Binary tensor clustering

Summary

Jiaxin Hu   Zhuoyan Xu

Date: 2019.3.18

1 Nation's data analysis

Take the nation data as a binary tensor Y and consider the model:

$$logit(\mathbb{E}Y) = U = G \times_1 A \times_2 B \times_3 C$$

In this section, we use *glm* upgrading G, A, B, C to find the best set of them which let the model achieves the maximum of log-likelihood. The initial set of G, A, B, C is given from *tucker*, which is also used to orthogonalize the factor matrices in each step upgrading step.

The maximum log-likelihood the algorithm can find is related to the rank used for *tucker*. Higher the rank, higher the log-likelihood. Here gives a bunch of converged maximum log-likelihood with different choice of the rank:

k =(2,2,2)	k =(2,2,3)	k =(3,3,3)	k =(3,3,4)
-3345.471	-3085.918	-2953.060	-2700.114
k=(4,4,4)	k=(4,4,5)	k=(5,5,5)	k=(5,5,6)
-2579.720	-2297.204	-2128.060	-2039.276

Table 1: Log-Likelihood convergence results choosing different ranks of tucker; The algorithm stops when the relative increase of likelihood is lower than 0.005 or finishes 50 iterations.

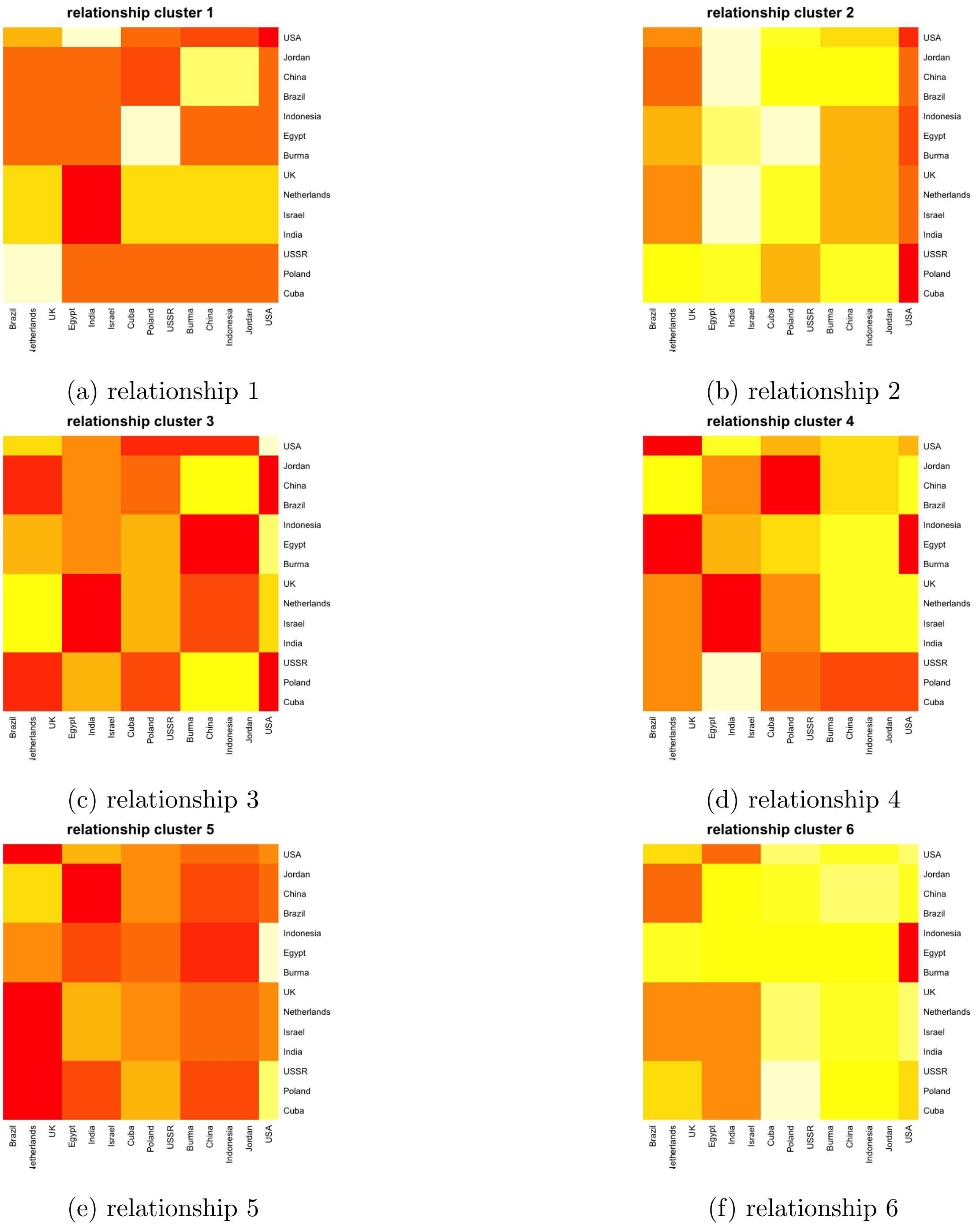
Next, we show the output when $k = (5, 5, 6)$ which gives the largest log-likelihood in our choices. For simplification, we only show the clustering result of the countries in two modes.

```
Mode1 Cluster 1: Cuba Poland USSR
Mode1 Cluster 2: India Israel Netherlands UK
Mode1 Cluster 3: Burma Egypt Indonesia
Mode1 Cluster 4: Brazil China Jordan
Mode1 Cluster 5: USA
```

```
Mode2 Cluster 1: Brazil Netherlands UK
Mode2 Cluster 2: Egypt India Israel
Mode2 Cluster 3: Cuba Poland USSR
Mode2 Cluster 4: Burma China Indonesia Jordan
Mode2 Cluster 5: USA
```

From the perspective of matrix, for there are 6 cluster of relationship, $Figure1$ plot the mean value of countries cluster in 6 relationship cluster.

From the perspective of tensor, the result of estimated U can be shown in $Figure2$. And through this plot we can see that most relationship are in one group except several relationships.



(a) relationship 1

(b) relationship 2

(c) relationship 3

(d) relationship 4

(e) relationship 5

(f) relationship 6

Figure 1: A clustering of countries in each relationship cluster, after reordering the country as the clustering result.

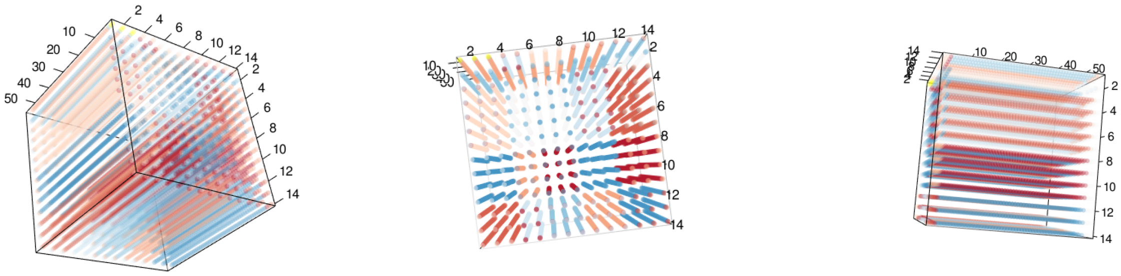


Figure 2: The estimated tensor U from three angle. Particularly, the second plot looks from mode1 and mode2; the third one looks from mode3.

2 Simulations

Then we apply a simulation: Consider

$$U^{20*30*40} = G^{3*4*5} \times_1 A^{20*3} \times_2 B^{30*4} \times_3 C^{40*5}$$

Where $G \sim N(0,1)$, A,B,C are membership matrices where elements are from {0,1}.

And

$$Y \sim Binomial(U)$$

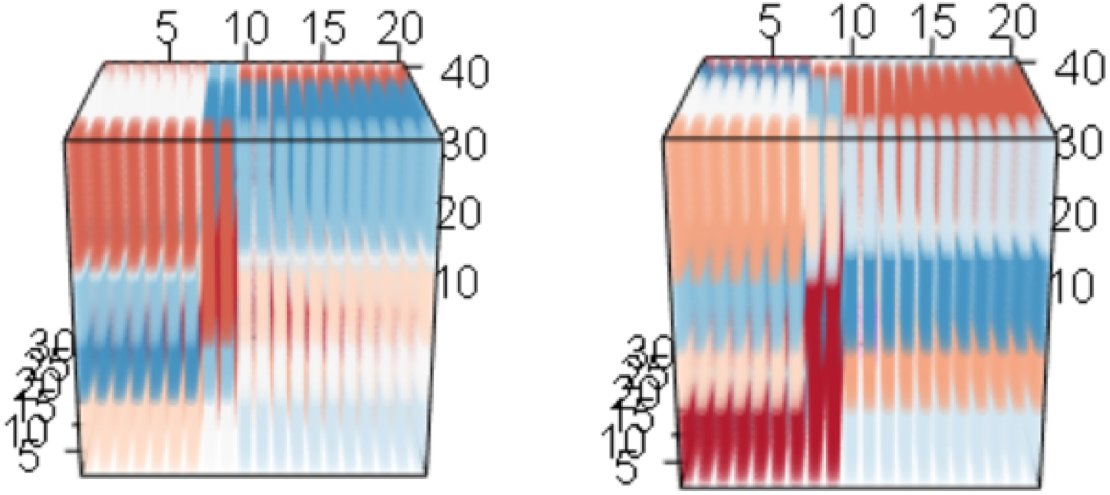
Then we use Y to estimate U, through $\hat{G}, \hat{A}, \hat{A}, \hat{A}$. The $MSE = ||\hat{U} - U||_F^2$ is calculated. This is the MSE though 10 simulations, which is shown below:

1	4364.219
2	4246.980
3	4767.271
4	4600.609
5	4664.368
6	4889.334
7	4418.455
8	4431.251
9	5006.387
10	4512.753

Table 2: MSE through 10 simulations

And, we also plot the estimated $\hat{U}$ and comparing it with ground truth.

As we can see below:



(a) Ground truth

(b) estimated U

Figure 3: The ground truth and estimated U