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

Youngseok Song\textsuperscript{5}, Wen Zhou\textsuperscript{6}, and Wen-Xin Zhou\textsuperscript{7}

\textsuperscript{5}Ecole Polytechnique Fédérale de Lausanne, \textsuperscript{6}Colorado State University
\textsuperscript{7}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

$$p_0^{-1}V^o(z^2) = \mathbb{P}(\chi_q^2 > z^2) + O_p(q^{1/2}n^{-\kappa_1} + n^{-1/2}q^{7/4} + q[n^{-1}\{\log(np) + d\}]^{\delta/(2+\delta)})$$

(A.1)
uniformly over \( z \geq 0 \) as \( n, p \to \infty \). Denote \( \mathbf{A}_j := \mathbf{C} \Sigma_j \mathbf{C}^T \in \mathbb{R}^{q \times q} \) the true covariance matrix of \( n^{1/2} \mathbf{C} (\tilde{\theta}_j - \theta_j) \), 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

\[
\mathbf{T}_j^0 = n^{1/2} \mathbf{A}_j^{-1/2} \mathbf{C} (\tilde{\theta}_j - \theta_j),
\]

where \( V_j^0 = \| \mathbf{T}_j^0 \|^2 \); 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 \( \mathbf{T}_j^0 \) can be approximated by a \( q \)-dimensional multivariate normal distribution, so that \( V_j^0 \) can be approximated by the \( \chi_q^2 \) distribution due to Lemma B.6. Let

\[
\mathbf{S}_j = n^{-1/2} (\mathbf{C} \Sigma^{-1}_Z \mathbf{C}^T)^{-1/2} \mathbf{C} \Sigma^{-1/2}_Z \sum_{i=1}^n \Sigma^{-1/2}_Z \left[ \ell'_j (\epsilon_{ij}) \mathbf{Z}_i - \mathbb{E} \{ \ell'_j (\epsilon_{ij}) \mathbf{Z}_i \} \right],
\]

\[
\mathbf{R}_j = n^{-1/2} (\mathbf{C} \Sigma^{-1}_Z \mathbf{C}^T)^{-1/2} \mathbf{C} \Sigma^{-1/2}_Z \sum_{i=1}^n \Sigma^{-1/2}_Z \mathbb{E} \{ \ell'_j (\epsilon_{ij}) \mathbf{Z}_i \}.
\]

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

\[
\| \mathbf{T}_j^0 - \sigma^{-1/2}_{\epsilon,jj} (\mathbf{S}_j + \mathbf{R}_j) \| \leq C_2 \tau_{0j} \frac{d + t}{(n \sigma_{\epsilon,jj})^{1/2}} \quad (A.4)
\]

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

\[
\sum_{j \in H_0} \mathbb{I} \{ \| \sigma^{-1/2}_{\epsilon,jj} \mathbf{S}_j \| \geq z + C_2 \tau_{0j} (n \sigma_{\epsilon,jj})^{-1/2} (d + t) \}
\leq V^0(z^2)
\]

\[
\leq \sum_{j \in H_0} \mathbb{I} \{ \| \sigma^{-1/2}_{\epsilon,jj} \mathbf{S}_j \| \geq z - C_2 \tau_{0j} (n \sigma_{\epsilon,jj})^{-1/2} (d + t) \}.
\]

(A.5)
with probability \( 1 - 3p \exp(-t) \) as long as \( n \geq C_3(d + t) \).

Define \( D_j = \mathbb{I}(\| \sigma_{r,j}^{-1/2} S_j \| \geq z) \) and \( \mathcal{P}_j = \mathbb{P}(\| \sigma_{r,j}^{-1/2} S_j \| \geq z) \) for \( j = 1, \ldots, p \) and \( z \geq 0 \). Under Condition 1, \( D_1, \ldots, D_p \) are weakly correlated. Recall that \( \mathcal{H}_0 = \{ j : 1 \leq j \leq p, H_{0j} \text{ is true} \} \), it holds
\[
\text{var} \left( p_0^{-1} \sum_{j \in \mathcal{H}_0} D_j \right) = \frac{1}{p_0^2} \sum_{j \in \mathcal{H}_0} \text{var}(D_j) + \frac{1}{p_0^2} \sum_{j,k \in \mathcal{H}_0, j \neq k} \text{cov}(D_j, D_k) \leq \frac{1}{4p_0} + \frac{1}{p_0^2} \sum_{j,k \in \mathcal{H}_0, j \neq k} \left( \mathbb{E}(D_j D_k) - \mathcal{P}_j \mathcal{P}_k \right). \tag{A.6}
\]

We first study \( \mathcal{P}_j \). Note that \( S_j \) is a sum of independent random vectors with \( \mathbb{E}(S_j) = 0 \) and \( \text{cov}(S_j) = s_j^2 I \) where \( s_j^2 = \mathbb{E}[\{ \ell_{r_j}^\prime(\epsilon_{ij}) \}^2] \). Let \( G \sim N(0, I) \) be a standard normal random vector. Lemmas B.2 and B.3 imply that
\[
\max_{1 \leq j \leq p} \left| \mathcal{P}_j - \mathbb{P}(\| G \| \geq z) \right| \leq n^{-1/2} q^{7/4} + q^{1/2} 2 \sqrt{\frac{2 \delta^2 + \delta}{\tau_{0j}}} \left( \frac{d + t}{n} \right)^{\frac{\delta}{2 + \delta}} \tag{A.7}
\]
holds uniformly over \( z \geq 0 \).

Following that, we consider \( \mathbb{E}(D_j D_k) \) for each pair \( (j, k) \) with \( 1 \leq j \neq k \leq p \). Set \( S = (s_j^{-1} S_j, s_k^{-1} S_k)^T \). Let \( G = (G_1, G_2) = (G_{11}, \ldots, G_{1q}, G_{21}, \ldots, G_{2q}) \in \mathbb{R}^{2q} \) be a Gaussian vector with \( \mathbb{E}(G) = 0 \) and \( \text{cov}(G) = \text{cov}(S) \). The block-structured matrix \( \text{cov}(S) \) has unit diagonal entries with \( \text{cov}(s_j^{-1} S_j) = I \). Also, \( \text{cov}(s_j^{-1} S_j, s_k^{-1} S_k) = (n s_j s_k)^{-1} \sum_{i=1}^n \text{cov}(\ell_{r_j}^\prime(\epsilon_{ij}), \ell_{r_k}^\prime(\epsilon_{ik})) I \) and
\[
\left| (n s_j s_k)^{-1} \sum_{i=1}^n \text{cov}(\ell_{r_j}^\prime(\epsilon_{ij}), \ell_{r_k}^\prime(\epsilon_{ik})) - r_{\epsilon,jk} \right| \leq \frac{\delta^2}{\tau_{0j}} \left( \frac{d + t}{n} \right)^{\frac{\delta}{2 + \delta}} \tag{A.8}
\]
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 \( 1 \leq j \neq k \leq p \). Putting together (A.8), Condition 1 (iv),
Corollary B.1, and Lemmas B.2 and B.5 with our choice on $\tau_j$ yield

$$|\mathbb{P}(\|\sigma_{\epsilon_{jj}}^{1/2} s_j G_1\| \geq x, \|\sigma_{\epsilon_{kk}}^{1/2} s_k G_2\| \geq x) - \mathbb{P}(\|Z_1\| \geq x) \mathbb{P}(\|Z_2\| \geq x)|$$

$$\leq \mathbb{P}(\|G_1\| \geq \sigma_{\epsilon_{jj}}^{1/2} s_j^{-1} x, \|G_2\| \geq \sigma_{\epsilon_{kk}}^{1/2} s_k^{-1} x) - \mathbb{P}(\|Z_1\| \geq x) \mathbb{P}(\|Z_2\| \geq x)$$

$$+ \mathbb{P}(\|\sigma_{\epsilon_{jj}}^{-1/2} s_j Z_1\| \geq x) \mathbb{P}(\|\sigma_{\epsilon_{kk}}^{-1/2} s_k Z_2\| \geq x) - \mathbb{P}(\|Z_1\| \geq x) \mathbb{P}(\|Z_2\| \geq x)$$

$$\lesssim q^{1/2} |r_{\epsilon,jk}| + q \{n^{-1}(d + t)\}^{\delta(2+\delta)}.$$

It follows that

$$|\mathbb{P}(\|G_1\| \geq s_j^{1/2} \sigma_{\epsilon_{jj}}^{-1/2} z, \|G_2\| \geq s_k^{1/2} \sigma_{\epsilon_{kk}}^{-1/2} z) - \mathbb{P}(\|Z\| \geq z)\}^2 | \lesssim q^{1/2} |r_{\epsilon,jk}| + q \left(\frac{d + t}{n}\right)^{\delta(2+\delta)}$$

(A.9)

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

$$\sup_{x, y \in \mathbb{R}} |\mathbb{P}(\|s_j^{-1} S_j\| \geq x, \|s_k^{-1} S_k\| \geq y) - \mathbb{P}(\|G_1\| \geq x, \|G_2\| \geq y)| \lesssim n^{-1/2} q^{7/4},$$

which implies

$$|\mathbb{E}(D_j D_k) - \mathbb{P}(\|G_1\| > s_j^{-1} \sigma_{\epsilon_{jj}}^{1/2} z, \|G_2\| > s_k^{-1} \sigma_{\epsilon_{kk}}^{1/2} z)| \lesssim n^{-1/2} q^{7/4}. \quad \text{(A.10)}$$

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

$$|\mathbb{E}(D_j D_k) - \{\mathbb{P}(\|Z_1\| > z)\}^2 | \lesssim q^{1/2} |r_{\epsilon,jk}| + n^{-1/2} q^{7/4} + q \left(\frac{d + t}{n}\right)^{\delta(2+\delta)}. \quad \text{(A.11)}$$

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

$$\mathbb{E}[\{P_0^{-1} V^o(z^2) - \{\mathbb{P}(\chi_q^2 > z^2)\}^2 \} \lesssim q^{1/2} p^{-n_1} + n^{-1/2} q^{7/4} + q \left(\frac{d + t}{n}\right)^{\delta(2+\delta)} \quad \text{(A.12)}$$

on $E_1(t)$. Recall that $\mathbb{P}\{E_1(t)\} = 1 - 3p \exp(-t)$ as long as $n \geq C_3(d + t)$. Taking $t = \log(np)$ in (A.5) and (A.12) proves (A.1).
S1.2 Proof of Theorem 1

For statistic $\tilde{V}_j = n(C\tilde{\theta}_j - c_{0j})^T(C\tilde{\Sigma}_jC^T)^{-1}(C\tilde{\theta}_j - c_{0j})$, it holds

$$|\tilde{V}_j - V^*_j| = |n(C\tilde{\theta}_j - c_{0j})^T(C\tilde{\Sigma}_jC^T)^{-1}(C\tilde{\theta}_j - c_{0j}) - n(C\tilde{\theta}_j - c_{0j})^T(C\Sigma_jC^T)^{-1}(C\tilde{\theta}_j - c_{0j})|$$

$$\leq n\|\Sigma_{Z_j}^{1/2}(\tilde{\theta}_j - \theta_j)\|^2\Sigma_{Z_j}^{-1/2}C^T\{(C\tilde{\Sigma}_jC^T)^{-1} - (C\Sigma_jC^T)^{-1}\}C\Sigma_{Z_j}^{-1/2}\|$$

$$\leq n\|\Sigma_{Z_j}^{1/2}(\tilde{\theta}_j - \theta_j)\|^2|C\Sigma_{Z_j}^{-1/2}((C\tilde{\Sigma}_jC^T)^{-1} - (C\Sigma_jC^T)^{-1}|$$

$$\leq n\|\Sigma_{Z_j}^{1/2}(\tilde{\theta}_j - \theta_j)\|^2|C\Sigma_{Z_j}^{-1/2} - (C\Sigma_jC^T)^{-1}|.$$  

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

$$\max_{j \not\in \mathcal{H}_0} |\tilde{V}_j - V^*_j| \lesssim \{\log(np) + d\} \max_{1 \leq j \leq p} ||\tilde{\Sigma}_j - \Sigma_j||$$

(A.13)

with probability $1 - 2n^{-1}$ and the right hand side of (A.13) is $o_P(1)$. Combining this with the proof of Proposition 1 and Condition 1, we obtain $p_0^{-1}\tilde{V}(z) = P(\chi^2_q > z) + o_P(1)$.

Similarly, $R(z)$ can be replaced by $\tilde{R}(z)$. Consequently, $|\text{FDP}(z) - \text{AFDP}(z)| = o_P(1)$ as $n, p \to \infty$.

S1.3 Proof of Theorem 2

Recall that $m_j = n^{-1}\sum_{i=1}^n I_{\tau_j}(e_{ij})$, $W_j = n^{-1}\sum_{i=1}^n I_{\tau_j}(e_{ij})Z_iZ_i^T$, and $K_j = 1 + (nm_j)^{-1}(d + 1)(1 - m_j)$. Denote $A_{jn} := \{W_j^{-1}(n^{-1}\sum_{i=1}^n Z_iZ_i^T)W_j^{-1}\}^{-1}$. For each
j = 1, \ldots, p,

\[ \| \hat{\Sigma}_j - \Sigma_j \| = \left\| \frac{1}{K_j} \left[ \frac{1}{n-d-1} \sum_{i=1}^{n} \{ \ell'_r(e_{ij}) \}^2 \right] \mathbf{A}_{jn}^{-1} - \sigma_{e,jj} \Sigma_z^{-1} \right\| \]

\[ \leq \left\| \frac{1}{K_j} \left[ \frac{1}{n-d-1} \sum_{i=1}^{n} \{ \ell'_r(e_{ij}) \}^2 \right] \mathbf{A}_{jn}^{-1} - \Sigma_z^{-1} \right\| + \left\| \frac{1}{K_j} \left[ \frac{1}{n-d-1} \sum_{i=1}^{n} \{ \ell'_r(e_{ij}) \}^2 \right] - \sigma_{e,jj} \right\| \| \Sigma_z^{-1} \|. \]  

(A.14)

Note \( n/(n-d-1) \to 1 \) as \( n \to \infty \). Denote \( K_j^{-1}n^{-1} \sum_{i=1}^{n} \{ \ell'_r(e_{ij}) \}^2 \) by \( f(m_j, y_j) \), where 

\( y_j = n^{-1} \sum_{i=1}^{n} \{ \ell'_r(e_{ij}) \}^2 \). That is, 

\[ f(m_j, n^{-1} \sum_{i=1}^{n} \{ \ell'_r(e_{ij}) \}^2) = K_j^{-1}n^{-1} \sum_{i=1}^{n} \{ \ell'_r(e_{ij}) \}^2 \]

with \( f(1, \sigma_{e,jj}) = \sigma_{e,jj} \). It holds that \( f(m_j, y_j) \) is twice differentiable at \((1, \sigma_{e,jj})\), where

\[ \frac{\partial}{\partial y_j} f(m_j, y_j) = K_j^{-1}, \quad \frac{\partial}{\partial y_j} f(m_j, y_j)|_{(m_j, y_j) = (1, \sigma_{e,jj})} = 1, \]

\[ \frac{\partial}{\partial m_j} f(m_j, y_j) = \frac{n(d+1)}{nm_j + (d+1)(1-m_j)} y_j, \quad \frac{\partial}{\partial m_j} f(m_j, y_j)|_{(m_j, y_j) = (1, \sigma_{e,jj})} = \frac{d+1}{n} \sigma_{e,jj} \]

and

\[ \frac{\partial^2}{\partial y_j^2} f(m_j, y_j) = 0, \quad \frac{\partial^2}{\partial m_j \partial y_j} f(m_j, y_j) = \frac{n(d+1)}{nm_j + (d+1)(1-m_j)} y_j, \]

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

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

\[ \frac{1}{K_j} \left[ \frac{1}{n} \sum_{i=1}^{n} \{ \ell'_r(e_{ij}) \}^2 \right] - \sigma_{e,jj} = \left[ \frac{1}{n} \sum_{i=1}^{n} \{ \ell'_r(e_{ij}) \}^2 - \sigma_{e,jj} \right] + \frac{d+1}{n} \sigma_{e,jj}(m_j - 1) \]

\[ + R_1(1 + h_1, \sigma_{e,jj}^2 + h_2), \]

where \( R_1(\cdot, \cdot) \) is the remainder and satisfies \( \lim_{h \to 0} R_1(1 + h_1, \sigma_{e,jj}^2 + h_2)/\| h \| = 0 \) for 

\( h = (h_1, h_2) = c \{ (m_j, y_j) - (1, \sigma_{e,jj}) \} \) and \( c \in (0, 1) \) since \( f \) is twice differentiable (for
example, see Rudin, 1976). On the event $A_\Delta$ defined in Lemma B.9 with $P(A_\Delta) \geq 1 - 4n^{-1}$, Lemmas B.10 and B.11 imply

$$\max_{1 \leq j \leq p} \left| \frac{1}{K_j} \left[ \frac{1}{n} \sum_{i=1}^{n} \{ \ell'_{ij}(e_{ij}) \} \right] - \sigma_{e_{jj}} \right| \leq C_5 \left\{ \frac{\log(np) + d}{n} \right\}^{\delta/(2+\delta)} + C_6 \frac{d + 1}{n} \max \left\{ \left\{ \frac{\log(np) + d}{n} \right\}^{1/2} \frac{\Delta}{h_n} \right\} + R_1 \left[ 1 + h_1, \sigma_{e_{jj}}^2 + h_2 \right]$$

(A.15)

\begin{align*}
\leq & C \left\{ \frac{\log(np) + d}{n} \right\}^{\delta/(2+\delta)} \\
\text{with probability at least } & 1 - 8n^{-1}, \text{ where } C \text{ is a constant depending on } A_0. \text{ Note that the remainder in (A.15) is dominated by other terms as long as } m_j - 1 \text{ and } n^{-1} \sum_{i=1}^{n} \{ \ell'_{ij}(e_{ij}) \}^2 - \sigma_{e_{jj}}^2 \text{ are small given } d \ll n. \end{align*}

By Lemma B.7, it follows that

$$\| A_{jn}^{-1} - \Sigma_Z^{-1} \| \leq \frac{\| \Sigma_Z^{-1} \|^2}{1 - \| \Sigma_Z^{-1} \| \| A_{jn} - \Sigma_Z \|} \| A_{jn} - \Sigma_Z \|$$

(A.16)

$$= \| \Sigma_Z^{-1} \|^2 \| A_{jn} - \Sigma_Z \| \sum_{k=0}^{\infty} (\| \Sigma_Z^{-1} \| \| A_{jn} - \Sigma_Z \|)^k. $$

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

$$A_{jn} = W_j \hat{\Sigma}_n^{-1} W_j = \hat{\Sigma}_n + 2(W_j - \hat{\Sigma}_n) + (W_j - \hat{\Sigma}_n) \hat{\Sigma}_n^{-1} (W_j - \hat{\Sigma}_n)$$
so that
\[ A_{jn} - \hat{\Sigma}_n = 2(W_j - \hat{\Sigma}_n) + (W_j - \hat{\Sigma}_n)\hat{\Sigma}_n^{-1}(W_j - \hat{\Sigma}_n) \]
\[ = 2(W_j - \hat{\Sigma}_n) + (W_j - \hat{\Sigma}_n)(\hat{\Sigma}_n^{-1} - \Sigma^{-1})(W_j - \hat{\Sigma}_n) \]
\[ + (W_j - \hat{\Sigma}_n)\Sigma^{-1}(W_j - \hat{\Sigma}_n). \]

Hence, the bound for the operator norm of \( A_{jn} - \hat{\Sigma}_n \) is obtained by bounding \( \|W_j - \hat{\Sigma}_n\| \) and \( \|\hat{\Sigma}_n^{-1} - \Sigma^{-1}\| \). By Lemmas B.7 and B.12,

\[
\|A_{jn} - \hat{\Sigma}_n\| \leq 2\|W_j - \hat{\Sigma}_n\| + \|W_j - \hat{\Sigma}_n\|^2 \left[ \|\hat{\Sigma}_n^{-1} - \Sigma^{-1}\| + \lambda_{\min}(\Sigma_Z)\right]^{-1} \leq C \max \left\{ \left\{ \frac{\log(np) + d}{n} \right\}^{1/2} \frac{\Delta}{h_n} \right\} \tag{A.17}
\]

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

\[
\|A_{jn} - \Sigma_Z\| \leq \|A_{jn} - \hat{\Sigma}_n\| + \|\hat{\Sigma}_n - \Sigma_Z\| \leq C \max \left\{ \left\{ \frac{\log(np) + d}{n} \right\}^{1/2} \frac{\Delta}{h_n} \right\} \tag{A.18}
\]

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

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

\[
\max_{1 \leq j \leq p} \|\hat{\Sigma}_j - \Sigma_j\| \leq C \max \left\{ \left\{ \frac{\log(np) + d}{n} \right\}^{\delta/(2+\delta)} \frac{\Delta}{h_n} \right\}
\]

with probability at least \( 1 - 16n^{-1} \) for some positive constant \( C \) depending only on \( \lambda_{\min}(\Sigma_Z), \lambda_{\max}(\Sigma_Z), A_0, \) and \( v_{j,\delta} \).
S2 Auxiliary results

Recall that the first order derivative of the Huber loss is
\[ \ell'_\tau(x) = \begin{cases} x & |x| \leq \tau, \\ \tau \text{sgn}(x) & |x| > \tau \end{cases} \]
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 \( \epsilon_j \) given the existence of higher moments of \( \epsilon_j \).

**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
\[
|\mathbb{E}\ell'_\tau(z)| \leq \min \left\{ \frac{\sigma^2}{\tau}, \frac{\mathbb{E}(|z|^{\kappa})}{\tau^{\kappa-1}} \right\} \quad \text{and} \quad |\mathbb{E}\{\ell'_\tau(z)\}^2 - \sigma^2| \leq \frac{2\mathbb{E}(|z|^{\kappa})}{(\kappa - 2)\tau^{\kappa-2}}.
\]

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

**Corollary B.1.** For \( 1 \leq j \leq p \) and \( v_{j,\delta} = \{\mathbb{E}(|\epsilon_j|^{2+\delta})\}^{1/(2+\delta)} < \infty \), it holds that
\[
\sigma_{\epsilon_{jj}} - \frac{2v_{j,\delta}^{2+\delta}}{\delta \tau_j^{\delta}} \leq s^2_j \leq \sigma_{\epsilon_{jj}}.
\]
Proof. Applying Proposition B.1 with $\kappa = 2 + \delta$ for some $\delta > 0$ yields the first inequality. The second inequality follows $\ell'_{\tau_j}(x)^2 \leq x^2$. \qed

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 $1 \leq j \neq k \leq p$ and $\delta > 0$. Then

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

Proof. By definition,

$$\text{cov}(\epsilon_j, \epsilon_k) = \mathbb{E}\{\epsilon_j\epsilon_k(\mid \epsilon_j \leq \tau_j, \mid \epsilon_k \leq \tau_k)\} + \mathbb{E}\{\epsilon_j\epsilon_k\mid (\mid \epsilon_j \leq \tau_j) + \mid (\mid \epsilon_k \leq \tau_k)\}\}

- \mathbb{E}\{\epsilon_j\epsilon_k\mid (\mid \epsilon_j > \tau_j, \mid \epsilon_k > \tau_k)\},$$

$$\mathbb{E}\{\ell'_{\tau_j}(\epsilon_j)\} = \mathbb{E}\{\epsilon_j\mid (\mid \epsilon_j \leq \tau_j)\} + \tau_j\mathbb{E}\{\epsilon_j\mid (\mid \epsilon_j > \tau_j)\},$$

and

$$\mathbb{E}\{\ell'_{\tau_j}(\epsilon_j)\ell'_{\tau_k}(\epsilon_k)\} = \mathbb{E}\{\epsilon_j\epsilon_k\mid (\mid \epsilon_j \leq \tau_j, \mid \epsilon_k \leq \tau_k)\} + \tau_k\mathbb{E}\{\epsilon_j\epsilon_k\mid (\mid \epsilon_j \leq \tau_j, \mid \epsilon_k > \tau_k)\}

+ \tau_j\mathbb{E}\{\epsilon_j\epsilon_k\mid (\mid \epsilon_j > \tau_j, \mid \epsilon_k \leq \tau_k)\} + \tau_j\tau_k\mathbb{E}\{\epsilon_j\epsilon_k\mid (\mid \epsilon_j > \tau_j, \mid \epsilon_k > \tau_k)\}

= \text{cov}(\epsilon_j, \epsilon_k) - \mathbb{E}\{\epsilon_j\epsilon_k\mid (\mid \epsilon_j > \tau_j) + \mid (\mid \epsilon_k > \tau_k)\}\}

+ \tau_k\mathbb{E}\{\epsilon_j\epsilon_k\mid (\mid \epsilon_j \leq \tau_j, \mid \epsilon_k > \tau_k)\} + \tau_j\mathbb{E}\{\epsilon_j\epsilon_k\mid (\mid \epsilon_j > \tau_j, \mid \epsilon_k \leq \tau_k)\}

+ \mathbb{E}\{\epsilon_j\epsilon_k + \tau_j\tau_k\epsilon_j\epsilon_k\mid (\mid \epsilon_j > \tau_j, \mid \epsilon_k > \tau_k)\}. $$
Note that
\[
| \mathbb{E} \{ \epsilon_j \mathbb{I}(|\epsilon_j| > \tau_j) \} | = | \mathbb{E} \{ \epsilon_j^{1+\delta} \epsilon_j^{-\delta} \mathbb{I}(|\epsilon_j| > \tau_j) \} | \\
\leq (\mathbb{E} |\epsilon_j^{1+\delta}|) \max |\epsilon_j^{-\delta} \mathbb{I}(|\epsilon_j| > \tau_j)| \\
\leq \tau_j^{-\delta} \{ \mathbb{E}(|\epsilon_j|^{2+\delta}) \}^{1/(2+\delta)} \{ \mathbb{E}(|\epsilon_k|^{2+\delta}) \}^{1/(2+\delta)} \\
\leq \tau_j^{-\delta} \eta_{\epsilon_j} \leq \tau^{-\delta} \eta_{\epsilon_j}.
\]

Similarly,
\[
| \mathbb{E} \{ \epsilon_j \mathbb{I}(|\epsilon_j| > \tau_j, |\epsilon_k| > \tau_k) \} | = | \mathbb{E} \{ \epsilon_j^{1+\delta/2} \epsilon_k^{1+\delta/2} \epsilon_j^{-\delta/2} \epsilon_k^{-\delta/2} | \mathbb{I}(|\epsilon_j| > \tau_j, |\epsilon_k| > \tau_k) \} | \\
\leq \mathbb{E}(|\epsilon_j|^{1+\delta/2} | \mathbb{E}(|\epsilon_j|^{2+\delta}) \}^{1/(2+\delta)} \{ \mathbb{E}(|\epsilon_k|^{2+\delta}) \}^{1/(2+\delta)} \\
\leq \tau_j^{-\delta/2} \tau_k^{-\delta/2} \mathbb{I}(|\epsilon_j| > \tau_j, |\epsilon_k| > \tau_k) \leq \tau^{-\delta} \eta_{\epsilon_j}. 
\]

and
\[
| \mathbb{E} \{ \mathbb{I}(|\epsilon_j| > \tau_j, |\epsilon_k| > \tau_k) \} | = | \mathbb{E} \{ \epsilon_j^{1+\delta} \epsilon_k^{-\delta} | \mathbb{I}(|\epsilon_j| > \tau_j, |\epsilon_k| > \tau_k) \} | \\
\leq \mathbb{E}(|\epsilon_j|^{1+\delta} | \mathbb{E}(|\epsilon_j|^{2+\delta}) \}^{1/(2+\delta)} \{ \mathbb{E}(|\epsilon_k|^{2+\delta}) \}^{1/(2+\delta)} \\
\leq \tau_k^{-1-\delta} \eta_{\epsilon_j} \leq \tau^{-1-\delta} \eta_{\epsilon_j}, 
\]

Therefore, we have
\[
| \mathbb{E} \{ \ell_{\tau_j}(\epsilon_j) \ell_{\tau_k}(\epsilon_k) \} - \text{cov}(\epsilon_j, \epsilon_k) | \lesssim \tau^{-\delta} \eta_{\epsilon_j}
\]
as \( \max_{1 \leq j \leq p} v_{j,\delta} \leq C \). Hence, it yields

\[
| \text{cov}\{\ell_{\tau_j}(\epsilon_j), \ell_{\tau_k}(\epsilon_k)\} - \text{cov}(\epsilon_j, \epsilon_k) | \leq | \mathbb{E}\{\ell_{\tau_j}(\epsilon_j)\ell_{\tau_k}(\epsilon_k)\} - \text{cov}(\epsilon_j, \epsilon_k) | + | \mathbb{E}\{\ell_{\tau_j}(\epsilon_j)\} \mathbb{E}\{\ell_{\tau_k}(\epsilon_k)\} | \\
\leq \tau^{-\delta} v_{jk} + \min\left\{ \frac{\sigma_{\epsilon,jj}}{\tau_j}, \frac{\mathbb{E}\{|\epsilon_j|^{2+\delta}\}}{\tau_j^{1+\delta}} \right\} \min\left\{ \frac{\sigma_{\epsilon,kk}}{\tau_k}, \frac{\mathbb{E}\{|\epsilon_k|^{2+\delta}\}}{\tau_k^{1+\delta}} \right\} \\
\leq \max(\tau^{-\delta} v_{jk}, \tau^{-2-2\delta} v_{jk}^2)
\]

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 \( \hat{\theta}_j \)'s with adaptive robustification parameter \( \tau_j \), and also gives a non-asymptotic Bahadur representation under the finite moment condition on the errors.

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

\[
P[\|\Sigma^{1/2}_{Z}(\hat{\theta}_j - \theta_j)\| \geq C_1 \tau_{0j} \{n^{-1}(d+t)^{1/2}\}] \leq 2e^{-t}
\]

and

\[
P[\|\Sigma^{1/2}_{Z}(\hat{\theta}_j - \theta_j) - \frac{1}{n} \sum_{i=1}^{n} \ell_{\tau}(\epsilon_i) \Sigma^{-1/2}_{Z}Z_i\| \geq C_2 \tau_{0j} n^{-1}(d+t)] \leq 3e^{-t}
\]

as long as \( n \geq C_3(d+t) \), 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 $T_j$, $S_j$, and $R_j$ in (A.2) and (A.3), respectively, it holds

$$\|T_j - \sigma_{\epsilon,jj}^{-1/2}(S_j + R_j)\| \leq C_2\tau_{0j} \frac{d + t}{(n\sigma_{\epsilon,jj})^{1/2}}$$

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

**Proof.** By Lemma B.1,

$$\|T_j - \sigma_{\epsilon,jj}^{-1/2}(S_j + R_j)\| \leq (n\sigma_{\epsilon,jj}^{-1})^{1/2} \left\| (C\Sigma_Z^{-1}C^T)^{-1/2}C\Sigma_Z^{-1/2}\left\{ \Sigma_Z^{1/2}(\hat{\theta}_j - \theta_j) - \frac{1}{n} \sum_{i=1}^n \ell'_r(\epsilon_{ij})\Sigma_Z^{-1/2}Z_i \right\} \right\|$$

(B.1)

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 $N(0, I)$. We decompose $|\mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2}S_j\| \geq x) - \mathbb{P}(\|G\| \geq x)|$ into two parts. Lemma B.2 quantifies the difference between the cumulative distribution functions of $\|\sigma_{\epsilon,jj}^{-1/2}s_jG\|$ and $\|G\|$, and Lemma B.3 characterizes the difference between the cumulative distribution functions of $\|\sigma_{\epsilon,jj}^{-1/2}s_jG\|$ and $\|\sigma_{\epsilon,jj}^{-1/2}S_j\|$.

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

$$\sup_{x \in \mathbb{R}^+} \left| \mathbb{P}(\|\sigma_{\epsilon,jj}^{-1/2}s_jG\| \geq x) - \mathbb{P}(\|G\| \geq x) \right| \leq q^{1/2} \frac{2v_{\delta,j\delta}}{\delta\sigma_{\epsilon,jj} \tau_{0j}} \left( \frac{d + t}{n} \right)^{\delta/(2 + \delta)}.$$
Proof. It holds
\[
\|\sigma^{-1}_{\epsilon,jj}s_j^2 I - I\| = |\sigma^{-1}_{\epsilon,jj}s_j^2 - 1| \leq \frac{2}{\delta \sigma_{\epsilon,jj}} \frac{v_j^{2+\delta}}{\tau_j^{\delta}}.
\] (B.2)

\[
\text{tr}\{(\sigma^{-1}_{\epsilon,jj}s_j I - I)^2\} \leq q \left( \frac{2}{\delta \sigma_{\epsilon,jj}} \frac{v_j^{2+\delta}}{\tau_j^{\delta}} \right)^2.
\]

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 \(n \geq C_3(d + t)\). 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 \(G \sim N(0, I) \in \mathbb{R}^q\).

\[
\sup_{x \in \mathbb{R}^q} \left| \mathbb{P}(\|\epsilon^{-1/2} S_j \| \geq x) - \mathbb{P}(\|\epsilon^{-1/2} s_j G \| \geq x) \right| \leq n^{-1/2} q^{7/4}.
\] (B.3)

Proof. Denote \(\mathcal{C}\) the class of convex subsets of \(\mathbb{R}^q\). Recall that \(\text{cov}(S_j) = \sigma^{-1}_{\epsilon,jj}s_j^2 I\) for \(S_j\) in (A.3). By Theorem 1.1 in Bentkus (2005),

\[
\sup_{A \in \mathcal{C}} \left| \mathbb{P}(\|\epsilon^{-1/2} S_j \| \in A) - \mathbb{P}(\|\epsilon^{-1/2} s_j G \| \in A) \right| \leq q^{1/4} \sum_{i=1}^n \mathbb{E} \|n^{-1/2} \tau_j(\epsilon_{ij}) A_j^{-1/2} \Sigma_{ij}^{-1/2} Z_i \|^3
\]

\[
= q^{1/4} \frac{\mathbb{E} \{|\epsilon_{ij}|^3\}}{\sigma_{\epsilon,jj}} \mathbb{E} \|A_j^{-1/2} \Sigma_{ij}^{-1/2} Z_i \|^3
\]

\[
\leq \frac{q^{7/4}}{n^{1/2}}.
\]

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

The following coupling result compares \(\mathbb{P}(\|\epsilon^{-1/2} S_j \| > x, \|\epsilon^{-1/2} S_k \| > y)\) and its Gaussian counterpart for each \((j, k)\) pair.

Lemma B.4. Assume Condition 1 holds. Let \(G = (G_1, G_2) \in \mathbb{R}^{2q}\) be a Gaussian vector with \(\mathbb{E}(G) = 0\) and \(\text{cov}(G) = \text{cov}(S)\), which is given in Section S1. It satisfies
that

$$\sup_{x, y \in \mathbb{R}} \left| \mathbb{P}(\|s_j^{-1}S_j\| > x, \|s_k^{-1}S_k\| > y) - \mathbb{P}(\|G_1\| > x, \|G_2\| > y) \right| \lesssim n^{-1/2}q^{7/4}.$$ 

Proof. Notice that

$$\mathbb{P}(\|s_j^{-1}S_j\| > x, \|s_k^{-1}S_k\| > y) = 1 - \mathbb{P}(\|s_j^{-1}S_j\| \leq x) - \mathbb{P}(\|s_k^{-1}S_k\| \leq y)$$

$$+ \mathbb{P}(\|s_j^{-1}S_j\| \leq x, \|s_k^{-1}S_k\| \leq y)$$

and

$$\mathbb{P}(\|G_1\| > x, \|G_2\| > y) = 1 - \mathbb{P}(\|G_1\| \leq x) - \mathbb{P}(\|G_2\| \leq y)$$

$$+ \mathbb{P}(\|G_1\| \leq x, \|G_2\| \leq y).$$

Take $A(x, y) = \{v = (v_1, v_2)^T \in \mathbb{R}^{2q}, v_1, v_2 \in \mathbb{R}^q : \|v_1\| \leq x$ and $\|v_2\| \leq y$ and $x, y \in \mathbb{R}^+ \cup \{\infty\} \}$ in Theorem 1.1 in Bentkus (2005), we have

$$\sup_{x, y} \left| \mathbb{P}(\|s_j^{-1}S_j\| > x, \|s_k^{-1}S_k\| > y) - \mathbb{P}(\|G_1\| > x, \|G_2\| > y) \right| \lesssim \frac{q^{7/4}}{n^{1/2}},$$

which is the desired result. \qed

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

Lemma B.5. Let $G = (G_1, G_2) \in \mathbb{R}^{2q}$ be a Gaussian vector with $\mathbb{E}(G) = 0$, $\text{cov}(G_i) = I$ for $i = 1, 2$, and $\text{corr}(G_1, G_2) = r I$ where $|r| \leq k_0 < 1$. Let $Z_1, Z_2 \sim N(0, I)$ be independent and identically distributed $q$-dimensional standard normal vectors. Then

$$|\mathbb{P}(\|G_1\| > z, \|G_2\| > z) - \mathbb{P}(\|Z_1\| > z)^2| \leq C_q|r|$$

for some constant $C_q > 0$ only depending on $q$. 
Proof. For $q = 1$, it holds
\[
\mathbb{P}(|G_1| > x, |G_2| > x) = \mathbb{P}(|Z_1| > x) \mathbb{P}(|Z_2| > x)
\]
\[
= \{\mathbb{P}(G_1 < -x, G_2 < -x) - \mathbb{P}(Z_1 < -x) \mathbb{P}(Z_2 < -x)\}
- \{\mathbb{P}(G_1 < -x, G_2 < x) - \mathbb{P}(Z_1 < -x) \mathbb{P}(Z_2 < x)\}
- \{\mathbb{P}(G_1 < x, G_2 < -x) - \mathbb{P}(Z_1 < x) \mathbb{P}(Z_2 < -x)\}
+ \{\mathbb{P}(G_1 < x, G_2 < x) - \mathbb{P}(Z_1 < x) \mathbb{P}(Z_2 < x)\},
\]
so that $|\mathbb{P}(|G_1| > x, |G_2| > x) - \mathbb{P}(|Z_1| > x) \mathbb{P}(|Z_2| > x)| \leq |r|$ follows Corollary 2.1 from Li and Shao (2002). Set $C_1 = 1$.

For $q \geq 2$, let $\Sigma_1 = I$ and $\Sigma_2$ be the covariance matrices of $(Z_1, Z_2)$ and $(G_1, G_2)$, respectively. By the definition of the total variation distance and Theorem 1.1 in Devroye, Mehrabian, and Reddad (2018), it follows that
\[
\mathbb{P}(\|G_1\| > z, \|G_2\| > z) - \mathbb{P}(\|Z_1\| > z)^2
\]
\[
\leq \sup_{A: \text{measurable sets}} |\mathbb{P}\{(G_1, G_2) \in A\} - \mathbb{P}\{(Z_1, Z_2) \in A\}|
\leq \frac{3}{2} \|\Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2} - I\|_F
\leq \frac{3}{2} \left\|\begin{bmatrix} 0 & rI \\ rI & 0 \end{bmatrix}\right\|_F = \frac{3}{2} (2q)^{1/2} |r| = \mathbb{P}(|G_1| > x, |G_2| > x) \leq \mathbb{P}(|Z_1| > x) \mathbb{P}(|Z_2| > x)
\]
\[
\leq \frac{3}{2} (2q)^{1/2} |r| = \mathbb{P}(|G_1| > x, |G_2| > x) \leq \mathbb{P}(|Z_1| > x) \mathbb{P}(|Z_2| > x)
\]
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 $X \sim N_k(0, \Sigma)$. Let $\lambda_1, \ldots, \lambda_k$ be the eigenvalues of $\Sigma$. Then $\|X\|^2$ 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 $A$ such that $A \Sigma A^T = \text{diag}(\lambda_1, \ldots, \lambda_k)$. Then $AX$ is $N_k(0, \text{diag}(\lambda_1, \ldots, \lambda_k))$. Hence, $\|AX\|^2 = \|X\|^2$ implies that $\|X\|^2$ has the same distribution as $\sum_{i=1}^k \lambda_i Z_i^2$. \qed

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 $A$ and $B = A + \Delta A$ such that $\rho(A^{-1} \Delta A) < 1$ where $\rho(\cdot)$ is the spectral radius, it follows that

$$\|A^{-1} - B^{-1}\| \leq \frac{\|A^{-1} \Delta A\|}{1 - \|A^{-1} \Delta A\|} \|A^{-1}\| \leq \frac{\|A^{-1}\|^2}{1 - \|A^{-1} \Delta A\|} \|\Delta A\|. \quad (B.4)$$

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

$$\|A^{-1} - B^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \|\Delta A\|} \|A^{-1}\| \|\Delta A\|. \quad (B.5)$$

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

$$\|A^{-1} - B^{-1}\| = \|A^{-1} \Delta A B^{-1}\| \leq \|A^{-1} \Delta A\| \|B^{-1}\|.$$
Since \( \|B^{-1}\| \leq \|A^{-1}\| + \|A^{-1}\Delta A\|\|B^{-1}\| \), it holds
\[
\|B^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\Delta A\|}.
\]
Hence, we prove the first part of (B.4). The second inequality of (B.4) is straightforward from the first part since \( \|A^{-1}\Delta A\| \leq \|A^{-1}\|\|\Delta A\| \). Inequality (B.5) is a direct result from (B.4) and \( \|A^{-1}\Delta A\| \leq \|A^{-1}\|\|\Delta A\| \).

**Lemma B.8.** Under the conditions in Theorem 1,
\[
\max_{1 \leq j \leq p} \|(C\Sigma_j C^T)^{-1} - (C\Sigma_j C^T)^{-1}\| \approx \max_{1 \leq j \leq p} \|\Sigma_j - \tilde{\Sigma}_j\|.
\]

**Proof.** Without loss of generality, as \( \text{rank}(C) = q \), we assume \( CC^T = I_{q \times q} \) by considering \( (CC^T)^{-1/2}C \) in (2.2). It follows that
\[
\|(C\Sigma C^T)^{-1}\|(C\Sigma_j C^T - C\Sigma_j C^T) \| \leq \|(C\Sigma_j C^T)^{-1}\|\|\Sigma_j - \tilde{\Sigma}_j\|\|CC^T\|
\leq \frac{\|\Sigma_j - \tilde{\Sigma}_j\|}{\lambda_{\min}(C\Sigma_j C^T)} \cdot 1
\leq \frac{\|\Sigma_j - \tilde{\Sigma}_j\|}{\lambda_{\min}(\Sigma_j)}
\leq \frac{\|\Sigma_j - \tilde{\Sigma}_j\|}{\sigma_{\epsilon, jj}} \|\Sigma_j\|
\leq \frac{\|\Sigma_j - \tilde{\Sigma}_j\|}{\epsilon_c^2} \|\Sigma_j\|_F
\leq \|\tilde{\Sigma}_j - \Sigma_j\|d^{1/2},
\]
where \( \| \cdot \|_F \) is the Frobenius norm. The last line in the above inequality is less than 1 as long as \( \|\tilde{\Sigma}_j - \Sigma_j\| \leq d^{-1/2} \), 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,

\[ \| C \tilde{\Sigma}_j C^T - C \Sigma_j C^T \| \leq \| \tilde{\Sigma}_j - \Sigma_j \|. \]  

(B.6)

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

\[ \| (C \tilde{\Sigma}_j C^T)^{-1} - (C \Sigma_j C^T)^{-1} \| \leq \frac{\| (C \Sigma_j C^T)^{-1} \|^2 \| C \tilde{\Sigma}_j C^T - C \Sigma_j C^T \|}{1 - \| (C \Sigma_j C^T)^{-1} \|^2 \| (C \tilde{\Sigma}_j C^T - C \Sigma_j C^T) \|} \]

\[ \leq \frac{\| (C \Sigma_j C^T)^{-1} \|^2 \| C \tilde{\Sigma}_j C^T - C \Sigma_j C^T \|}{1 - \| (C \Sigma_j C^T)^{-1} \|^2 \| \tilde{\Sigma}_j - \Sigma_j \| \| \tilde{\Sigma}_j - \Sigma_j \|} \]

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 \( \tau_j = \tau_0 \left[ n/ \{ \log(n p) + d \} \right]^{1/(2+\delta)} \) and \( \tau_0 \geq v_{j,\delta} \) with \( \delta \in (0,2] \). Under Conditions 2,

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

with probability at least \( 1 - \frac{4}{n} \), where \( C_4 \) is a constant depending on \( v_{j,\delta} \) and \( A_0 \) only.
Proof. For each \(i\) and \(j\), it follows that

\[
\max_{1 \leq i \leq n, 1 \leq j \leq p} |e_{ij} - \epsilon_{ij}| = \max_{1 \leq i \leq n, 1 \leq j \leq p} |Z^T_i (\hat{\theta}_j - \theta_j)| \\
\leq \max_{1 \leq i \leq n} \|\hat{Z}_i\| \left\{ \max_{1 \leq j \leq p} \|\Sigma_z^{1/2} (\hat{\theta}_j - \theta_j)\| \right\} \\
\leq C_4(d^{1/2} + t^{1/2}) \left( \frac{d + s}{n} \right)^{1/2}
\]

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

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

with probability at least \(1 - 4n^{-1}\). \(\Box\)

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

**Lemma B.10.** Let \(\tau_j = \tau_0 \left\lfloor n \{ \log(np) + d \}^{-1} \right\rfloor^{1/(2+\delta)}\), where \(\tau_0 \geq v_{j,\delta}\) for \(\delta \in (0, 2]\). On the event \(A_\Delta\),

\[
\max_{1 \leq j \leq p} \left\{ \frac{1}{n} \sum_{i=1}^n \{ \ell'_{\tau_j}(e_{ij}) \}^2 - \sigma_{\epsilon_{ij}} \right\} \leq C_5 \left\{ \log (np) + d \right\}^{\delta/(2+\delta)}
\]

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,

\[
\left| \frac{1}{n} \sum_{i=1}^{n} \{ \ell'_{\tau_j}(\epsilon_{ij}) \}^2 - \sigma_{\epsilon,jj} \right| \leq \left| \frac{1}{n} \sum_{i=1}^{n} \{ \ell'_{\tau_j}(\epsilon_{ij}) \}^2 - \{ \ell'_{\tau_j}(\epsilon_{ij}) \}^2 \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \{ \ell'_{\tau_j}(\epsilon_{ij}) \}^2 - s_j^2 \right| + |s_j^2 - \sigma_{\epsilon,jj}|.
\]

(B.8)

The last term on the right hand side of (B.8) is bounded using Corollary B.1. That is, 

\[ \sigma_{\epsilon,jj} - s_j^2 \leq 2\delta^{-1} \tau_j^{-\delta} \epsilon_{ij}^{2+\delta}. \]

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

\[
\frac{1}{n} \sum_{i=1}^{n} \{ \ell'_{\tau_j}(\epsilon_{ij}) \}^2 = \frac{1}{n} \sum_{i=1}^{n} \{ \ell'_{\tau_j}(\epsilon_{ij}) \}^2 + \frac{2}{n} \sum_{i=1}^{n} \{ \epsilon_{ij} \mathbb{I}(|\epsilon_{ij}| \leq \tau_j) \} (\epsilon_{ij} - \epsilon_{ij})
\]

\[ + \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(|\epsilon_{ij}| \leq \tau_j) (\epsilon_{ij} - \epsilon_{ij})^2 + \frac{1}{n} \sum_{i=1}^{n} R_{2j}(\epsilon_{ij}) \]

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

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

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

\[
\frac{1}{n} \sum_{i=1}^{n} \{ \ell'_{\tau_j}(\epsilon_{ij}) \}^2 - \frac{1}{n} \sum_{i=1}^{n} \{ \ell'_{\tau_j}(\epsilon_{ij}) \}^2
\]

\[ \leq \frac{2}{n} \sum_{i=1}^{n} \{ \epsilon_{ij} \mathbb{I}(|\epsilon_{ij}| \leq \tau_j) \} (\epsilon_{ij} - \epsilon_{ij}) + \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(|\epsilon_{ij}| \leq \tau_j) (\epsilon_{ij} - \epsilon_{ij})^2 + \frac{1}{n} \sum_{i=1}^{n} R_{2j}(\epsilon_{ij})
\]

\[ \leq 2\Delta \left[ \frac{1}{n} \sum_{i=1}^{n} \{ \epsilon_{ij} \mathbb{I}(|\epsilon_{ij}| \leq \tau_j) \} \right] + \Delta^2 \left[ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(|\epsilon_{ij}| \leq \tau_j) \right] + \frac{1}{n} \sum_{i=1}^{n} R_{2j}(\epsilon_{ij})
\]

\[ \leq 2\Delta \left[ \mathbb{E} \{ \epsilon_{ij} \mathbb{I}(|\epsilon_{ij}| \leq \tau_j) \} \right] + \tau_j \left\{ \log(np) / 2n \right\}^{1/2} \quad \text{(B.9)}
\]
+ \Delta^2 \left[ \mathbb{P}(|\epsilon_{ij}| \leq \tau_j) + \left\{ \frac{\log(np)}{2n} \right\}^{1/2} \right] + \left| \frac{1}{n} \sum_{i=1}^{n} R_{2j}(\epsilon_{ij}) \right|

\leq 2\Delta \left[ \mathbb{E}\{\ell_{\tau_j}(\epsilon_{ij})\} + \tau_j \left\{ \frac{\log(np)}{2n} \right\}^{1/2} \right] + \Delta^2 \left[ \mathbb{P}(|\epsilon_{ij}| \leq \tau_j) + \left\{ \frac{\log(np)}{2n} \right\}^{1/2} \right] + \left| \frac{1}{n} \sum_{i=1}^{n} R_{2j}(\epsilon_{ij}) \right|

\leq C\left\{ d^{1/2} + (2 \log n)^{1/2} \right\} \left\{ \frac{\log(np) + d}{n} \right\}^{(1+\delta)/(2+\delta)}

(B.10)

holds with probability at least 1 - 4n^{-1} as long as n \geq 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} \leq \frac{v_{ij}^4}{s_j^4}

and

\mathbb{E}(Q_{ij}^{2k}) \leq \frac{v_{ij}^4}{s_j^4} \left( \frac{\tau_j^2}{s_j^2} \right)^{k-2}

for all k \geq 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| \leq \left( \frac{v_{ij}^4}{s_j^4} \right)^{1/2} \left( \frac{2t}{n} \right)^{1/2} + \frac{\tau_j^2 t}{s_j^2 n}

(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

\[
\max_{1 \leq j \leq p} \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \ell'_{\tau_j}(e_{ij}) \right\}^2 - \sigma_{e,jj} \right| 
\leq C \left\{ d^{1/2} + (2 \log n)^{1/2} \right\} \left\{ \frac{\log(np) + d}{n} \right\}^{(1+\delta)/(2+\delta)} 
+ v_j^2 \left\{ \frac{2 \log(np)}{n} \right\}^{1/2} + \tau_{0j}^2 \frac{t}{n} \left\{ \frac{n}{\log(np) + d} \right\}^{2/(2+\delta)} + \frac{2v_j^2}{\delta} \left\{ \frac{\log(np) + d}{n} \right\}^{\delta/(2+\delta)}
\leq C_5 \left\{ \frac{\log(np) + d}{n} \right\}^{\delta/(2+\delta)}
\]

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

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

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

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

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 $A_{\Delta}$, $|\mathbb{I}_{\tau_j}^* (e_{ij}) - \mathbb{I}_{\tau_j}^* (e_{ij})| \leq \Delta h_n^{-1}$ due to the Lipschitz continuity of $\mathbb{I}_{\tau_j}^* (x)$. It follows that

\[
\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\tau_j}^* (e_{ij}) - 1 \right| \leq \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\tau_j}^* (e_{ij}) - 1 \right| + \frac{\Delta}{\tilde{h}_n}
\]

For the first term on the right hand side of (B.12), it follow Hoeffding’s inequality
and Markov’s inequality that
\[ \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\tau_j}^* (\epsilon_{ij}) - 1 \right| \leq \mathbb{E} \{ 1 - \mathbb{I}_{\tau_j}^* (\epsilon_{ij}) \} + \left( \frac{t}{2n} \right)^{1/2} \]
\[ \leq \mathbb{P} \{ |\epsilon_{ij}| \geq \tau_j \} + \left( \frac{t}{2n} \right)^{1/2} \]
\[ \leq \frac{v_{j,\delta}^{2+\delta}}{\tau_j^{2+\delta}} + \left( \frac{t}{2n} \right)^{1/2} \]
with probability at least \( 1 - 2 \exp(-t) \). Therefore,
\[ \max_{1 \leq j \leq p} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\tau_j}^* (\epsilon_{ij}) - 1 \right| \leq \frac{v_{j,\delta}^{2+\delta} \log(np) + d}{\tau_{ij}} + \left\{ \frac{\log(np)}{2n} \right\}^{1/2} \] (B.13)
with probability at least \( 1 - 2n^{-1} \). The lemma is therefore proved.

**Lemma B.12.** Let \( \tau_j = \tau_{ij} [n \{ \log(np) + d \}]^{-1/2} \) where \( \tau_{ij} \geq v_{j,\delta} \) for \( \delta \in (0, 2] \). On the event \( A_{\Delta} \), we have
\[ \max_{1 \leq j \leq p} \left\| W_j - \frac{1}{n} \sum_{i=1}^{n} Z_i Z_i^T \right\| \leq C_7 \max \left\{ \frac{\log(np) + d}{n} \right\}^{1/2} \left( \frac{\Delta}{h_n} \right) \]
with probability at least \( 1 - 2n^{-1} \), where \( C_7 > 0 \) is a constant depending only on \( \lambda_{\max}(\Sigma_Z) \), \( A_0 \), and \( v_{j,\delta} \) as long as \( n \geq C_3 \{ \log(np) + d \} \).

**Proof.** For \( \tilde{Z}_i = \Sigma_Z^{-1/2} Z_i \), we have
\[ \left\| W_j - \frac{1}{n} \sum_{i=1}^{n} Z_i Z_i^T \right\| = \left\| \Sigma_Z^{1/2} \left( \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\tau_j}^* (\epsilon_{ij}) - 1 \right) \tilde{Z}_i \tilde{Z}_i^T \right\| \Sigma_Z^{1/2} \]
\[ \leq \left\| \Sigma_Z \right\| \left\| \left( \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\tau_j}^* (\epsilon_{ij}) - 1 \right) \tilde{Z}_i \tilde{Z}_i^T \right\| . \]
On the event \( A_{\Delta} \), for each unit vector \( u \in \mathbb{R}^{d+1} \),
\[ \left| u^T \left[ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\tau_j}^* (\epsilon_{ij}) - 1 \right] \tilde{Z}_i \tilde{Z}_i^T \right| u \leq \frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbb{I} \{|\epsilon_{ij}| \geq \tau_j \} + \frac{\Delta}{h_n} \right\} \left| u^T \tilde{Z}_i \right|^2 \]
\[ \leq \mathbb{E} \left( |u^T \tilde{Z}_i|^2 \right) \left[ \mathbb{E} \{ \mathbb{I} \{|\epsilon_{ij}| \geq \tau_j \} \} + \frac{\Delta}{h_n} + C \max(\rho, \rho^2) \right] \]
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), \(\mathbb{E}(|u^T \tilde{Z}_i|^k) \leq A_1^k (ek/2) \Gamma(k/2)\) for all \(k \geq 1\), where \(A_1 \geq e^{-1/2}\) is a constant depending only on \(A_0\). Thus, 
\[
\mathbb{E}(\langle u, \tilde{Z}_i \rangle^2) \leq A_1^2 e
\]
and
\[
\mathbb{E}\{I(|e_{ij}| \geq \tau_j)\} = \mathbb{P}(|e_{ij}| \geq \tau_j) \leq \frac{v_{j,\delta}^{2+\delta}}{\tau_{0j}^{2+\delta}} \frac{d + t}{n}.
\]

Take \(t = \log(np)\). Putting together the obtained bounds yields
\[
\left\| \frac{1}{n} \sum_{i=1}^{n} \left\{ I_{\tau_j}(e_{ij}) - 1 \right\} Z_i Z_i^T \right\| \leq C_7 \max \left[ \left\{ \frac{\log(np) + d}{n} \right\}^{1/2}, \frac{\Delta}{h_n} \right]
\]
with probability at least \(1 - 2n^{-1}\) as long as \(n \geq C_3 \{\log(np) + d\}\), where \(C_7\) is a constant depending on \(A_0\), \(\lambda_{\max}(\Sigma_Z)\), and \(v_{j,\delta}\). □

### S3 Testing hypotheses of the linear combinations of \(\theta_j\)'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 \(j = 1, \ldots, p\), \(c \in \mathbb{R}^{d+1}\), and \(c_{0j} \in \mathbb{R}\), the two-sided and one-sided hypotheses of interest
are
\[ H_{0j} : \mathbf{c}^T\mathbf{\theta}_j = c_{0j} \text{ versus } H_{1j} : \mathbf{c}^T\mathbf{\theta}_j \neq c_{0j}, \quad (C.1) \]
and
\[ H_{0j} : \mathbf{c}^T\mathbf{\theta}_j \leq (\geq) c_{0j} \text{ versus } H_{1j} : \mathbf{c}^T\mathbf{\theta}_j > (<) c_{0j}, \quad (C.2) \]
respectively. For each \( j \), define
\[
U_j = n^{1/2}(\mathbf{c}^T\hat{\Sigma}_j\mathbf{c})^{-1/2}(\mathbf{c}^T\hat{\theta}_j - c_{0j}),
\]
where \( \hat{\theta}_j \) and \( \hat{\Sigma}_j \) 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
\[
\hat{V}(z) = \begin{cases} 
2p\Phi(-z) & \text{(two-sided)}, \\
p\Phi(-z) & \text{(one-sided)}.
\end{cases}
\]
Let the number of discoveries by \( R(z) = \sum_{j=1}^{p} \mathbb{I}(U_j \geq z) \). Then, we compute
\[
\hat{z}_\alpha = \inf \{ z \geq 0 : \text{AFDP}(z) \leq \alpha \},
\]
where \( \text{AFDP}(z) = \hat{V}(z)/R(z) \). For \( j = 1, \ldots, p \), \( H_{0j} \) in (C.1) or (C.2) is rejected whenever \( U_j \geq \hat{z}_\alpha \).

**S3.2 Theoretical guarantees**

The following result for testing (C.1) is a straightforward corollary of Proposition 1.

Denote \( U_j^\circ = n^{1/2}(\mathbf{c}^T\Sigma_j\mathbf{c})^{-1/2}(\mathbf{c}^T\hat{\theta}_j - c_{0j}) \) with known covariance \( \Sigma_j \). For \( \mathcal{H}_0 = \{ j : \)}
1 \leq j \leq p, \text{ } H_{0j} \text{ is true}}, \text{ let } V^0(z) = \sum_{j \in \mathcal{J}_0} \mathbb{I}(U^0_j \geq z) \text{ and } R^0(z) = \sum_{j=1}^p \mathbb{I}(U^0_j \geq z).

Define \( \text{AFDP}^0_{c1}(z) = 2p_0 \Phi(-z)/R^0(z) \) to be the counterpart of (3.1).

**Theorem C.1.** Consider testing (C.1). Assume Conditions 1 and 2 hold, and \( p_0 \geq p \) for some \( a \in (0, 1) \). Let \( \tau_j = \tau_{0j} n^{1/(2+\delta)} \{ \log(np) + d \}^{-1/(2+\delta)} \) with \( \tau_{0j} \geq \nu_{j,\delta} \) and \( \delta \in (0, 2] \).

Then, for any \( z \geq 0 \), \(| \text{FDP}^0(z) - \text{AFDP}^0_{c1}(z) | = o_p(1) \) as \( n, p \to \infty \).

Next, we provide the corresponding result for testing (C.2). Similarly, let \( \text{AFDP}^0_{c2}(z) = p_0 \Phi(-z)/R^0(z) \).

**Theorem C.2.** Consider testing (C.2). Assume Conditions 1 and 2 hold, and \( p_0 \geq p \) for some \( a \in (0, 1) \). Let \( \tau_j = \tau_{0j} n^{1/(2+\delta)} \{ \log(np) + d \}^{-1/(2+\delta)} \) with \( \tau_{0j} \geq \nu_{j,\delta} \) and \( \delta \in (0, 2] \).

Then, for any \( z \geq 0 \), \(| \text{FDP}^0(z) - \text{AFDP}^0_{c2}(z) | = o_p(1) \) as \( n, p \to \infty \).

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

\[
p_0^{-1} V^0(z) = \Phi(-z) + O_p(p^{-\kappa_1} + n^{-1/2} + [n^{-1} \{ \log(np) + d \}]^{\delta/(2+\delta)}) \tag{C.3}
\]

which leads to the conclusion immediately.

Let \( \sigma_j^2 = c^T \Sigma_j c = \sigma_{c,jj}(c^T \Sigma_Z^{-1} c) \in \mathbb{R} \), \( U^0_j = n^{1/2} \sigma_j^{-1} (c^T \tilde{\theta}_j - c^T \theta_j) \), and

\[
S_j = n^{-1/2} || c^T \Sigma_Z^{-1/2} c ||^{-1} c^T \Sigma_Z^{-1} \sum_{i=1}^n \{ \ell'_r(\epsilon_{ij}) Z_{ii} - \mathbb{E}\{ \ell'_r(\epsilon_{ij}) Z_{ii} \} \},
\]

\[
R_j = n^{-1/2} || c^T \Sigma_Z^{-1/2} c ||^{-1} c^T \Sigma_Z^{-1} \sum_{i=1}^n \mathbb{E}\{ \ell'_r(\epsilon_{ij}) Z_{ii} \}.
\]
For every $j \in H_0j$ and $t \geq 1$, it follows from Lemma B.1 that
\[
|U_j^o - \sigma_{\epsilon,jj}^{-1/2} (S_j + R_j)| = \left| n^{1/2} \sigma_0^{-1} (c^T \hat\theta_j - c^T \theta_j) - n^{-1/2} \sigma_0^{-1} c^T \Sigma^{-1} \sum_{i=1}^n \ell'_i (\epsilon_{ij}) Z_i \right|
\leq n^{1/2} \sigma_{\epsilon,jj}^{-1/2} \left| \Sigma^{-1/2} (\hat\theta_j - \theta_j) - \frac{1}{n} \sum_{i=1}^n \ell'_i (\epsilon_{ij}) \Sigma^{-1/2} Z_i \right|,
\]
with probability greater than $1 - 3 \exp(-t)$ as long as $n \geq C_3(d + t)$ with $\tau_j = \tau_0 \{ n(d + t)^{-1} \}^{1/2+\delta}$.

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 $E_1$,
\[
\sum_{j \in H_{0j}} \mathbb{I} \left\{ \sigma_{\epsilon,jj}^{-1/2} S_j \geq z + C_2 \frac{\tau_0 (d + t)}{(n \sigma_{\epsilon,jj})^{1/2}} \right\} \leq V(z) \leq \sum_{j \in H_{0j}} \mathbb{I} \left\{ \sigma_{\epsilon,jj}^{-1/2} S_j \geq z - C_2 \frac{\tau_0 (d + t)}{(n \sigma_{\epsilon,jj})^{1/2}} \right\}
\]
with probability $1 - 3pe^{-t}$. For $x \in \mathbb{R}$, define
\[
V^+(x) = \sum_{j \in H_{0j}} \mathbb{I} (\sigma_{\epsilon,jj}^{-1/2} S_j \geq x).
\]
Hence, (C.5) can be written as
\[
p_0^{-1} V^+ \left\{ z + C_2 \frac{\tau_0 (d + t)}{(n \sigma_{\epsilon,jj})^{1/2}} \right\} \leq p_0^{-1} V^0(z) \leq p_0^{-1} V^+ \left\{ z - C_2 \frac{\tau_0 (d + t)}{(n \sigma_{\epsilon,jj})^{1/2}} \right\}.
\]
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
\[
\mathbb{E}[\{p_0^{-1} V^+(z) - \Phi(-z)\}^2] \lesssim p^{-\kappa_1} + n^{-1/2} + \left( \frac{d + t}{n} \right)^{\delta/(2+\delta)}.
\]
Recall that $P(A_1) \leq 1 - 3pe^{-t}$ whenever $n \geq d + t$. Taking $t = \log(np)$ in (C.7) and (C.8) proves (C.3).

\[\square\]

**Remark C.1.** For testing $H_{0j} : c^T \theta_j \geq c_0$ versus $H_{1j} : c^T \theta_j < c_0$, we can use the same argument with (C.5) and (C.6) replaced by

\[
\sum_{j \in \mathcal{I}_{o,j}} \mathbb{I}\left\{ \sigma_{\epsilon, j}^{-1/2} S_j \leq -z - C_2 \frac{\tau_0 (d + t)}{n \sigma_{\epsilon, j}^2} \right\} \leq V^0(z) \leq \sum_{j \in \mathcal{I}_{o,j}} \mathbb{I}\left\{ \sigma_{\epsilon, j}^{-1/2} S_j \leq -z + C_2 \frac{\tau_0 (d + t)}{n \sigma_{\epsilon, j}^2} \right\}
\]

and $V^-(x) = \sum_{j \in \mathcal{I}_{o,j}} \mathbb{I}(\sigma_{\epsilon, j}^{-1/2} S_j \leq -x)$, 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 $z_i^T = \begin{bmatrix} 1 & x_i^T \end{bmatrix}$ the $i$th row of design matrix $Z$. We first impose the following regularity condition, which is similar to that in Sun et al. (2020).

**Condition D.1.** The Gram matrix $S_n = n^{-1} \sum_{i=1}^{n} z_i z_i^T$ is positive definite and there exist constants $c_l$ and $c_u$ such that $c_l \leq \lambda_{\min}(S_n) \leq \lambda_{\max}(S_n) \leq c_u$. As $n \to \infty$, $S_n \to \Sigma_Z$ 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 $\tilde{z}_i = S_n^{-1/2} z_i$,

\[
\sup_{u \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (u^T \tilde{z}_i)^4 \exp(\kappa |u^T \tilde{z}_i|^2) \leq M.
\]
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 \( p_0 \geq ap \) for some \( a \in (0, 1) \). Let \( \tau_j = \tau_{0j}\{n/\log(np)\}^{1/(2+\delta)} \) where \( \tau_{0j} \geq v_{j,\delta} \) for some \( \delta \in (0, 2] \). Then, for any \( z \geq 0 \), \(|\FDP^\circ(z) - \AFDP^\circ(z)| = o_\P(1)\) 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

\[
\mathbf{s}_j = n^{-1/2}(\mathbf{CS}_n^{-1}\mathbf{C}^T)^{-1/2}\mathbf{CS}_n^{-1/2}\sum_{i=1}^{n} \mathbf{S}_n^{-1/2} \mathbf{Z}_i[\ell'_{\tau}(\epsilon_{ij}) - \E\{\ell'_{\tau}(\epsilon_{ij})\}],
\]

\[
\mathbf{r}_j = n^{-1/2}(\mathbf{CS}_n^{-1}\mathbf{bC}^T)^{-1/2}\mathbf{CS}_n^{-1/2}\sum_{i=1}^{n} \mathbf{S}_n^{-1/2} \mathbf{Z}_i \E\{\ell'_{\tau}(\epsilon_{ij})\}.
\]

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

\[
\|\mathbf{T}_j^\circ - \sigma^{-1/2}_{\epsilon,ij}(\mathbf{s}_j + \mathbf{r}_j)\| \leq A\tau_{0j}(d + t)^{1/2}\left(\frac{dt}{\sigma_{\epsilon,ij}n}\right)^{1/2}
\]

with probability \( 1 - 2d\exp(-t) \) as long as \( n \geq \max\{32L^4d^2t, 2\kappa^{-2}(2d + t)\} \), where \( \Delta_{n,\delta} = n^{-1}\sum_{i=1}^{n} v_{\delta}\tilde{\mathbf{Z}}_i\tilde{\mathbf{Z}}_i^T \). 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
\[ P\{ E_1(t) \} = 1 - 2dp \exp(-t) \]. One can obtain the counterparts of (A.5) and (A.12),

\[
\sum_{j \in \mathcal{H}_0} \mathbb{I} \left[ \left\| \sigma_{\epsilon,j}^{-1/2} s_j \right\| \geq z + A\tau_0(d + t)^{1/2} \left( \frac{dt}{\sigma_{\epsilon,j}^2 n} \right)^{1/2} \right] \leq V^2(z^2)
\]

\[
\leq \sum_{j \in \mathcal{H}_0} \mathbb{I} \left[ \left\| \sigma_{\epsilon,j}^{-1/2} s_j \right\| \geq z - A\tau_0(d + t)^{1/2} \left( \frac{dt}{\sigma_{\epsilon,j}^2 n} \right)^{1/2} \right]
\]

and

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

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 \( \mathcal{L}_\tau(\theta) := n^{-1} \sum_{i=1}^n \ell_\tau(y_i - z_i^T \theta) \). Lemma D.1 provides the lower bound of
\( \lambda_{\min}\{S_n^{-1/2}\nabla^2 L_{\tau}(\tilde{\theta})S_n^{-1/2}\} \), 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 \( v_{\delta} := \{\mathbb{E}(|\varepsilon_i|^{2+\delta})\}^{1/(2+\delta)} < \infty \) for \( \delta \in (0.2] \). Then for any \( t, r > 0 \), the matrix \( S_n^{-1/2}\nabla^2 L_{\tau}(\tilde{\theta})S_n^{-1/2} \) with \( \tau > 2L_2r \) satisfies that

\[
\min_{\theta \in \mathbb{R}^{d+1}, ||S_n^{1/2}(\tilde{\theta} - \theta)|| \leq r} \lambda_{\min}\{S_n^{-1/2}\nabla^2 L_{\tau}(\tilde{\theta})S_n^{-1/2}\} \\
\geq 1 - (2L_2r/\tau)^2 - L_2^2\{(2v_\delta/\tau)^{2+\delta} + (2n)^{-1/2}t^{1/2}\},
\]

with probability at least \( 1 - \exp(-t) \) where \( L_2 = \max_{1 \leq i \leq n} \|\hat{Z}_i\| \).

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 \( \epsilon_i \) 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 \( \tau_0 \geq v_{\delta} \), the adaptive Huber regression estimator \( \tilde{\theta} = (\tilde{\mu}, \tilde{\beta}^T)^T \in \mathbb{R}^{d+1} \) in (2.3) with \( \tau = \tau_0(n/t)^{1/(2+\delta)} \) satisfies

\[
\|S_n^{1/2}(\tilde{\theta} - \theta)\| \leq C(L_{\infty}, \delta, v_{\delta})d^{1/2}\left(\frac{t}{n}\right)^{1/2}
\]

with probability at least \( 1 - (2d + 3)\exp(-t) \) as long as \( n \geq 32L_4^4d^2t \), where \( L_\infty = \max_{1 \leq i \leq n} \|\hat{Z}_i\|_\infty \) and \( C(L_{\infty}, \delta, v_{\delta}) \) is a constant only depending on \( L_{\infty}, \delta, \) and \( v_{\delta} \).
Proof. Recall that \( \tau = \tau_0(n/t)^{1/(2+\delta)} \). Let \( \hat{\theta}_\eta = \theta + \eta(\hat{\theta} - \theta) \) with \( \eta \in (0, 1] \) so that \( \|S_n^{1/2}(\hat{\theta}_\eta - \theta)\| \leq r \). Lemma 2 from Sun et al. (2020) gives

\[
\langle \nabla L_\tau(\hat{\theta}_\eta) - \nabla L_\tau(\theta), \hat{\theta}_\eta - \theta \rangle \leq \eta \langle \nabla L_\tau(\hat{\theta}) - \nabla L_\tau(\theta), \hat{\theta} - \theta \rangle,
\]

where \( \nabla L_\tau(\hat{\theta}) = 0 \) by the Karush-Kuhn-Tucker condition. By the mean value theorem for vector-valued functions, the equality

\[
\nabla L_\tau(\hat{\theta}_\eta) - \nabla L_\tau(\theta) = \left[ \int_0^1 \nabla^2 L_\tau\{ (1-t)\theta + t\hat{\theta}_\eta \} dt \right] (\hat{\theta}_\eta - \theta)
\]

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

\[
\min_{\tilde{a} \in \mathbb{R}^{d+1}, \|S_n^{1/2}(\tilde{a} - \theta)\| \leq r} \lambda_{\min}(S_n^{-1/2}\nabla^2 L_\tau(\tilde{a})S_n^{-1/2}) \geq a_0,
\]

(D.3)

then

\[
a_0\|S_n^{1/2}(\hat{\theta}_\eta - \theta)\|^2_2 \\
\leq \lambda_{\min}\left[ S_n^{-1/2}\nabla^2 L_\tau\{ (1-t)\theta + t\hat{\theta}_\eta \} S_n^{-1/2} \right] \|S_n^{1/2}(\hat{\theta}_\eta - \theta)\|^2_2 \\
= \frac{(\hat{\theta}_\eta - \theta)^T S_n^{1/2}}{\|S_n^{1/2}(\hat{\theta}_\eta - \theta)\|_2} \left[ S_n^{-1/2}\nabla^2 L_\tau\{ (1-t)\theta + t\hat{\theta}_\eta \} S_n^{-1/2} \right] \frac{S_n^{1/2}(\hat{\theta}_\eta - \theta)}{\|S_n^{1/2}(\hat{\theta}_\eta - \theta)\|_2} \|S_n^{1/2}(\hat{\theta}_\eta - \theta)\|^2_2 \\
= (\hat{\theta}_\eta - \theta)^T \left[ \nabla^2 L_\tau\{ (1-t)\theta + t\hat{\theta}_\eta \} \right] (\hat{\theta}_\eta - \theta)
\]

and

\[
a_0\|S_n^{1/2}(\hat{\theta}_\eta - \theta)\|^2_2 \leq \int_0^1 (\hat{\theta}_\eta - \theta)^T \left[ \nabla^2 L_\tau\{ (1-t)\theta + t\hat{\theta}_\eta \} \right] (\hat{\theta}_\eta - \theta) dt \\
= (\hat{\theta}_\eta - \theta)^T \left\{ \nabla L_\tau(\hat{\theta}_\eta) - \nabla L_\tau(\theta) \right\} \\
\leq \eta (\hat{\theta} - \theta)^T \left\{ -\nabla L_\tau(\theta) \right\} \\
\leq \eta \|S_n^{-1/2}\nabla L_\tau(\theta)\| \|S_n^{1/2}(\hat{\theta} - \theta)\|.
\]
by putting together all the results above. Setting $\eta = 1$ yields

$$a_0\|S_n^{1/2}(\widehat{\theta} - \theta)\|_2 \leqslant \|S_n^{-1/2}\nabla \mathcal{L}_\tau(\theta)\|_2.$$  \hspace{1cm} (D.4)

Denote the $k$th entry of $\xi = S_n^{-1/2}\nabla \mathcal{L}_\tau(\theta)$ 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 $n \geqslant (\sigma^2v_\delta^{2+\delta})^{(2+\delta)/\delta}t$, it follows that $|\mathbb{E}\ell'_\tau(\epsilon)| \leqslant \tau^{-(1+\delta)}v_\delta^{2+\delta}$. By the definition of $\ell'_\tau(\cdot)$,

$$|\mathbb{E}\{n^{-1}\ell'_\tau(\epsilon_i)\widetilde{z}_{ik}\}| \leqslant \frac{v_\delta^{2+\delta}}{n^{1+\delta}}L_\infty,$$

$$\left|\frac{1}{n}\ell'_\tau(\epsilon_i)\widetilde{z}_{ik} - \mathbb{E}\left\{\frac{1}{n}\ell'_\tau(\epsilon_i)\widetilde{z}_{ik}\right\}\right| \leqslant |\widetilde{z}_{ik}|\left(\frac{\tau}{n} + \frac{v_\delta^{2+\delta}}{n^{1+\delta}}\right) \leqslant L_\infty\left(\frac{\tau}{n} + \frac{v_\delta^{2+\delta}}{n^{1+\delta}}\right),$$

and

$$\mathbb{E}\left[\{n^{-1}\ell'_\tau(\epsilon_i)\widetilde{z}_{ik} - n^{-1}\mathbb{E}\{\ell'_\tau(\epsilon_i)\}\widetilde{z}_{ik}\}^2\right] \leqslant n^{-2}s^2\widetilde{z}_{ik}^2$$

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

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

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 $L_\infty$, $\delta$ and $v_\delta$, and

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

$$\leqslant \frac{L_\infty}{3} \frac{t}{n} \left(\tau + \frac{v_\delta^{2+\delta}}{\tau^{1+\delta}}\right) + \frac{L_\infty}{3} \left\{\frac{1}{9} \frac{t^2}{n^2} \left(\tau + \frac{v_\delta^{2+\delta}}{\tau^{1+\delta}}\right)^2 + 18s^2 \frac{t}{n}\right\}^{1/2}\frac{1}{n} \leqslant \left(\frac{t}{n}\right)^{1/2}. $$
Then, for any $t > 0$,
\[
\mathbb{P}(\|\xi\|_2 \geq C(d + 1)^{1/2} n^{-1/2} t^{1/2}) \leq \mathbb{P}(\|\xi\|_\infty \geq C n^{-1/2} t^{1/2}) \\
\leq \sum_{k=1}^{d+1} \mathbb{P}(|\xi_k| \geq C n^{-1/2} t^{1/2}) \quad \text{(D.5)} \\
\leq 2(d + 1) \exp(-t).
\]

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
\[
\min_{|S_n^{1/2}(\hat{\theta} - \theta)| \leq \tau/(4L_2)} \lambda_{\min}(S_n^{-1/2} \nabla^2 \mathcal{L}_n(\hat{\theta}) S_n^{-1/2}) \geq \frac{3}{4} - L_2^2 \left\{ \left( \frac{2v_\delta}{\tau_0} \right)^{2+\delta} \frac{t}{n} + \left( \frac{t}{2n} \right)^{1/2} \right\} \\
\geq \frac{1}{2},
\]
holds as long as $n \geq \max(32L_2^4, 2^{5+\delta} L_2^2) t = 32 \max(L_2^4, 2^\delta L_2^2) t$. By (D.4) and (D.5), we have
\[
\|S_n^{1/2}(\hat{\theta} - \theta)\| \leq 2Cd^{1/2} n^{-1/2} t^{1/2}
\]
with probability at least $1 - (2d + 3) \exp(-t)$. \hfill \Box

Lemma D.3 provides a nonasymptotic Bahadur representation under the fixed design, and it implies that $\sqrt{n}S_n^{1/2}(\hat{\theta} - \theta)$ 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 $\tau_0 \geq v_\delta$, the estimator $\hat{\theta}$ given in (2.3) with $\tau = \tau_0(n/t)^{1/(2+\delta)}$
satisfies that

\[
P\left\{ \left\| S_n^{1/2}(\hat{\theta} - \theta) - \frac{1}{n} \sum_{i=1}^{n} c_{\tau}(\epsilon_i) S_n^{-1/2}Z_i \right\| \geq A\tau_0(d + t)^{1/2} \frac{(dt)^{1/2}}{n} \right\} \leq 2(d + 2)e^{-t},
\]

whenever \( n \geq \max\{32L_\infty^4d^2t, 2\kappa^{-2}(2d + t)\} \), 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 \( \Delta_{n,\delta} = n^{-1} \sum_{i=1}^{n} v_\delta^2 z_i z_i^T \).

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 \( T_j^0 \) and its Bahadur representation in (D.1), it holds

\[
\|T_j^0 - \sigma_{\epsilon,ij}^{-1/2}(s_j + r_j)\| \leq A\tau_0(d + t)^{1/2} \frac{dt}{(n\sigma_{\epsilon,ij})^{1/2}}
\]

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 \( N(0, \sigma_{\epsilon,ij}^2I) \). We decompose \( \|P(\sigma_{\epsilon,ij}^{-1/2}s_j) \geq x\) - \( P(\|G\| \geq x)\) into two parts. Lemma D.4 quantifies the difference between the cumulative distribution functions of \( \|\sigma_{\epsilon,ij}^{-1/2}s_jG\| \) and \( \|G\| \), and Lemma D.5 quantifies the distinction between the cumulative distribution functions of \( \|\sigma_{\epsilon,ij}^{-1/2}s_jG\| \) and \( \|\sigma_{\epsilon,ij}^{-1/2}s_j\| \). Their proofs are similar to those of Lemmas B.2-B.3 and therefore are omitted.
Lemma D.4. Let $G \sim N(0, I) \in \mathbb{R}^q$. For $\tau_j = \tau_{0j} (n/t)^{1/(2+\delta)}$ for some $\delta \in (0, 2]$ where $\tau_{0j} \geq v_{j,\delta}$,

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

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

$$\sup_{x \in \mathbb{R}^q} \left| \mathbb{P}(\|\sigma_{\epsilon,j\epsilon}^{-1/2}s_j G\| \geq x) - \mathbb{P}(\|\sigma_{\epsilon,j\epsilon}^{-1/2}s_j G\| \geq x) \right| \leq n^{-1/2} q^{\gamma/4}.$$ 

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 $\Xi = 100 R_{\epsilon}^{1/2} \mathbf{1}$, where the correlation matrix $R_{\epsilon}$ has one of the following three structures:
• **Model 1**, $R_\epsilon$ is the identity matrix;

• **Model 2**, $R_\epsilon = (r_{\epsilon, jk})_{1 \leq j, k \leq p}$ 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**, $R_\epsilon = (r_{\epsilon, jk})_{1 \leq j, k \leq p}$ with $r_{\epsilon, jj} = 1$, $r_{\epsilon, j+1, j} = 0.3$, $r_{\epsilon, j+2, j} = 0.1$, and $r_{\epsilon, j+k} = r_{\epsilon, j+k,j} = 0$ for $k \geq 3$.

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

• **Hypothesis 1**, $H_{0j} : 1^T \beta_j = 0$ versus $H_{aj} : 1^T \beta_j \neq 0$, where $q = 1$, and

• **Hypothesis 2**, $H_{0j} : \beta_j = 0 \in \mathbb{R}^d$ versus $H_{aj} : \beta_j \neq 0$ ($j = 1, \ldots, p$), where $q = d$.

For **Hypothesis 1**, we let $\beta_{jk} \sim \text{Unif}(-150, 150)$ for $1 \leq j \leq p$ and $1 \leq k \leq d - 1$, $\beta_{jd} = -\sum_{k=1}^{d-1} \beta_{jk}$ for $1 \leq j \leq p - p_1$ so that $1^T \beta_j = 0$, and $\beta_{jd} = \delta d^{1/2} W_j - \sum_{k=1}^{d-1} \beta_{jk}$ for $p - p_1 + 1 \leq j \leq p$, where $W_j$ are Rademacher random variables. For **Hypothesis 2**, let $\beta_j = 0$ for $1 \leq j \leq p - p_1$, and $\beta_{jk} = (2d^{-1})^{1/2} \delta W_{jk}$ for $p - p_1 + 1 \leq j \leq p$ and $1 \leq k \leq d$, 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 $\tau_j$ 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 \texttt{edgeR} and \texttt{limma}. Similar observations are made when the dependence
is strong (\textit{Model 3}) and $d$ is large.

Similar to Figure 3 in the main paper, Figure S12 compares the powers of different
methods for testing \textit{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 $\tau_j$’s outperforms \texttt{limma}
and \texttt{edgeR} for all error settings.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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 $Y$. For the Box-Cox transformation, we select the transformation exponent $\lambda_j$ for the $j$th coordinate of response $Y$ 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 $j$th 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 $Y$ (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 $h_n \in \{0.25n^{-1/4}, 0.5n^{-1/4}, n^{-1/4}, 2n^{-1/4}, 4n^{-1/4}, \ldots\}$.
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 \( \tau_{0j} \) is determined as in \( \tau_j \) 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 matrix (with a subset of 25 words), which is relatively dense and will be used as $R_\epsilon$ 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.
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 \( \sigma_{\epsilon_{ij}} \) in Fan et al. (2019) fails to provide FDP control when \( d \geq 1. \) 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,

\[
\hat{\alpha}_j := \arg\min_{\alpha} \sum_{i=1}^{n} \ell_{\tau_{j1}} \{(Y_{ij} - X_i^T \hat{\Theta}_j) - \alpha\}
\]

\[
\hat{\gamma}_j := \arg\min_{\gamma \geq \hat{\alpha}_j^2} \sum_{i=1}^{n} \ell_{\tau_{j2}} \{(Y_{ij} - X_i^T \hat{\Theta}_j)^2 - \gamma\}
\]

\[
\tilde{\sigma}_{\epsilon,jj} := \hat{\gamma}_j - \hat{\alpha}_j^2,
\]

where \( \tau_{j1} \) and \( \tau_{j2} \) 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 \( \tilde{\Sigma}_j \) is estimated by \( \tilde{\sigma}_{\epsilon,jj} \hat{\Sigma}_Z^{-1} \) (FARM-1), which is an OLS-like covariance estimator with \( \tilde{\sigma}_j^2 \) replaced by \( \tilde{\sigma}_{\epsilon,jj} \) from (E.1), or \( \tilde{\sigma}_{\epsilon,jj} W^{-1} \hat{\Sigma}_Z W^{-1} \) (FARM-2), which is the covariance in (2.6) with \( \tilde{\sigma}_{\epsilon,jj} \) replaced by \( \tilde{\sigma}_{\epsilon,jj} \). 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 $\tilde{\sigma}_{r,jj} \hat{\Sigma}_Z^{-1}$ (FARM-1, ■), and our procedure with $\tilde{\sigma}_{r,jj} W^{-1} \hat{\Sigma}_Z W^{-1}$ (FARM-2, ♦), where $\tilde{\sigma}_{r,jj}$ 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
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
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


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

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.