March 15 tensor clustering

Yuchen Zeng

March 2019

1 New approach to estimate the accuracy of the model

When our true data is not sparse, we use CER to estimate the accuracy of our model. But now, if our true data tensor is sparse, CER would not be the best estimator. Consider the two 2-dimensional cases, $\begin{bmatrix} 0 & * \\ \# & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & 0 \\ * & \# \end{bmatrix}$. First we generate the data (already has partitions on both rows and columns: 4 biclusters) and then let the mean of some biclusters equal to 0. Now some of the data may come from different biclusters but they have the same mean: 0. Look at the first case, if our model is accurate enough, we would get both row CER and column CER equal to 1. But in the second case, it is pretty reasonable for a accurate model to conclude that there only exist 3 biclusters. However, in this case, CER would not closed to 1. So, here instead we use error rate (actually I think it is a monotone decreasing function of lambda), correct zero rate and wrong zero rate which are defined as follows:

$$\begin{aligned} error\ rate &= \frac{||\hat{X} - X||_F}{||X||_F} \\ correct\ zero\ rate &= \frac{|Sparse(X) \cap Sparse(\hat{X})|}{|Sparse(X)|} \\ wrong\ zero\ rate &= \frac{|Sparse(X)^c \cap Sparse(\hat{X})|}{|Sparse(X)^c|} \end{aligned}$$

Notation: Sparse(X) = {the 0 elements in tensor X}.

And using the true k,r,l, we obtained $\hat{X}$ using different λ and got the simulation table as table 1.

2 the problem in choosing λ by using BIC

The BIC is always the smallest only when $\lambda = 0$ (may resulting from some errors in my code).

λ	error rate	correct zero rate	wrong zero rate
0.1	0.053(0.041)	0.384(0.211)	0.042(0.063)
0.2	0.052(0.040)	0.782(0.170)	0.092(0.060)
0.3	0.033(0.031)	0.893(0.096)	0.094(0.064)
0.4	0.044(0.041)	0.953(0.352)	0.115(0.067)

Table 1: n=30,p=30,q=30,k=5,r=3,l=6,sparse rate=0.3

3 Verification of the method of choosing k,r,l

There are several simulation results of choosing k,r,l. I think it still need to be improved, but I cannot come up with a better idea to estimating k,r,l.

Simulation 1

```
out <- sim.choosekrl(20,40,30,2,4,3)

[[1]]
     [,1] [,2] [,3]
[1,]    2    4    3

[[2]]
     [,1] [,2] [,3]
[1,]    2    4    3

[[3]]
     [,1] [,2] [,3]
[1,]    2    3    3

[[4]]
     [,1] [,2] [,3]
[1,]    2    4    3

[[5]]
     [,1] [,2] [,3]
[1,]    2    3    3
```

Simulation 2

```
out <- sim.choosekrl(30,30,30,3,3,3)

[[1]]
     [,1] [,2] [,3]
[1,]    3    3    3

[[2]]
     [,1] [,2] [,3]
[1,]    3    3    3

[[3]]
     [,1] [,2] [,3]
[1,]    3    3    3

[[4]]
     [,1] [,2] [,3]
[1,]    3    3    3

[[5]]
     [,1] [,2] [,3]
[1,]    3    3    3
```

Simulation 3

```
out <- sim.choosekrl(20,20,20,2,2,4)

[[1]]
     [,1] [,2] [,3]
[1,]    2    2    3

[[2]]
```

```
     [,1] [,2] [,3]
[1,]    2    2    3

[[3]]
     [,1] [,2] [,3]
[1,]    2    2    3

[[4]]
     [,1] [,2] [,3]
[1,]    2    2    3

[[5]]
     [,1] [,2] [,3]
[1,]    2    2    3
```

Simulation 4

```
out <- sim.choosekrl(20,20,20,4,3,4)

[[1]]
     [,1] [,2] [,3]
[1,]    3    3    4

[[2]]
     [,1] [,2] [,3]
[1,]    3    3    4

[[3]]
     [,1] [,2] [,3]
[1,]    4    2    4

[[4]]
     [,1] [,2] [,3]
[1,]    4    3    3
[2,]    3    3    4

[[5]]
     [,1] [,2] [,3]
[1,]    4    4    3
[2,]    4    3    4
```