Homework 3 16.11.4

Chuan Lu, 13300180056, chuanlu13@fudan.edu.cn

Problem 1. Perform PCA to assign31.csv.

Result. Since there are many NAs in assign31.csv, we changed those NA to 0.

The results are as follows: The rate in the table is the rate of information reserved after PCA.

k	1	2	3	5	10	14
Rate	0.2139058	0.396026	0.5165856	0.6697862	0.920202	1

Table 1: Reserved information rate of different orders in PCA.

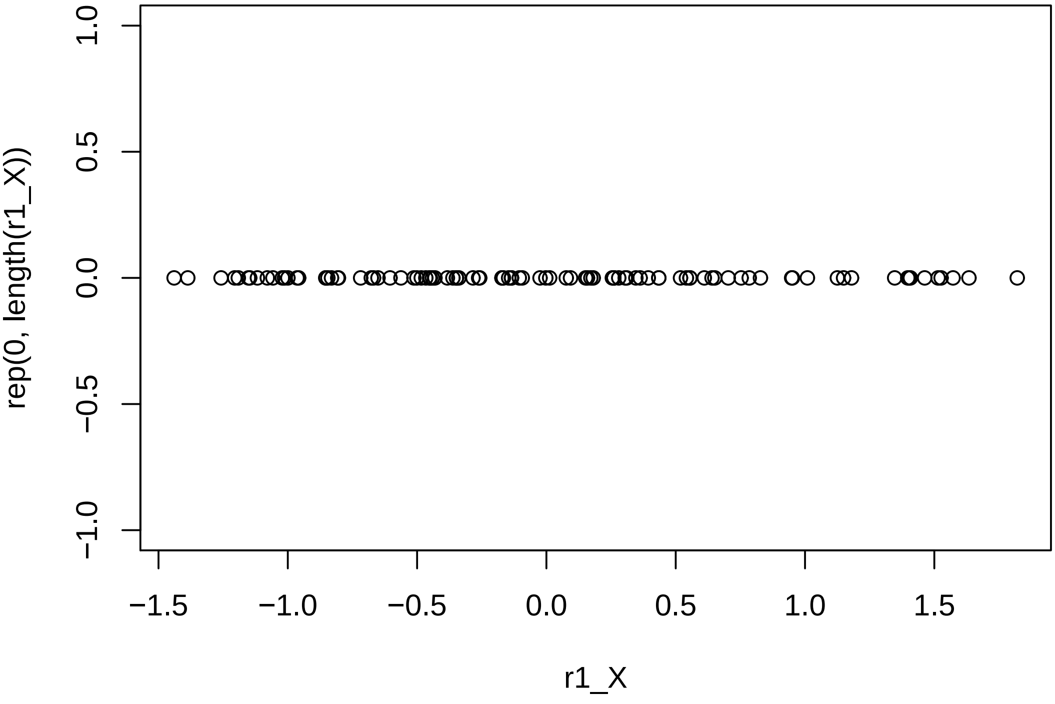


Figure 1: PCA, k = 1

From these figures(when $k = 1, 2, 3$) we can find that after PCA the data points still seem to be random and messy. When we calculate the rate of reserved information we can know that not until $k = 10$ the rate is less than 0.9, which is thought to be unsatisfying.

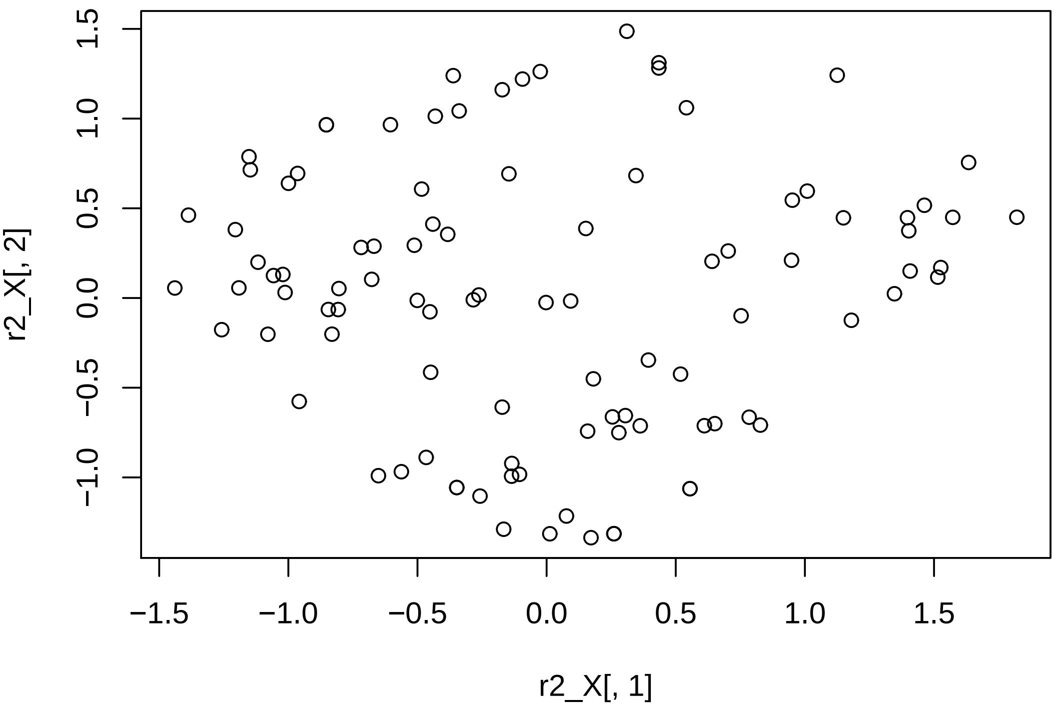


Figure 2: PCA, k = 2

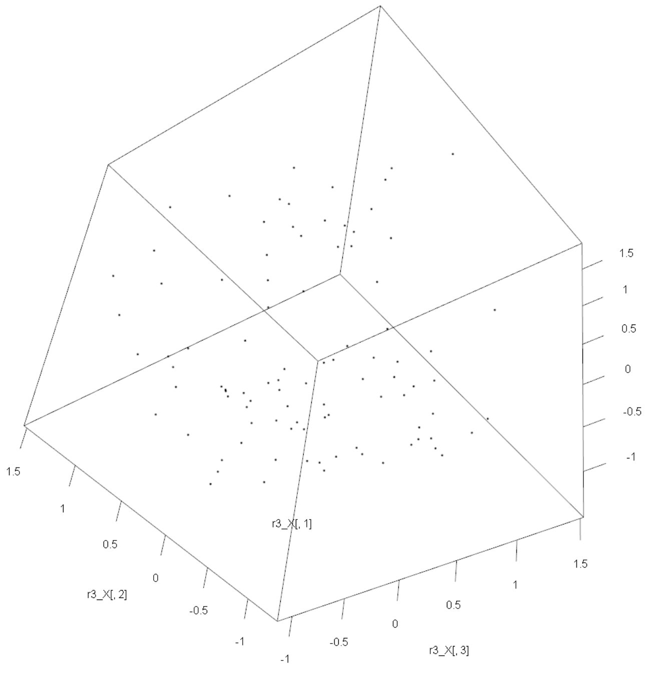


Figure 3: PCA, k = 3

Problem 2. Perform PCA and NMF to assign2.csv, and explain the difference.

Result. For PCA, the result is as follows: From figure(k = 1) we can assert nothing; from figure(k = 2) it seems the data can be simulated by two parallel lines. From the table we can find when k = 3, the rate of reserved information is over 0.9, which means k = 3 is a good estimation.

k	1	2	3
Rate	0.4326317	0.7648243	0.9298461

Table 2: Reserved information rate of different orders in PCA.

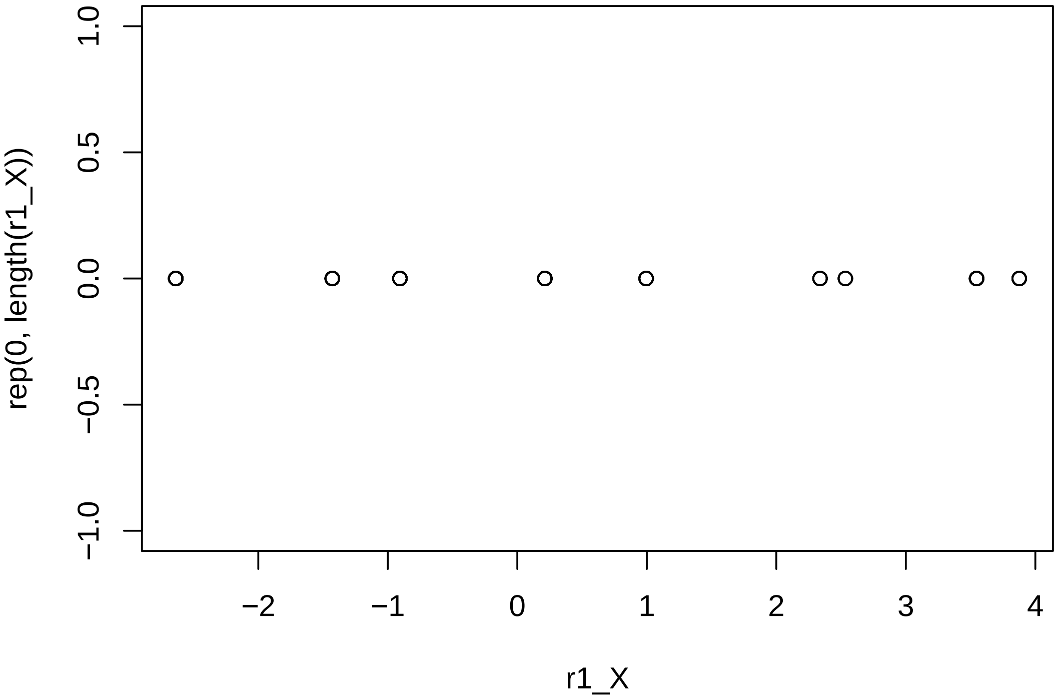


Figure 4: PCA, k = 1

For NMF, the result of error rate is as follows:

$$rate = \frac{\|A - WH\|_F}{\|A\|_F}$$

It must be wrong somewhere that with k being larger, the estimation becoming worse. I

k	1	2	3	5	10
Rate	0.6764784	0.6904373	2.629497	4.993591	14.36354

Table 3: Error rates of different orders in NMF.

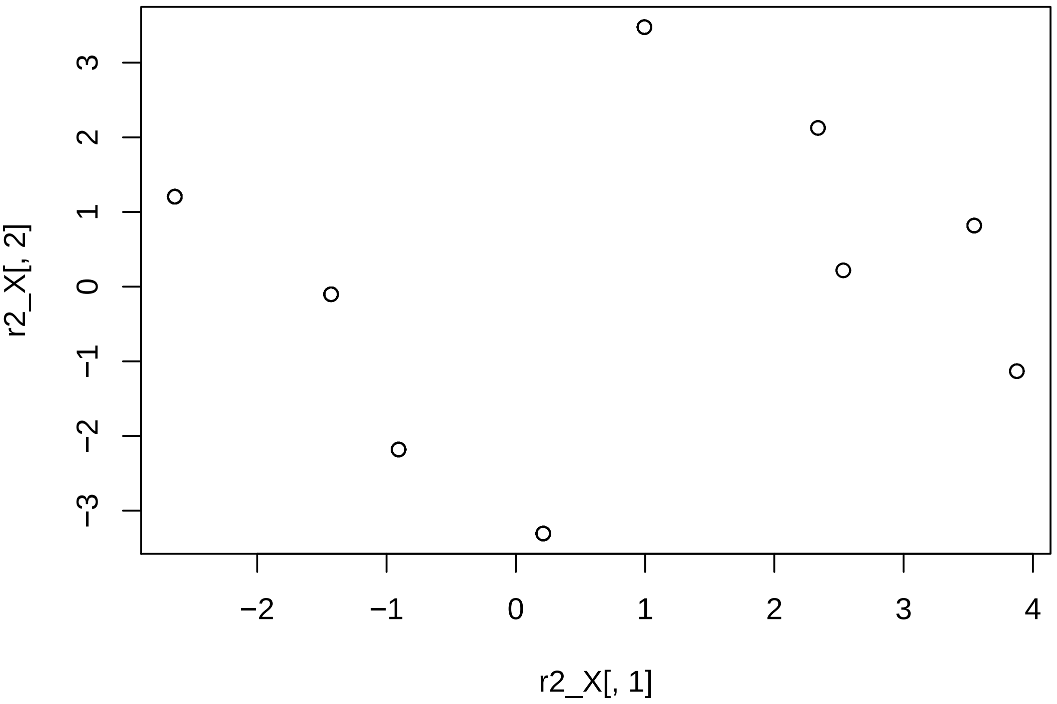


Figure 5: PCA, k = 2

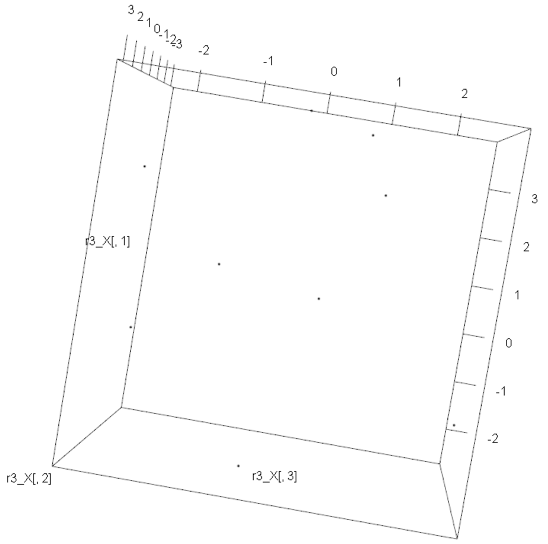


Figure 6: PCA, k = 3

think it should be the number of iterations, which is fixed at 1000.

Problem 3. Perform PCA and nmf to test2.csv

Result. For PCA, the result is shown as follows:

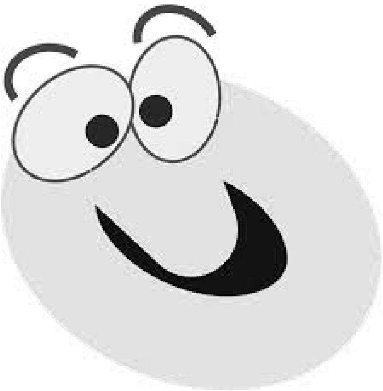
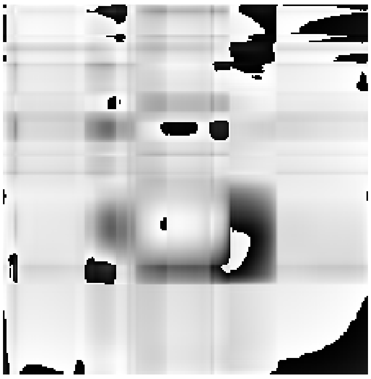


Figure 7: Original Picture

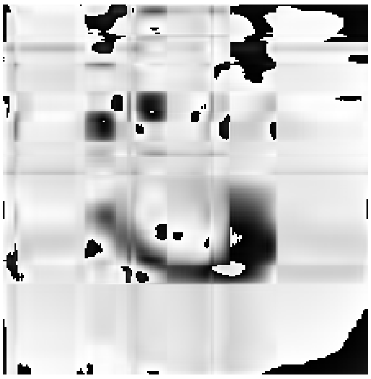
(a) k = 1

(b) k = 2

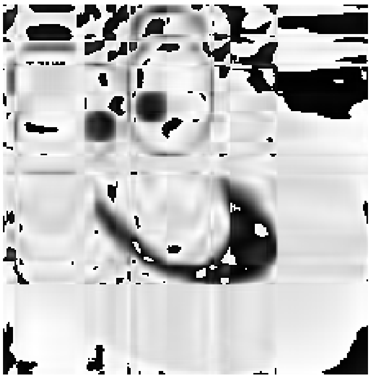


(c) k = 3

(a) k = 4

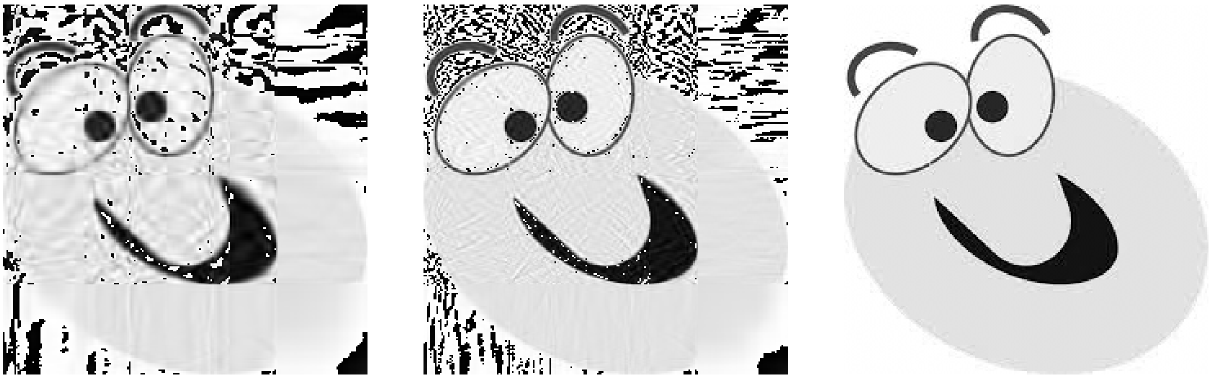


(b) k = 5



(c) k = 10

For NMF,



(a) k = 20

(b) k = 50

(c) k = 200