

# Graphic Lasso: Self-Consistency

Jiabin Hu

February 16, 2021

## 1 Noiseless case

Consider the noiseless case

$$\mathcal{Y} = f(\Theta),$$

where  $\Theta = \mathcal{C} \times_1 \mathbf{M}_1 \times_2 \cdots \times_K \mathbf{M}_K$ , and  $f(\cdot)$  is an entry-wise link function. Suppose we have the following optimization problem.

$$\max_{\Theta = \mathcal{C} \times_1 \mathbf{M}_1 \times_2 \cdots \times_K \mathbf{M}_K} \mathcal{L}_{\mathcal{Y}}(\Theta) = \langle \mathcal{Y}, \Theta \rangle - \sum_{i_1, \dots, i_K} g(\Theta_{i_1, \dots, i_K}). \quad (1)$$

**Lemma 1** (Noiseless estimation). *Let  $\{\mathcal{C}, \mathbf{M}_k\}$  denote the true parameters and  $\{\hat{\mathcal{C}}, \hat{\mathbf{M}}_k\}$  are the estimation which maximizes the loss function. Suppose  $g(\cdot)$  is a convex function with bounded second derivative  $\sup_x g''(x) \leq a$ , and  $\max_{r_1, \dots, r_K} |(g')^{-1}(f(c_{r_1, \dots, r_K}))| \leq C$ , where  $C$  is a positive constant depends on  $\mathcal{C}$ . Assume the minimal gap between blocks is strictly larger than 0, i.e.,  $\delta > 0$ . Then, for any  $\epsilon > 0$ , we have*

$$\mathbb{P}(MCR(\hat{\mathbf{M}}_k, \mathbf{M}_k) \geq \epsilon) = 0.$$

*Proof.* We prove the accuracy in following steps.

1. With given membership matrix  $\hat{\mathbf{M}}_k$ , the estimate  $\hat{\mathcal{C}}$  is

$$\hat{c}_{r_1, \dots, r_K}(\hat{\mathbf{M}}_k) = (g')^{-1} \left( \frac{1}{\prod_k d_k \prod_k \hat{p}_{r_k}^{(k)}} [f(\mathcal{C}) \times_1 \mathbf{M}_1 \hat{\mathbf{M}}_1^T \times_2 \cdots \times_K \mathbf{M}_K \mathbf{M}_K^T]_{r_1, \dots, r_K} \right).$$

Note that the estimation  $\hat{\mathcal{C}}$  depends on  $\hat{\mathbf{M}}_k$ . Therefore, we denote the estimation as  $\hat{\mathcal{C}}(\hat{\mathbf{M}}_k) = \llbracket \hat{c}_{r_1, \dots, r_K}(\hat{\mathbf{M}}_k) \rrbracket$ .

2. We define some useful functions. First, we define

$$F(\hat{\mathbf{M}}_k) = \mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{C}}(\hat{\mathbf{M}}_k), \hat{\mathbf{M}}_k) = \sum_{r_1, \dots, r_K} \prod_k d_k \prod_k \hat{p}_{r_k}^{(k)} h(g'(\hat{c}_{r_1, \dots, r_K}(\hat{\mathbf{M}}_k))),$$

where  $h(x) = x(g')^{-1}(x) - g((g')^{-1}(x))$ .

Note that  $\hat{\mathcal{C}}(\hat{\mathbf{M}}_k)$  does not include the randomness. Thus, we have  $g'(\hat{c}_{r_1, \dots, r_K}(\hat{\mathbf{M}}_k)) = \mathbb{E} \left[ g'(\hat{c}_{r_1, \dots, r_K}(\hat{\mathbf{M}}_k)) \right]$ , and

$$G(\hat{\mathbf{M}}_k) = \sum_{r_1, \dots, r_K} \prod_k d_k \prod_k \hat{p}_{r_k}^{(k)} h(\mathbb{E} \left[ g'(\hat{c}_{r_1, \dots, r_K}(\hat{\mathbf{M}}_k)) \right]) = F(\hat{\mathbf{M}}_k),$$

which implies that there does not exist the estimation error.

**Note that for true membership, we have**

$$F(\mathbf{M}_k) = G(\mathbf{M}_k) = \mathcal{L}_Y(\hat{\mathcal{C}}(\mathbf{M}_k), \mathbf{M}_k),$$

where  $\hat{\mathcal{C}}(\mathbf{M}_k) = (g')^{-1}(f(\mathcal{C}))$  **is not equal to the true core tensor  $\mathcal{C}$ .**

3. We only need to consider the classification error. Under the assumptions of the positive minimal gap and the boundedness of the second derivative of  $g$ , when  $MCR(\hat{\mathbf{M}}_k, \mathbf{M}_k) \geq \epsilon$  for any  $\epsilon > 0$ , we have

$$G(\hat{\mathbf{M}}_k) - G(\mathbf{M}_k) \leq -\frac{\epsilon}{4a} \tau^{K-1} \delta.$$

4. Since  $\{\hat{\mathcal{C}}\hat{\mathbf{M}}_k, \hat{\mathbf{M}}_k\}$  is the maximizer of the loss function, we have

$$0 \leq F(\hat{\mathbf{M}}_k) - F(\mathbf{M}_k) = G(\hat{\mathbf{M}}_k) - G(\mathbf{M}_k).$$

Therefore, we obtain that

$$\mathbb{P}(MCR(\hat{\mathbf{M}}_k, \mathbf{M}_k) \geq \epsilon) = \mathbb{P}(G(\hat{\mathbf{M}}_k) - G(\mathbf{M}_k) \leq -\frac{\epsilon}{4a} \tau^{K-1} \delta) = 0.$$

□

**Remark 1.** The lemma 1 implies that the true membership  $\mathbf{M}_k$  is the maximizer of the function  $G(\mathbf{M}'_k)$ . Due to the noiselessness,  $G(\mathbf{M}'_k) = \mathcal{L}_Y(\hat{\mathcal{C}}(\mathbf{M}'_k), \mathbf{M}'_k)$ , and  $\{\hat{\mathcal{C}}(\mathbf{M}_k), \mathbf{M}_k\}$  is the maximizer of the noiseless loss function. However, the true parameter  $\{\mathcal{C}, \mathbf{M}_k\}$  is not the maximizer of the noiseless loss function, since  $\hat{\mathcal{C}}(\mathbf{M}_k) \neq \mathcal{C}$ . Therefore, we conclude that the loss function (1) is **self-consistent to  $\{\hat{\mathcal{C}}(\mathbf{M}_k), \mathbf{M}_k\}$  but not self-consistent to  $\Theta$ .**

**Remark 2.** Define

$$\hat{\Theta} = \hat{\mathcal{C}}(\mathbf{M}_k) \times_1 \mathbf{M}_1 \times_1 \cdots \times_K \mathbf{M}_K.$$

Then,  $\hat{\Theta}$  is an unbiased estimate of  $\Theta$  if and only if  $g' = f$ .

**Remark 3.** Which assumption in the noisy case corresponds to the self-consistency of  $\mathbf{M}_k$ ?

Note that in the noisy case, we have

$$\begin{aligned} G_{noise}(\hat{\mathbf{M}}_k) &= \sum_{r_1, \dots, r_K} \prod_k d_k \prod_k \hat{p}_{r_k}^{(k)} h(\mathbb{E} [g'(\hat{c}_{r_1, \dots, r_K}(\hat{\mathbf{M}}_k))]) \\ &= \langle f(\mathcal{C}) \times_1 \mathbf{M}_1 \hat{\mathbf{M}}_1^T \times_2 \cdots \times_K \mathbf{M}_K \mathbf{M}_K^T, (g')^{-1} [f(\mathcal{C}) \times_1 \mathbf{M}_1 \hat{\mathbf{M}}_1^T \times_2 \cdots \times_K \mathbf{M}_K \mathbf{M}_K^T] \rangle \\ &\quad - \sum_{i_1, \dots, i_K} g \left( (g')^{-1} \left[ f(\mathcal{C}) \times_1 \mathbf{M}_1 \hat{\mathbf{M}}_1^T \times_2 \cdots \times_K \mathbf{M}_K \mathbf{M}_K^T \right] \times_1 \mathbf{M}_1 \times_2 \cdots \times_K \mathbf{M}_K \right)_{i_1, \dots, i_K} \\ &= F_{noiseless}(\hat{\mathbf{M}}_k). \end{aligned}$$

Therefore, we use the self-consistency when we derive the misclassification error. Note that the result that when  $MCR(\hat{\mathbf{M}}_k, \mathbf{M}_k) \geq \epsilon$ ,

$$G_{noise}(\hat{\mathbf{M}}_k) - G_{noise}(\mathbf{M}_k) \leq -\frac{\epsilon}{4a} \tau^{K-1} \delta \tag{2}$$

implies the self-consistency of  $\mathbf{M}_k$ . To obtain the result (2), we require

1. the convexity of  $g$  and  $\sup_x g''(x) \geq a$ ;
2. minimal gap strictly larger than 0, i.e.,  $\delta > 0$ .

## 2 General loss function

Consider the model

$$\mathbb{E}[\mathcal{Y}] = f(\Theta), \quad \text{where } \Theta = \mathcal{C} \times_1 \mathbf{M}_1 \times_2 \cdots \times_K \mathbf{M}_K.$$

**Theorem 2.1** (General property for loss function to guarantee the clustering accuracy). *Let  $\{\mathcal{C}, \mathbf{M}_k\}$  denote the true parameters, and  $\mathcal{L}_{\mathcal{Y}}(\mathcal{C}', \mathbf{M}'_k)$  denote the sample-based loss function. Define the sample-based loss function with respect to  $\mathbf{M}'_k$  as*

$$F(\mathbf{M}'_k) = \mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{C}}(\mathbf{M}'_k), \mathbf{M}'_k),$$

where

$$\hat{\mathcal{C}}(\mathbf{M}'_k) = \arg \max_{\mathcal{C}} \mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \mathbf{M}'_k).$$

Correspondingly, define the population-based loss function with respect to  $\mathbf{M}'_k$  as

$$G(\mathbf{M}'_k) = l(\tilde{\mathcal{C}}(\mathbf{M}'_k), \mathbf{M}'_k),$$

where

$$l(\mathcal{C}, \mathbf{M}_k) = \mathbb{E}_{\mathcal{Y}}[\mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \mathbf{M}_k)], \quad \text{and} \quad \tilde{\mathcal{C}}(\mathbf{M}'_k) = \arg \max_{\mathcal{C}} l(\mathcal{C}, \mathbf{M}'_k).$$

Suppose the loss function satisfies the following properties

1. (Self-consistency to  $\mathbf{M}_k$ ) Suppose  $MCR(\mathbf{M}'_k, \mathbf{M}_k) \geq \epsilon$  for  $\epsilon > 0$ . We have

$$G(\mathbf{M}'_k) - G(\mathbf{M}_k) \leq -C(\epsilon), \tag{3}$$

where  $C(\cdot)$  takes positive values.

2. (Bounded difference between sample- and population-based loss) The difference between sample-based and population-based loss function is bounded in probability, i.e.,

$$p(t) = \mathbb{P}(|F(\mathbf{M}'_k) - G(\mathbf{M}'_k)| \geq t) \rightarrow 0, \quad \text{as } t \rightarrow \infty. \tag{4}$$

Let  $\{\hat{\mathbf{M}}_k\}$  be the maximizer of  $F(\mathbf{M}_k)$ . Then, we have the following clustering accuracy, for any  $\epsilon > 0$ ,

$$\mathbb{P}(MCR(\hat{\mathbf{M}}_k, \mathbf{M}_k) \geq \epsilon) \leq p\left(\frac{C(\epsilon)}{2}\right).$$

*Proof.* Since  $\{\hat{\mathcal{C}}, \hat{\mathbf{M}}_k\}$  is the maximizer of the population-based objective function  $\mathcal{L}_{\mathcal{Y}}$ , we have

$$\begin{aligned} 0 &\leq F(\hat{\mathbf{M}}_k) - F(\mathbf{M}_k) \\ &= F(\hat{\mathbf{M}}_k) - G(\hat{\mathbf{M}}_k) + G(\hat{\mathbf{M}}_k) - G(\mathbf{M}_k) + G(\mathbf{M}_k) - F(\mathbf{M}_k). \end{aligned}$$

Suppose  $MCR(\hat{\mathbf{M}}_k, \mathbf{M}_k) \geq \epsilon$ . By the property (3), we have

$$0 \leq 2r - C(\epsilon),$$

where  $r = \sup_{\mathbf{M}'_k} |F(\mathbf{M}'_k) - G(\mathbf{M}'_k)|$ . Therefore, we have

$$\begin{aligned}\mathbb{P}(MCR(\hat{\mathbf{M}}_k, \mathbf{M}_k) \geq \epsilon) &= \mathbb{P}(G(\mathbf{M}'_k) - G(\mathbf{M}_k) \leq -C(\epsilon)) \\ &\leq \mathbb{P}(C(\epsilon) \leq 2r) \\ &= p\left(\frac{C(\epsilon)}{2}\right),\end{aligned}$$

where the last equation follows the second property (4).  $\square$

**Remark 4.** For the model in Tensor Block model, we have

$$C(\epsilon) = \frac{\epsilon}{4a} \tau^{K-1} \delta,$$

where  $a$  is the upper bound of  $g''(x)$ ,  $\tau$  is minimal proportion of the cluster, and  $\delta$  is the minimal gap between blocks. By the sub-Gaussianity of  $\mathcal{Y}$  and Hoeffding's inequality, we have

$$\begin{aligned}p(t) &\leq \mathbb{P}(C_1 \|g'(\hat{c}_{r_1, \dots, r_K}) - \mathbb{E}[g'(\hat{c}_{r_1, \dots, r_K})]\|_{\max} \geq t) \\ &\leq \mathbb{P}\left(\sup_{I_{r_1, \dots, r_K}} \frac{|\sum_{(i_1, \dots, i_K) \in I_{r_1, \dots, r_K}} \mathcal{Y}_{i_1, \dots, i_K} - \mathbb{E}[\mathcal{Y}_{i_1, \dots, i_K}]|}{|I_{r_1, \dots, r_K}|} \geq \frac{t}{C_1}\right) \\ &\leq 2^{1+\sum_k d_k} \exp\left(-\frac{t^2 L}{C_1^2}\right),\end{aligned}$$

where  $C_1$  is a positive constant related to the true core tensor  $\mathcal{C}$ ,  $I_{r_1, \dots, r_K}$  is the index set of the block  $(r_1, \dots, r_K)$  based on the estimate membership  $\hat{\mathbf{M}}_k$ , and  $L = \inf |I_{r_1, \dots, r_K}| \geq \tau^K \prod_k d_k$ .

**Remark 5.** When  $\tilde{\mathcal{C}}(\mathbf{M}_k) = \mathcal{C}$ , i.e.,  $g' = f$  in the tensor block model, the self-consistency to  $\mathbf{M}_k$  implies the self-consistency to  $\{\mathcal{C}, \mathbf{M}_k\}$  or  $\Theta = \mathcal{C} \times_1 \mathbf{M}_1 \times_2 \dots \times_K \mathbf{M}_K$ .

### 3 Precision matrix model

The precision model is stated as

$$\mathbb{E}[S^k] = \Omega^k = \sum_{l=1}^r u_{kl} \Theta^l, \quad k \in [K].$$

**Without the sparsity penalty**, we consider the optimization problem

$$\max_{\mathbf{U}, \Theta^l} \mathcal{L}_S(\mathbf{U}, \Theta^l) = - \sum_{k=1}^K \text{tr}(S^k \Omega^k) + \log \det(\Omega^k),$$

where  $\mathbf{U}$  is a membership matrix, and  $\{\Theta^l\}$  are irreducible and invertible.

**Proposition 1.** *The loss function  $\mathcal{L}_S$  satisfies the conditions for Theorem 2.1, and thus the clustering accuracy for precision matrix model is guaranteed.*

*Proof.* First, we define the functions  $F(\cdot)$  and  $G(\cdot)$  in the Theorem 2.1 under the precision matrix context.

Given the membership matrix  $\mathbf{U}'$ , we want to find the estimate  $\hat{\Theta}^l(\mathbf{U}') = \arg \max_{\Theta^l} \mathcal{L}_S(\mathbf{U}', \Theta^l)$ . Note that the  $\mathcal{L}_S(\mathbf{U}', \Theta^l)$  is concave respect to  $\Theta^l$ . Then, by the first order condition, we have

$$\hat{\Theta}^l = \left( \frac{\sum_{k \in I'_l} S^k}{|I'_l|} \right)^{-1},$$

where  $I'_l = \{k : u_{kl} = 1\}, l \in [r]$ . Thus, we obtain the function  $F(\mathbf{U}') = \mathcal{L}_S(\mathbf{U}', \hat{\Theta}^l(\mathbf{U}'))$ , which is

$$F(\mathbf{U}') = - \sum_{l=1}^r |I'_l| p + |I'_l| \log \det \left( \frac{\sum_{k \in I'_l} S^k}{|I'_l|} \right)^{-1}.$$

Note that

$$l(\mathbf{U}', \Theta^l) = \mathbb{E}_S[\mathcal{L}_S(\mathbf{U}', \Theta^l)] = - \sum_{k=1}^K \text{tr}(\Sigma^k \Omega^k) + \log \det(\Omega^k).$$

Therefore, we have

$$\tilde{\Theta}^l(\mathbf{U}') = \left( \frac{\sum_{k \in I'_l} \Sigma^k}{|I'_l|} \right)^{-1},$$

and

$$G(\mathbf{U}') = l(\mathbf{U}', \tilde{\Theta}^l(\mathbf{U}')) = - \sum_{l=1}^r |I'_l| p + |I'_l| \log \det \left( \frac{\sum_{a=1}^r D_{al} \Sigma^a}{|I'_l|} \right)^{-1},$$

where  $D_{al}$  is the elements of the confusion matrix.

Next, we verify the functions  $F(\cdot)$  and  $G(\cdot)$  satisfy the conditions in the Theorem 2.1. Let  $\{\mathbf{U}, \Theta^l\}$  denote the true membership and precision matrices, and  $\hat{\mathbf{U}}$  denote the estimated  $\mathbf{U}$  which maximizes  $F(\mathbf{U})$ .

#### 1. (Self-consistency to $\mathbf{U}$ )

Consider the subtraction

$$G(\hat{\mathbf{U}}) - G(\mathbf{U}) = - \sum_{l=1}^r \log \det \left( \frac{\sum_{a=1}^r D_{al} \Sigma^a}{|\hat{I}_l|} \right) + \sum_{l=1}^r \left( \frac{\sum_{a=1}^r D_{al} \log \det(\Sigma^a)}{|\hat{I}_l|} \right).$$

Since  $MCR(\hat{\mathbf{U}}, \mathbf{U}) \geq \epsilon$ , there exist  $l, k \neq k' \in [r]$  such that  $\min\{D_{kl}, D_{k'l}\} \geq \epsilon$ . Let  $\tilde{\Sigma} = \frac{\sum_{a=1}^r D_{al} \Sigma^a}{|\hat{I}_l|}$ . Consider the function  $f(t) = \log \det(\tilde{\Sigma} + t\Delta)$ , where  $\Delta = \Sigma - \tilde{\Sigma}$ . By Taylor Expansion, we have

$$\log \det(\Sigma) - \log \det(\tilde{\Sigma}) = f(1) - f(0) = f'(0) + \frac{f''(\xi)}{2}, \quad \text{for some } \xi \in [0, 1],$$

where

$$f'(0) = \langle (\tilde{\Sigma})^{-1}, \Delta \rangle, \quad \text{and} \quad f''(\xi) = -\text{vec}(\Delta)^T (\tilde{\Sigma} + \xi\Delta)^{-1} \otimes (\tilde{\Sigma} + \xi\Delta)^{-1} \text{vec}(\Delta) \leq -s \|\Delta\|_F^2, \quad (5)$$

where  $s$  is a positive constant which  $s \leq \varphi_{\max}^{-2}(\tilde{\Sigma} + \xi\Delta)$ .

Let  $\Delta^l = \Sigma^l - \tilde{\Sigma}, l \in [r]$ . With the Taylor Expansion (5), we have

$$\begin{aligned}
\left( \frac{\sum_{a=1}^r D_{al} \log \det(\Sigma^a)}{|\hat{I}_l|} \right) - \log \det(\tilde{\Sigma}) &= \sum_{a=1}^l \frac{D_{al}}{|\hat{I}_l|} \left[ \log \det(\Sigma^a) - \log \det(\tilde{\Sigma}) \right] \\
&\leq \sum_{a=1}^r \frac{D_{al}}{|\hat{I}_l|} \left( \langle (\tilde{\Sigma})^{-1}, \Delta^a \rangle - \frac{1}{2} s \|\Delta^a\|_F^2 \right) \\
&\leq -\frac{D_{kl}}{2|\hat{I}_l|} s \|\Delta^k\|_F^2 - \frac{D_{k'l}}{2|\hat{I}_l|} s \|\Delta^{k'}\|_F^2,
\end{aligned}$$

where the last inequality follows by the fact that  $\sum_{a=1}^r \frac{D_{al}}{|\hat{I}_l|} \langle \tilde{\Sigma}, \Delta^a \rangle = 0$ . By the inequality  $\frac{1}{2} \|A + B\|_F^2 \leq \|A\|_F^2 + \|B\|_F^2$ , we obtain that

$$\left( \frac{\sum_{a=1}^r D_{al} \log \det(\Sigma^a)}{|\hat{I}_l|} \right) - \log \det(\tilde{\Sigma}) \leq -\frac{\min\{D_{kl}, D_{k'l}\} s}{|\hat{I}_l|} \|\Sigma^k - \Sigma^{k'}\|_F^2 \leq -\frac{\epsilon}{4s|\hat{I}_l|} \delta.$$

For other  $l' \in [r]/l$ , since  $-\log \det(\cdot)$  is a convex function, by Jensen's inequality, we have

$$\left( \frac{\sum_{a=1}^r D_{al'} \log \det(\Sigma^a)}{|\hat{I}_{l'}|} \right) - \log \det \left( \frac{\sum_{a=1}^r D_{al'} \Sigma^a}{|\hat{I}_{l'}|} \right) \leq 0.$$

Then, we have

$$G(\hat{\mathbf{U}}) - G(\mathbf{U}) \leq -\frac{\epsilon}{4s} \delta,$$

which implies the self-consistency holds.

## 2. (Bounded difference between sample- and population-based loss)

For arbitrary  $\mathbf{U}$ , consider the absolute subtraction

$$|F(\mathbf{U}) - G(\mathbf{U})| \leq \sum_{l=1}^r |I_l| \left| \log \det \left( \frac{\sum_{k \in I_l} S^k}{|I_l|} \right) - \log \det \left( \mathbb{E} \left[ \frac{\sum_{k \in I_l} S^k}{|I_l|} \right] \right) \right|.$$

Consider the function  $f(t) = -\log \det \left( \frac{\sum_{k \in I_l} S^k}{|I_l|} + t\Delta \right)$ , where  $\Delta = \mathbb{E} \left[ \frac{\sum_{k \in I_l} S^k}{|I_l|} \right] - \frac{\sum_{k \in I_l} S^k}{|I_l|}$ .

By the previous calculation (5), we know that  $f(t)$  is a convex function. Then, the function is locally Lipschitz with  $L = \sup_t |f'(t)|$ . Therefore, we have

$$\begin{aligned}
|F(\mathbf{U}) - G(\mathbf{U})| &\leq \sum_{l=1}^r |I_l| |f(1) - f(0)| \\
&\leq \sum_{l=1}^r |I_l| |f'(1)| \\
&\leq K \sup \left| \left\langle \left( \mathbb{E} \left[ \frac{\sum_{k \in I_l} S^k}{|I_l|} \right] \right)^{-1}, \frac{\sum_{k \in I_l} S^k}{|I_l|} - \mathbb{E} \left[ \frac{\sum_{k \in I_l} S^k}{|I_l|} \right] \right\rangle \right| \\
&\leq K p^2 \max_{l \in [r]} \left\| \Theta^l \right\| \max_{k, (i,j)} |S_{(i,j)}^k - \mathbb{E}[S_{(i,j)}^k]|.
\end{aligned}$$

Therefore, by Lemma 2, we have

$$\begin{aligned}
p(t) &= \mathbb{P}(|F(\mathbf{U}) - G(\mathbf{U})| \geq t) \\
&\leq \mathbb{P}\left(Kp^2 \max_{l \in [r]} \|\Theta^l\|_{\max} \max_{k, (i,j)} |S_{(i,j)}^k - \mathbb{E}[S_{(i,j)}^k]| \geq t\right) \\
&\leq C_1 \exp\left(-C_2 \frac{\min_{k \in [K]} n_k t^2}{K^2 p^4 \max_{l \in [r]} \|\Theta^l\|_{\max}^2}\right),
\end{aligned}$$

where  $C_1, C_2$  are two constants. □

**Remark 6.** The above proof does not consider the sparsity constraints. Recall the general tensor block model. The convexity of  $g$  and boundedness of  $g''$  (as well as irreducibility of  $\mathcal{C}$ ) ensures the self-consistency of  $\mathbf{M}_k$ . In precision matrix model, if we add a convex sparsity penalty  $R(\Theta^l)$  (e.g.  $L_1, L_0$  norm) to the objective function, the nonlinear term  $-\log \det(\Omega^k) + R(\Theta^l)$  still keeps convex, which can be considered as the function “ $g$ ” in the precision matrix context. Therefore, my conjecture is that **the sparsity penalty to the objective function won’t affect the self-consistency to  $\mathbf{U}$** . Meanwhile, the difference between sample- and population-based is independent with the penalty. Thus, the loss function with sparsity penalty guarantees the clustering accuracy.

**Lemma 2.** Let  $Z_i \sim_{i.i.d.} \mathcal{N}(0, \Sigma)$  and  $\varphi_{\max}(\Sigma) \leq \tau < \infty$ . Let  $\Sigma = \llbracket \Sigma_{ij} \rrbracket$ , then

$$P\left(\left|\sum_{i=1}^n Z_{ij} Z_{ik} - n \Sigma_{jk}\right| \geq n\nu\right) \leq c_1 e^{-c_2 n \nu^2}, \quad \text{for } |\nu| \leq \delta,$$

where  $c_1, c_2, \delta$  depends on  $\tau$  only.

*Proof.* See Lemma 1 of Rothman et.al. □