

# Graphic Lasso: Expectation Issue

Jiixin Hu

January 23, 2021

## 1 Outside Expectation

Suppose  $\mathcal{Y}$  is a binary tensor, where  $\mathcal{Y}_{i_1, \dots, i_K} \sim \text{Ber}(c_{r_1, \dots, r_K})$  independently.

### 1.1 Least squared

Suppose the objective function is the least squared function. With given membership  $\{\mathbf{M}_k\}$ , the estimation of core tensor is

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

The function  $F(\mathbf{M}_k)$  and  $G(\mathbf{M}_k)$  are defined as following.

$$F(\mathbf{M}_k) = \sum_{r_1, \dots, r_K} \prod_k p_{r_k}^{(k)} \hat{c}_{r_1, \dots, r_K}^2, \quad G(\mathbf{M}_k) = \mathbb{E}[F(\mathbf{M}_k)] = \sum_{r_1, \dots, r_K} \prod_k p_{r_k}^{(k)} \mathbb{E}(\hat{c}_{r_1, \dots, r_K}^2). \quad (1)$$

Let  $\mu_{r_1, \dots, r_K} = \mathbb{E}[\hat{c}_{r_1, \dots, r_K}]$ . We have

$$\mu_{r_1, \dots, r_K} = \mathbb{E}[\hat{c}_{r_1, \dots, r_K}] = \frac{1}{\prod_k p_{r_k}^{(k)}} [\mathcal{C} \times_1 \mathbf{D}^{(1),T} \times_2 \dots \times_K \mathbf{D}^{(K),T}]_{r_1, \dots, r_K}.$$

Note that  $\text{Var}(\mathcal{Y}) = V(\mathcal{C}) \times_1 \mathbf{M}_1 \times_2 \dots \times_K \mathbf{M}_K$ , where  $V(c_{r_1, \dots, r_K}) = c_{r_1, \dots, r_K}(1 - c_{r_1, \dots, r_K})$ . Therefore, we have

$$\begin{aligned} \mathbb{E}[\hat{c}_{r_1, \dots, r_K}^2] &= \text{Var}(\hat{c}_{r_1, \dots, r_K}) + [\mathbb{E}(\hat{c}_{r_1, \dots, r_K})]^2 \\ &= \frac{1}{[\prod_k d_k][\prod_k p_{r_k}^{(k)}]^2} [V(\mathcal{C}) \times_1 \mathbf{D}^{(1),T} \times_2 \dots \times_K \mathbf{D}^{(K),T}]_{r_1, \dots, r_K} + \mu_{r_1, \dots, r_K}^2. \end{aligned} \quad (2)$$

Plugging the equation (2) into the definition of  $G(\mathbf{M}_k)$  (1), we have

$$G(\mathbf{M}_k) = \sum_{r_1, \dots, r_K} \prod_k p_{r_k}^{(k)} \mu_{r_1, \dots, r_K}^2 + \sum_{r_1, \dots, r_K} \frac{1}{[\prod_k d_k][\prod_k p_{r_k}^{(k)}]} [V(\mathcal{C}) \times_1 \mathbf{D}^{(1),T} \times_2 \dots \times_K \mathbf{D}^{(K),T}]_{r_1, \dots, r_K}.$$

Since the estimation of the core tensor  $\mathcal{C}$  is related to both the mean of  $\mathcal{Y}$  but also the variance of  $\mathcal{Y}$ , the error for misclassification can be separated into two parts.

Suppose we have the true membership, the first term  $I_1 = \sum_{r_1, \dots, r_K} \prod_k p_{r_k}^{(k)} \mu_{r_1, \dots, r_K}^2$  is maximized with true membership. However, the variance term  $I_2 = \sum_{r_1, \dots, r_K} \frac{1}{[\prod_k d_k][\prod_k p_{r_k}^{(k)}]} [V(\mathcal{C}) \times_1$

$\mathbf{D}^{(1),T} \times_2 \cdots \times_K \mathbf{D}^{(K),T}]_{r_1, \dots, r_K}$  may not achieve the maxima with the true membership. Here is a counter-example.

### Counter Example

Let  $\{\mathbf{M}_k\}$  denote the true membership, and  $\{\hat{\mathbf{M}}_k\}$  denote the estimation of the membership. Then we have

$$I_2(\mathbf{M}_k) = \sum_{r_1, \dots, r_K} \frac{1}{\prod_k d_k} V(\mathcal{C})_{r_1, \dots, r_K},$$

and

$$I_2(\hat{\mathbf{M}}_k) = \sum_{r_1, \dots, r_K} \frac{1}{[\prod_k d_k][\prod_k \hat{p}_{r_k}^{(k)}]} [V(\mathcal{C}) \times_1 \mathbf{D}^{(1),T} \times_2 \cdots \times_K \mathbf{D}^{(K),T}]_{r_1, \dots, r_K}.$$

Consider a special case. Suppose  $\hat{\mathbf{M}}_k = \mathbf{M}_k$  for  $k = 2, \dots, K$ . Then  $\mathbf{D}^{(k)}$  are diagonal matrices and  $\hat{p}_{r_k}^{(k)} = p_{r_k}^{(k)}$  for  $k = 2, \dots, K$ . Since the misclassification happens only on the first mode, let  $MCR(\hat{\mathbf{M}}_1, \mathbf{M}_1) = \epsilon$ . Fixed  $r_2, \dots, r_K$ , assume the classification happens only between blocks  $(r_1, r_2, \dots, r_K)$  and  $(r'_1, r_2, \dots, r_K)$  keeping the cluster proportions the same as true proportion  $\hat{p}_{r_1}^{(1)} = p_{r_1}^{(1)}, \hat{p}_{r'_1}^{(1)} = p_{r'_1}^{(1)}$ . This setting describes the case that  $d_1\epsilon$  elements are switched from block  $r_1$  to  $r'_1$ , where other membership on mode 1 are correct. For simplicity, let  $V(\mathcal{C})_{r_1, \dots, r_K} = V(c)$  and  $V(\mathcal{C})_{r'_1, \dots, r_K} = V(c')$ . Therefore, we only need to compare

$$\tilde{I}_2(\mathbf{M}_k) = V(c) + V(c'),$$

and

$$\tilde{I}_2(\hat{\mathbf{M}}_k) = \frac{d_1 D_{r_1, r_1}^{(1)} V(c) + d_1 \epsilon V(c')}{d_1 p_{r_1}^{(1)}} + \frac{d_1 D_{r'_1, r'_1}^{(0)} V(c') + d_1 \epsilon V(c)}{d_1 p_{r'_1}^{(1)}},$$

where  $D_{r_1, r_1}^{(1)} + \epsilon = p_{r_1}^{(1)}$  and  $D_{r'_1, r'_1}^{(0)} + \epsilon = p_{r'_1}^{(1)}$ .

By a straight forward calculation, we have

$$\tilde{I}_2(\hat{\mathbf{M}}_k) - \tilde{I}_2(\mathbf{M}_k) = \left( \frac{\epsilon}{p_{r'_1}^{(1)}} - \frac{\epsilon}{p_{r_1}^{(1)}} \right) V(c) + \left( \frac{\epsilon}{p_{r_1}^{(1)}} - \frac{\epsilon}{p_{r'_1}^{(1)}} \right) V(c').$$

Note that the proportion of cluster and the magnitude of variance are independent. Let  $p_{r_1}^{(1)} < p_{r'_1}^{(1)}$ . The subtraction  $\tilde{I}_2(\hat{\mathbf{M}}_k) - \tilde{I}_2(\mathbf{M}_k) < 0$  when  $V(c) > V(c')$ , and  $\tilde{I}_2(\hat{\mathbf{M}}_k) - \tilde{I}_2(\mathbf{M}_k) > 0$  when  $V(c) < V(c')$ . This implies that the term  $I_2$  may not achieve the maxima with true membership.