

# Tensor Block Model- Paper Review

Jiaxin Hu

04/20/2020

First edition: 04/07/2020; Second edition: 04/15/2020

## SUMMARY

Wang's and Lei's paper both aim to identify the block structure from a noisy tensor. Tensor block model (TBM) proposed by Wang can handle clustering without order limitation while tensor stochastic model (TSBM) proposed by Lei concentrates on the clustering on 2D network data across multi-layers. TBM can be considered as a high-order generalization of TBSM.

## 1 COMPARISON OF TWO MODELS

### 1.1 Model

#### 1.1.1 TBM

Let  $\mathcal{Y} = \llbracket y_{i_1, \dots, i_K} \rrbracket \in \mathbb{R}^{d_1 \times \dots \times d_K}$  be an order- $K$  observation tensor. Assume there are  $R_k$  clusters on the  $k$ th mode of tensor, where  $k = 1, \dots, K$ . We propose the following tensor block model,

$$\mathcal{Y} = \mathcal{C} \times_1 M_1 \times_2 \dots \times_K M_K + \mathcal{E},$$

where  $\mathcal{C} = \llbracket c_{r_1, \dots, r_K} \rrbracket \in \mathbb{R}^{R_1 \times \dots \times R_K}$  is the core tensor,  $M_k = \llbracket m_{i,j} \rrbracket \in \{0, 1\}^{d_k \times R_k}$  is the membership matrix on mode  $k$  and  $\mathcal{E} = \llbracket \epsilon_{i_1, \dots, i_K} \rrbracket \in \mathbb{R}^{d_1 \times \dots \times d_K}$  is the noise tensor where  $\epsilon_{i_1, \dots, i_K} \sim^{i.i.d} sG(\sigma^2)$ . Given the observation  $\mathcal{Y}$  and number of clusters  $R_1, \dots, R_K$ , the least-square estimator is:

$$(\hat{\mathcal{C}}, \hat{M}_1, \dots, \hat{M}_K) = \arg \min_{\mathcal{C}, M_1, \dots, M_K} \|\mathcal{Y} - \mathcal{C} \times_1 M_1 \times_2 \dots \times_K M_K\|^2,$$

where we define  $\Theta = \mathcal{C} \times_1 M_1 \times_2 \dots \times_K M_K$  and its estimate  $\hat{\Theta} = \hat{\mathcal{C}} \times_1 \hat{M}_1 \dots \times_K \hat{M}_K$ .

#### 1.1.2 TSBM

Let  $\mathcal{Y} = \llbracket y_{i_1, i_2, i_3} \rrbracket \in \{0, 1\}^{m \times n \times n}$  be a symmetric observation tensor where  $y_{ijk} = y_{ikj}$ ,  $j \neq k$ . Assume there are  $R$  clusters on the second and third mode. Let  $\mathbf{g} = \llbracket g_i \rrbracket \in \{1, \dots, R\}^n$  be the marginal partition vector on the second and third mode, where  $g_i \in [R]$  refers to the cluster for the  $i$ -th column. We propose the following tensor stochastic block model,

$$\mathbb{P}(\mathcal{Y} = 1) = \mathcal{P} = \mathcal{B} \times_2 G \times_3 G,$$

where  $\mathcal{B} = \llbracket b_{i_1, r_2, r_3} \rrbracket \in \mathbb{R}^{m \times R \times R}$  is the core tensor and  $G \in \{0, 1\}^{n \times R}$  is the membership matrix corresponds to partition  $\mathbf{g}$ . In  $G$ ,  $G_{ij} = 1$  if  $j = g_i$ , otherwise  $G_{ij} = 0$ . Noticed that  $\mathcal{Y}_{ijk} \sim^{i.i.d} \text{Ber}(\mathcal{P}_{ijk})$ ,  $\mathcal{Y}_{ijk} - \mathcal{P}_{ijk} \sim sG(1/4)$ , TSBM can be written as TBM:

$$\mathcal{Y} = \mathcal{B} \times_2 G \times_3 G + \mathcal{E}$$

where  $\mathcal{E} \in \mathbb{R}^{m \times n \times n}$  is a sub-Gaussian noise tensor. Given the observation  $\mathcal{Y}$  and the number of clusters  $R$ , the least square estimate is:

$$\begin{aligned} (\hat{g}, \hat{B}) &= \arg \min_{g, \tilde{B}} \sum_{i=1}^m \omega_i \sum_{1 \leq j \neq l \leq n} (\mathcal{Y}_{ijk} - \tilde{B}_{i,g_j,g_l})^2 \\ \iff (\hat{G}, \hat{B}) &= \arg \min_{g, \tilde{B}} \sum_{i=1}^m \omega_i \|\mathcal{Y}_{i..} - G\tilde{B}_{i..}G^T\|^2 \end{aligned}$$

where  $\omega_i = (\omega_1, \dots, \omega_m)$  is the user-defined weights for each layer. Note that  $\mathcal{Y}_{i..}$  has no diagonal observation.

### 1.1.3 Discussion of the Model

In general, TSBM is a special case of TBM. TBM can handle continuous, binary and even hybrid data on higher-order ( $\geq 3$ ) tensor with membership on multiple ( $\geq 3$ ) mode.

The sparsity settings are quite different. TBM adds regularization term on the objective function while TSBM uses  $\rho_n$  in  $B = \rho_n B^0$ , where  $\|B^0\|_{max} = 1$ , to control the sparsity.

## 1.2 Assumption

### 1.2.1 TBM

- A1. An irreducible core is required. Let  $\delta_{min}$  be the minimal gap between blocks, where  $\delta_{min} = \min_k \delta^{(k)}$ ,  $\delta^{(k)} = \min_{r_k \neq r'_k} \max_{r_1, \dots, r_{k-1}, r_{k+1}, \dots, r_K} (c_{r_1, \dots, r_k, \dots, r_K} - c_{r_1, \dots, r'_k, \dots, r_K})^2$ .
- A2. The signal for each block should be finite and constant. The core tensor  $\mathcal{C}$  is a constant tensor and the largest element of  $\mathcal{C}$  is smaller than the infinity,  $\|\mathcal{C}\|_{max} < +\infty$ .
- A3. The cluster proportion should be lower bounded. Define the minimum cluster proportion as  $\tau = \min_k \min_r \frac{1}{d_k} \sum_i^{d_k} \mathbb{I}[m_{ir}^{(k)} = 1]$ . Then  $\tau$  should be larger than 0,  $\tau > 0$ .

### 1.2.2 TBSM

- B1. Communities should be separated. Define the community separation as  $\delta^2 = \min_{1 \leq j \neq j' \leq K} \|B_{.j} - B_{.j'}\|^2$ . Then the separation should be larger than 0,  $\delta^2 > 0$ .
- B2. The network is sparse. Let  $\mathcal{B} = \rho_0 \mathcal{B}^0$ , where the entries of  $\mathcal{B}^0$  are of constant order and  $\|\mathcal{B}^0\|_{max} = 1$ . The parameter  $\rho$  controls the sparsity of the network.
- B3. The number of layers  $m$  should not go to the infinity faster than the number of network nodes  $n$ , i.e.  $m \leq cn$  for some constant  $c$ .
- B4. Community size should be lower bounded. Let  $n_{min}$  be the smallest community size in  $\mathbf{g}$  and then  $n_{min} > 0$ .

## 1.3 Asymptotic Result

### 1.3.1 TBM

**Theorem 1** (Convergence rate of MSE).

$$MSE(\Theta_{true}, \hat{\Theta}) \leq \frac{C_1 \sigma^2}{\prod_k d_k} (\prod_k R_k + \sum_k d_k \log R_k)$$

with high probability goes to 1 when  $\sum_k d_k \log R_k \rightarrow +\infty$ , where  $C_1$  is a constant.

The first term,  $\prod_k R_k$ , is the parameter number in  $C$  and the second term,  $\sum_k d_k \log R_k$ , is the penalty to estimate the membership. The bound follows the heuristics in matrix sample complexity as below:

$$\frac{(\# \text{ of paras}) + \log(\text{complexity of models})}{\# \text{ of samples}}.$$

**Theorem 2** (Convergence rate of MCR). *Suppose  $MCR(\hat{M}_k, M_{k,true}) = \max_{r \in [R_k], a \neq a' \in [R_k]} \min\{D_{a,r}^{(k)}, D_{a',r}^{(k)}\}$ , where  $D_{r,r'}^{(k)} = \frac{1}{d_k} \sum_{i=1}^{d_k} \mathbb{I}[m_{i,r}^{(k)} = \hat{m}_{i,r'}^{(k)} = 1]$ ,  $r, r' \in [R_k]$ :*

$$P(MCR(\hat{M}_k, M_{k,true}) \geq \epsilon) \leq 2^{1+\sum_k d_k} \exp\left(-\frac{C\epsilon^2 \delta_{min}^2 \tau^{3K-2} \prod_{k=1}^K d_k}{\sigma^2 \|C\|_{max}^2}\right)$$

where  $C$  is a positive constant and  $\tau$  is the cluster proportion lower bound for both estimated and true membership matrix.

Suppose  $d_1 = \dots = d_K = d$  choose  $\epsilon = C' \frac{\sigma \|C\|_{max}}{\delta_{min} \tau^{(3K-2)/2}} d^{-(K-1)/2}$ , we get:

$$MCR(\hat{M}_k, M_{k,true}) \leq C'' \frac{\sigma \|C\|_{max}}{\delta_{min} \tau^{(3K-2)/2}} d^{-(K-1)/2}$$

with probability goes to 1 when  $d \rightarrow +\infty$ .

**Proof Sketch:**

**Thm 1:** Known  $\|\hat{\Theta}_{ols} - \mathcal{Y}\|^2 \leq \|\Theta_{true} - \mathcal{Y}\|^2$ .

1. Prove  $\|\hat{\Theta} - \Theta_{true}\|_F \leq 2 \sup_{\Theta, \Theta'} \langle \frac{\Theta - \Theta'}{\|\Theta - \Theta'\|}, \epsilon \rangle$
2. Use the property of the Gaussian Width.

**Thm 2:**

1. Transfer the optimization problem to  $(\hat{M}_1, \dots, \hat{M}_K) = \arg \max_{M_k} F(M_1, \dots, M_K)$ .
2. Decompose the stochastic deviation:

$$\begin{aligned} F(M_1, \dots, M_K) - F(M_{1,true}, \dots, M_{K,true}) &= F(M_1, \dots, M_K) - G(D^{(1)}, \dots, D^{(K)}) + \\ &\quad G(D^{(1)}, \dots, D^{(K)}) - G(M_{1,true}, \dots, M_{K,true}) + \\ &\quad G(M_{1,true}, \dots, M_{K,true}) - F(M_{1,true}, \dots, M_{K,true}) \end{aligned}$$

First and third deviation are from estimation, second is from label mismatching.

3. Discover the relationship between MCR and mismatching deviation:

$$MCR(\hat{M}_k, M_{k,true}) \geq \epsilon \Rightarrow G(D^{(1)}, \dots, D^{(K)}) - G(M_{1,true}, \dots, M_{K,true}) \leq -\frac{1}{4} \epsilon \tau^{K-1} \delta_{min}$$

4. Known  $F(\hat{M}_{1,ols}, \dots, \hat{M}_{K,ols}) \geq F(\hat{M}_{1,true}, \dots, \hat{M}_{K,true})$  and the upper bound of mismatching deviation, use the property of estimation deviation, i.e. sub-Gaussian, to solve the desire probability for MCR:

$$P(MCR(\hat{M}_k, M_{k,true}) \geq \epsilon) \leq P(\sup |F(M_1, \dots, M_K) - G(D^{(1)}, \dots, D^{(K)})| \geq \frac{\epsilon \tau^{K-1} \delta_{min}}{8})$$

### 1.3.2 TBSM

**Theorem 3** (Main theorem). *The MCR for TBSM,  $\eta_h$ , can be re-written using the definition in TBM. Let  $D$  denote the confusion matrix for the partition  $h$ . Assume  $H$  is the membership matrix corresponds to  $h$  and  $\hat{h}, \hat{H}$  are the estimates. Then  $D_{r,r'} = \frac{1}{n} \sum_i^n \mathbb{I}[H_{i,r} = \hat{H}_{i,r'} = 1], r, r' \in [R]$  and  $\eta_h = \frac{n}{n_{min}} MCR(H, \hat{H}) = \frac{n}{n_{min}} \max_{r \in [R], a \neq a' \in [R]} \min\{D_{a,r}, D_{a',r}\}$ .*

In sparse case  $n\rho_n < \log n$ :

$$\eta_h \leq CK \left(\frac{n}{n_{min}}\right)^2 \left(\frac{m}{\delta^2}\right) \left(\frac{R(\log n)^{3/2}}{n\rho_n m^{1/2}}\right) \left(1 + \frac{R(\log n)^{3/2}}{n\rho_n m^{1/2}}\right)$$

In moderately dense case,  $n\rho_n \geq \log n$ :

$$\eta_h \leq CK \left(\frac{n}{n_{min}}\right)^2 \left(\frac{m}{\delta^2}\right) \left(\frac{R(\log n)}{(n\rho_n m)^{1/2}}\right) \left(1 + \frac{R(\log n)}{(n\rho_n m)^{1/2}}\right)$$

**Proof Sketch:**

1. Transfer the optimization problem to following maximization for partition  $h$  :

$$\text{maximize } f(h; Y) = \sum_{k=1}^R C_{n_k(h)}^2 \left\| \frac{Y * (\omega \circ H_k \circ H_k)}{n_k(h)(n_k(h) - 1)} \right\|^2 + \sum_{1 \leq j < k \leq R} n_j(h)n_k(h) \left\| \frac{Y * (\omega \circ H_j \circ H_k)}{n_j(h)n_k(h)} \right\|^2$$

where  $f(h; Y)$  can be considered as between-group variance.

2. The uniquely optimizer of  $f(h; P)$  is  $h = g$ , up to permutation.

3. Decompose the deviation:  $f(h; Y) - f(g; Y) = f(h; Y) - f(h; P) + f(h; P) - f(g; P) + f(g; P) - f(h; Y)$ . The first and third deviation are from sampling/estimation error. The second deviation comes from mismatching.

4. Find the upper bound of sampling error:  $\sup_h |f(h; Y) - f(h; P)| \leq c_1 \kappa_n$ , with probability tending to 1 as  $n \rightarrow \infty$ . This is from the main technical component of this paper: tensor concentration result.

5. Find the relationship between MCR and mismatching deviation:  $f(g; P) - f(h; P) \geq c_2 \eta_h n_{min}^2 \rho^2 \delta^2 R^{-1}$ . Combine the above results then get desired convergence rate.

### 1.3.3 Discussion on Asymptotic results

Let consider the special case where  $d_1, \dots, d_K = d = m, n$  and  $R_1, \dots, R_K = R$ . In typical balanced community case where the sizes of clusters are roughly equal, the minimum cluster size satisfies  $n_{min} \asymp d$ , and the community separation satisfies  $\delta^2 \asymp m$ , in TBSM. That means  $MCR_{TBSM} = \eta_h$  will vanish as  $d \rightarrow \infty$ . In TBM,  $\tau, \delta_{min}$  can be consider as constant because  $\tau$  refers to the proportion which lies in  $[0, 1]$  and  $\delta_{min}$  is the entry-wise separation. That implies the convergence rate will also be dominate by  $d$  when  $K = 3$  and  $MCR_{TBM}$  will vanish as  $d \rightarrow \infty$ . To explain the similar performances, we can consider the membership matrix for the first mode in TBM is an identical matrix. The convergence rate is still of order  $O(d^{-1})$ , though  $\tau$  will be pretty small.

Another thing needs to be noticed is the number of clusters  $R$ . In TBM,  $R_1, \dots, R_k$  are always fixed while TBSM allow  $R = O(1)$ , which means  $R$  can be bounded by a very large constant. If we allow  $R$  growing,  $MCR(\hat{M}_k, M_{k,true})$  in TBM will get smaller inherently while  $\eta_h = \frac{n}{n_{min}} MCR(\hat{H}, H_{true})$  in

TBSM won't vanish as TBM because  $n_{min} \asymp n/R$ . The relationship  $\eta_h \geq MCR(\hat{M}_k, M_{k,true})$  can also be identified. When  $R$  is fixed, these two criteria are on the same level. In both model,  $R$  should not be too large because growing  $R$  will lead inflated coefficient in the convergence rate.

#### 1.4 Comments on the proofs

**Optimality** The proofs for both paper rely on the optimality of the true partition in population model. That means the  $G(M_1, \dots, M_K) = \mathbb{E}(F(M_1, \dots, M_K))$  in Wang's paper and  $f(h; \mathcal{P})$  in Lei's paper are uniquely maximized only when the partition is equal to the true partition up to label permutation. However, both of them do not require the optimality in the samples. That implies the least square estimate minimized the objective function with observation  $\mathcal{Y}$ ,  $F(\{\hat{M}_k\}) \geq F(\{M_{k,true}\})$  and  $f(\hat{h}; \mathcal{Y}) \geq f(h_{true}; \mathcal{Y})$ , where  $F(M_1, \dots, M_K) = -f(\hat{\mathcal{C}}, M_1, \dots, M_K)$  is a stochastic function depending on samples through  $\hat{\mathcal{C}}$ .

#### Typos and clarification

- In Lei's proof for Lemma 2, around line 420 in the supplement, it is confusing to understand the two inequalities, (1):  $\|P * (\omega \circ H_k \circ H_k)\| \lesssim n_k^2(h)m^{1/2}p_{max}$  and (2):  $\|(Y - P) * (\omega \circ H_k \circ H_k)\| \lesssim n_k(h)(np_{max}) \vee \log n^{1/2}$ .

For inequality (1), we should notice that  $*$  is not inner product. For tensor  $A, B \in \mathbb{R}^{m \times n \times n}$ ,  $A * B$  is the  $m \times 1$  vector obtained by taking entry-wise product of  $A$  and  $B$  and then summing over the second and third mode. Therefore, we can get the left hand,

$$\|P * (\omega \circ H_k \circ H_k)\| \leq \|(n_k^2(h)p_{max}, \dots, n_k^2(h)p_{max})_{m \times 1}\| = \sqrt{m \times (n_k^2(h)p_{max})^2} = m^{1/2} \times n_k^2(h)p_{max}.$$

For inequality (2), it is a direct result of Theorem 2 introduced in Lei's paper. However, the right hand should times  $\log n$  according to Thm 2. This typo leads the rest of the proof incorrect. All the term  $(\log n)^{1/2}$  should be replaced by  $(\log n)^{3/2}$ . Though, their result presented in the section 3 is correct.

- In Wang's proof for Theorem 2, section **A.4.2** in the supplement, the objective function for minimization,  $f(\mathcal{C}, \{M_k\}) = \langle \mathcal{Y}, \Theta \rangle - \frac{\|\Theta\|_2}{2}$  should times a negative sign. Then the problem can transfer to maximize  $F\{M_k\}$  smoothly.

#### 1.5 Algorithm

Algorithms in two paper both are high-ordered extension of the ordinary  $k$ -means clustering. Both of them can not guarantee to converge to a global minimum. However, they are good enough to find estimates with good performance when choosing marginal  $k$ -means result as an initial point.

## 2 QUESTIONS&EXTENSIONS

### 2.1 How to interpret the sparsity parameter $\rho_n$ in Lei's paper?

In Lei's paper, the sparsity setting for outcome tensor  $\mathcal{Y}$  is  $\mathcal{B} = \rho_n \mathcal{B}^0$ , where  $\|\mathcal{B}^0\|_{\max} = 1$ . Intuitively,  $\rho_n$  can be considered as the signal level of the model. Here we can interpret  $\rho_n$  as sparsity control parameter using the following model:

$$\mathcal{Y} = \begin{cases} \text{Ber}(\mathcal{B}^0) & \text{if } Z = 1 \\ 0 & \text{if } Z = 0 \end{cases}, \text{ where } Z \sim \text{Ber}(\rho_n).$$

Therefore, we can get  $\mathbb{E}(\mathcal{Y}) = \rho_n \mathcal{B}^0$  and the sparsity in  $\mathcal{Y}$  is controlled by  $\rho_n$ .

### 2.2 What about Likelihood-based estimate?

Try a likelihood based estimate in TBM:  $\hat{\Theta}_{MLE} = \arg \max_{\Theta} \mathcal{L}_{\mathcal{Y}}(\Theta) = \sum_{i_1, \dots, i_K} f(\theta_{i_1, \dots, i_K})$ . Let  $l(\Theta) = \mathbb{E}[\mathcal{L}_{\mathcal{Y}}(\Theta)]$  be the expectation of log-likelihood function. We know that  $\mathcal{L}_{\mathcal{Y}}(\hat{\Theta}_{MLE}) \geq \mathcal{L}_{\mathcal{Y}}(\Theta_{true})$ . Similar to the proof in tensor regression, we can get following conclusions:

1.  $|\mathcal{L}_{\mathcal{Y}}(\Theta) - l(\Theta)| = |\langle \mathcal{E}, \Theta \rangle|$
2.  $l(\hat{\Theta}_{MLE}) - l(\Theta_{true}) \leq -\frac{L}{2} \|\hat{\Theta}_{MLE} - \Theta_{true}\|_F^2$

where  $L$  is the lower bound for the Hessian matrix. That means  $\frac{\partial^2 l(\Theta)}{\partial \Theta^2} = \text{Var}(\mathcal{Y}|\Theta) = b''(\Theta) \geq L$ . In stochastic model,  $b(\theta) = \log(1 + \exp(\theta))$ ,  $b''(\theta) = \frac{1}{\exp(-\theta) + \exp(\theta) + 2}$ . That means, we should upper bound the signal level  $\|\Theta\|_{\max} < \alpha$ , where  $\alpha$  is a constant.

With bounded signal, we can go through the rest proof as tensor regression.

$$\begin{aligned} 0 &\leq \mathcal{L}_{\mathcal{Y}}(\hat{\Theta}_{MLE}) - \mathcal{L}_{\mathcal{Y}}(\Theta_{true}) \\ &\leq \mathcal{L}_{\mathcal{Y}}(\hat{\Theta}_{MLE}) - l(\hat{\Theta}_{MLE}) - (\mathcal{L}_{\mathcal{Y}}(\Theta_{true}) - l(\Theta_{true})) + (l(\hat{\Theta}_{MLE}) - l(\Theta_{true})) \\ &\leq \langle \mathcal{E}, \hat{\Theta}_{MLE} - \Theta_{true} \rangle - \frac{L}{2} \|\hat{\Theta}_{MLE} - \Theta_{true}\|_F^2 \end{aligned}$$

Let  $M \in \mathcal{M}$  denote the collection of  $K$  membership matrix and  $\mathcal{C}$  denote the core tensor. It is easy to know that  $|\mathcal{M}| \leq \prod_k R_k^{d_k}$ . Assume  $R = \prod_k R_k$ . Then for any  $t > 0$ , we have:

$$\begin{aligned} \mathbb{P}(\|\hat{\Theta}_{MLE} - \Theta_{true}\|_F \geq t) &= \mathbb{P}\left(\sup_{M, M' \in \mathcal{M}} \sup_{\mathcal{C}, \mathcal{C}' \in \mathbb{R}^R} \left| \frac{\Theta(M, \mathcal{C}) - \Theta(M', \mathcal{C}')}{\|\Theta(M, \mathcal{C}) - \Theta(M', \mathcal{C}')\|_F}, \mathcal{E} \right| \geq \frac{Lt}{2} \right) \\ &\leq \exp(2 \sum_k d_k \log R_k + C_1 \prod_k R_k - \frac{C_2 L^2 t^2}{32 \sigma^2}). \end{aligned}$$

Choose  $t = C\sigma\sqrt{\prod_k R_k + \sum_k d_k \log R_k}$ , we can get the same convergence rate as TBM.

#### 2.2.1 Questions to be answered

- May I generalize a higher-order version of TSBM?

Let  $g \in [R]$  be the partition on second and third mode with corresponding membership matrix  $G$ . Assume there is also a block structure on the first model with partition  $h \in [R_0]$  and

its membership matrix  $H$ . Then the least square estimate can be obtained by maximizing the following new  $f(g, h; \mathcal{Y})$ .

$$f(g, h; \mathcal{Y}) = \sum_{k=1}^R \sum_{l=1}^{R_0} \frac{n_k(g)(n_k(g))m_l(h)}{2} \left\| \frac{\langle \mathcal{Y}, (H_l \circ G_k \circ G_k) \rangle}{n_k(g)(n_k(g) - 1)m_l(h)} \right\|^2 +$$

$$\sum_{1 \leq j < k \leq R} \sum_{l=1}^{R_0} n_j(g)n_k(g)m_l(h) \left\| \frac{\langle \mathcal{Y}, (H_l \circ G_j \circ G_k) \rangle}{n_j(g)n_k(g)m_l(h)} \right\|^2$$

### Conjectures:

1. Lemma 1 is still satisfies for  $g, h$ .

**Proof:** Since the maximization problem for  $f(g, h; \mathcal{Y})$  is equal to the minimization problem:

$$\min_{g, h, \tilde{B}} \sum_{1 \leq i \leq m, 1 \leq j \neq l} (\mathcal{Y}_{ijk} - \tilde{B}_{h_i g_k g_l})^2.$$

Replace  $\mathcal{Y}$  by  $\mathcal{P}$  and assume  $(g, h)$  is the true partition which leads to the objective function is 0. If there are different partition  $(g', h')$  (ignoring permutation), there are at least one pairs of point s.t.  $\tilde{B}_{h_i g_k g_l} = \tilde{B}_{h_{i1} g_{k1} g_{l1}}$  while  $\tilde{B}_{h'_i g'_k g'_l} \neq \tilde{B}_{h'_{i1} g'_{k1} g'_{l1}}$ . Then  $\mathcal{P}_{ijk} - \tilde{B}_{h'_i g'_k g'_l}$  and  $\mathcal{P}_{i1j1k1} = \mathcal{P}_{ijk} - \tilde{B}_{h'_{i1} g'_{k1} g'_{l1}}$  can not be both are 0. The objective function for  $(g', h')$  must be strictly positive. Therefore, ignoring the permutation, the optimizer of the objective function is unique.

2. Lemma 2 is still hold on similar level. (Can be proved)

**Proof:**

$$2(f(g, h; \mathcal{Y}) - f(g, h; \mathcal{P})) = \sum_{k=1}^R \sum_{l=1}^{R_0} \frac{\|\langle \mathcal{Y}, (H_l \circ G_k \circ G_k) \rangle\|^2 - \|\langle \mathcal{P}, (H_l \circ G_k \circ G_k) \rangle\|^2}{n_k(g)(n_k(g) - 1)m_l(h)} +$$

$$\sum_{1 \leq j \neq k \leq R} \sum_{l=1}^{R_0} \frac{\|\langle \mathcal{Y}, (H_l \circ G_j \circ G_k) \rangle\|^2 - \|\langle \mathcal{P}, (H_l \circ G_j \circ G_k) \rangle\|^2}{n_j(g)n_k(g)m_l(h)}$$

For the first term, we have:

$$\frac{\|\langle \mathcal{Y}, (H_l \circ G_k \circ G_k) \rangle\|^2 - \|\langle \mathcal{P}, (H_l \circ G_k \circ G_k) \rangle\|^2}{n_k(g)(n_k(g) - 1)m_l(h)}$$

$$= \frac{\langle \mathcal{Y} - \mathcal{P}, (H_l \circ G_k \circ G_k) \rangle^2 + |\langle 2\mathcal{P}, (H_l \circ G_k \circ G_k) \rangle \langle \mathcal{Y} - \mathcal{P}, (H_l \circ G_k \circ G_k) \rangle|}{n_k(g)(n_k(g) - 1)m_l(h)}$$

First, we can get  $|\langle \mathcal{P}, (H_l \circ G_k \circ G_k) \rangle| \lesssim m_l(h)n_k^2(g)p_{max}$ . For  $|\langle \mathcal{Y} - \mathcal{P}, (H_l \circ G_k \circ G_k) \rangle|$ , with Cauchy Schwarz inequality:

$$\langle \mathcal{Y} - \mathcal{P}, (H_l \circ G_k \circ G_k) \rangle^2 = \left( \sum_{i \in h_l} \left( \sum_{j \in G_k} \mathcal{Y}_{ijj} - \mathcal{P}_{ijj} \right) \right)^2 \leq \left( \sum_{i \in h_l} \left( \sum_{j \in G_k} \mathcal{Y}_{ijj} - \mathcal{P}_{ijj} \right)^2 \right) m_l(h)$$

$$= m_l(h) \|(\mathcal{Y} - \mathcal{P}) * (H_l \circ G_k \circ G_k)\|^2$$

Then we can get:

$$\begin{aligned}
\sum_{l=1}^{R_0} \frac{\langle \mathcal{Y} - \mathcal{P}, (H_l \circ G_k \circ G_k) \rangle^2}{n_k(g)(n_k(g) - 1)m_l(h)} &\leq \sum_{l=1}^{R_0} \frac{m_l(h) \|(\mathcal{Y} - \mathcal{P}) * (H_l \circ G_k \circ G_k)\|^2}{n_k(g)(n_k(g) - 1)m_l(h)} \\
&= \frac{\sum_{l=1}^{R_0} \|(\mathcal{Y} - \mathcal{P}) * (H_l \circ G_k \circ G_k)\|^2}{n_k(g)(n_k(g))} \\
&= \frac{\|(\mathcal{Y} - \mathcal{P}) * (\omega \circ G_k \circ G_k)\|^2}{n_k(g)(n_k(g) - 1)} \\
&\lesssim \frac{n_k^2(g) \log^2(n) \{(np_{max}) \vee \log_n\}}{n_k^2(g)}
\end{aligned}$$

And:

$$\begin{aligned}
&\sum_{l=1}^{R_0} \frac{|\langle 2\mathcal{P}, (H_l \circ G_k \circ G_k) \rangle \langle \mathcal{Y} - \mathcal{P}, (H_l \circ G_k \circ G_k) \rangle|}{n_k(g)(n_k(g) - 1)m_l(h)} \\
&\lesssim \sum_{l=1}^{R_0} \frac{m_l(h) n_k^2(g) p_{max} (m_l(h))^{1/2} \|(\mathcal{Y} - \mathcal{P}) * (H_l \circ G_k \circ G_k)\|}{n_k(g)(n_k(g) - 1)m_l(h)} \\
&\lesssim p_{max} \sum_{l=1}^{R_0} (m_l(h))^{1/2} \|(\mathcal{Y} - \mathcal{P}) * (H_l \circ G_k \circ G_k)\| \\
&\leq \sqrt{m} \|(\mathcal{Y} - \mathcal{P}) * (\omega \circ G_k \circ G_k)\| = \sqrt{m} n_k(g) \log(n) \{(np_{max}) \vee \log_n\}^{1/2}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sum_{k=1}^R \sum_{l=1}^{R_0} \frac{\| \langle \mathcal{Y}, (H_l \circ G_k \circ G_k) \rangle \|^2 - \| \langle \mathcal{P}, (H_l \circ G_k \circ G_k) \rangle \|^2}{n_k(g)(n_k(g) - 1)m_l(h)} \\
&\lesssim \sum_{k=1}^R \log^2(n) \{(np_{max}) \vee \log_n\} + n_k(g) \sqrt{m} p_{max} \log(n) \{(np_{max}) \vee \log_n\}^{1/2}
\end{aligned}$$

Similar result can get for the second term in  $2(f(g, h; \mathcal{Y}) - f(g, h; \mathcal{P}))$ . That would lead lemma 2 still holds for the extended problem. 3. Lemma 3 will hold after changing some definition such as  $n_{min}$  and  $\delta$ .

- Can we extend TBM for growing clusters if change the criteria?
- How will MLE perform on MCR?