Exact Clustering

Jiaxin Hu

July 5, 2021

1 Hard constraint

Model

Suppose we have K categories following multivariate normal distribution with precision matrix $\Omega_k \in \mathbb{R}^{p\times p}$ belonging to R groups. Suppose S_k denote the sample covariance matrices for k-th group with n sample size, and $\Sigma_k = \Omega_k^{-1}$ denote the true covariance matrices. Consider the model

$$\Omega_k = \Theta_0 + u_k \Theta_{z(k)},$$

where Θ_0 is the intercept matrix, $\Theta_r, r \in [R]$ denote the factor matrices, $z = (z(1), ..., z(K)) \in [R]^K$ denote the label vector, and $u = (u_1, ..., u_K) \in \mathbb{R}^K$ denote the degree-corrected parameter vector for K categories. Let $M \in \mathbb{R}^{K\times R}$ denote the membership matrix generated by z, and $U = \text{diag}(u) \in \mathbb{R}^{K\times K}$. Rewrite the model in matrix form,

$$\boldsymbol{\Omega} = \boldsymbol{\Theta_0} + UM\boldsymbol{\Theta},$$

where

$$\boldsymbol{\Omega} = \begin{bmatrix} \text{vec}(\Omega_1) \\ \vdots \\ \text{vec}(\Omega_K) \end{bmatrix}, \quad \boldsymbol{\Theta_0} = \begin{bmatrix} \text{vec}(\Theta_0) \\ \vdots \\ \text{vec}(\Theta_0) \end{bmatrix}, \quad \boldsymbol{\Theta} = \begin{bmatrix} \text{vec}(\Theta_1) \\ \vdots \\ \text{vec}(\Theta_R). \end{bmatrix}$$

Our goal is to find the good estimation of $(U, M, \{\Theta_r\})$.

Notations

1. Let $U^*, u^*, M^*, z^*, \{\Theta_r^*\}_{r=0}^R$ denote the true parameters.
2. Let $I_r = \{k \in [K] : z(k) = r\}$ collects the categories that belong to group r with given membership z, and $I_{ar} = \{k \in [K] : z(k) = r, z^*(k) = a\}$ collects the categories that belong to group r based on z and true group a.
3. Let $D_{ar} = |I_{ar}|$ and $MCR(\hat{M}, M^*) = \max_{r,a,a'\in[R]} \min\{D_{ar}, D_{a'r}\}$.

Parameter Space

Suppose the true parameters $(U^*, M^*, \{\Theta_r^*\})$ belongs to the space $\mathcal{P}^*$, where

$$
\begin{aligned}
\mathcal{P}^* = \Bigg\{(U, M, \{\Theta_r\}): \quad & \Theta_r \text{ is positive definite for all } r = \{0\} \cup [R]; \\
& 0 < \tau_1 < \min_{r \in \{0\}\cup[R]} \varphi_{\min}(\Theta_r) \le \max_{r \in \{0\}\cup[R]} \varphi_{\max}(\Theta_r) < \tau_2; \\
& \max_{r, r' \in [R]} \cos(\Theta_r, \Theta_{r'}) < \delta < 1; \\
& \min_{r \in [R]} |I_r| \ge 1; \quad \min_{k \in [K]} |u_k| > m > 0; \\
& \sum_{k \in I_r} u_k^2 = 1, \quad \sum_{k \in I_r} u_k = 0, \text{ for all } r \in [R] \Bigg\}.
\end{aligned}
$$

Suppose we find the estimate in a larger space $\mathcal{P}$, where

$$
\begin{aligned}
\mathcal{P} = \Bigg\{(U, M, \{\Theta_r\}): \quad & \Theta_r \text{ is positive definite for all } r = \{0\} \cup [R]; \\
& \max_{r, r' \in [R]} \cos(\Theta_r, \Theta_{r'}) < \delta < 1; \\
& \min_{r \in [R]} |I_r| \ge 1; \quad \min_{k \in [K]} |u_k| > m > 0; \\
& \sum_{k \in I_r} u_k^2 = 1, \quad \sum_{k \in I_r} u_k = 0, \text{ for all } r \in [R] \Bigg\}.
\end{aligned}
$$

Estimator

We consider the constrained MLE, denoted by $(\hat{U}, \hat{M}, \{\hat{\Theta}_r\})$, where

$$(\hat{U}, \hat{M}, \{\hat{\Theta}_r\}) = \mathop{\arg\min}_{(U, M, \Theta_r) \in \mathcal{P}} \mathcal{Q}(U, M, \Theta_r),$$

and

$$\mathcal{Q}(U, M, \Theta_r) = \sum_{k \in [K]} \langle S_k, \Theta_0 + u_k \Theta_{z(k)} \rangle - \log\det\left(\Theta_0 + u_k \Theta_{z(k)}\right).$$

Exact Clustering rate

Here we show the exact clustering rate of the MLE $(\hat{U}, \hat{M}, \hat{\{\Theta_r\}})$, i.e., the rate for $MCR(\hat{M}, M^*) = 0$.

Lemma 1 (Exact clustering rate for MLE)**.** *For the MLE $(\hat{U}, \hat{M}, \hat{\{\Theta_r\}})$, we have*

$$\mathbb{P}\left(MCR(\hat{M}, M^*) = 0\right) \ge 1 - \sum_{\epsilon \in [\varepsilon]} K^R \left\{1 - \left[1 - C_1 \exp\left(-C_2 n \frac{m^2 F^2 \epsilon}{32 \tau_2^4 p^2 K}\right)\right]^K\right\},$$

where $\varepsilon = \lceil \frac{K-R+1}{2} \rceil$ and $F^2 = 2m^2\tau_1^2 - \frac{2\delta\tau_2^2}{m^2}$.

Remark 1. Lemma (1) implies a exponential rate on n for the exact clustering, i.e., $\mathbb{P}\left(MCR(\hat{M}, M^*) = 0\right) = 1 - \mathcal{O}(\exp(-n))$.

Proof. We write the probability

$$
\begin{aligned}
\mathbb{P}\left(MCR(\hat{M}, M^*) = 0\right) &= \mathbb{P}\left((\hat{U}, \hat{M}, \{\hat{\Theta}_r\}) = \underset{(U,M,\Theta_r)\in\mathcal{P}}{\arg\min}\ \mathcal{Q}(U, M, \Theta_r), \quad MCR(\hat{M}, M^*) = 0\right) \\
&= 1 - \sum_{\epsilon\in[\varepsilon]} \mathbb{P}\left((\hat{U}, \hat{M}, \{\hat{\Theta}_r\}) = \underset{(U,M,\Theta_r)\in\mathcal{P}}{\arg\min}\ \mathcal{Q}(U, M, \Theta_r), \quad MCR(\hat{M}, M^*) = \epsilon\right) \\
&\geq 1 - \sum_{\epsilon\in[\varepsilon]} \sum_{\tilde{M}:MCR(\tilde{M},M^*)=\epsilon} \mathbb{P}\left(0 \geq \mathcal{Q}(\tilde{U}, \tilde{M}, \tilde{\Theta}_r) - \mathcal{Q}(U^*, M^*, \Theta_r^*)\right), \qquad (1)
\end{aligned}
$$

where the probability is taken with respect to the random samples S_k, and $\tilde{U}, \{\tilde{\Theta}_r\}$ are optimizer of $\mathcal{Q}(U, \tilde{M}, \Theta_r)$ with given membership $\tilde{M}$, and $\varepsilon = \lceil \frac{K-R+1}{2} \rceil$ is the largest possible value of MCR. Then, we only need to find the upper bound for probability $\mathbb{P}\left(0 \geq \mathcal{Q}(\tilde{U}, \tilde{M}, \tilde{\Theta}_r) - \mathcal{Q}(U^*, M^*, \Theta_r^*)\right)$ for some $\tilde{M}$ such that $MCR(\tilde{M}, M^*) = \epsilon$, and combine the upper bounds together.

Before the proof, we introduce few notations. Let $\Delta_0 = \tilde{\Theta}_0 - \Theta_0$ and $\Delta_{k,ar} = \Delta_0 + \tilde{u}_k\tilde{\Theta}_r - u_k^*\Theta_a^*$ for $k \in I_{ar}$.

Step I: Upper bound

Note that

$$
\begin{aligned}
\mathcal{Q}(\tilde{U}, \tilde{M}, \tilde{\Theta}_r) - \mathcal{Q}(U^*, M^*, \Theta_r^*) &\geq \sum_{r,a\in[R]} \sum_{k\in I_{ar}} \left[\frac{1}{4\tau_2^2} \|\Delta_{k,ar}\|_F^2 + \langle S_k - \Sigma_k, \Delta_{k,ar}\rangle\right] \\
&\geq \sum_{r,a\in[R]} \sum_{k\in I_{ar}} \left[\frac{1}{4\tau_2^2} \|\Delta_{k,ar}\|_F^2 - \|S_k - \Sigma_k\|_{\max} p \|\Delta_{k,ar}\|_F\right].
\end{aligned}
$$

Note that

$$
\sum_{a,r\in[R]} \sum_{k\in I_{ar}} \|\Delta_{k,ar}\|_F \leq \sqrt{K} \sqrt{\sum_{a,r\in[R]} \sum_{k\in I_{ar}} \|\Delta_{k,ar}\|_F^2}.
$$

Then, we have

$$
\begin{aligned}
&\mathbb{P}\left(0 \geq \mathcal{Q}(\tilde{U}, \tilde{M}, \tilde{\Theta}_r) - \mathcal{Q}(U^*, M^*, \Theta_r^*)\right) \\
&\leq \mathbb{P}\left(\frac{1}{4\tau_2^2} \sum_{r,a\in[R]} \sum_{k\in I_{ar}} \|\Delta_{k,ar}\|_F^2 \leq \max_{k\in[K]} \|S_k - \Sigma_k\|_{\max} p\sqrt{K} \sqrt{\sum_{a,r\in[R]} \sum_{k\in I_{ar}} \|\Delta_{k,ar}\|_F^2}\right) \\
&= \mathbb{P}\left(\frac{1}{4\tau_2^2 p\sqrt{K}} \sqrt{\sum_{a,r\in[R]} \sum_{k\in I_{ar}} \|\Delta_{k,ar}\|_F^2} \leq \max_{k\in[K]} \|S_k - \Sigma_k\|_{\max}\right). \qquad (2)
\end{aligned}
$$

For simplicity, let $V_k = \|S_k - \Sigma_k\|_{\max}$. Since $MCR(\tilde{M}, M^*) = \epsilon$, there exists r_0, a_1, a_2 such that $\min\{D_{a_1,r_0}, D_{a_2,r_0}\} = \epsilon$. Note that

$$\begin{aligned}
\sum_{a,r\in[R]}\sum_{k\in I_{ar}} \|\Delta_{k,ar}\|_F^2 &= K\|\Delta_0\|_F^2 + \sum_{a,r\in[R]}\sum_{k\in I_{ar}} \left\|\tilde{u}_k\tilde{\Theta}_r - u_k^*\Theta_a^*\right\|_F^2 \\
&\geq \sum_{k\in I_{a_1r_0}} \left\|\tilde{u}_k\tilde{\Theta}_{r_0} - u_k^*\Theta_{a_1}^*\right\|_F^2 + \sum_{k\in I_{a_2r_0}} \left\|\tilde{u}_k\tilde{\Theta}_{r_0} - u_k^*\Theta_{a_2}^*\right\|_F^2 \\
&\geq \frac{\epsilon}{2}\max_{k\in I_{a_1r_0}, k\in I_{a_2r_0}} \left[\left\|\tilde{u}_k\tilde{\Theta}_{r_0} - u_k^*\Theta_{a_1}^*\right\|_F + \left\|\tilde{u}_{k'}\tilde{\Theta}_{r_0} - u_{k'}^*\Theta_{a_2}^*\right\|_F\right]^2,
\end{aligned}$$

and

$$\begin{aligned}
\left[\left\|\tilde{u}_k\tilde{\Theta}_{r_0} - u_k^*\Theta_{a_1}^*\right\|_F + \left\|\tilde{u}_{k'}\tilde{\Theta}_{r_0} - u_{k'}^*\Theta_{a_2}^*\right\|_F\right]^2 &\geq m^2\left[\left\|\tilde{\Theta}_{r_0} - \frac{u_k^*}{\tilde{u}_k}\Theta_{a_1}^*\right\|_F + \left\|\tilde{\Theta}_{r_0} - \frac{u_{k'}^*}{\tilde{u}_{k'}}\Theta_{a_2}^*\right\|_F\right]^2 \\
&\geq m^2\left\|\frac{u_k^*}{\tilde{u}_k}\Theta_{a_1}^* - \frac{u_{k'}^*}{\tilde{u}_{k'}}\Theta_{a_2}^*\right\|_F^2 \\
&\geq m^2F^2,
\end{aligned}$$

where $F^2 = 2m^2\tau_1^2 - \frac{2\delta\tau_2^2}{m^2}$, and the inequalities follows by the inequality (2) in note 0629. Then, we have

$$\mathbb{P}\left(\frac{mF\sqrt{\epsilon}}{4\sqrt{2}\tau_2^2 p\sqrt{K}} \leq \max_{k\in[K]}\|S_k - \Sigma_k\|_{\max}\right) = 1 - \left[\mathbb{P}\left(\frac{mF\sqrt{\epsilon}}{4\sqrt{2}\tau_2^2 p\sqrt{K}} \geq \|S_k - \Sigma_k\|_{\max}\right)\right]^K,$$

with

$$\begin{aligned}
\mathbb{P}\left(\frac{mF\sqrt{\epsilon}}{4\sqrt{2}\tau_2^2 p\sqrt{K}} \geq \|S_k - \Sigma_k\|_{\max}\right) &= 1 - \mathbb{P}\left(\frac{mF\sqrt{\epsilon}}{4\sqrt{2}\tau_2^2 p\sqrt{K}} \leq \|S_k - \Sigma_k\|_{\max}\right) \\
&\geq 1 - C_1\exp\left(-C_2 n\frac{m^2F^2\epsilon}{32\tau_2^4p^2K}\right),
\end{aligned}$$

where the last inequality follows by the Lemma 1 in (Rothman et al., 2008). Hence, the probability (2) becomes

$$\mathbb{P}\left(0 \geq \mathcal{Q}(\tilde{U}, \tilde{M}, \tilde{\Theta}_r) - \mathcal{Q}(U^*, M^*, \Theta_r^*)\right) \leq 1 - \left[1 - C_1\exp\left(-C_2 n\frac{m^2F^2\epsilon}{32\tau_2^4p^2K}\right)\right]^K. \tag{3}$$

Step II: Combine Plugging the upper bound (3) into the probability (1), we have

$$\mathbb{P}\left(MCR(\hat{M}, M^*) = 0\right) \geq 1 - \sum_{\epsilon\in[\varepsilon]} K^R\left\{1 - \left[1 - C_1\exp\left(-C_2 n\frac{m^2F^2\epsilon}{32\tau_2^4p^2K}\right)\right]^K\right\}.$$

□

References

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. Electronic Journal of Statistics, 2:494–515.