Tensor Bclustering and Factorization

Zhuoyan Xu

February 28, 2019

1 Biclustering: Some Concepts

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:

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

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 $F_1, \ldots, F_G$.

2.1 Gaussian assumption

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

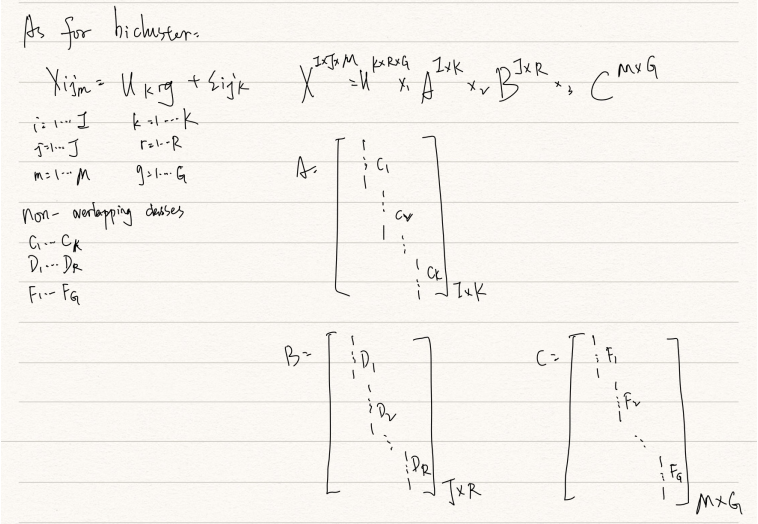


Figure 1: handwritten

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

Minimizing the log-likelihood is equivalent to

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

In matrix form:

$$\min_{U,A,B,C} ||X - U \times_1 A \times_2 B \times_3 C||_F^2$$

2.2 Bernoulli assumption

Under Bernoulli assumption, where response X is binary, the transformation need to be used on mean(or expectation) of response.

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

Similarly, we use logit transformation:

$$logit(\mathbb{E}[X_{I\times J\times M}]) = log(\frac{\mathbb{E}[X_{I\times J\times M}]}{1-\mathbb{E}[X_{I\times J\times M}]}) = U \times_1 A \times_2 B \times_3 C$$

the log-likelihood function would be:

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

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:

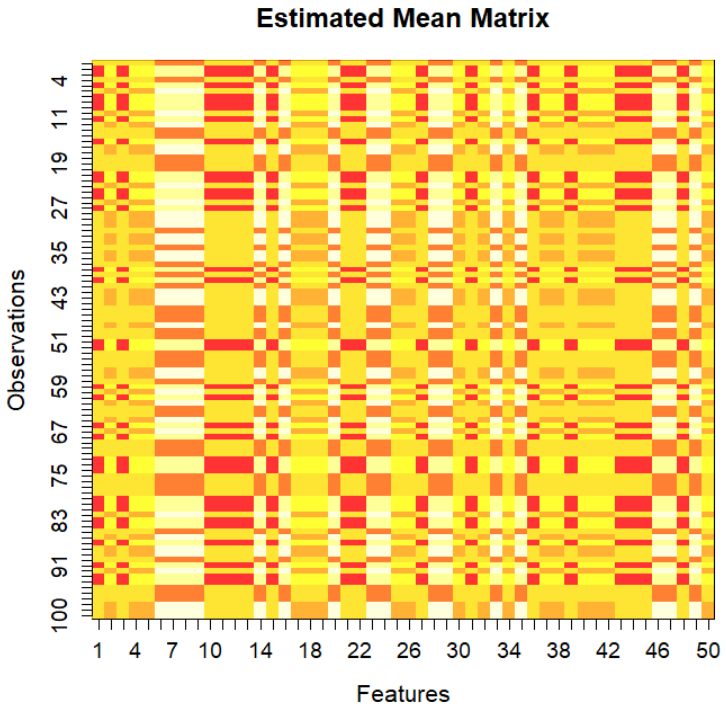
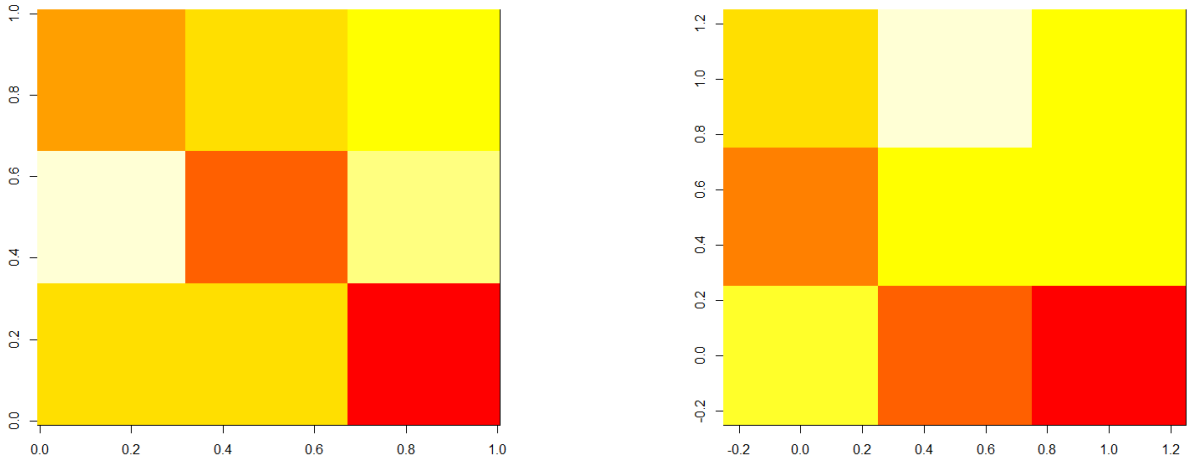


Figure 2: Estimated X

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.

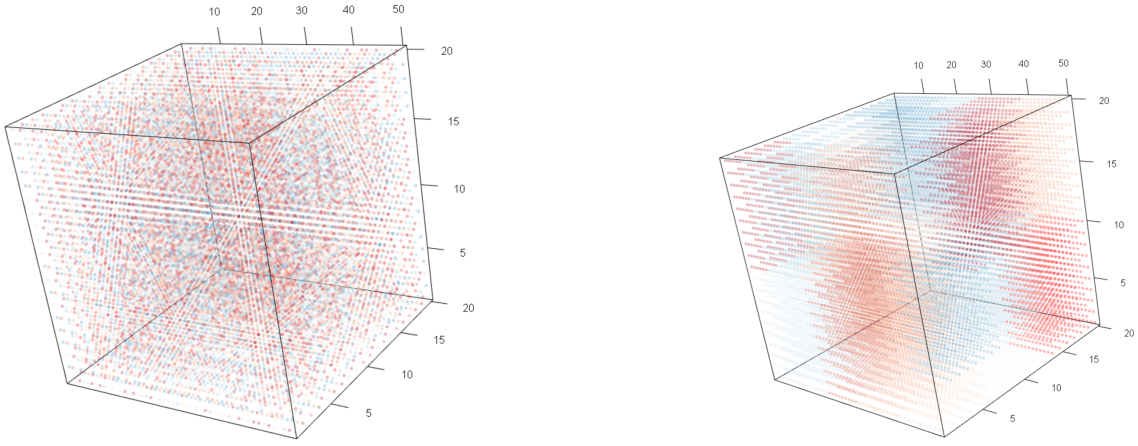


(a) clutering result (b) true value μ

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

5 Some Views of Tensor Biclustering

5.1 Gaussian case

Consider loss function in matrix form:

$$\min_{U,A,B,C} ||X - U \times_1 A \times_2 B \times_3 C||_F^2$$

Where $X \in R^{d_1 \times d_2 \times d_3}$, $U \in R^{r_1 \times r_2 \times r_3}$.

Degree of Freedom

The degree of freedom of this problem is :

$$r_1 r_2 r_3 + r_1^{d_1} + r_2^{d_2} + r_3^{d_3} - r_1! - r_2! - r_3!$$

In this case:

- $r_1 r_2 r_3$ comes from the μs in core matrix U.
- For mode-1, each observations need to be assigned to one of the r_1 clusters, and there are total d_1 observations in mode 1, which makes $r_1^{d_1}$..
- Since for each membership matrix, when shuffling two groups and the corresponding estimated μ in U makes no change(for e.g. $UA = (UP^{-1})(PA)$), which makes $-r1!$.

Alleviate Constrain

Consider matrix form, rank(A) = r_1, rank(B) = r_2, rank(C) = r_3. For membership matrix A(similar as B,C) There are 3 constrains:

- $A \in R^{d_1 \times r_1}$.
- A = $[a_{ij}]$, $a_{ij} \in \{0, 1\}$.
- $A^T A = I$

Under full constrains, A is membership matrix. If we exclude second constrain(0,1 constrain), we got sparse matrix. If we exclude sparse constrain, we got SVD orthogonal matrix.

5.2 Binary case

loss function/criterion

Why not use

$$\min_{U,A,B,C} ||X - U \times_1 A \times_2 B \times_3 C||_F^2$$

1. Since the y is binary, cross entropy is a better loss function than MSE.
2. logit transformation make sure the output probability(after sigmoid or softmax) is in [0,1].
3. MLE estimate is asymptotically most efficient and robust. In Gaussian case, MLE estimate is also the optimal solution of MSE.

6 Extension to include covariate

Consider the following scenario:

Suppose $Y \in R^{d_1 \times d_2 \times d_3}$ is a 0,1 tensor contains the information about scholars went conferences and their keyword. The mode-1 is individuals, mode-2 is conference, mode-3 is keyword. Y = $[y_{ijk}]$, and:

$$y_{ijk} = \begin{cases} 1 & individual\ i\ publish\ at\ conference\ j\ about\ keyword\ k \\ 0 & otherwise \end{cases}$$

If elements in A,B,C are all binary(A,B,C is membership matrix), we call it **hard clustering**. Otherwise we call it **soft clustering**.

Suppose $X \in R^{d_1 \times p}$ is the additional information about individuals(mode-1 of Y), such as the partition or university or age or gender of these individuals. As for the information about university or gender, the X can be binary use Onehot encoder to denote it, as for the information age or other, the X can be continuous. This case we call it **supervised clustering**

6.1 Gaussian case

Consider the question might be hard to solve, we begin with Gaussian case, which makes elements in Y is continuous and Gaussian distribution. We still use the MSE loss, there are two ideas to solve it.

Regression

Under this condition, we just replace A with X:

$$\min_{U,A,B,C} ||Y - U \times_1 X \times_2 B \times_3 C||_F^2$$

Where $U \in R^{P \times r_2 \times r_3}$, $X \in R^{d_1 \times P}$. We just use iteration method to get result.

Penalty

Under this condition, we add a penalty:

$$\min_{U,A,B,C} ||Y - U \times_1 A \times_2 B \times_3 C||_F^2 + \alpha||A - X||_F^2 + \lambda||A||_{2,1}$$

Where α, λ is penalty parameter, the last penalty is regularization requires $A \in R^{d_1 \times p}$ to be sparse.

More complexity

If we still keep the dimension of A is $R^{d_1 \times r_1}$ If we suppose $W \in R^{d_1 \times C}$(divide individuals into C classes), there is a matrix coefficients $W \in R^{r_1 \times p}$ connect A and X such as:

$$X = AW$$

we have:

$$\min_{U,A,B,C} ||Y - U \times_1 A \times_2 B \times_3 C||_F^2 + \alpha||AW - X||_F^2 + \lambda||W||_{2,1}$$

This is concretely discussed in [Bokai Cao, Chun-Ta Lu,*Semi-supervised Tensor Factorization for Brain Network Analysis*]. This paper use ADMM to compute the result.

6.2 Binary case

Classification

Under this condition, we just replace A with X:

$$logit(\mathbb{E}[Y]) = log(\frac{\mathbb{E}[Y]}{1 - \mathbb{E}[Y]}) = U \times_1 X \times_2 B \times_3 C$$

the log-likelihood function would be:

$$l(\mu) = \sum_{i=1}^{d_1}\sum_{k=1}^{d_2}\sum_{k=1}^{d_3}\{Y_{ijk}([UXBC]_{ijk}) - log(1 + \exp[UXBC]_{ijk})\}$$

Where $U \in R^{P \times r_2 \times r_3}$, $X \in R^{d_1 \times P}$. We just use iteration method to get result.

Penalty

Under this condition, we add a penalty:

$$\max_{U,A,B,C} \sum_{i=1}^{d_1}\sum_{k=1}^{d_2}\sum_{k=1}^{d_3}\{Y_{ijk}([UABC]_{ijk}) - log(1 + \exp[UABC]_{ijk})\}$$
$$subject\ to\ A = X$$

Where $A \in R^{d_1 \times p}$ to be sparse.

More complexity

If we still keep the dimension of A is $R^{d_1 \times r_1}$ If we suppose $W \in R^{d_1 \times C}$(divide individuals into C classes), there is a matrix coefficients $W \in R^{r_1 \times p}$ connect A and X such as:

$$X = AW$$

we have:

$$\max_{U,A,B,C} \sum_{i=1}^{d_1}\sum_{k=1}^{d_2}\sum_{k=1}^{d_3}\{Y_{ijk}([UABC]_{ijk}) - log(1 + \exp[UABC]_{ijk})\}$$
$$subject\ to\ AW = X$$

,