Supplemental Material for "Distributed Algorithms for High-Dimensional Statistical Inference and Structure Learning with Heterogeneous Data"

Hongru Zhao and **Xiaotong Shen**
School of Statistics, University of Minnesota, Twin Cities

S1 Heterogeneous Linear Regression with Site-Specific Heteroskedastic Errors

We extend the heterogeneous linear model in Remark 1 to allow each site j its own error variance $N_{n_j}\big(0, \sigma_j^2 I_{n_j}\big)$.

The negative log-likelihood (up to constant) is $\frac{n_j}{2}\log\sigma_j^2 + \frac{1}{2\sigma_j^2}\big\|\mathbf{Y}_j - \mathbf{X}_j\boldsymbol{\beta}_0 - \mathbf{W}_j\boldsymbol{\beta}_j\big\|_2^2$. Profiling out σ_j^2 or equivalently reparameterizing via weighted least squares, one obtains the site-wise loss

$$L_j(\boldsymbol{\beta}_0, \boldsymbol{\beta}_j) = \tfrac{1}{2}\big\|\mathbf{Y}'_j - \mathbf{X}'_j\boldsymbol{\beta}_0 - \mathbf{W}'_j\boldsymbol{\beta}_j\big\|_2^2 \tag{S1.1}$$

with $\mathbf{Y}'_j = \frac{\mathbf{Y}_j}{\sigma_j}, \mathbf{X}'_j = \frac{\mathbf{X}_j}{\sigma_j}, \mathbf{W}'_j = \frac{\mathbf{W}_j}{\sigma_j}$. This weighted least squares loss simply replaces the loss L_j in (5.25). To compute the CMLE for the parameters $\{\boldsymbol{\beta}_j\}_{j=0}^K$ and variances $\{\sigma_j^2\}_{j=1}^K$, we use a block coordinate descent (BCD) algorithm, see Algorithm S1.

By treating the number of sites K as a fixed parameter, the constrained likelihood–ratio testing procedure and Theorem 2 remain valid without modification.

S2 Proof of Theorem 1

Proof. We will show that if $\kappa \geq |A^0_{H_0}|$, then $\{i \in [p]\backslash B : \beta_i^0 \neq 0\} \subset \{i \in [p]\backslash B : \widehat{\Gamma}_i \neq 0\}$ almost surely, where the estimators $\widehat{\boldsymbol{\Gamma}}_{H_0}$ and $\widehat{\boldsymbol{\Gamma}}_{H_1}$ are obtained from Algorithm 3. For H_0, let $A^0 = \{i \in [p]\backslash B : \beta_i^0 \neq 0\}$ and $A^{[t]} = \{i \in [p]\backslash B : |\widetilde{\Gamma}_i^{[t]}| \geq \tau\}$. For H_1, let $A^0 = \{i \in [p]\backslash B : \beta_i^0 \neq 0\} \cup B$ and $A^{[t]} = \{i \in [p]\backslash B : |\widetilde{\Gamma}_i^{[t]}| \geq \tau\} \cup B$. Set $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{ol}$, where $\widehat{\boldsymbol{\beta}}^{ol}$ is the

Algorithm S1 BCD for CMLE in Heterogeneous Linear Regression with Heteroskedastic Errors

Initialize: Set $t = 0$ and $\sigma_j^{2(0)} = 1$ for $j = 1, \dots, K$.
while not converged **do**
Parameter update: Treat (S1.1) (with the current $\{\widehat{\sigma}_j^{2\,(t)}\}$) as the site-wise loss and run Algorithm 2 to obtain $\{\widehat{\boldsymbol{\beta}}_j^{(t+1)}\}_{j=0}^{K}$.
Variance update: For each $j = 1, \dots, K$,

$$\widehat{\sigma}_j^{2\,(t+1)} = \frac{1}{n_j} \big\| \mathbf{Y}_j - \mathbf{X}_j\, \widehat{\boldsymbol{\beta}}_0^{(t+1)} - \mathbf{W}_j\, \widehat{\boldsymbol{\beta}}_j^{(t+1)} \big\|_2^2.$$

$t \leftarrow t + 1$
end while
Output: $\{\widehat{\boldsymbol{\beta}}_j\}_{j=0}^{K}$ and $\{\widehat{\sigma}_j^2\}_{j=1}^{K}$.

oracle estimate that minimizes $\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ under the constraint $\boldsymbol{\beta}_{(A^0)^c} = 0$. Set $E = \left\{ \left\| \mathbf{X}^T \widehat{\boldsymbol{\varepsilon}}/n \right\|_\infty \leq 0.5\lambda\tau \right\} \bigcap \left\{ \left\| \boldsymbol{\beta}^0 - \widehat{\boldsymbol{\beta}}^{ol} \right\|_\infty \leq 0.5\tau \right\}$.

We will show that $A^0 \triangle A^{[t]}$ is eventually empty set on event E, which has a probability tending to 1, where $\triangle$ denotes the symmetric difference. By the optimality criterion Lee and Lee (2005) for (4.12) and (4.13), we have

$$\langle \widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\boldsymbol{\Gamma}}^{[t]}, -\mathbf{X}^T(\boldsymbol{Y} - \mathbf{X}\widetilde{\boldsymbol{\Gamma}}^{[t]})/n + \lambda\tau\nabla \left\| \widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]})^c} \right\|_1 \rangle \geq 0. \tag{S2.2}$$

Rearranging the terms, we have

$$\begin{aligned}
& \left\| \mathbf{X}(\widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\boldsymbol{\Gamma}}^{[t]}) \right\|_2^2 / n \\
\leq & \langle \widetilde{\boldsymbol{\Gamma}}^{[t]} - \widehat{\boldsymbol{\beta}}^{ol}, \mathbf{X}^T \widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla \left\| \widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]})^c} \right\|_1 \rangle \\
= & \langle \widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^0 \setminus A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0 \setminus A^{[t-1]}}, \mathbf{X}^T \widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla \left\| \widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]})^c} \right\|_1 \rangle \\
& + \langle \widetilde{\boldsymbol{\Gamma}}_{A^{[t-1]} \setminus A^0} - \widehat{\boldsymbol{\beta}}^{ol}_{A^{[t-1]} \setminus A^0}, \mathbf{X}^T \widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla \left\| \widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]})^c} \right\|_1 \rangle \\
& + \langle \widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]} \cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]} \cup A^0)^c}, \mathbf{X}^T \widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla \left\| \widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]})^c} \right\|_1 \rangle \\
& + \langle \widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^{[t-1]} \cap A^0} - \widehat{\boldsymbol{\beta}}^{ol}_{A^{[t-1]} \cap A^0}, \mathbf{X}^T \widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla \left\| \widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]})^c} \right\|_1 \rangle.
\end{aligned} \tag{S2.3}$$

Let $\mathbf{X}_{A^0}$ denote the matrix, whose A^0 columns are the same as the A^0 columns of $\mathbf{X}$ and all other columns are zeros. Similarly, we can define $\mathbf{X}_{(A^0)^c}$. Notice that $\mathbf{X} = \mathbf{X}_{A^0} + \mathbf{X}_{(A^0)^c}$, $\widehat{\boldsymbol{\beta}}^{ol}_{(A^0)^c} = \mathbf{0}$, $\widehat{\boldsymbol{\beta}}^{ol} = (\mathbf{X}_{A^0}^T \mathbf{X}_{A^0})^{\dagger} \mathbf{X}_{A^0}^T \boldsymbol{Y}$,

$\widehat{\boldsymbol{Y}} = \mathbf{X}_{A^0}(\mathbf{X}_{A^0}^T\mathbf{X}_{A^0})^{\dagger}\mathbf{X}_{A^0}^T\boldsymbol{Y}$, $P_{A^0} = \mathbf{X}_{A^0}(\mathbf{X}_{A^0}^T\mathbf{X}_{A^0})^{\dagger}\mathbf{X}_{A^0}^T$, as well as $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (I - P_{A^0})\boldsymbol{Y} = (I - P_{A^0})\boldsymbol{\varepsilon}$. Note that $\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}} = \mathbf{X}_{(A^0)^c}^T\widehat{\boldsymbol{\varepsilon}}$.

Thus, we know that the fourth term in the last equation of (S2.3) vanishes, since

$$\begin{aligned}
&\langle \widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^{[t-1]}\cap A^0} - \widehat{\boldsymbol{\beta}}^{ol}_{A^{[t-1]}\cap A^0}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla \left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]})^c}\right\|_1 \rangle \\
=&\langle \widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^{[t-1]}\cap A^0} - \widehat{\boldsymbol{\beta}}^{ol}_{A^{[t-1]}\cap A^0}, \mathbf{X}^T_{(A^0)^c}\widehat{\boldsymbol{\varepsilon}}/n \rangle \\
=&\langle \mathbf{X}_{(A^0)^c}(\widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^{[t-1]}\cap A^0} - \widehat{\boldsymbol{\beta}}^{ol}_{A^{[t-1]}\cap A^0}), \widehat{\boldsymbol{\varepsilon}}/n \rangle = 0.
\end{aligned} \tag{S2.4}$$

Note that the third term in the last equation of (S2.3) satisfies

$$\langle \widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}, \nabla \left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]})^c}\right\|_1 \rangle = \left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}\right\|_1, \tag{S2.5}$$

because $\widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c} = 0$.

Recalling (S2.3), we have

$$\begin{aligned}
0 \le& \left\|\mathbf{X}(\widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\boldsymbol{\Gamma}}^{[t]})\right\|_2^2/n \\
\le& \langle \widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^0\setminus A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0\setminus A^{[t-1]}}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla \left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^0\setminus A^{[t-1]}}\right\|_1 \rangle \\
&+ \langle \widetilde{\boldsymbol{\Gamma}}_{A^{[t-1]}\setminus A^0} - \widehat{\boldsymbol{\beta}}^{ol}_{A^{[t-1]}\setminus A^0}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n \rangle \\
&+ \langle \widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}, \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}/n - \lambda\tau\nabla \left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c}\right\|_1 \rangle \\
\le& \left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^0\triangle A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0\triangle A^{[t-1]}}\right\|_1 \cdot (\left\|\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}\right\|_\infty/n + \lambda\tau) \\
&+ \left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}\right\|_1 \cdot (\left\|\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}\right\|_\infty/n - \lambda\tau).
\end{aligned} \tag{S2.6}$$

Rearranging inequality (S2.6) implies that

$$\begin{aligned}
&\left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}\right\|_1 \cdot (\lambda\tau - \left\|\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}\right\|_\infty/n) \\
\le& \left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^0\triangle A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0\triangle A^{[t-1]}}\right\|_1 \cdot (\left\|\mathbf{X}^T\widehat{\boldsymbol{\varepsilon}}\right\|_\infty/n + \lambda\tau).
\end{aligned}$$

Over event E, we have

$$\begin{aligned}
\left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{(A^{[t-1]}\cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]}\cup A^0)^c}\right\|_1 &\le 3\left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^0\triangle A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0\triangle A^{[t-1]}}\right\|_1 \\
&\le 3\left\|\widetilde{\boldsymbol{\Gamma}}^{[t]}_{A^0\cup A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0\cup A^{[t-1]}}\right\|_1.
\end{aligned}$$

Recall that $\kappa_1 = \left|\{i \in [p]\backslash B; |\widetilde{\Gamma}_i^{[0]}| \geq \tau\}\right|$ and $\kappa_{max} = \max\{\kappa, \kappa_1\}$. Without loss of generality, we can assume that $\widetilde{\mathbf{\Gamma}}_B^{[0]} = \mathbf{0}$. For the base case, we know that

$$\left|A^0 \triangle A^{[0]}\right| \leq |A^0 \backslash B| + |A^{[0]} \backslash A^0| \leq 2\kappa_{max}.$$

For the induction step, assume $\left|A^0 \triangle A^{[t-1]}\right| \leq 2\kappa_{max}$ on event E, for $t \geq 1$. We aim to show that $\left|A^0 \triangle A^{[t]}\right| \leq 2\kappa_{max}$ over event E.

Applying Assumption 1 and (S2.6), over event E, we have

$$\begin{aligned}
&c_1 \left\|\widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\mathbf{\Gamma}}^{[t]}\right\|_2^2 \leq \left\|\mathbf{X}(\widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\mathbf{\Gamma}}^{[t]})\right\|_2^2 / n \\
\leq &\left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{A^0 \triangle A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0 \triangle A^{[t-1]}}\right\|_1 \cdot (\left\|\mathbf{X}^T \widehat{\boldsymbol{\varepsilon}}\right\|_\infty / n + \lambda\tau) \\
&+ \left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{(A^{[t-1]} \cup A^0)^c} - \widehat{\boldsymbol{\beta}}^{ol}_{(A^{[t-1]} \cup A^0)^c}\right\|_1 \cdot (\left\|\mathbf{X}^T \widehat{\boldsymbol{\varepsilon}}\right\|_\infty / n - \lambda\tau) \\
\leq &\frac{3}{2}\lambda\tau \left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{A^0 \triangle A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0 \triangle A^{[t-1]}}\right\|_1.
\end{aligned} \tag{S2.7}$$

By Cauchy-Schwarz inequality,

$$\left\|\widetilde{\mathbf{\Gamma}}^{[t]}_{A^0 \triangle A^{[t-1]}} - \widehat{\boldsymbol{\beta}}^{ol}_{A^0 \triangle A^{[t-1]}}\right\|_1^2 \leq \left|A^0 \triangle A^{[t-1]}\right| \left\|\widetilde{\mathbf{\Gamma}}^{[t]} - \widehat{\boldsymbol{\beta}}^{ol}\right\|_2^2. \tag{S2.8}$$

Combining (S2.7), (S2.8) and the third condition in Theorem 1,

$$\left\|\widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\mathbf{\Gamma}}^{[t]}\right\|_2 \leq \frac{3\lambda\tau}{2c_1}\sqrt{|A^0 \triangle A^{[t-1]}|} \leq \frac{\tau}{4}\sqrt{|A^0 \triangle A^{[t-1]}|}. \tag{S2.9}$$

Applying (S2.9) and induction assumption $\left|A^0 \triangle A^{[t-1]}\right| \leq 2\kappa_{max}$, we have

$$\left\|\widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\mathbf{\Gamma}}^{[t]}\right\|_2 / \tau \leq \frac{1}{4}\sqrt{2\kappa_{max}} \leq \frac{1}{2}\sqrt{\kappa_{max}} \leq \sqrt{\kappa_{max}}. \tag{S2.10}$$

Because for any $i \in A^{[t]} \backslash A^0$, $|\widetilde{\Gamma}_i^{[t]} - \widehat{\beta}_i^{ol}| = |\widetilde{\Gamma}_i^{[t]}| \geq \tau$, we have

$$\sqrt{|A^{[t]} \backslash A^0|} \leq \left\|\widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\mathbf{\Gamma}}^{[t]}\right\|_2 / \tau \leq \sqrt{\kappa_{\max}}.$$

Thus,

$$\left|A^0 \triangle A^{[t]}\right| \leq |A^0 \backslash B| + |A^{[t]} \backslash A^0| \leq 2\kappa_{max}.$$

By induction, we already showed that over event E, $\left|A^0 \triangle A^{[t]}\right| \leq 2\kappa_{max}$ for any $0 \leq t \leq t_{max}$, where t_{max} denotes the total number of the iteration of Algorithm 3.

Note that Algorithm 3 is terminated at t if

$$\text{supp}\{\widetilde{\mathbf{\Gamma}}^{[t]}\}\backslash B = \text{supp}\{\widetilde{\mathbf{\Gamma}}^{[t-1]}\}\backslash B.$$

We will show that over event E, for any $t \geq 1$,

$$\sqrt{|A^0 \triangle A^{[t]}|} \leq \frac{1}{2}\sqrt{|A^0 \triangle A^{[t-1]}|}.$$

If $i \in A^0 \backslash A^{[t]}$, then we know that $\beta_i^0 \neq 0$ and $\widetilde{\Gamma}_i^{[t]} = 0$. Over event E, $|\beta_i^0 - \widehat{\beta}_i^{ol}| \leq 0.5\tau$. According to assumption $\kappa \geq |A_{H_0}^0|$ in Theorem 1, we know that

$$|\widetilde{\Gamma}_i^{[t]} - \widehat{\beta}_i^{ol}| \geq |\widetilde{\Gamma}_i^{[t]} - \beta_i^0| - |\beta_i^0 - \widehat{\beta}_i^{ol}| \geq \tau - 0.5\tau = 0.5\tau. \tag{S2.11}$$

If $i \in A^{[t]} \backslash A^0$, then we know that $\beta_i^0 = \widehat{\beta}_i^{ol} = 0$ and $|\widetilde{\Gamma}_i^{[t]}| \geq \tau$, which implies $|\widetilde{\Gamma}_i^{[t]} - \widehat{\beta}_i^{ol}| = |\widetilde{\Gamma}_i^{[t]}| \geq \tau$.

Over event E, we obtain that for any $i \in A^0 \triangle A^{[t]}$, $|\widetilde{\Gamma}_i^{[t]} - \widehat{\beta}_i^{ol}| \geq \tau - 0.5\tau = 0.5\tau$. Combining (S2.9), we obtain

$$\sqrt{|A^0 \triangle A^{[t]}|} \leq \frac{1}{0.5\tau}\left\|\widehat{\boldsymbol{\beta}}^{ol} - \widetilde{\mathbf{\Gamma}}^{[t]}\right\|_2 \leq \frac{1}{2}\sqrt{|A^0 \triangle A^{[t-1]}|}. \tag{S2.12}$$

In conclusion, over event E, we obtain

$$\sqrt{|A^0 \triangle A^{[t]}|} \leq \frac{1}{2^t}\sqrt{2\kappa_{max}}. \tag{S2.13}$$

If $t \geq \lceil \frac{\log(2\kappa_{max})}{\log 4} \rceil$, $\sqrt{|A^0 \triangle A^{[t]}|} < 1$, i.e., $A^0 \triangle A^{[t]} = \emptyset$.

Set $t_{max} = \lceil \frac{\log(2\kappa_{max})}{\log 4} \rceil$. We already showed that over event E,

$$\{i \in [p]\backslash B : |\widetilde{\Gamma}_i^{[t_{max}]}| \geq \tau\} = \{i \in [p]\backslash B : \beta_i^0 \neq 0\}.$$

This means that for any $j \notin \{i \in [p]\backslash B : |\widetilde{\Gamma}_i^{[t_{max}]}| \geq \tau\}$, $|\widetilde{\Gamma}_j^{[t_{max}]}| < \tau$, and for any $k \in \{i \in [p]\backslash B : \widetilde{\Gamma}_i^{[t_{max}]} \geq \tau\}$, $|\widetilde{\Gamma}_k^{[t_{max}]}| \geq \tau$. Thus, over the event E

$$\begin{aligned}\text{supp}\{\boldsymbol{\beta}^0\}\backslash B &= \{i \in [p]\backslash B : \beta_i^0 \neq 0\} \\ &= \{i \in [p]\backslash B : \left|\widetilde{\Gamma}_i^{[t_{max}]}\right| \geq \tau\} \subset \text{supp}\{\widehat{\mathbf{\Gamma}}\}\backslash B.\end{aligned} \tag{S2.14}$$

Next, we aim to bound the probability of event E. Let $\boldsymbol{a} := \mathbf{X}^T\widehat{\boldsymbol{\varepsilon}} = (a_1, \cdots, a_p)^T$. It is obvious that $\boldsymbol{a} = \mathbf{X}^T(I - P_{A^0})\boldsymbol{\varepsilon} = \mathbf{X}_{(A^0)^c}^T(I - P_{A^0})\boldsymbol{\varepsilon}$.

Recall the error $\boldsymbol{\varepsilon} \sim N(0, \Sigma)$ with $\Sigma = \sigma^2 I$. By Assumption 2, for any $1 \leq l \leq p$,

$$\mathrm{Var}(a_l) = \boldsymbol{e}_l^T \mathbf{X}^T (I - P_{A^0}) \Sigma (I - P_{A^0}) \mathbf{X} \boldsymbol{e}_l \leq \sigma^2 (\mathbf{X}^T (I - P_{A^0}) \mathbf{X})_{ll} \leq n\sigma^2 c_2^2. \tag{S2.15}$$

By the upper-tail inequality for sub-gaussian distribution and the third condition in Theorem 1, we have

$$\begin{aligned} &\mathbb{P}(\left\|\mathbf{X}^T \widehat{\boldsymbol{\varepsilon}}\right\|_\infty / n > 0.5\lambda\tau) = \mathbb{P}(\|\boldsymbol{a}\|_\infty / n > 0.5\lambda\tau) \\ \leq &\sum_{l=1}^{p} \mathbb{P}(|\boldsymbol{a}|_l > 0.5 n\lambda\tau) \leq 2p \exp\left(-\frac{1}{8}\frac{n\lambda^2\tau^2}{\sigma^2 c_2^2}\right) \leq \frac{2}{p^3 n^4}. \end{aligned} \tag{S2.16}$$

Set $\boldsymbol{b} := \widehat{\boldsymbol{\beta}}^{ol} - \boldsymbol{\beta}^0$. A direct calculation implies that

$$\boldsymbol{b} = (\mathbf{X}_{A^0}^T \mathbf{X}_{A^0})^{\dagger} \mathbf{X}_{A^0}^T \boldsymbol{\varepsilon} = (b_1, \cdots, b_p)^T.$$

By Assumption 2, for any $1 \leq l \leq p$,

$$\begin{aligned} \mathrm{Var}(b_l) =& \boldsymbol{e}_l^T (\mathbf{X}_{A^0}^T \mathbf{X}_{A^0})^{\dagger} \mathbf{X}_{A^0}^T \Sigma \mathbf{X}_{A^0} (\mathbf{X}_{A^0}^T \mathbf{X}_{A^0})^{\dagger} \boldsymbol{e}_l \\ \leq& \sigma^2 \big((\mathbf{X}_{A^0}^T \mathbf{X}_{A^0})^{\dagger}\big)_{ll} \leq \frac{1}{n}\sigma^2 c_3^2 I(l \in A^0). \end{aligned}$$

By Assumptions 2 and (4.17), we have

$$\begin{aligned} &\mathbb{P}\left(\left\|\widehat{\boldsymbol{\beta}}^{ol} - \boldsymbol{\beta}^0\right\|_\infty > 0.5\tau\right) \leq \sum_{i \in A^0} \mathbb{P}(|\boldsymbol{b}|_i > 0.5\tau) \\ \leq& 2|A^0| \exp\left(-\frac{1}{8}\frac{n\tau^2}{\sigma^2 c_3^2}\right) \leq \frac{2(\kappa_{H_0}^0 + |B|)}{p^4 n^4} = o\Big(\frac{1}{p^4 n^3}\Big). \end{aligned} \tag{S2.17}$$

Case 1: $\kappa = |A_{H_0}^0|$. In this case, (S2.14) implies that over the event E

$$\mathrm{supp}\{\boldsymbol{\beta}^0\} \backslash B = \mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\} \backslash B. \tag{S2.18}$$

Thus, under the requirement for (τ, λ) in Theorem 1,

$$\begin{aligned} &\mathbb{P}(\mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\} \backslash B \neq A^0 \backslash B) \\ =& \mathbb{P}(\{\mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\} \backslash B \neq A^0 \backslash B\} \cap E) + \mathbb{P}(\{\mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\} \backslash B \neq A^0 \backslash B\} \cap E^c). \end{aligned} \tag{S2.19}$$

Notice that $E \subset \{\mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\} \backslash B = A^{[t_{max}]} \backslash B\}$, and $E \subset \{A^{[t_{max}]} \backslash B = A^0 \backslash B\}$. Thus,

$$\mathbb{P}(\{\mathrm{supp}\{\widehat{\boldsymbol{\Gamma}}\} \backslash B \neq A^0 \backslash B\} \cap E) = 0, \text{ and}$$

$$\mathbb{P}(\text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B \neq A^0\backslash B) \leq \mathbb{P}(E^c) \leq \frac{C'}{n^3}, \tag{S2.20}$$

where C' is a absolute constant.

Note that $\{\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\beta}}^{ol}\} = \{\text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B = A^0\backslash B\}$. By Borel-Cantelli lemma, we have $\{\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\beta}}^{ol}\}$ almost surely as $n \to \infty$.

It remains to show that $\widehat{\boldsymbol{\beta}}^{ol}$ is a global minimizer of (4.8) or (4.9) with high probability.

Applying Theorem 2 of Shen et al. (2013) and its proof, with the degree of separation condition 4, we obtain

$$\begin{aligned}\mathbb{P}\left(\widehat{\boldsymbol{\beta}}^{\ell_0} \neq \widehat{\boldsymbol{\beta}}^{ol}\right) &\leq \frac{e+1}{e-1}\exp\left(-\frac{n}{18\sigma^2}\left(C_{\min} - 36\frac{\log p + \log n}{n}\sigma^2\right)\right)\\ &\leq \frac{e+1}{e-1}\frac{1}{p^2n^2},\end{aligned}$$

where $\widehat{\boldsymbol{\beta}}^{\ell_0}$ denotes the global minimizer of (4.8) or (4.9). By Borel-Cantelli lemma, we have $\{\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\beta}}^{ol} = \widehat{\boldsymbol{\beta}}^{\ell_0}\}$ almost surely as $n \to \infty$.

Case 2: $\kappa \geq |A^0_{H_0}|$. Thus, under the requirement for (τ, λ) in Theorem 1,

$$\begin{aligned}&\mathbb{P}(\text{supp}\{\boldsymbol{\beta}^0\}\backslash B \not\subset \text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B)\\ =&\mathbb{P}(\{\text{supp}\{\boldsymbol{\beta}^0\}\backslash B \not\subset \text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\} \cap E)\\ &+ \mathbb{P}(\{\text{supp}\{\boldsymbol{\beta}^0\}\backslash B \not\subset \text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\} \cap E^c).\end{aligned}$$

By (S2.14), we know that $E \subset \{\text{supp}\{\boldsymbol{\beta}^0\}\backslash B \subset \text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\}$. Thus,

$$\mathbb{P}(\{\text{supp}\{\boldsymbol{\beta}^0\}\backslash B \not\subset \text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\} \cap E) = 0 \text{ and}$$

$$\mathbb{P}\big(\text{supp}\{\boldsymbol{\beta}^0\}\backslash B \not\subset \text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\big) \leq \mathbb{P}(E^c) \leq \frac{C'}{n^3}. \tag{S2.21}$$

By the Borel-Cantelli lemma, we have $\{\text{supp}\{\boldsymbol{\beta}^0\}\backslash B \subset \text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\}$ almost surely as $n \to \infty$.

Under H_0, we know that $\text{supp}\{\widehat{\boldsymbol{\beta}}^{ol}_{H_0}\} \subset \text{supp}\{\boldsymbol{\beta}^0\}\backslash B$, which implies that $\{\text{supp}\{\widehat{\boldsymbol{\beta}}^{ol}_{H_0}\} \subset \text{supp}\{\widehat{\boldsymbol{\Gamma}}\}\backslash B\}$ almost surely as $n \to \infty$.

Under H_1, we know that $\text{supp}\{\widehat{\boldsymbol{\beta}}^{ol}_{H_1}\} \subset \text{supp}\{\boldsymbol{\beta}^0\} \cup B$, which implies that $\{\text{supp}\{\widehat{\boldsymbol{\beta}}^{ol}_{H_1}\} \subset \text{supp}\{\widehat{\boldsymbol{\Gamma}}\} \cup B\}$ almost surely as $n \to \infty$. $\square$

S3 Proof of Theorem 2

Proof. By Theorem 1, we have

$$\lim_{n\to\infty} \mathbb{P}(\{\widehat{\mathbf{\Gamma}}^{(0)} = \widehat{\boldsymbol{\beta}}^{ls}_{H_0}\} \cap \{\widehat{\mathbf{\Gamma}}^{(1)} = \widehat{\boldsymbol{\beta}}^{ls}_{H_1}\}) = 1, \tag{S3.22}$$

where $\widehat{\mathbf{\Gamma}}^{(0)}$ and $\widehat{\mathbf{\Gamma}}^{(1)}$ are obtained from Algorithm 3, and $\widehat{\boldsymbol{\beta}}^{ls}_{H_0}$ and $\widehat{\boldsymbol{\beta}}^{ls}_{H_1}$ are obtained from (4.10) and (4.11), respectively.

The remainder of the proof follows the argument in Theorem 2 of Zhu et al. (2020), which establishes the sampling distribution of $\Lambda_n(B)$.

□

S4 Proof of Theorem 3

Proof. By Theorem 1, we have

$$\lim_{n\to\infty} \mathbb{P}(\widehat{\mathbf{\Gamma}}^{(1)} = \widehat{\boldsymbol{\beta}}^{ls}) = 1, \tag{S4.23}$$

where $\widehat{\mathbf{\Gamma}}^{(1)}$ is obtained from Algorithm 3 and $\widehat{\boldsymbol{\beta}}^{ls} = \widehat{\boldsymbol{\beta}}^{ls}_{H_1}$ is obtained from (4.11), supported on $A^0 \cup B$.

Thus, to show (5.26), it suffices to show

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}^{ls}_B - \boldsymbol{\beta}^0_B) \xrightarrow{d} N(0, \Sigma), \tag{S4.24}$$

as $n \to \infty$.

Let $t \in \mathbb{R}^B$ and $u \in \mathbb{R}^{A^0\cup B}$ such that $u_B = t$ and $u_{B^c} = 0$. Note that

$$\widehat{\boldsymbol{\beta}}^{ls}_{A^0\cup B} = (\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})^\dagger(\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})\boldsymbol{\beta}^0_{A^0\cup B} + (\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})^\dagger\mathbf{X}^T_{A^0\cup B}\boldsymbol{\varepsilon}, \tag{S4.25}$$

where $\mathbf{X}_{A^0\cup B}$ denotes the sub-matrix with columns of $A^0 \cup B$.

Due to $u \in \{\boldsymbol{\xi} \in \mathbb{R}^{A^0\cup B}; \boldsymbol{\xi}_i = 0, i \notin B\} \subset \mathcal{R}(\mathbf{X}^T_{A^0\cup B})$, we know that there exists vector v such that $u = \mathbf{X}^T_{A^0\cup B}v$ and

$$\begin{aligned}&(\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})(\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})^\dagger u\\ =&\mathbf{X}^T_{A^0\cup B}\Big(\mathbf{X}_{A^0\cup B}(\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})^\dagger\mathbf{X}^T_{A^0\cup B}\Big)v = \mathbf{X}^T_{A^0\cup B}v = u.\end{aligned}$$

By direct calculation of the characteristic function of $\sqrt{n}(\widehat{\boldsymbol{\beta}}^{ls}_B - \boldsymbol{\beta}^0_B)$, we know that

$$
\begin{aligned}
&\mathbb{E}e^{it^T\sqrt{n}(\widehat{\boldsymbol{\beta}}^{ls}_B-\boldsymbol{\beta}^0_B)} = \mathbb{E}e^{iu^T\sqrt{n}(\widehat{\boldsymbol{\beta}}^{ls}_{A^0\cup B}-\boldsymbol{\beta}^0_{A^0\cup B})}\\
=&\mathbb{E}_{\mathbf{X}}\mathbb{E}_{\boldsymbol{\varepsilon}}e^{iu^T\sqrt{n}(\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})^{\dagger}\mathbf{X}^T_{A^0\cup B}\boldsymbol{\varepsilon}}\\
=&\mathbb{E}_{\mathbf{X}}\exp\{-1/2\cdot\sigma^2u^Tn(\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})^{\dagger}\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B}(\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})^{\dagger}u\}\\
=&\mathbb{E}_{\mathbf{X}}\exp\{-1/2\cdot\sigma^2u^T(1/n\cdot\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})^{\dagger}u\}\\
=&\mathbb{E}_{\mathbf{X}}\exp\{-1/2\cdot\sigma^2t^T(1/n\cdot\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})^{\dagger}_{B,B}t\}.
\end{aligned}
$$

Under the assumption of Moore–Penrose inverse $\sigma^2\left(\frac{1}{n}\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B}\right)^{\dagger}_{B,B}$ converges in distribution to some positive semi-definite matrix Σ, and by applying the continuous mapping theorem, we obtain that

$$
\mathbb{E}_{\mathbf{X}}\exp\{-1/2\cdot\sigma^2t^T(1/n\cdot\mathbf{X}^T_{A^0\cup B}\mathbf{X}_{A^0\cup B})^{\dagger}_{B,B}t\}\to e^{-1/2t^T\Sigma t}, \tag{S4.26}
$$

as $n\to\infty$, for any $t\in\mathbb{R}^B$.

By Lévy's continuity theorem, we complete the proof of weak convergence (S4.24). Therefore, the proof of Theorem 3 is completed. $\square$

S5 Top-κ index set Selection

Algorithm S2 Threshold-Based Top-κ index set Selection

Inputs:

- Current parameter estimates $\{\widetilde{\Gamma}^{[t]}_{k,j}\}_{(k,j)\in\mathcal{S}}$. Target sparsity level κ.
- Initial thresholds a, b such that

$$\left|\left\{(k,j) : |\widetilde{\Gamma}^{[t]}_{k,j}| \le a\right\}\right| < \kappa \quad \text{and} \quad \left|\left\{(k,j) : |\widetilde{\Gamma}^{[t]}_{k,j}| \le b\right\}\right| > \kappa.$$

Initialize: Set $c \leftarrow \frac{a+b}{2}$.

while $b - a$ is not sufficiently small **do**

Site-Level Operations: Each site counts how many local parameters satisfy $|\widetilde{\Gamma}^{[t]}_{k,j}| \le c$ and sends *only* this count to a central location.

Aggregation and Update:

- Sum the local counts with the center count to get the total count

$$\left|\left\{(k,j) : |\widetilde{\Gamma}^{[t]}_{k,j}| \le c\right\}\right|.$$

- If the total is exactly κ, **break** the loop.
- If the total is $< \kappa$, set $a \leftarrow c$. Otherwise, set $b \leftarrow c$. Update $c \leftarrow \frac{a+b}{2}$.

end while

Finalize Threshold c^*: Let $c^* \leftarrow c$ if the loop ended early due to an exact match, or set $c^* \leftarrow a$ if we finished the binary search without an exact match.

Construct the ℓ_0 Projected Set:

- Each site identifies which local parameters exceed $|\widetilde{\Gamma}^{[t]}_{k,j}| > c^*$.
- Let C be the union of those parameter indices across all sites,

$$C = \left\{(k,j) : |\widetilde{\Gamma}^{[t]}_{k,j}| \ge \text{threshold } c^*\right\}.$$

- If $|C| < \kappa$, select additional $\kappa - |C|$ parameters with $|\widetilde{\Gamma}^{[t]}_{k,j}| \in (a, b)$ until $|C| = \kappa$.

Output: The projected index set C.

References

Lee, G. M. and K. B. Lee (2005). Vector variational inequalities for nondifferentiable convex vector optimization problems. *Journal of Global Optimization 32*, 597–612.

Shen, X., W. Pan, Y. Zhu, and H. Zhou (2013). On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics 65*(5), 807–832.

Zhu, Y., X. Shen, and W. Pan (2020). On high-dimensional constrained maximum likelihood inference. *Journal of the American Statistical Association 115*(529), 217–230.