

# Graphic Lasso: Miscellaneous

Jiaxin Hu

February 6, 2021

## 1 Weakest assumption for the TBM clustering accuracy

Consider the model

$$\mathbb{E}[\mathcal{Y}] = f(\Theta),$$

where  $\Theta = \mathcal{C} \times \mathbf{M}_1 \times_2 \cdots \times_K \mathbf{M}_K$ . Define the misclassification rate on the  $k$ -th mode as

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

where  $D^{(k)} \in \mathbb{R}^{R_k \times R_k}$  is the confusion matrix on the  $k$ -th, and  $D_{rr'}^{(k)} = \frac{1}{d_k} \sum_{i=1}^{d_k} \mathbf{I}\{\mathbf{M}_{k,ir_k} = \hat{\mathbf{M}}_{k,ir_k} = 1\}$ .

**Theorem 1.1.** *Consider the optimization problem*

$$\max_{\Theta} \mathcal{L}_{\mathcal{Y}}(\Theta) = \langle \mathcal{Y}, \Theta \rangle - \sum_{(i_1, \dots, i_K)} g(\Theta_{i_1, \dots, i_K}). \quad (1)$$

*The weakest sufficient conditions for maximizer to (1) satisfies the following upper bound with high probability*

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

are

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

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

3. The function  $h(x) = xf^{-1}(x) - g(f^{-1}(x))$  is convex,  $\sup_{x \in \mathcal{S}} |h'(x)| \leq p(\mathcal{C})$ , where  $p(\mathcal{C})$  is a term related to  $\mathcal{C}$ , and  $\sup_{x \in \mathcal{S}} h''(x)$  is lower bounded by a positive constant  $a$ , where  $\mathcal{S}$  is the convex hull of the entries of  $f(\mathcal{C})$ .
4. 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  $\sigma$ .

*Proof.* **With condition 1**, we are able to find the unique maximizer of  $\mathcal{C} = \llbracket c_{r_1, \dots, r_K} \rrbracket$  with given membership  $\{\mathbf{M}_k\}$ , which is

$$\hat{\mathcal{C}} = (g')^{-1}(\mathcal{Y} \times_1 \mathbf{D}_1 \times_2 \cdots \times_K \mathbf{D}_K).$$

Then, we construct the unique functions

$$F(\mathbf{M}_k) = \mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{C}}, \mathbf{M}_k) \sum_{r_1, \dots, r_K} \prod_k p_{r_k}^{(k)} h(f(\hat{c}_{r_1, \dots, r_K})),$$

and the population version of  $F(\mathbf{M}_k)$

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

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

**With condition 3**, the gap between sample- and population-version of the objective function are upper bounded by a term related to the residual tensor  $\mathcal{Y} - \mathbb{E}[\mathcal{Y}]$ . That is

$$\begin{aligned} |F(\mathbf{M}_k) - G(\mathbf{M}_k)| &\leq \sum_{r_1, \dots, r_K} \prod_k p_{r_k}^{(k)} |h(f(\hat{c}_{r_1, \dots, r_K})) - h(\mathbb{E}[f(\hat{c}_{r_1, \dots, r_K})])| \\ &\leq \sup_{x \in \mathcal{S}} |h'(x)| \|f(\hat{c}_{r_1, \dots, r_K}) - \mathbb{E}[f(\hat{c}_{r_1, \dots, r_K})]\|_{\max}, \\ &\leq p(\mathcal{C}) \|f(\hat{c}_{r_1, \dots, r_K}) - \mathbb{E}[f(\hat{c}_{r_1, \dots, r_K})]\|_{\max}, \end{aligned}$$

where the second inequality follows by the fact that  $h$  is convex and thus  $h$  is local Lipschitz with  $L = \sup_{x \in \mathcal{S}} |h'(x)|$ , and the third inequality follows the condition 3.

**With condition 2,3**, we satisfy the assumptions of Lemma 1, then for any  $\epsilon > 0$ , the misclassification  $MCR(\hat{\mathbf{M}}_k, \mathbf{M}_k) \geq \epsilon$  for some  $k \in [K]$  implies

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

With the optimality of  $\hat{\mathbf{M}}_k$ , we have  $F(\hat{\mathbf{M}}_k) \geq F(\mathbf{M}_k)$ . Then, the probability for the misclassification rate changes to the probability for the residual. That is

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

where  $I_{r_1, \dots, r_K} = \{(i_1, \dots, i_K) | \mathbf{M}_{k, i_k r_k} = 1, k \in [K]\}$  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 **with condition 4**, and  $L = \min |I_{r_1, \dots, r_K}| \geq \tau^K \prod_k d_k$ .  $\square$

## 2 Mixed membership clustering

### 2.1 Vector version

Table 1 summaries the mixed membership models in Ji Zhu's paper of matrix and vector versions.

	Matrix version	Vector version
Observation	The (symmetric) adjacency matrix $A = \llbracket A_{ij} \rrbracket \in \{0, 1\}^{n \times n}$ . The entry $A_{ij} = 1$ implies there exists a correlation between node $i$ and $j$ , otherwise $A_{ij} = 0$ .	The mean vector $A = \llbracket A_i \rrbracket \in \{0, 1\}^n$ .
Distribution	Assume $A_{ij} \sim \text{Ber}(p_{ij}),$ independently.	Assume $A_i \sim \text{Ber}(p_i),$ independently.
Model	Let $W = \mathbb{E}[A] \in \mathbb{R}^{n \times n}$ . Consider the model $W = \alpha_n \Theta Z B Z^T \Theta,$ where $\alpha_n \rightarrow 0$ , $\Theta = \text{diag}(\theta_1, \dots, \theta_n)$ , $Z \in \mathbb{R}^{n \times K}$ is the mixed membership matrix, and $B \in \mathbb{R}^{K \times K}$ represents the probabilities between pure nodes.	Let $W = \mathbb{E}[A] \in \mathbb{R}^n$ . Consider the model $W = \alpha_n \Theta Z B,$ where $\alpha_n \rightarrow 0$ , $\Theta = \text{diag}(\theta_1, \dots, \theta_n)$ , $Z \in \mathbb{R}^{n \times K}$ is the mixed membership matrix, and $B \in \mathbb{R}^K$ represents the probabilities of pure nodes.
Identifiability	Under following conditions, the parameters $(\alpha_n, \Theta, Z, B)$ are identifiable. <ol style="list-style-type: none"> <li>1. <math>B</math> full rank and strictly positive definite, with <math>B_{kk} = 1, k \in [K]</math>.</li> <li>2. All <math>Z_{ik} \geq 0</math>, <math>\ Z_{i.}\  = 1, i \in [n]</math>, and for each <math>k \in [K]</math> there exists an <math>i</math> such that <math>Z_{ik} = 1</math>.</li> <li>3. The degree parameters <math>\theta_i \geq 0</math> and <math>\frac{1}{n} \sum_{i=1}^n \theta_i = 1</math>.</li> </ol>	<b>(Conjecture)</b> Under following conditions, the parameters $(\alpha_n, \Theta, Z, B)$ are identifiable. <ol style="list-style-type: none"> <li>1. <math>\min_{i \neq j}  B_i - B_j  &gt; 0</math>, with <math>0 &lt; B_k \leq 1, k \in [K]</math>.</li> <li>2. All <math>Z_{ik} \geq 0</math>, <math>\sum_{k=1}^K Z_{ik} = 1</math>, and for each <math>k \in [K]</math> there exists an <math>i</math> such that <math>Z_{ik} = 1</math>.</li> <li>3. The degree parameters <math>\theta_i \geq 0</math> and <math>\frac{1}{n} \sum_{i=1}^n \theta_i = 1</math>.</li> </ol>

Table 1: Matrix and vector version of mixed membership model.

## 2.2 Connection to precision matrix model

Table 2 indicates a possible model for the precision matrix clustering based on the vector version of Zhu's model.

	Precision matrix
Observation	<p>Vectorized sample covariance matrix</p> $A = \begin{bmatrix} A_i \\ \vdots \\ A_n \end{bmatrix} = \begin{bmatrix} \text{vec}(S_1) \\ \vdots \\ \text{vec}(S_n) \end{bmatrix} \in \mathbb{R}^{n \times p^2},$ <p>where <math>S_i</math> is the sample covariance matrix for <math>i</math>-th category.</p>
Distribution	<p>Assume <math>X_{ij} \sim \mathcal{N}_p(0, \Sigma_i), i \in [n], j \in [m]</math>, independently. We have</p> $\mathbb{E}[A] = W = \begin{bmatrix} \text{vec}(\Sigma_1) \\ \vdots \\ \text{vec}(\Sigma_n) \end{bmatrix}.$
Model	<p>Consider the model</p> $W = \alpha_n \Theta Z B,$ <p>where <math>\alpha_n \rightarrow 0</math>, <math>\Theta = \text{diag}(\theta_1, \dots, \theta_n)</math>, <math>Z \in \mathbb{R}^{n \times K}</math> is the mixed membership matrix, and <math>B \in \mathbb{R}^{K \times p^2}</math> vectorized parameter matrix for pure categories, i.e.,</p> $B = \begin{bmatrix} \text{vec}(\Omega_1^{-1}) \\ \vdots \\ \text{vec}(\Omega_K^{-1}) \end{bmatrix}.$
Identifiability	<p><b>(Conjecture)</b> Under the following conditions, the parameter set <math>(\alpha_n, \Theta, Z, \{\Omega_k\})</math> are identifiable.</p> <ol style="list-style-type: none"> <li>1. <math>\text{rank}(B) = K</math>.</li> <li>2. All <math>Z_{ik} \geq 0</math>, <math>\sum_{k=1}^K Z_{ik} = 1</math>, and for each <math>k \in [K]</math> there exists an <math>i</math> such that <math>Z_{ik} = 1</math>.</li> <li>3. The degree parameters <math>\theta_i \geq 0</math> and <math>\frac{1}{n} \sum_{i=1}^n \theta_i = 1</math>.</li> </ol>

Table 2: Possible precision matrix model with mixed membership.