Graphic Lasso: Miscellaneous

Jiaxin Hu

February 11, 2021

1 Corrected proof for sufficient condition

Consider the

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

and the optimization problem

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

Theorem 1.1 (Sufficient condition)**.** *Let* $\{\mathcal{C}, \boldsymbol{M}_k\}$ *denote the true parameters and* $\{\hat{\mathcal{C}}, \hat{\boldsymbol{M}}_k\}$ *denote the maximizer of the objective function* (1)*. The minimal sufficient conditions to obtain the clustering accuracy in form of*

notation conflicts. p is also used in your assumption 1.

$$\mathbb{P}(MCR(\hat{\boldsymbol{M}}_k, \boldsymbol{M}_k) \geq \epsilon) \leq p(\epsilon, \delta), \quad \textit{where} \quad p(\epsilon,\delta) \to 0, \quad \textit{as} \quad \epsilon \to 1$$

include

Is y a function of Theta, or a function of C?

I do not understand what restriction this assumption actually imposes.

1. *The function* g *is convex,* $\sup_{x=f(c_{r_1,\ldots,r_K})}(g')^{-1}(x) \leq p(\mathcal{C})$*, where* $p(\cdot)$ *is a function of the true parameter* $\mathcal{C}$*, and* $\sup_x g''(x) \leq a$*, where* a *is a positive constant.*

The current statement reads like left-hand-side is always a function of c

2. *The minimal gap between blocks is strictly larger than 0, i.e.,* $\delta = \min_k \delta^{(k)} > 0$ *, where*

$$\delta^{(k)} = \min_{r_k \neq r_k'} \max_{r_1,\ldots,r_{k-1},r_{k+1},\ldots,r_K} (f(c_{r_1,\ldots,r_k,\ldots,r_K}) - f(c_{r_1,\ldots,r_k',\ldots,r_K}))^2.$$

3. *The observation satisfies the assumptions for Hoeffding's inequality, i.e., each entry of* $\mathcal{Y}$ ~~*is bounded in* $[a, b]$ *or*~~ *sub-Gaussian with parameter* σ*.*

bounded r.v. must be sub-Gaussian

Proof. We proof the sufficiency in following steps:

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

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

The estimate is unique since g is convex.

2. We define following useful functions. First, let $F(\hat{\boldsymbol{M}}_k) = \mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{C}}, \hat{\boldsymbol{M}}_k)$, where $\hat{\mathcal{C}} = \hat{\mathcal{C}}(\hat{\boldsymbol{M}}_k)$ is the estimate depends on $\hat{\boldsymbol{M}}_k$. We have

$$\begin{aligned} F(\hat{\boldsymbol{M}}_k) &= \langle \mathcal{Y} \times_1 \hat{\boldsymbol{M}}_1^T \times_2 \cdots \times_K \boldsymbol{M}_K^T, \hat{\mathcal{C}} \rangle - \sum_{i_1,\ldots,i_K} g([\hat{\mathcal{C}} \times_1 \hat{\boldsymbol{M}}_1 \times_2 \cdots \times_K \boldsymbol{M}_K]_{i_1,\ldots,i_K}) \\ &= \sum_{r_1,\ldots,r_K} \prod_k d_k \prod_k \hat{p}_{r_k}^{(k)} [g'(\hat{c}_{r_1,\ldots,r_K}) \hat{c}_{r_1,\ldots,r_K} - g(\hat{c}_{r_1,\ldots,r_K})] \\ &= \sum_{r_1,\ldots,r_K} \prod_k d_k \prod_k \hat{p}_{r_k}^{(k)} h(g'(\hat{c}_{r_1,\ldots,r_K})), \end{aligned}$$

where $h(x) = x(g')^{-1}(x) - g((g')^{-1}(x))$. Correspondingly, we define

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

where $\mathbb{E}[g'(\hat{c}_{r_1,\ldots,r_K})] = \frac{1}{\prod_k \hat{p}_{r_k}^{(k)}} [f(\mathcal{C}) \times_1 D_1^T \times_2 \cdots \times_K D_K^T]$.

Then, for the true parameters $\{\mathcal{C}, \boldsymbol{M}_k\}$, we have

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

and

$$G(\boldsymbol{M}_k) = \sum_{r_1,\ldots,r_K} \prod_k d_k \prod_k p_{r_k}^{(k)} h(\mathbb{E}[g'(\hat{c}_{r_1,\ldots,r_K})]) = \sum_{r_1,\ldots,r_K} \prod_k d_k \prod_k p_{r_k}^{(k)} h(f(c_{r_1,\ldots,r_K})).$$

3. Consider the difference between $F(\boldsymbol{M}_k')$ and $G(\boldsymbol{M}_k')$. Since g is convex and $h''(x) = \frac{1}{g''((g')^{-1}(x))} > 0$, then the function h is convex and thus h is local Lipschitz. Note that $h'(x) = (g')^{-1}(x)$. Therefore, we have

$$|F(\boldsymbol{M}_k') - G(\boldsymbol{M}_k')| \leq p(\mathcal{C}) \left\| g'(\hat{c}_{r_1,\ldots,r_K}) - \mathbb{E}[g'(\hat{c}_{r_1,\ldots,r_K})] \right\|_{\max}. \tag{2}$$

4. Consider the misclassification error. With assumption 1,2, we satisfy the condition for Lemma 1. Therefore, we have

$$G(\hat{\boldsymbol{M}}_k) - G(\boldsymbol{M}_k) \leq -\frac{\epsilon}{4a} \tau^{K-1} \delta. \tag{3}$$

5. Combining step (2) with step (3), we obtain the accuracy

$$\begin{aligned} \mathbb{P}(MCR(\hat{\boldsymbol{M}}_k, \boldsymbol{M}_k) \geq \epsilon) &\leq \mathbb{P}\left(\sup_{\{\boldsymbol{M}_k\}} \left\| g'(\hat{c}_{r_1,\ldots,r_K}) - \mathbb{E}[g'(\hat{c}_{r_1,\ldots,r_K})] \right\|_{\max} \geq \frac{\epsilon}{8ap(\mathcal{C})} \tau^{K-1} \delta \right) \\ &\leq \mathbb{P}\left(\sup_{I_{r_1,\ldots,r_K}} \frac{\sum_{(i_1,\ldots,i_K) \in I_{r_1,\ldots,r_K}} \mathcal{Y}_{i_1,\ldots,i_K} - \mathbb{E}[\mathcal{Y}_{i_1,\ldots,i_K}]}{|I_{r_1,\ldots,r_K}|} \geq \frac{\epsilon}{8ap(\mathcal{C})} \tau^{K-1} \delta \right) \\ &\leq 2^{1+\sum d_k} \exp\left(-\frac{\epsilon^2 \tau^{2K-2} \delta^2 L}{C\sigma^2 ap(\mathcal{C})^2} \right). \end{aligned}$$

does convexity of h implies the self-consistency?

□

Remark 1. Note that the proof does not utilize the self-consistency property. For misclassification error, we have

$$G(\hat{\boldsymbol{M}}_k) = \sum_{r_1,\ldots,r_K} \prod_k d_k \prod_k \hat{p}_{r_k}^{(k)} h\left(\frac{1}{\prod_k \hat{p}_{r_k}^{(k)}} [f(\mathcal{C}) \times_1 D_1^T \times_2 \cdots \times_K D_K^T]_{r_1,\ldots,r_K} \right),$$

and

$$\begin{aligned} G(\boldsymbol{M}_k) &= \sum_{r_1,\ldots,r_K} \prod_k d_k \prod_k p_{r_k}^{(k)} h(f(c_{r_1,\ldots,r_K})) \\ &= \sum_{r_1,\ldots,r_K} \prod_k d_k \prod_k \hat{p}_{r_k}^{(k)} \frac{1}{\prod \hat{p}_{r_k}^{(k)}} [h\left(f(\mathcal{C})\right) \times_1 D_1^T \times_2 \cdots \times_K D_K^T]_{r_1,\ldots,r_K}. \end{aligned}$$

The true parameter $\{\boldsymbol{M}_k\}$ is the maximizer of $G(\boldsymbol{M}_k)$ because of the convexity of h. The linearity of $g'(\hat{c}_{r_1,\ldots,r_K})$ is also crucial to take the advantage of Jensen's inequality.

where is it mentioned in the current Assumption?

Lemma 1. *Suppose minimal gap between blocks is strictly larger than 0, i.e., $\delta = \min_k \delta^{(k)} > 0$, and $h''(x) \geq \frac{1}{a}$. For an fixed $\epsilon > 0$, suppose $MCR(\hat{\boldsymbol{M}}_k, \boldsymbol{M}_k) \geq \epsilon$ for some $k \in [K]$. We have*

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

Proof. We provide the proof for $k = 1$. The proof for other $k \in [K]$ is similar. Since $MCR(\hat{\boldsymbol{M}}_1, \boldsymbol{M}_1) \geq \epsilon$, there exist some $r_1 \in [R_1]$ and $a_1 \neq a_1'$ such that $\min\{D_{a_1,r_1}^{(1)}, D_{a_1',r_1}^{(1)}\} \geq \epsilon$. Let $\mathcal{N} = [\![h(g'(c_{r_1,\ldots,r_K}))]\!]$ and $W = \prod_k \hat{p}_{r_k}^{(k)}$. Then, there exists c^* such that

$$\begin{aligned} &[\mathcal{N} \times_1 \boldsymbol{D}^{(1),T} \times_2 \cdots \times_K \boldsymbol{D}^{(K),T}]_{r_1,\ldots,r_K} \\ &= D_{a_1,r_1}^{(1)} D_{a_2,r_2}^{(2)} \cdots D_{a_K,r_K}^{(K)} h(g'(c_{a_1,\ldots,a_K})) + D_{a_1',r_1}^{(1)} D_{a_2,r_2}^{(2)} \cdots D_{a_K,r_K}^{(K)} h(g'(c_{a_1',\ldots,a_K})) \\ &+ (W - D_{a_1,r_1}^{(1)} D_{a_2,r_2}^{(2)} \cdots D_{a_K,r_K}^{(K)} - D_{a_1',r_1}^{(1)} D_{a_2,r_2}^{(2)} \cdots D_{a_K,r_K}^{(K)}) c^*. \end{aligned}$$

Define $\mu_{r_1,\ldots,r_K} = \frac{1}{\prod_k \hat{p}_{r_k}^{(k)}} [f(\mathcal{C}) \times_1 D_1^T \times_2 \cdots \times_K D_K^T]$. Then, by Taylor Expansion of function $h(\cdot)$ at the point $\mu_{r_1,\ldots,r_K}$, we have

$$\begin{aligned} &\frac{1}{W}[\mathcal{N} \times_1 \boldsymbol{D}^{(1),T} \times_2 \cdots \times_K \boldsymbol{D}^{(K),T}]_{r_1,\ldots,r_K} - h(\mu_{r_1,\ldots,r_K}) \\ &\geq \frac{1}{2W} D_{a_1,r_1}^{(1)} D_{a_2,r_2}^{(2)} \cdots D_{a_K,r_K}^{(K)} h''(\mu_{r_1,\ldots,r_K})(g'(c_{a_1,\ldots,a_K}) - \mu_{r_1,\ldots,r_K})^2 \\ &+ \frac{1}{2W} D_{a_1,r_1}^{(1)} D_{a_2,r_2}^{(2)} \cdots D_{a_K,r_K}^{(K)} h''(\mu_{r_1,\ldots,r_K})(g'(c_{a_1',\ldots,a_K}) - \mu_{r_1,\ldots,r_K})^2 \\ &+ \frac{1}{2W} (W - D_{a_1,r_1}^{(1)} D_{a_2,r_2}^{(2)} \cdots D_{a_K,r_K}^{(K)} - D_{a_1',r_1}^{(1)} D_{a_2,r_2}^{(2)} \cdots D_{a_K,r_K}^{(K)}) h''(\mu_{r_1,\ldots,r_K})(c^* - \mu_{r_1,\ldots,r_K})^2, \end{aligned}$$

where $h''(x) = \frac{1}{g''(g'^{,-1}(x))} \frac{1}{a}$. By the inequality $a^2 + b^2 \geq \frac{(a+b)^2}{2}$, we obtain that

$$\begin{aligned} &\frac{1}{W}[\mathcal{N} \times_1 \boldsymbol{D}^{(1),T} \times_2 \cdots \times_K \boldsymbol{D}^{(K),T}]_{r_1,\ldots,r_K} - h(\mu_{r_1,\ldots,r_K}) \\ &\geq \frac{1}{a4W} \min\{D_{a_1,r_1}^{(1)}, D_{a_1',r_1}^{(1)}\} D_{a_2,r_2}^{(2)} \cdots D_{a_K,r_K}^{(K)} (g'(c_{a_1,\ldots,a_K}) - g'(c_{a_1',\ldots,a_K}))^2. \end{aligned} \tag{4}$$

Noted $h(\cdot)$ is a convex function, for other $r_1' \in [R_1]/\{r_1\}$, by Jensen's inequality, we have

$$\frac{1}{W}[\mathcal{N} \times_1 \boldsymbol{D}^{(1),T} \times_2 \cdots \times_K \boldsymbol{D}^{(K),T}]_{r_1',\ldots,r_K} - h(\mu_{r_1',\ldots,r_K}) \geq 0. \tag{5}$$

Combing the inequality (4) and (5), we obtain that

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

where the inequality follows by the fact that $\sum_{r_k} D^{(k)}_{a_k r_k} = p^{(k)}_{a_k} \geq \tau$.

□

2 General Loss Function

Consider the model

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

Theorem 2.1 (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 to estimate $\{\mathcal{C}, \boldsymbol{M}_k\}$. Define the population-based loss function as*

$$l(\mathcal{C}', \boldsymbol{M}_k') = \mathbb{E}_{\mathcal{Y}}[\mathcal{L}_{\mathcal{Y}}(\mathcal{C}', \boldsymbol{M}_k')].$$

Good!

For all $\{\mathcal{C}', \boldsymbol{M}_k'\}$ in the parameter space, suppose the sample-based and population-based satisfies the following properties

In practice, we need to find the explicit form of C and p.
1) p depends on the sub-Gaussianity.
2) C depends on the specific problem.
What is the explicit form of C and p in your earlier tensor clustering example?
How about the precision matrix problem?

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

$$l(\mathcal{C}', \boldsymbol{M}_k') - l(\mathcal{C}, \boldsymbol{M}_k) \geq -C(\epsilon), \tag{6}$$

   where $C(\cdot)$ is the function of ϵ which takes positive value.

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.,*

$$p(t) = \mathbb{P}(|\mathcal{L}_{\mathcal{Y}}(\mathcal{C}', \boldsymbol{M}_k') - l(\mathcal{C}', \boldsymbol{M}_k')| \geq t) \to 0, \quad as \quad t \to \infty. \tag{7}$$

Let $\{\hat{\mathcal{C}}, \hat{\boldsymbol{M}}_k\}$ denote the maximizer of the $\mathcal{L}_{\mathcal{Y}}$. Then, we obtain the clustering accuracy, for any $\epsilon > 0$,

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

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

$$\begin{aligned} 0 &\leq \mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{C}}, \hat{\boldsymbol{M}}_k) - \mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \boldsymbol{M}_k) \\ &= \mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{C}}, \hat{\boldsymbol{M}}_k) - l(\hat{\mathcal{C}}, \hat{\boldsymbol{M}}_k) + l(\hat{\mathcal{C}}, \hat{\boldsymbol{M}}_k) - l(\mathcal{C}, \boldsymbol{M}_k) + l(\mathcal{C}, \boldsymbol{M}_k) - \mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \boldsymbol{M}_k). \end{aligned}$$

Suppose $MCR(\hat{\boldsymbol{M}}_k, \boldsymbol{M}_k) \geq \epsilon$. By the property (6), we have

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

where $r = \sup_{\mathcal{C}', \boldsymbol{M}_k'} |\mathcal{L}_{\mathcal{Y}}(\mathcal{C}', \boldsymbol{M}_k') - l(\mathcal{C}', \boldsymbol{M}_k')|$. Therefore, we have

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

where the last equation follows the second property (7). □