Graphic Lasso: Clustering accuracy for precision matrix model

Jiaxin Hu

February 17, 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

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

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

Proposition 1. *The loss function $\mathcal{L}_S$ satisfies the conditions for Theorem 1.1, and thus the clustering accuracy for precision matrix model is guaranteed.*

Proof. First, we introduce some useful notations.

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

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

where $I_l' = \{k : u_{kl}' \neq 0\}$ is the index set for the l-th group based on the membership $\boldsymbol{U}'$. The sample-based loss is defined as

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

Correspondingly, define the population-based loss function as

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

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

$$\tilde{\Theta}^l(\boldsymbol{U}') = \arg\max_{\Theta} -\sum_{k\in I_l'} \langle \Sigma^k, \Theta\rangle + |I_l'| \log\det(\Theta) + \lambda |I_l'| \left\|\Theta\right\|_1. \tag{1}$$

Then, the population-based loss is defined as

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

Note that $\hat{\Theta}^l(\boldsymbol{U}')$ and $\tilde{\Theta}^l(\boldsymbol{U}')$ 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 1.1. Let $\{\boldsymbol{U}, \Theta^l\}$ denote the true membership and precision matrices, and define $\hat{\boldsymbol{U}} = \arg\max_{\boldsymbol{U}} F(\boldsymbol{U})$. We also define the confusion matrix $D = [\![D_{ij}]\!] \in \mathbb{R}^{r\times r}$, where $D_{ij} = \sum_{k=1}^K \boldsymbol{I}\{u_{ki} = \hat{u}_{kj} = 1\}$.

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

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

and

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

Define the function

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

By the definition (1), we know that

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

Therefore, **we have the self-consistency** of $\boldsymbol{U}$, i.e., $G(\boldsymbol{U}') \le G(\boldsymbol{U})$.

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

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

Then, we have

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

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

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

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

Conjecture:

For M_1,

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

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

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

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

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

Therefore, we bound the difference as

$$|F(\boldsymbol{U})-G(\boldsymbol{U})|\leq C'(\boldsymbol{U},\Theta^l,p,\lambda)C''(\max_{k,(ij)}|\Sigma^k_{ij}-S^k_{ij}|),$$

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

□

Theorem 1.1 (General property for loss function to guarantee the clustering accuracy)**.** *Let $\{\mathcal{C},\boldsymbol{M}_k\}$ denote the true parameters, and $\mathcal{L}_{\mathcal{Y}}(\mathcal{C}',\boldsymbol{M}'_k)$ denote the sample-based loss function. Define the sample-based loss function with respect to $\boldsymbol{M}'_k$ as*

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

where

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

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

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

where

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

Suppose the loss function satisfies the following properties

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

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

   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.,*

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

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

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