# Graphic Lasso: Self-Consistency

Jiaxin Hu

February 15, 2021

## 1 Noiseless case

Consider the noiseless case

$$\mathcal{Y} = f(\Theta),$$

where $\Theta = \mathcal{C} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K$, and $f(\cdot)$ is an entry-wise link function. Suppose we have the following optimization problem.

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

**Lemma 1** (Noiseless estimation)**.** *Let $\{\mathcal{C}, \boldsymbol{M}_k\}$ denote the true parameters and $\{\hat{\mathcal{C}}, \hat{\boldsymbol{M}}_k\}$ are the estimation which maximizes the loss function. Suppose $g(\cdot)$ is a convex function with bounded second derivative $\sup_x g''(x) \leq a$, and $\max_{r_1,\ldots,r_K} |(g')^{-1}(f(c_{r_1,\ldots,r_K}))| \leq C$, where $C$ is a positive constant depends on $\mathcal{C}$. Assume the minimal gap between blocks is strictly larger than 0, i.e., $\delta > 0$. Then, for any $\epsilon > 0$, we have*

$$\mathbb{P}(MCR(\hat{\boldsymbol{M}}_k, \boldsymbol{M}_k) \geq \epsilon) = 0.$$

*Proof.* We prove the accuracy in following steps.

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

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

   Note that the estimation $\hat{\mathcal{C}}$ depends on $\hat{\boldsymbol{M}}_k$. Therefore, we denote the estimation as $\hat{\mathcal{C}}(\hat{\boldsymbol{M}}_k) = [\![\hat{c}_{r_1,\ldots,r_K}(\hat{\boldsymbol{M}}_k)]\!]$.

2. We define some useful functions. First, we define

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

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

   Note that $\hat{\mathcal{C}}(\hat{\boldsymbol{M}}_k)$ does not include the randomness. Thus, we have $g'(\hat{c}_{r_1,\ldots,r_K}(\hat{\boldsymbol{M}}_k)) = \mathbb{E}\left[g'(\hat{c}_{r_1,\ldots,r_K}(\hat{\boldsymbol{M}}_k))\right]$, and

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

which implies that there does not exist the estimation error.

**Note that for true membership, we have**

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

where $\hat{\mathcal{C}}(\boldsymbol{M}_k) = (g')^{-1}(f(\mathcal{C}))$ **is not equal to the true core tensor** $\mathcal{C}$.

3. We only need to consider the classification error. Under the assumptions of the positive minimal gap and the boundedness of the second derivative of $g$, when $MCR(\hat{\boldsymbol{M}}_k, \boldsymbol{M}_k) \geq \epsilon$ for any $\epsilon > 0$, we have
$$G(\hat{\boldsymbol{M}}_k) - G(\boldsymbol{M}_k) \leq -\frac{\epsilon}{4a}\tau^{K-1}\delta.$$

4. Since $\{\hat{\mathcal{C}}\hat{\boldsymbol{M}}_k, \hat{\boldsymbol{M}}_k\}$ is the maximizer of the loss function, we have
$$0 \leq F(\hat{\boldsymbol{M}}_k) - F(\boldsymbol{M}_k) = G(\hat{\boldsymbol{M}}_k) - G(\boldsymbol{M}_k).$$
Therefore, we obtain that
$$\mathbb{P}(MCR(\hat{\boldsymbol{M}}_k, \boldsymbol{M}_k) \geq \epsilon) = \mathbb{P}(G(\hat{\boldsymbol{M}}_k) - G(\boldsymbol{M}_k) \leq -\frac{\epsilon}{4a}\tau^{K-1}\delta) = 0.$$

□

**Remark 1.** The lemma 1 implies that the true membership $\boldsymbol{M}_k$ is the maximizer of the function $G(\boldsymbol{M}'_k)$. Due to the noiselessness, $G(\boldsymbol{M}'_k) = \mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{C}}(\boldsymbol{M}'_k), \boldsymbol{M}'_k)$, and $\{\hat{\mathcal{C}}(\boldsymbol{M}_k), \boldsymbol{M}_k\}$ is the maximizer of the noiseless loss function. However, the true parameter $\{\mathcal{C}, \boldsymbol{M}_k\}$ is not the maximizer of the noiseless loss function, since $\hat{\mathcal{C}}(\boldsymbol{M}_k) \neq \mathcal{C}$. Therefore, we conclude that the loss function (1) is **self-consistent to** $\{\hat{\mathcal{C}}(\boldsymbol{M}_k), \boldsymbol{M}_k\}$ but not **self-consistent to** $\Theta$.

**Remark 2.** Define
$$\hat{\Theta} = \hat{\mathcal{C}}(\boldsymbol{M}_k) \times_1 \boldsymbol{M}_1 \times_1 \cdots \times_K \boldsymbol{M}_K.$$
Then, $\hat{\Theta}$ is an unbiased estimate of $\Theta$ if and only if $g' = f$.

**Remark 3. Which assumption in the noisy case corresponds to the self-consistency of $\boldsymbol{M}_k$?**

Note that in the noisy case, we have

$$\begin{aligned}
G_{noise}(\hat{\boldsymbol{M}}_k) &= \sum_{r_1,\ldots,r_K} \prod_k d_k \prod_k \hat{p}^{(k)}_{r_k} h(\mathbb{E}\left[g'(\hat{c}_{r_1,\ldots,r_K}(\hat{\boldsymbol{M}}_k))\right]) \\
&= \langle f(\mathcal{C}) \times_1 \boldsymbol{M}_1\hat{\boldsymbol{M}}_1^T \times_2 \cdots \times_K \boldsymbol{M}_K\boldsymbol{M}_K^T, (g')^{-1}\left[f(\mathcal{C}) \times_1 \boldsymbol{M}_1\hat{\boldsymbol{M}}_1^T \times_2 \cdots \times_K \boldsymbol{M}_K\boldsymbol{M}_K^T\right]\rangle \\
&\qquad - \sum_{i_1,\ldots,i_K} g\left((g')^{-1}\left[f(\mathcal{C}) \times_1 \boldsymbol{M}_1\hat{\boldsymbol{M}}_1^T \times_2 \cdots \times_K \boldsymbol{M}_K\boldsymbol{M}_K^T\right] \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K\right)_{i_1,\ldots,i_K} \\
&= F_{noiseless}(\hat{\boldsymbol{M}}_k).
\end{aligned}$$

Therefore, we use the self-consistency when we derive the misclassification error. Note that the result that when $MCR(\hat{\boldsymbol{M}}_k, \boldsymbol{M}_k) \geq \epsilon$,

$$G_{noise}(\hat{\boldsymbol{M}}_k) - G_{noise}(\boldsymbol{M}_k) \leq -\frac{\epsilon}{4a}\tau^{K-1}\delta \tag{2}$$

implies the self-consistency of $\boldsymbol{M}_k$. To obtain the result (2), we require

1. the convexity of $g$ and $\sup_x g''(x) \geq a$;
2. minimal gap strictly larger than 0, i.e., $\delta > 0$.

## 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. 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*

$$G(\boldsymbol{M}_k') - G(\boldsymbol{M}_k) \leq -C(\epsilon), \tag{3}$$

   *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.,*

$$p(t) = \mathbb{P}(|F(\boldsymbol{M}_k') - G(\boldsymbol{M}_k')| \geq t) \to 0, \quad \textit{as} \quad t \to \infty. \tag{4}$$

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

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

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

$$\begin{aligned}0 &\leq F(\hat{\boldsymbol{M}}_k) - F(\boldsymbol{M}_k)\\ &= F(\hat{\boldsymbol{M}}_k) - G(\hat{\boldsymbol{M}}_k) + G(\hat{\boldsymbol{M}}_k) - G(\boldsymbol{M}_k) + G(\boldsymbol{M}_k) - F(\boldsymbol{M}_k).\end{aligned}$$

Suppose $MCR(\hat{\boldsymbol{M}}_k, \boldsymbol{M}_k) \geq \epsilon$. By the property (3), we have

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

where $r = \sup_{\boldsymbol{M}'_k} |F(\boldsymbol{M}'_k) - G(\boldsymbol{M}'_k)|$. Therefore, we have

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

where the last equation follows the second property (4). □

**Remark 4.** For the model in Tensor Block model, we have

$$C(\epsilon) = \frac{\epsilon}{4a}\tau^{K-1}\delta,$$

where $a$ is the upper bound of $g''(x)$, $\tau$ is minimal proportion of the cluster, and $\delta$ is the minimal gap between blocks. By the sub-Gaussianity of $\mathcal{Y}$ and Hoeffding's inequality, we have

$$\begin{aligned}
p(t) &\leq \mathbb{P}(C_1 \left\| g'(\hat{c}_{r_1,...,r_K}) - \mathbb{E}[g'(\hat{c}_{r_1,...,r_K})] \right\|_{\max} \geq t) \\
&\leq \mathbb{P}\left(\sup_{I_{r_1,...,r_K}} \frac{|\sum_{(i_1,...,i_K)\in I_{r_1,...,r_K}} \mathcal{Y}_{i_1,...,i_K} - \mathbb{E}[\mathcal{Y}_{i_1,...,i_K}]|}{|I_{r_1,...,r_K}|} \geq \frac{t}{C_1}\right) \\
&\leq 2^{1+\sum_k d_k} \exp\left(-\frac{t^2 L}{C_1^2}\right),
\end{aligned}$$

where $C_1$ is a positive constant related to the true core tensor $\mathcal{C}$, $I_{r_1,...,r_K}$ is the index set of the block $(r_1, ..., r_K)$ based on the estimate membership $\hat{\boldsymbol{M}}_k$, and $L = \inf |I_{r_1,...,r_K}| \geq \tau^K \prod_k d_k$.

**Remark 5.** When $\tilde{\mathcal{C}}(\boldsymbol{M}_k) = \mathcal{C}$, i.e., $g' = f$ in the tensor block model, the self-consistency to $\boldsymbol{M}_k$ implies the self-consistency to $\{\mathcal{C}, \boldsymbol{M}_k\}$ or $\Theta = \mathcal{C} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K$.