Graphic Lasso: Miscellaneous

Jiaxin Hu

February 11, 2021

1 Correct

1.1 Definitions

Consider the optimization

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

Some key definitions in the previous notes are not correct. Here is the correctness.

1. With given membership $\boldsymbol{M}_k$, the estimation of $\mathcal{C}=[\![c_{r_1,\ldots,r_K}]\!]$ is of form

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

2. We define the function $F(\boldsymbol{M}_k)=\mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{C}},\boldsymbol{M}_k)$, i.e.,

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

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

3. Correspondingly, we define $G(\boldsymbol{M}_k)$ as

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

1.2 Loss functions

The sample-based loss function is

$$\mathcal{L}_{\mathcal{Y}}(\mathcal{C}',\boldsymbol{M}_k')=\langle\mathcal{Y}\times_1(\boldsymbol{M}_1')^T\times_2\cdots\times_K(\boldsymbol{M}_K')^T,\mathcal{C}'\rangle-\sum_{(i_1,\ldots,i_K)}g(\mathcal{C}'\times_1\boldsymbol{M}_1'\times_2\cdots\times_K\boldsymbol{M}_k')_{i_1,\ldots,i_K}.$$

Correspondingly, the population-based loss function is

$$
\begin{aligned}
l(\mathcal{C}', \boldsymbol{M}'_k) &= \mathbb{E}_{\mathcal{Y}}[\mathcal{L}_{\mathcal{Y}}(\mathcal{C}', \boldsymbol{M}'_k)] \\
&= \langle f(\mathcal{C}) \times_1 D_1^T \times_2 \cdots \times_K D_K^T, \mathcal{C}' \rangle - \sum_{(i_1,\dots,i_K)} g(\mathcal{C}' \times_1 \boldsymbol{M}'_1 \times_2 \cdots \times_K \boldsymbol{M}'_k)_{i_1,\dots,i_K}.
\end{aligned}
$$

Therefore, we know that

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

where $\hat{\mathcal{C}}$ is the estimate depends on $\boldsymbol{M}_k$. On the other hand,the explicit form of $G(\boldsymbol{M}_k)$ is

$$
\begin{aligned}
G(\hat{\boldsymbol{M}}_k) &= \sum_{r_1,\dots,r_K} \prod_k p_{r_k}^{(k)} h\left(\frac{1}{\prod_k d_k \prod_k p_{r_k}^{(k)}} f(\mathcal{C}) \times_1 D_1^T \times_2 \cdots \times_K D_K^T \right) \\
&= \langle f(\mathcal{C}) \times_1 D_1^T \times_2 \cdots \times_K D_K^T, \tilde{\mathcal{C}} \rangle - \sum_{(i_1,\dots,i_K)} g(\tilde{\mathcal{C}} \times_1 \hat{\boldsymbol{M}}_1 \times_2 \cdots \times_K \hat{\boldsymbol{M}}_k)_{i_1,\dots,i_K}, \\
&= l(\tilde{\mathcal{C}}, \hat{\boldsymbol{M}}_k).
\end{aligned}
$$

where

$$
\tilde{c}_{r_1,\dots,r_K} = (g')^{-1} \left(\frac{1}{\prod_k d_k \prod_k p_{r_k}^{(k)}} f(\mathcal{C}) \times_1 D_1^T \times_2 \cdots \times_K D_K^T \right)_{r_1,\dots,r_K}.
$$

Also, with the true membership $\{\boldsymbol{M}_k\}$, the previous definition for $G(\boldsymbol{M}_k)$ is

$$
\begin{aligned}
G(\boldsymbol{M}_k) &= \sum_{r_1,\dots,r_K} \prod_k p_{r_k}^{(k)} h(g'(c_{r_1,\dots,r_K})) \\
&= \langle g'(\mathcal{C}) \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K, \mathcal{C} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K \rangle - \sum_{i_1,\dots,i_K} g(\mathcal{C} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K),
\end{aligned}
$$

which is not equal to the $l(\mathcal{C}, \boldsymbol{M}_k)$ because the first term in the inner product should be $f(\mathcal{C})$.

1.3 Self-consistency (conjecture)

My conjecture is that the upper bound for misclassification rate uses the self-consistency property implicitly.

Suppose we have self-consistency property. We have

$$
(\mathcal{C}, \boldsymbol{M}_k) = \underset{(\mathcal{C}', \boldsymbol{M}'_k)}{\arg\max}\, l(\mathcal{C}', \boldsymbol{M}'_k),
$$

which implies $f = g'$. Then, based on the definition of $G(\boldsymbol{M}_k)$, we have $G(\boldsymbol{M}_k) = l(\mathcal{C}, \boldsymbol{M}_k)$. Thus, it is natural that

$$
G(\hat{\boldsymbol{M}}_k) - G(\boldsymbol{M}_k) = l(\tilde{\mathcal{C}}, \hat{\boldsymbol{M}}_k) - l(\mathcal{C}, \boldsymbol{M}_k) < 0.
$$

If we do not have self-consistency, $G(\cdot)$ is not the population-based function for the true parameter, and thus the true membership $\{\boldsymbol{M}_k\}$ may not be the maximizer of $G(\cdot)$.

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*

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

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

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

$$l(\mathcal{C}', \boldsymbol{M}_k') - l(\mathcal{C}, \boldsymbol{M}_k) \leq -C(\epsilon), \tag{1}$$

   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.,*

$$p(t) = \mathbb{P}(|\mathcal{L}_{\mathcal{Y}}(\mathcal{C}', \boldsymbol{M}_k') - l(\mathcal{C}', \boldsymbol{M}_k')| \geq t) \to 0, \quad \text{as} \quad t \to \infty. \tag{2}$$

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 (1), 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 (2). $\square$