Research Note 1

Jiaxin Hu

Email: jhu267@wisc.edu

Date: 2019.2.14

1 Clustering in Matrix and Tensor with Normal Settings

1.1 Biclustering in Matrix with Normal setting

Settings

- Suppose we have a $n \times p$ data matrix $\mathbf{Y}$ which has n rows and p columns

- Suppose there is a partition of rows $C_1, \ldots, C_K$ which separates the rows into K subgroups. The partition C_k satisfies:
  1. $C_k \subseteq \{1, 2, \ldots, n\},\ \forall k \in \{1, 2, \ldots, K\}$
  2. $C_i \cap C_j = \emptyset,\ \forall i \neq j\ and\ i, j \in \{1, 2, \ldots, K\}$
  3. $C_1 \cup C_2 \cup \cdots \cup C_K = \{1, 2, \ldots, n\}$

- Suppose there is a partition of columns $D_1, \ldots, D_R$ which separates the columns into R subgroups. The partition D_r satisfies:
  1. $D_r \subseteq \{1, 2, \ldots, p\},\ \forall r \in \{1, 2, \ldots, R\}$
  2. $D_i \cap D_j = \emptyset,\ \forall i \neq j\ and\ i, j \in \{1, 2, \ldots, R\}$
  3. $D_1 \cup D_2 \cup \cdots \cup D_R = \{1, 2, \ldots, p\}$

- There are $R \times K$ biclusters of the data matrix in total.

Model and Assumptions

- Construct a model:

$$Y_{ij} = \mu_{kr} + \epsilon_{ij},\ \epsilon_{ij} \sim_{i.i.d} N(0, \sigma^2),\ i = 1, \ldots, n,\ j = 1, \ldots, p$$

  where refers to $\mathbb{E}Y_{ij} = \mu_{kr}$ if Y_{ij} belongs to the bicluster B_{kr}.

- The linear model can be rewritten in a matrix form:

$$\mathbf{Y}_{n \times p} = A_{n \times K} \mu_{K \times R} B^T_{p \times R} + \epsilon_{n \times p},\ \epsilon_{ij} \sim_{i.i.d} N(0, \sigma^2),\ i = 1, \ldots, n,\ j = 1, \ldots, p$$

where A, B are membership matrix:

$$A = (a_{ik}) \begin{cases} 1 & row\ i\ is\ in\ group\ C_k \\ 0 & otherwise \end{cases}$$

$$B = (b_{jr}) \begin{cases} 1 & column\ j\ is\ in\ group\ D_r \\ 0 & otherwise \end{cases}$$

Loss Our goal is to find the biculsters of the data matrix i.e. to find the partitions C_k, D_r with unknown parameters K, R, μ_{kr} and λ in a sparse clustering case which can minimize the loss function.

- Loss function in non-sparse case:

$$\min_{\{C_k\}_K, \{D_r\}_R, \mu \in \mathbb{R}^{K \times R}} \left\{ \sum_{k=1}^{K} \sum_{r=1}^{R} \sum_{i \in C_k} \sum_{j \in D_r} (Y_{ij} - \mu_{kr})^2 \right\}$$

- Loss function in sparse case:

$$\min_{\{C_k\}_K, \{D_r\}_R, \mu \in \mathbb{R}^{K \times R}} \left\{ \frac{1}{2} \sum_{k=1}^{K} \sum_{r=1}^{R} \sum_{i \in C_k} \sum_{j \in D_r} (Y_{ij} - \mu_{kr})^2 + \lambda \sum_{k=1}^{K} \sum_{r=1}^{R} |\mu_{kr}| \right\}$$

  which add a *lasso* or l_1 penalty to the original least square loss function. The larger the λ is, the sparser the μ is.

1.2 Clustering in Tensor with Normal setting

Settings

- Suppose we have a $n_1 \times n_2 \times \cdots \times n_D$ data tensor $Y_{i_1, i_2, \ldots, i_D}$ which has D modes. And in each mode M_d, the data is n_d dimensional.

- Suppose there is a partition in each mode $M_{d1}, \ldots, M_{dK_d}, d = 1, \ldots, D$ which separates the mode M_d into K_d subgroups. The partition $M_{dk_d}, d = 1, \ldots, D$ satisfies:
  1. $M_{dk_d} \subseteq \{1, 2, \ldots, n_d\},\ \forall k_d \in \{1, 2, \ldots, K_d\}$
  2. $M_{di} \cap M_{dj} = \emptyset,\ \forall i \neq j\ and\ i, j \in \{1, 2, \ldots, K_d\}$
  3. $M_{d1} \cup M_{d2} \cup \cdots \cup M_{dK_d} = \{1, 2, \ldots, n_d\}$

- There are $K_1 \times K_2 \times \cdots \times K_D$ clusters of the data tensor in total.

Model and Assumptions

- Construct a model:

$$Y_{i_1,i_2,\dots,i_D} = \mu_{k_1,k_2,\dots,k_D} + \epsilon_{i_1,i_2,\dots,i_D},\ \ \epsilon_{i_1,i_2,\dots,i_D} \sim_{i.i.d} N(0,\sigma^2)$$
$$i_d = 1,2,\dots,n_d;\ \ k_d = 1,2,\dots,K_d;\ \ d = 1,2,\dots,D$$

  where refers to $\mathbb{E}Y_{i_1,i_2,\dots,i_D} = \mu_{k_1,k_2,\dots,k_D}$ if $Y_{i_1,i_2,\dots,i_D}$ belongs to the cluster $B_{k_1,k_2,\dots,k_D}$.

- Write the model in a tensor form:

$$\mathbf{Y}_{n_1\times n_2\times\cdots\times n_D} = \mathbf{U}_{K_1\times K_2\times\dots K_D} \times_1 A_1^T \times_2 A_2^T \times_3 \cdots \times_D A_D^T + \epsilon_{n_1\times n_2\times\cdots\times n_D}$$
$$\epsilon_{i_1,i_2,\dots,i_D} \sim_{i.i.d} N(0,\sigma^2),\ \ i_d = 1,2,\dots,n_d;\ \ d = 1,2,\dots,D$$

  where A_d would be the membership matrix of d mode with dimension $n_d \times K_d$:

$$A_d = (a_{i_d k_d}) \begin{cases} 1 & if\ the\ i_d th\ element\ in\ mode\ d\ belongs\ to\ partition\ M_{dk_d} \\ 0 & otherwise \end{cases}$$

  And in the model, $\times_d$ refers to the multiplication between a matrix and a tensor through the d mode. $\mathbf{U}$ refers to the tensor of $\mu_{k_1,\dots,k_d}$

Loss Function Similar with biclustering, our goal is to find the partition that minimize the loss function with unknown parameter $K_1,\dots,K_D,\mathbf{U}$ and penalty parameter λ. In the tensor case, the loss function is in the similar formula of matrix and in this case I just give the loss function in sparse case.

- Loss function in sparse case:

$$\min_{\{M_{1k_1}\}_{K_1},\dots\{M_{Dk_d}\}_{K_D},\mathbf{U}\in\mathbb{R}^{K_1\times\cdots\times K_D}} \sum_{k_1=1}^{K_1}\cdots\sum_{k_D=1}^{K_D}\sum_{i_1\in M_{1k_1}}\cdots\sum_{i_D\in M_{Dk_D}}(Y_{i_1,\dots,i_D}-\mu_{k_1,\dots,k_D})^2 \\ + \lambda\sum_{k_1=1}^{K_1}\cdots\sum_{k_D=1}^{K_D}|\mu_{k_1,\dots,k_D}|$$

2 Clustering in Matrix and Tensor with Binary Settings

2.1 Biclustering in Matrix with binary setting

Settings

- The partition setting of $C_1, \dots, C_K$ and $D_1, \dots, D_R$ would be the same with the normal setting. There are $R \times K$ biclusters in total.

- The elements in the $n \times p$ data matrix $\mathbf{Y}$ belong to $\{0, 1\}$.

Model and Assumption

- For $\mathbf{Y}$ is a binary data matrix, we assume Y_{ij} follows Bernoulli distribution:

$$Y_{ij} \sim Ber(p_{ij});\ \mathbb{E}Y_{ij} = p_{ij};\ logit(\mathbb{E}Y_{ij}) = log\frac{p_{ij}}{1-p_{ij}} = \mu_{kr}$$
$$i = 1, \dots, n;\ j = 1, \dots, p;\ Y_{ij}\ \textit{are independent.}$$

- Rewrite the model in a matrix form:

$$logit(\mathbb{E}\mathbf{Y})_{n\times p} = log(\mathbf{P})_{n\times p} = A_{n\times K}\mathbf{U}_{K\times R}B^T_{p\times R}$$

  where $P_{ij} = \frac{p_{ij}}{1-p_{ij}}$, $U_{kr} = \mu_{kr}$ and A, B are membership matrices defined above.

Optimization Maximum Likelihood?

2.2 Clustering in Tensor with Binary setting

Settings

- The partition and the modes settings of the data tensor are the same with normal setting. The data $\mathbf{Y}_{n_1\times\cdots\times n_D}$ has D modes and on each mode we have a partition $M_{d1}, \dots M_{dK_d}, d = 1, \dots, D$. There are $K_1 \times \dots K_D$ clusters in total.

- Every element in the data belongs to $\{0, 1\}$.

Model and Assumption

- For the data is binary, we assume the data follows a Bernoulli distribution and construct a logistic model:

$$
\begin{aligned}
&Y_{i_1,\dots,i_D} \sim Ber(p_{i_1,\dots,i_D});\ \mathbb{E}Y_{i_1,\dots,i_D} = p_{i_1,\dots,i_D};\ i_d = 1,\dots,n_d \\
&logit(\mathbb{E}Y_{i_1,\dots,i_D}) = log\frac{p_{i_1,\dots,i_D}}{1-p_{i_1,\dots,i_D}} = \mu_{k_1,\dots,k_D};\ k_d = 1,\dots,K_d \\
&Y_{i_1,\dots,i_D}\ are\ independent
\end{aligned}
$$

- Rewrite the model in tensor form:

$$
logit(\mathbb{E}\mathbf{Y})_{n_1\times\cdots\times n_D} = log(\mathbf{P})_{n_1\times\cdots\times n_D} = \mathbf{U}_{K_1\times K_2\times\dots K_D} \times_1 A_1^T \times_2 A_2^T \times_3 \cdots \times_D A_D^T
$$

where $P_{i_1,\dots,i_D} = \frac{p_{i_1,\dots,i_D}}{1-p_{i_1,\dots,i_D}}$, $U_{k_1,\dots,k_D} = \mu_{k_1,\dots,k_D}$ and $A_1,\dots,A_D$ are membership matrices defined above.

Optimization

3 Additional Simulation of Table 3

Simulation settings In this simulation case, $n = 100, p = 50, K = 3, R = 3$ and $Y_{ij} \sim_{i.i.d} N(\mu_{kr}, 4^2)$, $\mu_{kr} \sim Unif(-2, 2)$. Replication time $S = 50$.

	rowCER	colCER	sd of rCER	sd of cCER	sparse rate	selected λ
kmeans	0.2921	0.2541	0.0985	0.1256	-	
Bicluster $\lambda = 0$	0.2663	0.2326	0.1014	0.1257	-	
Bicluster $\lambda = 200$	0.2768	0.2354	0.0973	0.1290	0.4132	
Bicluster $\lambda = 400$	0.2921	0.2516	0.0902	0.1389	0.4342	
Bicluster $\lambda = 800$	0.3875	0.3432	0.1593	0.1804	0.3224	
Bicluster λ selected	0.2659	0.2332	0.1003	0.1355	0	51

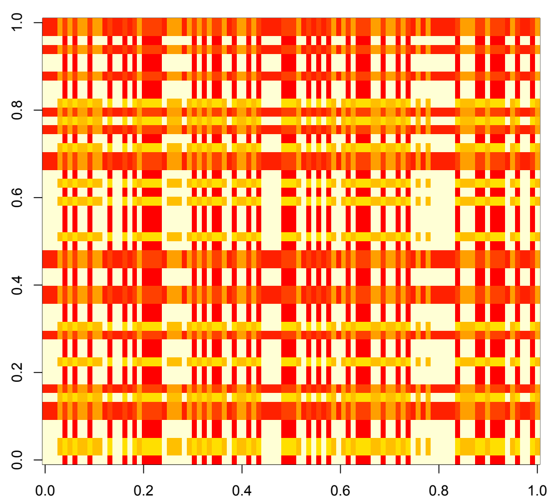

Figure 1: ground truth

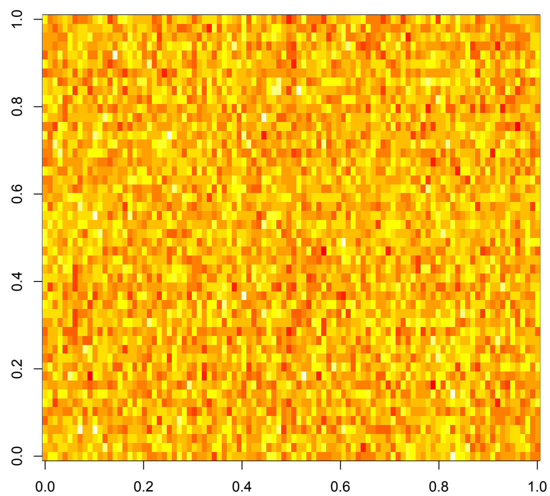

Figure 2: origin data matrix

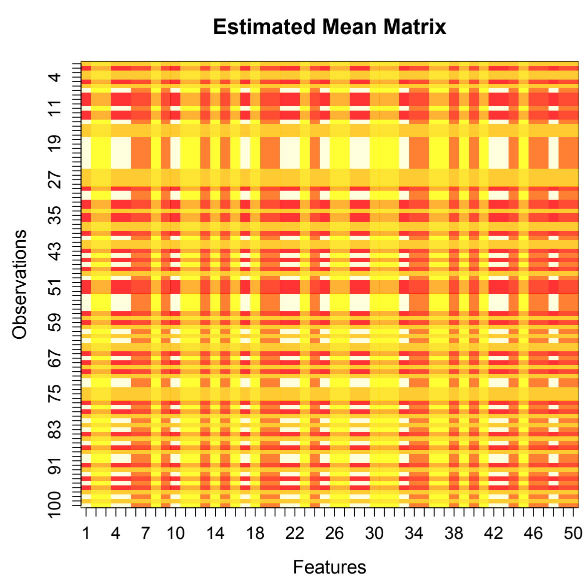

Figure 3: $\lambda = 0$

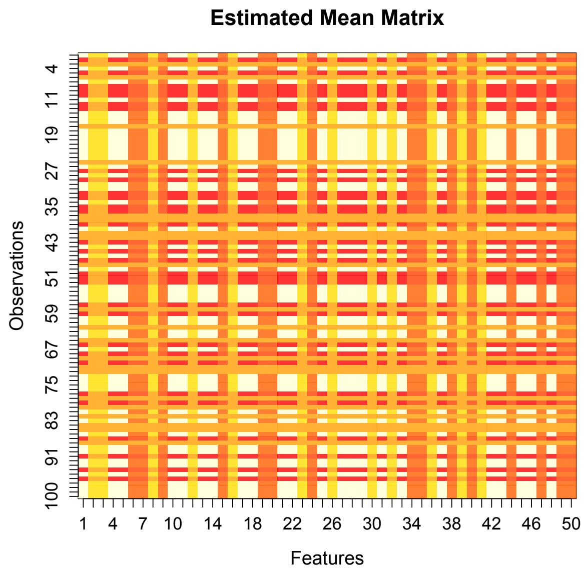

Figure 4: $\lambda = 200$

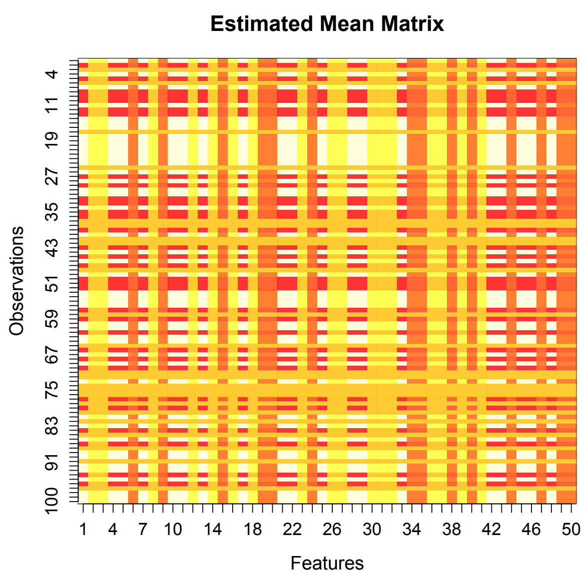

Figure 5: λ selected