Graphic Lasso: Micellaneous

Jiaxin Hu

February 6, 2021

1 Weakest assumption for the TBM clustering accuracy

Consider the model

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

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

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

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

Theorem 1.1. *Consider the optimization problem*

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

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

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

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*

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

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 σ.*

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

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

Then, we construct the unique functions

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

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

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

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

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

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 $MCR(\hat{\boldsymbol{M}}_k, \boldsymbol{M}_k) \geq \epsilon$ for some $k \in [K]$ implies

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

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

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

where $I_{r_1,...,r_K} = \{(i_1, ..., i_K) | \boldsymbol{M}_{k,i_k r_k} = 1, k \in [K]\}$ is the collection of the indices of the elements belong to the cluster $(r_1, ..., r_K)$, the last inequality follows by the Hoeffding's inequality **with condition 4**, and $L = \min |I_{r_1,...,r_K}| \geq \tau^K \prod_k d_k$. □

2 Mixed membership clustering

2.1 Vector verison

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 $A = [\![A_{ij}]\!] \in \{0,1\}^{n\times n}$. The entry $A_{ij} = 1$ implies there exists a correlation between node i and j, otherwise $A_{ij} = 0$.	The mean vector $A = [\![A_i]\!] \in \{0,1\}^n$.
Distribution	Assume $$A_{ij} \sim Ber(p_{ij}),$$ independently.	Assume $$A_i \sim Ber(p_i),$$ 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 $\alpha_n \to 0$, $\Theta = diag(\theta_1, ..., \theta_n)$, $Z \in \mathbb{R}^{n\times K}$ 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 $\alpha_n \to 0$, $\Theta = diag(\theta_1, ..., \theta_n)$, $Z \in \mathbb{R}^{n\times K}$ is the mixed membership matrix, and $B \in \mathbb{R}^K$ represents the probabilities of pure nodes.
Identifiability	Under following conditions, the parameters (α_n, Θ, Z, B) are identifiable. 1. B full rank and strictly positive definite, with $B_{kk} = 1, k \in [K]$. 2. All $Z_{ik} \geq 0$, $\Vert Z_{i\cdot}\Vert = 1, i \in [n]$, and for each $k \in [k]$ there exists an i such that $Z_{ik} = 1$. 3. The degree parameters $\theta_i \geq 0$ and $\frac{1}{n}\sum_{i=1}^n \theta_i = 1$.	(Conjecture) Under following conditions, the parameters (α_n, Θ, Z, B) are identifiable. 1. $\min_{i\neq j} \lvert B_i - B_j\rvert > 0$, with $0 < B_k \leq 1, k \in [K]$. 2. All $Z_{ik} \geq 0$, $\sum_{k=1}^K Z_{ik} = 1$, and for each $k \in [K]$ there exists an i such that $Z_{ik} = 1$. 3. The degree parameters $\theta_i \geq 0$ and $\frac{1}{n}\sum_{i=1}^n \theta_i = 1$.

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	Vectorized sample covariance matrix $$A = \begin{bmatrix} A_i \\ \vdots \\ A_n \end{bmatrix} = \begin{bmatrix} vec(S_1) \\ \vdots \\ vec(S_n) \end{bmatrix} \in \mathbb{R}^{n \times p^2},$$ where S_i is the sample covariance matrix for i-th category.
Distribution	Assume $X_{ij} \sim \mathcal{N}_p(0, \Sigma_i), i \in [n], j \in [m]$, independently. We have $$\mathbb{E}[A] = W = \begin{bmatrix} vec(\Sigma_1) \\ \vdots \\ vec(\Sigma_n) \end{bmatrix}.$$
Model	Consider the model $$W = \alpha_n \Theta Z B,$$ where $\alpha_n \to 0$, $\Theta = diag(\theta_1, ..., \theta_n)$, $Z \in \mathbb{R}^{n \times K}$ is the mixed membership matrix, and $B \in \mathbb{R}^{K \times p^2}$ vectorized parameter matrix for pure categories, i.e., $$B = \begin{bmatrix} vec(\Omega_1^{-1}) \\ \vdots \\ vec(\Omega_K^{-1}) \end{bmatrix}.$$
Identifiability	(Conjecture) Under the following conditions, the parameter set $(\alpha_n, \Theta, Z, \{\Omega_k\})$ are identifiable. 1. $rank(B) = K$. 2. All $Z_{ik} \geq 0$, $\sum_{k=1}^{K} Z_{ik} = 1$, and for each $k \in [K]$ there exists an i such that $Z_{ik} = 1$. 3. The degree parameters $\theta_i \geq 0$ and $\frac{1}{n} \sum_{i=1}^{n} \theta_i = 1$.

Table 2: Possible precision matrix model with mixed membership.