Graphic Lasso: Clustering accuracy

Jiaxin Hu

January 23, 2021

1 Accuracy

Consider the model

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

where $\mathcal{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, $\mathcal{C} = [\![c_{r_1,...,r_K}]\!] \in \mathbb{R}^{R_1 \times \cdots \times R_K}$, $\boldsymbol{M}_k = \in \{0,1\}^{d_k \times r_k}$ for all $k \in [K]$ are membership matrices, and $f(\cdot)$ is the link function. Define the misclassification rate on the k-th mode as

$$MCR(\hat{\boldsymbol{M}}_k, \boldsymbol{M}_k) = \max_{r \in [R_k], a \neq a' \in [R_k]} \min\{D_{ar}^{(k)}, D_{a'r}^{(k)}, \}$$

where $D^{(k)} \in \mathbb{R}^{R_k \times R_k}$ is the confusion matrix on the k-th, and $D_{rr'}^{(k)} = \frac{1}{d_k} \sum_{i=1}^{d_k} \boldsymbol{I}\{\boldsymbol{M}_{k,ir_k} = \hat{\boldsymbol{M}}_{k,ir_k} = 1\}$. Define the minimal gap between blocks as $\delta = \min_k \delta^{(k)}$, where

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

Theorem 1.1. *Let $\{\mathcal{C}, \boldsymbol{M}_k\}$ denote the true parameters, and $\Theta = [\![\Theta_{i_1,...,i_K}]\!] = \mathcal{C} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K$. Suppose $0 < a_1 < Var(\mathcal{Y}_{i_1,...,i_K} | \Theta_{i_1,...,i_K}) < a_2 < \infty$. Let σ denote the sub-Gaussian parameter of $\mathcal{Y}$. For any $\epsilon \in [0,1]$, the MLE estimator $\{\hat{\boldsymbol{M}}_k\}$ satisfies the following bound*

$$\mathbb{P}(MCR(\hat{\boldsymbol{M}}_k, \boldsymbol{M}_k) \geq \epsilon) \leq 2^{1+\sum_k d_k} \exp\left(- \frac{C \epsilon^2 \tau^{3K-2} \delta^2 \prod_k d_k}{\sigma^2 a_2^2 \|\mathcal{C}\|_{\max}^2} \right),$$

where $\tau > 0$ is the lower bound the cluster proporition.

Proof. Recall the objective function in our model is

$$\mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \{\boldsymbol{M}_k\}) = \langle \mathcal{Y}, \Theta \rangle + \sum_{i_1,...,i_K} b(\Theta_{i_1,...,i_K}), \tag{1}$$

where $\Theta = \mathcal{C} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K$, and $b'(\cdot) = f(\cdot)$. The deviation between the MLE $\{\hat{\mathcal{C}}, \hat{\boldsymbol{M}}_k\}$ and the true parameters $\{\mathcal{C}, \boldsymbol{M}_k\}$ comes from two aspects: the label assignment and the estimation of the core tensor. We tease apart these two parts.

1. First, we suppose the membership $\{\boldsymbol{M}_k\}$ are given. We now assess the stochastic error due to the estimation of $\mathcal{C}$, conditional on $\{\boldsymbol{M}_k\}$. Noted that the objective function is a convex function, the MLE of $\mathcal{C}$ satisfies the first-order condition. Then, for each $(r_1, ..., r_K), r_k \in [R_k], k = 1, ..., K$ we have

$$\hat{c}_{r_1,...,r_K} = (b')^{-1} \left(\frac{1}{d_1 \cdots d_K p_{r_1}^{(1)} \cdots p_{r_K}^{(K)}} \left[\mathcal{Y} \times_1 \boldsymbol{M}_1^T \times_2 \cdots \times_K \boldsymbol{M}_K^T \right]_{r_1,...,r_K} \right), \tag{2}$$

where $p_{r_k}^{(k)} = \frac{1}{d_k}\sum_{i=1}^{d_k} \boldsymbol{I}\{M_{k,ir_k} = 1\}$ is the portion of the r_k-th cluster on the k-th mode.

Consider the function $F(\boldsymbol{M}_k) = \mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{C}}, \{\boldsymbol{M}_k\})$, where $\hat{\mathcal{C}} = [\![\hat{c}_{r_1,...,r_K}]\!]$ is the estimation (2). The function $F(\boldsymbol{M}_k)$ is of form

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

where $h(x) = x(b')^{-1}(x) - b((b')^{-1}(x))$. Define the function $G(\boldsymbol{M}_k)$ as following.

$$G(\boldsymbol{M}_k) = \sum_{r_1,...,r_K} \prod_k p_{r_k}^{(k)} h(\mu_{r_1,...,r_K}),$$

where

$$\mu_{r_1,...,r_K} = \mathbb{E}[b'(\hat{c}_{r_1,...,r_K})] = \frac{1}{\prod_k p_{r_k}^{(k)}}[b'(\mathcal{C}) \times_1 \boldsymbol{D}^{(1),T} \times_2 \cdots \times_K \boldsymbol{D}^{(K),T}]_{r_1,...,r_K}. \tag{3}$$

The function $G(\boldsymbol{M}_k)$ can be considered as the population version of the function $F(\boldsymbol{M}_k)$. Therefore, the deviation $F(\boldsymbol{M}_k) - G(\boldsymbol{M}_k)$ quantifies the stochastic error due to the estimation of $\mathcal{C}$. Further, we define the residual tensor for block means, $\mathcal{R}(\boldsymbol{M}_k) = [\![R_{r_1,...,r_K}]\!]$, where

$$R_{r_1,...,r_K} = b'(\hat{c}_{r_1,...,r_K}) - \mathbb{E}[b'(\hat{c}_{r_1,...,r_K})].$$

2. Next, we free $\{\boldsymbol{M}_k\}$ and quantify the total deviation. Considering the MLE $\{\hat{\boldsymbol{M}}_k\}$, we have

$$(\hat{\boldsymbol{M}}_1, ..., \hat{\boldsymbol{M}}_K) = \arg\max_{\{\boldsymbol{M}_k\}} F(\boldsymbol{M}_k).$$

The expectation with respect to $\mathcal{C}$ of the objective function at the true parameter is

$$G(\boldsymbol{M}_k) = \sum_{r_1,...,r_K} \prod_k p_{r_k}^{(k)} h(b'(c_{r_1,...,r_K})).$$

Correspondingly, the expected objective function at the MLE is

$$G(\hat{\boldsymbol{M}}_k) = \sum_{r_1,...,r_K} \prod_k \hat{p}_{r_k}^{(k)} h(\mu_{r_1,...,r_K}),$$

where $\mu_{r_1,...,r_K}$ is defined in (3) and $\hat{p}_{r_k}^{(k)}$ are obtained by $\hat{\boldsymbol{M}}_k$. We use $G(\hat{\boldsymbol{M}}_k) - G(\boldsymbol{M}_k)$ to measure the stochastic deviation caused by mismatch in label assignment.

The Lemma 1 indicates that, if there is non-negligible mismatch between $\boldsymbol{M}_k$ and $\hat{\boldsymbol{M}}_k$, the estimate $\hat{\boldsymbol{M}}_k$ can not be the global optimizer to the objective function (1).

Back to probability of misclassification rate. By Lemma 1, we have

$$\mathbb{P}(MCR(\hat{\boldsymbol{M}}_k, \boldsymbol{M}_k) \geq \epsilon) \leq \mathbb{P}\left(G(\hat{\boldsymbol{M}}_k) - G(\boldsymbol{M}_k) \leq -\frac{\epsilon}{4a_2}\tau^{K-1}\delta'\right). \tag{4}$$

Notice that total deviation between $\{\boldsymbol{M}_k\}$ and $\hat{\boldsymbol{M}}_k$ is decomposed in three parts.

$$
\begin{aligned}
F(\hat{\boldsymbol{M}}_k) - F(\boldsymbol{M}_k) &= F(\hat{\boldsymbol{M}}_k) - G(\hat{\boldsymbol{M}}_k) + G(\hat{\boldsymbol{M}}_k) - G(\boldsymbol{M}_k) + G(\boldsymbol{M}_k) - F(\boldsymbol{M}_k) \\
&\leq 2r - \frac{\epsilon}{4a_2}\tau^{K-1}\delta, \qquad (5)
\end{aligned}
$$

where the last inequality follows the triangle inequality, and $r = \sup_{\{\boldsymbol{M}_k\}} |F(\boldsymbol{M}_k) - G(\boldsymbol{M}_k)|$. Since $\{\hat{\boldsymbol{M}}_k\}$ is MLE, the left hand side of the inequality (5) is larger or equal than 0. Plugging the decomposition (5) in to the probability (4), we obtain that

$$
\begin{aligned}
\mathbb{P}(MCR(\hat{\boldsymbol{M}}_k, \boldsymbol{M}_k) \geq \epsilon) &\leq \mathbb{P}\left(F(\hat{\boldsymbol{M}}_k) - F(\boldsymbol{M}_k) \leq 2r - \frac{\epsilon}{4\alpha W}\tau^{K-1}\delta\right) \\
&\leq \mathbb{P}\left(r \geq \frac{\epsilon}{8\alpha}\tau^{K-1}\delta\right) \qquad (6)
\end{aligned}
$$

Now, the problem transfers to a find a probability of r. Consider the term r, we have

$$
\begin{aligned}
|F(\boldsymbol{M}_k) - G(\boldsymbol{M}_k)| &\leq \sum_{r_1,\dots,r_K} \prod_k p_{r_k}^{(k)} |h(b'(\hat{c}_{r_1,\dots,r_K})) - h(\mu_{r_1,..,r_K})| \qquad (7) \\
&\leq \|\mathcal{C}\|_{\max} \|R(\boldsymbol{M}_k)\|_{\max},
\end{aligned}
$$

where the last inequality follows by the Taylor Expansion

$$
|h(b'(\hat{c}_{r_1,\dots,r_K})) - h(\mu_{r_1,\dots,r_K})| \leq \sup_{x=b'(c_{r_1,\dots,r_k})} |h'(x)| \, \|\mathcal{R}(\boldsymbol{M}_k)\|_{\max}
$$

, and $\sup_{x=b'(c_{r_1,\dots,r_k})} |h'(x)| = \sup_{x=b'(c_{r_1,\dots,r_k})} |(b')^{-1}(x)| = \sup_{c_{r_1,\dots,r_K}} |c_{r_1,\dots,r_K}| \leq \|\mathcal{C}\|_{\max}$.
Combining the probability (6) with the upper bound (7), we obtain the accuracy of MCR

$$
\begin{aligned}
\mathbb{P}(MCR(\hat{\boldsymbol{M}}_k, \boldsymbol{M}_k) \geq \epsilon) &\leq \mathbb{P}\left(\sup_{\{\boldsymbol{M}_k\}} \|\mathcal{R}\|_{\max} \geq \frac{\epsilon}{8\alpha \|\mathcal{C}\|_{\max}}\tau^{K-1}\delta\right) \\
&\leq \mathbb{P}\left(\sup_{I_{r_1,\dots,r_K}} \frac{\sum_{(i_1,\dots,i_K)\in I_{r_1,\dots,r_K}} \mathcal{Y}_{i_1,\dots,i_K} - \mathbb{E}[\mathcal{Y}_{i_1,\dots,i_K}]}{|I_{r_1,\dots,r_K}|} \geq \frac{\epsilon}{8a_2 \|\mathcal{C}\|_{\max}}\tau^{K-1}\delta\right) \\
&\leq 2^{1+\sum d_k} \exp\left(-\frac{\epsilon^2\tau^{2K-2}\delta^2 L}{C\sigma^2\alpha^2 \|\mathcal{C}\|_{\max}^2}\right),
\end{aligned}
$$

where $I_{r_1,\dots,r_K} = \{(i_1,\dots,i_K) | \boldsymbol{M}_{k,i_k r_k} = 1, k \in [K]\}$ is the collection of the indices of the elements belong to the cluster $(r_1,\dots,r_K)$, the last inequality follows by the Hoeffding's inequality, and $L = \min |I_{r_1,\dots,r_K}| \geq \tau^K \prod_k d_k$. □

Lemma 1. *For an fixed $\epsilon > 0$, suppose $MCR(\hat{\boldsymbol{M}}_k, \boldsymbol{M}_k) \geq \epsilon$ for some $k \in [K]$. We have*

$$
G(\hat{\boldsymbol{M}}_k) - G(\boldsymbol{M}_k) \leq -\frac{\epsilon}{4a_2}\tau^{K-1}\delta.
$$

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

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

Recall the definition of $\mu_{r_1,..,r_K}$ in (3). Then, by Taylor Expansion of function $h(\cdot)$ at the point $\mu_{r_1,\ldots,r_K}$, we have

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

where $h''(x) = \frac{1}{b''(b'^{,-1}(x))}$, and $\inf_{x=b'(c_{r_1,\ldots,r_K})} h''(x) = \inf_{c_{r_1,\ldots,r_K}} \frac{1}{b''(c_{r_1,\ldots,r_K})} \geq \frac{1}{\mathrm{Var}(Y_{i_1,\ldots,i_K})} \geq \frac{1}{a_2}$. By the inequality $a^2 + b^2 \geq \frac{(a+b)^2}{2}$, we obtain that

$$
\begin{aligned}
&\frac{1}{W}[\mathcal{N} \times_1 \boldsymbol{D}^{(1),T} \times_2 \cdots \times_K \boldsymbol{D}^{(K),T}]_{r_1,\ldots,r_K} - h(\mu_{r_1,\ldots,r_K}) \\
&\quad \geq \frac{1}{a_2 4W} \min\{D^{(1)}_{a_1,r_1}, D^{(1)}_{a_1',r_1}\} D^{(2)}_{a_2,r_2} \cdots D^{(K)}_{a_K,r_K} (b'(c_{a_1,\ldots,a_K}) - b'(c_{a_1',\ldots,a_K}))^2. \qquad (8)
\end{aligned}
$$

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

$$
\frac{1}{W}[\mathcal{N} \times_1 \boldsymbol{D}^{(1),T} \times_2 \cdots \times_K \boldsymbol{D}^{(K),T}]_{r_1',\ldots,r_K} - h(\mu_{r_1',\ldots,r_K}) \geq 0. \qquad (9)
$$

Combing the inequality (8) and (9), we obtain that

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

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

$\square$

2 Discussion

Following is the discussion about the definition of $G(\boldsymbol{M}_k) = \mathbb{E}[F(\boldsymbol{M}_k)]$, which is the expectation of $F(\boldsymbol{M}_k)$ with respect to $\hat{\mathcal{C}} = [\![\hat{c}_{r_1,\ldots,r_K}]\!]$.

2.1 Least Squared model

In the least squared model, with given membership $\{\boldsymbol{M}_k\}$, the estimation of the core tensor is

$$
\hat{c}_{r_1,\ldots,r_K} = \frac{1}{d_1 \ldots d_K p^{(1)}_{r_1} \cdots p^{(K)}_{r_K}} [\mathcal{Y} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K]_{r_1,\ldots,r_K}.
$$

We define the function $F(\boldsymbol{M}_k) = \mathcal{L}_{\mathcal{Y}}(\mathcal{C}, \{\boldsymbol{M}_k\})$. A straightforward calculation shows that

$$F(\boldsymbol{M}_k) = \sum_{r_1,\dots,r_K} \prod_k p_{r_k}^{(k)} \hat{c}_{r_1,\dots,r_K}^2.$$

Let $G(\boldsymbol{M}_k) = \mathbb{E}[F(\boldsymbol{M}_k)]$ denote the expectation of $F(\boldsymbol{M}_k)$ with respect to $\hat{\mathcal{C}}$. We have

$$G(\boldsymbol{M}_k) = \mathbb{E}[F(\boldsymbol{M}_k)] = \sum_{r_1,\dots,r_K} \prod_k p_{r_k}^{(k)} \mathbb{E}(\hat{c}_{r_1,\dots,r_K}^2). \tag{10}$$

Notice that $\mathbb{E}[\hat{c}_{r_1,\dots,r_K}^2] = \mathrm{Var}(\hat{c}_{r_1,\dots,r_K}) + (\mathbb{E}[\hat{c}_{r_1,\dots,r_K}])^2$, and $(\mathbb{E}[\hat{c}_{r_1,\dots,r_K}])^2 = \mu_{r_1,\dots,r_K}^2$ where $\mu_{r_1,\dots,r_K}$ is defined in (3). Since for each entry $\mathrm{Var}(\mathcal{Y}_{i_1,\dots,i_K}) = \mathrm{Var}(\epsilon_{i_1,\dots,i_K}) = \sigma_0^2$, and $\epsilon_{i_1,\dots,i_K}$ are i.i.d., the variance is equal to

$$\mathrm{Var}(\hat{c}_{r_1,\dots,r_K}) = \frac{1}{\prod_k d_k^2 \prod_k [p_{r_k}^{(k)}]^2} \prod_k d_k \prod_k p_{r_k}^{(k)} \sigma_0^2 = \frac{1}{\prod_k d_k \prod_k p_{r_k}^{(k)}} \sigma_0^2. \tag{11}$$

Plugging the variance (11) into the definition (10), we have

$$G(\boldsymbol{M}_k) = \sum_{r_1,\dots,r_K} \frac{1}{\prod_k d_k} \sigma_0^2 + \sum_{r_1,\dots,r_K} \prod_k p_{r_k}^{(k)} \mu_{r_1,\dots,r_K}^2.$$

Since the first term is independent with the membership $\{\boldsymbol{M}_k\}$, we can ignore the first term, and the conclusion won't change.

2.2 Exponential Family model

In the exponential family model, with given membership $\{\boldsymbol{M}_k\}$, the estimation of the core tensor is

$$\hat{c}_{r_1,\dots,r_K} = (b')^{-1} \frac{1}{d_1 \dots d_K p_{r_1}^{(1)} \cdots p_{r_K}^{(K)}} [\mathcal{Y} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K]_{r_1,\dots,r_K}.$$

Then, the corresponding function $F(\boldsymbol{M}_k)$ is of form

$$F(\boldsymbol{M}_k) = \sum_{r_1,\dots,r_K} \prod_k p_{r_k}^{(k)} h(b'(\hat{c}_{r_1,\dots,r_K})),$$

where $h(x) = x(b')^{-1}(x) - b((b')^{-1}(x))$. The expectation $G(\boldsymbol{M}_k)$ is

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

Consider the Taylor Expansion of $h(\cdot)$ at the point $\mathbb{E}[b'(\hat{c}_{r_1,\dots,r_K})]$. For simplicity, let $\hat{c}$ denote $\hat{c}_{r_1,\dots,r_K}$. We have

$$h(b'(\hat{c})) = h(\mathbb{E}[b'(\hat{c})]) + h'(\mathbb{E}[b'(\hat{c})])(b'(\hat{c}) - \mathbb{E}[b'(\hat{c})]) + \frac{h''(\alpha b'(\hat{c}) + (1-\alpha)\mathbb{E}(b'(\hat{c})))}{2}(b'(\hat{c}) - \mathbb{E}[b'(\hat{c})])^2,$$

for some $\alpha \in [0,1]$. Since the expectation of the first term $\mathbb{E}\left[h'(\mathbb{E}[b'(\hat{c})])(b'(\hat{c}) - \mathbb{E}[b'(\hat{c})])\right] = 0$, we only need to prove that $\mathbb{E}\left[\frac{h''(\alpha b'(\hat{c}) + (1-\alpha)\mathbb{E}(b'(\hat{c})))}{2}(b'(\hat{c}) - \mathbb{E}[b'(\hat{c})])^2\right]$ is not related to $\{\boldsymbol{M}_k\}$.

Below is just my thoughts.

Note that $h''(x) = \frac{1}{b''(b'^{,-1}(x))}$, and $\alpha b'(\hat{c}) + (1-\alpha)\mathbb{E}(b'(\hat{c}))$ is a linear combination of all entries of $b'(\mathcal{C})$ and $\mathcal{Y} = b'(\mathcal{C}) \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K + \mathcal{E}$, where $\mathcal{E} = [\![\epsilon_{i_1,\ldots,i_K}]\!]$ is a sub-gaussian mean-zero noise tensor. Recall the assumption that $0 < a_1 < \text{Var}(\mathcal{Y}_{i_1,\ldots,i_K} | c_{r_1,\ldots,r_K}) = b''(c_{r_1,\ldots,r_K}) < a_2 < \infty$. We have

$$\inf_{r_1,\ldots,r_K} \frac{1}{\text{Var}(\mathcal{Y}_{i_1,\ldots,i_K} | c_{r_1,\ldots,r_K} + \epsilon_{i_1,\ldots,i_K})} \leq h''(\alpha b'(\hat{c}) + (1-\alpha)\mathbb{E}(b'(\hat{c}))) \leq \sup_{r_1,\ldots,r_K} \frac{1}{\text{Var}(\mathcal{Y}_{i_1,\ldots,i_K} | c_{r_1,\ldots,r_K})},$$

which implies that

$$\frac{1}{4a_2} \leq h''(\alpha b'(\hat{c}) + (1-\alpha)\mathbb{E}(b'(\hat{c}))) \leq \frac{1}{a_1}.$$

2.3 Weakest Condition under which the clustering accuracy holds?

The following are three necessary conditions:

1. Irreducibility. This implies the minimal gap between blocks $\delta = \min_k \delta^{(k)} > 0$, where

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

   Otherwise, two blocks with the same mean are not able to be distinguished.

2. Convex loss function. The loss function should be convex, otherwise the estimation of $\mathcal{C}$ would not be unique with given membership $\{\boldsymbol{M}_k\}$, and thus the stochastic error for estimation will be hard to measure.

3. Bounded variance. This implies there exist two positive constants a_1, a_2 such that $0 < a_1 < \text{Var}(\mathcal{Y}_{i_1,\ldots,i_K} | \Theta_{i_1,\ldots,i_K}) < a_2 < \infty$.

The bounded variance condition may not be sufficient. Because in $G(\boldsymbol{M}_k)$, the rest parts other than $\sum_{r_1,\ldots,r_K} \prod_k p_{r_k}^{(k)} h(\mu_{r_1,\ldots,r_K})$ are also related to the membership $\{\boldsymbol{M}_k\}$.