Supplementary Appendix for High-dimensional Subgroup Regression Analysis

Fei Jiang, Lu Tian, Jian Kang, and Lexin Li

In this appendix, we first present a number of supporting lemmas. We then present the proofs of Theorems 1 and 2 of the paper.

S.1 Supporting lemmas

Lemma S1. *Let U be the uniformity of vector $\boldsymbol{x}$ defined as the ratio between the smallest nonzero entry and the largest one. Then*

$$\sqrt{\|\boldsymbol{x}\|_0} \leq \frac{1+U}{2\sqrt{U}} \frac{\|\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_2}$$

Proof: The lemma follows Lemma III.3 in Yin et al. (2014).

Lemma S2. *Let $\boldsymbol{A} = \boldsymbol{W}^{\top}\boldsymbol{W}/n$, and $\mathcal{M}$ and $\mathcal{N}$ denote the non-overlapping indices in $(1, \ldots, pq)$. Then,*

$$\|\boldsymbol{A}_{\mathcal{M},\mathcal{N}}\|_2 \leq \sqrt{\{\rho_+(\mathcal{M}) - \rho_-(\mathcal{M}\cup\mathcal{N})\}\{\rho_+(\mathcal{N}) - \rho_-(\mathcal{M}\cup\mathcal{N})\}}.$$

PROOF: A direct application of Lemma D.1 of Huang and Zhang (2010) completes the proof of Lemma S2. □

Lemma S3. *Denote $\boldsymbol{\Delta} = \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0$, and let Q_r and Q_c denote the sets of indices of the nonzero rows and columns of $\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0$. Let $R_j = \{(k-1)p + j, k = 1, \ldots, q\}$ be the ordered set of indices of vec$(\boldsymbol{\Theta}_0)$ corresponding to the elements in the jth row of $\boldsymbol{\Theta}_0$, and $C_k = \{(k-1)p + j, j = 1, \ldots, p\}$ be the set of indices of vec$(\boldsymbol{\Theta}_0)$ corresponding to the elements in kth column of $\boldsymbol{\Theta}_0$. Furthermore, we define $R_j(Q)$ and $C_k(Q)$ to be the subset of R_j and C_k with elements in a index set Q. For a given $m \geq 1$, let $S_{r0}, S_{r1}, \ldots$ and $S_{c0}, S_{c1}, \ldots$ be the consecutive blocks containing indices $j > g_r$, $k > g_c$, such that $2b_1 \leq |R_{S_{ru}}(Q_c)| \leq 2b_1 + m$, and $2b_1 \leq |C_{S_{cv}}(Q_r)| \leq 2b_1 + m$, where $R_{S_{ru}}(Q_c) = \{R_j(Q_c), j \in S_{ru}\}$, and $R_{S_{cv}}(Q_r) = \{R_j(Q_r), j \in S_{cv}\}$, $u, v = 0, 1, 2, \ldots$, respectively. Let $\bar{G}_r = \cup_{j\in S_r\cup S_{r0}} R_j(Q_c)$, $\bar{G}_c = \cup_{k\in S_c\cup S_{c0}} C_k(Q_r)$, and $G = \bar{G}_r \cup \bar{G}_c$. Define $\widetilde{\rho}_+(a,b) = ([\{\rho_+(a) - \rho_-(a+b)\}][\{\rho_+(b) - \rho_-(a+b)\}])^{1/2}$. Then,*

$$\sum_{j\notin S_r\cup S_{r0}, k\notin S_c\cup S_{c0}} \|\boldsymbol{\Delta}_{R_j\cap C_k}\|_2^2 \leq \min_{u,v}(4|S_{ru}|, 4|S_{cv}|)^{-1}\left(\sum_{j\notin S_r}\|\boldsymbol{\Delta}_{R_j}\|_2 + \sum_{k\notin S_c}\|\boldsymbol{\Delta}_{C_k}\|_2\right)^2,$$

$$\frac{1}{n}\left|\sum_{j\notin S_r\cup S_{r0},k\notin S_c\cup S_{c0}} \boldsymbol{\Delta}_G^\top \boldsymbol{W}_G^\top \boldsymbol{W}_{R_j\cap C_k}\boldsymbol{\Delta}_{R_j\cap C_k}\right|$$
$$\leq \widetilde{\rho}_+(|G|,2b_1+m)\|\boldsymbol{\Delta}_G\|_2\left\{\min_{u,v}(|S_{ru}|,|S_{cv}|)\right\}^{-1/2}\left\{\sum_{j\notin S_r}\|\boldsymbol{\Delta}_{R_j}\|_2+\sum_{k\notin S_c}\|\boldsymbol{\Delta}_{C_k}\|_2\right\}/2.$$

PROOF: We have that,

$$\begin{aligned}
&\sum_{j\notin S_r\cup S_{r0},k\notin S_c\cup S_{c0}}\|\boldsymbol{\Delta}_{R_j\cap C_k}\|_2^2\\
\leq\;&\sum_{j\notin S_r\cup S_{r0}}\|\boldsymbol{\Delta}_{R_j}\|_2^2/2+\sum_{k\notin S_c\cup S_{c0}}\|\boldsymbol{\Delta}_{C_k}\|_2^2/2\\
\leq\;&\sum_{j\notin S_r\cup S_{r0}}\|\boldsymbol{\Delta}_{R_j}\|_2\max_{j\notin S_r\cup S_{r0}}\|\boldsymbol{\Delta}_{R_j}\|_2/2+\sum_{k\notin S_c\cup S_{c0}}\|\boldsymbol{\Delta}_{C_k}\|_2\max_{k\notin S_c\cup S_{c0}}\|\boldsymbol{\Delta}_{C_k}\|_2/2\\
\leq\;&\sum_{j\notin S_r\cup S_{r0}}\|\boldsymbol{\Delta}_{R_j}\|_2\min_{j\in S_{r0}}\|\boldsymbol{\Delta}_{R_j}\|_2/2+\sum_{k\notin S_c\cup S_{c0}}\|\boldsymbol{\Delta}_{C_k}\|_2\min_{k\in S_{c0}}\|\boldsymbol{\Delta}_{C_k}\|_2/2\\
\leq\;&\sum_{j\notin S_r\cup S_{r0}}\|\boldsymbol{\Delta}_{R_j}\|_2\sum_{j\in S_{r0}}\|\boldsymbol{\Delta}_{R_j}\|_2/(2|S_{r0}|)\\
&+\sum_{k\notin S_c\cup S_{c0}}\|\boldsymbol{\Delta}_{C_k}\|_2\sum_{k\in S_{c0}}\|\boldsymbol{\Delta}_{C_k}\|_2/(2|S_{c0}|)\\
\leq\;&\min_{u,v}(4|S_{ru}|,4|S_{cv}|)^{-1}\left(\sum_{j\notin S_r}\|\boldsymbol{\Delta}_{R_j}\|_2+\sum_{k\notin S_c}\|\boldsymbol{\Delta}_{C_k}\|_2\right)^2.
\end{aligned}$$

The last inequality holds by the fact that $ab\leq(a+b)^2/2$ for any positive a,b.

Furthermore, we have that,

$$\begin{aligned}
&\sum_{u\geq1,v\geq1}\|\boldsymbol{\Delta}_{R_{S_{ru}}\cap C_{S_{cv}}}\|_2=\sum_{u\geq1,v\geq1}\sqrt{\sum_{j\in S_{ru},k\in S_{cv}}\|\boldsymbol{\Delta}_{R_{rj}\cap C_{ck}}\|_2^2}\\
\leq\;&\sum_{u\geq1,v\geq1}\sqrt{\sum_{j\in S_{ru},k\in S_{cv}}\|\boldsymbol{\Delta}_{R_j\cap C_k}\|_2^2}\\
\leq\;&\sum_{u\geq1,v\geq1}\sqrt{\sum_{j\in S_{ru}}\|\boldsymbol{\Delta}_{R_j}\|_2\max_{j\in S_{ru}}\|\boldsymbol{\Delta}_{R_j}\|_2+\sum_{k\in S_{cv}}\|\boldsymbol{\Delta}_{C_k}\|_2\max_{k\in S_{cv}}\|\boldsymbol{\Delta}_{C_k}\|_2}\\
\leq\;&\sum_{u\geq1,v\geq1}\sqrt{\sum_{j\in S_{ru}}\|\boldsymbol{\Delta}_{R_j}\|_2\min_{j\in S_{ru-1}}\|\boldsymbol{\Delta}_{R_j}\|_2+\sum_{k\in S_{cv}}\|\boldsymbol{\Delta}_{C_k}\|_2\min_{k\in S_{cv-1}}\|\boldsymbol{\Delta}_{C_k}\|_2}\\
\leq\;&\sum_{u\geq1,v\geq1}\left\{\sum_{j\in S_{ru}}\|\boldsymbol{\Delta}_{R_j}\|_2\sum_{j\in S_{ru-1}}\|\boldsymbol{\Delta}_{R_j}\|_2/|S_{ru-1}|+\sum_{k\in S_{cv}}\|\boldsymbol{\Delta}_{C_k}\|_2\sum_{j\in S_{cv-1}}\|\boldsymbol{\Delta}_{C_k}\|_2/|S_{cv-1}|\right\}^{1/2}
\end{aligned}$$

$$
\begin{aligned}
&\leq \{\min_{u,v}(|S_{ru}|,|S_{cv}|)\}^{-1/2}1/2 \sum_{u\geq 1,v\geq 1}\left\{\left(\sum_{j\in S_{ru}}\|\boldsymbol{\Delta}_{R_j}\|_2+\sum_{j\in S_{ru-1}}\|\boldsymbol{\Delta}_{R_j}\|_2\right)^2\right.\\
&\quad\left.+\left(\sum_{j\in S_{cv}}\|\boldsymbol{\Delta}_{C_k}|_2+\sum_{k\in S_{cv-1}}\|\boldsymbol{\Delta}_{C_k}\|_2\right)^2\right\}^{1/2}\\
&\leq \{\min_{u,v}(|S_{ru}|,|S_{cv}|)\}^{-1/2}1/2 \sum_{u\geq 1,v\geq 1}\left(\sum_{j\in S_{ru}}\|\boldsymbol{\Delta}_{R_j}\|_2+\sum_{j\in S_{ru-1}}\|\boldsymbol{\Delta}_{R_j}\|_2\right.\\
&\quad\left.+\sum_{j\in S_{cv}}\|\boldsymbol{\Delta}_{C_k}|_2+\sum_{k\in S_{cv-1}}\|\boldsymbol{\Delta}_{C_k}\|_2\right)\\
&\leq \{\min_{u,v}(|S_{ru}|,|S_{cv}|)\}^{-1/2}1/2\left(\sum_{u\geq 0}\sum_{j\in S_{ru}}\|\boldsymbol{\Delta}_{R_j}\|_2+\sum_{v\geq 0}\sum_{k\in S_{cv}}\|\boldsymbol{\Delta}_{C_k}\|_2\right)\\
&= \{\min_{u,v}(4|S_{ru}|,4|S_{cv}|)\}^{-1/2}\left(\sum_{j\notin S_r}\|\boldsymbol{\Delta}_{R_j}\|_2+\sum_{k\notin S_c}\|\boldsymbol{\Delta}_{C_k}\|_2\right).
\end{aligned}
$$

Therefore, we have that,

$$
\begin{aligned}
&\frac{1}{n}\left|\sum_{j\notin S_r\cup S_{r0},k\notin S_c\cup S_{c0}}\boldsymbol{\Delta}_G^{\intercal}\boldsymbol{W}_G^{\intercal}\boldsymbol{W}_{R_j\cap C_k}\boldsymbol{\Delta}_{R_j\cap C_k}\right|\\
&\leq \frac{1}{n}\sum_{u\geq 1,v\geq 1}\left|\boldsymbol{\Delta}_G^{\intercal}\boldsymbol{W}_G^{\intercal}\boldsymbol{W}_{R_{S_{ru}}\cap C_{S_{cv}}}\boldsymbol{\Delta}_{R_{S_{ru}}\cap C_{S_{cv}}}\right|\\
&\leq \frac{1}{n}\sum_{u\geq 1,v\geq 1}\|\boldsymbol{W}_G^{\intercal}\boldsymbol{W}_{R_{S_{ru}}(Q_c)\cap R_{S_{cv}}(Q_r)}\|_2\|\boldsymbol{\Delta}_G\|_2\|\boldsymbol{\Delta}_{R_{S_{ru}}(Q_c)\cap C_{S_{cv}}(Q_r)}\|_2\\
&\leq \widetilde{\rho}_+(|G|,2b_1+m)\|\boldsymbol{\Delta}_G\|_2\sum_{u\geq 1,v\geq 1}\|\boldsymbol{\Delta}_{G_{S_{ru}}\cap G_{S_{cv}}}\|_2\\
&\leq \widetilde{\rho}_+(|G|,2b_1+m)\|\boldsymbol{\Delta}_G\|_2\left\{\min_{u,v}(|S_{ru}|,|S_{cv}|)\right\}^{-1/2}\left\{\sum_{j\notin S_r}\|\boldsymbol{\Delta}_{R_j}\|_2+\sum_{k\notin S_c}\|\boldsymbol{\Delta}_{C_k}\|_2\right\}/2.
\end{aligned}
$$

The third inequality holds due to Lemma S2. This completes the proof of Lemma S3. □

Lemma S4. *Suppose that* $\lambda_r\sqrt{q}\geq 2n^{-1}\max_{j=1,\ldots,p,j\in Q_r}\|\boldsymbol{W}_{R_j(Q_c)}^{\intercal}\boldsymbol{\epsilon}\|_2$*, and* $\lambda_c\sqrt{p}\geq 2n^{-1}\max_{k=1,\ldots,q,k\in Q_c}\|\boldsymbol{W}_{C_k(Q_r)}^{\intercal}\boldsymbol{\epsilon}\|_2$*. Then, the solution for (2) satisfies that,*

$$
\lambda_c\sqrt{p}\sum_{k=2,k\notin S_c}^{q}\|\{vec(\widehat{\boldsymbol{\Theta}})\}_{C_k}\|_2+\lambda_r\sqrt{q}\sum_{j=1,j\notin S_r}^{p}\|\{vec(\widehat{\boldsymbol{\Theta}})\}_{R_j}\|_2
$$

$$\leq \ 3\lambda_c\sqrt{p}\sum_{k=2,k\in S_c}^{q}\|\{vec(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)\}_{C_k}\|_2+3\lambda_r\sqrt{q}\sum_{j=1,j\in S_r}^{p}\|\{vec(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)\}_{R_j}\|_2.$$

PROOF: From (2), we obtain the subgradient condition for the minimizer that

$$\boldsymbol{W}^\intercal\boldsymbol{W}\text{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)-\boldsymbol{W}^\intercal\boldsymbol{\epsilon}+n\lambda_c\sqrt{p}\sum_{k=2}^{q}\boldsymbol{\nu}_k+n\lambda_r\sqrt{q}\sum_{j=1}^{p}\boldsymbol{\mu}_j=\mathbf{0}, \tag{S.1}$$

where $\boldsymbol{\mu}$ is the subgradient of $\|\mathbf{\Theta}_{j\cdot}\|_2$ with respect to $\text{vec}(\mathbf{\Theta})$ evaluated at $\widehat{\mathbf{\Theta}}$; i.e.,

$$\begin{aligned}\boldsymbol{\mu}_{jR_j^c} &= \mathbf{0},\\ \boldsymbol{\mu}_{jR_j} &= \begin{cases}\frac{\widehat{\mathbf{\Theta}}_{j\cdot}}{\|\widehat{\mathbf{\Theta}}_{j\cdot}\|_2}, & \text{if } \widehat{\mathbf{\Theta}}_{j\cdot}\neq\mathbf{0},\\ \forall\boldsymbol{\mu}, \|\boldsymbol{\mu}\|_2\leq 1, & \text{if } \widehat{\mathbf{\Theta}}_{j\cdot}=\mathbf{0}.\end{cases}\end{aligned}$$

and $\boldsymbol{\nu}$ is the subgradient of $\|\mathbf{\Theta}_{\cdot k}\|_2$ with respect to $\text{vec}(\mathbf{\Theta})$ evaluated at $\widehat{\mathbf{\Theta}}$; i.e.,

$$\begin{aligned}\boldsymbol{\nu}_{kC_k^c} &= \mathbf{0},\\ \boldsymbol{\nu}_{kC_k} &= \begin{cases}\frac{\widehat{\mathbf{\Theta}}_{\cdot k}}{\|\widehat{\mathbf{\Theta}}_{\cdot k}\|_2}, & \text{if } \widehat{\mathbf{\Theta}}_{\cdot k}\neq\mathbf{0},\\ \forall\boldsymbol{\nu} : \|\boldsymbol{\nu}\|_2\leq 1, & \text{if } \widehat{\mathbf{\Theta}}_{\cdot k}=\mathbf{0}.\end{cases}\end{aligned}$$

By definition, we then have

$$\begin{aligned}&\text{vec}(\widehat{\mathbf{\Theta}})^\intercal\boldsymbol{\mu}_j=\|\text{vec}(\widehat{\mathbf{\Theta}})_{R_j}\|_2, \quad \text{vec}(\widehat{\mathbf{\Theta}})^\intercal\boldsymbol{\nu}_k=\|\text{vec}(\widehat{\mathbf{\Theta}})_{C_k}\|_2,\\ &|\text{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)^\intercal\boldsymbol{\mu}_j|\leq\|\text{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)_{R_j}\|_2, \quad \text{and } |\text{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)^\intercal\boldsymbol{\nu}_k|\leq\|\text{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)_{C_k}\|_2.\end{aligned}$$

Multiplying both sides of (S.1) by $\text{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)$, we obtain that

$$\begin{aligned}0 \ \geq\ & -\text{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)^\intercal\boldsymbol{W}^\intercal\boldsymbol{W}\text{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)=-\text{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)^\intercal\boldsymbol{W}^\intercal\boldsymbol{\epsilon}\\ &+n\lambda_c\sqrt{p}\sum_{k=2}^{q}\text{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)^\intercal\boldsymbol{\nu}_k+n\lambda_r\sqrt{q}\sum_{j=1}^{p}\text{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)^\intercal\boldsymbol{\mu}_j\\ =\ & -\text{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)^\intercal\boldsymbol{W}^\intercal\boldsymbol{\epsilon}\\ &+n\lambda_c\sqrt{p}\sum_{k=2,k\in S_c}^{q}\text{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)^\intercal\boldsymbol{\nu}_k+n\lambda_r\sqrt{q}\sum_{j=1,j\in S_r}^{p}\text{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)^\intercal\boldsymbol{\mu}_j\\ &+n\lambda_c\sqrt{p}\sum_{k=2,k\notin S_c}^{q}\text{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)^\intercal\boldsymbol{\nu}_k+n\lambda_r\sqrt{q}\sum_{j=1,j\notin S_r}^{p}\text{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)^\intercal\boldsymbol{\mu}_j\\ \geq\ & -\text{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)^\intercal\boldsymbol{W}^\intercal\boldsymbol{\epsilon}\\ &-n\lambda_c\sqrt{p}\sum_{k=2,k\in S_c}^{q}\|\{\text{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)\}_{C_k}\|_2-n\lambda_r\sqrt{q}\sum_{j=1,j\in S_r}^{p}\|\{\text{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)\}_{R_j}\|_2\end{aligned}$$

$$+n\lambda_c\sqrt{p}\sum_{k=2,k\notin S_c}^{q}\|\{\mathrm{vec}(\widehat{\mathbf{\Theta}})\}_{C_k}\|_2+n\lambda_r\sqrt{q}\sum_{j=1,j\notin S_r}^{p}\|\{\mathrm{vec}(\widehat{\mathbf{\Theta}})\}_{R_j}\|_2,$$

which implies that

$$\begin{aligned}
& \lambda_c\sqrt{p}\sum_{k=2,k\notin S_c}^{q}\|\{\mathrm{vec}(\widehat{\mathbf{\Theta}})\}_{C_k}\|_2+\lambda_r\sqrt{q}\sum_{j=1,j\notin S_r}^{p}\|\{\mathrm{vec}(\widehat{\mathbf{\Theta}})\}_{R_j}\|_2 \\
\leq\ & \lambda_c\sqrt{p}\sum_{k=2,k\in S_c}^{q}\|\{\mathrm{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)\}_{C_k}\|_2+\lambda_r\sqrt{q}\sum_{j=1,j\in S_r}^{p}\|\{\mathrm{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)\}_{R_j}\|_2 \\
& +n^{-1}\mathrm{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)^{\intercal}\boldsymbol{W}^{\intercal}\boldsymbol{\epsilon} \\
\leq\ & \lambda_c\sqrt{p}\sum_{k=2,k\in S_c}^{q}\|\{\mathrm{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)\}_{C_k}\|_2+\lambda_r\sqrt{q}\sum_{j=1,j\in S_r}^{p}\|\{\mathrm{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)\}_{R_j}\|_2 \\
& +n^{-1}\sum_{j=1,j\in Q_r}^{p}\|\mathrm{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)_{R_j}\|_2\|\boldsymbol{W}_{R_j(Q_c)}^{\intercal}\boldsymbol{\epsilon}\|_2 \\
& +n^{-1}\sum_{k=2,k\in Q_c}^{q}\|\mathrm{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)_{C_k}\|_2\|\boldsymbol{W}_{C_k(Q_r)}^{\intercal}\boldsymbol{\epsilon}\|_2 \\
\leq\ & \lambda_c\sqrt{p}\sum_{k=2,k\in S_c}^{q}\|\{\mathrm{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)\}_{C_k}\|_2+\lambda_r\sqrt{q}\sum_{j=1,j\in S_r}^{p}\|\{\mathrm{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)\}_{R_j}\|_2 \\
& +0.5\lambda_r\sqrt{q}\sum_{j=1}^{p}\|\mathrm{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)_{R_j}\|_2+0.5\lambda_c\sqrt{p}\sum_{k=2}^{q}\|\mathrm{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)_{C_k}\|_2.
\end{aligned}$$

Therefore, we have obtained that

$$\begin{aligned}
& \lambda_c\sqrt{p}\sum_{k=2,k\notin S_c}^{q}\|\{\mathrm{vec}(\widehat{\mathbf{\Theta}})\}_{C_k}\|_2+\lambda_r\sqrt{q}\sum_{j=1,j\notin S_r}^{p}\|\{\mathrm{vec}(\widehat{\mathbf{\Theta}})\}_{R_j}\|_2 \\
\leq\ & 3\lambda_c\sqrt{p}\sum_{k=2,k\in S_c}^{q}\|\{\mathrm{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)\}_{C_k}\|_2+3\lambda_r\sqrt{q}\sum_{j=1,j\in S_r}^{p}\|\{\mathrm{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}_0)\}_{R_j}\|_2,
\end{aligned}$$

which completes the proof of Lemma S4. □

Lemma S5. *Suppose $\boldsymbol{X}\in\mathbb{R}^{m\times p}$ is zero-mean sub-Gaussian with the parameters $(\mathbf{\Sigma},\sigma^2)$. Let $\mathbb{K}(s)\equiv(\boldsymbol{v}\in\mathbb{R}^p|\|\boldsymbol{v}\|_2\leq 1,\|\boldsymbol{v}\|_0\leq s)$. Then, there exists a constant $c>0$, such that*

$$\Pr\left[\sup_{\boldsymbol{v}\in\mathbb{K}(2s)}|\frac{\|\boldsymbol{X}\boldsymbol{v}\|_2^2}{n}-E\left(\frac{\|\boldsymbol{X}\boldsymbol{v}\|_2^2}{n}\right)|\geq t\right]\leq 2\exp\left\{-cn\min\left(\frac{t^2}{\sigma^4},\frac{t}{\sigma^2}\right)+2s\log(p)\right\}.$$

PROOF: A direct application of Lemma 15 of Loh and Wainwright (2012) completes the proof of Lemma S5. □

Lemma S6. *Suppose $n > \log(p)$. There is a constant $c > 0$, such that*

$$\begin{aligned}
\Pr\left\{\max_{j=1,\dots,p,j\in Q_r} n^{-1}\|\boldsymbol{W}_{R_j(Q_c)}^{\intercal}\boldsymbol{\epsilon}\|_2 > \sqrt{4b_1\log(p)/(cn)}\right\} &\leq 12\exp\left\{-b_1\log(p)\right\}, \\
\Pr\left\{\max_{k=1,\dots,q,k\in Q_c} n^{-1}\|\boldsymbol{W}_{C_k(Q_r)}^{\intercal}\boldsymbol{\epsilon}\|_2 > \sqrt{4b_1\log(q)/(cn)}\right\} &\leq 12\exp\left\{-b_1\log(q)\right\}.
\end{aligned}$$

PROOF: First, we note that

$$\|\boldsymbol{W}_{C_k(Q_r)}^{\intercal}\boldsymbol{\epsilon}\|_2 \leq \sup_{\boldsymbol{v}\in\mathbb{K}(2b_1)} |\boldsymbol{\epsilon}^{\intercal}\boldsymbol{W}\boldsymbol{v}|.$$

Letting $\boldsymbol{\phi}(\boldsymbol{v}) = \|\boldsymbol{v}\|_2^2/n - E(\|\boldsymbol{v}\|_2^2/n)$, we have

$$\boldsymbol{\epsilon}^{\intercal}\boldsymbol{W}\boldsymbol{v}/n = \frac{1}{2}\left\{\boldsymbol{\phi}(\boldsymbol{W}\boldsymbol{v}+\boldsymbol{\epsilon}) - \boldsymbol{\phi}(\boldsymbol{W}\boldsymbol{v}) - \boldsymbol{\phi}(\boldsymbol{\epsilon})\right\}.$$

Therefore, by Lemma S5, for some universal constant c, we have

$$\begin{aligned}
\Pr(n^{-1}\|\boldsymbol{W}_{C_k(Q_r)}^{\intercal}\boldsymbol{\epsilon}\|_2 > t) &\leq \Pr\left\{\sup_{\boldsymbol{v}\in\mathbb{K}(2b_1)} |\boldsymbol{\epsilon}^{\intercal}\boldsymbol{W}\boldsymbol{v}/n| > t\right\} \\
&\leq 6\exp\left\{-cn\min(t^2,t) + 2b_1\log(q)\right\}.
\end{aligned}$$

Letting $t = \sqrt{4b_1\log(q)/(cn)}$, we have

$$\begin{aligned}
\Pr\left(\max_{k=1,\dots,q,k\in Q_c} n^{-1}\|\boldsymbol{W}_{C_k(Q_r)}^{\intercal}\boldsymbol{\epsilon}\|_2 > t\right) &\leq \Pr\left\{\max_{k=1,\dots,q,k\in Q_c}\sup_{v\in\mathbb{K}(2b_1)} |\boldsymbol{\epsilon}^{\intercal}\boldsymbol{W}\boldsymbol{v}/n| > t\right\} \\
&\leq 12b_1\exp\left\{-cn\min(t^2,t) + 2b_1\log(q)\right\} \\
&\leq 12\exp\left\{-cn\min(t^2,t) + 3b_1\log(q)\right\} \\
&= 12\exp\left\{-b_1\log(q)\right\}.
\end{aligned}$$

Using the same argument, we can show that

$$\Pr\left\{\max_{j=1,\dots,p,j\in Q_r} n^{-1}\|\boldsymbol{W}_{R_j(Q_c)}^{\intercal}\boldsymbol{\epsilon}\|_2 > \sqrt{4b_1\log(p)/(cn)}\right\} \leq 12\exp\left\{-b_1\log(p)\right\}.$$

This completes the proof of Lemma S6. □

Lemma S7. *Suppose* $\boldsymbol{a} = (\boldsymbol{a}_g, g = 1, \ldots, \tilde{G})^\intercal$ *be an arbitrary group center, such that* $\|\boldsymbol{a}_g\|_0 \le b_1$ *and the group size* $\tilde{G} < \infty$. *Let* $P_{\widehat{\boldsymbol{\Theta}}}$ *denote the probability measure induced by* $\widehat{\boldsymbol{\Theta}}\boldsymbol{Z}_i$. *Suppose the conditions in Theorem 1 hold. Then,*

$$\left| \int \min_{1\le g\le \tilde{G}} \|\boldsymbol{v} - \boldsymbol{a}_g\|_2^2 dP_{\widehat{\boldsymbol{\Theta}}}(\boldsymbol{v}) - \int \min_{1\le g\le \tilde{G}} \|\boldsymbol{v} - \boldsymbol{a}_g\|_2^2 dP_{\boldsymbol{\Theta}_0}(\boldsymbol{v}) \right| = O_p\left\{ b_1 \sqrt{\log(pq)/n} \right\}.$$

PROOF: Let k_r, k_c denote the number of nonzero rows and columns of $\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0$. By Lemma S1 and Condition ((A2)), we have that $k_r + k_c \le 2b_1$. Let $\boldsymbol{E}_R \in \mathbb{R}^{k_r \times p}$ and $\boldsymbol{E}_C \in \mathbb{R}^{q\times k_c}$ be the selection matrix, so that $\boldsymbol{E}_R\boldsymbol{\Theta}$ corresponds to the nonzero rows, and $\boldsymbol{\Theta}\boldsymbol{E}_C$ the nonzero columns of $\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0$. Also the dimensions of $\widehat{\boldsymbol{\Theta}}\boldsymbol{Z}_i$ and $\boldsymbol{\Theta}_0\boldsymbol{Z}_i$ are both smaller than b_1. Then, for any given center $\boldsymbol{a}$,

$$\begin{aligned}
&\left| \int \|\boldsymbol{v} - \boldsymbol{a}\|_2^2 dP_{\widehat{\boldsymbol{\Theta}}}(\boldsymbol{v}) - \int \|\boldsymbol{v} - \boldsymbol{a}\|_2^2 dP_{\boldsymbol{\Theta}_0}(\boldsymbol{v}) \right| \\
= & \left| \frac{\partial \int \|\boldsymbol{v} - \boldsymbol{a}\|_2^2 dP_{\boldsymbol{\Theta}^*}(\boldsymbol{v})}{\partial \mathrm{vec}(\boldsymbol{E}_R\boldsymbol{\Theta}\boldsymbol{E}_C)^\intercal} \mathrm{vec}\left\{ \boldsymbol{E}_R(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0)\boldsymbol{E}_C \right\} \right| \\
\le & \left\| \frac{\partial \int \|\boldsymbol{v} - \boldsymbol{a}\|_2^2 dP_{\boldsymbol{\Theta}^*}(\boldsymbol{v})}{\partial \mathrm{vec}(\boldsymbol{E}_R\boldsymbol{\Theta}\boldsymbol{E}_C)^\intercal} \right\|_2 \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F \\
= & \left\| \int \|\boldsymbol{v} - \boldsymbol{a}\|_2^2 \frac{\partial f_{\boldsymbol{\Theta}^*}(\boldsymbol{v})}{\partial \mathrm{vec}(\boldsymbol{E}_R\boldsymbol{\Theta}\boldsymbol{E}_C)^\intercal} d\boldsymbol{v} \right\|_2 \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F \\
= & \left\| \int \|\boldsymbol{v}\|_2^2 \frac{\partial f_{\boldsymbol{\Theta}^*}(\boldsymbol{v})}{\partial \mathrm{vec}(\boldsymbol{E}_R\boldsymbol{\Theta}\boldsymbol{E}_C)^\intercal} d\boldsymbol{v} \right\|_2 \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F \\
& + \left\| 2\boldsymbol{a}^\intercal \int \boldsymbol{v} \frac{\partial f_{\boldsymbol{\Theta}^*}(\boldsymbol{v})}{\partial \mathrm{vec}(\boldsymbol{E}_R\boldsymbol{\Theta}\boldsymbol{E}_C)^\intercal} d\boldsymbol{v} \right\|_2 \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F \\
& + \left\| \int \|\boldsymbol{a}\|_2^2 \frac{\partial f_{\boldsymbol{\Theta}^*}(\boldsymbol{v})}{\partial \mathrm{vec}(\boldsymbol{E}_R\boldsymbol{\Theta}\boldsymbol{E}_C)^\intercal} d\boldsymbol{v} \right\|_2 \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F \\
\le & \|\boldsymbol{a}\|_2^2 \left\| \int \frac{\partial f_{\boldsymbol{\Theta}^*}(\boldsymbol{v})}{\partial \mathrm{vec}(\boldsymbol{E}_R\boldsymbol{\Theta}\boldsymbol{E}_C)^\intercal} d\boldsymbol{v} \right\|_2 \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F \\
& + 2\|\boldsymbol{a}\|_2 \left\| \int \boldsymbol{v} \frac{\partial f_{\boldsymbol{\Theta}^*}(\boldsymbol{v})}{\partial \mathrm{vec}(\boldsymbol{E}_R\boldsymbol{\Theta}\boldsymbol{E}_C)^\intercal} d\boldsymbol{v} \right\|_{op} \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F \\
& + \|\boldsymbol{v}\|_2^2 \left\| \int \frac{\partial f_{\boldsymbol{\Theta}^*}(\boldsymbol{v})}{\partial \mathrm{vec}(\boldsymbol{E}_R\boldsymbol{\Theta}\boldsymbol{E}_C)^\intercal} d\boldsymbol{v} \right\|_2 \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F \\
= & O_p\{b_1\sqrt{\log(pq)/n}\},
\end{aligned}$$

where $\|\cdot\|_{op}$ is the matrix operator norm, $\boldsymbol{\Theta}^*$ is a point between $\widehat{\boldsymbol{\Theta}}$ and $\boldsymbol{\Theta}_0$. The last equality holds by Theorem 1, and Condition (B3) that $|\int \|\boldsymbol{v}\|_2^2 \partial f_{\boldsymbol{\Theta}^*}(\boldsymbol{v})/\partial\mathrm{vec}(\boldsymbol{\Theta})^\intercal \boldsymbol{e}_j d\boldsymbol{v}|$, $|\partial f_{\boldsymbol{\Theta}^*}(\boldsymbol{v})/\partial\mathrm{vec}(\boldsymbol{\Theta})^\intercal \boldsymbol{e}_j|$, and $\|\int \boldsymbol{v}\partial f_{\boldsymbol{\Theta}^*}(\boldsymbol{v})/\partial\mathrm{vec}(\boldsymbol{\Theta})^\intercal \boldsymbol{e}_j d\boldsymbol{v}\|_2$ are finite. This lead to $\int \|\boldsymbol{v}\|_2^2 \partial f_{\boldsymbol{\Theta}^*}(\boldsymbol{v})/\partial\mathrm{vec}(\boldsymbol{\Theta})^\intercal \boldsymbol{e}_j d\boldsymbol{v} = O(1)$, $\int \partial f_{\boldsymbol{\Theta}^*}(\boldsymbol{v})/\partial\mathrm{vec}(\boldsymbol{\Theta})^\intercal \boldsymbol{e}_j d\boldsymbol{v} = O(1)$, $\|\int \boldsymbol{v}\partial f_{\boldsymbol{\Theta}^*}(\boldsymbol{v})/\partial\mathrm{vec}(\boldsymbol{\Theta})^\intercal \boldsymbol{e}_j d\boldsymbol{v}\|_2 = O(1)$, and that $\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0$ has at most $2b_1$ nonzero entries by Condition (A2) and Lemma S1.

This completes the proof of Lemma S7. □

S.2 Proof of Theorem 1

We first present a more general theorem, which establishes the relationship between the convergence order of $\widehat{\boldsymbol{\Theta}}$, the choice of the penalty parameters, and $\rho_-(s)$. Then the result of Theorem 1 follows immediately from Theorem S1 and Lemma S6 for some specific choices of the penalty parameters and $\rho_-(s)$.

Theorem S1. *Suppose Conditions (A1), (A2) and (A3) hold. Furthermore, suppose that*

$$\lambda_r\sqrt{q} \geq \frac{2}{n}\max_{j\in Q_r}\|\boldsymbol{W}_{R_j(Q_c)}^{\intercal}\boldsymbol{\epsilon}\|_2, \quad \lambda_c\sqrt{p} \geq \frac{2}{n}\max_{k\in Q_c}\|\boldsymbol{W}_{C_k(Q_r)}^{\intercal}\boldsymbol{\epsilon}\|_2, \tag{S.2}$$

almost surely. Then,

$$\|vec(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0)\|_2 \leq 3\left\{2.25\frac{\left(\lambda_r\sqrt{q}g_r+\lambda_c\sqrt{p}g_c\right)^2}{\min\{\lambda_r\sqrt{q},\lambda_c\sqrt{p}\}^2}+1\right\}^{1/2}\times\frac{\lambda_r\sqrt{q}g_r+\lambda_c\sqrt{p}g_c}{\rho_-(s)}.$$

PROOF: Recall that $\boldsymbol{\Delta} = \text{vec}(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0)$. Multiplying both sides of (S.1) by a vector $\boldsymbol{v}\in\mathbb{R}^{pq}$, such that $\boldsymbol{v}_G = \boldsymbol{\Delta}_G$ and $\boldsymbol{v}_{G^c} = \boldsymbol{0}$, where G is defined in Lemma (S3), we obtain that,

$$\frac{1}{n}\boldsymbol{\Delta}_G^{\intercal}\boldsymbol{W}_G^{\intercal}\boldsymbol{W}\boldsymbol{\Delta}-\boldsymbol{\Delta}_G^{\intercal}\boldsymbol{W}_G^{\intercal}\boldsymbol{\epsilon}+\lambda_r\sqrt{q}\sum_{j\in S_r\cup S_{r0}}\boldsymbol{\Delta}_{R_j}^{\intercal}\boldsymbol{\mu}_{jR_j}+\lambda_c\sqrt{p}\sum_{k\geq 2,k\in S_c\cup S_{c0}}\boldsymbol{\Delta}_{C_k}^{\intercal}\boldsymbol{\nu}_{kC_k}=0.$$

By the definition of G, we have that $|G|\leq d_0+2b_1+m$. Using the same argument as those leading to (S.1), we obtain that,

$$\begin{aligned}
&2\frac{1}{n}\boldsymbol{\Delta}_G^{\intercal}\boldsymbol{W}_G^{\intercal}\boldsymbol{W}\boldsymbol{\Delta}+\lambda_r\sqrt{q}\sum_{j\in S_{r0}}\|\text{vec}(\widehat{\boldsymbol{\Theta}})_{R_j}\|_2+\lambda_c\sqrt{p}\sum_{k\geq 2,k\in S_{c0}}\|\text{vec}(\widehat{\boldsymbol{\Theta}})_{C_k}\|_2\\
\leq\ & 3\lambda_r\sqrt{q}\sum_{j\in S_r}\|\boldsymbol{\Delta}_{R_j}\|_2+3\lambda_c\sqrt{p}\sum_{k\geq 2,k\in S_c}\|\boldsymbol{\Delta}_{C_k}\|_2.
\end{aligned}\tag{S.3}$$

Furthermore, denote $s=d_0+2b_1+m$. Then $|G|\leq s$ and

$$\begin{aligned}
&n^{-1}\boldsymbol{\Delta}_G^{\intercal}\boldsymbol{W}_G^{\intercal}\boldsymbol{W}\boldsymbol{\Delta}=n^{-1}\boldsymbol{\Delta}_G^{\intercal}\boldsymbol{W}_G^{\intercal}\boldsymbol{W}_G\boldsymbol{\Delta}_G+n^{-1}\boldsymbol{\Delta}_G^{\intercal}\boldsymbol{W}_G^{\intercal}\boldsymbol{W}_{G^c}\boldsymbol{\Delta}_{G^c}\\
=\ & n^{-1}\boldsymbol{\Delta}_G^{\intercal}\boldsymbol{W}_G^{\intercal}\boldsymbol{W}_G\boldsymbol{\Delta}_G+n^{-1}\sum_{j\notin S_r\cup S_{r0},k\notin S_c\cup S_{c0}}\boldsymbol{\Delta}_G^{\intercal}\boldsymbol{W}_G^{\intercal}\boldsymbol{W}_{R_j\cap C_k}\boldsymbol{\Delta}_{R_j\cap C_k}\\
\geq\ & n^{-1}\boldsymbol{\Delta}_G^{\intercal}\boldsymbol{W}_G^{\intercal}\boldsymbol{W}_G\boldsymbol{\Delta}_G\\
&-\widetilde{\rho}_+(|G|,2b_1+m)\|\boldsymbol{\Delta}_G\|_2\left\{\min_{u,v}(|S_{ru}|,|S_{cv}|)\right\}^{-1/2}\left\{\sum_{j\notin S_r}\|\boldsymbol{\Delta}_{R_j}\|_2+\sum_{k\notin S_c}\|\boldsymbol{\Delta}_{C_k}\|_2\right\}/2
\end{aligned}$$

$$
\begin{aligned}
&\geq n^{-1}\boldsymbol{\Delta}_G^{\intercal}\boldsymbol{W}_G^{\intercal}\boldsymbol{W}_G\boldsymbol{\Delta}_G - \widetilde{\rho}_+(|G|, 2b_1+m)\|\boldsymbol{\Delta}_G\|_2\left\{\min_{u,v}(|S_{ru}|, |S_{cv}|)\right\}^{-1/2}/2 \\
&\quad \times \min(\lambda_r\sqrt{q}, \lambda_c\sqrt{p})^{-1}\left\{\lambda_r\sqrt{q}\sum_{j\notin S_r}\|\boldsymbol{\Delta}_{R_j}\|_2 + \lambda_c\sqrt{p}\sum_{k\notin S_c}\|\boldsymbol{\Delta}_{C_k}\|_2\right\} \\
&\geq \rho_-(s)\|\boldsymbol{\Delta}_G\|_2^2 - 3\widetilde{\rho}_+(s, s-d_0)\left\{\min_{u,v}(|S_{ru}|, |S_{cv}|)\right\}^{-1/2}/2 \\
&\quad \times \min(\lambda_r\sqrt{q}, \lambda_c\sqrt{p})^{-1}\left\{\lambda_r\sqrt{q}|S_r| + \lambda_c\sqrt{p}|S_c|\right\}\|\boldsymbol{\Delta}_G\|_2^2 \\
&\geq \rho_-(s)\|\boldsymbol{\Delta}_G\|_2^2 - 3\{\rho_+(s) - \rho_-(2s-d_0)\}^{1/2}\{\rho_+(s-d) - \rho_-(2s-d_0)\}^{1/2}/2 \\
&\quad \times \min(\lambda_r\sqrt{q}, \lambda_c\sqrt{p})^{-1}\left\{\lambda_r\sqrt{q}|S_r| + \lambda_c\sqrt{p}|S_c|\right\}\|\boldsymbol{\Delta}_G\|_2^2 \\
&\geq 0.5\rho_-(s)\|\boldsymbol{\Delta}_G\|_2^2.
\end{aligned}
$$

The third line holds due to Lemma S3. The fifth line holds due to Lemma S4, the equality in (S.1), and the fact that each $\|\boldsymbol{\Delta}_{R_j}\|_2$ and $\|\boldsymbol{\Delta}_{C_k}\|_2$ is smaller than $\|\boldsymbol{\Delta}_G\|_2$. The sixth line holds by the definition of $\widetilde{\rho}_+(a,b)$, and the fact that $|G| \leq s$, $\rho_-(s) \leq \rho_-(G)$. Besides, because $R_{S_{ru}}(Q_c)$ or $R_{S_{cv}}(Q_r)$ must include at least $2b_1$ elements, therefore, $\{\min_{u,v}(|S_{ru}|, |S_{cv}|)\}$ is lower bounded by 1. The last inequality holds because of the fact that m can be any positive constant and by Condition (A3).

Therefore, combined with (S.3), we obtain that,

$$
\begin{aligned}
\|\boldsymbol{\Delta}_G\|_2^2 &\leq 3\rho_-(s)^{-1}\lambda_r\sqrt{q}\sum_{j\in S_r}\|\boldsymbol{\Delta}_{R_j}\|_2 + 3\rho_-(s)^{-1}\lambda_c\sqrt{p}\sum_{k\geq 2, k\in S_c}\|\boldsymbol{\Delta}_{C_k}\|_2 \\
&\leq 3\rho_-(s)^{-1}(\lambda_r\sqrt{q}|S_r| + \lambda_c\sqrt{p}|S_c|)\|\boldsymbol{\Delta}_G\|_2,
\end{aligned}
$$

In turn, we have

$$
\|\boldsymbol{\Delta}_G\|_2^2 \leq 9\rho_-(s)^{-2}(\lambda_r\sqrt{q}|S_r| + \lambda_c\sqrt{p}|S_c|)^2. \tag{S.4}
$$

Now we have that,

$$
\begin{aligned}
&\|\boldsymbol{\Delta}\|_2^2 - \|\boldsymbol{\Delta}_G\|_2^2 \leq \|\boldsymbol{\Delta}_{G^c}\|_2^2 \leq \sum_{j\notin S_r\cup S_{r0}, k\notin S_c\cup S_{c0}}\|\boldsymbol{\Delta}_{R_j\cap C_k}\|_2^2 \\
&\leq \min_{u,v}(4|S_{ru}|, 4|S_{cv}|)^{-1}\left(\sum_{j\notin S_r}\|\boldsymbol{\Delta}_{R_j}\|_2 + \sum_{k\notin S_c}\|\boldsymbol{\Delta}_{C_k}\|_2\right)^2 \\
&\leq 2.25\min_{u,v}(|S_{ru}|, |S_{cv}|)^{-1}\min\{\lambda_r\sqrt{q}, \lambda_c\sqrt{p}\}^{-2}\left(\lambda_r\sqrt{q}\sum_{j\in S_r}\|\boldsymbol{\Delta}_{R_j}\|_2 + \lambda_c\sqrt{p}\sum_{k\in S_c}\|\boldsymbol{\Delta}_{C_k}\|_2\right)^2 \\
&\leq 2.25\min_{u,v}(|S_{ru}|, |S_{cv}|)^{-1}\min\{\lambda_r\sqrt{q}, \lambda_c\sqrt{p}\}^{-2}\left(\lambda_r\sqrt{q}|S_r| + \lambda_c\sqrt{p}|S_c|\right)^2\|\boldsymbol{\Delta}_G\|_2^2.
\end{aligned}
$$

The third inequality holds due to Lemma S3, and the fourth inequality holds due to Lemma S4. Combine with (S.4), we obtain that,

$$
\|\boldsymbol{\Delta}\|_2^2 \leq 9\{2.25\min_{u,v}(|S_{ru}|, |S_{cv}|)^{-1}\min\{\lambda_r\sqrt{q}, \lambda_c\sqrt{p}\}^{-2}\left(\lambda_r\sqrt{q}|S_r| + \lambda_c\sqrt{p}|S_c|\right)^2 + 1\}
$$

$$
\begin{aligned}
& \times\rho_-(s)^{-2}(\lambda_r\sqrt{q}|S_r|+\lambda_c\sqrt{p}|S_c|)^2 \\
\leq\ & 9\{2.25\min\{\lambda_r\sqrt{q},\lambda_c\sqrt{p}\}^{-2}\left(\lambda_r\sqrt{q}|S_r|+\lambda_c\sqrt{p}|S_c|\right)^2+1\} \\
& \times\rho_-(s)^{-2}(\lambda_r\sqrt{q}g_r+\lambda_c\sqrt{p}g_c)^2.
\end{aligned}
$$

This completes the proof of Theorem S1. □

S.3 Proof of Theorem 2

PROOF: Let $\boldsymbol{a} = (\boldsymbol{a}_1^\intercal, \ldots, \boldsymbol{a}_G^\intercal)^\intercal$ be a vector of G group centers with fewer than Gb_1 number of nonzero entries in probability. Define

$$
W(\boldsymbol{a}, P_{\boldsymbol{\Theta}}) = \int \min_{1\leq g\leq G} \|\boldsymbol{v}-\boldsymbol{a}_g\|_2^2 dP_{\boldsymbol{\Theta}},
$$

Then, by Lemma A of Pollard et al. (1982), we have that $W(\boldsymbol{a}, P_{\widehat{\boldsymbol{\Theta}}})$ and $W(\boldsymbol{a}, P_{\boldsymbol{\Theta}_0})$ are both differentiable functions with respect to $\boldsymbol{a}$. Furthermore, $W_n(\boldsymbol{a},\boldsymbol{\Theta}) = \sum_{i=1}^n\{\min_{1\leq g\leq G}\|\boldsymbol{\Theta}\boldsymbol{Z}_i - \boldsymbol{a}_g\|_2^2\}/n$. Let $\widetilde{\boldsymbol{a}}$ and $\boldsymbol{a}_0$ be the minimizer for $W(\boldsymbol{a}, P_{\widehat{\boldsymbol{\Theta}}})$, and $W(\boldsymbol{a}, P_{\boldsymbol{\Theta}_0})$, respectively. By Lemma S7, we have that $\|\widetilde{\boldsymbol{a}}-\boldsymbol{a}_0\|_2 = o_p(1)$. By the consistency of the K-means algorithm as shown in Pollard (1981), we have $\|\widehat{\boldsymbol{a}}-\widetilde{\boldsymbol{a}}\|_2 = o_p(1)$. Henceforth, $\|\widehat{\boldsymbol{a}}-\boldsymbol{a}_0\|_2 = o_p(1)$.

Next, because $\widehat{\boldsymbol{a}}$ is the minimizer for $W(\widehat{\boldsymbol{a}}, P_{\widehat{\boldsymbol{\Theta}},n})$, we have that,

$$
\begin{aligned}
0 \ \geq\ & W_n(\widehat{\boldsymbol{a}},\widehat{\boldsymbol{\Theta}}) - W_n(\boldsymbol{a}_0,\widehat{\boldsymbol{\Theta}}) \\
=\ & W_n(\widehat{\boldsymbol{a}},\boldsymbol{\Theta}_0) - W_n(\boldsymbol{a}_0,\boldsymbol{\Theta}_0) + W_n(\widehat{\boldsymbol{a}},\widehat{\boldsymbol{\Theta}}) - W_n(\widehat{\boldsymbol{a}},\boldsymbol{\Theta}_0) + W_n(\boldsymbol{a}_0,\widehat{\boldsymbol{\Theta}}) - W_n(\boldsymbol{a}_0,\boldsymbol{\Theta}_0) \\
=\ & W_n(\widehat{\boldsymbol{a}},\boldsymbol{\Theta}_0) - W_n(\boldsymbol{a}_0,\boldsymbol{\Theta}_0) + W(\widehat{\boldsymbol{a}},P_{\widehat{\boldsymbol{\Theta}}}) - W(\widehat{\boldsymbol{a}},P_{\boldsymbol{\Theta}_0}) + W(\boldsymbol{a}_0,P_{\widehat{\boldsymbol{\Theta}}}) - W(\boldsymbol{a}_0,P_{\boldsymbol{\Theta}_0}) \\
& + o_p(n^{-1/2}) \\
=\ & -n^{-1/2}\boldsymbol{Z}_n^\intercal(\widehat{\boldsymbol{a}}-\boldsymbol{a}_0) + 1/2(\widehat{\boldsymbol{a}}-\boldsymbol{a}_0)^\intercal\frac{\partial^2 W(\boldsymbol{a}_0,P_{\boldsymbol{\Theta}_0})}{\partial\boldsymbol{a}\boldsymbol{a}^\intercal}(\widehat{\boldsymbol{a}}-\boldsymbol{a}_0) + O_p\left\{b_1\sqrt{\log(pq)/n}\right\} \\
& + o_p(n^{-1/2}) \\
\geq\ & \|\widehat{\boldsymbol{a}}-\boldsymbol{a}_0\|_2^2 D_{\min} - n^{-1/2}\boldsymbol{Z}_n^\intercal(\widehat{\boldsymbol{a}}-\boldsymbol{a}_0) + O_p\left\{b_1\sqrt{\log(pq)/n}\right\},
\end{aligned}
$$

where $\boldsymbol{Z}_n$ is a random vector with at most $2Gb_1$ nonzero elements, and is asymptotically normally distributed (Gänssler and Stute, 1979), and $2D_{\min}$ is the smallest eigenvalue of the matrix $\partial^2 W(\boldsymbol{a}, P_{\boldsymbol{\Theta}})/\partial\boldsymbol{a}\boldsymbol{a}^\intercal|_{(\boldsymbol{a},\boldsymbol{\Theta})=(\boldsymbol{a}_0,\boldsymbol{\Theta}_0)}$. By Condition (B1) that $W(\boldsymbol{a}, P_{\boldsymbol{\Theta}_0})$ is convex with a unique minimizer, i.e., $D_{\min} > 0$. The third line holds by the central limit theorem. The fourth line hold by Lemma S7 that both $W(\widehat{\boldsymbol{a}}, P_{\widehat{\boldsymbol{\Theta}}} - P_{\boldsymbol{\Theta}_0})$ and $W(\boldsymbol{a}_0, P_{\widehat{\boldsymbol{\Theta}}} - P_{\boldsymbol{\Theta}_0})$ are of order $O_p\left\{b_1\sqrt{\log(pq)/n}\right\}$, and by Lemma D and Theorem of Pollard (1981).

Therefore, we obtain that,

$$
\|\widehat{\boldsymbol{a}}-\boldsymbol{a}_0\|_2^2 D_{\min} \leq n^{-1/2}\boldsymbol{Z}_n^\intercal(\widehat{\boldsymbol{a}}-\boldsymbol{a}_0) + O_p\left\{b_1\sqrt{\log(pq)/n}\right\} \leq D_1 n^{-1/2}(Gb_1)^{1/2}\|\widehat{\boldsymbol{a}}-\boldsymbol{a}_0\|_2
$$

$$+O_p\left\{b_1\sqrt{\log(pq)/n}\right\},$$

for some constant D_1, which implies that,

$$\{\|\widehat{\boldsymbol{a}}-\boldsymbol{a}_0\|_2 - D_1(4n)^{-1/2}(Gb_1)^{1/2}\}^2 \leq O_p\left\{b_1\sqrt{\log(pq)/n}\right\} + D_1^2(4n)^{-1}Gb_1.$$

Therefore,

$$\|\widehat{\boldsymbol{a}}-\boldsymbol{a}_0\|_2 = O_p\left[\{b_1^2\log(pq)/n\}^{1/4} + n^{-1/2}(Gb_1)^{1/2}\right].$$

This completes the proof of Theorem 2. □

References

Gänssler, P. and Stute, W. (1979). Empirical processes: a survey of results for independent and identically distributed random variables. *The Annals of Probability*, pages 193–243.

Huang, J. and Zhang, T. (2010). The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664.

Pollard, D. (1981). Strong consistency of k-means clustering. *The Annals of Statistics*, pages 135–140.

Pollard, D. et al. (1982). A central limit theorem for k-means clustering. *The Annals of Probability*, 10(4):919–926.

Yin, P., Esser, E., and Xin, J. (2014). Ratio and difference of l_1 and l_2 norms and sparse representation with coherent dictionaries. *Communications in Information and Systems*, 14(2):87–109.