Supplementary materials

Appendix A: Proofs of Theorems

For any two real number-valued $p \times q$ matrices $A = (A_{ij})$ and $B = (B_{ij})$, define their Frobenius inner product as $\langle A, B \rangle_F = \sum_{i,j} A_{ij} B_{ij}$. For any matrix $\boldsymbol{M}$, we use $\|\boldsymbol{M}\|_F$ and $\|\boldsymbol{M}\|_{op}$ to denote its Frobenius norm and operator norm, respectively. For the sake of clarity, we define $\Omega_{r,s} = \{\boldsymbol{C} : \text{rank}(\boldsymbol{C}) = r,\ \|\boldsymbol{C}\|_{2,0} \le s\}$ as the set of the low-rank and row-sparse matrices and $B_F(r_0, \boldsymbol{C}^*)$ as a ball centered at $\boldsymbol{C}^*$ with radius r_0, which will be used later.

A.1. Proof of Proposition 1

To begin with, we consider the second-order expansion of the loss function $L(\boldsymbol{C}; \mathcal{D})$ by applying Taylor's theorem with Lagrange remainder. That is, fix the current estimate $\boldsymbol{C}^m$, and suppose $\boldsymbol{C}$ is in a neighborhood of $\boldsymbol{C}^m$, then there exists $0 < \theta < 1$ such that

$$L(\boldsymbol{C}; \mathcal{D}) = L(\boldsymbol{C}^m; \mathcal{D}) - \frac{1}{n} \left\langle \boldsymbol{X}(\boldsymbol{C} - \boldsymbol{C}^m), (\tilde{\boldsymbol{Y}} - \tilde{P}^{\boldsymbol{C}^m}) \right\rangle_F + \frac{1}{2n} \sum_{i=1}^{n} \boldsymbol{x}_i^\top (\boldsymbol{C} - \boldsymbol{C}^m) \boldsymbol{H}_i^{\boldsymbol{C}'} (\boldsymbol{C} - \boldsymbol{C}^m)^\top \boldsymbol{x}_i,$$

where $\boldsymbol{H}_i^{\boldsymbol{C}'}$ is the within observation Hessian matrix for the coefficient matrix $\boldsymbol{C}' = \theta \boldsymbol{C} + (1-\theta)\boldsymbol{C}^m$, $i = 1, \ldots, n$. By the definition of $S(\boldsymbol{C}; \mathcal{D} | \boldsymbol{C}^{(m)})$, we have

$$S(\boldsymbol{C}; \mathcal{D} | \boldsymbol{C}^{(m)}) - L(\boldsymbol{C}; \mathcal{D}) = \frac{1}{2n} \sum_{i=1}^{n} \boldsymbol{x}_i^\top (\boldsymbol{C} - \boldsymbol{C}^m) \left(\frac{1}{2} I_{q-1} - \boldsymbol{H}_i^{\boldsymbol{C}'} \right) (\boldsymbol{C} - \boldsymbol{C}^m)^\top \boldsymbol{x}_i.$$

In Simon et al. (2013) Lemma 3.3 has proved, for any coefficient matrix $\boldsymbol{C}'$,

$$\boldsymbol{H}_i^{\boldsymbol{C}'} \preceq 2 \max_{ij} \{ \boldsymbol{P}_{ij}^{\boldsymbol{C}'} (1 - \boldsymbol{P}_{ij}^{\boldsymbol{C}'}) \} I \preceq \frac{1}{2} I_{q-1}.$$

Therefore, we have

$$S(\boldsymbol{C}; \mathcal{D} | \boldsymbol{C}^{(m)}) \ge L(\boldsymbol{C}; \mathcal{D})$$

for all $\boldsymbol{C}$ around the current estimate $\boldsymbol{C}^m$. Finally, the equality $S(\boldsymbol{C}^m; \mathcal{D} | \boldsymbol{C}^{(m)}) = L(\boldsymbol{C}^m; \mathcal{D})$ holds by applying the definition of $S(\boldsymbol{C}; \mathcal{D} | \boldsymbol{C}^{(m)})$.

A.2. Proof of Theorem 1

Recall that $\hat{\boldsymbol{C}}$ is the minimizer of (4), then by the definition of surrogate function, we have

$$\hat{\boldsymbol{C}} = \arg \min_{\boldsymbol{C} \in \Omega_r} S(\boldsymbol{C}; \mathcal{D} | \hat{\boldsymbol{C}}).$$

Therefore, we can derive an explicit expression for $\hat{\boldsymbol{C}}$ as follows.

$$\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} = (\boldsymbol{X}_{\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\hat{\mathcal{A}}})^{-1} \boldsymbol{X}_{\hat{\mathcal{A}}}^\top M^{\hat{\boldsymbol{C}}} \boldsymbol{V}^{\hat{\boldsymbol{C}}} (\boldsymbol{V}^{\hat{\boldsymbol{C}}})^\top,$$
$$\hat{\boldsymbol{C}}_{(\hat{\mathcal{A}})^c \cdot} = \boldsymbol{0},$$

where $M^{\hat{\boldsymbol{C}}} = \boldsymbol{X}\hat{\boldsymbol{C}} + 2(\tilde{Y} - \tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}})$ and $\boldsymbol{V}^{\hat{\boldsymbol{C}}} \in \mathbb{R}^{q\times r}$ consists of the eigenvectors that corresponds to the top r-th largest eigenvalues of $(M^{\hat{\boldsymbol{C}}})^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}}(\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top M^{\hat{\boldsymbol{C}}}$. It can be seen that $(\boldsymbol{V}^{\hat{\boldsymbol{C}}})^\top \boldsymbol{V}^{\hat{\boldsymbol{C}}} = I_r$. Firstly, we take difference between $\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot}$ and $\boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}$, i.e.,

$$\begin{aligned}
\|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}\|_F &= \|(\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top M^{\hat{\boldsymbol{C}}}\boldsymbol{V}^{\hat{\boldsymbol{C}}}(\boldsymbol{V}^{\hat{\boldsymbol{C}}})^\top - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}\|_F \\
&= \|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot}\boldsymbol{V}^{\hat{\boldsymbol{C}}}(\boldsymbol{V}^{\hat{\boldsymbol{C}}})^\top - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot} + (\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}})\boldsymbol{V}^{\hat{\boldsymbol{C}}}(\boldsymbol{V}^{\hat{\boldsymbol{C}}})^\top\|_F \\
&= \|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot}\boldsymbol{V}^{\hat{\boldsymbol{C}}} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}\boldsymbol{V}^{\hat{\boldsymbol{C}}} + (\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\ \boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}})\boldsymbol{V}^{\hat{\boldsymbol{C}}}\|_F \\
&= \|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot}\boldsymbol{V}^{\hat{\boldsymbol{C}}} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}\boldsymbol{V}^{\hat{\boldsymbol{C}}} + (\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*} + \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*} - \tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}})\boldsymbol{V}^{\hat{\boldsymbol{C}}}\|_F \\
&\le \|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot} + (\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*} + \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*} - \tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}})\|_F.
\end{aligned}$$

Using the triangle inequality gives

$$\|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}\| \le \|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot} - (\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2(\tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})\|_F + \|(\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})\|_F. \quad (1)$$

We know that the (i,k)-th element of $\tilde{\boldsymbol{P}}^{\boldsymbol{C}}$ is actually a function of $\boldsymbol{c}_k^\top \boldsymbol{x}_i$ with other elements fixed. Let $f(\cdot)$ denote this function, i.e., $f(x) = \frac{\exp(x)}{\alpha+\exp(x)}$, where α is a constant. Then it is easy to show that $f(x)$ has the Lipschitz continuity, that is,

$$\tilde{c}\|\boldsymbol{X}(\hat{\boldsymbol{C}} - \boldsymbol{C}^*)\|_F \le \|\tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*}\|_F \le \frac{1}{4}\|\boldsymbol{X}(\hat{\boldsymbol{C}} - \boldsymbol{C}^*)\|_F.$$

Therefore, each element of $\tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*}$, denoted by $\Delta_{i,j}$, satisfies $\frac{1}{4}|(\boldsymbol{X}\hat{\boldsymbol{C}} - \boldsymbol{X}\boldsymbol{C}^*)_{i,j}| \ge |\Delta_{i,j}| \ge \tilde{c}|(\boldsymbol{X}\hat{\boldsymbol{C}} - \boldsymbol{X}\boldsymbol{C}^*)_{i,j}|$. For convenience, we denote $\Delta = (\Delta_{i,j})$ with $\Delta_{i,j} = \alpha_{i,j}(\boldsymbol{X}\hat{\boldsymbol{C}} - \boldsymbol{X}\boldsymbol{C}^*)_{i,j}$ and $\alpha_{i,j}$ lies in $[\tilde{c}, \frac{1}{4}]$.

Next we divide the columns of $\boldsymbol{X}$ by the sets of $\hat{\mathcal{A}}$ and $\mathcal{A}^*$. In particular, define three instrumental matrices $\boldsymbol{X}_1, \boldsymbol{X}_2$ and $\boldsymbol{X}_3$ as following

$$\begin{aligned}
&(\boldsymbol{X}_1)_{\cdot\hat{\mathcal{A}}} = \boldsymbol{X}_{\cdot\hat{\mathcal{A}}}, \\
&(\boldsymbol{X}_1)_{\cdot\hat{\mathcal{A}}^c} = \boldsymbol{0}, \\
&(\boldsymbol{X}_2)_{\cdot\mathcal{A}^*-\hat{\mathcal{A}}} = \boldsymbol{X}_{\cdot\mathcal{A}^*-\hat{\mathcal{A}}}, \\
&(\boldsymbol{X}_2)_{\cdot(\mathcal{A}^*-\hat{\mathcal{A}})^c} = \boldsymbol{0}, \\
&(\boldsymbol{X}_3)_{\cdot(\mathcal{A}^*\cup\hat{\mathcal{A}})^c} = \boldsymbol{X}_{\cdot(\mathcal{A}^*-\hat{\mathcal{A}})^c}, \\
&(\boldsymbol{X}_3)_{\cdot\mathcal{A}^*\cup\hat{\mathcal{A}}} = \boldsymbol{0}.
\end{aligned}$$

Similarly, we can separate the matrix Δ into three instrumental matrices, Δ_1, Δ_2, and Δ_3, in the following way.

$$\begin{aligned}
(\Delta_1)_{i,j} &= \alpha_{i,j}(\boldsymbol{X}_1\hat{\boldsymbol{C}} - \boldsymbol{X}_1\boldsymbol{C}^*)_{i,j}, \\
(\Delta_2)_{i,j} &= \alpha_{i,j}(\boldsymbol{X}_2\hat{\boldsymbol{C}} - \boldsymbol{X}_2\boldsymbol{C}^*)_{i,j}, \\
(\Delta_3)_{i,j} &= \alpha_{i,j}(\boldsymbol{X}_3\hat{\boldsymbol{C}} - \boldsymbol{X}_3\boldsymbol{C}^*)_{i,j}.
\end{aligned}$$

Notice that $\Delta_3 \equiv 0$ and we define Δ_3 to provide a symmetric definition for Δ_1. Since $\Delta = \Delta_1 + \Delta_2 + \Delta_3$, we can rewrite the inequality in (1) as

$$\begin{aligned}
\|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}\| &\le \|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot} - (\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2(\Delta_1 + \Delta_2 + \Delta_3)\|_F \\
&\quad + \|(\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})\|_F \\
&\le \|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot} - (\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2\Delta_1\|_F + \|(\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2\Delta_2\|_F \\
&\quad + \|(\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2\Delta_3\|_F + \|(\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})\|_F,
\end{aligned} \quad (2)$$

where a triangle inequality is used in the last inequality. For the first term, by the definition Δ_1, we know that each element of the matrix $(\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \Delta_1$ is actually $\alpha_{i,j}$. Thus, we can its upper bound by

$$\|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot} - (\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2\Delta_1\|_F \le (1-2\tilde{c})\|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}\|_F.$$

For the second term, applying Conditions 3.1-3.2 as well as the definition of Δ_2, we can obtain

$$\begin{aligned}
&\|(\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2\Delta_2\|_F \\
&\le 2\times\frac{1}{4}\|(\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot A^*-\hat{\mathcal{A}}}(\hat{\boldsymbol{C}}_{A^*-\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{A^*-\hat{\mathcal{A}}\cdot})\|_F \\
&\le \frac{\theta_s}{2c_-(s)}\|\hat{\boldsymbol{C}}_{A^*-\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{A^*-\hat{\mathcal{A}}\cdot}\|_F.
\end{aligned}$$

Therefore, the inequality (2) becomes,

$$\begin{aligned}
\|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}\| &\le (1-2\tilde{c})\|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}\|_F + \frac{\theta_s}{2c_-(s)}\|\hat{\boldsymbol{C}}_{A^*-\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{A^*-\hat{\mathcal{A}}\cdot}\|_F + \|(\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})\|_F. \\
&\le (1-2\tilde{c}+\bar{c}\frac{\theta_s}{2c_-(s)})\|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}\|_F + +\|(\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})\|_F,
\end{aligned}$$

where the last inequality holds due to the fact that $\hat{\boldsymbol{C}}$ lies in the cone set $\Lambda(s)$. Based on the above result, we can obtain the upper bound of $\|\hat{\boldsymbol{C}} - \boldsymbol{C}^*\|_F^2$ as follows.

$$\begin{aligned}
\|\hat{\boldsymbol{C}} - \boldsymbol{C}^*\|_F^2 &= \|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}\|_F^2 + \|\hat{\boldsymbol{C}}_{A^*-\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{A^*-\hat{\mathcal{A}}\cdot}\|_F^2 \\
&\le (1-2\tilde{c}+\bar{c}\frac{\theta_s}{2c_-(s)})^2\|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}\|_F^2 + \|\hat{\boldsymbol{C}}_{A^*-\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{A^*-\hat{\mathcal{A}}\cdot}\|_F^2 \\
&+ (1-2\tilde{c}+\bar{c}\frac{\theta_s}{2c_-(s)})\|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}\|_F\|(\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})\|_F \\
&+ \|(\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})\|_F^2 \\
&\le (1-2\tilde{c}+\bar{c}\frac{\theta_s}{2c_-(s)})^2\|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}\|_F^2 + \|\hat{\boldsymbol{C}}_{A^*-\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{A^*-\hat{\mathcal{A}}\cdot}\|_F^2 \\
&+ (1-2\tilde{c}+\bar{c}\frac{\theta_s}{2c_-(s)})\frac{\bar{c}}{\sqrt{1+\bar{c}^2}}\|\hat{\boldsymbol{C}} - \boldsymbol{C}^*\|_F\|(\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})\|_F \\
&+ \|(\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})\|_F^2
\end{aligned}$$

Since $1-2\tilde{c}+\bar{c}\frac{\theta_s}{2c_-(s)} < 1$, we can obtain

$$\begin{aligned}
&(1-2\tilde{c}+\bar{c}\frac{\theta_s}{2c_-(s)})^2\|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}\|_F^2 + \|\hat{\boldsymbol{C}}_{A^*-\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{A^*-\hat{\mathcal{A}}\cdot}\|_F^2 \\
&< \|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}\|_F^2 + \|\hat{\boldsymbol{C}}_{A^*-\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{A^*-\hat{\mathcal{A}}\cdot}\|_F^2 \\
&= \|\hat{\boldsymbol{C}} - \boldsymbol{C}^*\|_F^2.
\end{aligned}$$

Let $\alpha_d = \frac{(1-2\tilde{c}+\bar{c}\frac{\theta_s}{2c_-(s)})^2\|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}\|_F^2 + \|\hat{\boldsymbol{C}}_{A^*-\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{A^*-\hat{\mathcal{A}}\cdot}\|_F^2}{\|\hat{\boldsymbol{C}} - \boldsymbol{C}^*\|_F^2} < 1$. After some simplifications, we get

$$(\alpha_d - (1-2\tilde{c}+\bar{c}\frac{\theta_s}{2c_-(s)})^2)\|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}\|_F = (1-\alpha_d)\|\hat{\boldsymbol{C}}_{A^*-\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{A^*-\hat{\mathcal{A}}\cdot}\|_F \le \bar{c}(1-\alpha_d)\|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}\|_F.$$

Canceling the same terms $\|\hat{\boldsymbol{C}}_{\hat{\mathcal{A}}\cdot} - \boldsymbol{C}^*_{\hat{\mathcal{A}}\cdot}\|_F$, we get

$$\alpha_d \le \frac{\bar{c} + (1-2\tilde{c}+\bar{c}\frac{\theta_s}{2c_-(s)})^2}{\bar{c}+1} < 1.$$

Without loss of generality, we set

$$\alpha_d = \frac{\bar{c} + (1-2\tilde{c}+\bar{c}\frac{\theta_s}{2c_-(s)})^2}{\bar{c}+1}.$$

Therefore,

$$\|\hat{\boldsymbol{C}}-\boldsymbol{C}^*\|_F^2 \le \alpha_d\|\hat{\boldsymbol{C}}-\boldsymbol{C}^*\|_F^2 + (1-2\tilde{c}+\bar{c}\frac{\theta_s}{2c_-(s)})\frac{\bar{c}}{\sqrt{1+\bar{c}^2}}\|\hat{\boldsymbol{C}}-\boldsymbol{C}^*\|_F\|(\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top\boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})\|_F$$
$$+\|(\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top\boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})\|_F^2.$$

Solving the above quadratic inequality in $\|\hat{\boldsymbol{C}}-\boldsymbol{C}^*\|_F$ yields

$$\|\hat{\boldsymbol{C}}-\boldsymbol{C}^*\|_F \le \frac{-(1-2\tilde{c}+\bar{c}\frac{\theta_s}{2c_-(s)})+\sqrt{(1-2\tilde{c}+\bar{c}\frac{\theta_s}{2c_-(s)})^2+4(1-\alpha_d)}}{2(1-\alpha_d)}\left\|(\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top\boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})\right\|_F,$$
$$\le \frac{-(1-2\tilde{c}+\bar{c}\frac{\theta_s}{2c_-(s)})+\sqrt{(1-2\tilde{c}+\bar{c}\frac{\theta_s}{2c_-(s)})^2+4(1-\alpha_d)}}{2(1-\alpha_d)}\sqrt{r}\left\|(\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top\boldsymbol{X}_{\cdot\hat{\mathcal{A}}})^{-1}\boldsymbol{X}_{\cdot\hat{\mathcal{A}}}^\top 2(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})\right\|_{op},$$

Then applying Condition 3.1 and Lemma 5, with probability at least $1-\delta$ for any $\delta>0$, we have

$$\|\hat{\boldsymbol{C}}-\boldsymbol{C}^*\|_F \le \frac{-(1-2\tilde{c}+\bar{c}\frac{\theta_s}{2c_-(s)})+\sqrt{(1-2\tilde{c}+\bar{c}\frac{\theta_s}{2c_-(s)})^2+4(1-\alpha_d)}}{(1-\alpha_d)c_-(s)}\sqrt{\frac{3rqsP^2}{n}\log\frac{2pq}{\delta}}.$$

This completes the proof.

A.3. Proof of Theorem 2

Recall that $\boldsymbol{C}^*$ is the true coefficient matrix, and $\boldsymbol{C}^{m,k}$ is the estimate at the k-th iteration of the inner loop of the m-th step of the outer loop in **Algorithm 2**. To prove Theorem 2, we first define several immediate variables. Let $\hat{\boldsymbol{C}}^0=\boldsymbol{C}^0$ denote the initialized value of $\boldsymbol{C}$ in **Algorithm 2**. Then at the m-th step of **Algorithm 2** ($m\ge 1$), define

$$\hat{\boldsymbol{C}}^m = \arg\min_{\boldsymbol{C}\in\Omega_r} S(\boldsymbol{C};\mathcal{D}|\hat{\boldsymbol{C}}^{m-1}).$$

Denote the minimizer of (4) as $\hat{\boldsymbol{C}}$, then by the definition of surrogate function, we have

$$\hat{\boldsymbol{C}} = \arg\min_{\boldsymbol{C}\in\Omega_r} S(\boldsymbol{C};\mathcal{D}|\hat{\boldsymbol{C}}).$$

Next we define the noise-free version of the surrogate function. For any coefficient matrix $\boldsymbol{C}'$ in Ω_r, define the noise-free surrogate function as

$$\tilde{S}(\boldsymbol{C};\mathcal{D}|\boldsymbol{C}') = L(\boldsymbol{C}';\mathcal{D}) + \frac{1}{4n}\|\boldsymbol{X}(\boldsymbol{C}-\boldsymbol{C}') - 2(\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*}-\tilde{\boldsymbol{P}}^{\boldsymbol{C}'})\|_F^2 - \frac{1}{n}\|\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*}-\tilde{\boldsymbol{P}}^{\boldsymbol{C}'}\|_F. \tag{3}$$

It is easy to show that the true coefficient matrix $\boldsymbol{C}^*$ is actually the minimizer of the noise-free surrogate function around $\boldsymbol{C}^*$, that is,

$$\boldsymbol{C}^* = \arg\min_{\boldsymbol{C}\in\Omega_r} \tilde{S}(\boldsymbol{C};\mathcal{D}|\boldsymbol{C}^*).$$

Similarly, we can define the "true" coefficient matrix $\boldsymbol{C}^{m*}$ at the m-th step of **Algorithm 2**,

$$\boldsymbol{C}^{m*} = \arg\min_{\boldsymbol{C}\in\Omega_r} \tilde{S}(\boldsymbol{C};\mathcal{D}|\hat{\boldsymbol{C}}^{m-1}).$$

The proof of Theorem 2 is based on three following lemmas.

LEMMA 1. *Assume that Conditions 3.1-3.2 hold. Then for the output estimate $\boldsymbol{C}^{m,k}$ at the k-th inner loop iteration of the m-th step of **Algorithm 2** with $s\ge s^*$, $r\ge r^*$, we have*

$$\begin{aligned}\|\boldsymbol{C}^{m,k}-\boldsymbol{C}^{m*}\|_{op} &\le \alpha_1\gamma^k\|\boldsymbol{C}^{m*}\|_{op} + (\alpha_2+\alpha_3\sqrt{r})h(s),\\ \|\boldsymbol{C}^{m,k}-\boldsymbol{C}^{m*}\|_F &\le \alpha_1\sqrt{2r}\gamma^k\|\boldsymbol{C}^{m*}\|_{op} + \sqrt{2r}(\alpha_2+\alpha_3\sqrt{r})h(s),\end{aligned} \tag{4}$$

where

$$h(s) = \frac{1}{n}\max_{|\mathcal{A}|\le s}\|\boldsymbol{X}_{\cdot\mathcal{A}}^{\top} Err^m\|_{op},$$
$$Err^m = 2(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}^{m-1}}) + \boldsymbol{X}(\hat{\boldsymbol{C}}^{m-1} - \boldsymbol{C}^{m*}).$$

In addition, with probability at least $1-\delta$ *for any constant* $\alpha > 0$ *and sufficiently large* n*, we have*

$$h(s) \le 5\frac{c_+(s)^2}{c_-(s)}(1-2\zeta(1-\zeta))^{m-1}(\|\hat{\boldsymbol{C}}^0 - \hat{\boldsymbol{C}}\|_F + (1-2\zeta(1-\zeta))\|\hat{\boldsymbol{C}}^0 - \boldsymbol{C}^*\|_F) + 5c_+(s)\|\boldsymbol{C}^* - \hat{\boldsymbol{C}}\|_F + 2\sqrt{\frac{3qsP^2}{n}\log\frac{2pq}{\delta}}.$$

LEMMA 2. *Assume conditions 3.1-3.2 hold, we have*

$$\|\hat{\boldsymbol{C}}^m - \hat{\boldsymbol{C}}\|_F \le \big(1-2\zeta(1-\zeta)\big)^m \frac{c_+(s)}{c_-(s)}\|\hat{\boldsymbol{C}}^0 - \hat{\boldsymbol{C}}\|_F,$$

where

$$\zeta = \min_{\boldsymbol{C}\in B_F(r_0,\boldsymbol{C}^*),\boldsymbol{C}\in\Omega_r,i,j} \tilde{\boldsymbol{P}}^{\boldsymbol{C}}_{i,j}.$$

is a constant lying between 0 and 1.

LEMMA 3. *Assume conditions 3.1-3.2 hold, we have*

$$\|\boldsymbol{C}^{m*} - \boldsymbol{C}^*\|_F \le \big(1-2\zeta(1-\zeta)\big)^m \frac{c_+(s)}{c_-(s)}\|\hat{\boldsymbol{C}}^0 - \boldsymbol{C}^*\|_F + \big(1-2\zeta(1-\zeta)\big)\|\hat{\boldsymbol{C}} - \boldsymbol{C}^*\|_F,$$

where ζ *is defined as Lemma 2.*

It follows from Lemma 1 and the triangle inequality that

$$\begin{aligned}\|\boldsymbol{C}^{m,k} - \boldsymbol{C}^*\|_F &\le \|\boldsymbol{C}^{m,k} - \boldsymbol{C}^{m*}\|_F + \|\boldsymbol{C}^{m*} - \boldsymbol{C}^*\|_F \\ &\le \alpha_1\sqrt{2r}\gamma^k\|\boldsymbol{C}^{m*}\|_{op} + \sqrt{2r}(\alpha_2 + \alpha_3\sqrt{r})h(s) + \|\boldsymbol{C}^{m*} - \boldsymbol{C}^*\|_F \\ &\le \alpha_1\sqrt{2r}\gamma^k\|\boldsymbol{C}^*\|_{op} + \sqrt{2r}(\alpha_2 + \alpha_3\sqrt{r})h(s) + (\alpha_1\sqrt{2r}\gamma^{k+1} + 1)\|\boldsymbol{C}^{m*} - \boldsymbol{C}^*\|_F.\end{aligned}$$

Now, we consider bounding the last two terms in the right hand size of the above inequality. In specific, we bound the second stochastic term by Lemma 1, and give an upper bound for the third term by Lemma 3. Thus, with probability at least $1-\delta$ for any constant $\alpha > 0$ and sufficiently large n, we have

$$\begin{aligned}\|\boldsymbol{C}^{m,k} - \boldsymbol{C}^*\|_F \le{}& \alpha_1\sqrt{2r}\gamma^{k+1}\|\boldsymbol{C}^*\|_{op} + 5\sqrt{2r}(\alpha_2 + \alpha_3\sqrt{r})\frac{c_+(s)^2}{c_-(s)}(1-2\zeta(1-\zeta))^{m-1}\|\hat{\boldsymbol{C}}^0 - \hat{\boldsymbol{C}}\|_F \\ &+ \left(5\sqrt{2r}(\alpha_2 + \alpha_3\sqrt{r})c_+(s) + \alpha_1\sqrt{2r}\gamma^k + 1\right)\frac{c_+(s)}{c_-(s)}(1-2\zeta(1-\zeta))^m\|\hat{\boldsymbol{C}}^0 - \boldsymbol{C}^*\|_F \qquad (5) \\ &+ 5\sqrt{2r}(\alpha_2 + \alpha_3\sqrt{r})c_+(s)\|\boldsymbol{C}^* - \hat{\boldsymbol{C}}\|_F + 2\sqrt{2r}(\alpha_2 + \alpha_3\sqrt{r})\sqrt{\frac{3qsP^2}{n}\log\frac{2pq}{\delta}}.\end{aligned}$$

For the term $\|\hat{\boldsymbol{C}}^0 - \hat{\boldsymbol{C}}\|_F$, applying the triangle inequality and Theorem 1 gives the following result,

$$\begin{aligned}\|\hat{\boldsymbol{C}}^0 - \hat{\boldsymbol{C}}\|_F &\le \|\hat{\boldsymbol{C}}^0 - \boldsymbol{C}^*\|_F + \|\hat{\boldsymbol{C}} - \boldsymbol{C}^*\|_F \\ &\le \|\hat{\boldsymbol{C}}^0 - \boldsymbol{C}^*\|_F + \frac{-(1-2\tilde{c}+\bar{c}\frac{\theta_s}{2c_-(s)}) + \sqrt{(1-2\tilde{c}+\bar{c}\frac{\theta_s}{2c_-(s)})^2 + 4(1-\alpha_d)}}{(1-\alpha_d)c_-(s)}\sqrt{\frac{3qsP^2}{n}\log\left(\frac{2pq}{\delta}\right)}\end{aligned}$$

with probability at least $1-\delta$ for any constant $\alpha>0$ and sufficiently large n. Substituting the bound into (5), we have

$$\begin{aligned}
\|\boldsymbol{C}^{m,k}-\boldsymbol{C}^*\|_F &\le \alpha_1\sqrt{2r}\gamma^{k+1}\|\boldsymbol{C}^*\|_{op}+\Bigg[5\sqrt{2}(\frac{\alpha_2}{\sqrt{r}}+\alpha_3)\frac{c_+(s)^2}{c_-(s)}(1-2\zeta(1-\zeta))^{m-1}\\
&+\left(5\sqrt{2}(\frac{\alpha_2}{\sqrt{r}}+\alpha_3)c_+(s)+\alpha_1\frac{\sqrt{2}}{\sqrt{r}}\gamma^k+\frac{1}{r}\right)\frac{c_+(s)}{c_-(s)}(1-2\zeta(1-\zeta))^m\Bigg]r\|\hat{\boldsymbol{C}}^0-\boldsymbol{C}^*\|_F\\
&+\left(2+(\frac{c_+(s)^2}{c_-(s)}(1-2\zeta(1-\zeta))^{m-1}+c_+(s))5\alpha\right)\sqrt{2}(\frac{\alpha_2}{\sqrt{r}}+\alpha_3)r\sqrt{\frac{3qsP^2}{n}\log\frac{2pq}{\delta}}\\
&\le \alpha_1\sqrt{2r}\gamma^{k+1}\|\boldsymbol{C}^*\|_{op}+\Bigg[5\sqrt{2}(\alpha_2+\alpha_3)\frac{c_+(s)^2}{c_-(s)}(1-2\zeta(1-\zeta))^{m-1}\\
&+\left(5\sqrt{2}(\alpha_2+\alpha_3)c_+(s)+\alpha_1\sqrt{2}+1\right)\frac{c_+(s)}{c_-(s)}(1-2\zeta(1-\zeta))^m\Bigg]r\|\hat{\boldsymbol{C}}^0-\boldsymbol{C}^*\|_F\\
&+\left(2+(\frac{c_+(s)^2}{c_-(s)}(1-2\zeta(1-\zeta))^{m-1}+c_+(s))5\alpha\right)\sqrt{2}(\alpha_2+\alpha_3)r\sqrt{\frac{3qsP^2}{n}\log\frac{2pq}{\delta}}.
\end{aligned}$$

Taking $c_2=\sqrt{2}\alpha_1$, $c_3=5\sqrt{2}(\alpha_2+\alpha_3)\frac{c_+(s)^2}{c_-(s)}(1-2\zeta(1-\zeta))^{m-1}+\left(5\sqrt{2}(\alpha_2+\alpha_3)c_+(s)+\alpha_1\sqrt{2}+1\right)\frac{c_+(s)}{c_-(s)}(1-2\zeta(1-\zeta))^m$, and leaving $c_4=\left(2+(\frac{c_+(s)^2}{c_-(s)}(1-2\zeta(1-\zeta))^{m-1}+c_+(s))5\alpha\right)\sqrt{2}(\alpha_2+\alpha_3)\sqrt{3\log(2)P^2}$ gives Theorem 2.

A.4. Proof of Corollary 1

The first part of Corollary 1 follows immediately from Theorem 2. In particular, when $k\ge\log_\gamma\frac{c_4\sqrt{r}}{2c_2\|\boldsymbol{C}^*\|_F}\sqrt{\frac{qs\log pq}{n}}$, we have

$$\gamma^k\le\frac{c_4\sqrt{r}}{2c_2\|\boldsymbol{C}^*\|_F}\sqrt{\frac{qs\log pq}{n}},$$

which is equal to

$$\gamma^k c_2\|\boldsymbol{C}^*\|_F\le\frac{c_4\sqrt{r}}{2}\sqrt{\frac{qs\log pq}{n}}.$$

Similarly, when $m\ge\log_{\zeta'}\frac{c_4 r}{2c_3\|\boldsymbol{C}^0-\boldsymbol{C}^*\|_F}\sqrt{\frac{qs\log pq}{n}}$, we have

$$(\zeta')^m\le\frac{c_4}{2c_3\|\boldsymbol{C}^0-\boldsymbol{C}^*\|_F}\sqrt{\frac{qs\log pq}{n}}.$$

which is equal to

$$(1-2\zeta(1-\zeta))^m c_3 r\|\boldsymbol{C}^0-\boldsymbol{C}^*\|_F\le\frac{c_4 r}{2}\sqrt{\frac{qs\log pq}{n}}.$$

Next, combining the inequality $\min_{i\in\mathrm{supp}(\boldsymbol{C}^*)}\|\boldsymbol{C}^*_{i\cdot}\|_2\ge 2c_4 r\sqrt{\frac{qs\log(pq)}{n}}$, we have

$$\mathrm{supp}(\boldsymbol{C}^*)\subseteq\mathrm{supp}(\boldsymbol{C}^{m,k}).$$

If not, we have at least one true nonzero being missed in the implementation. Without loss of generality, we assume that the i-th row is screened out, i.e., $i\in\mathrm{supp}(\boldsymbol{C}^*)\setminus\mathrm{supp}(\boldsymbol{C}^{m,k})$. Then

$$\|\boldsymbol{C}^*-\boldsymbol{C}^{m,k}\|_F\ge\|\boldsymbol{C}^*_{i\cdot}\|_2\ge 2c_4 r\sqrt{\frac{qs\log(pq)}{n}},$$

which leads to a contradiction to the inequality in the first part.

Owing to the same size of $\mathrm{supp}(\boldsymbol{C}^*)$ and $\mathrm{supp}(\boldsymbol{C}^{m,k})$, we conclude that

$$\mathrm{supp}(\boldsymbol{C}^*)=\mathrm{supp}(\boldsymbol{C}^{m,k}).$$

Appendix B: Proofs of lemmas

B.1. Proof of Lemma 1

The proof of Lemma 1 follows similar lines of the proof of Theorem 3.1 of Wen et al. (2022), with some minor modifications. Thus, we introduce the main difference and present an outline of the proof for the first part of Lemma 1. The main difference is the model setting: while a reduced rank regression model is considered in Wen et al. (2022), our method is built on the reduced rank multinomial logistic regression model. This difference leads to the difference in error term. In particular, the error term Err^m here is defined as

$$Err^m = 2(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}^{m-1}}) + \boldsymbol{X}(\hat{\boldsymbol{C}}^{m-1} - \boldsymbol{C}^{m*}).$$

In contrast, the error term in Wen et al. (2022) is $\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{C}^*$. The involvement of $\hat{\boldsymbol{C}}^{m-1}$ in the definition of error leads a much more complicated expression.

Before presenting the proof, we introduce some notations and lemmas needed. Let the row support set of $\boldsymbol{C}^{m,k}$ be $\mathcal{A}^{m,k}$, and the set of zero rows be $\mathcal{I}^{m,k} = (\mathcal{A}^{m,k})^c$. We define the auxiliary index sets by

$$\mathcal{A}_{11}^{m,k} = \mathcal{A}^* \cap \mathcal{A}^{m,k} \cap \mathcal{I}^{m,k+1}, \quad \mathcal{A}_{22}^{m,k} = \mathcal{A}^* \cap \mathcal{I}^{m,k} \cap \mathcal{I}^{m,k+1},$$
$$\mathcal{I}_{11}^{m,k} = \mathcal{I}^* \cap \mathcal{A}^{m,k} \cap \mathcal{A}^{m,k+1}, \quad \mathcal{I}_{22}^{m,k} = \mathcal{I}^* \cap \mathcal{I}^{m,k} \cap \mathcal{A}^{m,k+1}.$$

We can see that $\mathcal{A}_{11}^{m,k}$ is the subset of the true nonzero rows that are kicked out at the $(m,k+1)$-th iteration and selected at (m,k)-th iteration at the same time, and $\mathcal{A}_{11}^{m,k}$ is the subset of the true nonzero rows that are kicked out at both the $(m,k+1)$-th iteration and the (m,k)-th iteration. Similarly, $\mathcal{A}_{11}^{m,k}$ is the subset of the true nonzero rows that are kicked out at the $(m,k+1)$-th iteration and selected at (m,k)-th iteration at the same time, and $\mathcal{A}_{11}^{m,k}$ is the subset of the true nonzero rows that are kicked out at both the $(m,k+1)$-th iteration and the (m,k)-th iteration. In fact, they are generalization of false positive and false negative for the switch from the (m,k)-th to the $(m,k+1)$-th iteration. We further define the difference in $\boldsymbol{C}^*$ as $D(\mathcal{A}^k) = \|\boldsymbol{C}^*_{\mathcal{A}^* \cap \mathcal{I}^{m,k}}\|_{op}$. For convenience, we introduce some more definition as follows.

$$\Delta(\boldsymbol{C}^{m,k+1}) = \boldsymbol{C}^{m,k+1}_{A^{m,k}_{\cdot}} - \boldsymbol{C}^*_{A^{m,k}_{\cdot}}, \quad \tilde{\Gamma}^{k+1}_{\mathcal{A}_{22}^{m,k}} = \frac{1}{n} X^\top_{\cdot \mathcal{A}_{22}^{m,k}} (2(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}^{m-1}}) + \boldsymbol{X}\hat{\boldsymbol{C}}^{m-1} - \boldsymbol{X}\boldsymbol{C}^{m,k+1}).$$

An outline of the proof is given below.

1. Construct inequality to bound $\|\boldsymbol{C}^{m,k} - \boldsymbol{C}^{m*}\|_{op}$ by $D(\mathcal{A}^k)$ as follows.

$$\|\boldsymbol{C}^{m,k+1}_{\mathcal{A}^{m,k}_{\cdot}} - \boldsymbol{C}^{m*}_{\mathcal{A}^{m,k}_{\cdot}}\|_{op} \le \frac{\theta_s}{c_-(s)} D(\mathcal{A}^k) + \frac{2}{nc_-(s)} \|\boldsymbol{X}_{\cdot \mathcal{A}^{m,k}} Err^m\|_{op},$$
$$\|\boldsymbol{C}^{m,k+1} - \boldsymbol{C}^{m*}\|_{op} \le (1 + \frac{\theta_s}{c_-(s)}) D(\mathcal{A}^k) + \frac{2}{nc_-(s)} \|\boldsymbol{X}_{\cdot \mathcal{A}^{m,k}} Err^m\|_{op}.$$

   This is the same with Lemma B.1 in Wen et al. (2022) except the definition of Err^m.

2. Develop the iterative inequality for $D(\mathcal{A}^{k+1})$,

$$D(\mathcal{A}^{k+1}) \le \gamma D(\mathcal{A}^k) + \alpha h(s).$$

   To get the above result, we first need to prove three following results. The first is Lemma B.2 of Wen et al. (2022), which are listed below.

$$D(A^{k+1}) \le \|\boldsymbol{C}^{m*}_{A^k_{11}}\|_{op} + \|\boldsymbol{C}^{m*}_{\mathcal{A}^{m,k}_{22}}\|_{op},$$
$$\|\boldsymbol{C}^{m*}_{\mathcal{A}^{m,k}_{11}}\|_{op} \le \|\Delta(\boldsymbol{C}^{m,k+1})_{\mathcal{A}^{m,k}_{11}}\|_{op} + \|\boldsymbol{C}^{m,k+1}_{\mathcal{A}^{m,k}_{11}}\|_{op},$$
$$\|\boldsymbol{C}^{m*}_{\mathcal{A}^{m,k}_{22}}\|_{op} \le \frac{1}{c_-(s)} (\|\tilde{\Gamma}_{\mathcal{A}^{m,k}_{22}}\|_{op} + \theta_s \|\Delta(\boldsymbol{C}^{m,k+1})\|_{op} + \theta_s D(\mathcal{A}^k) + h(s)).$$

The second is Lemma B.3 of Wen et al. (2022) stated as

$$\|\boldsymbol{B}^{k+1}_{\mathcal{A}^{m,k}_{11}\cdot}\|_{op}+\|\Gamma^{k+1}_{\mathcal{A}^{m,k}_{22}\cdot}\|_{op}\leq\sqrt{r}(\|\boldsymbol{C}^{k+1}_{\mathcal{I}^{m,k}_{11}\cdot}\|_{op}+\|\tilde{\Gamma}^{k+1}_{\mathcal{I}^{m,k}_{22}\cdot}\|_{op}).$$

The last one is Lemma 4, which is a modified version of Lemma B.4 of Wen et al. (2022). The proof of Lemma 4 is presented in Appendix B.4.

LEMMA 4. *Suppose Condition 3.1 and 3.2 hold, we have*

$$\|\tilde{\Gamma}^{k+1}_{\mathcal{A}^{m,k}_{22}\cdot}\|_{op}\leq\|\Gamma^{k+1}_{\mathcal{A}^{m,k}_{22}\cdot}\|_{op}+\frac{\alpha_0(1+c_-(s)c_+(s))+1}{n}\|\boldsymbol{X}^{\top}_{\cdot A^k}Err^m\|_{op},\tag{6}$$

where $\alpha_0>0$ is some constant.

3. Recursively apply the above iterative inequality to derive

$$D(\mathcal{A}^{k+1})\leq\gamma^{k+1}\|\boldsymbol{C}^{m*}\|_{op}+\alpha h(s).$$

4. Combining the results in Step 1 and Step 3, we have the results in Lemma 1.

Next, we prove the results regarding of $h(s)$. By definition, for any $\mathcal{A}$ satisfying $|\mathcal{A}|\leq s$,

$$\begin{aligned}\frac{1}{n}\|\boldsymbol{X}^{\top}_{\cdot\mathcal{A}}Err^m\|_{op}&=\frac{1}{n}\|\boldsymbol{X}^{\top}_{\cdot\mathcal{A}}\left(2(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}^{m-1}})+\boldsymbol{X}(\hat{\boldsymbol{C}}^{m-1}-\boldsymbol{C}^{m*})\right)\|_{op}\\&=\frac{1}{n}\|\boldsymbol{X}^{\top}_{\cdot\mathcal{A}}\left(2(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})+2(\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*}-\tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}^{m-1}})+\boldsymbol{X}(\hat{\boldsymbol{C}}^{m-1}-\boldsymbol{C}^{m*})\right)\|_{op}\\&\leq\frac{1}{n}\|2\boldsymbol{X}^{\top}_{\cdot\mathcal{A}}(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})\|_{op}+\frac{1}{n}\|2\boldsymbol{X}^{\top}_{\cdot\mathcal{A}}(\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*}-\tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}^{m-1}})+\boldsymbol{X}^{\top}_{\cdot\mathcal{A}}\boldsymbol{X}(\hat{\boldsymbol{C}}^{m-1}-\boldsymbol{C}^{m*})\|_{op}.\end{aligned}\tag{7}$$

where the last inequality holds by applying the triangle inequality. From Lemma 5, the first term of the right hand size of (7) is bounded above as

$$\max_{\mathcal{A}:|\mathcal{A}|\leq s}\frac{1}{n}\|(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})^{\top}\boldsymbol{X}_{\cdot\mathcal{A}}\|_{op}\leq\sqrt{\frac{3qsP^2}{n}\log\frac{2pq}{\delta}}.$$

with probability at least $1-\delta$.

As for the second term of (7), using the fact that $\|A\|_{op}\leq\|A\|_F$ and the triangle inequality gives

$$\begin{aligned}&\frac{1}{n}\|2\boldsymbol{X}^{\top}_{\cdot\mathcal{A}}(\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*}-\tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}^{m-1}})+\boldsymbol{X}^{\top}_{\cdot\mathcal{A}}\boldsymbol{X}(\hat{\boldsymbol{C}}^{m-1}-\boldsymbol{C}^{m*})\|_{op}\\\leq&\frac{1}{n}\|2\boldsymbol{X}^{\top}_{\cdot\mathcal{A}}(\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*}-\tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}^{m-1}})+\boldsymbol{X}^{\top}_{\cdot\mathcal{A}}\boldsymbol{X}(\hat{\boldsymbol{C}}^{m-1}-\boldsymbol{C}^{m*})\|_{F}\\\leq&\frac{1}{n}\|2\boldsymbol{X}^{\top}_{\cdot\mathcal{A}}(\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*}-\tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}^{m-1}})\|_{F}+\frac{1}{n}\|\boldsymbol{X}^{\top}_{\cdot\mathcal{A}}\boldsymbol{X}(\hat{\boldsymbol{C}}^{m-1}-\boldsymbol{C}^{m*})\|_{F}\end{aligned}\tag{8}$$

To bound the term $\|2\boldsymbol{X}^{\top}_{\cdot\mathcal{A}}(\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*}-\tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}^{m-1}})\|_F/n$, we bound the absolute value of each element in terms of $\boldsymbol{XC}$. In particular, for any $\boldsymbol{C}$ and other elements fixed, the (i,k)-th element of $\tilde{\boldsymbol{P}}^{\boldsymbol{C}}$ is actually a function of $\boldsymbol{c}_k^{\top}\boldsymbol{x}_i$. Let $f(\cdot)$ denote this function, i.e., $f(x)=\frac{\exp(x)}{\alpha+\exp(x)}$, where α is a constant. With simple calculation, we know that $f(\cdot)$ is a contraction function with the Lipschitz constant less than 1. Hence,

$$\begin{aligned}&\frac{1}{n}\|2\boldsymbol{X}^{\top}_{\cdot\mathcal{A}}(\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*}-\tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}^{m-1}})\|_{F}+\frac{1}{n}\|\boldsymbol{X}^{\top}_{\cdot\mathcal{A}}\boldsymbol{X}(\hat{\boldsymbol{C}}^{m-1}-\boldsymbol{C}^{m*})\|_{F}\\\leq&\frac{2}{n}\|\boldsymbol{X}^{\top}_{\cdot\mathcal{A}}\boldsymbol{X}(\boldsymbol{C}^{*}-\hat{\boldsymbol{C}}^{m-1})\|_{F}+\frac{1}{n}\|\boldsymbol{X}^{\top}_{\cdot\mathcal{A}}\boldsymbol{X}(\hat{\boldsymbol{C}}^{m-1}-\boldsymbol{C}^{m*})\|_{F}.\end{aligned}\tag{9}$$

The two terms of the right hand size of (9) are similar. Thus, we only present the proof for the upper bound of the first term, and that of the second term is omitted. Denote the row support set of $\boldsymbol{C}^*$ and $\hat{\boldsymbol{C}}^{m-1}$ as $\mathcal{A}^*$ and $\hat{\mathcal{A}}^{m-1}$, respectively. Applying the triangle inequality again gives

$$\begin{aligned}
&\frac{2}{n}\|\boldsymbol{X}_{\cdot\mathcal{A}}^\top \boldsymbol{X}(\boldsymbol{C}^* - \hat{\boldsymbol{C}}^{m-1})\|_F \\
=& \frac{2}{n}\|\boldsymbol{X}_{\cdot\mathcal{A}}^\top \boldsymbol{X}_{\cdot\mathcal{A}^*\cup\hat{\mathcal{A}}^{m-1}}(\boldsymbol{C}^*_{\mathcal{A}^*\cup\hat{\mathcal{A}}^{m-1}\cdot} - \hat{\boldsymbol{C}}^{m-1}_{\mathcal{A}^*\cup\hat{\mathcal{A}}^{m-1}\cdot})\|_F \\
=& \frac{2}{n}\left\|\boldsymbol{X}_{\cdot\mathcal{A}}^\top \left(\boldsymbol{X}_{\cdot\mathcal{A}^*}(\boldsymbol{C}^*_{\mathcal{A}^*\cdot} - \hat{\boldsymbol{C}}^{m-1}_{\mathcal{A}^*\cdot}) + \boldsymbol{X}_{\cdot\hat{\mathcal{A}}^{m-1}\setminus\mathcal{A}^*}(\boldsymbol{C}^*_{\hat{\mathcal{A}}^{m-1}\setminus\mathcal{A}^*\cdot} - \hat{\boldsymbol{C}}^{m-1}_{\hat{\mathcal{A}}^{m-1}\setminus\mathcal{A}^*\cdot})\right)\right\|_F \\
\le& \frac{2}{n}\|\boldsymbol{X}_{\cdot\mathcal{A}}^\top \boldsymbol{X}_{\cdot A^*}(\boldsymbol{C}^*_{\mathcal{A}^*\cdot} - \hat{\boldsymbol{C}}^{m-1}_{\mathcal{A}^*\cdot})\|_F + \frac{2}{n}\|\boldsymbol{X}_{\cdot\mathcal{A}}^\top \boldsymbol{X}_{\cdot\hat{\mathcal{A}}^{m-1}\setminus\mathcal{A}^*}(\boldsymbol{C}^*_{\hat{\mathcal{A}}^{m-1}\setminus\mathcal{A}^*\cdot} - \hat{\boldsymbol{C}}^{m-1}_{\hat{\mathcal{A}}^{m-1}\setminus\mathcal{A}^*\cdot})\|_F \\
\le& \frac{2}{n}\|\boldsymbol{X}_{\cdot\mathcal{A}}^\top\|_F\|\boldsymbol{X}_{\cdot\mathcal{A}^*}\|_F\|\boldsymbol{C}^*_{\mathcal{A}^*\cdot} - \hat{\boldsymbol{C}}^{m-1}_{\mathcal{A}^*\cdot}\|_F + \frac{2}{n}\|\boldsymbol{X}_{\cdot\mathcal{A}}^\top\|_F\|\boldsymbol{X}_{\cdot\hat{\mathcal{A}}^{m-1}\setminus\mathcal{A}^*}\|_F\|\boldsymbol{C}^*_{\hat{\mathcal{A}}^{m-1}\setminus\mathcal{A}^*\cdot} - \hat{\boldsymbol{C}}^{m-1}_{\hat{\mathcal{A}}^{m-1}\setminus\mathcal{A}^*\cdot}\|_F \\
\le& 4c_+(s)\|\boldsymbol{C}^* - \hat{\boldsymbol{C}}^{m-1}\|_F,
\end{aligned}$$

where the last equality holds by using Condition 3.1. Similar procedure can be forwarded to show that

$$\frac{1}{n}\|\boldsymbol{X}_{\cdot\mathcal{A}}^\top \boldsymbol{X}(\hat{\boldsymbol{C}}^{m-1} - \boldsymbol{C}^{m*})\|_F \le c_+(s)\|\boldsymbol{C}^{m*} - \hat{\boldsymbol{C}}^{m-1}\|_F.$$

Substituting the above inequalities into (8) and (9), we can obtain

$$\begin{aligned}
&\frac{1}{n}\|2\boldsymbol{X}_{\cdot\mathcal{A}}^\top(\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*} - \tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}^{m-1}}) + \boldsymbol{X}_{\cdot\mathcal{A}}^\top \boldsymbol{X}(\hat{\boldsymbol{C}}^{m-1} - \boldsymbol{C}^{m*})\|_{op} \\
\le& 4c_+(s)\|\boldsymbol{C}^* - \hat{\boldsymbol{C}}^{m-1}\|_F + c_+(s)\|\boldsymbol{C}^{m*} - \hat{\boldsymbol{C}}^{m-1}\|_F \\
\le& 5c_+(s)\|\boldsymbol{C}^* - \hat{\boldsymbol{C}}\|_F + 5c_+(s)\|\hat{\boldsymbol{C}}^{m-1} - \hat{\boldsymbol{C}}\|_F + c_+(s)\|\boldsymbol{C}^{m*} - \boldsymbol{C}^*\|_F.
\end{aligned}$$

Then we can bound above the right hand size by applying Lemmas 2-3,

$$\begin{aligned}
&\frac{1}{n}\|2\boldsymbol{X}_{\cdot\mathcal{A}}^\top(\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*} - \tilde{\boldsymbol{P}}^{\hat{\boldsymbol{C}}^{m-1}}) + \boldsymbol{X}_{\cdot\mathcal{A}}^\top \boldsymbol{X}(\hat{\boldsymbol{C}}^{m-1} - \boldsymbol{C}^{m*})\|_{op} \\
\le& 5\frac{c_+(s)^2}{c_-(s)}(1-2\zeta(1-\zeta))^{m-1}(\|\hat{\boldsymbol{C}}^0 - \hat{\boldsymbol{C}}\|_F + (1-2\zeta(1-\zeta))\|\hat{\boldsymbol{C}}^0 - \boldsymbol{C}^*\|_F) + 5c_+(s)\|\boldsymbol{C}^* - \hat{\boldsymbol{C}}\|_F
\end{aligned}$$

Therefore, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\max_{\mathcal{A}:|\mathcal{A}|\le s}\frac{1}{n}\|\boldsymbol{X}_{\cdot\mathcal{A}}^\top Err^m\|_{op} \le 5\frac{c_+(s)^2}{c_-(s)}(1-2\zeta(1-\zeta))^{m-1}(\|\hat{\boldsymbol{C}}^0 - \hat{\boldsymbol{C}}\|_F + (1-2\zeta(1-\zeta))\|\hat{\boldsymbol{C}}^0 - \boldsymbol{C}^*\|_F) \\
&+ 5c_+(s)\|\boldsymbol{C}^* - \hat{\boldsymbol{C}}\|_F + 2\sqrt{\frac{3qsP^2}{n}\log\frac{2pq}{\delta}},
\end{aligned}$$

which finishes the proof.

B.2. Proof of Lemma 2

The first part of the proof proceeds along similar lines of the proof of Theorem 1 of Balakrishnan et al. (2017). Denote $\boldsymbol{G} = \boldsymbol{X}\boldsymbol{C}$ and $\boldsymbol{G}' = X\boldsymbol{C}'$. Then, we can rewrite the surrogate function $S(\boldsymbol{C};\mathcal{D}|\boldsymbol{C}')$ with respect to G as $S_X(\boldsymbol{G}|\boldsymbol{G}')$. For convenience, we introduce the mapping M given by

$$M(G) := \arg\min_{\boldsymbol{G}'} S_X(\boldsymbol{G}'|\boldsymbol{G}). \tag{10}$$

Let $\hat{\boldsymbol{G}} = \boldsymbol{X}\hat{\boldsymbol{C}}$, then we know $\hat{\boldsymbol{G}} = M(\hat{\boldsymbol{G}})$, by the definition of $\hat{\boldsymbol{C}}$ and equation (10). In the first part, we show that the mapping M, in $B_F(r_0, \boldsymbol{C}^*)$ satisfies

$$\|M(\boldsymbol{G}) - \hat{\boldsymbol{G}}\|_F \le (1-2\zeta(1-\zeta))\|\boldsymbol{G} - \hat{\boldsymbol{G}}\|_F \quad \text{for all } \boldsymbol{G}, \tag{11}$$

where $0<\zeta<1, \zeta=\min_{\boldsymbol{C}\in B_F(r_0,\boldsymbol{C}^*),\boldsymbol{C}\in\Omega_r,i,j}\tilde{\boldsymbol{P}}^{\boldsymbol{C}}_{i,j}$. By the fact that $1/4\geq c_{\boldsymbol{C}^*}>0$ (Condition 3.3), we know that $1/4\geq\zeta(1-\zeta)>0$.

Denote the gradient of $S_X(\boldsymbol{G}|\boldsymbol{G}')$ with respect to $\boldsymbol{G}$ as $\nabla S_X(\boldsymbol{G}|\boldsymbol{G}')$. Then by the definition of surrogate function, we have

$$\nabla S_X(\boldsymbol{G}|\boldsymbol{G}')=\frac{1}{2n}\left(\boldsymbol{G}-\boldsymbol{G}'-2(\tilde{Y}-\tilde{\boldsymbol{P}}^{\boldsymbol{G}'})\right),$$

where $\tilde{\boldsymbol{P}}^{\boldsymbol{G}'}$ represents the term $\tilde{\boldsymbol{P}}^{\boldsymbol{C}'}$ being rewritten as a function of $\boldsymbol{G}'=\boldsymbol{X}\boldsymbol{C}'$. Making use the fact that

$$\nabla S_X(\hat{\boldsymbol{G}}|\hat{\boldsymbol{G}})-\nabla S_X(M(\boldsymbol{G})|\hat{\boldsymbol{G}})=\frac{1}{2n}\left(\hat{\boldsymbol{G}}-M(\boldsymbol{G})\right),$$

we can show that the following equation hold for all $\boldsymbol{G}$,

$$\langle\nabla S_X(\hat{\boldsymbol{G}}|\hat{\boldsymbol{G}})-\nabla S_X(M(\boldsymbol{G})|\hat{\boldsymbol{G}}),\hat{\boldsymbol{G}}-M(\boldsymbol{G})\rangle_F=\beta\|\hat{\boldsymbol{G}}-M(\boldsymbol{G})\|_F^2 \tag{12}$$

with $\beta=\frac{1}{2n}$.

Next, we show that the following inequality hold for all $\boldsymbol{G}$,

$$\|\nabla S_X(M(\boldsymbol{G})|\hat{\boldsymbol{G}})-\nabla S_X(M(\boldsymbol{G})|\boldsymbol{G})\|_F\leq\tau\|\boldsymbol{G}-\hat{\boldsymbol{G}}\|_F, \tag{13}$$

where $\tau=\frac{1}{2n}\big(1-2\zeta(1-\zeta)\big)$ and $0<\zeta<1$ is a constant related to $\hat{\boldsymbol{G}}$. Noting that

$$\nabla S_X(M(\boldsymbol{G})|\hat{\boldsymbol{G}})-\nabla S_X(M(\boldsymbol{G})|\boldsymbol{G})=\frac{1}{2n}\left(\boldsymbol{G}-\hat{\boldsymbol{G}}-2(\tilde{\boldsymbol{P}}^{\boldsymbol{G}}-\tilde{\boldsymbol{P}}^{\hat{\boldsymbol{G}}})\right)\triangleq f(\hat{\boldsymbol{G}}).$$

Taking gradient of $f(\cdot)$ with respect of $\hat{\boldsymbol{G}}$ leads to

$$\nabla f(\hat{\boldsymbol{G}})=\frac{1}{2n}\left(1-2\tilde{\boldsymbol{P}}^{\hat{\boldsymbol{G}}}(1-\tilde{\boldsymbol{P}}^{\hat{\boldsymbol{G}}})\right).$$

Let $P^{\hat{\boldsymbol{G}}}_{\min}$ denote the minimal element in the matrix $\tilde{\boldsymbol{P}}^{\hat{\boldsymbol{G}}}$ and assume that $P^{\boldsymbol{G}}_{\min}>0$. By the convexity of function f, we can bound the Frobenius norm of $f(\hat{\boldsymbol{G}})$ as

$$\|\nabla S_X(M(\boldsymbol{G})|\hat{\boldsymbol{G}})-\nabla S_X(M(\boldsymbol{G})|\boldsymbol{G})\|_F\leq\|\nabla f(\hat{\boldsymbol{G}})(\boldsymbol{G}-\hat{\boldsymbol{G}})\|_F\leq\tau\|\boldsymbol{G}-\hat{\boldsymbol{G}}\|_F$$

with $\tau=\frac{1}{2n}\left(1-2P^{\hat{\boldsymbol{G}}}_{\min}(1-P^{\hat{\boldsymbol{G}}}_{\min})\right)$ in $B_F(r_0,\boldsymbol{C}^*)$.

Finally, the Cauchy-Schwartz inequality implies that the left had side of (12) is upper bounded by

$$\langle\nabla S_X(\hat{\boldsymbol{G}}|\hat{\boldsymbol{G}})-\nabla S_X(M(\boldsymbol{G})|\hat{\boldsymbol{G}}),\hat{\boldsymbol{G}}-M(\boldsymbol{G})\rangle_F\leq\|\nabla S_X(M(\boldsymbol{G})|\hat{\boldsymbol{G}})-\nabla S_X(M(\boldsymbol{G})|\boldsymbol{G})\|_F\|\boldsymbol{G}-\hat{\boldsymbol{G}}\|_F.$$

Substituting the above inequality into (12) and (13) yields

$$\beta\|\hat{\boldsymbol{C}}-M(\boldsymbol{C})\|_F^2\leq\tau\|\hat{\boldsymbol{C}}-M(\boldsymbol{C})\|_F\|\boldsymbol{C}-\hat{\boldsymbol{C}}\|_F.$$

Canceling the term $\beta\|\hat{\boldsymbol{C}}-M(\boldsymbol{C})\|_F$ on both sides and $\zeta\leq\frac{\tau}{\beta},\forall\boldsymbol{C}\in B_F(r_0,\boldsymbol{C}^*),\boldsymbol{C}\in\omega_r$, we can obtain the inequality (11).

Let $\hat{\boldsymbol{G}}^m=\boldsymbol{X}\hat{\boldsymbol{C}}^m$, then we have $\hat{\boldsymbol{G}}^m=M(\hat{\boldsymbol{G}}^{m-1})$. By iteratively applying (11), we have for all $m=1,2,\ldots,$

$$\begin{aligned}\|\hat{\boldsymbol{G}}^m-\hat{\boldsymbol{G}}\|_F&\leq(1-2\zeta(1-\zeta))\,\|\hat{\boldsymbol{G}}^{m-1}-\hat{\boldsymbol{G}}\|_F\\&\leq\cdots\\&\leq(1-2\zeta(1-\zeta))^m\,\|\hat{\boldsymbol{G}}^0-\hat{\boldsymbol{G}}\|_F.\end{aligned} \tag{14}$$

Applying Condition 3.1, we can obtain the following inequalities

$$\|\hat{\boldsymbol{G}}^{(m)}-\hat{\boldsymbol{G}}\|_F=\|\boldsymbol{X}\hat{\boldsymbol{C}}^{(m)}-\boldsymbol{X}\hat{\boldsymbol{C}}\|_F\geq\sqrt{n}c_-(s)\|\hat{\boldsymbol{C}}^{(m)}-\hat{\boldsymbol{C}}\|_F,$$

$$\|\hat{\boldsymbol{G}}^{(0)}-\hat{\boldsymbol{G}}\|_F=\|\boldsymbol{X}\boldsymbol{C}^0-\boldsymbol{X}\hat{\boldsymbol{C}}\|_F\leq\sqrt{n}c_+(s)\|\hat{\boldsymbol{C}}^{(0)}-\hat{\boldsymbol{C}}\|_F.$$

Substituting these two inequalities into (14), we have

$$\|\hat{\boldsymbol{C}}^{(m)}-\hat{\boldsymbol{C}}\|_F\leq\big(1-2\zeta(1-\zeta)\big)^m\frac{c_+(s)}{c_-(s)}\|\hat{\boldsymbol{C}}^0-\hat{\boldsymbol{C}}\|_F.$$

B.3. Proof of Lemma 3

Rewrite the noise-free surrogate function $\tilde{S}(\boldsymbol{C};\mathcal{D}|\boldsymbol{C}')$ with respect to G as $\tilde{S}_X(\boldsymbol{G}|\boldsymbol{G}')$. For convenience, we introduce the mapping $\tilde{M}$ given by

$$\tilde{M}(G) := \arg\min_{\boldsymbol{G}'} \tilde{S}_X(\boldsymbol{G}'|\boldsymbol{G}). \tag{15}$$

The remaining proof follows similar lines of Lemma 2, and the details are omitted here.

B.4. Proof of Lemma 4

Recall that $\boldsymbol{C}^{m,k}$ is the output coefficient matrix at the k-th iteration of the inner loop and the m-th iteration of the outer loop in **Algorithm 2**. Let $\mathcal{A}^{m,k}$ be the corresponding row support set, and $\mathcal{I}^{m,k} = (\mathcal{A}^{m,k})^c$. Similarly, denote $\boldsymbol{B}^{m,k}$, $\Gamma^{m,k}$, and $\boldsymbol{V}^{m,k}$ as the intermediate matrices at the k-th iteration of **Algorithm 1** within the m-th iteration of **Algorithm 2**. From step 5 in **Algorithm 1**, we have

$$\Gamma^{m,k}_{\mathcal{I}^{m,k}} = \boldsymbol{X}^\top_{\cdot\mathcal{I}^{m,k}}(\boldsymbol{M}\boldsymbol{V}^{m,k} - \boldsymbol{X}\boldsymbol{B}^{m,k})/n.$$

From the definition of $\tilde{\Gamma}^{k+1}_{A^k_{22}\cdot}$ *and* $\Gamma^{k+1}_{A^k_{22}\cdot}$, we know

$$\tilde{\Gamma}^{k+1}_{A^k_{22}\cdot} = \Gamma^{k+1}_{A^k_{22}\cdot}(\boldsymbol{V}^{k+1})^\top + \boldsymbol{X}^\top_{\cdot A^k}\boldsymbol{M}^{(m)}(\mathbf{I} - \boldsymbol{P}_{\boldsymbol{V}})/n,$$

after some simplifications. Here we denote $\boldsymbol{P}_{\boldsymbol{V}} = \boldsymbol{V}^{k+1}(\boldsymbol{V}^{k+1})^\top$ for convenience. Using Lemma B.5 in Wen et al. (2022), we can get

$$\|\tilde{\Gamma}^{k+1}_{A^k_{22}\cdot}\|_{op} \le \|\Gamma^{k+1}_{A^k_{22}\cdot}(\boldsymbol{V}^{k+1})^\top\|_{op} + n^{-1}\|\boldsymbol{X}^\top_{\cdot A^k_{22}}\boldsymbol{M}^{(m)}(\mathbf{I} - \boldsymbol{P}_{\boldsymbol{V}})\|_{op}. \tag{16}$$

Assume that $rank(\boldsymbol{X}^\top_{\cdot A^k}\boldsymbol{X}\boldsymbol{C}^{(m)*}) = r^*$ without loss of the generality. Moreover, we have $rank(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{C}^{(m)*}) = r^*$. Thus there exists a linear transformer $\mathbb{T} \in \mathbb{R}^{|A^k_{22}|\times|A^k|}$ satisfying $\boldsymbol{X}^\top_{\cdot A^k_{22}}\boldsymbol{X}\boldsymbol{C}^{(m)*} = \mathbb{T}(\boldsymbol{X}^\top_{\cdot A^k})\boldsymbol{X}\boldsymbol{C}^{(m)*}$. Because $\boldsymbol{X}^\top_{\cdot A^k_{22}}\boldsymbol{X}\boldsymbol{C}^{(m)*}$ is a submatrix of the original matrix. It ensures that

$$\begin{aligned} n^{-1}\|\boldsymbol{X}^\top_{\cdot A^k_{22}}\boldsymbol{X}\boldsymbol{C}^{(m)*}(\mathbf{I} - \boldsymbol{P}_{\boldsymbol{V}})\|_{op} &\le \frac{\|\mathbb{T}\|_{op}}{n}\|\boldsymbol{X}^\top_{\cdot A^k}\boldsymbol{X}\boldsymbol{C}^{(m)*}(\mathbf{I} - \boldsymbol{P}_{\boldsymbol{V}})\|_{op} \\ &\le \frac{\alpha_0}{n}\|\boldsymbol{X}^\top_{\cdot A^k}\boldsymbol{X}\boldsymbol{C}^{(m)*}(\mathbf{I} - \boldsymbol{P}_{\boldsymbol{V}})\|_{op}. \end{aligned}$$

And we also have

$$\begin{aligned} \boldsymbol{X}^\top_{\cdot A^k}\boldsymbol{X}\boldsymbol{C}^*(\mathbf{I} - \boldsymbol{P}_{\boldsymbol{V}}) &= \boldsymbol{X}^\top_{\cdot A^k}M^m(\mathbf{I} - \boldsymbol{P}_{\boldsymbol{V}}) - \boldsymbol{X}^\top_{\cdot A^k}Err^m(\mathbf{I} - \boldsymbol{P}_{\boldsymbol{V}}) \\ &= (\boldsymbol{X}^\top_{\cdot A^k}\boldsymbol{X}_{\cdot A^k})^{1/2}\sum_{l=1}^{q-1}\sigma_l^{k+1}\boldsymbol{u}_l^{k+1}(\boldsymbol{v}_l^{k+1})^\top(\mathbf{I} - \boldsymbol{P}_{\boldsymbol{V}}) - \boldsymbol{X}^\top_{\cdot A^k}Err^m(\mathbf{I} - \boldsymbol{P}_{\boldsymbol{V}}). \end{aligned} \tag{17}$$

The decomposition above comes from

$$(\boldsymbol{X}^\top_{\cdot A^k}\boldsymbol{X}_{\cdot A^k})^{-1/2}\boldsymbol{X}^\top_{\cdot A^k}M^m = \sum_{l=1}^{q-1}\sigma_l^{k+1}\boldsymbol{u}_l^{k+1}(\boldsymbol{v}_l^{k+1})^\top,$$

where $\sigma_l^{k+1}, \boldsymbol{u}_l^{k+1}$ are the l-th largest singular value and corresponding left singular vector, respectively. Due to the orthogonality of $\boldsymbol{v}_l^{k+1}$ and $\boldsymbol{P}_{\boldsymbol{V}}$, we have

$$\sum_{l=1}^{q-1}\sigma_l^{k+1}\boldsymbol{u}_l^{k+1}(\boldsymbol{v}_l^{k+1})^\top(\mathbf{I} - \boldsymbol{P}_{\boldsymbol{V}}) = \sum_{l=r+1}^{q-1}\sigma_l^{k+1}\boldsymbol{u}_l^{k+1}(\boldsymbol{v}_l^{k+1})^\top(\mathbf{I} - \boldsymbol{P}_{\boldsymbol{V}})$$

Meanwhile, by the definition of σ_{r+1}^{k+1}, we know that it is the $(r+1)$-th largest singular value of

$$(\boldsymbol{X}_{\cdot A^k}^\top \boldsymbol{X}_{\cdot A^k})^{-1/2}\boldsymbol{X}_{\cdot A^k}^\top M^m = (\boldsymbol{X}_{\cdot A^k}^\top \boldsymbol{X}_{\cdot A^k})^{-1/2}\boldsymbol{X}_{\cdot A^k}^\top \boldsymbol{X}\boldsymbol{C}^{m*} + (\boldsymbol{X}_{\cdot A^k}^\top \boldsymbol{X}_{\cdot A^k})^{-1/2}\boldsymbol{X}_{\cdot A^k}^\top Err^m.$$

Therefore, by the inequality (17), we have

$$\begin{aligned}
\|\boldsymbol{X}_{\cdot A^k}^\top \boldsymbol{X}\boldsymbol{C}^*(\mathbf{I}-\boldsymbol{P}_{\boldsymbol{V}})\|_F &\le \|(\boldsymbol{X}_{\cdot A^k}^\top \boldsymbol{X}_{\cdot A^k})^{1/2}\sum_{l=r+1}^{q-1}\sigma_l^{k+1}\boldsymbol{u}_l^{k+1}(\boldsymbol{v}_l^{k+1})^\top(\mathbf{I}-\boldsymbol{P}_{\boldsymbol{V}})\|_F + \|\boldsymbol{X}_{\cdot A^k}^\top Err^m\|_F \\
&\le \sigma_{r+1}^{k+1}\|(\boldsymbol{X}_{\cdot A^k}^\top \boldsymbol{X}_{\cdot A^k})^{1/2}\|_F + \|\boldsymbol{X}_{\cdot A^k}^\top Err^m\|_F \\
&\le \|\boldsymbol{X}_{\cdot A^k}^\top Err^{(m)}\|_{op}\|(\boldsymbol{X}_{\cdot A^k}^\top \boldsymbol{X}_{\cdot A^k})^{-1/2}\|_{op}\|(\boldsymbol{X}_{\cdot A^k}^\top \boldsymbol{X}_{\cdot A^k})^{1/2}\|_{op} + \|\boldsymbol{X}_{\cdot A^k}^\top Err^m\|_F.
\end{aligned}$$

Combining (16) and applying the triangle inequality again, we can obtain

$$\begin{aligned}
\|\tilde{\Gamma}_{A_{22}^k}^{k+1}\|_{op} &\le \|\Gamma_{A_{22}^k\cdot}^{k+1}\|_{op} + n^{-1}\|\boldsymbol{X}_{\cdot A_{22}^k}^\top \boldsymbol{X}\boldsymbol{C}^{(m)*}(\mathbf{I}-\boldsymbol{P}_{\boldsymbol{V}})\|_{op} + n^{-1}\|\boldsymbol{X}_{\cdot A^k}^\top Err^m\|_{op} \\
&\le \|\Gamma_{A_{22}^k\cdot}^{k+1}\|_{op} + \frac{\alpha_0}{n}\|\boldsymbol{X}_{\cdot A^k}^\top \boldsymbol{X}\boldsymbol{C}^{(m)*}(\mathbf{I}-\boldsymbol{P}_{\boldsymbol{V}})\|_{op} + n^{-1}\|\boldsymbol{X}_{\cdot A^k}^\top Err^m\|_{op} \\
&\le \|\Gamma_{A_{22}^k\cdot}^{k+1}\|_{op} + \frac{\alpha_0}{n}\|\boldsymbol{X}_{\cdot A^k}^\top Err^{(m)}\|_{op}\|(\boldsymbol{X}_{\cdot A^k}^\top \boldsymbol{X}_{\cdot A^k})^{-1/2}\|_{op}\|(\boldsymbol{X}_{\cdot A^k}^\top \boldsymbol{X}_{\cdot A^k})^{1/2}\|_{op} \\
&+ \frac{\alpha_0}{n}\|\boldsymbol{X}_{\cdot A^k}^\top Err^{(m)}\|_{op} + n^{-1}\|\boldsymbol{X}_{\cdot A^k}^\top Err^m\|_{op} \\
&\le \|\Gamma_{A_{22}^k\cdot}^{k+1}\|_{op} + \frac{\alpha_0(1+c_-(s)c_+(s))+1}{n}\|\boldsymbol{X}_{\cdot A^k}^\top Err^m\|_{op},
\end{aligned}$$

where α_0 is the operator norm of $\mathbb{T}$ and the last inequality comes from condition 3.1.

B.5. Lemma 5 and its proof

LEMMA 5. *Suppose each row of $\boldsymbol{Y}$ comes from the multinomial logistic model with a deterministic $n\times p$ design matrix $\boldsymbol{X}$ and an $p\times q$ coefficient matrix $\boldsymbol{C}$. Denote $P^2 = \max_j diag(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X})\|\tilde{\boldsymbol{P}}_{j\cdot}^{\boldsymbol{C}^*}\|_F$ and the probability matrix as $\boldsymbol{P}^{\boldsymbol{C}^*}$. Let $\tilde{\boldsymbol{Y}}$ and $\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*}$ denote the matrix by eliminating the first column of $\boldsymbol{Y}$ and $\boldsymbol{P}^{\boldsymbol{C}^*}$. Then for any $\delta\in(0,1)$, with probability at least $1-\delta$,*

$$\max_{\mathcal{A}:|\mathcal{A}|\le s}\frac{1}{n}\|(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})^\top\boldsymbol{X}_{\cdot\mathcal{A}}\|_{op} \le \sqrt{\frac{3qsP^2}{n}\log\frac{2pq}{\delta}}.$$

Proof For any $\zeta>0$, by definition,

$$\begin{aligned}
Pr\left(\max_{\mathcal{A}:|\mathcal{A}|\le s}\frac{1}{n}\|(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})^\top\boldsymbol{X}_{\cdot\mathcal{A}}\|_{op}\ge\zeta\right) &= Pr\left(\max_{\mathcal{A}:|\mathcal{A}|\le s}\frac{1}{n^2}\|(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})^\top\boldsymbol{X}_{\cdot\mathcal{A}}\|_{op}^2\ge\zeta^2\right) \\
&\le Pr\left(\sup_{\|\boldsymbol{u}\|_2=1,\|\boldsymbol{u}\|_0\le s} n^{-2}\|(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})^\top\boldsymbol{X}\boldsymbol{u}\|_2^2\ge\zeta^2\right) \\
&\le Pr\left(\sum_{i=1}^q\sup_{\|\boldsymbol{u}\|_2=1,\|u\|_0\le s} n^{-2}((\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})_{\cdot i}^\top\boldsymbol{X}\boldsymbol{u})^2\ge\zeta^2\right),
\end{aligned}$$

where $(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})_{\cdot i}$ denotes the i-th column of matrix $\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*}$. Applying the union bound yields

$$\begin{aligned}
&Pr\left(\sum_{i=1}^{q} \sup_{\|\boldsymbol{u}\|_2=1, \|u\|_0 \le s} \frac{1}{n^2}((\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})_{\cdot i}^{\top} \boldsymbol{X}\boldsymbol{u})^2 \ge \zeta^2\right) \\
\le &\sum_{i=1}^{q} Pr\left(\sup_{\|\boldsymbol{u}\|_2=1, \|u\|_0 \le s} \frac{1}{n^2}\left((\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})_{\cdot i}^{\top} \boldsymbol{X}\boldsymbol{u}\right)^2 \ge \frac{\zeta^2}{q}\right) \\
= &\sum_{i=1}^{q} Pr\left(\sup_{\|\boldsymbol{u}\|_2=1, \|u\|_0 \le s} \frac{1}{n}\left|(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})_{\cdot i}^{\top} \boldsymbol{X}\boldsymbol{u}\right| \ge \frac{\zeta}{\sqrt{q}}\right) \\
\le &\sum_{i=1}^{q} Pr\left(\max_{1 \le j \le p} \frac{1}{n}|(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})_{\cdot i}^{\top} \boldsymbol{X}_{\cdot j}| \ge \frac{\zeta}{\sqrt{qs}}\right) \\
\le &\sum_{i=1}^{p}\sum_{i=1}^{q} Pr\left(\frac{1}{\sqrt{n}}|(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{P}}^{\boldsymbol{C}^*})_{\cdot i}^{\top} \boldsymbol{X}_{\cdot j}| \ge \zeta\sqrt{\frac{n}{qs}}\right) \\
\le &2pq \exp\left(-\frac{n\zeta^2}{3qsP^2}\right),
\end{aligned}$$

where the Bernoulli Chernoff bound (Vershynin 2018) is applied in the last inequality. By taking $\delta = 2pq\exp(-\frac{n\zeta^2}{2qsP^2})$, we obtain $\zeta = \sqrt{\frac{3qsP^2}{n}\log\frac{2pq}{\delta}}$, which concludes the proof.

Appendix C: HAM10000 data analysis

The HAM10000 dataset consists of 10,015 dermatoscopic images from different populations, acquired and stored by different modalities. It includes a representative collection of 7 important diagnostic categories in the realm of pigmented lesions: Actinic keratoses and intraepithelial carcinoma/Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (solar lentigines/seborrheic keratoses and lichen-planus like keratoses, bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, vasc). Figure 1 shows five samples of the original images.

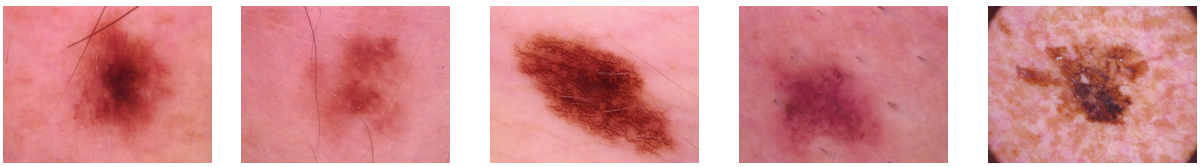

Figure 1 5 examples of HAM10000 dataset

Table 1 reports the results for each method. Notice that the minus log-likelihood value Pred is calculated on the whole data set since we do not have an independent test data set here. We can see that our method yields a sparse reduced rank model with the lowest sparsity. In contrast, the mLasso method identifies much more predictors and a higher Pred value. NNET has very poor performance for this dataset due to its high sparsity, spending most of the time wandering randomly. Due to the sensitivity and instability to data, RRVGLM does't work for this specific data.

To compare the prediction accuracy and stability for these methods, we randomly split the whole data set into a training set with size 900 and a testing set with the rest samples. The optimal pair of rank and

Table 1 **The results on the whole data set.**

Method	Pred	$\hat{r}$	$\lvert\hat{\mathcal{A}}\rvert$
SRRMLR	16018.99	3	31
mLasso	18481.05	-	182
NNET	18611.97	-	192
RRVGLM	-	-	-

'-', these results can not be obtained.

Table 2 **The average results on the test set of HAM10000 with the standard deviation in parentheses.**

Method	Pred	$\hat{r}$	$\lvert\hat{\mathcal{A}}\rvert$
SRRMLR	4919.94 (57.89)	3.84 (1.04)	29.09 (5.69)
mLasso	5646.84 (87.18)	-	173.59 (5.14)
NNET	6492.99 (232.37)	-	192 (0)
RRVGLM	-	-	-

'-', these results can not be obtained.

sparsity is determined by the strategy discussed above, but the 5-fold cross-validation procedure is carried out using the training set. According to the upper bounds developed in the simulation setup, we set the maximum range $(r_{\max}, s_{\max})$ of these two tuning parameters to 6 and 38, respectively. In addition, we use the testing set to calibrate the predictive performance of each estimator $\hat{\boldsymbol{C}}$ by Pred in (11). The random splitting process is implemented 100 times to yields the average performance of the above methods.

Summary results are presented in Table 2. Figure 2 displays line charts of the 100 replications for the measurements of Pred and $|\hat{\mathcal{A}}|$. Overall, the results are accordance with those in Table 1. In particular, the proposed SRRMLR method yields significant better performance compared to other methods in all of the measurements. The mLasso achieves competitive prediction accuracy, but it produces a much larger model size with too many selected predictors. The RRVGLM cannot even be applied to this data set. The NNET leads to models with the largest model size and the highest value in Pred. Furthermore, the models from NNET are unstable in terms of predictive accuracy, which can be also seen from Figure 2. We conclude that the SRRMLR is feasible and effective when handling data in practice with the consideration of both prediction accuracy and model interpretability.

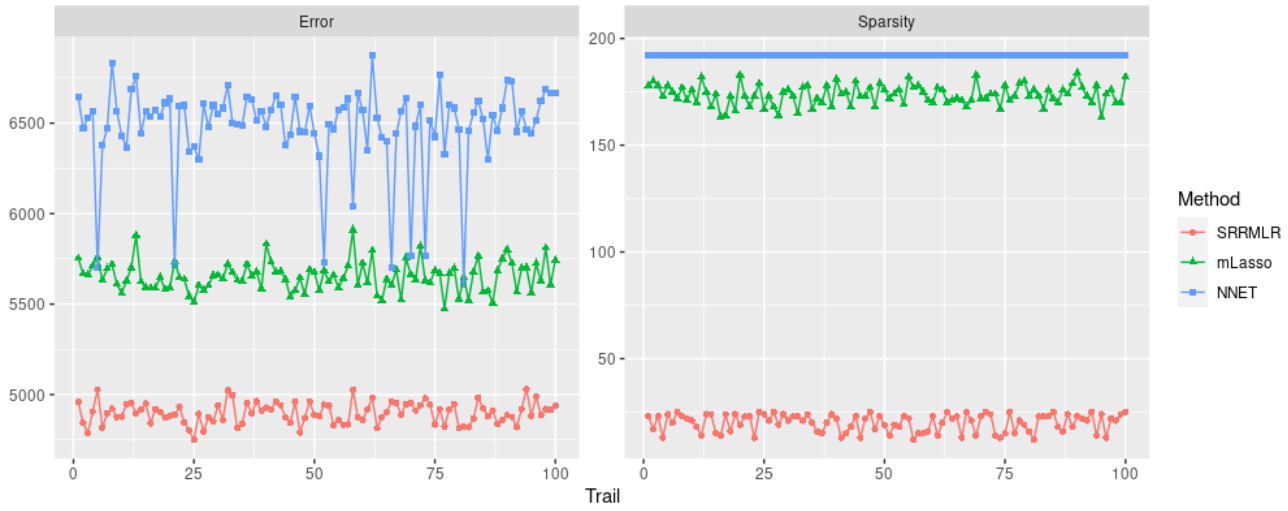

Figure 2 **The trails of prediction error (left) and row sparsity (right) among 100 replications.**

References

Balakrishnan S, Wainwright MJ, Yu B (2017) Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics* 45(1):77 – 120.

Simon N, Friedman J, Hastie T (2013) A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. *arXiv preprint arXiv:1311.6529* .

Vershynin R (2018) *High-dimensional probability: An introduction with applications in data science*, volume 47 (Cambridge university press).

Wen C, Dong R, Wang X, Li W, Zhang H (2022) Simultaneous best subset selection and dimension reduction via primal-dual iterations. *arXiv:2211.15889* .