Feb 14 meeting

Zhuoyan Xu

February 21, 2019

1 Biclustering

Clustering One way clustering, clustering the n observations on the basis of p features or vice versa.

Transposable data Characterized by the fact that both rows and columns are of scientific interested and may contain clusters or other structure.

When we encountered transposable data, one way clustering cannot reflect the fact that both rows and columns are of scientific interest. Then we use biclustering to simultaneously clustering the rows and columns of a data matrix. Bicluster is a subset of the data matrix, which are usually three distinct types.

The simplest one is a constant bicluster, in which all elements take on approximately a constant value. Consider a large $n \times p$ matrix $X = [x_{ij}]$ with n observations and features. We assume n observations belong to K unknown and non-overlapping classes $C_1, \ldots, C_K$, and p features belong to R unknown and non-overlapping classes $D_1, \ldots, D_R$.

Under Gaussian assumption, our model can be written as:

$$X_{ij} = \mu_{kr} + \epsilon_{ij} \quad , \ \epsilon_{ij} \sim i.i.d \ N(0, \sigma^2)$$

Where μ_{kr} is the mean value in $|C_k||D_r|$.

In a matrix way it can be shown as:

$$X_{n\times p} = A_{n\times K} U_{K\times R} B_{R\times p} + [\epsilon_{ij}]$$

where A is the matrix contains dummy variables indicates section in rows, and B is the matrix contains dummy variables indicates section in columns. rank(A) = K,rank(B) = R. Minimizing the log-likelihood is equivalent to

$$\min_{C_1,\ldots,C_K,D_1,\ldots,D_R,\mu\in\mathbb{R}^{K\times R}} \{\sum_{k=1}^{K}\sum_{p=1}^{P}\sum_{i\in C_k}\sum_{j\in D_r}(X_{ij}-\mu_{kr})^2\}$$

The computation is straightforward:

1. Given U,B, compute A using kmeans.

2. Given A,B, compute U as the mean of that block.

3. Given A,U, compute B using kmeans.

Under Bernoulli assumption, where response X is binary, the transformation need to be used on response.

$$X_{ij} \sim \quad i.i.d\ Bernoulli(\mu_{kr})$$

Usually we use logit transformation:

$$logit(\mathbb{E}[X_{n\times p}]) = log(\frac{\mathbb{E}[X_{n\times p}]}{1-\mathbb{E}[X_{n\times p}]}) = A_{n\times K}U_{K\times R}X_{R\times p}$$

The likelihood function $L(\mu) = f(X_{11}, \ldots, X_{np})$ is:

$$f(X_{11}, \ldots, X_{np}) = \prod_{k=1}^{K}\prod_{p=1}^{P}\prod_{i\in C_k}\prod_{j\in D_r}\mu_{kr}^{X_{ij}}(1-\mu_{kr})^{1-X_{ij}}$$

Take logarithm, we have:

$$\begin{aligned} l(\mu) =& \sum_{k=1}^{K}\sum_{p=1}^{P}\sum_{i\in C_k}\sum_{j\in D_r}\{X_{ij}log(\frac{\mu_{kr}}{1-\mu_{kr}}) + log(1-\mu_{kr})\} \\ =& \sum_{k=1}^{K}\sum_{p=1}^{P}\sum_{i\in C_k}\sum_{j\in D_r}\{X_{ij}log([AUB]_{ij}) - log(1+\exp[AUB]_{ij}) \end{aligned}$$

2 Tensor Biclustering

In tensor language, it's almost the same as matrix. onsider a large $I \times J \times M$ tensor $X = [x_{ijm}]$. We assume values in mode-1 belong to K unknown and non-overlapping classes $C_1, \ldots, C_K$, and values in mode-2 belong to R unknown and non-overlapping classes $D_1, \ldots, D_R$, values in mode-3 belong to G unknown and non-overlapping classes $F_1, \ldots, F_G$. Under Gaussian assumption, our model can be written as:

$$X_{ijm} = \mu_{krg} + \epsilon_{ijm} \quad , \ \epsilon_{ijm} \sim i.i.d\ N(0, \sigma^2)$$

Where μ_{krg} is the mean value in $|C_k||D_r||F_g|$.

In a matrix way it can be shown as:

$$X^{I\times J\times M} = U^{K\times R\times G} \times_1 A^{I\times K} \times_2 B^{J\times R} \times_3 C^{M\times G} + [\epsilon_{ijm}]$$

where A is the matrix contains dummy variables indicates section in mode-1 fibers(columns), and B is the matrix contains dummy variables indicates section in mode-2 fibers(rows), and C is the matrix contains dummy variables indicates section in mode-3 fibers(tubes). rank(A) = K,rank(B) = R, rank(C) = G.

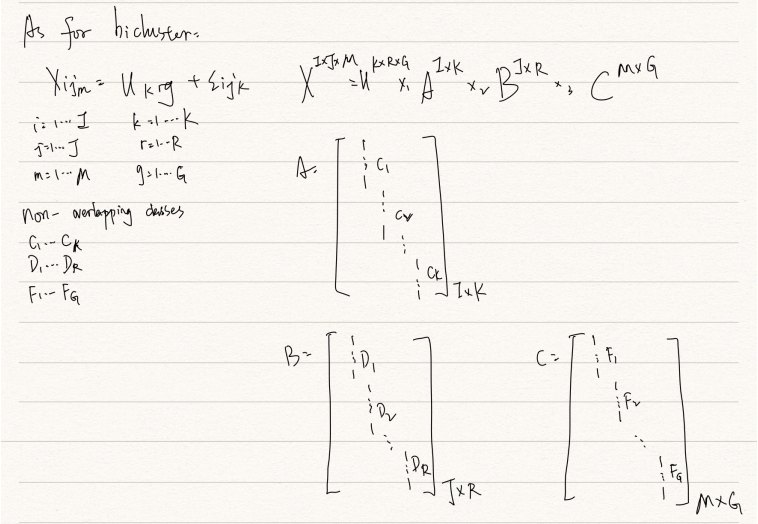


Figure 1: handwritten

Minimizing the log-likelihood is equivalent to

$$\min_{C_1,\ldots,C_K,D_1,\ldots,D_R,F_1,\ldots,F_G,\mu\in\mathbb{R}^{K\times R\times G}} \{\sum_{k=1}^{K}\sum_{p=1}^{P}\sum_{g=1}^{G}\sum_{i\in C_k}\sum_{j\in D_r}\sum_{m\in F_g}(X_{ijm}-\mu_{krg})^2\}$$

Under Bernoulli assumption, where response X is binary, the transformation need to be used on response.

$$X_{ijm} \sim \quad i.i.d\ Bernoulli(\mu_{krg})$$

Similarly, we use logit transformation, the log-likelihood function would be:

$$l(\mu) = \sum_{k=1}^{K}\sum_{p=1}^{P}\sum_{g=1}^{G}\sum_{i\in C_k}\sum_{j\in D_r}\sum_{m\in F_g}\{X_{ijm}log([UABC]_{ijm}) - log(1+\exp[UABC]_{ijm})\}$$

3 Simulation

3.1 simulation1

When I try out the simulation1 in section 6.2 in Tan and Witten(2014). I use n = 100, p = 50, K = 3, R = 3. The result is shown below:

The first one is when I set λ is 0.

	method	rowCER	colCER	sparsity_rate
1	kmean	0.18(0.10)	0.19(0.11)	0.00(0.00)
2	sparseBC	0.12(0.09)	0.10(0.10)	0.00(0.00)

Then I use BIC criterion to select λ as described in section 5.2 in Tan and Witten(2014). The result shows when λ is zero, it has the lowest BIC.

The estimated mean matrix is shown below:

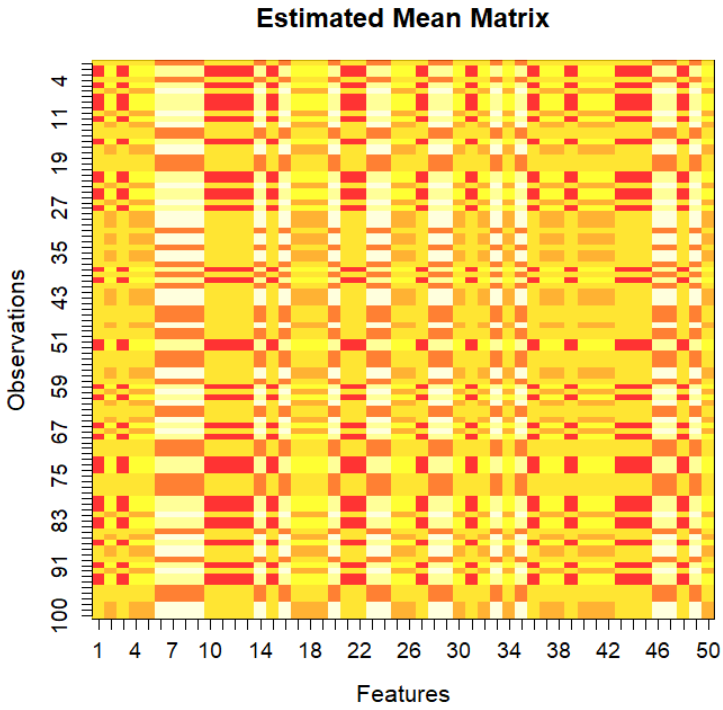
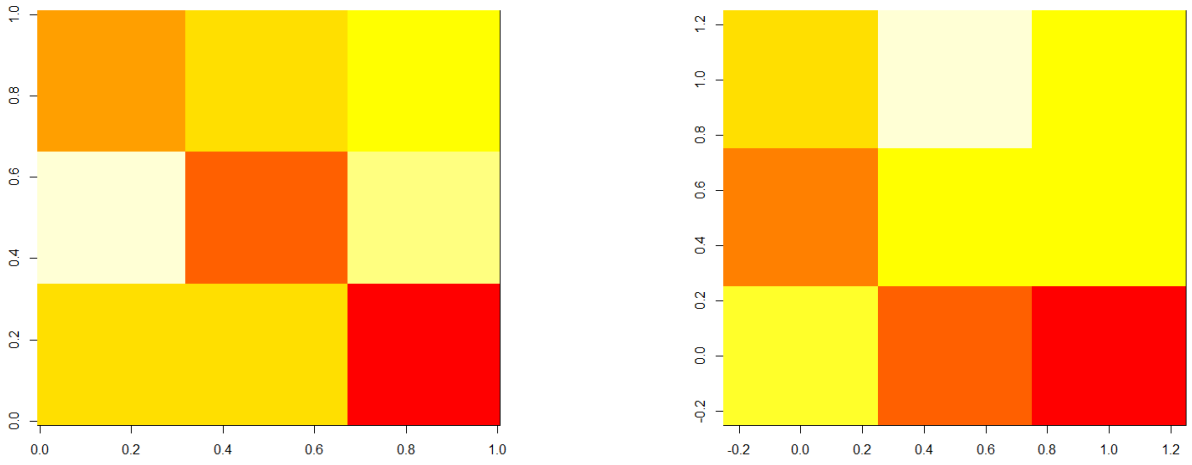


Figure 2: Estimated X

After resorted and put the datapoints that in same cluster together, we have subfigure a, and the true mean of each cluster(true μ) is shown in subfigure b.



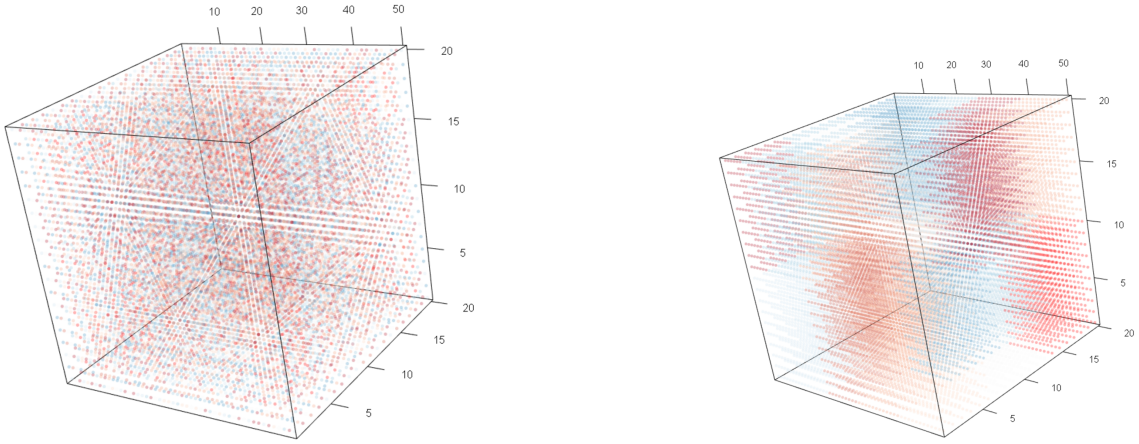
(a) clutering result

(b) true value μ

Figure 3: The comparison

Thus, we have the same results.

4 tensor plot



(a) true raw tensor

(b) true ordered tensor

Figure 4: The comparison