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, I 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.

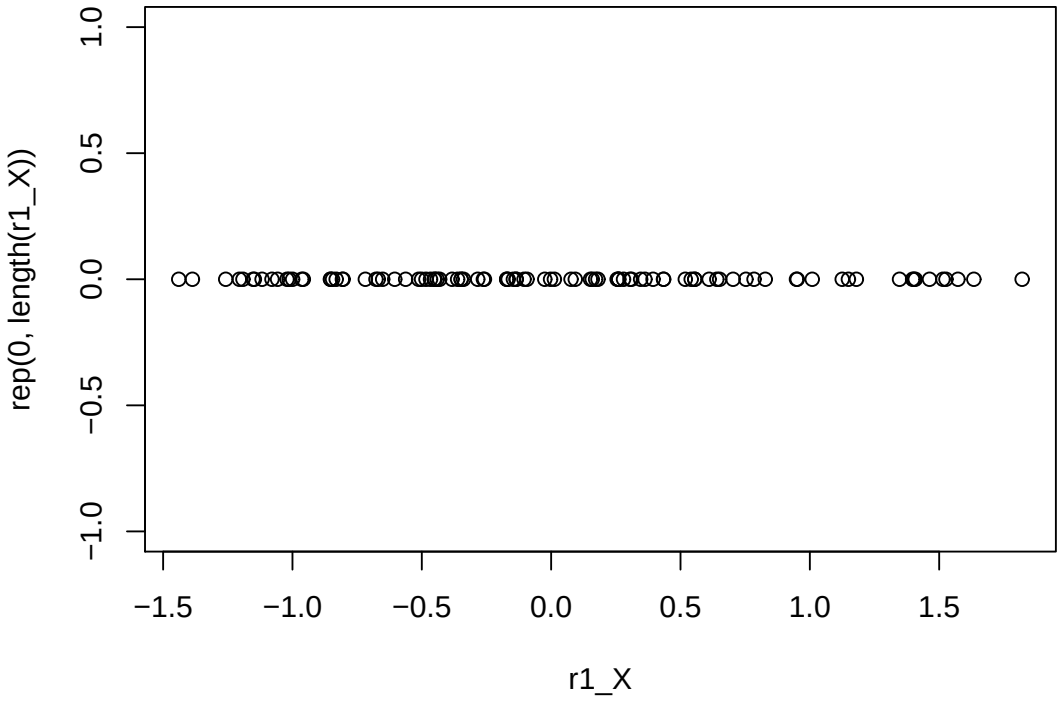


Figure 1: PCA, k = 1

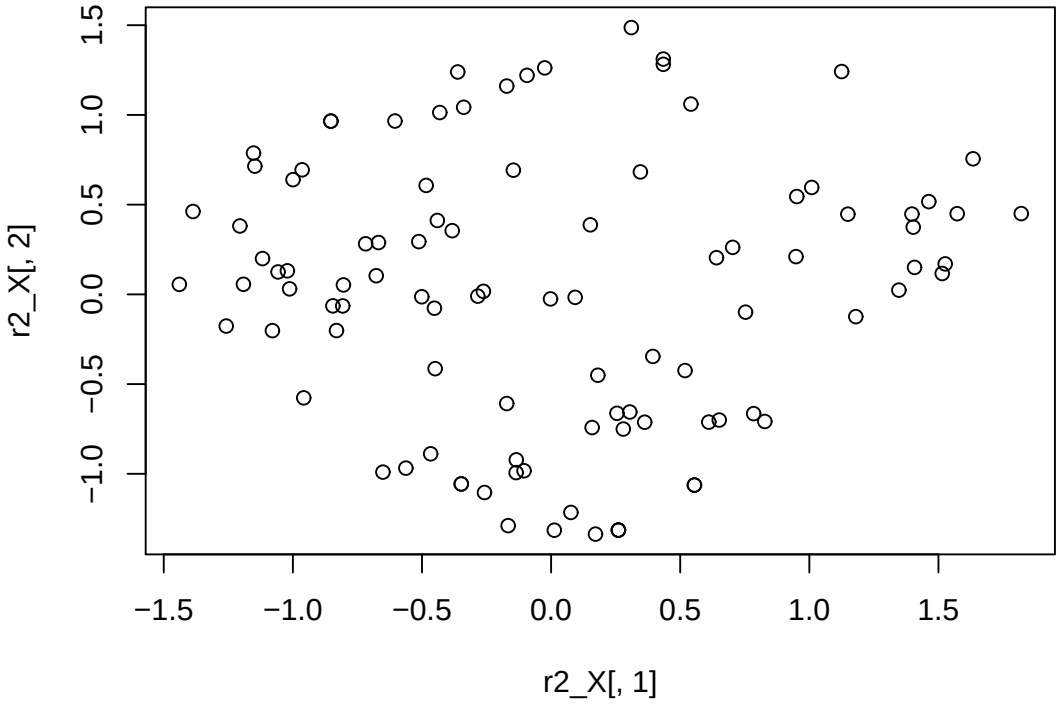


Figure 2: PCA, k = 2

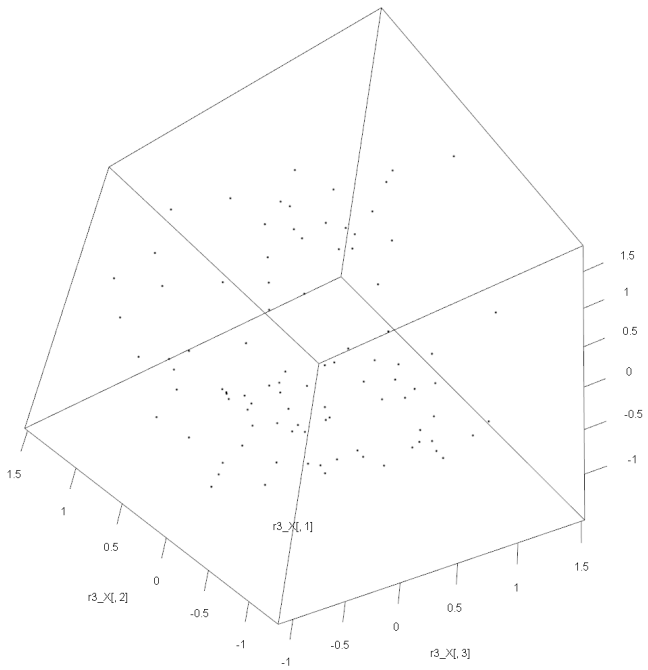


Figure 3: PCA, k = 3

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.

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

Result. For PCA, the result is as follows:

k	1	2	3
Rate	0.4326317	0.7648243	0.9298461

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

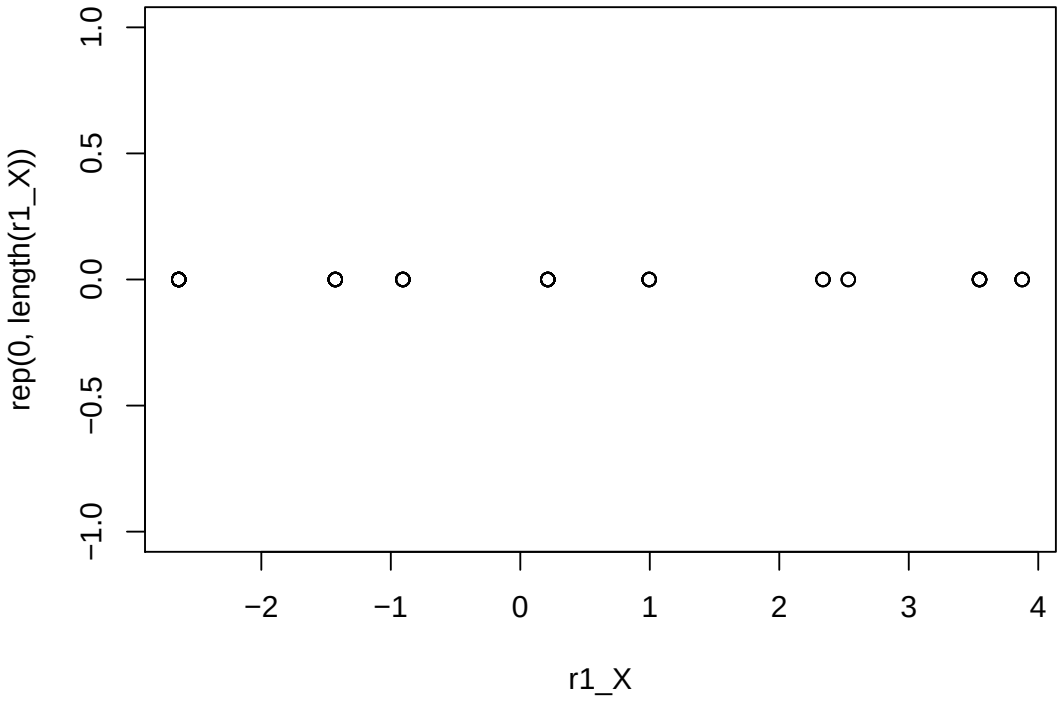


Figure 4: PCA, k = 1

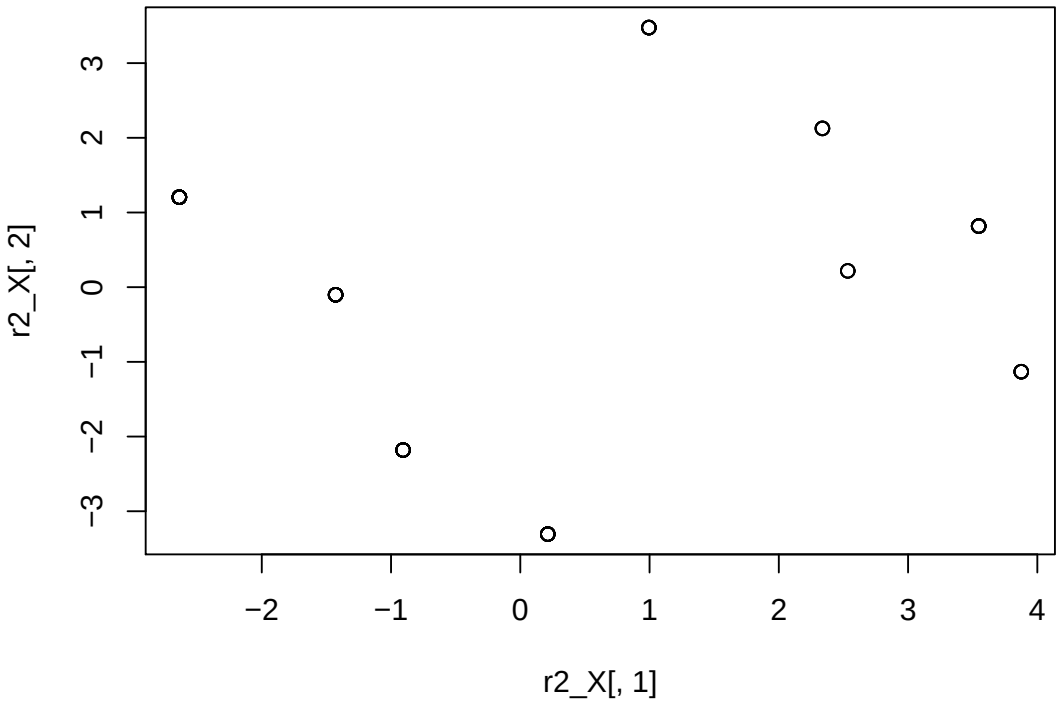


Figure 5: PCA, k = 2

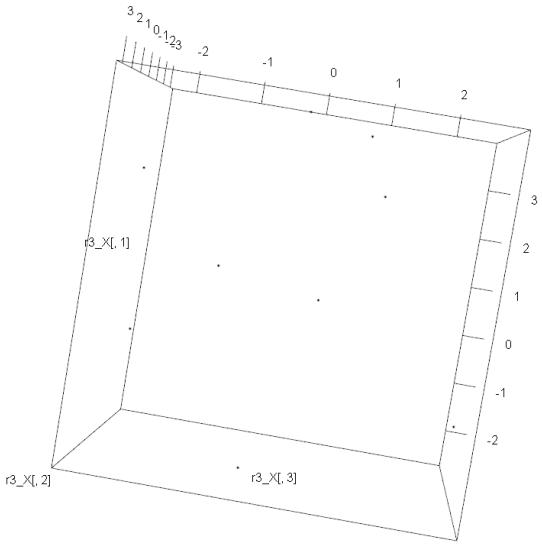


Figure 6: PCA, k = 3

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.

For NMF, the result is as follows:

Firstly, we use the Divergence objective function

$$min \quad \Sigma_{i,j}(A_{ij}log(\frac{A_{ij}}{WH_{ij}}) - A_{ij} + WH_{ij})$$

which assumes the elements in A obey a Poisson distribution. The error rate is defined as follows for both Divergence objective function and Mean squared obj. function.

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

The basis of NMF(divergence obj func) are as follows:

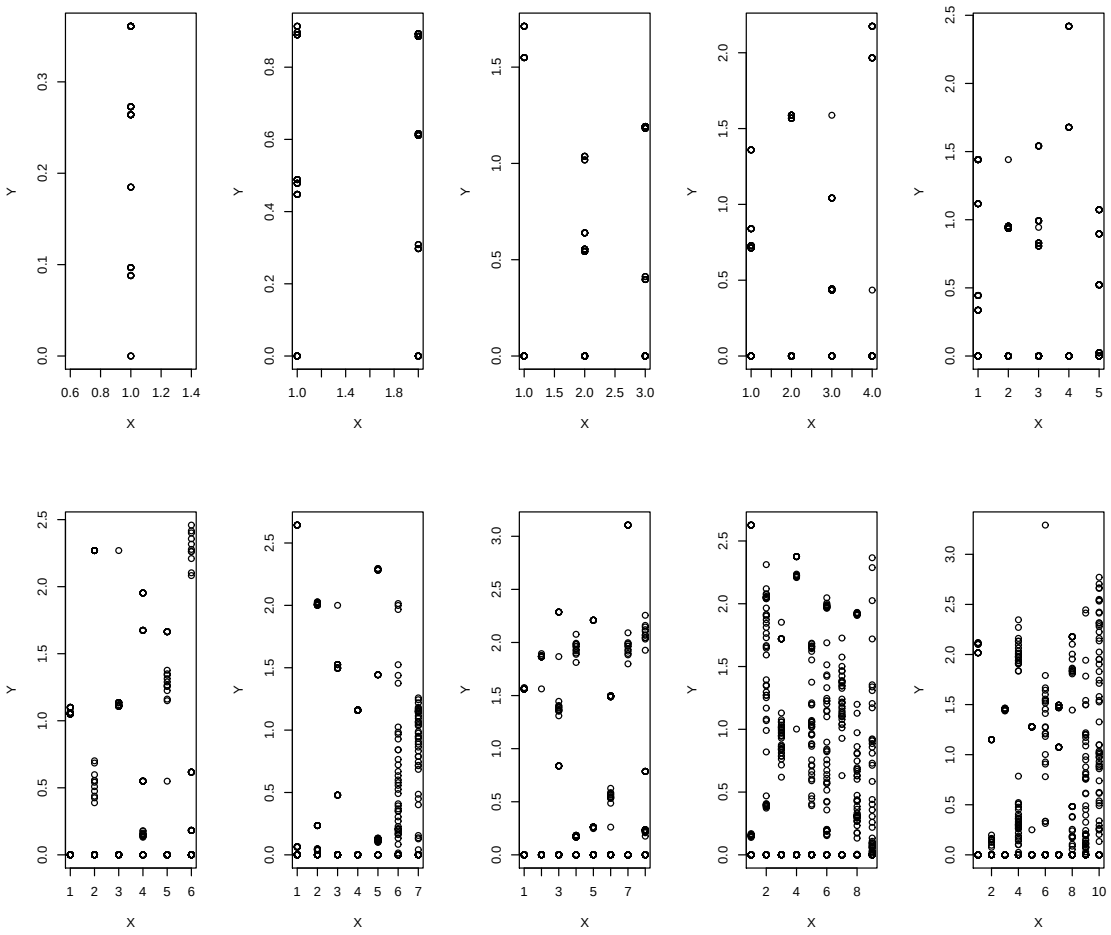


Figure 7: NMF, D, basis, n=1:10

And the error rate is as follows:

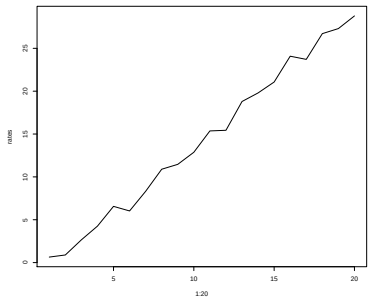


Figure 8: NMF, error rate, n=1:20

From the curve of error rates we can know, that in this case the assumption that entries obeys Poisson distribution is not acceptable. So I tried Mean squared error objective function instead. In the meantime, MSE converged much quicker than D.

When using Mean squared error objective function

$$min \quad \|A - WH\|_F^2$$

The basis is as follows:

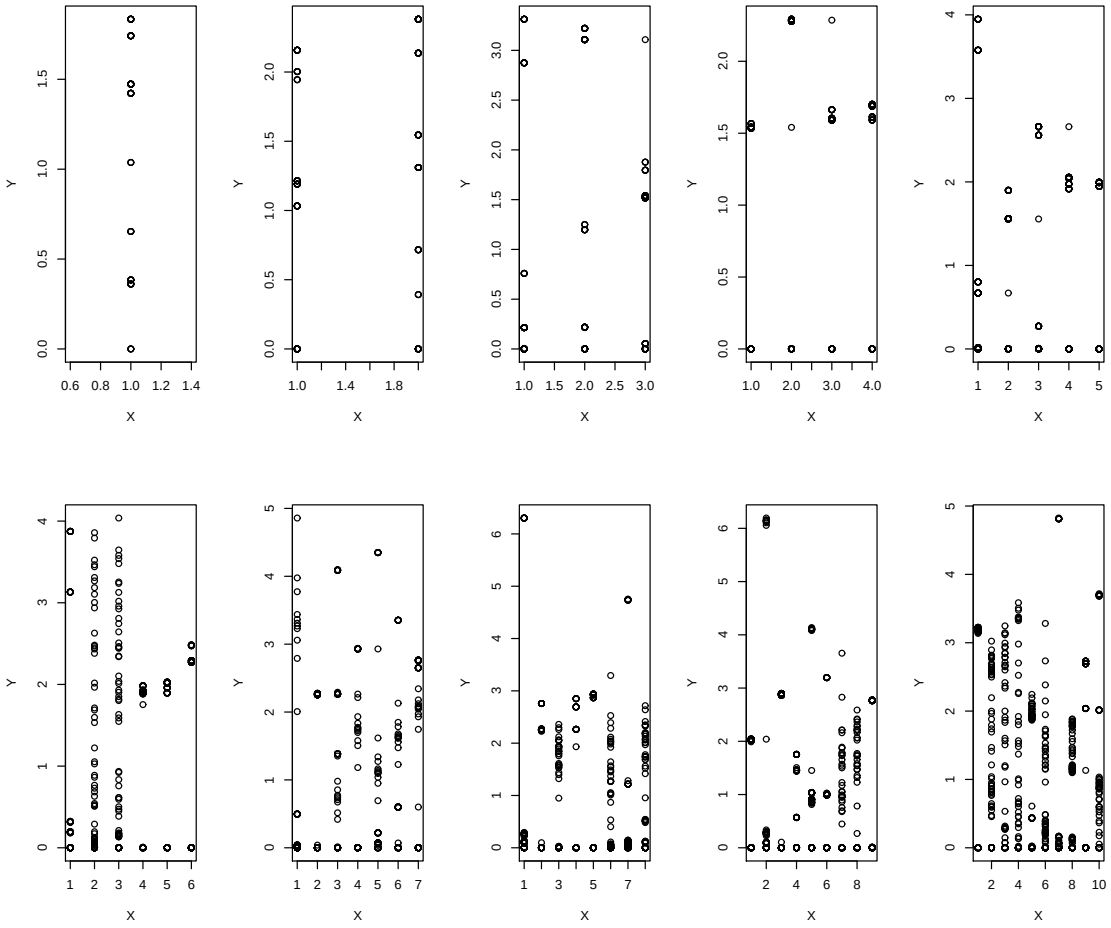


Figure 9: NMF, M, basis, n=1:10

And the error rate is as follows:

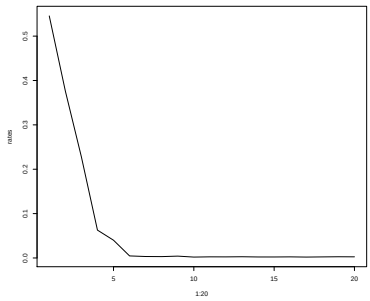


Figure 10: NMF, M, basis, n=1:20

One can tell from the error curve, that when we choose $k = 6$, the result is good enough to represent the data from A.

Problem 3. Perform PCA and NMF to test2.csv

Analysis. When we do PCA to an image, we assume each row being a data point. Then PCA can be presented as

$$Y = XP$$

where P is the matrix constructed by eigenvectors of $X^\top X$. Then

$$X^\top X = P\Sigma P^\top$$

where Σ is a diagonal matrix. So choosing the first k principle conponents is equivalent to choosing the first k columns of P. We denote it as P_1.

If we let $A = XP_1P_1^\top$, then

$$A^\top A = P_1P_1^\top X^\top XP_1P_1^\top = P_1P_1^\top PDP^\top P_1P_1^\top$$
$$= P_1D_{1:k}P_1^\top$$

So

$$\hat{X} = XP_1P_1^\top$$

is a approximation of X, and it is equivalent to restoring X with its first k singular values.

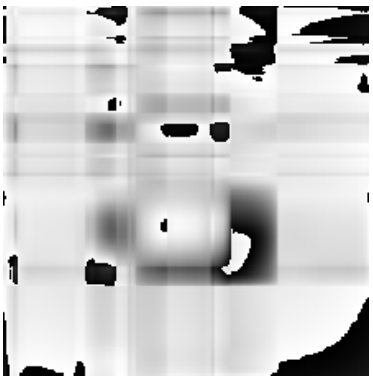
Result. For PCA, the result is shown as follows:

(I used Python-OpenCV to draw these images.)

Figure 11: Original Picture

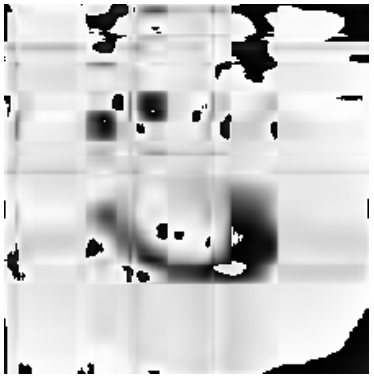
(a) k = 1

(b) k = 2



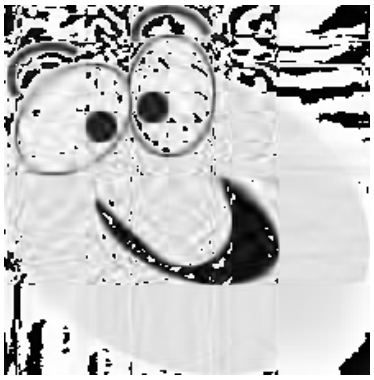
(c) k = 3

(d) k = 4

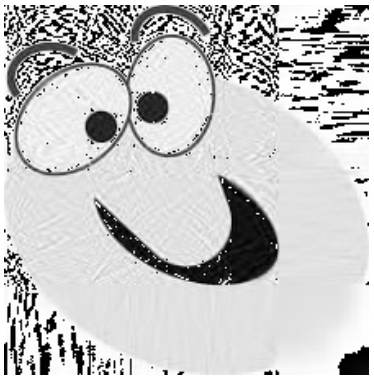


(e) k = 5

(f) k = 10



(g) k = 20



(h) k = 50

(i) k = 200

For NMF, when we use Mean squared error obj func, the results are as follows:

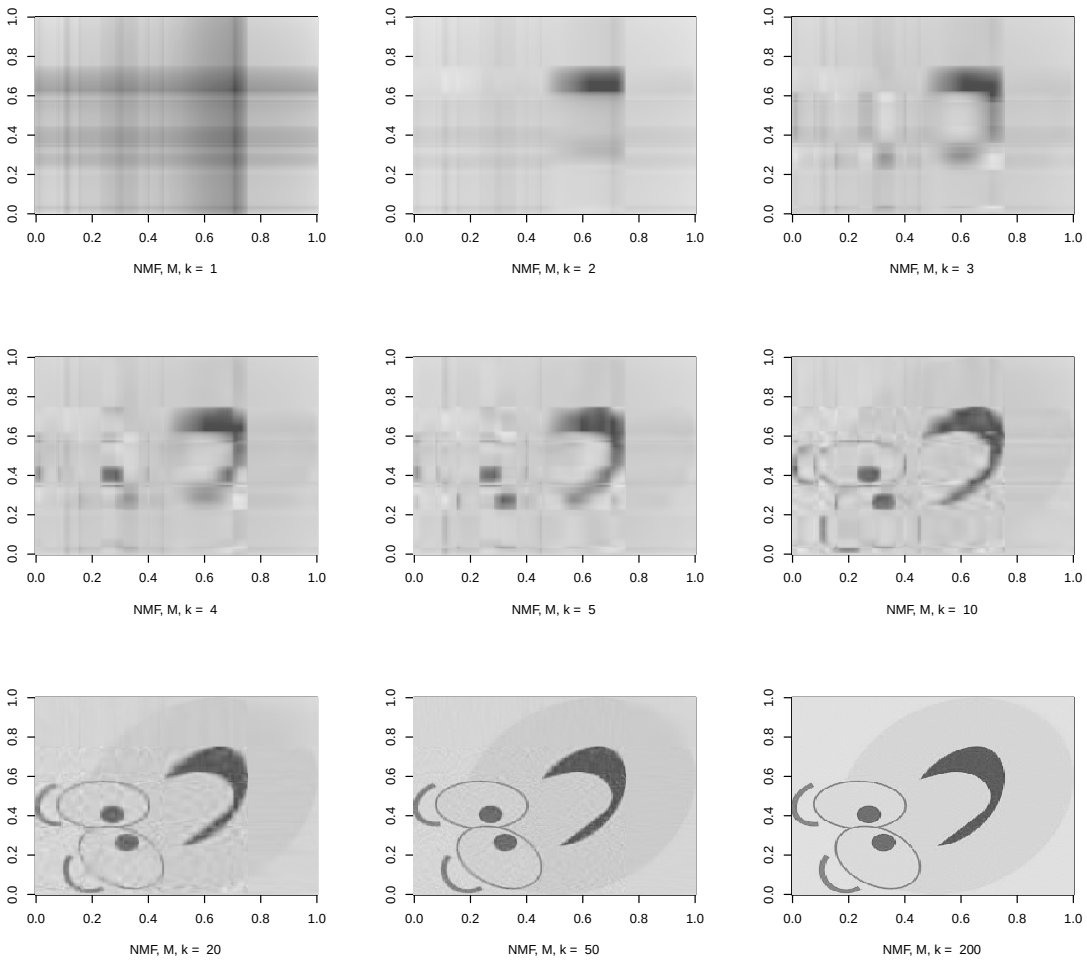


Figure 12: NMF, using MSE

When we use Divergence squared error obj func, the results are as follows:

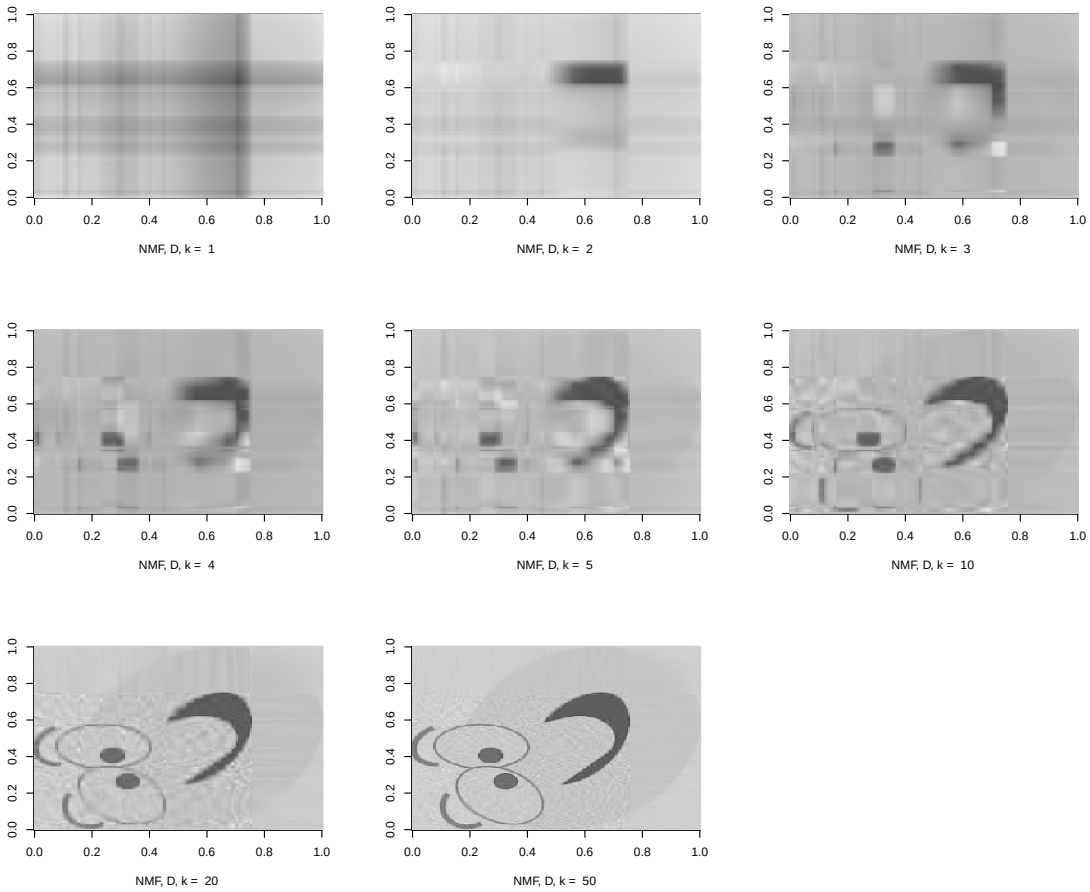


Figure 13: NMF, using Divergence

One can tell from the images that using mean squared error is almost the same as using divergence obj func, and their results are better than using PCA. I think it is because entries in a image are nonnegative, so if we use NMF, then the new basis are nonnegative which could be more appropriate. However, since PCA can be performed easily using SVD, the speed of PCA is hundreds faster than NMF.