Extension from the unified regularized estimation framework

Jiaxin Hu

June 11, 2021

Suppose we have K categories of n multivariate normal samples with sample covariance matrices S_k. Let Σ_k denote the true covariance matrices, and $\Omega_k^* = \Sigma_k^{-1}$ denote the true precision matrices. To estimate the precision matrix, we purpose the optimization problem

$$\min_{\Omega^k} \quad \mathcal{L}(\Omega_k, S_k) + \lambda \mathcal{R}(\Omega_k), \tag{1}$$

where $\mathcal{L}(\Omega_k, S_k) = \sum_{k=1}^K \langle S_k, \Omega_k \rangle - \log\det(\Omega_k)$ denotes the loss function. Applying different structures on Ω_k, we assume different parameter spaces. Assuming K categories can be clustered by R groups based on the precision matrices, we also tackle the clustering problem in the model. Here, **we assume the true clustering membership $\boldsymbol{U} \in \mathbb{R}^{K\times R}$ are known.** Particularly, we have following models.

1. Suppose the K categories have a common precision structure, i.e., $\Omega_k^* = \Theta$. Then, we have

$$\mathcal{L}(\Theta, S_k) = \sum_{k=1}^K \langle S_k, \Theta \rangle - K \log\det(\Theta), \quad \mathcal{R}(\Theta) = K \left\| \Theta \right\|_1. \tag{2}$$

2. Suppose the K categories are clustered in R groups based on the magnitude of precision matrix, i.e., $\Omega_k^* = \sum_{r=1}^R u_{kr}\Theta_r^*$ and $u_{kr} = 1$ if k-th category belongs to the r-th group and $u_{kr} = 0$ otherwise. Let $I_r = \{k \in [K] : u_{kr} = 1\}$ and $\sum_{r=1}^R |I_r| = K$. Then, we have

$$\mathcal{L}(\Theta_1, ..., \Theta_R, S_k) = \sum_{r=1}^R \mathcal{L}_r(\Theta_r, S_k), \quad \mathcal{R}(\Theta_1, ..., \Theta_R) = \sum_{r=1}^R \mathcal{R}_r(\Theta_r), \tag{3}$$

where

$$\mathcal{L}_r(\Theta_r, S_k) = \sum_{k\in I_r} \langle S_k, \Theta_r \rangle - |I_r| \log\det(\Theta_r), \quad \mathcal{R}_r(\Theta_r) = |I_r| \left\| \Theta_r \right\|_1.$$

3. Suppose the K categories are clustered in R groups based on the space of precision matrix, i.e., $\Omega_k^* = \sum_{r=1}^R u_{kr}\Theta_r^*$ and $u_{kr} > 0$ if k-th category belongs to the r-th group and $u_{kr} = 0$ otherwise. For identifiability, we have $\left\| u_{\cdot r} \right\|_F = 1, r \in [R]$. Then, we have

$$\mathcal{L}(\Theta_1, ..., \Theta_R, S_k) = \sum_{r=1}^R \mathcal{L}_r(\Theta_r, S_k), \quad \mathcal{R}(\Theta_1, ..., \Theta_R) = \sum_{r=1}^R \mathcal{R}_r(\Theta_r),$$

where

$$\mathcal{L}_r(\Theta_r, S_k) = \sum_{k\in I_r} \langle S_k, u_{kr}\Theta_r \rangle - |I_r| \log\det(u_{kr}\Theta_r), \quad \mathcal{R}_r(\Theta_r) = |I_r| \left\| \Theta_r \right\|_1.$$

4. Suppose the K categories are clustered in R groups based on the space of precision matrix with an intercept matrix, i.e., $\Omega_k^* = \Theta_0 + \sum_{r=1}^R u_{kr}\Theta_r^*$ and $u_{kr} \neq 0$ if k-th category belongs to the r-th group and $u_{kr} = 0$ otherwise. For identifiability, we have $\|u_{.r}\|_F = 1$ and $\sum_{k=1}^K u_{kr} = 0, r \in [R]$. Note that we allow $r = 0$ in this case and thus $I_0 = \{k \in [K] : u_{kr} = 0, \text{for all } r \in [R]\}$ and $\sum_{r=0}^R |I_r| = K$. Then, we have

$$\begin{aligned}\mathcal{L}(\Theta_0, \Theta_1, ..., \Theta_R, S_k) &= \sum_{r=1}^R \sum_{k \in I_r} \langle S_k, \Theta_0 + u_{kr}\Theta_r\rangle - |I_r| \log\det(\Theta_0 + u_{kr}\Theta_r) \\ &\quad + \sum_{k \in I_0} \langle S_k, \Theta_0\rangle - |I_0| \log\det(\Theta_0), \\ \mathcal{R}(\Theta_0, \Theta_1, ..., \Theta_R) &= \sum_{r=1}^R |I_r| \|\Theta_r\|_1 + K \|\Theta_0\|_1 .\end{aligned}$$

1 Case 1

Corollary 1. Suppose $\|\Theta^*\|_0 = s$ and $\lambda \geq C'\sqrt{\frac{\log p}{nK}}$. Let $\hat{\Theta}_\lambda$ denote the optimal solution to (1) and $\hat{\Delta} = \hat{\Theta}_\lambda - \Theta^*$. With high probability tends to 1, the optimal solution satisfies the bound

$$\left\|\hat{\Theta}_\lambda - \Theta^*\right\|_F \leq C_1\tau^2\sqrt{\frac{s\log p}{nK}}.$$

Proof. Let $\Delta = \hat{\Theta} - \Theta^*$, where $\hat{\Theta}$ is an arbitrary estimate. Define the function

$$\mathcal{F}(\Delta) = \mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) + \lambda\left[\mathcal{R}(\Theta^* + \Delta) - \mathcal{R}(\Theta^*)\right], \tag{4}$$

where $\mathcal{L}, \mathcal{R}$ are in definition (2). Note that $\mathcal{F}(\Delta)$ includes two parts: 1) the difference between the loss function and 2) the difference between the regularizer term. We deal with these two parts by the RSC property and decomposability respectively.

For the difference between the regularization term, we define the model subspace

$$\mathcal{M} = \{\Theta \in \mathbb{R}^{p\times p} | \Theta_{ij} \neq 0, (i,j) \notin T\}, \quad T = \{(i,j) | \Theta_{ij}^* \neq 0\},$$

where $|T| = s$. Then, we know that $\mathcal{R}(\Theta)$ is decomposable with $\mathcal{M}$, and the dual norm $\mathcal{R}^*(\Theta) = \frac{1}{K}\|\Theta\|_{\max}$. Besides, the subspace compatbility constant with respect to the pair $(\|\cdot\|_1, \|\cdot\|_F)$ is

$$\Psi(\mathcal{M}) = \sup_{A \in \mathcal{M}/\{0\}} \frac{K\|A\|_1}{\|A\|_F} = K\sqrt{s}.$$

Then, by the Lemma 3 in the Supplement of (Negahban et al., 2012), we have

$$\mathcal{R}(\Theta^* + \Delta) - \mathcal{R}(\Theta^*) \geq \mathcal{R}(\Delta_{\mathcal{M}^\perp}) - \mathcal{R}(\Delta_{\mathcal{M}}). \tag{5}$$

For the difference between loss function, we have

$$\begin{aligned}\mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) &= \sum_{k=1}^K \langle S_k, \Delta\rangle - K\left[\log\det(\Theta^* + \Delta) - \log\det(\Theta^*)\right] \\ &\geq \sum_{k=1}^K \langle S_k - \Sigma, \Delta\rangle + \frac{K}{4\tau^2}\|\Delta\|_F^2,\end{aligned} \tag{6}$$

where τ is the largest singular value of Θ^*, and the last inequality follows by the Lemma A1 in (Guo et al., 2011). Note that

$$|\sum_{k=1}^{K}\langle S_k - \Sigma, \Delta\rangle| = |\langle\sum_{k=1}^{K} S_k - K\Sigma, \Delta\rangle| \leq \mathcal{R}^*\left(\sum_{k=1}^{K} S_k - K\Sigma\right)\mathcal{R}(\Delta),$$

where

$$\mathcal{R}^*\left(\sum_{k=1}^{K} S_k - K\Sigma\right) = \left\|\frac{1}{K}\sum_{k=1}^{K} S_k - \Sigma\right\|_{\max} \leq C\sqrt{\frac{\log p}{nK}},$$

with high probability by the Lemma 1 of (Rothman et al., 2009). Since $\lambda \geq C'\sqrt{\frac{\log p}{nK}}$, we have $\lambda \geq 2\mathcal{R}^*\left(\nabla\mathcal{L}(\Theta^*)\right)$, for C' large enough.

Plugging the inequality (5) and (6) into the function (4), with high probability, we have

$$\begin{aligned}\mathcal{F}(\Delta) &\geq \frac{K}{4\tau^2}\|\Delta\|_F^2 + \lambda\left[\mathcal{R}(\Delta_{\mathcal{M}^\perp}) - \mathcal{R}(\Delta_{\mathcal{M}})\right] - \frac{\lambda}{2}\mathcal{R}(\Delta)\\ &\geq \frac{K}{4\tau^2}\|\Delta\|_F^2 - \frac{3\lambda}{2}\mathcal{R}(\Delta_{\mathcal{M}}),\\ &\geq \frac{K}{4\tau^2}\|\Delta\|_F^2 - \frac{3\lambda}{2}\Psi(\mathcal{M})\|\Delta\|_F,\end{aligned}$$

where the second the inequality follows by the triangle inequality $\mathcal{R}(\Delta) \leq \mathcal{R}(\Delta_{\mathcal{M}^\perp}) + \mathcal{R}(\Delta_{\mathcal{M}})$, and the third inequality follows by the definition of subspace compatibility constant.

Note that $\mathcal{F}(\Delta) > 0$ with high probability for all Δ satisfying

$$\|\Delta\|_F \geq \frac{3\lambda\Psi(\mathcal{M})4\tau^2}{2K} = C_1\tau^2\sqrt{\frac{s\log p}{nK}},$$

for some positive constant C_1. Therefore, we know that

$$\left\|\hat{\Delta}\right\|_F = \left\|\hat{\Theta}_\lambda - \Theta^*\right\|_F \leq C_1\tau^2\sqrt{\frac{s\log p}{nK}},$$

with high probability. □

2 Case 2

Corollary 2. Suppose $\|\Theta_r^*\|_0 \leq s$ and $\lambda \geq \max_r C'\sqrt{\frac{\log p}{n|I_r|}}$. Let $\hat{\Theta}_{r,\lambda}$ denote the optimal solution to (1), and $\hat{\Delta}_r = \hat{\Theta}_{r,\lambda} - \Theta_r^*, r \in [R]$. With high probability tends to 1, the optimal solution satisfies the bound

$$\sum_{k=1}^{K}\left\|\hat{\Omega}_{k,\lambda} - \Omega_k^*\right\|_F = \sum_{r=1}^{R}|I_r|\left\|\hat{\Delta}_r\right\|_F \leq C\tau^2\sum_{r=1}^{R}\sqrt{\frac{s\log p|I_r|}{n}}.$$

Proof. Let $\Delta_r = \hat{\Theta}_r - \Theta_r^*$, where $\hat{\Theta}_r$ are arbitrary estimates. By the definition in (3), we define the function

$$\mathcal{F}(\Delta_1, ..., \Delta_R) = \sum_{r=1}^{R}\mathcal{F}_r(\Delta_r),$$

where

$$\mathcal{F}_r(\Delta_r) = \mathcal{L}_r(\Theta_r^* + \Delta_r) - \mathcal{L}_r(\Theta_r^*) - \lambda \left[\mathcal{R}_r(\Theta_r^* + \Delta_r) - \mathcal{R}_r(\Theta_r^*)\right].$$

By Case 1, with $\lambda \geq \max_r C' \sqrt{\frac{\log p}{n|I_r|}}$, we know that $\mathcal{F}_r(\Delta_r) > 0$ with high probability for all Δ_r satisfying

$$\|\Delta_r\|_F \geq C_r \tau^2 \sqrt{\frac{s \log p}{n|I_r|}},$$

where τ is the largest singular value of $\Theta_r, r \in [R]$. To let $\mathcal{F}(\Delta_1, ..., \Delta_R) > 0$, the differences $\Delta_r, r \in [R]$ satisfying

$$(\Delta_1, ..., \Delta_R) \in \left\{\|\Delta_1\|_F \geq C_1 \tau^2 \sqrt{\frac{s \log p}{n|I_1|}}\right\} \times \cdots \times \left\{\|\Delta_R\|_F \geq C_R \tau^2 \sqrt{\frac{s \log p}{n|I_R|}}\right\},$$

which implies that

$$(\hat{\Delta}_1, ..., \hat{\Delta}_R) \in \left\{\|\Delta_1\|_F \leq C_1 \tau^2 \sqrt{\frac{s \log p}{n|I_1|}}\right\} \times \cdots \times \left\{\|\Delta_R\|_F \leq C_R \tau^2 \sqrt{\frac{s \log p}{n|I_R|}}\right\}.$$

Therefore, we have

$$\sum_{k=1}^{K} \left\|\hat{\Omega}_{k,\lambda} - \Omega_k^*\right\|_F = \sum_{r=1}^{R} |I_r| \left\|\hat{\Delta}_r\right\|_F \leq C\tau^2 \sum_{r=1}^{R} \sqrt{\frac{s \log p |I_r|}{n}}.$$

□

3 Case 3

Corollary 3. Suppose $\|\Theta_r^*\|_0 \leq s$ and $\lambda \geq \max_r C' \sqrt{\frac{\log p}{n|I_r|}}$. Let $\hat{\Theta}_{r,\lambda}$ denote the optimal solution to (1), and $\hat{\Delta}_r = \hat{\Theta}_{r,\lambda} - \Theta_r^*, r \in [R]$. With high probability tends to 1, the optimal solution satisfies the bound

$$\sum_{k=1}^{K} \left\|\hat{\Omega}_{k,\lambda} - \Omega_k^*\right\|_F = \sum_{r=1}^{R} \sum_{k \in I_r} u_{kr} \left\|\hat{\Delta}_r\right\|_F \leq C\tau^2 R \sqrt{\frac{s \log p}{n}}.$$

Proof. Let $\Delta_r = \hat{\Theta}_r - \Theta_r^*$, where $\hat{\Theta}_r$ are arbitrary estimates. By the definition in (3), we define the function

$$\mathcal{F}(\Delta_1, ..., \Delta_R) = \sum_{r=1}^{R} \mathcal{F}_r(\Delta_r),$$

where

$$\mathcal{F}_r(\Delta_r) = \mathcal{L}_r(\Theta_r^* + \Delta_r) - \mathcal{L}_r(\Theta_r^*) - \lambda \left[\mathcal{R}_r(\Theta_r^* + \Delta_r) - \mathcal{R}_r(\Theta_r^*)\right].$$

Note that the difference in loss function is different with previous cases due to the continuous u_{kr}. Specifically,

$$\mathcal{L}_r(\Theta_r^* + \Delta_r) - \mathcal{L}_r(\Theta_r^*) = \sum_{k \in I_r} \langle S_k, u_{kr}\Delta_r \rangle - |I_r| \left[\log\det(\Theta_r^* + \Delta_r) + \log u_{kr} - \log\det(\Theta_r^*) - \log u_{kr}\right]$$

$$\geq \sum_{k \in I_r} \langle u_{kr} S_k - \Sigma_r, \Delta_r \rangle + \frac{|I_r|}{4\tau^2} \|\Delta_r\|_F^2 ,$$

where the second inequality follows by the Lemma A1 in (Guo et al., 2011) and the fact that S_k corresponds to the true covariance matrix $\frac{1}{u_{kr}}\Sigma_r$. Note that the dual norm $\mathcal{R}_r^*(\Theta) = \frac{1}{|I_r|}\|\Theta\|_{\max}$. By Cauchy Schwartz inequality, we have

$$|\sum_{k \in I_r} \langle u_{kr} S_k - \Sigma_r, \Delta_r \rangle| = |\langle \sum_{k \in I_r} u_{kr} S_k - |I_r|\Sigma_r, \Delta_r \rangle| \leq \mathcal{R}_r^* \left(\sum_{k \in I_r} u_{kr} S_k - |I_r|\Sigma_r \right) \mathcal{R}_r(\Delta_r),$$

where

$$\mathcal{R}_r^* \left(\sum_{k \in I_r} u_{kr} S_k - |I_r|\Sigma_r \right) = \left\| \frac{1}{|I_r|} \sum_{k \in I_r} u_{kr} S_k - \Sigma_r \right\|_{\max} \leq C \sqrt{\frac{\log p}{n|I_r|}},$$

with high probability by equation (2) in note 0323.

The other parts keep the same with Case 2, and thus we have

$$(\hat{\Delta}_1, ..., \hat{\Delta}_R) \in \left\{ \|\Delta_1\|_F \leq C_1 \tau^2 \sqrt{\frac{s \log p}{n|I_1|}} \right\} \times \cdots \times \left\{ \|\Delta_R\|_F \leq C_R \tau^2 \sqrt{\frac{s \log p}{n|I_R|}} \right\}.$$

Therefore, we have

$$\sum_{k=1}^{K} \left\| \hat{\Omega}_{k,\lambda} - \Omega_k^* \right\|_F = \sum_{r=1}^{R} \sum_{k \in I_r} u_{kr} \left\| \hat{\Delta}_r \right\|_F \leq C\tau^2 R \sqrt{\frac{s \log p}{n}},$$

by the fact that $\sum_{k \in I_r} u_{kr} \leq \sqrt{|I_r| \sum_{k \in I_r} u_{kr}^2} = \sqrt{I_r}$.

□

References

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B., et al. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical science*, 27(4):538–557.

Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186.