Appendix of "Optimal Sparse Linear Prediction for Block-missing Multi-modality Data without Imputation"

December 30, 2018

Proof of Proposition 1: By changing variables, the optimization problem can be rewritten as

$$\min_{\alpha_1,\alpha_2,\gamma} E[\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F^2]$$
$$s.t. \quad \hat{\mathbf{\Sigma}} = \alpha_1\tilde{\mathbf{\Sigma}}_I + \alpha_2\tilde{\mathbf{\Sigma}}_C + (1-\alpha_1)\gamma\mathbf{I_p}.$$

Using the facts that $\mathbf{\Sigma} = \mathbf{\Sigma}_I + \mathbf{\Sigma}_C$ and $E(\tilde{\mathbf{\Sigma}}_I) = \mathbf{\Sigma}_I$, we can rewrite the objective function as

$$\begin{aligned}
&E[\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F^2] = E[\|\alpha_1\tilde{\mathbf{\Sigma}}_I + \alpha_2\tilde{\mathbf{\Sigma}}_C + (1-\alpha_1)\gamma\mathbf{I_p} - \mathbf{\Sigma}\|_F^2] \\
=&E[\|\alpha_1\tilde{\mathbf{\Sigma}}_I + \alpha_2\tilde{\mathbf{\Sigma}}_C + (1-\alpha_1)\gamma\mathbf{I_p} - \alpha_1\mathbf{\Sigma}_I - (1-\alpha_1)\mathbf{\Sigma}_I - \mathbf{\Sigma}_C\|_F^2] \\
=&E[\|\alpha_1(\tilde{\mathbf{\Sigma}}_I - \mathbf{\Sigma}_I) + (1-\alpha_1)(\gamma\mathbf{I_p} - \mathbf{\Sigma}_I)\|_F^2] + E[\|\alpha_2\tilde{\mathbf{\Sigma}}_C - \mathbf{\Sigma}_C\|_F^2] \\
=&\alpha_1^2 E[\|\tilde{\mathbf{\Sigma}}_I - \mathbf{\Sigma}_I\|_F^2] + (1-\alpha_1)^2\|\gamma\mathbf{I_p} - \mathbf{\Sigma}_I\|_F^2 + E[\|\alpha_2\tilde{\mathbf{\Sigma}}_C - \mathbf{\Sigma}_C\|_F^2].
\end{aligned}$$

Therefore, the optimal value of γ can be obtained by minimizing $\|\gamma\mathbf{I_p} - \mathbf{\Sigma}_I\|_F^2$. Thus, the optimal value is $\gamma^* = \text{tr}(\mathbf{\Sigma}_I)/p = \text{tr}(\mathbf{\Sigma})/p$. The optimal value of α_2 can be obtained by minimizing $E[\|\alpha_2\tilde{\mathbf{\Sigma}}_C - \mathbf{\Sigma}_C\|_F^2]$. The optimal value is $\alpha_2^* = \frac{\|\mathbf{\Sigma}_C\|_F^2}{\|\mathbf{\Sigma}_C\|_F^2 + \delta_C^2}$. Replacing γ by its optimal value γ^* in the objective function and taking the derivative of the objective function with respect to α_1, we can find that the optimal value of α_1 is $\alpha_1^* = \frac{\theta^2}{\theta^2 + \delta_I^2}$. Thus, the optimal value of α_3 is $\alpha_3^* = \gamma^*(1 - \alpha_1^*) = \frac{\gamma^* \delta_I^2}{\theta^2 + \delta_I^2}$.

At the optimum, the value of the objective function is equal to $\frac{\delta_I^2 \theta^2}{\delta_I^2 + \theta^2} + \frac{\delta_C^2 \|\mathbf{\Sigma}_C\|_F^2}{\delta_C^2 + \|\mathbf{\Sigma}_C\|_F^2}$, which is less than $\delta_I^2 + \delta_C^2$. Since $E[\|\tilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F^2] = \delta_I^2 + \delta_C^2$, we have $E[\|\tilde{\mathbf{\Sigma}}^* - \mathbf{\Sigma}\|_F^2] \leq E[\|\tilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F^2]$. $\square$

To show the proof of Theorem 1, we first show the definition of the sub-Gaussian distribution and Lemma 1 of Ravikumar et al. (2011).

Definition 1: A zero-mean random variable Z is sub-Gaussian with parameter $L > 0$ if

$$E(\exp(tZ) \leq \exp(L^2 t^2/2)) \text{ for all } t \in R.$$

Lemma 1 (Ravikumar et al. (2011)) Consider a zero-mean random vector $(X_1, \ldots, X_p)$ with covariance $\mathbf{\Sigma}$ such that each $X_j/\sqrt{\sigma_{jj}}$ is sub-Gaussian with parameter L. Given n i.i.d. samples, the associated sample covariance $\hat{\mathbf{\Sigma}}$ satisfies the tail bound

$$P(|\hat{\sigma}_{jt} - \sigma_{jt}| \geq \delta) \leq 4\exp\{-\frac{n\delta^2}{128(1+4L^2)^2 \max_j(\sigma_{jj})^2}\},$$

for all $\delta \in (0, 8\max_j(\sigma_{jj})(1 + 4L^2))$.

Proof of Theorem 1: In our theoretical study, we assume that $\sigma_{jj} = 1$ for each j. Under the condition (**A1**), we know that $X_j/\sqrt{\sigma_{jj}}$ is sub-Gaussian with parameter L for each $j = 1, 2, \ldots, p$. The random variable $y/\sqrt{\text{var}(y)}$ is sub-Gaussian with parameter $L/\sqrt{\text{var}(y)}$.

We use the above **Lemma 1** and let $\delta = \nu_1\sqrt{\frac{\log p}{n_{jt}}}$. If

$$\nu_1\sqrt{(\log p)/n_{jt}} < 8\max_j(\sigma_{jj})(1 + 4L^2),$$

we can use the above ***Lemma 1*** and have the following result

$$
\begin{aligned}
P(|\tilde{\sigma}_{jt} - \sigma_{jt}| \geq \delta) &\leq 4\exp\{-\frac{n_{jt}\delta^2}{128(1+4L^2)^2 \max_j(\sigma_{jj})^2}\} \\
&= 4\exp\{-\frac{n_{jt}\cdot \nu_1^2 \frac{\log p}{n_{jt}}}{128(1+4L^2)^2 \max_j(\sigma_{jj})^2}\} \\
&= 4\exp\{-\frac{\nu_1^2}{128(1+4L^2)^2 \max_j(\sigma_{jj})^2}\log p\} \\
&= 4p^{-\frac{\nu_1^2}{128(1+4L^2)^2 \max_j(\sigma_{jj})^2}}.
\end{aligned}
$$

In our theoretical studies, we choose $\nu_1 = 8\sqrt{6}(1+4L^2)\max_j(\sigma_{jj}) = 8\sqrt{6}(1+4L^2)$ and $\nu_2 = 4$. If $\min_{j,t} n_{jt} > 6\log p$, we can check that

$$
\delta = 8\sqrt{6}(1+4L^2)\sqrt{\frac{\log p}{n_{jt}}}\max_j(\sigma_{jj}) < 8(1+4L^2)\max_j(\sigma_{jj}),
$$

and

$$
P(|\tilde{\sigma}_{jt} - \sigma_{jt}| \geq \nu_1\sqrt{\frac{\log p}{n_{jt}}}) \leq 4p^{-\frac{384(1+4L^2)^2 \max_j(\sigma_{jj})^2}{128(1+4L^2)^2 \max_j(\sigma_{jj})^2}} = \frac{4}{p^3} = \frac{\nu_2}{p^3},
$$

for any $j, t \in \{1, 2, \ldots, p\}$.

Hence, under condition (**A1**) and the condition $\min_{j,t} n_{jt} > 6\log p$, we have

$$
\begin{aligned}
\max_{j,t} P(|\tilde{\sigma}_{jt} - \sigma_{jt}| \geq \nu_1\sqrt{\frac{\log p}{n_{jt}}}) &\leq \frac{\nu_2}{p^3} \\
P(\|\tilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{max} \geq \nu_1\sqrt{\frac{\log p}{\min_{j,t} n_{jt}}}) &\leq \frac{\nu_2}{p^3}\cdot p^2 = \frac{\nu_2}{p},
\end{aligned}
$$

where the constants $\nu_1 = 8\sqrt{6}(1+4L^2)$ and $\nu_2 = 4$.

In addition, we know that each $X_j/\sqrt{\sigma_{jj}}$ and the random variable $y/\sqrt{\text{var}(y)}$ are also sub-Gaussian with parameter $\frac{L}{\min\{1,\sqrt{\text{var}(y)}\}}$. Let $\nu_3 = 16(1+4\frac{L^2}{\min\{\text{var}(y),1\}})\max\{\text{var}(y),1\}$

and $\nu_4 = 4$. If $\min_{j,t} n_{jt} > 6\log p$, we can check that

$$\min_j n_j \geq \min_{j,t} n_{jt} > 6\log p > 4\log p,$$
$$\nu_3\sqrt{\frac{\log p}{n_j}} < 8(1+4\frac{L^2}{\min\{\text{var}(y),1\}})\max\{\text{var}(y),1\} \text{ for each } j=1,2,\ldots,p.$$

Using the above ***Lemma 1***, we have

$$\max_j P(|\tilde{c}_j - c_j| \geq \nu_3\sqrt{\frac{\log p}{n_j}}) \leq \frac{\nu_4}{p^2},$$
$$P(\|\tilde{C} - C\|_{max} \geq \nu_3\sqrt{\frac{\log p}{\min_j n_j}}) \leq \frac{\nu_4}{p^2}\cdot p = \frac{\nu_4}{p}.$$

□

Proof of Theorem 2: Denote $\hat{\mathbf{\Sigma}} = \alpha_1\tilde{\mathbf{\Sigma}}_I + \alpha_2\tilde{\mathbf{\Sigma}}_C + (1-\alpha_1)\mathbf{I_p}$, where $1-\alpha_1 = O(\sqrt{\log p/\min_j n_j})$ and $1-\alpha_2 = O(\sqrt{\log p/\min_{j,t} n_{jt}})$. We first show the convergence rate of $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{max}$. Based on the definition of $\hat{\mathbf{\Sigma}} = (\hat{\sigma}_{jt})_{j,t=1}^p$, we know that

$$\hat{\sigma}_{jt} - \sigma_{jt} = \begin{cases} 0 & \text{if } j = t; \\ \alpha_1\tilde{\sigma}_{jt} - \sigma_{jt} & \text{if } j \neq t \text{ (} j \text{ and } t \text{ are in the same modality);} \\ \alpha_2\tilde{\sigma}_{jt} - \sigma_{jt} & \text{if } j \neq t \text{ (} j \text{ and } t \text{ are in different modalities).} \end{cases}$$

Thus, if $j \neq t$ and the predictors j and t are in the same modality, with probability at least $1-\nu_2/p^3$, we have

$$\begin{aligned} |\hat{\sigma}_{jt} - \sigma_{jt}| = |\alpha_1\tilde{\sigma}_{jt} - \sigma_{jt}| &\leq \alpha_1|\tilde{\sigma}_{jt} - \sigma_{jt}| + (1-\alpha_1)|\sigma_{jt}| \\ &\leq \alpha_1|\tilde{\sigma}_{jt} - \sigma_{jt}| + 1 - \alpha_1 \\ \text{(by Theorem 1)} &\leq \alpha_1\nu_1\sqrt{\log p/\min_j n_j} + 1 - \alpha_1 \\ &\leq \nu_1\sqrt{\log p/\min_j n_j} + 1 - \alpha_1. \end{aligned}$$

Similarly, if $j \neq t$ and the predictors j and t are in different modalities, with probability at least $1 - \nu_2/p^3$, we have

$$|\hat{\sigma}_{jt} - \sigma_{jt}| = |\alpha_2 \tilde{\sigma}_{jt} - \sigma_{jt}| \leq \nu_1 \sqrt{\log p / \min_{j,t} n_{jt}} + 1 - \alpha_2.$$

Therefore, there exists two constants ν_1' and ν_2 such that

$$P(\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{max} \geq \nu_1' \sqrt{\log p / \min_{j,t} n_{jt}}) \leq \nu_2/p.$$

Denote events $\mathcal{A} = \{\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{max} \leq \nu_1' \sqrt{\log p / \min_{j,t} n_{jt}}\}$ and $\mathcal{B} = \{\|\tilde{C} - C\|_{max} \leq \nu_3 \sqrt{\log p / \min_j n_j}\}$. From Theorem 1 and the above convergence rate of $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{max}$, we have $P(\mathcal{A} \cap \mathcal{B}) \geq 1 - (\nu_2 + \nu_4)/p$. In events $\mathcal{A}$ and $\mathcal{B}$, we have

$$\begin{aligned} \|\tilde{C} - \hat{\mathbf{\Sigma}}\beta^0\|_{max} &\leq \|\tilde{C} - C\|_{max} + \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{max} \|\beta^0\|_1 \\ &\leq (\nu_3 + \nu_1' \|\beta^0\|_1) \sqrt{\log p / \min_{j,t} n_{jt}}. \end{aligned}$$

Therefore, we have $\|\tilde{C} - \hat{\mathbf{\Sigma}}\beta^0\|_{max} = O_p(\|\beta^0\|_1 \sqrt{\log p / \min_{j,t} n_{jt}})$. Next, we show that $\|\tilde{\beta} - \beta^0\|_2 = O_p(\sqrt{s}\lambda)$. By the KKT condition, the solution $\tilde{\beta}$ of (4) satisfies

$$\|\hat{\mathbf{\Sigma}}\tilde{\beta} - \tilde{C}\|_{max} \leq \lambda.$$

On the other hand, since $\tilde{\beta}$ is the solution to (4), we have

$$\frac{1}{2}\tilde{\beta}^T \hat{\mathbf{\Sigma}} \tilde{\beta} - \tilde{C}^T \tilde{\beta} + \lambda \|\tilde{\beta}\|_1 \leq \frac{1}{2} \beta^{0^T} \hat{\mathbf{\Sigma}} \beta^0 - \tilde{C}^T \beta^0 + \lambda \|\beta^0\|_1.$$

Therefore,

$$\begin{aligned}
2\lambda\|\tilde{\beta}\|_1 &\leq {\beta^0}^T\hat{\mathbf{\Sigma}}\beta^0 - \tilde{\beta}^T\hat{\mathbf{\Sigma}}\tilde{\beta} + 2\tilde{C}^T(\tilde{\beta}-\beta^0) + 2\lambda\|\beta^0\|_1 \\
&= (\beta^0-\tilde{\beta})^T\hat{\mathbf{\Sigma}}\beta^0 + \tilde{\beta}^T\hat{\mathbf{\Sigma}}(\beta^0-\tilde{\beta}) + 2\tilde{C}^T(\tilde{\beta}-\beta^0) + 2\lambda\|\beta^0\|_1 \\
&= (2\tilde{C} - \hat{\mathbf{\Sigma}}\tilde{\beta} - \hat{\mathbf{\Sigma}}\beta^0)^T(\tilde{\beta}-\beta^0) + 2\lambda\|\beta^0\|_1 \\
&\leq (\|\tilde{C} - \hat{\mathbf{\Sigma}}\tilde{\beta}\|_{max} + \|\tilde{C} - \hat{\mathbf{\Sigma}}\beta^0\|_{max}) \cdot \|\beta^0-\tilde{\beta}\|_1 + 2\lambda\|\beta^0\|_1 \\
&\leq (\lambda + \|\tilde{C} - \hat{\mathbf{\Sigma}}\beta^0\|_{max}) \cdot \|\beta^0-\tilde{\beta}\|_1 + 2\lambda\|\beta^0\|_1.
\end{aligned}$$

If the tuning parameter $\lambda = 2\|\tilde{C} - \hat{\mathbf{\Sigma}}\beta^0\|_{max}$, we have

$$2\lambda\|\tilde{\beta}\|_1 \leq 1.5\lambda\|\beta^0-\tilde{\beta}\|_1 + 2\lambda\|\beta^0\|_1.$$

Hence,

$$\|\tilde{\beta}-\beta^0\|_1 \leq 4(\|\tilde{\beta}-\beta^0\|_1 + \|\beta^0\|_1 - \|\tilde{\beta}\|_1).$$

Let $\delta = \tilde{\beta} - \beta^0$. Since for $j \in J^c$, $|\tilde{\beta}_j - \beta^0_j| + |\beta^0_j| - |\tilde{\beta}_j| = 0$, it holds that

$$\|\tilde{\beta}-\beta^0\|_1 = \|\delta_J\|_1 + \|\delta_{J^c}\|_1 \leq 4(\|\delta_J\|_1 + \|\beta^0_J\|_1 - \|\tilde{\beta}_J\|_1) \leq 8\|\delta_J\|_1.$$

Hence, $\|\delta_{J^c}\|_1 \leq 7\|\delta_J\|_1$. Furthermore, under the condition (**A2**), if the event $\mathcal{A}$ occurs, we have

$$\frac{\delta^T\hat{\mathbf{\Sigma}}\delta}{\delta^T\delta} = \frac{\delta^T\mathbf{\Sigma}\delta}{\delta^T\delta} + \frac{\delta^T(\hat{\mathbf{\Sigma}}-\mathbf{\Sigma})\delta}{\delta^T\delta} \geq m - 64s\|\hat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|_{max} \geq m - 64s\nu_1'\sqrt{\log p/\min_{j,t} n_{jt}}.$$

If we assume that $s\nu_1'\sqrt{\log p/\min_{j,t} n_{jt}} = o(1)$ or $\min_{j,t} n_{jt} > (128\nu_1'/m)^2(s^2\log p)$, we have

$$\frac{\delta^T\hat{\mathbf{\Sigma}}\delta}{\delta^T\delta} \geq m - m/2 = m/2 > 0, \tag{S1}$$

for sufficiently large s, p, and $\min_{jt} n_{jt}$. On the other hand, we have

$$\delta^T \hat{\mathbf{\Sigma}} \delta \leq \|\hat{\mathbf{\Sigma}}(\tilde{\beta} - \beta^0)\|_{max} \|\delta\|_1 \leq (\|\hat{\mathbf{\Sigma}}\tilde{\beta} - \tilde{C}\|_{max} + \|\tilde{C} - \hat{\mathbf{\Sigma}}\beta^0\|_{max})\|\delta\|_1$$
$$\leq (\lambda + \|\tilde{C} - \hat{\mathbf{\Sigma}}\beta^0\|_{max})\|\delta\|_1 = 1.5\lambda\|\delta\|_1. \quad \text{(S2)}$$

Therefore, by (S1) and (S2), we have $\frac{m}{2}\|\delta\|_2^2 \leq \delta^T \hat{\mathbf{\Sigma}} \delta \leq 1.5\lambda\|\delta\|_1 \leq 12\lambda\|\delta_J\|_1 \leq 12\lambda\sqrt{s}\|\delta\|_2$. Hence, $\|\delta\|_2 \leq 24\lambda\sqrt{s}/m$.

Therefore $\|\tilde{\beta} - \beta^0\|_2 = O_p(\sqrt{s}\lambda) = O_p(\|\beta^0\|_1 \sqrt{s \log p / \min_{j,t} n_{jt}})$. This completes the proof. □

Proof of Theorem 3: For each $j, t \in \{1, 2, \ldots, p\}$, we know that $\breve{\sigma}_{jt} = \sigma_{jt} = 1$ if $j = t$, and it equals to the solution to $\sum_{i \in S_{jt}} \psi_{H_{jt}}(x_{ij}x_{it} - \mu) = 0$ otherwise. Under condition (**A3**) and the assumption that $\min_{j,t} n_{jt} \geq 24 \log p$, let $H_{jt} = \frac{Q_1}{12}\sqrt{n_{jt}/\log p}$ for each $j, t \in \{1, 2, \ldots, p\}$, by Theorem 5 in Fan et al. (2016), we know that for all $j, t \in \{1, 2, \ldots, p\}$,

$$P(|\breve{\sigma}_{jt} - \sigma_{jt}| \geq Q_1\sqrt{\frac{\log p}{n_{jt}}}) \leq \frac{2}{p^3}.$$

Therefore, we have

$$\max_{j,t} P(|\breve{\sigma}_{jt} - \sigma_{jt}| \geq Q_1\sqrt{\frac{\log p}{n_{jt}}}) \leq \frac{2}{p^3}.$$

Furthermore, for each $j, t \in \{1, 2, \ldots, p\}$, we have

$$P(|\breve{\sigma}_{jt} - \sigma_{jt}| \geq Q_1\sqrt{\frac{\log p}{\min_{j,t} n_{jt}}}) \leq P(|\breve{\sigma}_{jt} - \sigma_{jt}| \geq Q_1\sqrt{\frac{\log p}{n_{jt}}}) \leq \frac{2}{p^3}.$$

Thus,

$$P(\|\breve{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{max} \geq Q_1\sqrt{\frac{\log p}{\min_{j,t} n_{jt}}}) = P(\max_{j,t} |\breve{\sigma}_{jt} - \sigma_{jt}| \geq Q_1\sqrt{\frac{\log p}{\min_{j,t} n_{jt}}}) \leq \frac{2}{p}.$$

In addition, for each $j \in \{1, 2, \ldots, p\}$ and $i \in S_j$, we have

$$Var(x_{ij}y_i) \le E(x_{ij}^2 y_i^2) = \sqrt{E(x_{ij}^4)E(y_i^4)} \le 2(Q_1 + Q_2)^2.$$

Let $H_j = (Q_1 + Q_2)\sqrt{n_j/\log p}$ for each $j \in \{1, 2 \ldots, p\}$. Using Theorem 5 in Fan et al. (2016), we know that for each $j \in \{1, 2, \ldots, p\}$,

$$P(|\breve{c}_j - c_j| \ge 8(Q_1 + Q_2)\sqrt{\frac{\log p}{n_j}}) \le \frac{2}{p^2}.$$

Therefore, we have

$$\max_j P(|\breve{c}_j - c_j| \ge 8(Q_1 + Q_2)\sqrt{\frac{\log p}{n_j}}) \le \frac{2}{p^2}.$$

Furthermore, since

$$P(|\breve{c}_j - c_j| \ge 8(Q_1 + Q_2)\sqrt{\frac{\log p}{\min_j n_j}}) \le P(|\breve{c}_j - c_j| \ge 8(Q_1 + Q_2)\sqrt{\frac{\log p}{n_j}}),$$

we know that

$$P(\|\breve{C} - C\|_{max} \ge 8(Q_1 + Q_2)\sqrt{\frac{\log p}{\min_j n_j}}) \le p \cdot \frac{2}{p^2} = \frac{2}{p}.$$

This completes the proof. □

Proof of Theorem 4: Under the conditions $\max_{1 \le j \le p} \mathrm{E}(|X_j|^{4\ell}) \le T$ and $\mathrm{E}(\epsilon_1^{2\ell}) = \mathrm{E}(\epsilon_2^{2\ell}) = \cdots = \mathrm{E}(\epsilon_n^{2\ell}) \le T$, by Theorem 2 in Whittle (1960), we have

$$\max_{j,t} P(|\tilde{\sigma}_{jt} - \sigma_{jt}| \ge \frac{d_1}{2T}\sqrt{\frac{p}{n_{jt}}}) \le \frac{d_2}{p^{2h}}.$$

Furthermore, for each $j, t \in \{1, 2, \ldots, p\}$, we have

$$P(|\tilde{\sigma}_{jt} - \sigma_{jt}| \geq \frac{d_1}{2T}\sqrt{\frac{p}{\min_{j,t} n_{jt}}}) \leq P(|\tilde{\sigma}_{jt} - \sigma_{jt}| \geq \frac{d_1}{2T}\sqrt{\frac{p}{n_{jt}}}) \leq \frac{d_2}{p^{2h}}.$$

Thus,

$$\begin{aligned} P(\|\tilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{max} \geq \frac{d_1}{2T}\sqrt{\frac{p}{\min_{j,t} n_{jt}}}) &= P(\max_{j,t} |\tilde{\sigma}_{jt} - \sigma_{jt}| \geq \frac{d_1}{2T}\sqrt{\frac{p}{\min_{j,t} n_{jt}}}) \\ &\leq \frac{d_2}{p^{2h-2}}. \end{aligned}$$

In addition, since $E(\tilde{C}) = C$, using the Theorem 2 in Whittle (1960), for all j, we have

$$P(|\tilde{c}_j - c_j| \geq \frac{d_3}{2T}\sqrt{\frac{p}{n_j}}) \leq \frac{d_4}{p^{2h-1}},$$

where d_3 and d_4 are two positive constants.

Therefore, we have

$$\max_j P(|\tilde{c}_j - c_j| \geq \frac{d_3}{2T}\sqrt{\frac{p}{n_j}}) \leq \frac{d_4}{p^{2h-1}}.$$

Furthermore, since

$$P(|\tilde{c}_j - c_j| \geq \frac{d_3}{2T}\sqrt{\frac{p}{\min_j n_j}}) \leq P(|\tilde{c}_j - c_j| \geq \frac{d_3}{2T}\sqrt{\frac{p}{n_j}}),$$

we know that

$$P(\|\tilde{C} - C\|_{max} \geq \frac{d_3}{2T}\sqrt{\frac{p}{\min_j n_j}}) \leq p \cdot \frac{d_4}{p^{2h-1}} = \frac{d_4}{p^{2h-2}}.$$

This completes the proof. □

Proof of Theorem 5: The proof is almost the same as the proof of Theorem 2.

We use the same technique to show that

$$P(\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{max} \geq Q_1' \sqrt{\log p / \min_{j,t} n_{jt}}) \leq 2/p.$$

Then, in events $\mathcal{A} = \{\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{max} \leq Q_1' \sqrt{\log p / \min_{j,t} n_{jt}})\}$ and $\mathcal{B} = \{\|\breve{C} - C\|_{max} \leq 8(Q_1 + Q_2)\sqrt{\log p / \min_j n_j}\}$, for sufficiently large s, p, and $\min_{j,t} n_{jt}$, we show that $\|\breve{\beta} - \beta^0\|_2 \leq 24\lambda\sqrt{s}/m$ with probability at least $1 - 4/p$. This completes the proof. □

Proof of Theorem 6: By the KKT condition, we know that $\tilde{\beta}$ is a solution to (4) if and only if there exists a subgradient $\gamma \in R^p$ such that

$$\tilde{C} - \hat{\mathbf{\Sigma}}\tilde{\beta} = \lambda\gamma,$$

where for each $j \in \{1, 2, \ldots, p\}$, $\gamma_j = \text{sign}(\tilde{\beta}_j)$ if $\tilde{\beta}_j \neq 0$, and $\gamma_j \in [-1, 1]$ if $\tilde{\beta}_j = 0$.

We can construct a point $\tilde{\beta} \in R^p$ by letting $\tilde{\beta}_J = (\hat{\mathbf{\Sigma}}_{JJ})^{-1}\tilde{C}_J - \lambda(\hat{\mathbf{\Sigma}}_{JJ})^{-1} \cdot \text{sign}(\beta_J^0)$ and $\tilde{\beta}_{J^c} = 0$. Define events $\mathcal{A}_1 = \{\|\tilde{\beta}_J - \beta_J^0\|_{max} < \beta_{min}^0\}$ and $\mathcal{A}_2 = \{\|\tilde{C}_{J^c} - \hat{\mathbf{\Sigma}}_{J^cJ}\tilde{\beta}_J\|_{max} \leq \lambda\}$. If events $\mathcal{A}_1$ and $\mathcal{A}_2$ hold, we can check that $\tilde{\beta}$ is a solution to (4) and $\text{sign}(\tilde{\beta}) = \text{sign}(\beta^0)$. To prove the theorem, we only need to show that $P(\mathcal{A}_1) \longrightarrow 1$ and $P(\mathcal{A}_2) \longrightarrow 1$, as $\min_{j,t} n_{jt} \to \infty$ and $p \to \infty$.

Step 1: show the upper bound of $\|(\hat{\mathbf{\Sigma}}_{JJ})^{-1}\|_\infty$.

Denote $V = \|(\mathbf{\Sigma}_{JJ})^{-1}\|_\infty$. Since

$$\begin{aligned}
&\|(\hat{\mathbf{\Sigma}}_{JJ})^{-1} - (\mathbf{\Sigma}_{JJ})^{-1}\|_\infty \leq \|(\mathbf{\Sigma}_{JJ})^{-1}\|_\infty \cdot \|(\hat{\mathbf{\Sigma}}_{JJ})^{-1}\|_\infty \cdot \|\hat{\mathbf{\Sigma}}_{JJ} - \mathbf{\Sigma}_{JJ}\|_\infty \\
&\leq \|(\mathbf{\Sigma}_{JJ})^{-1}\|_\infty \cdot (\|(\mathbf{\Sigma}_{JJ})^{-1}\|_\infty + \|(\hat{\mathbf{\Sigma}}_{JJ})^{-1} - (\mathbf{\Sigma}_{JJ})^{-1}\|_\infty) \cdot \|\hat{\mathbf{\Sigma}}_{JJ} - \mathbf{\Sigma}_{JJ}\|_\infty \\
&= V(V + \|(\hat{\mathbf{\Sigma}}_{JJ})^{-1} - (\mathbf{\Sigma}_{JJ})^{-1}\|_\infty) \cdot \|\hat{\mathbf{\Sigma}}_{JJ} - \mathbf{\Sigma}_{JJ}\|_\infty,
\end{aligned}$$

we have,

$$\|(\hat{\mathbf{\Sigma}}_{JJ})^{-1} - (\mathbf{\Sigma}_{JJ})^{-1}\|_\infty \leq \frac{V^2\|\hat{\mathbf{\Sigma}}_{JJ} - \mathbf{\Sigma}_{JJ}\|_\infty}{1 - V\|\hat{\mathbf{\Sigma}}_{JJ} - \mathbf{\Sigma}_{JJ}\|_\infty} \leq \frac{sV^2\|\hat{\mathbf{\Sigma}}_{JJ} - \mathbf{\Sigma}_{JJ}\|_{max}}{1 - sV\|\hat{\mathbf{\Sigma}}_{JJ} - \mathbf{\Sigma}_{JJ}\|_{max}},$$

and

$$\|(\hat{\mathbf{\Sigma}}_{JJ})^{-1}\|_\infty \le V + \frac{sV^2\|\hat{\mathbf{\Sigma}}_{JJ} - \mathbf{\Sigma}_{JJ}\|_{max}}{1 - sV\|\hat{\mathbf{\Sigma}}_{JJ} - \mathbf{\Sigma}_{JJ}\|_{max}} = \frac{V}{1 - sV\|\hat{\mathbf{\Sigma}}_{JJ} - \mathbf{\Sigma}_{JJ}\|_{max}}.$$

Step 2: show the upper bound of $\|\tilde{C}_J - \hat{\mathbf{\Sigma}}_{JJ}\beta_J^0\|_{max}$. We have

$$\begin{aligned}\|\tilde{C}_J - \hat{\mathbf{\Sigma}}_{JJ}\beta_J^0\|_{max} &\le \|\tilde{C}_J - C_J\|_{max} + \|(\mathbf{\Sigma}_{JJ} - \hat{\mathbf{\Sigma}}_{JJ})\beta_J^0\|_{max} \\ &\le \|\tilde{C}_J - C_J\|_{max} + \|\mathbf{\Sigma}_{JJ} - \hat{\mathbf{\Sigma}}_{JJ}\|_\infty \|\beta_J^0\|_{max} \\ &\le \|\tilde{C}_J - C_J\|_{max} + s\beta_{max}^0\|\hat{\mathbf{\Sigma}}_{JJ} - \mathbf{\Sigma}_{JJ}\|_{max}.\end{aligned}$$

Step 3: show that $P(\mathcal{A}_1) \longrightarrow 1$ as $\min_{j,t} n_{jt} \to \infty$ and $p \to \infty$.

Define events $\mathcal{A}_3 = \{\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{max} \le \nu_1' \sqrt{\log p / \min_{j,t} n_{jt}}$ and $\mathcal{A}_4 = \{\|\tilde{C} - C\|_{max} \le \nu_3 \sqrt{\log p / \min_j n_j}\}$. By Theorem 1 and Theorem 2, we know that $P(\mathcal{A}_3) \longrightarrow 1$ and $P(\mathcal{A}_4) \longrightarrow 1$ as $p \to \infty$. If events $\mathcal{A}_3$ and $\mathcal{A}_4$ occur, since $\frac{1+s\beta_{max}^0}{\lambda}\sqrt{\frac{\log p}{\min_{j,t} n_{jt}}} \longrightarrow 0$, if $sV\sqrt{\log p / \min_{j,t} n_{jt}} \longrightarrow 0$ or the following condition holds

$$\|(\mathbf{\Sigma}_{JJ})^{-1}\|_\infty \cdot \sqrt{\frac{s^2 \log p}{\min_{j,t} n_{jt}}} \le \frac{\eta}{\nu_1'(4+\eta)},$$

we have

$$\begin{aligned}&\|\tilde{\beta}_J - \beta_J^0\|_{max} = \|(\hat{\mathbf{\Sigma}}_{JJ})^{-1}\tilde{C}_J - \lambda(\hat{\mathbf{\Sigma}}_{JJ})^{-1} \cdot sign(\beta_J^0) - \beta_J^0\|_{max} \\ &\le \|(\hat{\mathbf{\Sigma}}_{JJ})^{-1}\tilde{C}_J - \beta_J^0\|_{max} + \lambda\|(\hat{\mathbf{\Sigma}}_{JJ})^{-1}\|_\infty \\ &\le (\|\tilde{C}_J - \hat{\mathbf{\Sigma}}_{JJ}\beta_J^0\|_{max} + \lambda) \cdot \|(\hat{\mathbf{\Sigma}}_{JJ})^{-1}\|_\infty \\ &\le (\|\tilde{C}_J - C_J\|_{max} + s\beta_{max}^0\|\hat{\mathbf{\Sigma}}_{JJ} - \mathbf{\Sigma}_{JJ}\|_{max} + \lambda) \cdot \frac{V}{1 - sV\|\hat{\mathbf{\Sigma}}_{JJ} - \mathbf{\Sigma}_{JJ}\|_{max}} \\ &\le \frac{2\lambda V}{1 - sV\|\hat{\mathbf{\Sigma}}_{JJ} - \mathbf{\Sigma}_{JJ}\|_{max}} \le 4\lambda V,\end{aligned}$$

for sufficiently large p and $\min_{j,t} n_{jt}$. Therefore, if $\lambda V / \beta_{min}^0 \longrightarrow 0$, we have $P(\mathcal{A}_1) = 1 - P(\{\|\tilde{\beta}_J - \beta_J^0\|_{max} \ge \beta_{min}^0\}) \longrightarrow 1$.

Step 4: show the upper bound of $\|\hat{\mathbf{\Sigma}}_{J^cJ}(\hat{\mathbf{\Sigma}}_{JJ})^{-1} - \mathbf{\Sigma}_{J^cJ}(\mathbf{\Sigma}_{JJ})^{-1}\|_\infty$. We have

$$\begin{aligned}
&\|\hat{\mathbf{\Sigma}}_{J^cJ}(\hat{\mathbf{\Sigma}}_{JJ})^{-1} - \mathbf{\Sigma}_{J^cJ}(\mathbf{\Sigma}_{JJ})^{-1}\|_\infty \\
&\leq \|\mathbf{\Sigma}_{J^cJ}((\hat{\mathbf{\Sigma}}_{JJ})^{-1} - (\mathbf{\Sigma}_{JJ})^{-1})\|_\infty + \|(\hat{\mathbf{\Sigma}}_{J^cJ} - \mathbf{\Sigma}_{J^cJ})(\hat{\mathbf{\Sigma}}_{JJ})^{-1}\|_\infty \\
&\leq \|\mathbf{\Sigma}_{J^cJ}(\mathbf{\Sigma}_{JJ})^{-1}\|_\infty \cdot \|\mathbf{\Sigma}_{JJ} - \hat{\mathbf{\Sigma}}_{JJ}\|_\infty \cdot \|(\hat{\mathbf{\Sigma}}_{JJ})^{-1}\|_\infty \\
&\qquad\qquad + \|(\hat{\mathbf{\Sigma}}_{JJ})^{-1}\|_\infty \cdot \|\hat{\mathbf{\Sigma}}_{J^cJ} - \mathbf{\Sigma}_{J^cJ}\|_\infty \\
&\leq \|(\hat{\mathbf{\Sigma}}_{JJ})^{-1}\|_\infty \cdot (\|\hat{\mathbf{\Sigma}}_{JJ} - \mathbf{\Sigma}_{JJ}\|_\infty + \|\hat{\mathbf{\Sigma}}_{J^cJ} - \mathbf{\Sigma}_{J^cJ}\|_\infty) \leq \frac{2sV\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{max}}{1 - sV\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{max}}.
\end{aligned}$$

Step 5: show that $P(\mathcal{A}_2) \longrightarrow 1$ as $\min_{j,t} n_{jt} \to \infty$ and $p \to \infty$.

Since $\tilde{\beta}_J = (\hat{\mathbf{\Sigma}}_{JJ})^{-1}\tilde{C}_J - \lambda(\hat{\mathbf{\Sigma}}_{JJ})^{-1} \cdot \text{sign}(\beta_J^0)$, we have

$$\begin{aligned}
&\|\tilde{C}_{J^c} - \hat{\mathbf{\Sigma}}_{J^cJ}\tilde{\beta}_J\|_{max} \leq \|\tilde{C}_{J^c} - \hat{\mathbf{\Sigma}}_{J^cJ}(\hat{\mathbf{\Sigma}}_{JJ})^{-1}\tilde{C}_J\|_{max} + \lambda\|\hat{\mathbf{\Sigma}}_{J^cJ}\hat{\mathbf{\Sigma}}_{JJ}^{-1}\|_\infty \\
&\quad \leq \|\tilde{C}_{J^c} - C_{J^c}\|_{max} + \|(\mathbf{\Sigma}_{J^cJ}(\mathbf{\Sigma}_{JJ})^{-1} - \hat{\mathbf{\Sigma}}_{J^cJ}(\hat{\mathbf{\Sigma}}_{JJ})^{-1})C_J\|_{max} \\
&\qquad + \|\hat{\mathbf{\Sigma}}_{J^cJ}(\hat{\mathbf{\Sigma}}_{JJ})^{-1}(C_J - \tilde{C}_J)\|_{max} + \lambda\|\hat{\mathbf{\Sigma}}_{J^cJ}(\hat{\mathbf{\Sigma}}_{JJ})^{-1}\|_\infty \\
&\quad \leq \underbrace{\|\tilde{C}_{J^c} - C_{J^c}\|_{max}}_{(I)} + \underbrace{s\beta^0_{max}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{max} \cdot (1 + \|\hat{\mathbf{\Sigma}}_{J^cJ}(\hat{\mathbf{\Sigma}}_{JJ})^{-1}\|_\infty)}_{(II)} \\
&\qquad + \underbrace{(\lambda + \|\tilde{C}_J - C_J\|_{max}) \cdot \|\hat{\mathbf{\Sigma}}_{J^cJ}(\hat{\mathbf{\Sigma}}_{JJ})^{-1}\|_\infty}_{(III)}.
\end{aligned}$$

If events $\mathcal{A}_3$ and $\mathcal{A}_4$ occur, we know that

$$\begin{aligned}
&(I) \leq \nu_3\sqrt{\log p / \min_j n_j} \leq \nu_3\sqrt{\log p / \min_{j,t} n_{jt}} \\
&(II) \leq \nu_1' s\beta^0_{max}\sqrt{\log p / \min_{j,t} n_{jt}} \cdot (2 - \eta + \frac{2sV\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{max}}{1 - sV\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{max}}) \\
&(III) \leq (\lambda + \nu_3\sqrt{\log p / \min_j n_j}) \cdot (1 - \eta + \frac{2sV\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{max}}{1 - sV\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{max}}).
\end{aligned}$$

Since $\frac{1+s\beta^0_{max}}{\lambda}\sqrt{\frac{\log p}{\min_{j,t} n_{jt}}} \longrightarrow 0$, if $sV\sqrt{\log p/\min_{j,t} n_{jt}} \longrightarrow 0$ or we assume that

$$\|(\mathbf{\Sigma}_{JJ})^{-1}\|_\infty \cdot \sqrt{\frac{s^2 \log p}{\min_{j,t} n_{jt}}} \le \frac{\eta}{\nu_1'(4+\eta)},$$

we have

$$\frac{\|\tilde{C}_{J^c} - \hat{\mathbf{\Sigma}}_{J^cJ}\tilde{\beta}_J\|_{max}}{\lambda} \le \frac{(I)}{\lambda} + \frac{(II)}{\lambda} + \frac{(III)}{\lambda} \le \frac{\eta}{4} + \frac{\eta}{4} + 1 - \frac{\eta}{2} = 1,$$

for sufficiently large p and $\min_{j,t} n_{jt}$.

Therefore, $P(\mathcal{A}_2) \longrightarrow 1$ as $\min_{j,t} n_{jt} \to \infty$ and $p \to \infty$. This completes the proof. □

Proof of Theorem 7: The proof is almost the same as the proof of Theorem 6. We define events $\mathcal{A}_5 = \{\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{max} \le Q_1'\sqrt{\log p/\min_{j,t} n_{jt}}$ and $\mathcal{A}_6 = \{\|\breve{C} - C\|_{max} \le 8(Q_1 + Q_2)\sqrt{\log p/\min_j n_j}\}$. By Theorem 3, we know that $P(\mathcal{A}_5) \longrightarrow 1$ and $P(\mathcal{A}_6) \longrightarrow 1$ as $p \to \infty$. If events $\mathcal{A}_5$ and $\mathcal{A}_6$ occur, we can show that $P(\{\|\breve{\beta}_J - \beta^0_J\|_{max} < \beta^0_{min}\}) \longrightarrow 1$ and $P(\{\|\breve{C}_{J^c} - \hat{\mathbf{\Sigma}}_{J^cJ}\breve{\beta}_J\|_{max} \le \lambda\}) \longrightarrow 1$, as $\min_{j,t} n_{jt} \to \infty$ and $p \to \infty$. This completes the proof. □

References

Fan, J., Li, Q., and Wang, Y. (2016), "Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, to appear.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011), "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electronic Journal of Statistics*, 5, 935–980.

Whittle, P. (1960), "Bounds for the moments of linear and quadratic forms in independent variables," *Theory of Probability & Its Applications*, 5, 302–305.