

# Graphic Lasso: Clustering accuracy for precision matrix model

Jiixin Hu

February 18, 2021

## 1 With convex penalty $L_1$ norm

The precision model is stated as

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

Consider the following penalized optimization problem

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

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 3.1, and thus the clustering accuracy for precision matrix model is guaranteed.*

*Proof.* First, we introduce some useful notations.

Given the membership  $\mathbf{U}'$ , let  $\hat{\Theta}^l(\mathbf{U}') = \arg \max_{\Theta^l} \mathcal{L}_S(\mathbf{U}', \Theta^l)$ . Particularly, for each  $l \in [r]$ , we have

$$\hat{\Theta}^l(\mathbf{U}') = \arg \max_{\Theta} - \sum_{k \in I'_l} \langle S^k, \Theta \rangle + |I'_l| \log \det(\Theta) + \lambda |I'_l| \|\Theta\|_1,$$

index l

where  $I'_l = \{k : u'_{kl} \neq 0\}$  is the index set for the  $l$ -th group based on the membership  $\mathbf{U}'$ . The sample-based loss is defined as

$$F(\mathbf{U}') = \mathcal{L}_S(\mathbf{U}', \hat{\Theta}^l(\mathbf{U}')).$$

Correspondingly, define the population-based loss function as

$$l(\mathbf{U}, \Theta^l) = \mathbb{E}_S[\mathcal{L}_S(\mathbf{U}, \Theta^l)] = - \sum_{k=1}^K \text{tr}(\Sigma^k \Omega^k) + \log \det(\Omega^k) + \lambda \sum_{k=1}^K \left\| \Omega^k \right\|_1.$$

~~Given the membership  $\mathbf{U}'$ , let  $\tilde{\Theta}^l(\mathbf{U}') = \arg \max_{\Theta^l} \mathcal{L}_S(\mathbf{U}', \Theta^l)$ . Particularly, for each  $l \in [r]$ , we have~~

~~$$\tilde{\Theta}^l(\mathbf{U}') = \arg \max_{\Theta} - \sum_{k \in I'_l} \langle \Sigma^k, \Theta \rangle + |I'_l| \log \det(\Theta) + \lambda |I'_l| \|\Theta\|_1. \quad (1)$$~~

Then, the population-based loss is defined as

$$G(\mathbf{U}') = l(\mathbf{U}', \tilde{\Theta}^l(\mathbf{U}')).$$

**Note that  $\hat{\Theta}^l(\mathbf{U}')$  and  $\tilde{\Theta}^l(\mathbf{U}')$  do not have closed forms. But both of them only utilize  $|I_l'|$  sample covariance(true covariance) matrices based on the membership.**

Next, we verify the functions  $F(\cdot)$  and  $G(\cdot)$  satisfy the conditions in the Theorem 3.1. Let  $\{\mathbf{U}, \Theta^l\}$  denote the true membership and precision matrices, and define  $\hat{\mathbf{U}} = \arg \max_{\mathbf{U}} F(\mathbf{U})$ . We also define the confusion matrix  $D = \llbracket D_{ij} \rrbracket \in \mathbb{R}^{r \times r}$ , where  $D_{ij} = \sum_{k=1}^K \mathbf{I}\{u_{ki} = \hat{u}_{kj} = 1\}$ .

1. (Self-consistency) First, we consider the explicit formulas for  $G(\hat{\mathbf{U}})$  and  $G(\mathbf{U})$ .

$$\begin{aligned} G(\hat{\mathbf{U}}) &= l(\hat{\mathbf{U}}, \tilde{\Theta}^l(\hat{\mathbf{U}})) \\ &= \sum_{l=1}^r \left[ \sum_{k \in \hat{I}_l} -\langle \Sigma^k, \tilde{\Theta}^l(\hat{\mathbf{U}}) \rangle + |\hat{I}_l| \log \det(\tilde{\Theta}^l(\hat{\mathbf{U}})) - \lambda |\hat{I}_l| \left\| \tilde{\Theta}^l(\hat{\mathbf{U}}) \right\|_1 \right] \\ &= \sum_{l=1}^r \left[ \sum_{a=1}^r D_{al} \left( -\langle \Sigma^a, \tilde{\Theta}^l(\hat{\mathbf{U}}) \rangle + \log \det(\tilde{\Theta}^l(\hat{\mathbf{U}})) - \lambda \left\| \tilde{\Theta}^l(\hat{\mathbf{U}}) \right\|_1 \right) \right], \end{aligned}$$

and

$$\begin{aligned} G(\mathbf{U}) &= l(\mathbf{U}, \tilde{\Theta}^l(\mathbf{U})) \\ &= \sum_{l=1}^r \left[ -|I_l| \langle \Sigma^k, \tilde{\Theta}^l(\mathbf{U}) \rangle + |I_l| \log \det(\tilde{\Theta}^l(\mathbf{U})) - \lambda |I_l| \left\| \tilde{\Theta}^l(\mathbf{U}) \right\|_1 \right] \\ &= \sum_{l=1}^r \left[ \sum_{a=1}^r D_{al} \left( -\langle \Sigma^a, \tilde{\Theta}^l(\mathbf{U}) \rangle + \log \det(\tilde{\Theta}^l(\mathbf{U})) - \lambda \left\| \tilde{\Theta}^l(\mathbf{U}) \right\|_1 \right) \right]. \end{aligned}$$

Define the function

$$h^k(\Theta) = -\langle \Sigma^k, \Theta \rangle + \log \det(\Theta) - \lambda \|\Theta\|_1.$$

By the definition (1), we know that

$$\tilde{\Theta}^k(\mathbf{U}) = \arg \max_{\Theta} h^k(\Theta), k = 1, \dots, r.$$

Therefore, we have the self-consistency of  $\mathbf{U}$ , i.e.,  $G(\hat{\mathbf{U}}) \leq G(\mathbf{U})$ .

Next, we want to find the function which links the subtraction  $G(\hat{\mathbf{U}}) - G(\mathbf{U})$  with the misclassification rate  $MCR(\hat{\mathbf{U}}, \mathbf{U})$ , where  $MCR(\hat{\mathbf{U}}, \mathbf{U}) = \max_{l, a \neq a' \in [r]} \min\{D_{al}, D_{a'l}\}$ .

Suppose  $MCR(\hat{\mathbf{U}}, \mathbf{U}) \geq \epsilon$ . There exist  $l, k \neq k' \in [r]$  such that  $\min\{D_{kl}, D_{k'l}\} \geq \epsilon$ .

Then, we have

$$\begin{aligned} G(\hat{\mathbf{U}}) - G(\mathbf{U}) &\leq D_{kl} \left( h^k(\tilde{\Theta}^l(\hat{\mathbf{U}})) - h^k(\tilde{\Theta}^k(\mathbf{U})) \right) + D_{k'l} \left( h^k(\tilde{\Theta}^l(\hat{\mathbf{U}})) - h^k(\tilde{\Theta}^{k'}(\mathbf{U})) \right) \\ &\leq \epsilon C(\mathbf{U}, \Theta^l, \lambda), \end{aligned}$$

where  $C$  is a function of the true parameters  $\{\mathbf{U}, \Theta^l\}$ . **Need to figure out the explicit form of  $C$  in next step.**

2. (Bounded difference between sample- and population-based loss) For arbitrary  $\mathbf{U}$ , consider the absolute subtraction

$$\begin{aligned} |F(\mathbf{U}) - G(\mathbf{U})| &= |\mathcal{L}_S(\mathbf{U}, \hat{\Theta}^l(\mathbf{U})) - l(\mathbf{U}, \tilde{\Theta}^l(\mathbf{U}))| \\ &\leq |\mathcal{L}_S(\mathbf{U}, \hat{\Theta}^l(\mathbf{U})) - l(\mathbf{U}, \hat{\Theta}^l(\mathbf{U}))| + |l(\mathbf{U}, \hat{\Theta}^l(\mathbf{U})) - l(\mathbf{U}, \tilde{\Theta}^l(\mathbf{U}))| \\ &= M_1 + M_2. \end{aligned}$$

**Conjecture:**

For  $M_1$ ,

$$M_1 = \left| \sum_{l=1}^r \sum_{k \in I_l} \langle \Sigma^k - S^k, \hat{\Theta}^l(\mathbf{U}) \rangle \right| = \max_{k, (ij)} |\Sigma_{ij}^k - S_{ij}^k| C_1(\mathbf{U}, \Theta^l, p),$$

use Cauchy-Schwarz inequality.  
multiple ways to bound by,  $l_1$ ,  $l_2$ ,  $l_\infty$  norm, etc.

where  $C_1$  is a function of the true parameters  $\{\mathbf{U}, \Theta^l\}$  and the dimension  $p$ .

For  $M_2$ , note that  $l(\mathbf{U}, \Theta)$  is a convex function of  $\Theta$  and thus  $l$  is local Lipschitz. We may have

$$M_2 \leq \max_{l \in [r]} \sup_{\Theta^l} \left| \frac{\partial}{\partial \Theta^l} l(\mathbf{U}, \Theta^l) \right| \left\| \hat{\Theta}^l(\mathbf{U}) - \tilde{\Theta}^l(\mathbf{U}) \right\|_{\max}$$

Also, we can consider  $\max_{l \in [r]} \sup_{\Theta^l} \left| \frac{\partial}{\partial \Theta^l} l(\mathbf{U}, \Theta^l) \right| = C_2(\mathbf{U}, \Theta^l, \lambda)$ , where  $C_2$  is the function of the true parameters  $\{\mathbf{U}, \Theta^l\}$  and tuning parameter  $\lambda$ . Since  $\hat{\Theta}^l$  is the sample-based estimation and  $\tilde{\Theta}^l$  is the population-based estimation, my conjecture is that  $\left\| \hat{\Theta}^l(\mathbf{U}) - \tilde{\Theta}^l(\mathbf{U}) \right\|_{\max} = C_3(\max_{k, (ij)} |\Sigma_{ij}^k - S_{ij}^k|)$ .

Therefore, we bound the difference as

$$|F(\mathbf{U}) - G(\mathbf{U})| \leq C'(\mathbf{U}, \Theta^l, p, \lambda) C''(\max_{k, (ij)} |\Sigma_{ij}^k - S_{ij}^k|),$$

and then we can utilize of residual to find a  $p(t) = \mathbb{P}(|F(\mathbf{U}) - G(\mathbf{U})| \geq t) \rightarrow 0$  as  $t \rightarrow \infty$ .

□

## 2 Misclassification error

We explore the perturbed version of the self-consistency in this section.

**Lemma 1** (Self-consistency of  $\mathbf{U}$ ). *Suppose  $MCR(\hat{\mathbf{U}}, \mathbf{U}) \geq \epsilon$  and the minimal gap between  $\{\Theta^l\}$  denoted  $\delta$  is positive. For  $\lambda \leq C' \left( \frac{\log p}{n} \right)^{1/2}$  with some constant  $C'$ , we have the perturbation version of the self-consistency.*

how? additional condition on  $\lambda$

$$G(\hat{\mathbf{U}}) - G(\mathbf{U}) \leq -\frac{\epsilon}{4\tau^2} \delta^2 + \epsilon \lambda \sqrt{p} C \left( \frac{p \log p}{n} \right)^{1/2} < 0.$$

*Proof.* Suppose  $MCR(\hat{\mathbf{U}}, \mathbf{U}) \geq \epsilon$ . Let  $\{\mathbf{U}, \Theta^l\}$  denote the true parameters, and  $\Theta^l = (\Sigma^l)^{-1}$ . Define the function

$$h^k(\Theta) = -\langle \Sigma^k, \Theta \rangle + \log \det(\Theta) - \lambda \|\Theta\|_1.$$

There exist  $l, k \neq k' \in [r]$  such that  $\min\{D_{kl}, D_{k'l}\} \geq \epsilon$ . Then, we have

$$\begin{aligned} G(\hat{\mathbf{U}}) - G(\mathbf{U}) &\leq D_{kl} \left( h^k(\tilde{\Theta}^l(\hat{\mathbf{U}})) - h^k(\tilde{\Theta}^k(\mathbf{U})) \right) + D_{k'l} \left( h^{k'}(\tilde{\Theta}^l(\hat{\mathbf{U}})) - h^{k'}(\tilde{\Theta}^{k'}(\mathbf{U})) \right) \\ &\leq D_{kl} \left( h^k(\tilde{\Theta}^l(\hat{\mathbf{U}})) - h^k(\Theta^k) \right) + D_{k'l} \left( h^{k'}(\tilde{\Theta}^l(\hat{\mathbf{U}})) - h^{k'}(\Theta^k) \right), \end{aligned} \quad (2)$$

where the second inequality follows the fact that  $h^k(\Theta^k) \leq h^k(\tilde{\Theta}^k(\mathbf{U}))$  since  $h^k(\tilde{\Theta}^k(\mathbf{U}))$  is the maximizer of  $h^k(\Theta)$  by the definition. For simplicity, let  $\hat{\Theta}$  denote  $\tilde{\Theta}^l(\hat{\mathbf{U}})$ . Define  $\Delta^k = \hat{\Theta} - \Theta^k$ . Consider the function

$$f^k(t) = \log \det(\Theta^k + t\Delta),$$

and by Taylor expansion we have

$$f^k(1) - f^k(0) = \langle \Sigma^k, \Delta^k \rangle - \text{vec}(\Delta^k)^T \int_0^1 (1-v)(\Theta^k + v\Delta^k)^{-1} \otimes (\Theta^k + v\Delta^k)^{-1} d\text{vvec}(\Delta^k).$$

Then, we have

$$\begin{aligned} h^k(\tilde{\Theta}^k) - h^k(\hat{\Theta}^k) &= \langle \Sigma^k, \Delta^k \rangle - f^k(1) + f^k(0) - \lambda \left( \|\Theta^k\|_1 - \|\hat{\Theta}\|_1 \right) \\ &\geq A_1 - |A_2|, \end{aligned}$$

where

$$\begin{aligned} A_1 &= \text{vec}(\Delta^k)^T \int_0^1 (1-v)(\Theta^k + v\Delta^k)^{-1} \otimes (\Theta^k + v\Delta^k)^{-1} d\text{vvec}(\Delta^k) \\ A_2 &= \lambda \left( \|\Theta^k\|_1 - \|\hat{\Theta}\|_1 \right). \end{aligned}$$

By Guo's paper, we know that

$$A_1 \geq \frac{1}{4\tau^2} \|\Delta^k\|_F^2, \quad (3)$$

where  $\max_{k \in [r]} \varphi_{\max}(\Theta^k) \leq \tau < \infty$ . Also note that

$$|A_2| \leq \lambda \|\Theta^k - \hat{\Theta}\|_1 \leq \lambda \sqrt{p} \|\Delta^k\|_F. \quad (4)$$

Plug the inequalities (3) and (4) in to the inequality (2), we obtain that

$$G(\hat{\mathbf{U}}) - G(\mathbf{U}) \leq D_{kl} \left( -\frac{1}{4\tau^2} \|\Delta^k\|_F^2 + \lambda \sqrt{p} \|\Delta^k\|_F \right) + D_{k'l} \left( -\frac{1}{4\tau^2} \|\Delta^{k'}\|_F^2 + \lambda \sqrt{p} \|\Delta^{k'}\|_F \right).$$

Intuitively, if we have  $\lambda$  very small, then we obtain the perturbation version of self-consistency. By a straightforward calculation, if we have

$$\lambda \leq \frac{1}{4\tau^2 \sqrt{p}} \min_{k \in [r]} \|\Delta^k\|_F, \quad (5)$$

then the perturbation version of self-consistency holds. Recall our previous conclusion for the  $\Omega$  estimation. If  $\lambda = \mathcal{O} \left( \left( \frac{\log p}{n} \right)^{1/2} \right)$ , we have

$$\min_{k \in [r]} \|\Delta^k\|_F \leq C \left( \frac{p \log p}{n} \right)^{1/2}$$

how to control other (k, l, k')

not directly applicable.  
Your earlier result (0115.pdf) for Omega is under constrained optimization, but not under penalized optimization. (?)

Extension should be easy though.

with high probability. This implies that when  $\lambda \leq C' \left( \frac{\log p}{n} \right)^{1/2}$ , the  $\lambda$  satisfies the condition (5) with high probability. Finally, we obtain the perturbation version of self-consistency,

$$\begin{aligned} G(\hat{\mathbf{U}}) - G(\mathbf{U}) &\leq -\frac{\epsilon}{4\tau^2} \left\| \Theta^k - \Theta^{k'} \right\|_F^2 + \epsilon\lambda\sqrt{p}C \left( \frac{p \log p}{n} \right)^{1/2} \\ &\leq -\frac{\epsilon}{4\tau^2} \delta^2 + \epsilon\lambda\sqrt{p}C \left( \frac{p \log p}{n} \right)^{1/2}, \end{aligned}$$

where  $\delta$  is the minimal gap between  $\Theta^l$ . □

**Remark 1.** When  $\lambda = 0$ , the subtraction  $G(\hat{\mathbf{U}}) - G(\mathbf{U}) \leq -\frac{1}{4\tau^2} \delta^2$  agrees with the result under the case without penalty.

**Remark 2.** The difficulty of the proof comes from that  $\tilde{\Theta}^l(\mathbf{U})$  does not have a closed form. In other literatures, they usually consider the true  $\Theta^l$  rather than  $\tilde{\Theta}^l(\mathbf{U})$  under the true membership. The possible reason is that the properties (such as singular value, minimal gap) of  $\Theta^l$  are easy to describe while it is hard to tell the properties of  $\tilde{\Theta}^l(\mathbf{U})$  (except it is an optimizer). Therefore, I introduce the true precision matrices in the proof in step (2). As a result, the upper bound becomes related with the precision matrices estimation  $\|\Delta\|_F = \|\hat{\Theta} - \Theta^k\|_F$ , and thus the control for  $\lambda$  is required.

### 3 Others

**Theorem 3.1** (General property for loss function to guarantee the clustering accuracy). *Let  $\{\mathcal{C}, \mathbf{M}_k\}$  denote the true parameters, and  $\mathcal{L}_Y(\mathcal{C}', \mathbf{M}'_k)$  denote the sample-based loss function. Define the sample-based loss function with respect to  $\mathbf{M}'_k$  as*

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

where

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

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

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

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

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

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.,

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

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

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