Graphic Lasso: Clustering accuracy

Jiaxin Hu

January 20, 2021

Consider the model

$$\mathbb{E}[\mathcal{Y}] = f(\mathcal{C} \times \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K),$$

where $\mathcal{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, $\mathcal{C} \in \mathbb{R}^{R_1 \times \cdots \times R_K}$, $\boldsymbol{M}_k \in \mathbb{R}^{d_k \times r_k}$ for all $k \in [K]$, and f is the link function. The quasi-likelihood function of $\{\mathcal{C}, \boldsymbol{M}_k\}$ is

$$\mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \boldsymbol{M}_k) = \langle \mathcal{Y}, \Theta \rangle + \sum_{i_1,\ldots,i_K} b(\Theta_{i_1,\ldots,i_K}),$$

where $\Theta = \mathcal{C} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K$, and $b'(\cdot) = f(\cdot)$.

Suppose we already know the membership $\{\boldsymbol{M}_k\}$. Let $I_{r_1,\ldots,r_K} = \{(i_1,\ldots,i_K) | \boldsymbol{M}_{k,i_k r_k} = 1, k \in [K]\}$ and $p^{(k)}_{r_k} = \frac{1}{d_k}\sum_{i=1}^{d_k} \boldsymbol{I}\{M_{k,ir_k} = 1\}$. Note that $I_{r_1,\ldots,r_K} = d_1 \cdots d_K p^{(1)}_{r_1} \cdots p^{(K)}_{r_K}$. The MLE of $\mathcal{C} = [\![c_{r_1,\ldots,r_K}]\!]$ satisfy the following equality,

$$\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial c_{r_1,\ldots,r_k}} = \sum_{(i_1,\ldots,i_K) \in I_{r_1,\ldots,r_K}} \mathcal{Y}_{i_1,\ldots,i_K} - |I_{r_1,\ldots,r_K}| b'(c_{r_1,\ldots,r_K}) = 0,$$

which implies that

$$\hat{c}_{r_1,\ldots,r_K} = (b')^{-1}\left(\frac{1}{d_1 \cdots d_K p^{(1)}_{r_1} \cdots p^{(K)}_{r_K}} \left[\mathcal{Y} \times_1 \boldsymbol{M}_1^T \times_2 \cdots \times_K \boldsymbol{M}_K^T \right]_{r_1,\ldots,r_K} \right).$$

Let $F(\boldsymbol{M}_k) = \mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{C}}, \boldsymbol{M}_k)$. Then, the function $F(\boldsymbol{M}_k)$ is of form

$$F(\boldsymbol{M}_k) = \sum_{r_1,\ldots,r_K} \prod_k p^{(k)}_{r_k} \left[b'(\hat{c}_{r_1,\ldots,r_K}) \hat{c}_{r_1,\ldots,r_K} - b(\hat{c}_{r_1,\ldots,r_K}) \right] = \sum_{r_1,\ldots,r_K} \prod_k p^{(k)}_{r_k} h(b'(\hat{c}_{r_1,\ldots,r_K})),$$

where $h(x) = x(b')^{-1}(x) - b((b')^{-1}(x))$. Let $G(\boldsymbol{M}_k) = \mathbb{E}[F(\boldsymbol{M}_k)]$ denote the expectation of $F(\boldsymbol{M}_k)$ with respect to $\hat{\mathcal{C}}$. Then, we have

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

In the paper, $G(\boldsymbol{M}_k) = \sum_{r_1,\ldots,r_K} \prod_k p^{(k)}_{r_k} h(\mathbb{E}[b'(\hat{c}_{r_1,\ldots,r_K})])$? In the arxiv version, equality 18 indicates $\mathbb{E}[\hat{c}^2_{r_1,\ldots,r_K}] = (\mathbb{E}(\hat{c}_{r_1,\ldots,r_K}))^2$. To go through the rest parts, we use the definition in red. Let $R(\boldsymbol{M}_k)$ denote the residual tensor, i.e.,

$$R(\boldsymbol{M}_k)_{r_1,\ldots,r_K} = b'(\hat{c}_{r_1,\ldots,r_K}) - \mathbb{E}[b'(\hat{c}_{r_1,\ldots,r_K})] \leq \frac{\sum_{(i_1,\ldots,i_K) \in I_{r_1,\ldots,r_K}} \mathcal{Y}_{i_1,\ldots,i_K} - \mathbb{E}[\mathcal{Y}_{i_1,\ldots,i_K}]}{|I_{r_1,\ldots,r_K}|}.$$

First, we have the upper bound of the estimation error

$$
\begin{aligned}
|F(\boldsymbol{M}_k) - G(\boldsymbol{M}_k)| &\leq \sum_{r_1,\ldots,r_K} \prod_k p_{r_k}^{(k)} |h(b'(\hat{c}_{r_1,\ldots,r_K})) - h(\mathbb{E}[(b'(\hat{c}_{r_1,\ldots,r_K})])| \\
&\leq \|\mathcal{C}\|_{\max} \|R(\boldsymbol{M}_k)\|_{\max},
\end{aligned}
$$

where the inequality follows $|h(b'(\hat{c}_{r_1,\ldots,r_K})) - h(\mathbb{E}[(b'(\hat{c}_{r_1,\ldots,r_K})])| \leq \sup_{x=b'(c_{r_1,\ldots,r_k})} h'(x) \|R(\boldsymbol{M}_k)\|_{\max}$ by Taylor Expansion, and $h'(x) = (b')^{-1}(x)$.

Next, we consider the upper bound of misclassification error

$$
G(\hat{\boldsymbol{M}}_k) - G(\boldsymbol{M}_k),
$$

where $\boldsymbol{M}_k$ denote the true membership,

$$
G(\hat{\boldsymbol{M}}_k) = \sum_{r_1,\ldots,r_K} \prod_k \hat{p}_{r_k}^{(k)} h(\mu_{r_1,\ldots,r_K}), \quad G(\boldsymbol{M}_k) = \sum_{r_1,\ldots,r_K} \prod_k p_{r_k}^{(k)} h(b'(c_{r_1,\ldots,r_K})),
$$

and

$$
\mu_{r_1,\ldots,r_K} = \mathbb{E}[b'(c_{r_1,\ldots,r_K})] = \frac{1}{\prod_k \hat{p}_{r_k}^{(k)}} [b'(\mathcal{C}) \times_1 \boldsymbol{D}^{(1),T} \times_2 \cdots \times_K \boldsymbol{D}^{(K),T}]_{r_1,\ldots,r_K}.
$$

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 $\mathcal{N} = [\![h(b'(c_{r_1,\ldots,r_K}))]\!]$ and $W = \prod_k \hat{p}_{r_k}^{(k)}$. Then, there exists c^* such that

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

Thus, by Taylor expansion for function $h(\cdot)$ at point $\mu_{r_1,\ldots,r_K}$, we have

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

Note that $h''(x) = \frac{1}{b''(b^{',-1}(x))}$, and $\text{Var}(Y_{i_1,\ldots,i_K}) = b''(b^{',-1}(c_{r_1,\ldots,r_K})) < \alpha$ for some $\alpha > 0$. This condition is not in Tensor Block model. Figure our whether we need it. By the inequality $a^2+b^2 \geq \frac{(a+b)^2}{2}$, we obtain that

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

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

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

Since $\mathcal{C}$ is irreducible, let δ_{min} denote the minimal gap between each cluster. Then, we have $(b'(c_{a_1,\ldots,a_K}) - b'(c_{a_1',\ldots,a_K}))^2 \geq \delta'$ for some $\delta' > 0$ because $b'(\cdot)$ is a one-to-one function. Therefore, we have

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

where the last line follows from $\sum_{r_k} D^{(k)}_{a_k r_k} = p^{(k)}_{a_k} \geq \tau$.

Back to the misclassification rate.

$$\mathbb{P}(MCR(\hat{\boldsymbol{M}}_k, \boldsymbol{M}_k) \geq \epsilon) \leq \mathbb{P}\left(G(\hat{\boldsymbol{M}}_k) - G(\boldsymbol{M}_k) \leq -\frac{\epsilon}{4\alpha}\tau^{K-1}\delta'\right). \tag{1}$$

Since $\{\hat{\boldsymbol{M}}_k\}$ is MLE, we have

$$0 \leq F(\hat{\boldsymbol{M}}_k) - F(\boldsymbol{M}_k) \leq 2r - \frac{\epsilon}{4\alpha}\tau^{K-1}\delta',$$

where $r = \sup|F(\boldsymbol{M}_k) - G(\boldsymbol{M}_k)|$. Plugging the above inequality into the probability (1), we have

$$\begin{aligned}&\mathbb{P}\left(G(\hat{\boldsymbol{M}}_k) - G(\boldsymbol{M}_k) \leq -\frac{\epsilon}{4\alpha}\tau^{K-1}\delta'\right)\\ &\leq \mathbb{P}\left(F(\hat{\boldsymbol{M}}_k) - F(\boldsymbol{M}_k) \leq 2r - \frac{\epsilon}{4\alpha W}\tau^{K-1}\delta'\right)\\ &\leq \mathbb{P}\left(r \geq \frac{\epsilon}{8\alpha}\tau^{K-1}\delta'\right)\\ &\leq \mathbb{P}\left(\sup_{I_{r_1,\ldots,r_K}} \frac{\sum_{(i_1,\ldots,i_K)\in I_{r_1,\ldots,r_K}} \mathcal{Y}_{i_1,\ldots,i_K} - \mathbb{E}[\mathcal{Y}_{i_1,\ldots,i_K}]}{|I_{r_1,\ldots,r_K}|} \geq \frac{\epsilon}{8\alpha\,\|\mathcal{C}\|_{\max}}\tau^{K-1}\delta'\right)\\ &\leq 2^{1+\sum d_k}\exp\left(-\frac{\epsilon^2\tau^{2K-2}\delta'^2 L}{C\sigma^2\alpha^2\,\|\mathcal{C}\|^2_{\max}}\right),\end{aligned}$$

where $L \geq \tau^K \prod_k d_k$, σ is the sub-Gaussian parameter.