.4 gives

$$
\sup_{x,y\in\mathbb{R}} |\,\mathbb{P}(\|s_j^{-1}\mathbf{S}_j\| \geqslant x, \|s_k^{-1}\mathbf{S}_k\| \geqslant y) - \mathbb{P}(\|\mathbf{G}_1\| \geqslant x, \|\mathbf{G}_2\| \geqslant y)| \lesssim n^{-1/2}q^{7/4},
$$

which implies

$$
|\,\mathbb{E}(D_jD_k) - \mathbb{P}(\|\mathbf{G}_1\| > s_j^{-1}\sigma_{\epsilon,jj}^{1/2}z, \|\mathbf{G}_2\| > s_k^{-1}\sigma_{\epsilon,kk}^{1/2}z)| \lesssim n^{-1/2}q^{7/4}. \tag{A.10}
$$

Putting (A.9), (A.10), and Lemma B.2 together, we obtain

$$
|\,\mathbb{E}(D_jD_k) - \{\mathbb{P}(\|\mathbf{Z}_1\| > z)\}^2| \lesssim q^{1/2}|r_{\epsilon,jk}| + n^{-1/2}q^{7/4} + q\left(\frac{d+t}{n}\right)^{\delta/(2+\delta)}. \tag{A.11}
$$

Consequently, it follows from (A.6), (A.7), (A.11), Condition 1, and Lemma B.6 that

$$
\mathbb{E}[\{p_0^{-1}V^\circ(z^2) - \{\mathbb{P}(\chi_q^2 > z^2)\}^2] \lesssim q^{1/2}p^{-\kappa_1} + n^{-1/2}q^{7/4} + q\left(\frac{d+t}{n}\right)^{\delta/(2+\delta)} \tag{A.12}
$$

on $E_1(t)$. Recall that $\mathbb{P}\{E_1(t)\} = 1 - 3p\exp(-t)$ as long as $n \geqslant C_3(d+t)$. Taking $t = \log(np)$ in (A.5) and (A.12) proves (A.1).

S1.2 Proof of Theorem 1

For statistic $\tilde{V}_j = n(\mathbf{C}\hat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})^{\mathrm{T}}(\mathbf{C}\tilde{\boldsymbol{\Sigma}}_j\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\hat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})$, it holds

$$
\begin{aligned}
|\tilde{V}_j - V_j^{\circ}| &= |n(\mathbf{C}\hat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})^{\mathrm{T}}(\mathbf{C}\tilde{\boldsymbol{\Sigma}}_j\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\hat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j}) - n(\mathbf{C}\hat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})^{\mathrm{T}}(\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\hat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})| \\
&\leqslant n\|\boldsymbol{\Sigma}_Z^{1/2}(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)\|^2\|\boldsymbol{\Sigma}_Z^{-1/2}\mathbf{C}^{\mathrm{T}}\{(\mathbf{C}\tilde{\boldsymbol{\Sigma}}_j\mathbf{C}^{\mathrm{T}})^{-1} - (\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\}\mathbf{C}\boldsymbol{\Sigma}_Z^{-1/2}\| \\
&\leqslant n\|\boldsymbol{\Sigma}_Z^{1/2}(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)\|^2\|\mathbf{C}\boldsymbol{\Sigma}_Z^{-1/2}\|^2\|(\mathbf{C}\tilde{\boldsymbol{\Sigma}}_j\mathbf{C}^{\mathrm{T}})^{-1} - (\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\| \\
&\lesssim n\|\boldsymbol{\Sigma}_Z^{1/2}(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)\|^2\|(\mathbf{C}\tilde{\boldsymbol{\Sigma}}_j\mathbf{C}^{\mathrm{T}})^{-1} - (\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\|.
\end{aligned}
$$

Given $\max_{1\leqslant j\leqslant p}\|\tilde{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\| = o_{\mathbb{P}}[\{\log(np) + d\}^{-1}]$, Lemmas B.1 and B.8 imply that

$$
\max_{j\in\mathscr{H}_0}\|\tilde{V}_j - V_j^{\circ}\| \lesssim \{\log(np) + d\}\max_{1\leqslant j\leqslant p}\|\tilde{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\| \tag{A.13}
$$

with probability $1-2n^{-1}$ and the right hand side of (A.13) is $o_{\mathbb{P}}(1)$. Combining this with the proof of Proposition 1 and Condition 1, we obtain $p_0^{-1}\tilde{V}(z) = \mathbb{P}(\chi_q^2 > z) + o_{\mathbb{P}}(1)$. Similarly, $R(z)$ can be replaced by $\tilde{R}(z)$. Consequently, $|\widetilde{\mathrm{FDP}}(z) - \widetilde{\mathrm{AFDP}}(z)| = o_{\mathbb{P}}(1)$ as $n, p \to \infty$.

S1.3 Proof of Theorem 2

Recall that $m_j = n^{-1}\sum_{i=1}^n \mathbb{I}_{\tau_j}^*(e_{ij})$, $\mathbf{W}_j = n^{-1}\sum_{i=1}^n\{\mathbb{I}_{\tau_j}^*(e_{ij})\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}\}$, and $K_j = 1 + (nm_j)^{-1}(d+1)(1-m_j)$. Denote $\mathbf{A}_{jn} := \{\mathbf{W}_j^{-1}(n^{-1}\sum_{i=1}^n \mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}})\mathbf{W}_j^{-1}\}^{-1}$. For each

$j = 1, \ldots, p$,

$$\begin{aligned}\|\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\| &= \left\| \frac{1}{K_j}\left[\frac{1}{n-d-1}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2\right]\mathbf{A}_{jn}^{-1} - \sigma_{\epsilon,jj}{\mathbf{\Sigma}_Z}^{-1}\right\| \\ &\leqslant \left|\frac{1}{K_j}\left[\frac{1}{n-d-1}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2\right]\right| \|\mathbf{A}_{jn}^{-1} - {\mathbf{\Sigma}_Z}^{-1}\| \\ &\quad + \left|\frac{1}{K_j}\left[\frac{1}{n-d-1}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2\right] - \sigma_{\epsilon,jj}\right| \|\mathbf{\Sigma}_Z^{-1}\|.\end{aligned} \tag{A.14}$$

Note $n/(n-d-1) \to 1$ as $n \to \infty$. Denote $K_j^{-1}n^{-1}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2$ by $f(m_j, y_j)$, where $y_j = n^{-1}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2$. That is, $f[m_j, n^{-1}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2] = K_j^{-1}n^{-1}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2$ with $f(1, \sigma_{\epsilon,jj}) = \sigma_{\epsilon,jj}$. It holds that $f(m_j, y_j)$ is twice differentiable at $(1, \sigma_{\epsilon,jj})$, where

$$\frac{\partial}{\partial y_j}f(m_j, y_j) = K_j^{-1},\ \frac{\partial}{\partial y_j}f(m_j, y_j)|_{(m_j,y_j)=(1,\sigma_{\epsilon,jj})} = 1,$$
$$\frac{\partial}{\partial m_j}f(m_j, y_j) = \frac{n(d+1)}{\{nm_j + (d+1)(1-m_j)\}^2}y_j,\ \frac{\partial}{\partial m_j}f(m_j, y_j)|_{(m_j,y_j)=(1,\sigma_{\epsilon,jj})} = \frac{d+1}{n}\sigma_{\epsilon,jj}$$

and

$$\frac{\partial^2}{\partial y_j^2}f(m_j, y_j) = 0,\ \frac{\partial^2}{\partial m_j \partial y_j}f(m_j, y_j) = \frac{n(d+1)}{\{nm_j + (d+1)(1-m_j)\}^2},$$
$$\frac{\partial^2}{\partial m_j^2}f(m_j, y_j) = \frac{-2n(d+1)(n-d-1)}{\{nm_j + (d+1)(1-m_j)\}^3}y_j.$$

Apply Taylor's theorem on $f[m_j, n^{-1}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2]$ with respect to m_j and $n^{-1}\sum\{\ell'_{\tau_j}(e_{ij})\}^2$ at 1 and $\sigma_{\epsilon,jj}$, it follows that

$$\begin{aligned}\frac{1}{K_j}\left[\frac{1}{n}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2\right] - \sigma_{\epsilon,jj} &= \left[\frac{1}{n}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2 - \sigma_{\epsilon,jj}\right] + \frac{d+1}{n}\sigma_{\epsilon,jj}(m_j - 1) \\ &\quad + R_1(1 + h_1, \sigma_{\epsilon,jj}^2 + h_2),\end{aligned}$$

where $R_1(\cdot,\cdot)$ is the remainder and satisfies $\lim_{\mathbf{h}\to\mathbf{0}} R_1(1 + h_1, \sigma_{\epsilon,jj}^2 + h_2)/\|\mathbf{h}\| = 0$ for $\mathbf{h} = (h_1, h_2) = c\{(m_j, y_j) - (1, \sigma_{\epsilon,jj})\}$ and $c \in (0, 1)$ since f is twice differentiable (for

example, see Rudin, 1976). On the event A_Δ defined in Lemma B.9 with $\mathbb{P}(A_\Delta) \geqslant 1 - 4n^{-1}$, Lemmas B.10 and B.11 imply

$$\begin{aligned}
&\max_{1\leqslant j\leqslant p}\left|\frac{1}{K_j}\left[\frac{1}{n}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2\right]-\sigma_{\epsilon,jj}\right| \\
\leqslant & C_5\left\{\frac{\log(np)+d}{n}\right\}^{\delta/(2+\delta)}+C_6\frac{d+1}{n}\max\left[\left\{\frac{\log(np)+d}{n}\right\}^{1/2},\frac{\Delta}{h_n}\right] \\
&+R_1\left[1+h_1,\sigma^2_{\epsilon,jj}+h_2\right] \\
\leqslant & C\left\{\frac{\log(np)+d}{n}\right\}^{\delta/(2+\delta)}
\end{aligned} \tag{A.15}$$

with probability at least $1 - 8n^{-1}$, where C is a constant depending on A_0. Note that the remainder in (A.15) is dominated by other terms as long as $m_j - 1$ and $n^{-1}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2 - \sigma^2_{\epsilon,jj}$ are small given $d \ll n$.

By Lemma B.7, it follows that

$$\begin{aligned}
\|\mathbf{A}_{jn}^{-1}-\mathbf{\Sigma}_Z^{-1}\| &\leqslant \frac{\|\mathbf{\Sigma}_Z^{-1}\|^2}{1-\|\mathbf{\Sigma}_Z^{-1}\|\|\mathbf{A}_{jn}-\mathbf{\Sigma}_Z\|}\|\mathbf{A}_{jn}-\mathbf{\Sigma}_Z\| \\
&= \|\mathbf{\Sigma}_Z^{-1}\|^2\|\mathbf{A}_{jn}-\mathbf{\Sigma}_Z\|\sum_{k=0}^{\infty}(\|\mathbf{\Sigma}_Z^{-1}\|\|\mathbf{A}_{jn}-\mathbf{\Sigma}_Z\|)^k.
\end{aligned} \tag{A.16}$$

Note that $\|\mathbf{\Sigma}_Z^{-1}\|\|\mathbf{A}_{jn}-\mathbf{\Sigma}_Z\| \ll 1$ as long as $\|\mathbf{A}_{jn}-\mathbf{\Sigma}_Z\| \ll \lambda_{\min}(\mathbf{\Sigma}_Z)$. Hence, we only need to focus on $\|\mathbf{A}_{jn}-\mathbf{\Sigma}_Z\|$. Denote the sample covariance of $\mathbf{Z}_i$ by $\hat{\mathbf{\Sigma}}_n = n^{-1}\sum_{i=1}^{n}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}$. Decompose $\mathbf{A}_{jn}$ as

$$\mathbf{A}_{jn} = \mathbf{W}_j\hat{\mathbf{\Sigma}}_n^{-1}\mathbf{W}_j = \hat{\mathbf{\Sigma}}_n + 2(\mathbf{W}_j-\hat{\mathbf{\Sigma}}_n) + (\mathbf{W}_j-\hat{\mathbf{\Sigma}}_n)\hat{\mathbf{\Sigma}}_n^{-1}(\mathbf{W}_j-\hat{\mathbf{\Sigma}}_n)$$

so that

$$\begin{aligned}\mathbf{A}_{jn} - \widehat{\mathbf{\Sigma}}_n &= 2(\mathbf{W}_j - \widehat{\mathbf{\Sigma}}_n) + (\mathbf{W}_j - \widehat{\mathbf{\Sigma}}_n)\widehat{\mathbf{\Sigma}}_n^{-1}(\mathbf{W}_j - \widehat{\mathbf{\Sigma}}_n) \\ &= 2(\mathbf{W}_j - \widehat{\mathbf{\Sigma}}_n) + (\mathbf{W}_j - \widehat{\mathbf{\Sigma}}_n)(\widehat{\mathbf{\Sigma}}_n^{-1} - \mathbf{\Sigma}_Z^{-1})(\mathbf{W}_j - \widehat{\mathbf{\Sigma}}_n) \\ &\quad + (\mathbf{W}_j - \widehat{\mathbf{\Sigma}}_n)\mathbf{\Sigma}_Z^{-1}(\mathbf{W}_j - \widehat{\mathbf{\Sigma}}_n).\end{aligned}$$

Hence, the bound for the operator norm of $\mathbf{A}_{jn} - \widehat{\mathbf{\Sigma}}_n$ is obtained by bounding $\|\mathbf{W}_j - \widehat{\mathbf{\Sigma}}_n\|$ and $\|\widehat{\mathbf{\Sigma}}_n^{-1} - \mathbf{\Sigma}_Z^{-1}\|$. By Lemmas B.7 and B.12,

$$\begin{aligned}\|\mathbf{A}_{jn} - \widehat{\mathbf{\Sigma}}_n\| &\leqslant 2\|\mathbf{W}_j - \widehat{\mathbf{\Sigma}}_n\| + \|\mathbf{W}_j - \widehat{\mathbf{\Sigma}}_n\|^2 \left[\|\widehat{\mathbf{\Sigma}}_n^{-1} - \mathbf{\Sigma}_Z^{-1}\| + \{\lambda_{\min}(\mathbf{\Sigma}_Z)\}^{-1}\right] \\ &\leqslant C \max\left[\left\{\frac{\log(np) + d}{n}\right\}^{1/2}, \frac{\Delta}{h_n}\right],\end{aligned} \tag{A.17}$$

where C is a constant depending on A_0, $\lambda_{\max}(\mathbf{\Sigma}_Z)$, and $v_{j,\delta}$ since the first term on the right hand side of the first inequality dominates the others as long as $n \geqslant C_3\{\log(np) + d\}$. By the triangle inequality, the concentration of sample covariance matrices (Vershynin, 2018, Exercise 4.7.3), and (A.17), it holds that

$$\|\mathbf{A}_{jn} - \mathbf{\Sigma}_Z\| \leqslant \|\mathbf{A}_{jn} - \widehat{\mathbf{\Sigma}}_n\| + \|\widehat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}_Z\| \leqslant C \max\left[\left\{\frac{\log(np) + d}{n}\right\}^{1/2}, \frac{\Delta}{h_n}\right] \tag{A.18}$$

with probability $1 - 4n^{-1}$ as long as $n \geqslant C_3\{\log(np) + d\}$ where C is a constant depending on $\lambda_{\max}(\mathbf{\Sigma}_Z)$, $v_{j,\delta}$, and A_0.

Putting together (A.14)-(A.18), for $\delta \in (0, 2]$, it follows that

$$\max_{1 \leqslant j \leqslant p} \|\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\| \leqslant C \max\left[\left\{\frac{\log(np) + d}{n}\right\}^{\delta/(2+\delta)}, \frac{\Delta}{h_n}\right]$$

with probability at least $1 - 16n^{-1}$ for some positive constant C depending only on $\lambda_{\min}(\mathbf{\Sigma}_Z)$, $\lambda_{\max}(\mathbf{\Sigma}_Z)$, A_0, and $v_{j,\delta}$.

S2 Auxiliary results

Recall that the first order derivative of the Huber loss is

$$\ell'_\tau(x) = \begin{cases} x & |x| \leqslant \tau, \\ \tau\,\mathrm{sgn}(x) & |x| > \tau \end{cases}$$

and its second order derivative is $\ell''_\tau(x) = \mathbb{I}(|x| < \tau)$ when $|x| \neq \tau$.

S2.1 Some auxiliary lemmas

We first state a few auxiliary lemmas. Proposition B.1 is Proposition A.2 from Zhou et al. (2018). It quantifies the difference between the first two moments of $\ell'_\tau(\epsilon_j)$ and ϵ_j given the existence of higher moments of ϵ_j.

Proposition B.1. *Let z be a real-valued random variable with $\mathbb{E}(z) = 0$ and $\sigma^2 = \mathbb{E}(z^2) > 0$. Assume that $\mathbb{E}(|z|^\kappa) < \infty$ for some $\kappa > 2$. Then*

$$|\,\mathbb{E}\,\ell'_\tau(z)| \leqslant \min\left\{\frac{\sigma^2}{\tau}, \frac{\mathbb{E}(|z|^\kappa)}{\tau^{\kappa-1}})\right\} \quad \textit{and} \quad |\,\mathbb{E}\{\ell'_\tau(z)\}^2 - \sigma^2| \leqslant \frac{2\,\mathbb{E}(|z|^\kappa)}{(\kappa-2)\tau^{\kappa-2}}.$$

The following corollary from Zhou et al. (2018) reveals the bias of $s_j^2 = \mathbb{E}\{\ell'_{\tau_j}(\epsilon_{ij})^2\}$ with respect to the true error variance $\sigma_{\epsilon,jj}$. It implies that, with the adaptive robustification parameter τ_j, $s_j^2 \to \sigma_{\epsilon,jj}$ as $n \to \infty$.

Corollary B.1. *For $1 \leqslant j \leqslant p$ and $v_{j,\delta} = \{\mathbb{E}(|\epsilon_j|^{2+\delta})\}^{1/(2+\delta)} < \infty$, it holds that*

$$\sigma_{\epsilon,jj} - \frac{2}{\delta}\frac{v_{j,\delta}^{2+\delta}}{\tau_j^\delta} \leqslant s_j^2 \leqslant \sigma_{\epsilon,jj}.$$

Proof. Applying Proposition B.1 with $\kappa = 2+\delta$ for some $\delta > 0$ yields the first inequality. The second inequality follows $\ell'_{\tau_j}(x)^2 \leqslant x^2$. □

Next, Proposition B.2 implies that the covariance of $\ell'_{\tau_j}(\epsilon_j)$ can be approximated by the covariance of true errors. It is employed to prove the main theorems.

Proposition B.2. *Assume* $\tau = \min(\tau_j, \tau_k)$ *and* $v_{jk} = \max\{\mathbb{E}(|\epsilon_j|^{2+\delta}), \mathbb{E}(|\epsilon_k|^{2+\delta})\} < \infty$ *for* $1 \leqslant j \neq k \leqslant p$ *and* $\delta > 0$. *Then*

$$|\operatorname{cov}\{\ell'_{\tau_j}(\epsilon_j), \ell'_{\tau_k}(\epsilon_k)\} - \operatorname{cov}(\epsilon_j, \epsilon_k)| \lesssim \max(\tau^{-\delta} v_{jk}, \tau^{-2-2\delta} v_{jk}^2).$$

Proof. By definition,

$$\begin{aligned}
\operatorname{cov}(\epsilon_j, \epsilon_k) =& \mathbb{E}\{\epsilon_j\epsilon_k \mathbb{I}(|\epsilon_j| \leqslant \tau_j, |\epsilon_k| \leqslant \tau_k)\} + \mathbb{E}[\epsilon_j\epsilon_k\{\mathbb{I}(|\epsilon_j| > \tau_j) + \mathbb{I}(|\epsilon_k| > \tau_k)\}] \\
&- \mathbb{E}\{\epsilon_j\epsilon_k \mathbb{I}(|\epsilon_j| > \tau_j, |\epsilon_k| > \tau_k)\}, \\
\mathbb{E}\{\ell'_{\tau_j}(\epsilon_j)\} =& \mathbb{E}\{\epsilon_j \mathbb{I}(|\epsilon_j| \leqslant \tau_j)\} + \tau_j \mathbb{E}\{\operatorname{sgn}(\epsilon_j)\mathbb{I}(|\epsilon_j| > \tau_j)\},
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\{\ell'_{\tau_j}(\epsilon_j)\ell'_{\tau_k}(\epsilon_k)\} =& \mathbb{E}\{\epsilon_j\epsilon_k \mathbb{I}(|\epsilon_j| \leqslant \tau_j, |\epsilon_k| \leqslant \tau_k)\} + \tau_k \mathbb{E}\{\epsilon_j \operatorname{sgn}(\epsilon_k)\mathbb{I}(|\epsilon_j| \leqslant \tau_j, |\epsilon_k| > \tau_k)\} \\
&+ \tau_j \mathbb{E}\{\operatorname{sgn}(\epsilon_j)\epsilon_k \mathbb{I}(|\epsilon_j| > \tau_j, |\epsilon_k| \leqslant \tau_k)\} + \tau_j\tau_k \mathbb{E}\{\operatorname{sgn}(\epsilon_j\epsilon_k)\mathbb{I}(|\epsilon_j| > \tau_j, |\epsilon_k| > \tau_k)\} \\
=& \operatorname{cov}(\epsilon_j, \epsilon_k) - \mathbb{E}[\epsilon_j\epsilon_k\{\mathbb{I}(|\epsilon_j| > \tau_j) + \mathbb{I}(|\epsilon_k| > \tau_k)\}] \\
&+ \tau_k \mathbb{E}\{\epsilon_j \operatorname{sgn}(\epsilon_k)\mathbb{I}(|\epsilon_j| \leqslant \tau_j, |\epsilon_k| > \tau_k)\} + \tau_j \mathbb{E}\{\operatorname{sgn}(\epsilon_j)\epsilon_k \mathbb{I}(|\epsilon_j| > \tau, |\epsilon_k| \leqslant \tau_k)\} \\
&+ \mathbb{E}[\{\epsilon_j\epsilon_k + \tau_j\tau_k \operatorname{sgn}(\epsilon_j\epsilon_k)\}\mathbb{I}(|\epsilon_j| > \tau_j, |\epsilon_k| > \tau_k)].
\end{aligned}$$

Note that

$$
\begin{aligned}
|\,\mathbb{E}\{\epsilon_j\epsilon_k\mathbb{I}(|\epsilon_j|>\tau_j\}| &= |\,\mathbb{E}\{\epsilon_j^{1+\delta}\epsilon_k\epsilon_j^{-\delta}\mathbb{I}(|\epsilon_j|>\tau_j)\}| \\
&\leqslant (\mathbb{E}\,|\epsilon_j^{1+\delta}\epsilon_k|)\max|\epsilon_j^{-\delta}\mathbb{I}(|\epsilon_j|>\tau_j)| \\
&\leqslant \tau_j^{-\delta}\{\mathbb{E}(|\epsilon_j|^{2+\delta})\}^{(1+\delta)/(2+\delta)}\{\mathbb{E}(|\epsilon_k|^{2+\delta})\}^{1/(2+\delta)} \\
&\leqslant \tau_j^{-\delta}v_{jk} \leqslant \tau^{-\delta}v_{jk}.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
|\,\mathbb{E}\{\epsilon_j\epsilon_k\mathbb{I}(|\epsilon_j|>\tau_j,|\epsilon_k|>\tau_k\}| &= |\,\mathbb{E}\{\epsilon_j^{1+\delta/2}\epsilon_k^{1+\delta/2}\epsilon_j^{-\delta/2}\epsilon_k^{-\delta/2}\mathbb{I}(|\epsilon_j|>\tau_j,|\epsilon_k|>\tau_k)\}| \\
&\leqslant \mathbb{E}(|\epsilon_j^{1+\delta/2}\epsilon_k^{1+\delta/2}|)\max|\epsilon_j^{-\delta/2}\epsilon_k^{-\delta/2}\mathbb{I}(|\epsilon_j|>\tau_j,|\epsilon_k|>\tau_k)| \\
&\leqslant \tau_j^{-\delta/2}\tau_k^{-\delta/2}\{\mathbb{E}(|\epsilon_j|^{2+\delta})\}^{(1+\delta)/(2+\delta)}\{\mathbb{E}(|\epsilon_k|^{2+\delta})\}^{1/(2+\delta)} \\
&\leqslant \tau_j^{-\delta/2}\tau_k^{-\delta/2}v_{jk} \leqslant \tau^{-\delta}v_{jk}.
\end{aligned}
$$

$$
\begin{aligned}
|\,\mathbb{E}\{\epsilon_j\mathbb{I}(|\epsilon_j|\leqslant\tau_j,|\epsilon_k|>\tau_k)\}| &= |\,\mathbb{E}\{\epsilon_j\epsilon_k^{1+\delta}\epsilon_k^{-1-\delta}\mathbb{I}(|\epsilon_j|\leqslant\tau_j,|\epsilon_k|>\tau_k\}| \\
&\leqslant \mathbb{E}(|\epsilon_j\epsilon_k^{1+\delta}|)\max|\epsilon_k^{-1-\delta}\mathbb{I}(|\epsilon_k|>\tau_k)| \\
&\leqslant \tau_k^{-1-\delta}v_{jk} \leqslant \tau^{-1-\delta}v_{jk},
\end{aligned}
$$

and

$$
\begin{aligned}
|\,\mathbb{E}\{\mathbb{I}(|\epsilon_j|>\tau_j,|\epsilon_k|>\tau_k\}| &= |\,\mathbb{E}\{\epsilon_j^{1+\delta/2}\epsilon_k^{1+\delta/2}\epsilon_j^{-1-\delta/2}\epsilon_k^{-1-\delta/2}\mathbb{I}(|\epsilon_j|>\tau_j,|\epsilon_k|>\tau_k)\}| \\
&\leqslant \mathbb{E}(|\epsilon_j^{1+\delta/2}\epsilon_k^{1+\delta/2}|)\max|\epsilon_j^{-1-\delta/2}\epsilon_k^{-1-\delta/2}\mathbb{I}(|\epsilon_j|>\tau_j,|\epsilon_k|>\tau_k)| \\
&\leqslant \tau_j^{-1-\delta/2}\tau_k^{-1-\delta/2}v_{jk} \leqslant \tau^{-2-\delta}v_{jk}.
\end{aligned}
$$

Therefore, we have

$$
|\,\mathbb{E}\{\ell'_{\tau_j}(\epsilon_j)\ell'_{\tau_k}(\epsilon_k)\} - \mathrm{cov}(\epsilon_j,\epsilon_k)| \lesssim \tau^{-\delta}v_{jk}
$$

as $\max_{1\leqslant j\leqslant p} v_{j,\delta} \leqslant C_\epsilon$. Hence, it yields

$$\begin{aligned}|\operatorname{cov}\{\ell'_{\tau_j}(\epsilon_j), \ell'_{\tau_k}(\epsilon_k)\} - \operatorname{cov}(\epsilon_j, \epsilon_k)| &\leqslant |\mathbb{E}\{\ell'_{\tau_j}(\epsilon_j)\ell'_{\tau_k}(\epsilon_k)\} - \operatorname{cov}(\epsilon_j, \epsilon_k)| + |\mathbb{E}\{\ell'_{\tau_j}(\epsilon_j)\}\,\mathbb{E}\{\ell'_{\tau_k}(\epsilon_k)\}| \\ &\lesssim \tau^{-\delta} v_{jk} + \min\left\{\frac{\sigma_{\epsilon,jj}}{\tau_j}, \frac{\mathbb{E}(|\epsilon_j|^{2+\delta})}{\tau_j^{1+\delta}}\right\} \min\left\{\frac{\sigma_{\epsilon,kk}}{\tau_k}, \frac{\mathbb{E}(|\epsilon_k|^{2+\delta})}{\tau_k^{1+\delta}}\right\} \\ &\lesssim \max(\tau^{-\delta} v_{jk}, \tau^{-2-2\delta} v_{jk}^2)\end{aligned}$$

by Proposition B.1. □

The following result on the non-asymptotic bound for the adaptive Huber regression estimator is borrowed from Theorem 7 in Sun et al. (2020). It provides an exponential-type concentration inequalities for $\widehat{\boldsymbol{\theta}}_j$'s with adaptive robustification parameter τ_j, and also gives a non-asymptotic Bahadur representation under the finite moment condition on the errors.

Lemma B.1. *Under Condition 2, for any $t > 0$, $\tau_{0j} \geqslant v_{j,\delta} := (\mathbb{E}\,|\epsilon|^{2+\delta})^{1/(2+\delta)}$, $j = 1, \ldots, p$, $\widehat{\boldsymbol{\theta}}_j$ with $\tau_j = \tau_{0j}\{n(d+t)^{-1}\}^{1/(2+\delta)}$ satisfies*

$$\mathbb{P}[\|\boldsymbol{\Sigma}_Z^{1/2}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)\| \geqslant C_1 \tau_{0j}\{n^{-1}(d+t)\}^{1/2}] \leqslant 2e^{-t}$$

and

$$\mathbb{P}[\|\boldsymbol{\Sigma}_Z^{1/2}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j) - \frac{1}{n}\sum_{i=1}^n \{\ell'_\tau(\epsilon_i)\boldsymbol{\Sigma}_Z^{-1/2}\mathbf{Z}_i\}\| \geqslant C_2 \tau_{0j} n^{-1}(d+t)] \leqslant 3e^{-t}$$

as long as $n \geqslant C_3(d+t)$, where C_1–C_3 are positive constants depending only on A_0 from Condition 2.

S2.2 Technical results for proving Proposition 1

From Lemma B.1, we expect the following approximation of our testing statistics.

Corollary B.2. *For* $\mathbf{T}_j^\circ$, $\mathbf{S}_j$, *and* $\mathbf{R}_j$ *in* (A.2) *and* (A.3), *respectively, it holds*

$$\|\mathbf{T}_j^\circ - \sigma_{\epsilon,jj}^{-1/2}(\mathbf{S}_j + \mathbf{R}_j)\| \leqslant C_2 \tau_{0j} \frac{d+t}{(n\sigma_{\epsilon,jj})^{1/2}}$$

with probability at least $1 - 2\exp(-t)$ *under the random design.*

Proof. By Lemma B.1,

$$\begin{aligned}
&\|\mathbf{T}_j^\circ - \sigma_{\epsilon,jj}^{-1/2}(\mathbf{S}_j + \mathbf{R}_j)\| \\
&\leqslant (n\sigma_{\epsilon,jj}{}^{-1})^{1/2} \left\| (\mathbf{C}\boldsymbol{\Sigma}_Z^{-1}\mathbf{C}^{\mathrm{T}})^{-1/2}\mathbf{C}\boldsymbol{\Sigma}_Z^{-1/2} \left\{ \boldsymbol{\Sigma}_Z^{1/2}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j) - \frac{1}{n}\sum_{i=1}^n \ell_\tau'(\epsilon_{ij})\boldsymbol{\Sigma}_Z^{-1/2}\mathbf{Z}_i \right\} \right\| \qquad \text{(B.1)} \\
&\leqslant C_2 \tau_{0j} \frac{d+t}{(n\sigma_{\epsilon,jj})^{1/2}}
\end{aligned}$$

with probability at least $1 - 3\exp(-t)$. □

The following results show that the distribution of the Bahadur representation in (B.1) is close to $N(\mathbf{0}, \mathbf{I})$. We decompose $|\,\mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{S}_j\| \geqslant x) - \mathbb{P}(\|\mathbf{G}\| \geqslant x)|$ into two parts. Lemma B.2 quantifies the difference between the cumulative distribution functions of $\|\sigma_{\epsilon,jj}^{-1/2}s_j\mathbf{G}\|$ and $\|\mathbf{G}\|$, and Lemma B.3 characterizes the difference between the cumulative distribution functions of $\|\sigma_{\epsilon,jj}^{-1/2}s_j\mathbf{G}\|$ and $\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{S}_j\|$.

Lemma B.2. *Let* $\mathbf{G} \sim N(\mathbf{0}, \mathbf{I}) \in \mathbb{R}^q$. *Let* $\tau_j = \tau_{0j}\{n(d+t)^{-1}\}^{1/(2+\delta)}$ *for some* $\delta > 0$ *where* $\tau_{0j} \geqslant v_{j,\delta}$. *Then, it holds that*

$$\sup_{x\in\mathbb{R}^+} \left| \mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2}s_j\mathbf{G}\| \geqslant x) - \mathbb{P}(\|\mathbf{G}\| \geqslant x) \right| \leqslant q^{1/2} \frac{2}{\delta\sigma_{\epsilon,jj}} \frac{v_{j,\delta}^{2+\delta}}{\tau_{0j}^{\delta}} \left(\frac{d+t}{n}\right)^{\delta/(2+\delta)}.$$

Proof. It holds

$$\begin{aligned} \|\sigma_{\epsilon,jj}^{-1}s_j^2\mathbf{I}-\mathbf{I}\| &= |\sigma_{\epsilon,jj}^{-1}s_j^2-1| \leqslant \frac{2}{\delta\sigma_{\epsilon,jj}}\frac{v_{j,\delta}^{2+\delta}}{\tau_j^{\delta}} \\ \mathrm{tr}\{(\sigma_{\epsilon,jj}^{-1/2}s_j\mathbf{I}-\mathbf{I})^2\} &\leqslant q\left(\frac{2}{\delta\sigma_{\epsilon,jj}}\frac{v_{j,\delta}^{2+\delta}}{\tau_j^{\delta}}\right)^2. \end{aligned} \tag{B.2}$$

With $\tau_j = \tau_{0j}\{n(d+t)^{-1}\}^{1/(2+\delta)}$, (B.2) satisfies the conditions of Lemma A.7 in the supplement of Spokoiny and Zhilova (2015) whenever $n \geqslant C_3(d+t)$. Combining Corollary B.1 with Lemma A.7 in the supplement of Spokoiny and Zhilova (2015), we get the desired result. □

Lemma B.3. *Let* $\mathbf{G} \sim N\mathbf{0},\mathbf{I}) \in \mathbb{R}^q$.

$$\sup_{x\in\mathbb{R}^+}\left|\mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{S}_j\| \geqslant x) - \mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2}s_j\mathbf{G}\| \geqslant x)\right| \lesssim n^{-1/2}q^{7/4}. \tag{B.3}$$

Proof. Denote $\mathscr{C}$ the class of convex subsets of $\mathbb{R}^q$. Recall that $\mathrm{cov}(\mathbf{S}_j) = \sigma_{\epsilon,jj}^{-1}s_j^2\mathbf{I}$ for $\mathbf{S}_j$ in (A.3). By Theorem 1.1 in Bentkus (2005),

$$\begin{aligned} \sup_{A\in\mathscr{C}}|\,\mathbb{P}(\sigma_{\epsilon,jj}^{-1/2}\mathbf{S}_j\in A) - \mathbb{P}(\sigma_{\epsilon,jj}^{-1/2}s_j\mathbf{G}\in A)| &\lesssim q^{1/4}\sum_{i=1}^n \mathbb{E}\,\|n^{-1/2}\ell'_{\tau_j}(\epsilon_{ij})\mathbf{A}_j^{-1/2}\mathbf{C}\mathbf{\Sigma}_Z^{-1/2}\widetilde{\mathbf{Z}}_i\|^3 \\ &= \frac{q^{1/4}}{n^{1/2}}\frac{\mathbb{E}\{|\ell'_{\tau_j}(\epsilon_{ij})|^3\}}{\sigma_{\epsilon,jj}}\,\mathbb{E}\,\|\mathbf{A}_j^{-1/2}\mathbf{C}\mathbf{\Sigma}_Z^{-1/2}\widetilde{\mathbf{Z}}_i\|^3 \\ &\lesssim \frac{q^{7/4}}{n^{1/2}}. \end{aligned}$$

Take $A = \{\mathbf{v}\in\mathbb{R}^{d+1} : \|\mathbf{v}\| \leqslant x, x>0\}$, we obtain (B.3). □

The following coupling result compares $\mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{S}_j\| > x, \|\sigma_{\epsilon,kk}^{-1/2}\mathbf{S}_k\| > y)$ and its Gaussian counterpart for each (j,k) pair.

Lemma B.4. *Assume Condition 1 holds. Let* $\mathbf{G} = (\mathbf{G}_1,\mathbf{G}_2) \in \mathbb{R}^{2q}$ *be a Gaussian vector with* $\mathbb{E}(\mathbf{G}) = \mathbf{0}$ *and* $\mathrm{cov}(\mathbf{G}) = \mathrm{cov}(\mathbf{S})$*, which is given in Section S1. It satisfies*

that

$$\sup_{x,y\in\mathbb{R}} |\,\mathbb{P}(\|s_j^{-1}\mathbf{S}_j\| > x, \|s_k^{-1}\mathbf{S}_k\| > y) - \mathbb{P}(\|\mathbf{G}_1\| > x, \|\mathbf{G}_2\| > y)| \lesssim n^{-1/2}q^{7/4}.$$

Proof. Notice that

$$\begin{aligned}\mathbb{P}(\|s_j^{-1}\mathbf{S}_j\| > x, \|s_k^{-1}\mathbf{S}_k\| > y) = 1 - \mathbb{P}(\|s_j^{-1}\mathbf{S}_j\| \leqslant x) - \mathbb{P}(\|s_k^{-1}\mathbf{S}_k\| \leqslant y)\\ + \mathbb{P}(\|s_j^{-1}\mathbf{S}_j\| \leqslant x, \|s_k^{-1}\mathbf{S}_k\| \leqslant y)\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(\|\mathbf{G}_1\| > x, \|\mathbf{G}_2\| > y) = 1 - \mathbb{P}(\|\mathbf{G}_1\| \leqslant x) - \mathbb{P}(\|\mathbf{G}_2\| \leqslant y)\\ + \mathbb{P}(\|\mathbf{G}_1\| \leqslant x, \|\mathbf{G}_2\| \leqslant y).\end{aligned}$$

Take $A(x,y) = \{\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2)^{\mathrm{T}} \in \mathbb{R}^{2q}, \mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^q : \|\mathbf{v}_1\| \leqslant x \text{ and } \|\mathbf{v}_2\| \leqslant y \text{ and } x, y \in \mathbb{R}^+ \bigcup \{\infty\}\}$ in Theorem 1.1 in Bentkus (2005), we have

$$\sup_{x,y} |\,\mathbb{P}(\|s_j^{-1}\mathbf{S}_j\| > x, \|s_k^{-1}\mathbf{S}_k\| > y) - \mathbb{P}(\|\mathbf{G}_1\| > x, \|\mathbf{G}_2\| > y)| \lesssim \frac{q^{7/4}}{n^{1/2}},$$

which is the desired result. □

Lemma B.5 below provides a coupling between multivariate normal distributions.

Lemma B.5. *Let* $\mathbf{G} = (\mathbf{G}_1, \mathbf{G}_2) \in \mathbb{R}^{2q}$ *be a Gaussian vector with* $\mathbb{E}(\mathbf{G}) = \mathbf{0}$, $\mathrm{cov}(\mathbf{G}_i) = \mathbf{I}$ *for* $i = 1, 2$, *and* $\mathrm{corr}(\mathbf{G}_1, \mathbf{G}_2) = r\mathbf{I}$ *where* $|r| \leqslant k_0 < 1$. *Let* $\mathbf{Z}_1, \mathbf{Z}_2 \sim N(\mathbf{0}, \mathbf{I})$ *be independent and identically distributed* q*-dimensional standard normal vectors. Then*

$$\left|\mathbb{P}(\|\mathbf{G}_1\| > z, \|\mathbf{G}_2\| > z) - \mathbb{P}(\|\mathbf{Z}_1\| > z)^2\right| \leqslant C_q|r|$$

for some constant $C_q > 0$ *only depending on* q.

Proof. For $q = 1$, it holds

$$\begin{aligned}
&\mathbb{P}(|G_1| > x, |G_2| > x) - \mathbb{P}(|Z_1| > x)\,\mathbb{P}(|Z_2| > x) \\
&= \{\mathbb{P}(G_1 < -x, G_2 < -x) - \mathbb{P}(Z_1 < -x)\,\mathbb{P}(Z_2 < -x)\} \\
&\quad - \{\mathbb{P}(G_1 < -x, G_2 < x) - \mathbb{P}(Z_1 < -x)\,\mathbb{P}(Z_2 < x)\} \\
&\quad - \{\mathbb{P}(G_1 < x, G_2 < -x) - \mathbb{P}(Z_1 < x)\,\mathbb{P}(Z_2 < -x)\} \\
&\quad + \{\mathbb{P}(G_1 < x, G_2 < x) - \mathbb{P}(Z_1 < x)\,\mathbb{P}(Z_2 < x)\},
\end{aligned}$$

so that $|\,\mathbb{P}(|G_1| > x, |G_2| > x) - \mathbb{P}(|Z_1| > x)\,\mathbb{P}(|Z_2| > x)| \leqslant |r|$ follows Corollary 2.1 from Li and Shao (2002). Set $C_1 = 1$.

For $q \geqslant 2$, let $\mathbf{\Sigma}_1 = \mathbf{I}$ and $\mathbf{\Sigma}_2$ be the covariance matrices of $(\mathbf{Z}_1, \mathbf{Z}_2)$ and $(\mathbf{G}_1, \mathbf{G}_2)$, respectively. By the definition of the total variation distance and Theorem 1.1 in Devroye, Mehrabian, and Reddad (2018), it follows that

$$\begin{aligned}
&\left|\mathbb{P}(\|\mathbf{G}_1\| > z, \|\mathbf{G}_2\| > z) - \mathbb{P}(\|\mathbf{Z}_1\| > z)^2\right| \\
&\leqslant \sup_{A:\text{measurable sets}} |\mathbb{P}\{(\mathbf{G}_1, \mathbf{G}_2) \in A\} - \mathbb{P}\{(\mathbf{Z}_1, \mathbf{Z}_2) \in A\}| \\
&\leqslant \frac{3}{2}\|\mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2} - \mathbf{I}\|_F \\
&= \frac{3}{2}\left\|\begin{bmatrix} \mathbf{0} & r\mathbf{I} \\ r\mathbf{I} & \mathbf{0} \end{bmatrix}\right\|_F = \frac{3}{2}(2q)^{1/2}|r| =: C_q|r| \lesssim q^{1/2}|r|
\end{aligned}$$

where $\|\cdot\|_F$ is the Frobenius norm. □

The following lemma characterizes the distribution of quadratic form of the multivariate normal distribution, which appears in many textbooks (for example, van der Vaart (2000)).

Lemma B.6. *Consider* $\mathbf{X} \sim N_k(\mathbf{0}, \mathbf{\Sigma})$. *Let* $\lambda_1, \ldots, \lambda_k$ *be the eigenvalues of* $\mathbf{\Sigma}$. *Then* $\|\mathbf{X}\|^2$ *is distributed as* $\sum_{i=1}^k \lambda_i Z_i^2$ *for independent and identically distributed random variables* $Z_i \sim N(0,1)$.

Proof. There exists an orthogonal matrix $\mathbf{A}$ such that $\mathbf{A}\mathbf{\Sigma}\mathbf{A}^{\mathrm{T}} = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$. Then $\mathbf{AX}$ is $N_k(\mathbf{0}, \mathrm{diag}(\lambda_1, \ldots, \lambda_k))$. Hence, $\|\mathbf{AX}\|^2 = \|\mathbf{X}\|^2$ implies that $\|\mathbf{X}\|^2$ has the same distribution as $\sum_{i=1}^k \lambda_i Z_i^2$. □

S2.3 Technical results for Section 3.2

Lemma B.7 implies that the operator norm of the difference between two inverse matrices is bounded by a non-decreasing function ($f(x) = x/(1-x)$) of the operator norm of the difference between two matrices. It appears in many linear algebra textbooks (for example, see Horn and Johnson (1990), Chapter 5.8), and is used in the proof of Theorem 1.

Lemma B.7. *For* $d \times d$ *invertible matrices* $\mathbf{A}$ *and* $\mathbf{B} = \mathbf{A} + \Delta\mathbf{A}$ *such that* $\rho(\mathbf{A}^{-1}\Delta\mathbf{A}) < 1$ *where* $\rho(\cdot)$ *is the spectral radius, it follows that*

$$\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\| \leqslant \frac{\|\mathbf{A}^{-1}\Delta\mathbf{A}\|}{1 - \|\mathbf{A}^{-1}\Delta\mathbf{A}\|}\|\mathbf{A}^{-1}\| \leqslant \frac{\|\mathbf{A}^{-1}\|^2}{1 - \|\mathbf{A}^{-1}\Delta\mathbf{A}\|}\|\Delta\mathbf{A}\|. \tag{B.4}$$

Moreover, if $\|\mathbf{A}^{-1}\|\|\Delta\mathbf{A}\| < 1$, *it follows that*

$$\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\| \leqslant \frac{\|\mathbf{A}^{-1}\|}{1 - \|\mathbf{A}^{-1}\|\|\Delta\mathbf{A}\|}\|\mathbf{A}^{-1}\|\|\Delta\mathbf{A}\|. \tag{B.5}$$

Proof. Notice that $\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1} = \mathbf{A}^{-1}\Delta\mathbf{A}\mathbf{B}^{-1}$, so

$$\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\| = \|\mathbf{A}^{-1}\Delta\mathbf{A}\mathbf{B}^{-1}\| \leqslant \|\mathbf{A}^{-1}\Delta\mathbf{A}\|\|\mathbf{B}^{-1}\|.$$

Since $\|\mathbf{B}^{-1}\| \leqslant \|\mathbf{A}^{-1}\| + \|\mathbf{A}^{-1}\Delta\mathbf{A}\|\|\mathbf{B}^{-1}\|$, it holds

$$\|\mathbf{B}^{-1}\| \leqslant \frac{\|\mathbf{A}^{-1}\|}{1 - \|\mathbf{A}^{-1}\Delta\mathbf{A}\|}.$$

Hence, we prove the first part of (B.4). The second inequality of (B.4) is straightforward from the first part since $\|\mathbf{A}^{-1}\Delta\mathbf{A}\| \leqslant \|\mathbf{A}^{-1}\|\|\Delta\mathbf{A}\|$. Inequality (B.5) is a direct result from (B.4) and $\|\mathbf{A}^{-1}\Delta\mathbf{A}\| \leqslant \|\mathbf{A}^{-1}\|\|\Delta\mathbf{A}\|$. □

Lemma B.8. *Under the conditions in Theorem 1,*

$$\max_{1\leqslant j\leqslant p} \|(\mathbf{C}\widetilde{\mathbf{\Sigma}}_j\mathbf{C}^{\mathrm{T}})^{-1} - (\mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\| \lesssim \max_{1\leqslant j\leqslant p} \|\widetilde{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\|.$$

Proof. Without loss of generality, as $\mathrm{rank}(\mathbf{C}) = q$, we assume $\mathbf{C}\mathbf{C}^{\mathrm{T}} = \mathbf{I}_{q\times q}$ by considering $(\mathbf{C}\mathbf{C}^{\mathrm{T}})^{-1/2}\mathbf{C}$ in (2.2). It follows that

$$\begin{aligned}
\|(\mathbf{C}\mathbf{\Sigma}\mathbf{C}^{\mathrm{T}})^{-1}\|\|(\mathbf{C}\widetilde{\mathbf{\Sigma}}_j\mathbf{C}^{\mathrm{T}} - \mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}})\| &\leqslant \|(\mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\|\|\widetilde{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\|\|\mathbf{C}\mathbf{C}^{\mathrm{T}}\| \\
&\leqslant \frac{\|\widetilde{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\|}{\lambda_{\min}(\mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}})} \cdot 1 \\
&\leqslant \frac{\|\widetilde{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\|}{\lambda_{\min}(\mathbf{\Sigma}_j)} \\
&\leqslant \frac{\|\widetilde{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\|}{\sigma_{\epsilon,jj}}\|\mathbf{\Sigma}_Z\| \\
&\leqslant \frac{\|\widetilde{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\|}{c_\epsilon^2}\|\mathbf{\Sigma}_Z\|_F \\
&\lesssim \|\widetilde{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\| d^{1/2},
\end{aligned}$$

where $\|\cdot\|_F$ is the Frobenius norm. The last line in the above inequality is less than 1 as long as $\|\widetilde{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\| \lesssim d^{-1/2}$, which holds for sufficiently large n and p by the assumptions in Theorem 1. Therefore, (B.5) in Lemma B.7 holds.

By the assumptions in Theorem 1,

$$\|\mathbf{C}\widetilde{\mathbf{\Sigma}}_j\mathbf{C}^{\mathrm{T}} - \mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}}\| \leqslant \|\widetilde{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\|. \tag{B.6}$$

Set $\mathbf{A} = \mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}}$, $\mathbf{B} = \mathbf{C}\widetilde{\mathbf{\Sigma}}_j\mathbf{C}^{\mathrm{T}}$, and $\Delta\mathbf{A} = \mathbf{C}\widetilde{\mathbf{\Sigma}}_j\mathbf{C}^{\mathrm{T}} - \mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}}$. Note that $\|(\mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\| = 1/\lambda_{\min}(\mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}})$ is bounded. Putting Lemma B.7 and (B.6) together,

$$\begin{aligned}
\|(\mathbf{C}\widetilde{\mathbf{\Sigma}}_j\mathbf{C}^{\mathrm{T}})^{-1} - (\mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\| &\leqslant \frac{\|(\mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\|^2\|\mathbf{C}\widetilde{\mathbf{\Sigma}}_j\mathbf{C}^{\mathrm{T}} - \mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}}\|}{1 - \|(\mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\widetilde{\mathbf{\Sigma}}_j\mathbf{C}^{\mathrm{T}} - \mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}})\|} \\
&\leqslant \frac{\|(\mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\|^2}{1 - \|(\mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\|\|(\mathbf{C}\widetilde{\mathbf{\Sigma}}_j\mathbf{C}^{\mathrm{T}} - \mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}})\|}\|\mathbf{C}\widetilde{\mathbf{\Sigma}}_j\mathbf{C}^{\mathrm{T}} - \mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}}\| \\
&\leqslant \frac{\|(\mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\|^2}{1 - \|(\mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}\|\|\widetilde{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\|}\|\widetilde{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\|,
\end{aligned}$$

so that the desired result is obtained. □

The following results characterize the non-asymptotic bounds for the proposed covariance estimators. We provide proofs under the random design only as those under the fixed design are almost identical. First, we study an event on which the maximum of $|e_{ij} - \epsilon_{ij}|$ over all $i = 1, \ldots, n$ and $j = 1, \ldots, p$ is small with an overwhelming probability.

Lemma B.9. *Let $\tau_j = \tau_{0j}[n/\{\log(np) + d\}]^{1/(2+\delta)}$ and $\tau_{0j} \geqslant v_{j,\delta}$ with $\delta \in (0, 2]$. Under Conditions 2,*

$$\max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p} |e_{ij} - \epsilon_{ij}| \leqslant C_4 \left\{d^{1/2} + (2\log n)^{1/2}\right\} \left\{\frac{\log(np) + d}{n}\right\}^{1/2}$$

with probability at least $1 - 4n^{-1}$, where C_4 is a constant depending on $v_{j,\delta}$ and A_0 only.

Proof. For each i and j, it follows that

$$\begin{aligned}\max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}|e_{ij}-\epsilon_{ij}| &= \max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}|Z_i^{\mathrm{T}}(\widehat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j)| \\ &\leqslant \max_{1\leqslant i\leqslant n}\|\widetilde{\mathbf{Z}}_i\|\left\{\max_{1\leqslant j\leqslant p}\|\boldsymbol{\Sigma}_Z^{1/2}(\widehat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j)\|\right\} \\ &\leqslant C_4(d^{1/2}+t^{1/2})\left(\frac{d+s}{n}\right)^{1/2}\end{aligned} \tag{B.7}$$

with probability at least $1-2n\exp(-t)-2p\exp(-s)$ for $t>0$ and $s>0$, where C_4 is a constant depending on $v_{j,\delta}$ and A_0. The third inequality above holds by Theorem 3.1 in Vershynin (2018) and Lemma B.1. Let $t=\log(n^2)$ and $s=\log(np)$, (B.7) implies that

$$\max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}|e_{ij}-\epsilon_{ij}| \leqslant C_4\left\{d^{1/2}+(2\log n)^{1/2}\right\}\left\{\frac{\log(np)+d}{n}\right\}^{1/2}$$

with probability at least $1-4n^{-1}$. □

Denote the event in Lemma B.9 by A_Δ, where $\Delta = C_4\left\{d^{1/2}+(2\log n)^{1/2}\right\}[n^{-1}\{\log(np)+d\}]^{1/2}$. Next, we derive the non-asymptotic bound of $\{\ell'_{\tau_j}(\epsilon_{ij})\}^2$.

Lemma B.10. *Let $\tau_j=\tau_{0j}[n\{\log(np)+d\}^{-1}]^{1/(2+\delta)}$, where $\tau_{0j}\geqslant v_{j,\delta}$ for $\delta\in(0,2]$. On the event A_Δ,*

$$\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=1}^n\{\ell'_{\tau_j}(e_{ij})\}^2-\sigma_{\epsilon,jj}\right| \leqslant C_5\left\{\frac{\log(np)+d}{n}\right\}^{\delta/(2+\delta)}$$

holds with probability at least $1-6n^{-1}$, where C_5 is a constant depending on A_0, $v_{j,\delta}$, and v_j.

Proof. First, by the triangle inequality,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2-\sigma_{\epsilon,jj}\right|\leqslant\left|\frac{1}{n}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2-\{\ell'_{\tau_j}(\epsilon_{ij})\}^2\right| + \left|\frac{1}{n}\sum_{i=1}^{n}\{\ell'_{\tau_j}(\epsilon_{ij})\}^2-s_j^2\right|+|s_j^2-\sigma_{\epsilon,jj}|. \tag{B.8}$$

The last term on the right hand side of (B.8) is bounded using Corollary B.1. That is, $\sigma_{\epsilon,jj}-s_j^2\leqslant 2\delta^{-1}\tau_j^{-\delta}v_{j,\delta}^{2+\delta}$.

For the first term on the right hand side of (B.8), it holds that

$$\frac{1}{n}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2=\frac{1}{n}\sum_{i=1}^{n}\{\ell'_{\tau_j}(\epsilon_{ij})\}^2+\frac{2}{n}\sum_{i=1}^{n}\{\epsilon_{ij}\mathbb{I}(|\epsilon_{ij}|\leqslant\tau_j)\}(e_{ij}-\epsilon_{ij}) + \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(|\epsilon_{ij}|\leqslant\tau_j)(e_{ij}-\epsilon_{ij})^2+\frac{1}{n}\sum_{i=1}^{n}R_{2j}(e_{ij})$$

for $\epsilon_{ij}\neq\pm\tau_j$ and $i=1,\ldots,n$ by Taylor's theorem with the Peano form of remainder

$$R_{2j}(x)=\{\ell'_{\tau_j}(x)\}^2-\{\ell'_{\tau_j}(\epsilon_{ij})\}^2-2\epsilon_{ij}\mathbb{I}(|\epsilon_{ij}|\leqslant\tau_j)(x-\epsilon_{ij})-\mathbb{I}(|\epsilon_{ij}|\leqslant\tau_j)(x-\epsilon_{ij})^2,$$

where $\lim_{x\to\epsilon_{ij}}\{R_{2j}(x)/(x-\epsilon_{ij})^2\}=0$. Thus, on the event A_Δ, the remainder is $o_{\mathbb{P}}(\Delta^2)$ and dominated by other terms. Then, by Hoeffding's inequality and Proposition B.1, on the event A_Δ,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2-\frac{1}{n}\sum_{i=1}^{n}\{\ell'_{\tau_j}(\epsilon_{ij})\}^2\right|\\
\leqslant&\left|\frac{2}{n}\sum_{i=1}^{n}\{\epsilon_{ij}\mathbb{I}(|\epsilon_{ij}|\leqslant\tau_j)\}(e_{ij}-\epsilon_{ij})\right|+\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(|\epsilon_{ij}|\leqslant\tau_j)(e_{ij}-\epsilon_{ij})^2\right|+\left|\frac{1}{n}\sum_{i=1}^{n}R_{2j}(e_{ij})\right|\\
\leqslant&2\Delta\left|\frac{1}{n}\sum_{i=1}^{n}\{\epsilon_{ij}\mathbb{I}(|\epsilon_{ij}|\leqslant\tau_j)\}\right|+\Delta^2\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(|\epsilon_{ij}|\leqslant\tau_j)\right|+\left|\frac{1}{n}\sum_{i=1}^{n}R_{2j}(e_{ij})\right|\\
\leqslant&2\Delta\left[|\mathbb{E}\{\epsilon_{ij}\mathbb{I}(|\epsilon_{ij}|\leqslant\tau_j)\}|+\tau_j\left\{\frac{\log(np)}{2n}\right\}^{1/2}\right]
\end{aligned} \tag{B.9}$$

$$
\begin{aligned}
&+\Delta^2\left[\mathbb{P}(|\epsilon_{ij}|\leqslant\tau_j)+\left\{\frac{\log(np)}{2n}\right\}^{1/2}\right]+\left|\frac{1}{n}\sum_{i=1}^n R_{2j}(e_{ij})\right| \\
\leqslant&2\Delta\left[\left|\mathbb{E}\{\ell'_{\tau_j}(\epsilon_{ij})\}\right|+\tau_j\left\{\frac{\log(np)}{2n}\right\}^{1/2}\right]+\Delta^2\left[\mathbb{P}(|\epsilon_{ij}|\leqslant\tau_j)+\left\{\frac{\log(np)}{2n}\right\}^{1/2}\right] \\
&+\left|\frac{1}{n}\sum_{i=1}^n R_{2j}(e_{ij})\right| \qquad\qquad (B.10)\\
\leqslant&C\{d^{1/2}+(2\log n)^{1/2}\}\left\{\frac{\log(np)+d}{n}\right\}^{(1+\delta)/(2+\delta)}
\end{aligned}
$$

holds with probability at least $1-4n^{-1}$ as long as $n\geqslant C_3\{\log(np)+d\}$, where C is a constant depending on A_0, $\sigma_{\epsilon,jj}$, $v_{j,\delta}$, and v_j.

For the second term, let $Q_{ij}=\ell'_{\tau_j}(\epsilon_{ij})/s_j$ with $\mathbb{E}(Q_{ij}^2)=1$. Then,

$$\mathbb{E}(Q_{ij}^4)=\frac{\mathbb{E}[\{\ell'_{\tau_j}(\epsilon_{ij})\}^4]}{s_j^4}\leqslant\frac{v_j^4}{s_j^4}$$

and

$$\mathbb{E}(Q_{ij}^{2k})\leqslant\frac{v_j^4}{s_j^4}\left(\frac{\tau_j^2}{s_j^2}\right)^{k-2}$$

for all $k\geqslant 3$. It follows from Bernstein's inequality that for any $t>0$,

$$\left|\frac{1}{n}\sum_{i=1}^n Q_{ij}^2-1\right|\leqslant\frac{(v_j^4)^{1/2}}{s_j^2}\left(\frac{2t}{n}\right)^{1/2}+\frac{\tau_j^2}{s_j^2}\frac{t}{n} \qquad (B.11)$$

with probability at least $1-2\exp(-t)$. Plugging (B.9), (B.11), and Corollary B.1 into

(B.8) with $\tau_j = \tau_{0j}[n\{\log(np)+d\}^{-1}]^{1/(2+\delta)}$ and $t = \log(np)$, we yield

$$\begin{aligned}
&\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\{\ell'_{\tau_j}(e_{ij})\}^2 - \sigma_{\epsilon,jj}\right| \\
&\leqslant C\left\{d^{1/2}+(2\log n)^{1/2}\right\}\left\{\frac{\log(np)+d}{n}\right\}^{(1+\delta)/(2+\delta)} \\
&\quad + v_j^2\left\{\frac{2\log(np)}{n}\right\}^{1/2} + \tau_{0j}^2\frac{t}{n}\left\{\frac{n}{\log(np)+d}\right\}^{2/(2+\delta)} + \frac{2v_{j,\delta}^{2+\delta}}{\delta}\left\{\frac{\log(np)+d}{n}\right\}^{\delta/(2+\delta)} \\
&\leqslant C_5\left\{\frac{\log(np)+d}{n}\right\}^{\delta/(2+\delta)}
\end{aligned}$$

as long as $n \geqslant C_3\{\log(np)+d\}$ with probability at least $1-6n^{-1}$ for $\delta \in (0,2]$. □

The next lemma provides the non-asymptotic bound for m_j.

Lemma B.11. *On the event A_Δ, for $\tau_j = \tau_{0j}[n\{\log(np)+d\}^{-1}]^{1/(2+\delta)}$ where $\tau_{0j} \geqslant v_{j,\delta}$ for $\delta \in (0,2]$,*

$$\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}^*_{\tau_j}(e_{ij}) - 1\right| \leqslant C_6 \max\left[\left\{\frac{\log(np)+d}{n}\right\}^{1/2}, \frac{\Delta}{h_n}\right]$$

holds with probability at least $1-2n^{-1}$, where C_6 is a constant only depending on $v_{j,\delta}$ and v_j.

Proof. On the event A_Δ, $|\mathbb{I}^*_{\tau_j}(e_{ij}) - \mathbb{I}^*_{\tau_j}(\epsilon_{ij})| \leqslant \Delta {h_n}^{-1}$ due to the Lipschitz continuity of $\mathbb{I}^*_{\tau_j}(x)$. It follows that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}^*_{\tau_j}(e_{ij}) - 1\right| \leqslant \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}^*_{\tau_j}(\epsilon_{ij}) - 1\right| + \frac{\Delta}{h_n} \tag{B.12}$$

For the first term on the right hand side of (B.12), it follow Hoeffding's inequality

and Markov's inequality that

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}^*_{\tau_j}(\epsilon_{ij})-1\right| &\leqslant \mathbb{E}\{1-\mathbb{I}^*_{\tau_j}(\epsilon_{ij})\}+\left(\frac{t}{2n}\right)^{1/2}\\ &\leqslant \mathbb{P}\left(|\epsilon_{ij})|\geqslant\tau_j\right)+\left(\frac{t}{2n}\right)^{1/2}\\ &\leqslant \frac{v_{j,\delta}^{2+\delta}}{\tau_j^{2+\delta}}+\left(\frac{t}{2n}\right)^{1/2}\end{aligned}$$

with probability at least $1-2\exp(-t)$. Therefore,

$$\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}^*_{\tau_j}(\epsilon_{ij})-1\right|\leqslant\frac{v_{j,\delta}^{2+\delta}}{\tau_{0j}}\frac{\log(np)+d}{n}+\left\{\frac{\log(np)}{2n}\right\}^{1/2} \tag{B.13}$$

with probability at least $1-2n^{-1}$. The lemma is therefore proved. □

Lemma B.12. *Let $\tau_j=\tau_{0j}[n\{\log(np)+d\}^{-1}]^{1/(2+\delta)}$ where $\tau_{0j}\geqslant v_{j,\delta}$ for $\delta\in(0,2]$. On the event A_Δ, we have*

$$\max_{1\leqslant j\leqslant p}\left\|\mathbf{W}_j-\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}\right\|\leqslant C_7\max\left[\left\{\frac{\log(np)+d}{n}\right\}^{1/2},\frac{\Delta}{h_n}\right]$$

with probability at least $1-2n^{-1}$, where $C_7>0$ is a constant depending only on $\lambda_{\max}(\mathbf{\Sigma}_Z)$, A_0, and $v_{j,\delta}$ as long as $n\geqslant C_3\{\log(np)+d\}$.

Proof. For $\widetilde{\mathbf{Z}}_i=\mathbf{\Sigma}_Z^{-1/2}\mathbf{Z}_i$, we have

$$\begin{aligned}\left\|\mathbf{W}_j-\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}\right\| &= \left\|\mathbf{\Sigma}_Z^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n}\{\mathbb{I}^*_{\tau_j}(e_{ij})-1\}\widetilde{\mathbf{Z}}_i\widetilde{\mathbf{Z}}_i^{\mathrm{T}}\right]\mathbf{\Sigma}_Z^{1/2}\right\|\\ &\leqslant \|\mathbf{\Sigma}_Z\|\left\|\frac{1}{n}\sum_{i=1}^{n}\{\mathbb{I}^*_{\tau_j}(e_{ij})-1\}\widetilde{\mathbf{Z}}_i\widetilde{\mathbf{Z}}_i^{\mathrm{T}}\right\|.\end{aligned}$$

On the event A_Δ, for each unit vector $\mathbf{u}\in\mathbb{R}^{d+1}$,

$$\begin{aligned}\left|\mathbf{u}^{\mathrm{T}}\left[\frac{1}{n}\sum_{i=1}^{n}\{\mathbb{I}^*_{\tau_j}(e_{ij})-1\}\widetilde{\mathbf{Z}}_i\widetilde{\mathbf{Z}}_i^{\mathrm{T}}\right]\mathbf{u}\right| &\leqslant \left|\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{I}\left(|\epsilon_{ij}|\geqslant\tau_j\right)+\frac{\Delta}{h_n}\right\}\langle\mathbf{u},\widetilde{\mathbf{Z}}_i\rangle^2\right|\\ &\leqslant \mathbb{E}(\langle\mathbf{u},\widetilde{\mathbf{Z}}_i\rangle^2)\left[\mathbb{E}\left\{\mathbb{I}\left(|\epsilon_{ij}|\geqslant\tau_j\right)\right\}+\frac{\Delta}{h_n}+C\max(\rho,\rho^2)\right]\end{aligned}$$

with probability at least $1 - 2\exp(-t)$ where $\rho = \{n^{-1}(d+t)\}^{1/2}$ and $C > 0$ is an absolute constant. From properties of the sub-Gaussian random variable (Vershynin, 2018), $\mathbb{E}(|\mathbf{u}^{\mathrm{T}}\widetilde{\mathbf{Z}}_i|^k) \leqslant A_1^k(ek/2)\Gamma(k/2)$ for all $k \geqslant 1$, where $A_1 \geqslant e^{-1/2}$ is a constant depending only on A_0. Thus,

$$\mathbb{E}(\langle \mathbf{u}, \widetilde{\mathbf{Z}}_i \rangle^2) \leqslant A_1^2 e$$

and

$$\mathbb{E}\left\{\mathbb{I}\left(|\epsilon_{ij}| \geqslant \tau_j\right)\right\} = \mathbb{P}\left(|\epsilon_{ij}| \geqslant \tau_j\right) \leqslant \frac{v_{j,\delta}^{2+\delta}}{\tau_{0j}^{2+\delta}} \frac{d+t}{n}.$$

Take $t = \log(np)$. Putting together the obtained bounds yields

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbb{I}_{\tau_j}^{*}(e_{ij}) - 1 \right\} \mathbf{Z}_i \mathbf{Z}_i^{\mathrm{T}} \right\| \leqslant C_7 \max\left[\left\{ \frac{\log(np) + d}{n} \right\}^{1/2}, \frac{\Delta}{h_n} \right]$$

with probability at least $1 - 2n^{-1}$ as long as $n \geqslant C_3\{\log(np) + d\}$, where C_7 is a constant depending on A_0, $\lambda_{\max}(\mathbf{\Sigma}_Z)$, and $v_{j,\delta}$. □

S3 Testing hypotheses of the linear combinations of $\boldsymbol{\theta}_j$'s

S3.1 Method

In this section, we briefly discuss testing hypotheses of the linear combinations of regression coefficients, which is a special case of (2.2) in the main article with $q = 1$. For $j = 1, \ldots, p$, $\mathbf{c} \in \mathbb{R}^{d+1}$, and $c_{0j} \in \mathbb{R}$, the two-sided and one-sided hypotheses of interest

are

$$H_{0j} : \mathbf{c}^{\mathrm{T}}\boldsymbol{\theta}_j = c_{0j} \quad \text{versus} \quad H_{1j} : \mathbf{c}^{\mathrm{T}}\boldsymbol{\theta}_j \neq c_{0j}, \tag{C.1}$$

and

$$H_{0j} : \mathbf{c}^{\mathrm{T}}\boldsymbol{\theta}_j \leqslant (\geqslant) c_{0j} \quad \text{versus} \quad H_{1j} : \mathbf{c}^{\mathrm{T}}\boldsymbol{\theta}_j > (<) c_{0j}, \tag{C.2}$$

respectively. For each j, define

$$U_j = n^{1/2}(\mathbf{c}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_j\mathbf{c})^{-1/2}(\mathbf{c}^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_j - c_{0j}),$$

where $\widehat{\boldsymbol{\theta}}_j$ and $\widehat{\boldsymbol{\Sigma}}_j$ are estimated by (2.3) and (2.6) in the main paper. Notice $U_j^2 = V_j$. For threshold $z > 0$, we estimate the number of false discoveries $V(z)$ by

$$\widehat{V}(z) = \begin{cases} 2p\Phi(-z) & \text{(two-sided)}, \\ p\Phi(-z) & \text{(one-sided)}. \end{cases}$$

Let the number of discoveries by $R(z) = \sum_{j=1}^{p} \mathbb{I}(U_j \geqslant z)$. Then, we compute

$$\widehat{z}_\alpha = \inf\left\{z \geqslant 0 : \mathrm{AFDP}(z) \leqslant \alpha\right\},$$

where $\mathrm{AFDP}(z) = \widehat{V}(z)/R(z)$. For $j = 1, \ldots, p$, H_{0j} in (C.1) or (C.2) is rejected whenever $U_j \geqslant \widehat{z}_\alpha$.

S3.2 Theoretical guarantees

The following result for testing (C.1) is a straightforward corollary of Proposition 1. Denote $U_j^\circ = n^{1/2}(\mathbf{c}^{\mathrm{T}}\boldsymbol{\Sigma}_j\mathbf{c})^{-1/2}(\mathbf{c}^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_j - c_{0j})$ with known covariance $\boldsymbol{\Sigma}_j$. For $\mathscr{H}_0 = \{j :$

$1 \leqslant j \leqslant p, H_{0j}$ is true$\}$, let $V^\circ(z) = \sum_{j\in\mathscr{H}_0} \mathbb{I}(U_j^\circ \geqslant z)$ and $R^\circ(z) = \sum_{j=1}^p \mathbb{I}(U_j^\circ \geqslant z)$. Define $\mathrm{AFDP}_{c1}^\circ(z) = 2p_0\Phi(-z)/R^\circ(z)$ to be the counterpart of (3.1).

Theorem C.1. *Consider testing* (C.1)*. Assume Conditions 1 and 2 hold, and* $p_0 \geqslant ap$ *for some* $a \in (0,1)$*. Let* $\tau_j = \tau_{0j} n^{1/(2+\delta)}\{\log(np)+d\}^{-1/(2+\delta)}$ *with* $\tau_{0j} \geqslant v_{j,\delta}$ *and* $\delta \in (0,2]$*. Then, for any* $z \geqslant 0$*,* $|\mathrm{FDP}^\circ(z) - \mathrm{AFDP}_{c1}^\circ(z)| = o_{\mathbb{P}}(1)$ *as* $n, p \to \infty$*.*

Next, we provide the corresponding result for testing (C.2). Similarly, let $\mathrm{AFDP}_{c2}^\circ(z) = p_0\Phi(-z)/R^\circ(z)$.

Theorem C.2. *Consider testing* (C.2)*. Assume Conditions 1 and 2 hold, and* $p_0 \geqslant ap$ *for some* $a \in (0,1)$*. Let* $\tau_j = \tau_{0j} n^{1/(2+\delta)}\{\log(np)+d\}^{-1/(2+\delta)}$ *with* $\tau_{0j} \geqslant v_{j,\delta}$ *and* $\delta \in (0,2]$*. Then, for any* $z \geqslant 0$*,* $|\mathrm{FDP}^\circ(z) - \mathrm{AFDP}_{c2}^\circ(z)| = o_{\mathbb{P}}(1)$ *as* $n, p \to \infty$*.*

Proof. The proof is similar to that of Proposition 1. Let $z \geqslant 0$. We will show the stronger result that on event $\{p^{-1}R^\circ(z) \geqslant c\}$ for some $c > 0$,

$$p_0^{-1}V^\circ(z) = \Phi(-z) + O_{\mathbb{P}}(p^{-\kappa_1} + n^{-1/2} + [n^{-1}\{\log(np)+d\}]^{\delta/(2+\delta)}), \tag{C.3}$$

which leads to the conclusion immediately.

Let $\sigma_j^2 = \mathbf{c}^{\mathrm{T}}\boldsymbol{\Sigma}_j\mathbf{c} = \sigma_{\epsilon,jj}(\mathbf{c}^{\mathrm{T}}\boldsymbol{\Sigma}_Z^{-1}\mathbf{c}) \in \mathbb{R}$, $U_j^\circ = n^{1/2}\sigma_j^{-1}(\mathbf{c}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_j - \mathbf{c}^{\mathrm{T}}\boldsymbol{\theta}_j)$, and

$$S_j = n^{-1/2}\|\mathbf{c}^{\mathrm{T}}\boldsymbol{\Sigma}_Z^{-1/2}\|^{-1}\mathbf{c}^{\mathrm{T}}\boldsymbol{\Sigma}_Z^{-1}\sum_{i=1}^n[\ell_\tau'(\epsilon_{ij})\mathbf{Z}_i - \mathbb{E}\{\ell_\tau'(\epsilon_{ij})\mathbf{Z}_i\}],$$
$$R_j = n^{-1/2}\|\mathbf{c}^{\mathrm{T}}\boldsymbol{\Sigma}_Z^{-1/2}\|^{-1}\mathbf{c}^{\mathrm{T}}\boldsymbol{\Sigma}_Z^{-1}\sum_{i=1}^n\mathbb{E}\{\ell_\tau'(\epsilon_{ij})\mathbf{Z}_i\}].$$

For every $j \in H_{0j}$ and $t \geqslant 1$, it follows from Lemma B.1 that

$$\begin{aligned} |U_j^\circ - \sigma_{\epsilon,jj}^{-1/2}(S_j + R_j)| &= \left| n^{1/2}\sigma_j^{-1}(\mathbf{c}^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}^{\mathrm{T}}\boldsymbol{\theta}_j) - n^{-1/2}\sigma_j^{-1}\mathbf{c}^{\mathrm{T}}\boldsymbol{\Sigma}_Z^{-1}\sum_{i=1}^n \ell'_\tau(\epsilon_{ij})\mathbf{Z}_i \right| \\ &\leqslant n^{1/2}\sigma_{\epsilon,jj}^{-1/2}\left\| \boldsymbol{\Sigma}_Z^{1/2}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j) - \frac{1}{n}\sum_{i=1}^n \ell'_{\tau_j}(\epsilon_{ij})\boldsymbol{\Sigma}_Z^{-1/2}\mathbf{Z}_i \right\| \\ &\leqslant C_2\frac{\tau_0(d+t)}{(n\sigma_{\epsilon,jj})^{1/2}} \end{aligned} \tag{C.4}$$

with probability greater than $1-3\exp(-t)$ as long as $n \geqslant C_3(d+t)$ with $\tau_j = \tau_{0j}\{n(d+t)^{-1}\}^{1/(2+\delta)}$.

For $j = 1, \ldots, p$, denote $E_{1j}(t)$ on which event (C.4) holds, and define $E_1 = \bigcap_{j=1}^p E_{1j}(t)$. On E_1,

$$\sum_{j\in H_{0j}} \mathbb{I}\left\{\sigma_{\epsilon,jj}^{-1/2}S_j \geqslant z + C_2\frac{\tau_0(d+t)}{(n\sigma_{\epsilon,jj})^{1/2}}\right\} \leqslant V(z) \leqslant \sum_{j\in H_{0j}} \mathbb{I}\left\{\sigma_{\epsilon,jj}^{-1/2}S_j \geqslant z - C_2\frac{\tau_0(d+t)}{(n\sigma_{\epsilon,jj})^{1/2}}\right\} \tag{C.5}$$

with probability $1-3pe^{-t}$. For $x \in \mathbb{R}$, define

$$V^+(x) = \sum_{j\in H_{0j}} \mathbb{I}(\sigma_{\epsilon,jj}^{-1/2}S_j \geqslant x). \tag{C.6}$$

Hence, (C.5) can be written as

$$p_0^{-1}V^+\left\{z + C_2\frac{\tau_0(d+t)}{(n\sigma_{\epsilon,jj})^{1/2}}\right\} \leqslant p_0^{-1}V^\circ(z) \leqslant p_0^{-1}V^+\left\{z - C_2\frac{\tau_0(d+t)}{(n\sigma_{\epsilon,jj})^{1/2}}\right\}. \tag{C.7}$$

Therefore, we only need to derive the orders of $V^+(x)$. The rest of the proof is almost identical to that of Proposition 1 by replacing Lemma B.5 with Lemma 2.1 from Li and Shao (2002). We can easily obtain a similar bound for $\mathbb{E}[\{p_0^{-1}V^+(z) - \Phi(-z)\}^2]$ that

$$\mathbb{E}[\{p_0^{-1}V^+(z) - \Phi(-z)\}^2] \lesssim p^{-\kappa_1} + n^{-1/2} + \left(\frac{d+t}{n}\right)^{\delta/(2+\delta)}. \tag{C.8}$$

Recall that $\mathbb{P}(A_1) \leqslant 1 - 3pe^{-t}$ whenever $n \gtrsim d + t$. Taking $t = \log(np)$ in (C.7) and (C.8) proves (C.3). $\square$

Remark C.1. For testing $H_{0j} : \mathbf{c}^{\mathrm{T}}\boldsymbol{\theta}_j \geqslant c_0$ versus $H_{1j} : \mathbf{c}^{\mathrm{T}}\boldsymbol{\theta}_j < c_0$, we can use the same argument with (C.5) and (C.6) replaced by

$$\sum_{j\in\mathscr{H}_{0j}} \mathbb{I}\left\{\sigma_{\epsilon,jj}^{-1/2} S_j \leqslant -z - C_2 \frac{\tau_0(d+t)}{(n\sigma_{\epsilon,jj})^{1/2}}\right\} \leqslant V^{\circ}(z) \leqslant \sum_{j\in\mathscr{H}_{0j}} \mathbb{I}\left\{\sigma_{\epsilon,jj}^{-1/2} S_j \leqslant -z + C_2 \frac{\tau_0(d+t)}{(n\sigma_{\epsilon,jj})^{1/2}}\right\}$$

and $V^{-}(x) = \sum_{j\in\mathscr{H}_{0j}} \mathbb{I}(\sigma_{\epsilon,jj}^{-1/2} S_j \leqslant -x)$, respectively.

S4 Results under the fixed design

In this section, we consider our testing procedure in Section 2 for model (2.1) under the fixed design. Denote $\mathbf{z}_i^{\mathrm{T}} = \begin{bmatrix} 1 & \mathbf{x}_i^{\mathrm{T}} \end{bmatrix}$ the ith row of design matrix $\mathbf{Z}$. We first impose the following regularity condition, which is similar to that in Sun et al. (2020).

Condition D.1. The Gram matrix $\mathbf{S}_n = n^{-1}\sum_{i=1}^n \mathbf{z}_i\mathbf{z}_i^{\mathrm{T}}$ is positive definite and there exist constants c_l and c_u such that $c_l \leqslant \lambda_{\min}(\mathbf{S}_n) \leqslant \lambda_{\max}(\mathbf{S}_n) \leqslant c_u$. As $n \to \infty$, $\mathbf{S}_n \to \boldsymbol{\Sigma}_Z$ which is also positive definite.

The following condition is similar to the finite fourth order moment condition under the random design.

Condition D.2. There exist constants $\kappa, M > 0$ such that, for $\widetilde{\mathbf{z}}_i = \mathbf{S}_n^{-1/2}\mathbf{z}_i$,

$$\sup_{\mathbf{u}\in\mathbb{S}^d} \frac{1}{n}\sum_{i=1}^n (\mathbf{u}^{\mathrm{T}}\widetilde{\mathbf{z}}_i)^4 \exp(\kappa|\mathbf{u}^{\mathrm{T}}\widetilde{\mathbf{z}}_i|^2) \leqslant M.$$

We start with the counterpart of Proposition 1 to show that our procedure controls the false discovery proportion given the covariances of regression coefficients under the fixed design.

Theorem D.1. *Assume Conditions 1, D.1, and D.2 hold, and* $p_0 \geqslant ap$ *for some* $a \in (0,1)$*. Let* $\tau_j = \tau_{0j}\{n/\log(np)\}^{1/(2+\delta)}$ *where* $\tau_{0j} \geqslant v_{j,\delta}$ *for some* $\delta \in (0,2]$*. Then, for any* $z \geqslant 0$*,* $|\mathrm{FDP}^\circ(z) - \mathrm{AFDP}^\circ(z)| = o_{\mathbb{P}}(1)$ *as* $n, p \to \infty$*.*

Proof. For ease of exposition, we use z^2 instead of z in this proof. The proof is similar to that of Proposition 1. First, define

$$\begin{aligned} \mathbf{s}_j &= n^{-1/2}(\mathbf{C}\mathbf{S}_n^{-1}\mathbf{C}^{\mathrm{T}})^{-1/2}\mathbf{C}\mathbf{S}_n^{-1/2}\sum_{i=1}^n \mathbf{S}_n^{-1/2}\mathbf{Z}_i[\ell'_\tau(\epsilon_{ij}) - \mathbb{E}\{\ell'_\tau(\epsilon_{ij})\}], \\ \mathbf{r}_j &= n^{-1/2}(\mathbf{C}\mathbf{S}_n^{-1}bC^{\mathrm{T}})^{-1/2}\mathbf{C}\mathbf{S}_n^{-1/2}\sum_{i=1}^n \mathbf{S}_n^{-1/2}\mathbf{Z}_i\,\mathbb{E}\{\ell'_{\tau_j}(\epsilon_{ij})\}. \end{aligned} \tag{D.1}$$

By Proposition B.1, $\|\mathbf{r}_j\|$ is a small order term. Together with Corollary D.1, it implies that

$$\|\mathbf{T}_j^\circ - \sigma_{\epsilon,jj}^{-1/2}(\mathbf{s}_j + \mathbf{r}_j)\| \leqslant A\tau_{0j}(d+t)^{1/2}\left(\frac{dt}{\sigma_{\epsilon,jj}n}\right)^{1/2} \tag{D.2}$$

with probability $1 - 2d\exp(-t)$ as long as $n \geqslant \max\{32L_\infty^4 d^2 t, 2\kappa^{-2}(2d+t)\}$, where $\Delta_{n,\delta} = n^{-1}\sum_{i=1}^n v_\delta \widetilde{\mathbf{z}}_i\widetilde{\mathbf{z}}_i^{\mathrm{T}}$. The rest of the proof are almost identical to that of Proposition 1. Define $E_{1j}(t)$ the event on which (D.2) holds and let $E_1(t) = \bigcap_{j=1}^p E_{1j}(t)$ where

$\mathbb{P}\{E_1(t)\} = 1 - 2dp\exp(-t)$. One can obtain the counterparts of (A.5) and (A.12),

$$\begin{aligned}
&\sum_{j\in\mathscr{H}_0} \mathbb{I}\left[\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{s}_j\| \geqslant z + A\tau_{0j}(d+t)^{1/2}\left(\frac{dt}{\sigma_{\epsilon,jj}n}\right)^{1/2}\right] \\
\leqslant\ & V^\circ(z^2) \\
\leqslant\ & \sum_{j\in\mathscr{H}_0} \mathbb{I}\left[\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{s}_j\| \geqslant z - A\tau_{0j}(d+t)^{1/2}\left(\frac{dt}{\sigma_{\epsilon,jj}n}\right)^{1/2}\right]
\end{aligned}$$

and

$$\mathbb{E}\{p_0^{-1}V^\circ(z^2) - \mathbb{P}(\chi_q^2 > z^2)\}^2 \lesssim q^{1/2}p^{-\kappa_1} + n^{-1/2}q^{7/4} + q(t/n)^{\delta/(2+\delta)}.$$

Using Lemmas D.4 and D.5, we can obtain the desired result by taking $t = \log(np)$. □

Counterparts of Theorems C.1 and C.2 remain true under the assumptions for Theorem D.1. Their statements and proofs are identical to those of Theorems C.1 and C.2 and therefore are omitted.

S4.1 Technical lemmas under the fixed design

In this subsection, for the sake of completeness, we collect some auxiliary lemmas used for proving Theorem D.1. Most proofs, except that for Lemma D.2, are omitted given their similarities to those in Section S2.2. We start with three technical lemmas, which are modified from results in Sun et al. (2020). Lemmas D.1-D.3 provide general conclusions for the adaptive Huber regression with dimension d under the fixed design, and we suppress index j in their statements for ease of presentation.

Let $\mathscr{L}_\tau(\boldsymbol{\theta}) := n^{-1}\sum_{i=1}^n \ell_\tau(y_i - \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\theta})$. Lemma D.1 provides the lower bound of

$\lambda_{\min}\{\mathbf{S}_n^{-1/2}\nabla^2\mathscr{L}_\tau(\widetilde{\boldsymbol{\theta}})\mathbf{S}_n^{-1/2}\}$, which can be shown by slightly modifying similar arguments in Sun et al. (2020) under Condition D.1.

Lemma D.1. *Assume Condition D.1 holds and* $v_\delta := \{\mathbb{E}(|\epsilon|^{2+\delta})\}^{1/(2+\delta)} < \infty$ *for* $\delta \in (0.2]$. *Then for any* $t, r > 0$, *the matrix* $\mathbf{S}_n^{-1/2}\nabla^2\mathscr{L}_\tau(\widetilde{\boldsymbol{\theta}})\mathbf{S}_n^{-1/2}$ *with* $\tau > 2L_2 r$ *satisfies that*

$$\min_{\widetilde{\boldsymbol{\theta}}\in\mathbb{R}^{d+1}:\|\mathbf{S}_n^{1/2}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta})\|\leqslant r} \lambda_{\min}\{\mathbf{S}_n^{-1/2}\nabla^2\mathscr{L}_\tau(\widetilde{\boldsymbol{\theta}})\mathbf{S}_n^{-1/2}\}$$
$$\geqslant 1-(2L_2 r/\tau)^2 - L_2^2\{(2v_\delta/\tau)^{2+\delta} + (2n)^{-1/2}t^{1/2}\},$$

with probability at least $1-\exp(-t)$ *where* $L_2 = \max_{1\leqslant i\leqslant n}\|\widetilde{\mathbf{z}}_i\|$.

The following lemma is a variation of Theorem 1 in Sun et al. (2020) under the fixed design. The original theorem assumes finite $(1+\delta)$ order moment of ϵ_i for some $\delta > 0$. Using Lemma D.1, the proof of Lemma D.2 is similar to that of Theorem 1 in Sun et al. (2020), while the major technical challenge focuses on deriving the sharp non-asymptotic rate using our adaptive robustification parameter.

Lemma D.2. *Assume Conditions 1 and D.1 hold and* $v_\delta < \infty$ *for* $\delta \in (0.2]$. *Then, for any* $t > 0$ *and* $\tau_0 \geqslant v_\delta$, *the adaptive Huber regression estimator* $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}, \widehat{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{d+1}$ *in* (2.3) *with* $\tau = \tau_0(n/t)^{1/(2+\delta)}$ *satisfies*

$$\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})\| \leqslant C(L_\infty, \delta, v_\delta)d^{1/2}\left(\frac{t}{n}\right)^{1/2}$$

with probability at least $1-(2d+3)\exp(-t)$ *as long as* $n \geqslant 32L_\infty^4 d^2 t$, *where* $L_\infty = \max_{1\leqslant i\leqslant n}\|\widetilde{\mathbf{z}}_i\|_\infty$ *and* $C(L_\infty, \delta, v_\delta)$ *is a constant only depending on* L_∞, δ, *and* v_δ.

Proof. Recall that $\tau = \tau_0(n/t)^{1/(2+\delta)}$. Let $\widehat{\boldsymbol{\theta}}_\eta = \boldsymbol{\theta} + \eta(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ with $\eta \in (0, 1]$ so that $\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta})\| \leqslant r$. Lemma 2 from Sun et al. (2020) gives

$$\langle \nabla \mathscr{L}_\tau(\widehat{\boldsymbol{\theta}}_\eta) - \nabla \mathscr{L}_\tau(\boldsymbol{\theta}), \widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta} \rangle \leqslant \eta \langle \nabla \mathscr{L}_\tau(\widehat{\boldsymbol{\theta}}) - \nabla \mathscr{L}_\tau(\boldsymbol{\theta}), \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \rangle,$$

where $\nabla \mathscr{L}_\tau(\widehat{\boldsymbol{\theta}}) = 0$ by the Karush-Kuhn-Tucker condition. By the mean value theorem for vector-valued functions, the equality

$$\nabla \mathscr{L}_\tau(\widehat{\boldsymbol{\theta}}_\eta) - \nabla \mathscr{L}_\tau(\boldsymbol{\theta}) = \left[\int_0^1 \nabla^2 \mathscr{L}_\tau\{(1-t)\boldsymbol{\theta} + t\widehat{\boldsymbol{\theta}}_\eta\} dt \right] (\widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta})$$

holds, where the integral of a matrix is component-wise integrals. If there exists a constant $a_0 > 0$ such that

$$\min_{\widetilde{\theta} \in \mathbb{R}^{d+1} : \|\mathbf{S}_n^{1/2}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\| \leqslant r} \lambda_{\min}\{\mathbf{S}_n^{-1/2} \nabla^2 \mathscr{L}_\tau(\widetilde{\boldsymbol{\theta}}) \mathbf{S}_n^{-1/2}\} \geqslant a_0, \tag{D.3}$$

then

$$\begin{aligned}
& a_0 \|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta})\|_2^2 \\
\leqslant\ & \lambda_{\min}\left[\mathbf{S}_n^{-1/2} \nabla^2 \mathscr{L}_\tau\{(1-t)\boldsymbol{\theta} + t\widehat{\boldsymbol{\theta}}_\eta\} \mathbf{S}_n^{-1/2} \right] \|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta})\|_2^2 \\
=\ & \frac{(\widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta})^{\mathrm{T}} \mathbf{S}_n^{1/2}}{\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta})\|_2} \left[\mathbf{S}_n^{-1/2} \nabla^2 \mathscr{L}_\tau\{(1-t)\boldsymbol{\theta} + t\widehat{\boldsymbol{\theta}}_\eta\} S_n^{-1/2} \right] \frac{\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta})}{\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta})\|_2} \|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta})\|_2^2 \\
=\ & (\widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta})^{\mathrm{T}} \left[\nabla^2 \mathscr{L}_\tau\{(1-t)\boldsymbol{\theta} + t\widehat{\boldsymbol{\theta}}_\eta\} \right] (\widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta})
\end{aligned}$$

and

$$\begin{aligned}
a_0 \|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta})\|_2^2 &\leqslant \int_0^1 (\widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta})^{\mathrm{T}} \left[\nabla^2 \mathscr{L}_\tau\{(1-t)\boldsymbol{\theta} + t\widehat{\boldsymbol{\theta}}_\eta\} \right] (\widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta}) dt \\
&= (\widehat{\boldsymbol{\theta}}_\eta - \boldsymbol{\theta})^{\mathrm{T}} \left\{ \nabla \mathscr{L}_\tau(\widehat{\boldsymbol{\theta}}_\eta) - \nabla \mathscr{L}_\tau(\boldsymbol{\theta}) \right\} \\
&\leqslant \eta (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\mathrm{T}} \{ -\nabla \mathscr{L}_\tau(\boldsymbol{\theta}) \} \\
&\leqslant \eta \|\mathbf{S}_n^{-1/2} \nabla \mathscr{L}_\tau(\boldsymbol{\theta})\| \|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})\|
\end{aligned}$$

by putting together all the results above. Setting $\eta = 1$ yields

$$a_0 \|\mathbf{S}_n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\|_2 \leqslant \|\mathbf{S}_n^{-1/2} \nabla \mathscr{L}_\tau(\boldsymbol{\theta})\|_2. \tag{D.4}$$

Denote the kth entry of $\xi = \mathbf{S}_n^{-1/2} \nabla \mathscr{L}_\tau(\boldsymbol{\theta})$ by $\xi_k = -n^{-1} \sum_{i=1}^n \ell'_\tau(\epsilon_i) \widetilde{z}_{ik}$. By the triangle inequality, $|\xi_k| \leqslant |\xi_k - \mathbb{E}(\xi_k)| + |\mathbb{E}(\xi_k)|$. By Proposition B.1, as long as $n \geqslant (\sigma^{-2} v_\delta^{2+\delta})^{(2+\delta)/\delta} t$, it follows that $|\mathbb{E}\, \ell'_\tau(\epsilon)| \leqslant \tau^{-(1+\delta)} v_\delta^{2+\delta}$. By the definition of $\ell'_\tau(\cdot)$,

$$\begin{aligned}
&\left|\mathbb{E}\{n^{-1} \ell'_\tau(\epsilon_i) \widetilde{z}_{ik}\}\right| \leqslant \frac{v_\delta^{2+\delta}}{n\tau^{1+\delta}} L_\infty, \\
&\left|\frac{1}{n} \ell'_\tau(\epsilon_i) \widetilde{z}_{ik} - \mathbb{E}\left\{\frac{1}{n} \ell'_\tau(\epsilon_i) \widetilde{z}_{ik}\right\}\right| \leqslant |\widetilde{z}_{ik}| \left(\frac{\tau}{n} + \frac{v_\delta^{2+\delta}}{n\tau^{1+\delta}}\right) \leqslant L_\infty \left(\frac{\tau}{n} + \frac{v_\delta^{2+\delta}}{n\tau^{1+\delta}}\right), \text{ and} \\
&\mathbb{E}\left[\{n^{-1} \ell'_\tau(\epsilon_i) \widetilde{z}_{ik} - n^{-1}\, \mathbb{E}\{\ell'_\tau(\epsilon_i)\} \widetilde{z}_{ik}\}^2\right] \leqslant n^{-2} s^2 \widetilde{z}_{ik}^2
\end{aligned}$$

By Bernstein's inequality (Vershynin, 2018, Theorem 2.8.4),

$$|\xi_k| \leqslant f(n,t) + \frac{v_\delta^{2+\delta}}{\tau^{1+\delta}} L_\infty \leqslant C \left(\frac{t}{n}\right)^{1/2}$$

with probability at least $1 - 2\exp(-t)$ as long as $\tau \geqslant v_\delta (n/t)^{1/(2+\delta)}$ and $n > t$ where C is a constant depending on L_∞, δ and v_δ, and

$$\begin{aligned}
f(n,t) &= \frac{L_\infty}{3} \frac{t}{n} \left(\tau + \frac{v_\delta^{2+\delta}}{\tau^{1+\delta}}\right) + \frac{1}{3} \left\{\frac{L_\infty^2}{9} \frac{t^2}{n^2} \left(\tau + \frac{v_\delta^{2+\delta}}{\tau^{1+\delta}}\right)^2 + 18 s^2 \frac{t}{n} \frac{1}{n} \sum_{i=1}^n \widetilde{z}_{ik}^2\right\}^{1/2} \\
&\leqslant \frac{L_\infty}{3} \frac{t}{n} \left(\tau + \frac{v_\delta^{2+\delta}}{\tau^{1+\delta}}\right) + \frac{L_\infty}{3} \left\{\frac{1}{9} \frac{t^2}{n^2} \left(\tau + \frac{v_\delta^{2+\delta}}{\tau^{1+\delta}}\right)^2 + 18 s^2 \frac{t}{n}\right\}^{1/2} \\
&\lesssim \left(\frac{t}{n}\right)^{1/2}.
\end{aligned}$$

Then, for any $t > 0$,

$$\begin{aligned}\mathbb{P}(\|\xi\|_2 \geqslant C(d+1)^{1/2}n^{-1/2}t^{1/2}) &\leqslant \mathbb{P}(\|\xi\|_\infty \geqslant Cn^{-1/2}t^{1/2}) \\ &\leqslant \sum_{k=1}^{d+1} \mathbb{P}(|\xi_k| \geqslant Cn^{-1/2}t^{1/2}) \\ &\leqslant 2(d+1)\exp(-t).\end{aligned} \tag{D.5}$$

By Lemma D.1, (D.3) holds for $a_0 = 1/2$ and $r = \tau/(4L_2)$ with probability at least $1 - \exp(-t)$ since

$$\begin{aligned}\min_{\|\mathbf{S}_n^{1/2}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta})\| \leqslant \tau/(4L_2)} \lambda_{\min}\{\mathbf{S}_n^{-1/2}\nabla^2\mathscr{L}_\tau(\tilde{\boldsymbol{\theta}})\mathbf{S}_n^{-1/2}\} &\geqslant \frac{3}{4} - L_2^2\left\{\left(\frac{2v_\delta}{\tau_0}\right)^{2+\delta}\frac{t}{n} + \left(\frac{t}{2n}\right)^{1/2}\right\} \\ &\geqslant \frac{1}{2},\end{aligned}$$

holds as long as $n \geqslant \max(32L_2^4, 2^{5+\delta}L_2^2)t = 32\max(L_2^4, 2^\delta L_2^2)t$. By (D.4) and (D.5), we have

$$\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})\| \leqslant 2Cd^{1/2}n^{-1/2}t^{1/2}$$

with probability at least $1 - (2d+3)\exp(-t)$. □

Lemma D.3 provides a nonasymptotic Bahadur representation under the fixed design, and it implies that $\sqrt{n}\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ can be approximated by the multivariate normal distribution. It is a variation of Theorem 3.3 in the first version of Sun et al. (2020), which is available on ArXiv:1706.06991v1. It can be proved using Lemma D.2 with $\tau = \tau_0(n/t)^{1/(2+\delta)}$.

Lemma D.3. *Assume that Conditions D.1 and D.2 hold, and that $v_\delta < \infty$ for $\delta \in (0.2]$. Then, for any $t > 0$ and $\tau_0 \geqslant v_\delta$, the estimator $\widehat{\boldsymbol{\theta}}$ given in (2.3) with $\tau = \tau_0(n/t)^{1/(2+\delta)}$*

satisfies that

$$\mathbb{P}\left\{\left\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})-\frac{1}{n}\sum_{i=1}^{n}\ell_\tau'(\epsilon_i)\mathbf{S}_n^{-1/2}\mathbf{Z}_i\right\|\geqslant A\tau_0(d+t)^{1/2}\frac{(dt)^{1/2}}{n}\right\}\leqslant 2(d+2)e^{-t},$$

whenever $n\geqslant\max\{32L_\infty^4d^2t,2\kappa^{-2}(2d+t)\}$, *where* $A>0$ *is a constant depending only on* M *in Condition D.2,* $C(L_\infty,\delta,v_\delta)$ *from Lemma D.2, and* $\tau_0^{-2}\|\Delta_{n,\delta}\|$ *with* $\Delta_{n,\delta}=n^{-1}\sum_{i=1}^n v_\delta^2\widetilde{z}_i\widetilde{z}_i^{\mathrm{T}}$.

We conclude this subsection with the counterparts of results in Section S2.2. From Lemma D.3, the adaptive Huber regression estimator is expected to be approximated by a Bahadur representation under the fixed design.

Corollary D.1. *For* $\mathbf{T}_j^\circ$ *and its Bahadur representation in* (D.1)*, it holds*

$$\|\mathbf{T}_j^\circ-\sigma_{\epsilon,jj}^{-1/2}(\mathbf{s}_j+\mathbf{r}_j)\|\leqslant A\tau_{0j}(d+t)^{1/2}\frac{dt}{(n\sigma_{\epsilon,jj})^{1/2}}$$

with probability at least $1-2(d+2)\exp(-t)$.

The following lemmas show that the distribution of the Bahadur representation in (D.1) is close to $N(\mathbf{0},\sigma_{\epsilon,jj}^2\mathbf{I})$. We decompose $|\,\mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{s}_j\|\geqslant x)-\mathbb{P}(\|\mathbf{G}\|\geqslant x)|$ into two parts. Lemma D.4 quantifies the difference between the cumulative distribution functions of $\|\sigma_{\epsilon,jj}^{-1/2}s_j\mathbf{G}\|$ and $\|\mathbf{G}\|$, and Lemma D.5 quantifies the distinction between the cumulative distribution functions of $\|\sigma_{\epsilon,jj}^{-1/2}s_j\mathbf{G}\|$ and $\|\sigma_{\epsilon,jj}^{-1/2}\mathbf{s}_j\|$. Their proofs are similar to those of Lemmas B.2-B.3 and therefore are omitted.

Lemma D.4. *Let* $\mathbf{G} \sim N(\mathbf{0}, \mathbf{I}) \in \mathbb{R}^q$. *For* $\tau_j = \tau_{0j}(n/t)^{1/(2+\delta)}$ *for some* $\delta \in (0, 2]$ *where* $\tau_{0j} \geqslant v_{j,\delta}$,

$$\sup_{x \in \mathbb{R}^+} \left| \mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2} s_j \mathbf{G}\| \geqslant x) - \mathbb{P}(\|\mathbf{G}\| \geqslant x) \right| \leqslant q^{1/2} \frac{v_\delta^{2+\delta}}{\delta \tau_{0j}^\delta \sigma_{\epsilon,jj}} \left(\frac{t}{n}\right)^{\delta/(2+\delta)}.$$

Lemma D.5. *Let* $\mathbf{G} \sim N(\mathbf{0}, \mathbf{I}) \in \mathbb{R}^q$.

$$\sup_{x \in \mathbb{R}^+} |\, \mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2} \mathbf{s}_j\| \geqslant x) - \mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2} s_j \mathbf{G}\| \geqslant x)| \lesssim n^{-1/2} q^{7/4}.$$

S5 Additional numerical studies and results

S5.1 Additional results from Section 4

In this section, we report additional numerical results for simulations detailed in Section 4 in the main paper. For ease of presentation, we revisit the simulation settings. We generate data from (2.1) in the main paper for $n = 85, 120, 150$, $p = 1000, 2000$, $p_1 = 50$, and $d = 6, 8$. We consider three heavy-tailed error distributions:

(a) Pareto distribution with shape parameter 4 and scale parameter 1,

(b) log-normal distribution with $\mu = 0$ and $\sigma = 1$, and

(c) a mixture of the log-normal distribution in (b) and the t_2 distribution with proportion 0.7 and 0.3 respectively.

To incorporate dependence, we set $\mathbf{\Xi} = 100\mathbf{R}_\epsilon^{1/2}\mathbf{E}$, where the correlation matrix $\mathbf{R}_\epsilon$ has one of the following three structures:

- *Model* 1, $\mathbf{R}_\epsilon$ is the identity matrix;
- *Model* 2, $\mathbf{R}_\epsilon = (r_{\epsilon,jk})_{1\leqslant j,k\leqslant p}$ is sparse with $r_{\epsilon,jj} = 1$ and $r_{\epsilon,ij} = r_{\epsilon,ji}$ independently drawn from $0.3 \times \text{Bernoulli}(0.1)$ for $i \neq j$; and
- *Model* 3, $\mathbf{R}_\epsilon = (r_{\epsilon,jk})_{1\leqslant j,k\leqslant p}$ with $r_{\epsilon,jj} = 1$, $r_{\epsilon,j,j+1} = r_{\epsilon,j+1,j} = 0.3$, $r_{\epsilon,j,j+2} = r_{\epsilon,j+2,j} = 0.1$, and $r_{\epsilon,j,j+k} = r_{\epsilon,j+k,j} = 0$ for $k \geqslant 3$.

For each $j = 1, \ldots, p$, we set $\mu_j = 5000$ and consider two hypotheses:

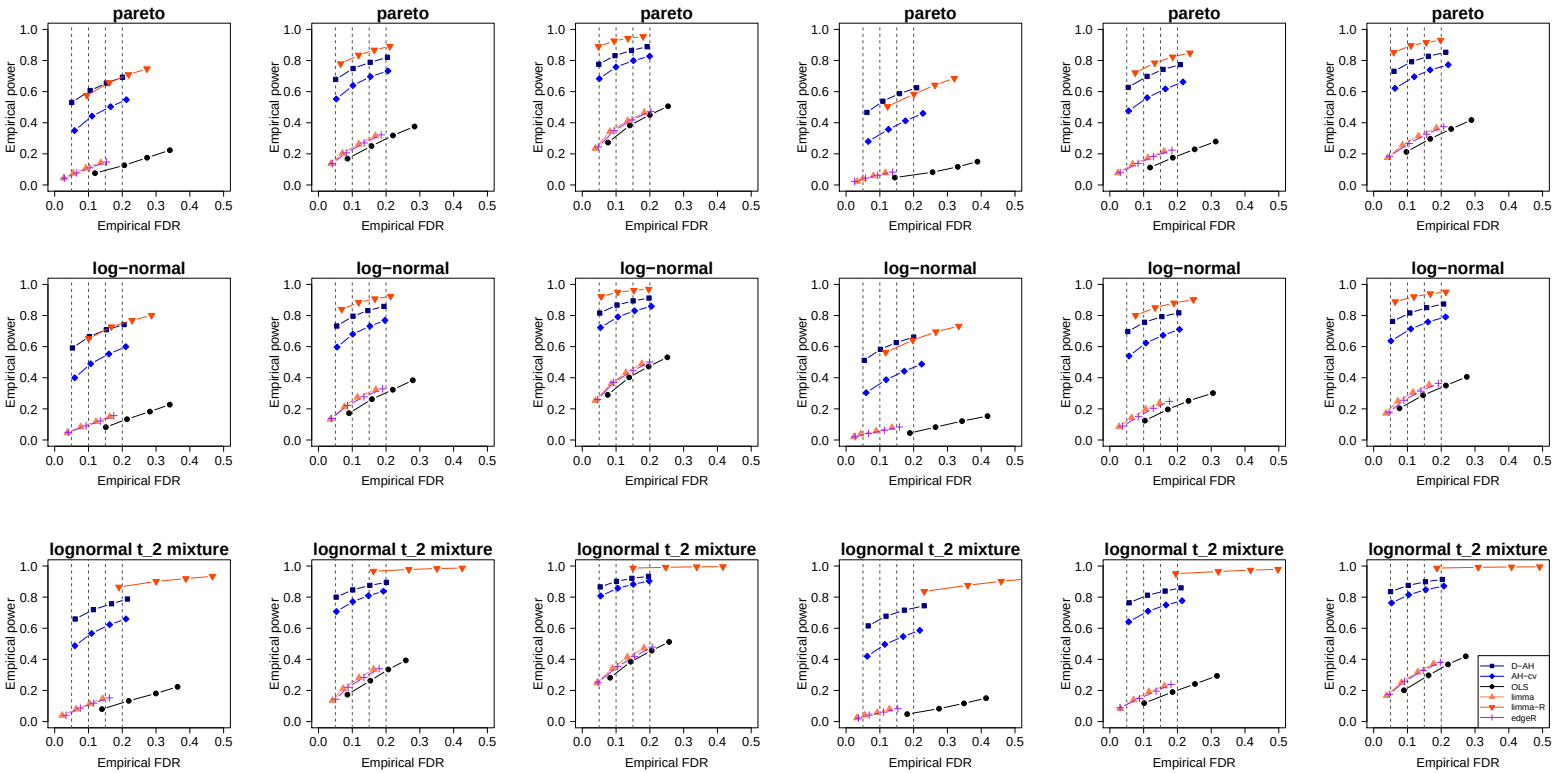
- *Hypothesis* 1, $H_{0j} : \mathbf{1}^{\mathrm{T}}\boldsymbol{\beta}_j = 0$ versus $H_{aj} : \mathbf{1}^{\mathrm{T}}\boldsymbol{\beta}_j \neq 0$, where $q = 1$, and
- *Hypothesis* 2, $H_{0j} : \boldsymbol{\beta}_j = 0 \in \mathbb{R}^d$ versus $H_{aj} : \boldsymbol{\beta}_j \neq 0$ $(j = 1, \ldots, p)$, where $q = d$.

For *Hypothesis* 1, we let $\beta_{jk} \sim \text{Unif}(-150, 150)$ for $1 \leqslant j \leqslant p$ and $1 \leqslant k \leqslant d-1$, $\beta_{jd} = -\sum_{k=1}^{d-1} \beta_{jk}$ for $1 \leqslant j \leqslant p - p_1$ so that $\mathbf{1}^{\mathrm{T}}\boldsymbol{\beta}_j = 0$, and $\beta_{jd} = \delta d^{1/2} W_j - \sum_{k=1}^{d-1} \beta_{jk}$ for $p - p_1 + 1 \leqslant j \leqslant p$, where W_j are Rademacher random variables. For *Hypothesis* 2, let $\boldsymbol{\beta}_j = 0$ for $1 \leqslant j \leqslant p - p_1$, and $\beta_{jk} = (2d^{-1})^{1/2}\delta W_{jk}$ for $p - p_1 + 1 \leqslant j \leqslant p$ and $1 \leqslant k \leqslant d$, where W_{jk} are Rademacher random variables. We take $\delta = 22.5$ for results in Figures S1 to S11.

Results for testing different hypotheses under *Model* 1 are presented in Figures S1-S4, and those under *Model* 3 are depicted in Figures S8-S11. The simulation results for *Model* 2 with different d's are displayed in Figures S5-S7. Similar observations to Section 4 are made from these extra numerical results. The proposed method that employs data-adaptive Huber regression or selects τ_j via five-fold cross-validation outperforms other competing methods in general with satisfactory control of the empirical

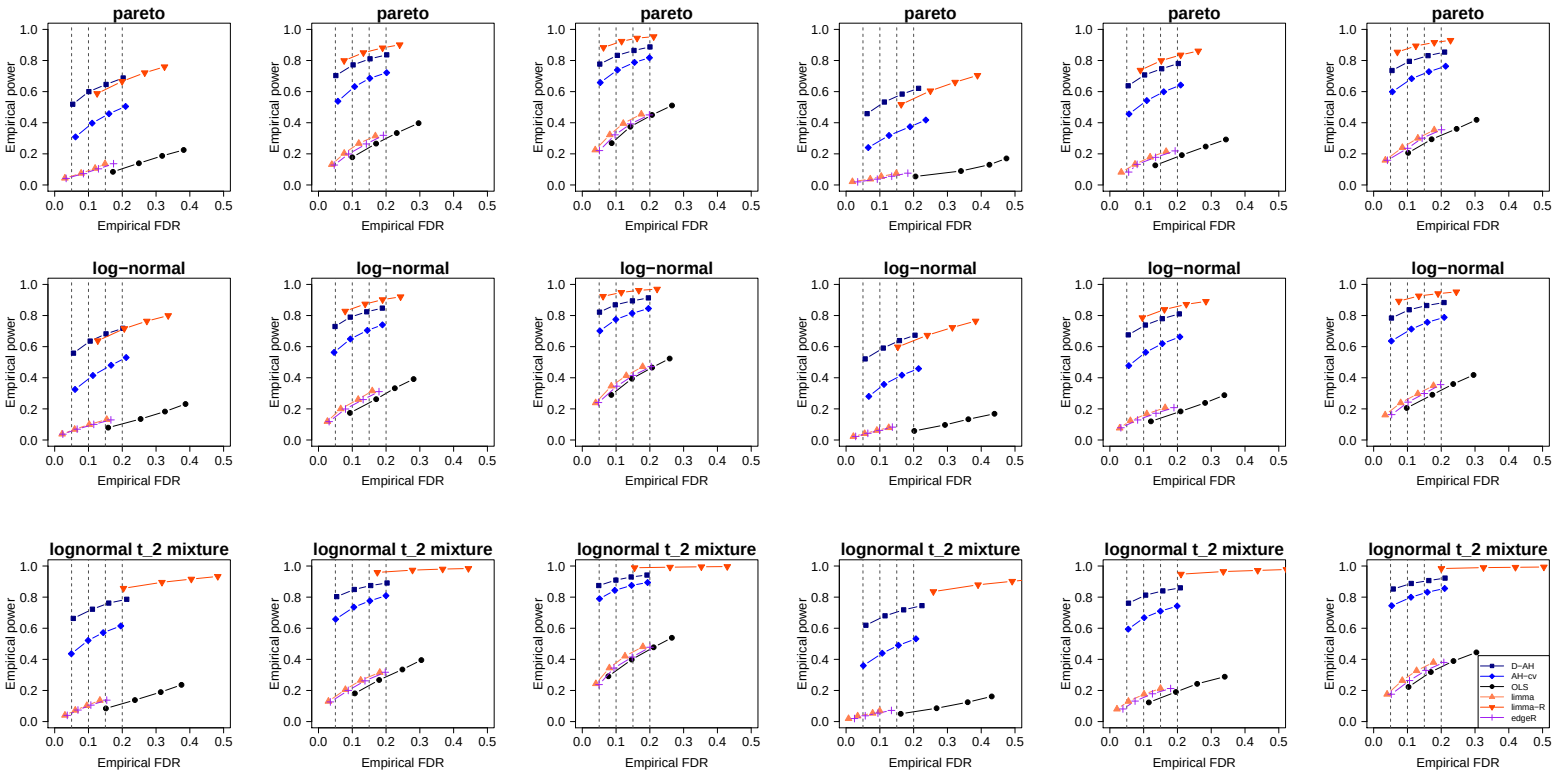
false discovery rate and good powers. When n is small and p is large (as $p = 2000$), the control of empirical false discovery rate is challenging for all methods. However, as n increases, our method preserves the nominal level of false discovery rate and is more powerful than `edgeR` and `limma`. Similar observations are made when the dependence is strong (*Model* 3) and d is large.

Similar to Figure 3 in the main paper, Figure S12 compares the powers of different methods for testing *Hypothesis 1*, the linear contrast, with varying signal strengths as defined in Section 4. Similar to Figure 3 in the main paper, the proposed method with either adaptive Huber regression or cross validation-selected τ_j's outperforms `limma` and `edgeR` for all error settings.



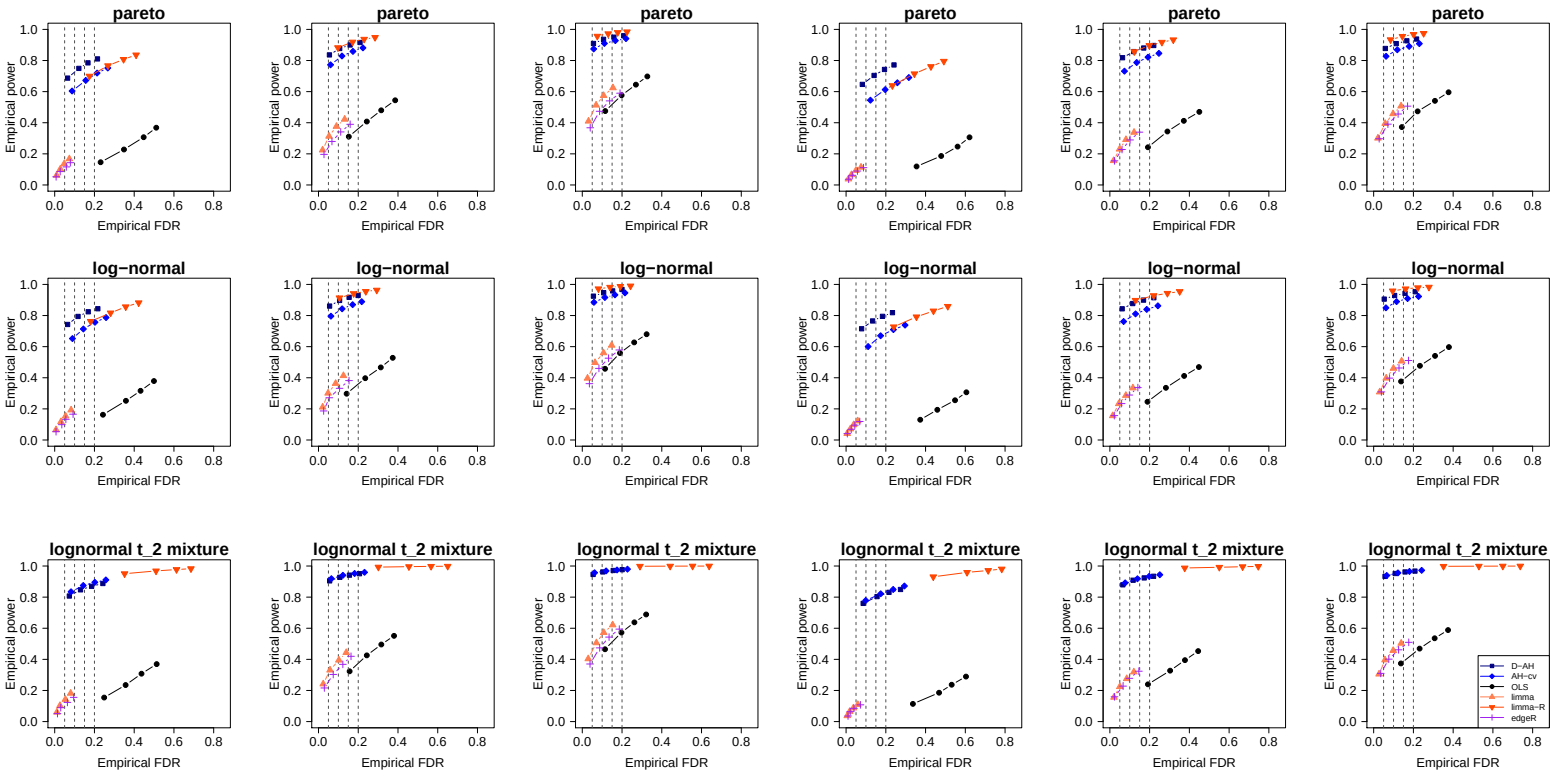
(a) $(n, p) = (85, 1000)$ (b) $(n, p) = (120, 1000)$ (c) $(n, p) = (150, 1000)$ (d) $(n, p) = (85, 2000)$ (e) $(n, p) = (120, 2000)$ (f) $(n, p) = (150, 2000)$

Figure S1: Empirical false discovery rate (FDR) and power for testing *Hypothesis* 1, a single contrast, under *Model* 1 (independent and identically distributed errors) with $d = 6$ by our procedure with data-adaptive Huber regression (D-AH, ■); our procedure with cross-validation (AH-cv, ♦); the ordinary least square estimator (OLS, •); `limma` (▲); `limma` with the robust regression (`limma`-R, ▾); and `edgeR` (+). Each point on the figures displays the empirical FDR and power of the corresponding method at a nominal FDR level, which is marked as a vertical gray dashed line. The pre-specified FDRs are $0.05, 0.1, 0.15, 0.2$.



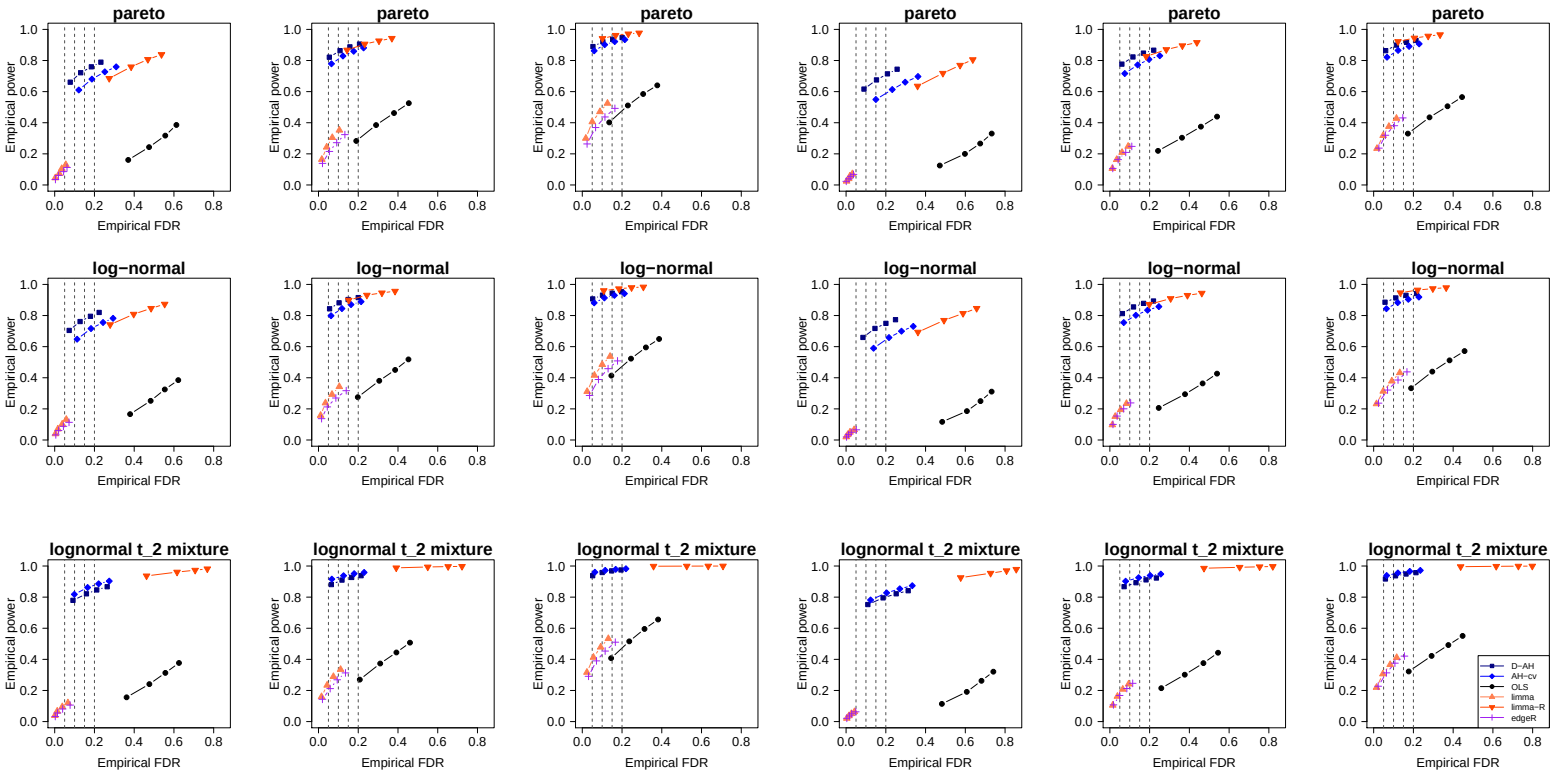
(a) $(n, p) = (85, 1000)$ (b) $(n, p) = (120, 1000)$ (c) $(n, p) = (150, 1000)$ (d) $(n, p) = (85, 2000)$ (e) $(n, p) = (120, 2000)$ (f) $(n, p) = (150, 2000)$

Figure S2: Empirical FDR and power for testing *Hypothesis* 1, a single contrast, under *Model* 1 (independent and identically distributed errors) with $d = 8$ by our procedure with data-adaptive Huber regression (D-AH, ■); our procedure with cross-validation (AH-cv, ⧫); the ordinary least square estimator (OLS, •); `limma` (▲); `limma` with the robust regression (`limma`-R, ▾); and `edgeR` (+). Each point on the figures displays the empirical FDR and power of the corresponding method at a nominal FDR level, which is marked as a vertical gray dashed line. The pre-specified FDRs are $0.05, 0.1, 0.15, 0.2$.



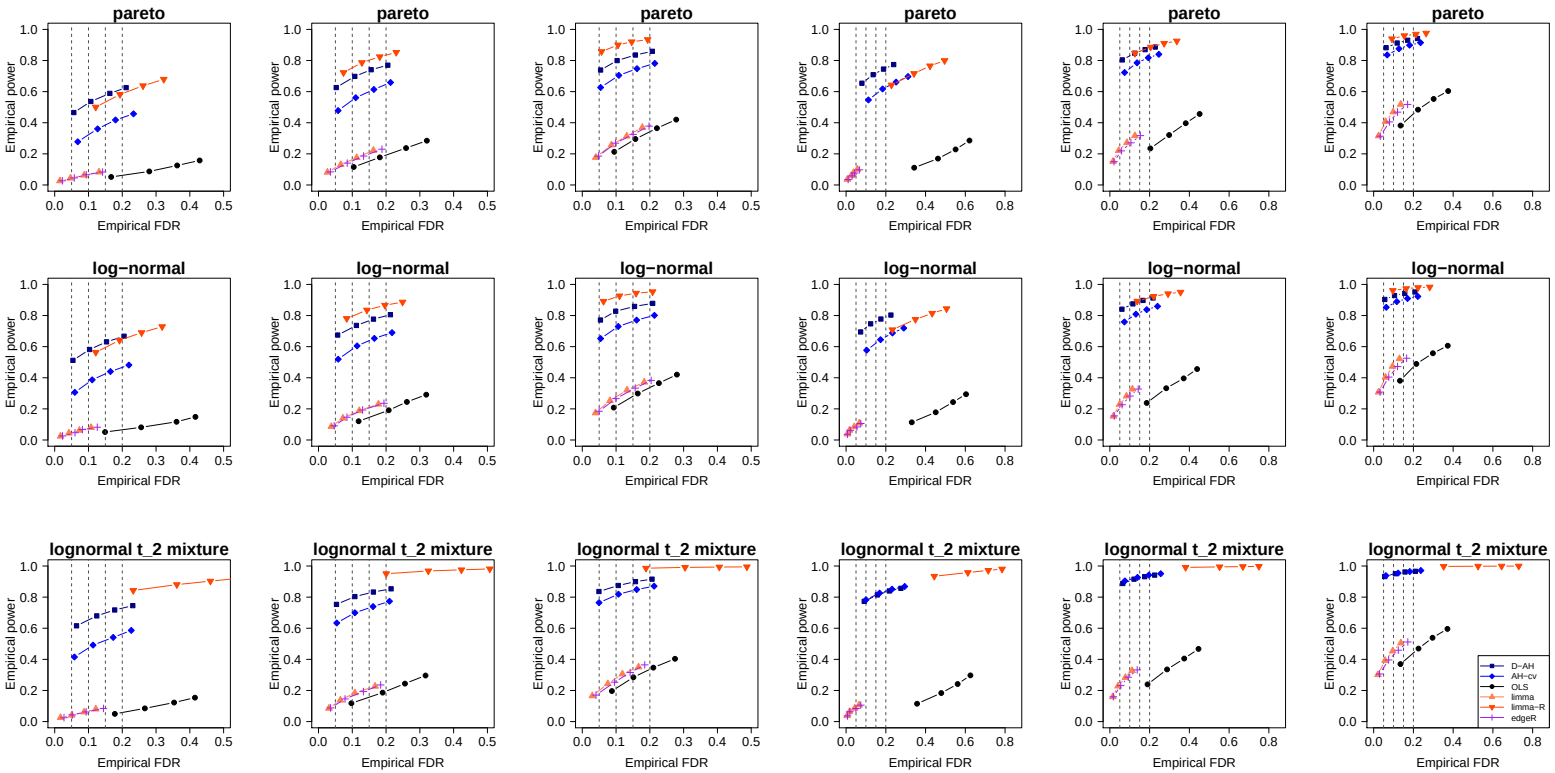
(a) $(n, p) = (85, 1000)$ (b) $(n, p) = (120, 1000)$ (c) $(n, p) = (150, 1000)$ (d) $(n, p) = (85, 2000)$ (e) $(n, p) = (120, 2000)$ (f) $(n, p) = (150, 2000)$

Figure S3: Empirical FDR and power for testing *Hypothesis* 2 under *Model* 1 (independent and identically distributed errors) with $d = 6$ by our procedure with data-adaptive Huber regression (D-AH, ■); our procedure with cross-validation (AH-cv, ⧫); the ordinary least square estimator (OLS, •); `limma` (▲); `limma` with the robust regression (`limma`-R, ▾); and `edgeR` (+). Each point on the figures displays the empirical FDR and power of the corresponding method at a nominal FDR level, which is marked as a vertical gray dashed line. The pre-specified FDRs are $0.05, 0.1, 0.15, 0.2$.



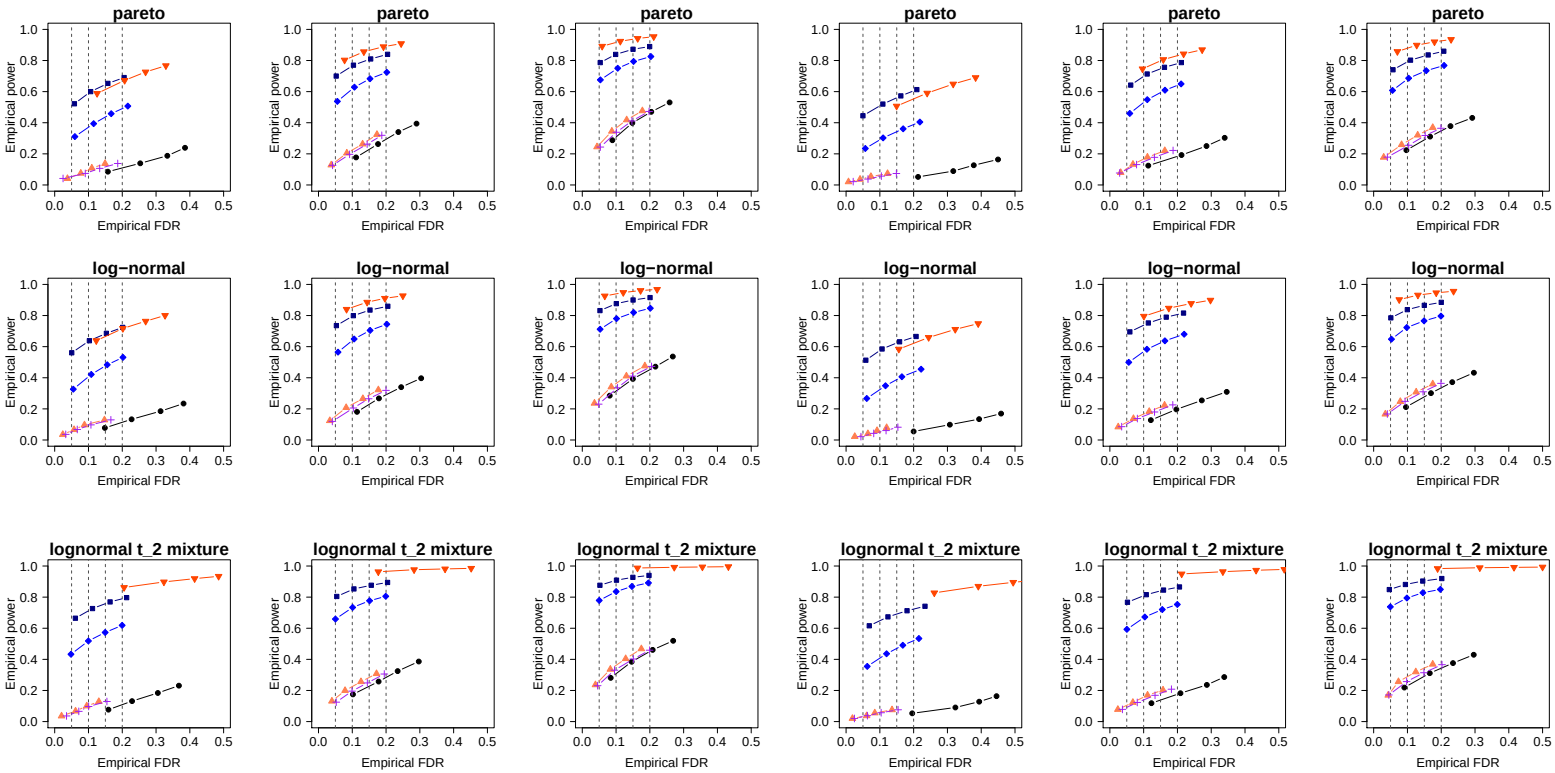
(a) $(n, p) = (85, 1000)$ (b) $(n, p) = (120, 1000)$ (c) $(n, p) = (150, 1000)$ (d) $(n, p) = (85, 2000)$ (e) $(n, p) = (120, 2000)$ (f) $(n, p) = (150, 2000)$

Figure S4: Empirical FDR and power for testing *Hypothesis* 2 under *Model* 1 (independent and identically distributed errors) with $d = 8$ by our procedure with data-adaptive Huber regression (D-AH, ■); our procedure with cross-validation (AH-cv, ⧫); the ordinary least square estimator (OLS, •); `limma` (▲); `limma` with the robust regression (`limma`-R, ▾); and `edgeR` (+). Each point on the figures displays the empirical FDR and power of the corresponding method at a nominal FDR level, which is marked as a vertical gray dashed line. The pre-specified FDRs are $0.05, 0.1, 0.15, 0.2$.



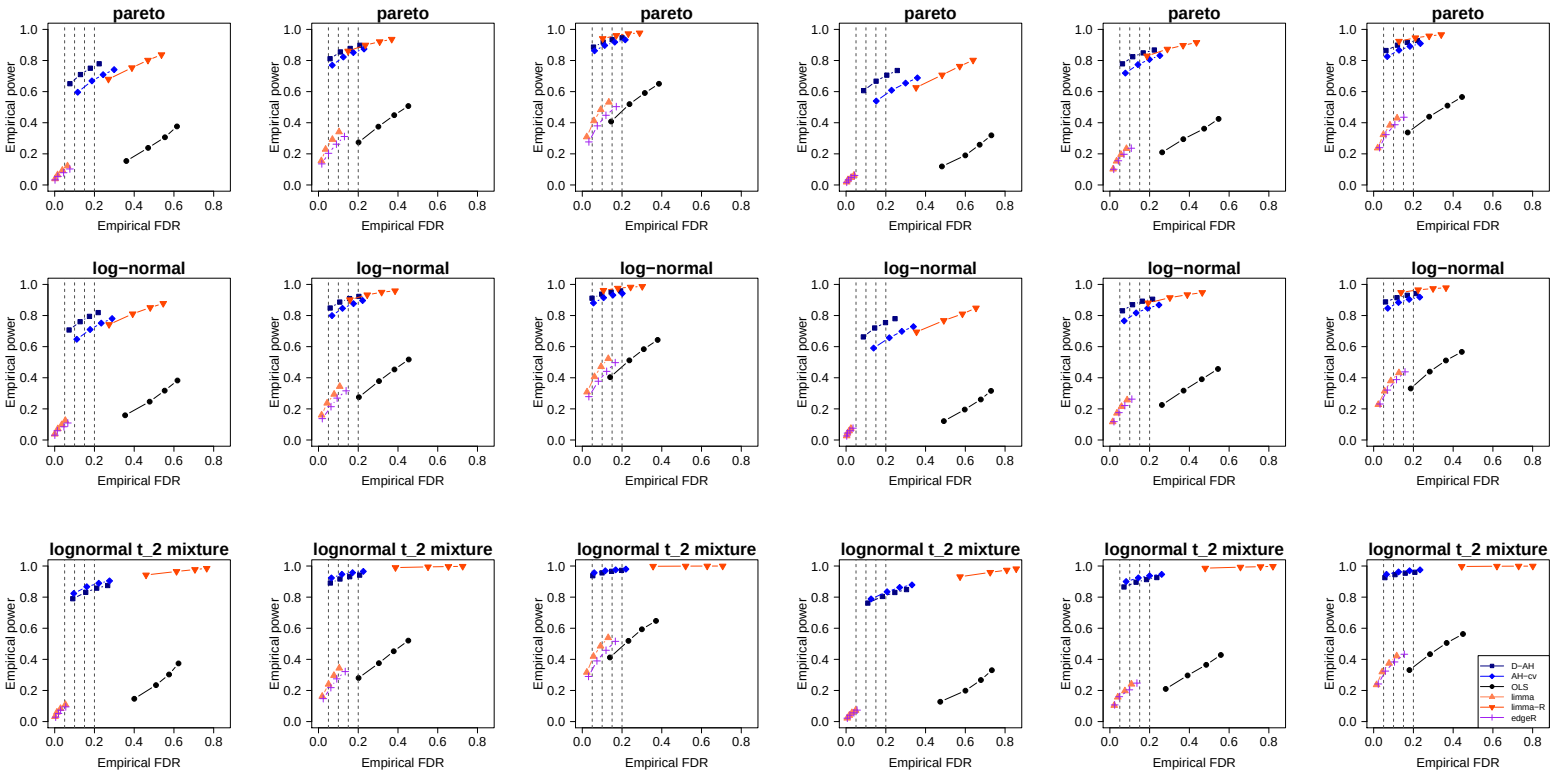
(a) H1, $n = 85$ (b) H1, $n = 120$ (c) H1, $n = 150$ (d) H2, $n = 85$ (e) H2, $n = 120$ (f) H2, $n = 150$

Figure S5: Empirical FDR and power for testing *Hypothesis* 1 (H1) and *Hypothesis* 2 (H2) under *Model* 2 (sparsely dependent errors) with $p = 2000$ and $d = 6$ by our procedure with data-adaptive Huber regression (D-AH, ■); our procedure with cross-validation (AH-cv, ⧫); the ordinary least square estimator (OLS, •); `limma` (▲); `limma` with the robust regression (`limma`-R, ▾); and `edgeR` (+). Each point on the figures displays the empirical FDR and power of the corresponding method at a nominal FDR level, which is marked as a vertical gray dashed line. The pre-specified FDRs are $0.05, 0.1, 0.15, 0.2$.



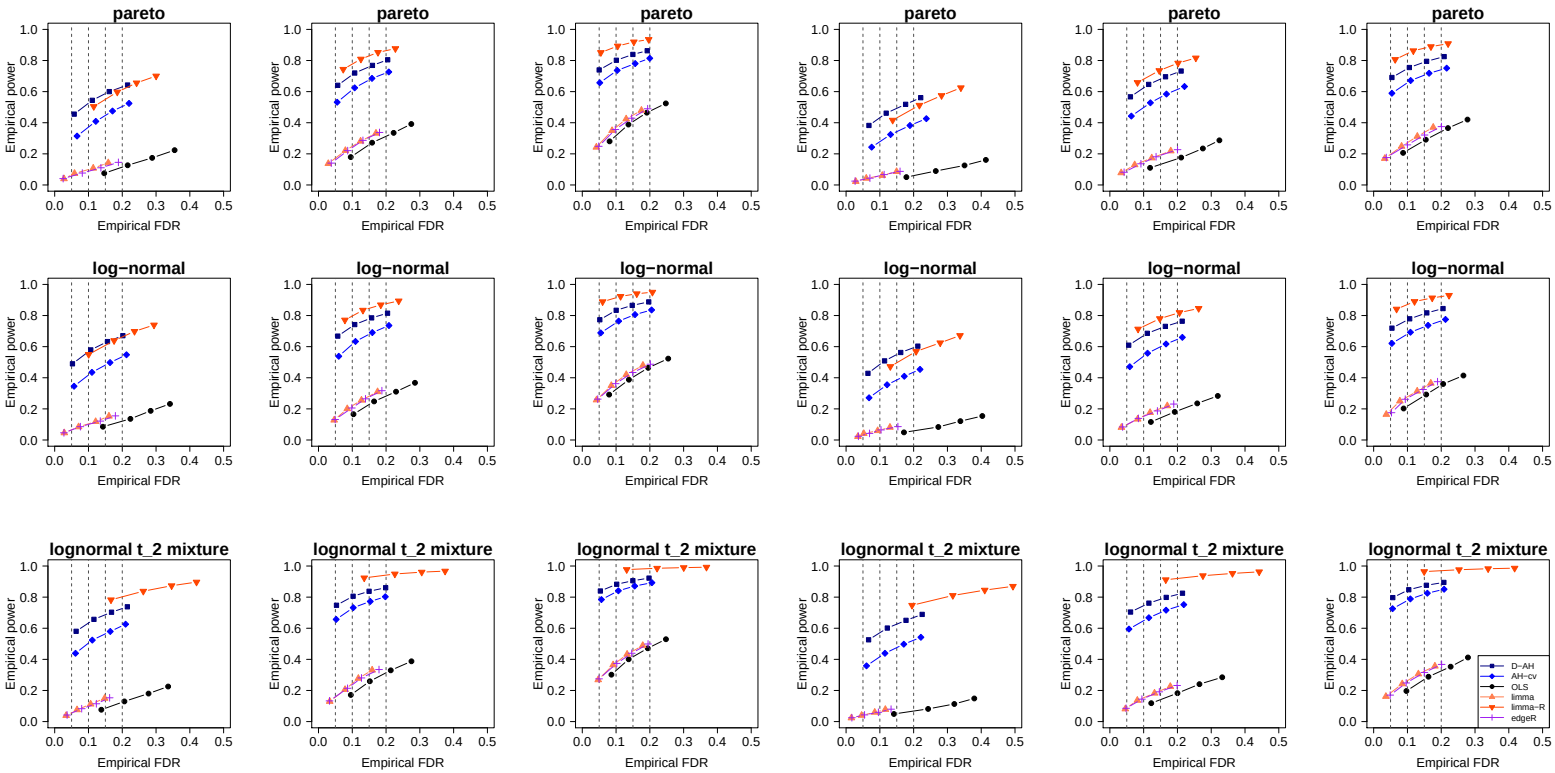
(a) $(n, p) = (85, 1000)$ (b) $(n, p) = (120, 1000)$ (c) $(n, p) = (150, 1000)$ (d) $(n, p) = (85, 2000)$ (e) $(n, p) = (120, 2000)$ (f) $(n, p) = (150, 2000)$

Figure S6: Empirical FDR and power for testing *Hypothesis* 1, a single contrast, under *Model* 2 (sparsely dependent errors) with $d = 8$ by our procedure with data-adaptive Huber regression (D-AH, ■); our procedure with cross-validation (AH-cv, ⧫); the ordinary least square estimator (OLS, •); `limma` (▲); `limma` with the robust regression (`limma`-R, ▾); and `edgeR` (+). Each point on the figures displays the empirical FDR and power of the corresponding method at a nominal FDR level, which is marked as a vertical gray dashed line. The pre-specified FDRs are $0.05, 0.1, 0.15, 0.2$.



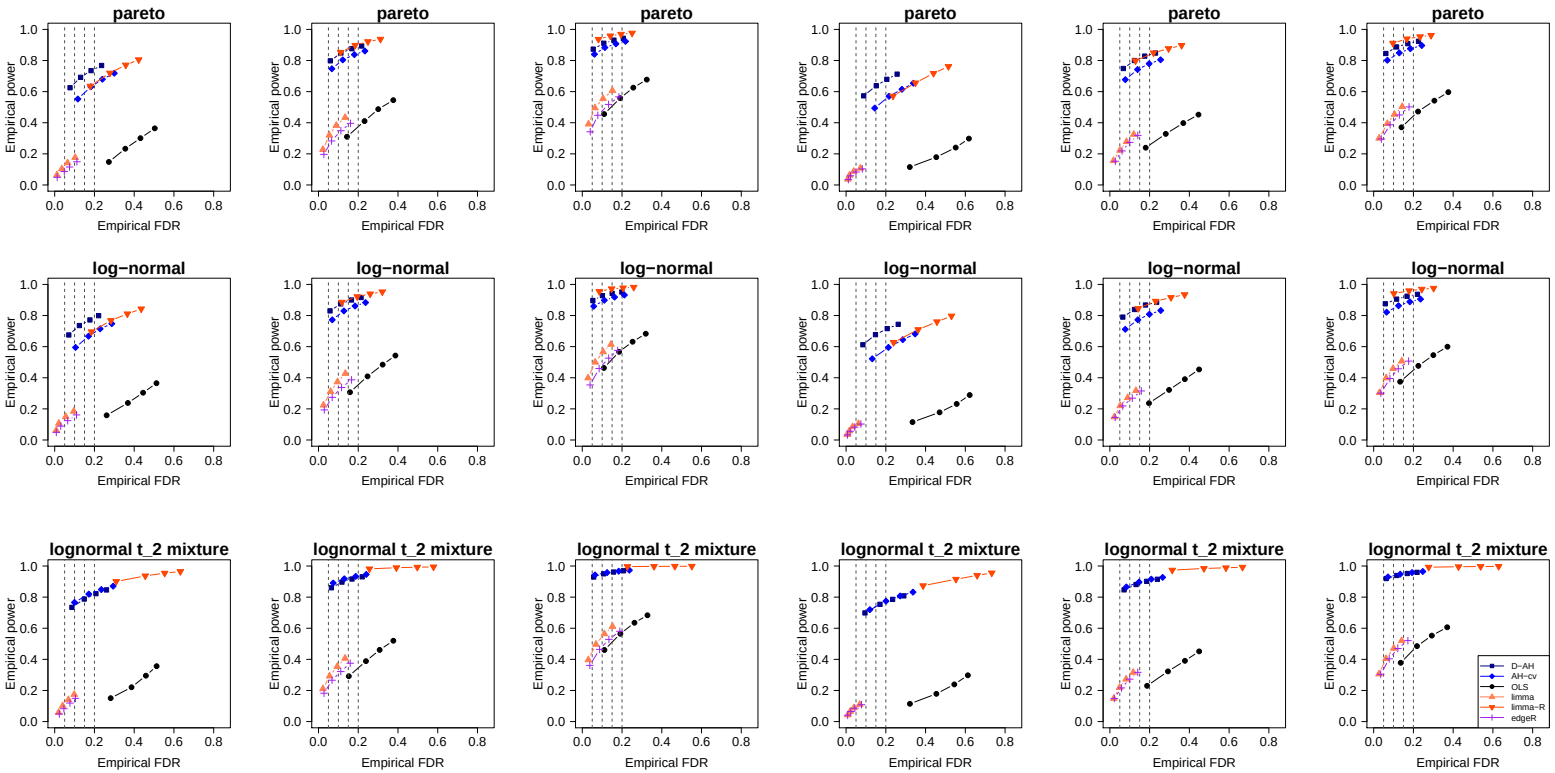
(a) $(n, p) = (85, 1000)$ (b) $(n, p) = (120, 1000)$ (c) $(n, p) = (150, 1000)$ (d) $(n, p) = (85, 2000)$ (e) $(n, p) = (120, 2000)$ (f) $(n, p) = (150, 2000)$

Figure S7: Empirical FDR and power for testing *Hypothesis* 2 under *Model* 2 (sparsely dependent errors) with $d = 8$ by our procedure with data-adaptive Huber regression (D-AH, ■); our procedure with cross-validation (AH-cv, ♦); the ordinary least square estimator (OLS, •); `limma` (▲); `limma` with the robust regression (`limma`-R, ▾); and `edgeR` (+). Each point on the figures displays the empirical FDR and power of the corresponding method at a nominal FDR level, which is marked as a vertical gray dashed line. The pre-specified FDRs are $0.05, 0.1, 0.15, 0.2$.



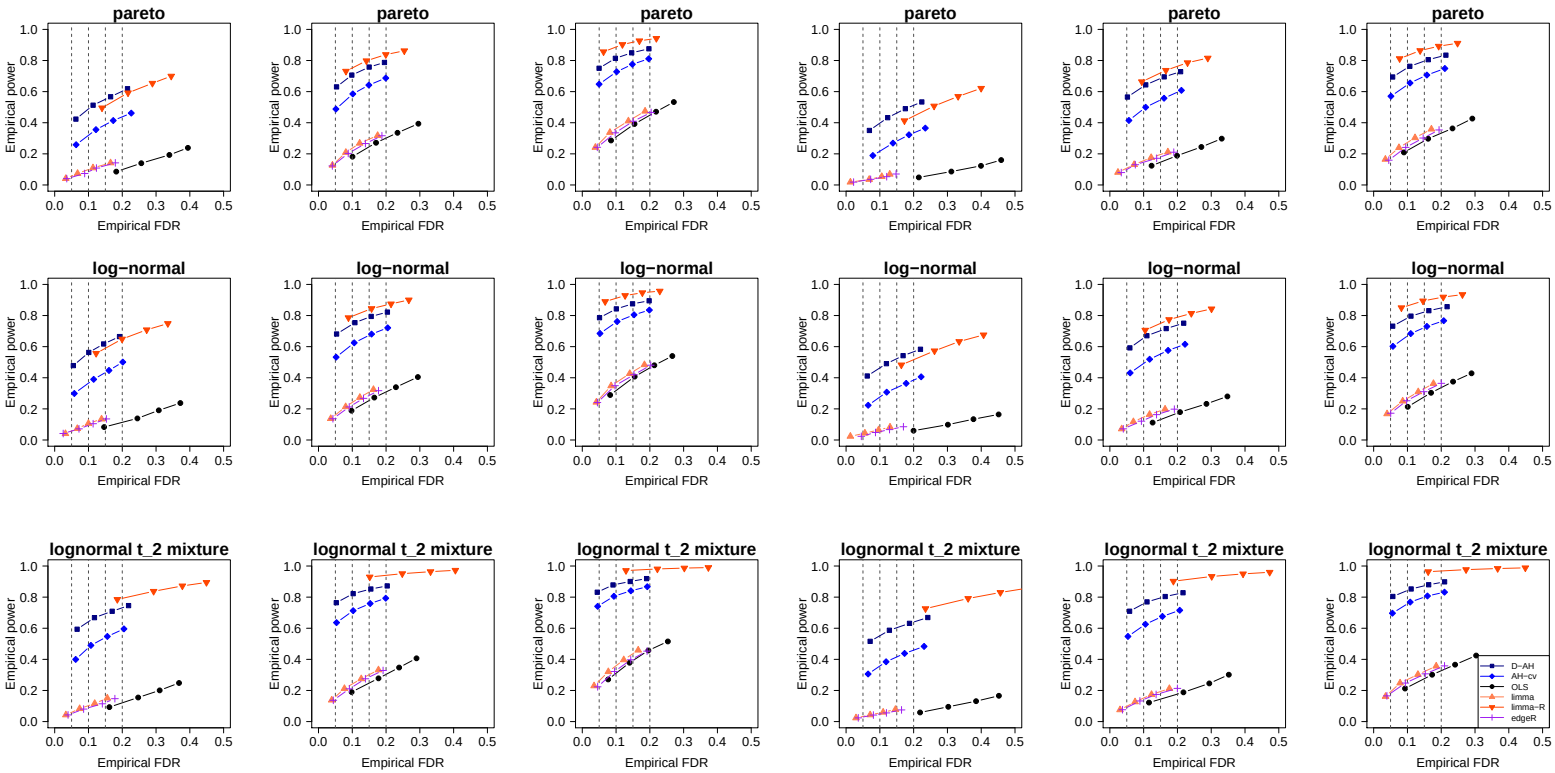
(a) $(n, p) = (85, 1000)$ (b) $(n, p) = (120, 1000)$ (c) $(n, p) = (150, 1000)$ (d) $(n, p) = (85, 2000)$ (e) $(n, p) = (120, 2000)$ (f) $(n, p) = (150, 2000)$

Figure S8: Empirical FDR and power for testing *Hypothesis* 1, a single contrast, under *Model* 3 (banding dependence in errors) with $d = 6$ by our procedure with data-adaptive Huber regression (D-AH, ■); our procedure with cross-validation (AH-cv, ⧫); the ordinary least square estimator (OLS, •); `limma` (▲); `limma` with the robust regression (`limma`-R, ▾); and `edgeR` (+). Each point on the figures displays the empirical FDR and power of the corresponding method at a nominal FDR level, which is marked as a vertical gray dashed line. The pre-specified FDRs are $0.05, 0.1, 0.15, 0.2$.



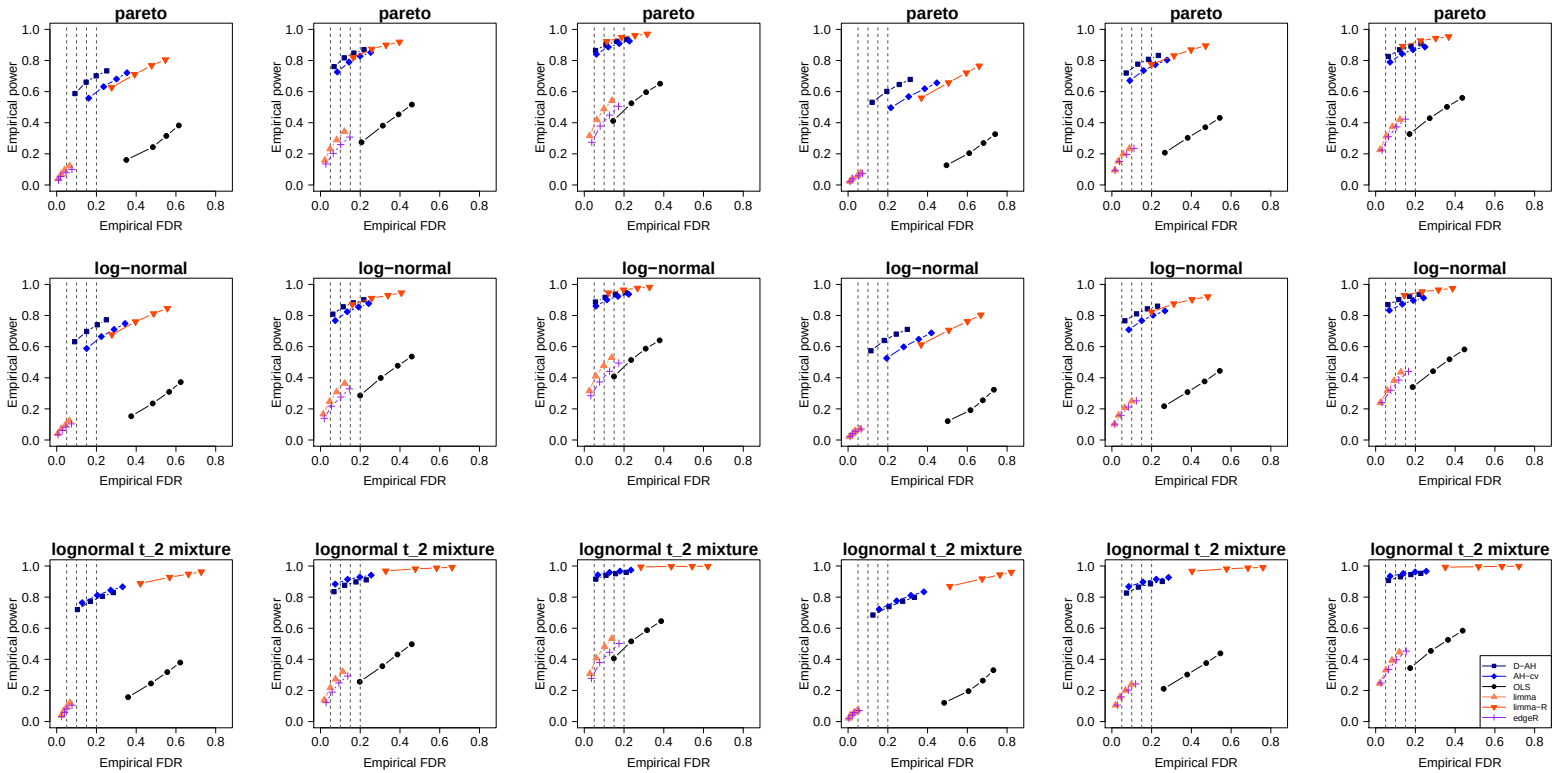
(a) $(n, p) = (85, 1000)$ (b) $(n, p) = (120, 1000)$ (c) $(n, p) = (150, 1000)$ (d) $(n, p) = (85, 2000)$ (e) $(n, p) = (120, 2000)$ (f) $(n, p) = (150, 2000)$

Figure S9: Empirical FDR and power for testing *Hypothesis* 2 under *Model* 3 (banding dependence in errors) with $d = 6$ by our procedure with data-adaptive Huber regression (D-AH, ■); our procedure with cross-validation (AH-cv, ♦); the ordinary least square estimator (OLS, •); `limma` (▲); `limma` with the robust regression (`limma`-R, ▾); and `edgeR` (+). Each point on the figures displays the empirical FDR and power of the corresponding method at a nominal FDR level, which is marked as a vertical gray dashed line. The pre-specified FDRs are $0.05, 0.1, 0.15, 0.2$.



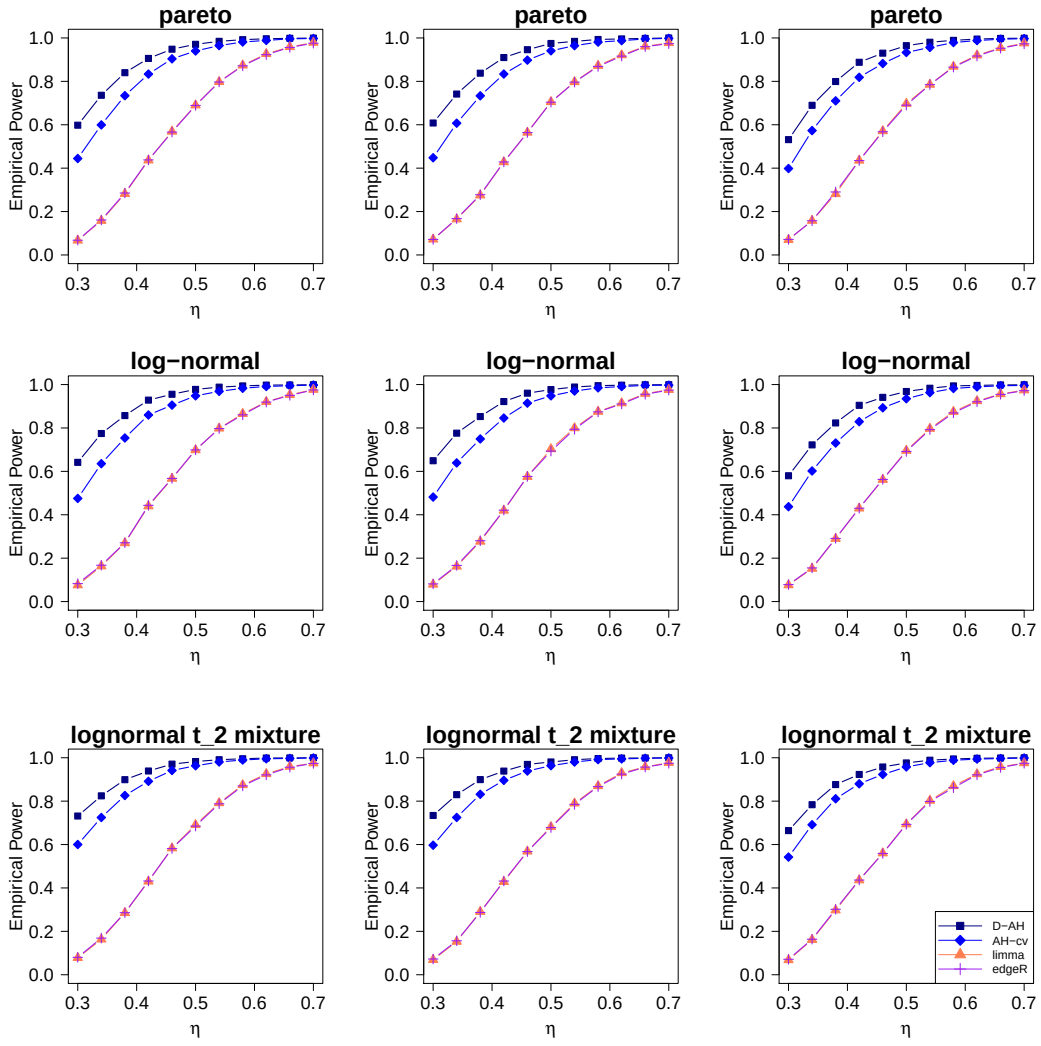
(a) $(n, p) = (85, 1000)$ (b) $(n, p) = (120, 1000)$ (c) $(n, p) = (150, 1000)$ (d) $(n, p) = (85, 2000)$ (e) $(n, p) = (120, 2000)$ (f) $(n, p) = (150, 2000)$

Figure S10: Empirical FDR and power for testing *Hypothesis* 1, a single contrast, under *Model* 3 (banding dependence in errors) with $d = 8$ by our procedure with data-adaptive Huber regression (D-AH, ■); our procedure with cross-validation (AH-cv, ♦); the ordinary least square estimator (OLS, •); `limma` (▲); `limma` with the robust regression (`limma`-R, ▾); and `edgeR` (+). Each point on the figures displays the empirical FDR and power of the corresponding method at a nominal FDR level, which is marked as a vertical gray dashed line. The pre-specified FDRs are $0.05, 0.1, 0.15, 0.2$.



(a) $(n, p) = (85, 1000)$ (b) $(n, p) = (120, 1000)$ (c) $(n, p) = (150, 1000)$ (d) $(n, p) = (85, 2000)$ (e) $(n, p) = (120, 2000)$ (f) $(n, p) = (150, 2000)$

Figure S11: Empirical FDR and power for testing *Hypothesis* 2 under *Model* 3 (banding dependence in errors) with $d = 8$ by our procedure with data-adaptive Huber regression (D-AH, ■); our procedure with cross-validation (AH-cv, ♦); the ordinary least square estimator (OLS, •); `limma` (▲); `limma` with the robust regression (`limma`-R, ▾); and `edgeR` (+). Each point on the figures displays the empirical FDR and power of the corresponding method at a nominal FDR level, which is marked as a vertical gray dashed line. The pre-specified FDRs are $0.05, 0.1, 0.15, 0.2$.



(a) *Model* 1 (b) *Model* 2 (c) *Model* 3

Figure S12: Empirical powers for testing *Hypothesis* 1, a single contrast, with $\eta = \{0.30, 0.34, \ldots, 0.66, 0.7\}$, $n = 100$, $d = 6$, and $p = 1000$ by our procedure with data-adaptive Huber regression (D-AH, ■) or with cross-validation (AH-cv, ♦); `limma` (▲); and `edgeR` (+).

S5.2 Simulations on comparing our method and naive approaches

In this section, we compare our method with some straightforward "naive" approaches that have been widely employed in practice. Specifically, the "naive" approaches are the ordinary least squares estimators with the log or the Box-Cox transformation on responses $\mathbf{Y}$. For the Box-Cox transformation, we select the transformation exponent λ_j for the jth coordinate of response $\mathbf{Y}$ from $\{-2, -1.9, \ldots, 1.9, 2\}$ using `MASS::boxcox()`. We consider $n = 120$, $p = 1000$, $d = 6$, for testing *Hypothesis* 2 under *Model* 1 (details are referred to Section S5.1). As displayed in Figure S13, the proposed method outperforms both "naive" approaches in terms of controlling the FDR and providing satisfactory power across all three error settings.

Next, we compare our method with the above "naive" approaches using the real data. Similar to the simulation studies, for the word-wise Box-Cox transformation, we select the transform exponent $\lambda_j \in \{-2, -1.9, \ldots, 1.9, 2\}$ for the jth word. For comparisons, we consider Hypothesis CDD1 (details are referred to Section 5 in the main paper). Figure S14 displays the number of differentially represented words by all three methods. Most of the differentially represented words identified by the "naive" approaches are also identified by our method. On the other hand, the proposed method identifies a large number of differentially represented words, which are missed by the "naive" approaches. Together with the simulation studies above, this reflects that the

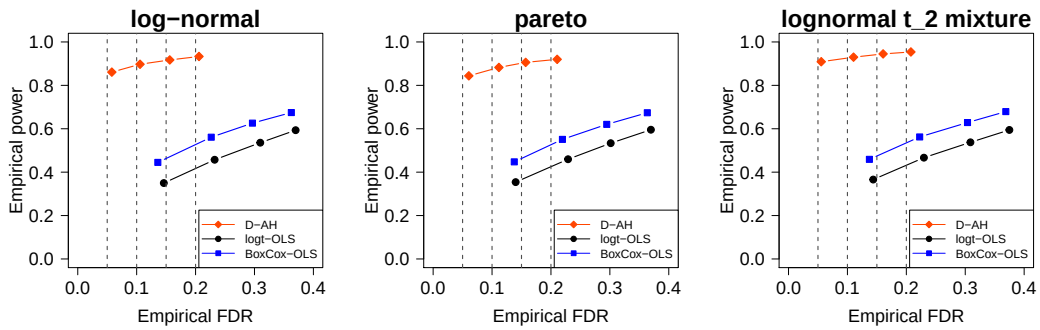


Figure S13: Empirical FDR and power for testing *Hypothesis* 2 under *Model* 1 with $n = 120$, $p = 1000$, and $d = 6$ by our procedure with data-adaptive Huber regression (D-AH, ⧫) and the ordinary least squares estimators with the log-transformed response (logt-OLS, •) or the (column-wise) Box-Cox transformed $\mathbf{Y}$ (BoxCox-OLS, ■). Each point on the figures displays the empirical FDR and power of the corresponding method at a nominal FDR, which is marked as a vertical gray dashed line. The pre-specified nominal FDRs are $0.05, 0.1, 0.15, 0.2$.

"naive" methods compromise powers in practice. From these two numerical studies, we observe that neither the log nor the Box-Cox transformation are able to fully address challenges brought by the heavy-tailed and/or skewed errors.

S5.3 Numerical studies on the sensitivity of our method on h_n

In this section, we conduct a small numerical experiment to study the sensitivity of the empirical FDR and power of our method on the choice of h_n. Specifically, we focus on $n = 120$, $p = 1000$, and $d = 6$ for testing *Hypothesis* 2 under *Model* 1 (details are referred to Section S5.1). We consider $h_n \in \{0.25n^{-1/4}, 0.5n^{-1/4}, n^{-1/4}, 2n^{-1/4}, 4n^{-1/4},$

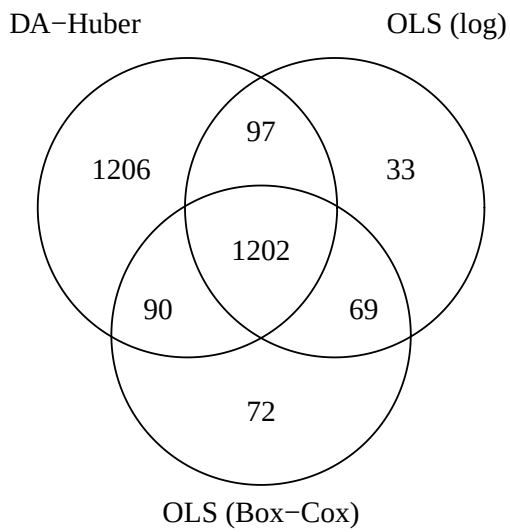


Figure S14: Comparing the proposed method with the "naive" approaches (OLS with log-transformed data or OLS with the Box-Cox transformed data) for testing Hypothesis CDD1 in Section 5 in the main paper. The Venn diagram displays the number of differentially represented words identified by each method with the nominal FDR controlled at 0.5%.

$0.1\tau_{0j}n^{-1/4}\}$, where τ_{0j} is determined as in τ_j by the adaptive Huber regression (Wang et al., 2021). Figure S15 displays both the empirical FDR and the power of our method with different choices of h_n. We observe that our method is reasonably stable with respect to a wide range of h_n, which suggests that in practice our method is not very sensitive to the choice of h_n.

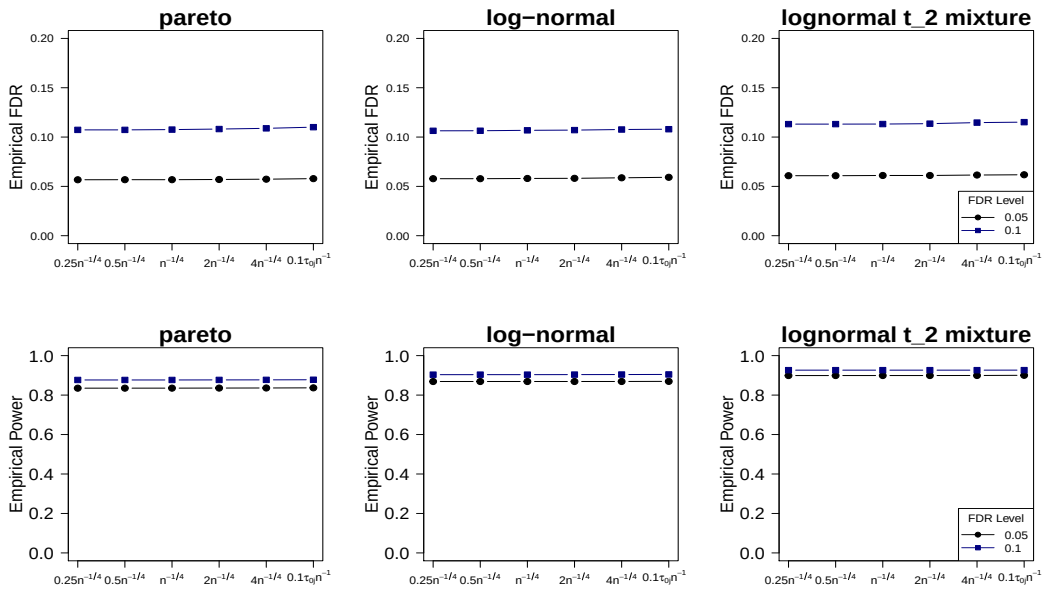


Figure S15: Empirical FDR and power for testing *Hypothesis* 2 under *Model* 1 with $n = 120$, $p = 1000$, and $d = 6$ by our procedure with data-adaptive Huber regression. Different choices of h_n are considered and the nominal FDR levels are 0.05 (•) and 0.1 (■). Each point displays the empirical FDR or power for the corresponding choice of h_n.

S5.4 Simulations with a "real correlation matrix"

We conclude this section with an extra small numerical experiment, where the dependence of errors is generated from the real data in Section 5. Using this setting, we study the performance of our method when Condition 1 (iv) is not necessarily satisfied. First, we randomly select 1000 words from the data set of Shakespeare's work discussed in Section 5 in the main paper. Upon adjusting for the effects of subject groups based on the 176 books, we use the fitted residuals (a 1000×176 residual matrix) to compute a

sample correlation matrix, which is further used as the input to the GLasso to estimate the dependence structure. Figure S16 shows a submatrix of the estimated correlation

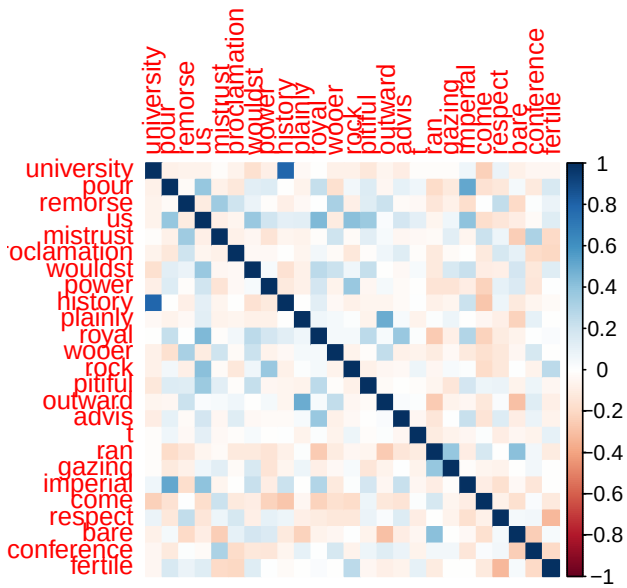


Figure S16: Heatmap of a submatrix (for 25 words) of the estimated correlation matrix for the 1000 randomly selected words from the data set of Shakespeare's work in Section 5 in the main paper.

matrix (with a subset of 25 words), which is relatively dense and will be used as $\mathbf{R}_\epsilon$ in settings described in Section 4 in the main paper. For the simulations, we consider $n = 150$, $p = 1000$, and $d = 6$ for testing *Hypothesis* 2 (details are referred to Section 4 in the main paper). As displayed in Figure S17, the proposed method still outperforms the competitors in terms of controlling the FDR and providing outstanding power simultaneously when the dependence of errors does not necessarily satisfy Condition 1

(iv). Together with simulation results for Model 2 in Section 4, this numerical experiment shows that our method is reliable in more general scenarios even Condition 1 (iv) is mildly violated.

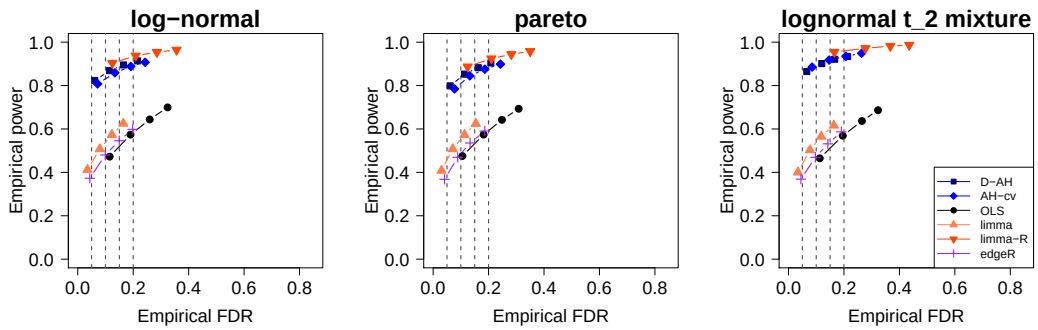


Figure S17: Empirical FDR and power for testing *Hypothesis* 2 under the correlation matrix estimated using the fitted residuals for 1000 randomly selected words from the data set of Shakespeare's work in Section 5 in the main paper. Here, $n = 150$, $p = 1000$, and $d = 6$. Methods for comparisons are: our procedure with data-adaptive Huber regression (D-AH, ■); our procedure with cross-validation (AH-cv, ⧫); the ordinary least square estimator (OLS, •); `limma` (▲); `limma` with the robust regression (`limma`-R, ▾); and `edgeR` (+). Each point on the figures displays the empirical FDR and power of the corresponding method at a nominal FDR, which is marked as a vertical gray dashed line. The pre-specified nominal FDR are $0.05, 0.1, 0.15, 0.2$.

S5.5 Numerical experiment for discussions in Section 2.3

In this section, we report a small simulation study to confirm our discussions in Section 2.3 in the main paper that the variance estimator of $\sigma_{\epsilon,jj}$ in Fan et al. (2019) fails to provide FDP control when $d \geqslant 1$. In simulations, we consider $n = 120$, $p = 1000$, $d = 6$,

for testing *Hypothesis* 2 under *Model* 1 (details are referred to Section S5.1). We adopt the idea from Section 6 of Fan et al. (2019) to estimate $\sigma_{\epsilon,jj}$, that is,

$$\begin{aligned} \widehat{\alpha}_j &:= \underset{\alpha}{\operatorname{argmin}} \sum_{i=1}^{n} \ell_{\tau_{j1}}\{(Y_{ij} - \mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_j) - \alpha\} \\ \widehat{\gamma}_j &:= \underset{\gamma \geqslant \widehat{\alpha}_j^2}{\operatorname{argmin}} \sum_{i=1}^{n} \ell_{\tau_{j2}}\{(Y_{ij} - \mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_j)^2 - \gamma\} \\ \widetilde{\sigma}_{\epsilon,jj} &:= \widehat{\gamma}_j - \widehat{\alpha}_j^{\,2}, \end{aligned} \tag{E.1}$$

where τ_{j1} and τ_{j2} are both selected by the data driven procedure in (Wang et al., 2021). Similar to Section 4 in the main paper, we compare both the empirical FDR and the power between our method and two alternatives, where $\widehat{\boldsymbol{\Sigma}}_j$ is estimated by $\widetilde{\sigma}_{\epsilon,jj}\widehat{\boldsymbol{\Sigma}}_Z^{-1}$ (*FARM-1*), which is an OLS-like covariance estimator with $\widehat{\sigma}_j^2$ replaced by $\widetilde{\sigma}_{\epsilon,jj}$ from (E.1), or $\widetilde{\sigma}_{\epsilon,jj}\mathbf{W}^{-1}\widehat{\boldsymbol{\Sigma}}_Z\mathbf{W}^{-1}$ (*FARM-2*), which is the covariance in (2.6) with $\widehat{\sigma}_{\epsilon,jj}$ replaced by $\widetilde{\sigma}_{\epsilon,jj}$. From Figure S18, we observe that our method provides much better finite sample performance than the alternatives for testing many general linear hypothesis under a variety of heavy-tailed/skewed error distributions. *FARM-1* and *FARM-2* either fail to control the FDR or substantially compromise the power.

S6 Addition results for the analysis on Project Gutenberg

In this section, we present addition details for analyzing data from the Standardized Project Gutenberg Corpus (SPGC) as described in Section 5 in the main article. Table

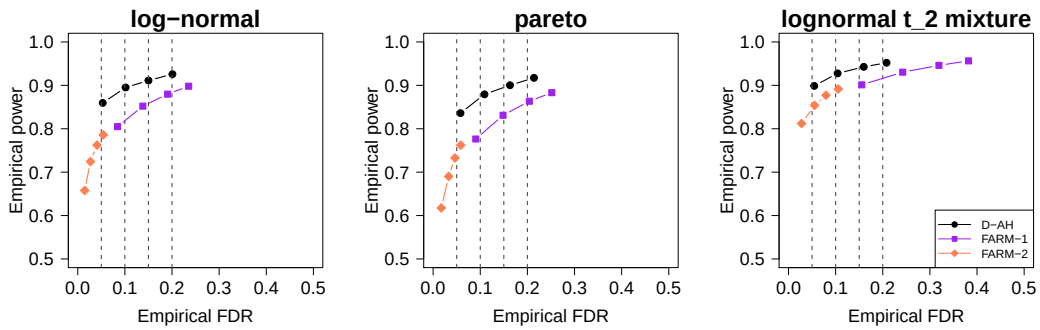
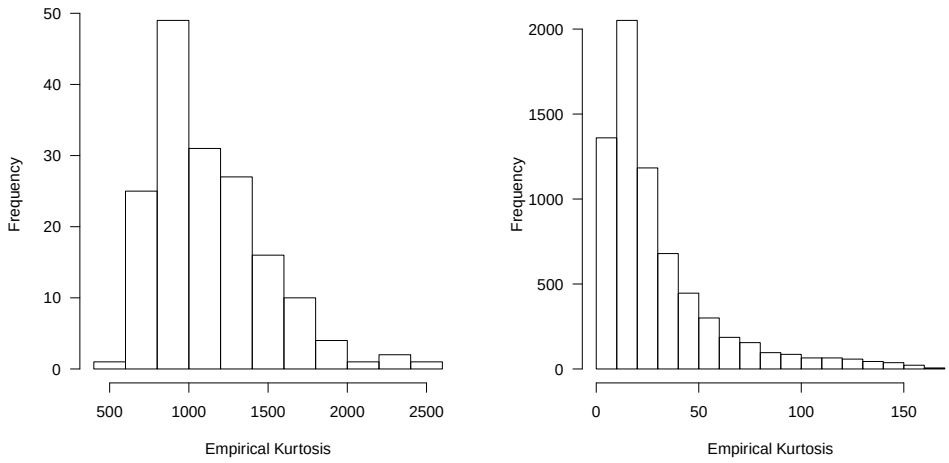


Figure S18: Empirical FDR and power for testing *Hypothesis* 2 under *Model* 1 with $n = 120$, $p = 1000$, and $d = 6$ by our procedure with data-adaptive Huber regression (D-AH, •), our procedure with $\widetilde{\sigma}_{\epsilon,jj}\widehat{\mathbf{\Sigma}}_Z^{-1}$ (*FARM-1*, ■), and our procedure with $\widetilde{\sigma}_{\epsilon,jj}\mathbf{W}^{-1}\widehat{\mathbf{\Sigma}}_Z\mathbf{W}^{-1}$ (*FARM-2*, ♦), where $\widetilde{\sigma}_{\epsilon,jj}$ is defined by (E.1). Each point on the figures displays the empirical FDR and power of the corresponding method at a nominal FDR level, which is marked as a vertical gray dashed line. The pre-specified nominal FDRs are $0.05, 0.1, 0.15, 0.2$. Error distributions are displayed in the plot captions.

S1 displays a snapshot of the raw word count data. The empirical kurtosis of the normalized data is reported in Figure S19, which provides the evidence of heavy tailedness of the data.

The word counts displayed in Table S1 agrees with the Zipf's law (Zipf, 1949), that is the frequency of a word in a corpus is inversely proportional to its rank in the frequency table. A few topic-related words or proper nouns are more frequently encountered in certain works. For example, *A Christmas Carol* has more "Christmas" than other books, and *Oliver Twist* has a substantially higher frequency of "Oliver" than others. In addition, the raw word count data matrix is sparse and consists of 62751



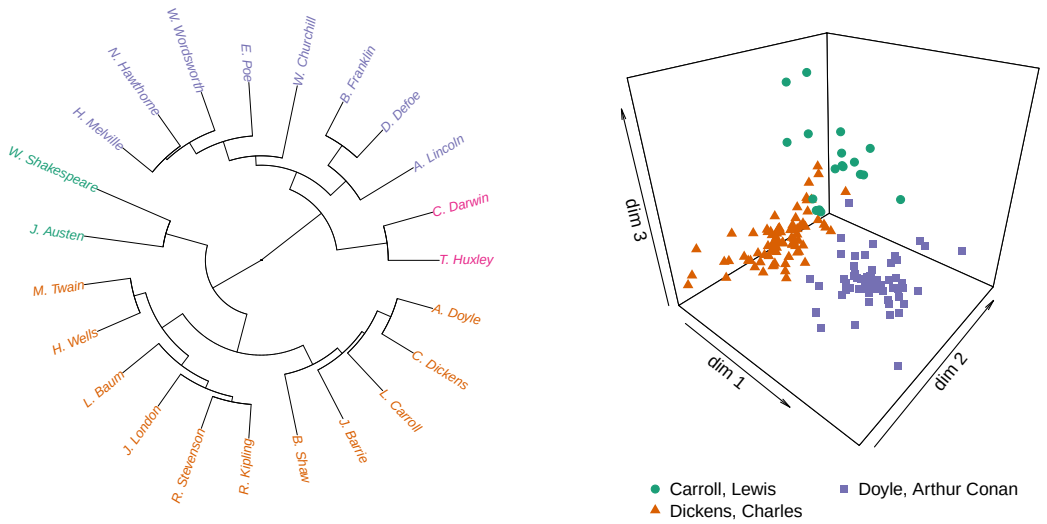
(a) The empirical kurtosis by books.

(b) The empirical kurtosis by words.

Figure S19: The empirical kurtosis of words counts for 167 books (panel (a)) and the empirical kurtosis of counts for 6839 words (panel (b)) from the works of Lewis Carroll, Charles Dickens, and Arthur Conan Doyle. The normalized counts are used.

unique English words. Most of the words have zero counts, 89% of them are removed by the filtering process in Section 5 in the main paper accordingly. Upon filtering, 51% of all the entries in the normalized count matrix are zero, and 82% of them are below 5.

Figure S20 (a) displays a hierarchical clustering result for 23 authors from U.K. and U.S. in the original SPGC data. We observe that Charles Darwin and Thomas H. Huxley were closely related and, as a matter of fact, they are both English biologists in the nineteenth century who focused on the evolution theory. Hence, in terms of the word count distributions, their writings are more similar to each other and distinguishable



(a) Hierarchical clustering for splitting 23 authors into four groups. Each color represent a group.

(b) Multidimensional scaling plot of works by Lewis Carroll, Arthur Conan Doyle, and Charles Dickens. Points represent books and symbols stands for authors. Data are post-processing.

Figure S20: Exploratory displays of the data.

compared to other authors. In addition, Lewis Carroll, Arthur Conan Doyle, and Charles Dickens are closely related from Figure S20 (a). From Figure S20 (b), we notice that the works among Lewis Carroll, Arthur Conan Doyle, and Charles Dickens are separated in general.

The Venn diagram in Figure S21 displays the number of differentially represented words for hypotheses considered in the first application in Section 5 in the main paper. For example, Dickens has 949 differentially represented words that distinguish him from the other two authors. Among those 949 words, “catch” and “curious” appear to be the

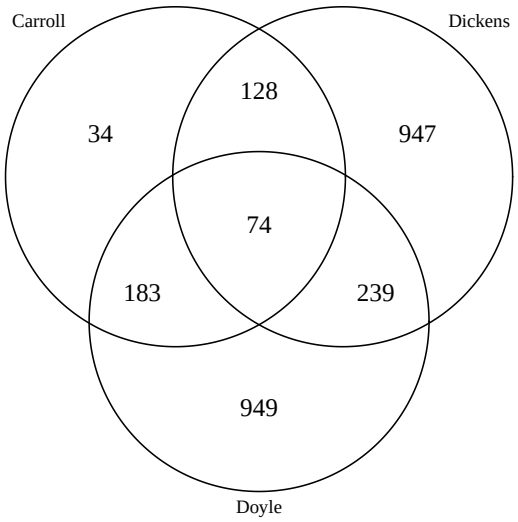


Figure S21: Comparing word counts of books of Lewis Carroll, Charles Dickens, and Arthur Conan Doyle by our method with the nominal false discovery rate controlled at 0.5%. The Venn diagram displays the number of differentially represented words for Hypothesis CDD2 (Carroll), Hypothesis CDD3 (Dickens), and Hypothesis CDD4 (Doyle).

most significant whereas “clock”, “horseback”, and “present” are the least significant ones. Further quantitative linguistic or literature investigations are required to uncover more insights on these identified differentially represented words.

References

Bentkus, V. (2005). A Lyapunov-type bound in $\mathbb{R}^d$. *Theory of Probability and Its Applications* **49**, 311–323.

Devroye, L., Mehrabian, A., and Reddad, T. (2018). The total variation distance between high-dimensional Gaussians. *Preprint*. https://arxiv.org/abs/1810.08693.

Horn, R. A. and Johnson, C. R. (1990). *Matrix Analysis*. Cambridge University Press, Cambridge, United Kingdom.

Li, W. V. and Shao, Q.-M. (2002). A normal comparison inequality and its applications. *Probability Theory and Related Fields* **122**, 494–508.

Rudin, W. (1976). *Principles of Mathematical Analysis*. McGraw-Hill, New York, NY.

Spokoiny, V. and Zhilova, M. (2015). Bootstrap confidence sets under model misspecification. *The Annals of Statistics* **43**, 2653–2675.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B* **64**, 479–498.

Sun, Q., Zhou, W.-X., and Fan, J. (2020). Adaptive Huber regression. *Journal of the American Statistical Association* **115**, 254–265.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press, Cambridge, United Kingdom.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, United Kingdom.

Wang, L., Zheng, C., Zhou, W., and Zhou, W.-X. (2021). A new principle for tuning-free Huber regression. *Statistica Sinica* **31**, 2153–2177.

Zhou, W.-X., Bose, K., Fan, J., and Liu, H. (2018). A new perspective on robust M-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *The Annals of Statistics* **46**, 1904–1931.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology.* Ravenio Books.

word/PG id	PG19033	PG730	PG1400	PG24022	PG244	PG2097	PG2852
the	636	9493	8143	1595	2569	2335	3330
and	337	5239	7071	1046	1368	1179	1628
of	201	3852	4433	696	1215	1122	1594
to	249	3852	5071	676	1093	1079	1408
a	277	3702	4040	709	1005	1092	1306
i	160	1357	6632	280	938	1219	1497
jolly	0	7	20	3	0	0	0
king	25	4	14	0	0	0	0
loss	0	20	12	1	4	3	4
colour	0	5	0	4	5	0	6
oliver	0	859	0	0	0	0	1
shaded	0	0	5	0	0	2	1
alice	172	0	0	0	5	0	0
murder	1	19	19	0	18	10	6
christmas	0	1	9	85	0	0	0

Table S1: Snapshot of the raw data in SPGC. PG id's represent different books: *Alice's Adventures in Wonderland* (PG19033) by Lewis Carroll, *Oliver Twist* (PG730), *Great Expectations* (PG1400), and *A Christmas Carol* (PG24022) by Charles Dickens, and *A Study in Scarlet* (PG244), *The Sign of the Four* (PG2097), and *The Hound of the Baskervilles* (PG2852) by Arthur Conan Doyle.