CEE 690 Health and Environmental Data Science, Spring 2019
Fang Feng
Homework 3

Answer to Problem 1

(1)

$$SSE = \sum_{i=1}^{n} min||X_i - C_i||_2^2$$

taking derivative:

$$\frac{\partial SSE}{\partial C_i} = -\sum_{j=1}^{k}\sum_{i=1}^{n}(2C_j - 2X_i) = 0$$

so:

$$\sum_{j=1}^{k}\sum_{i=1}^{n}(-X_i + C_j) = 0$$

then

$$\sum_{j=1}^{k}[nC_j - (X_1 + X_2 + ... + X_n)] = 0$$

where:

$$C_j = \frac{\sum_{i=1}^{n} X_i}{n}$$

which Update the cluster means

(2)

$$\sum_{i=1}^{n} min||X_i - C_j||_2^2$$

so suppose assignment of two clusters satisfy:

$$||X_i - C_1||_2^2 \geq ||X_i - C_2||_2^2$$

so

$$\sum_{i=1}^{n} min||X_i - C_2||_2^2 \leq \sum_{i=1}^{n} min||X_i - C_1||_2^2$$

which means the SSE decreases as we assign point to a closer center.

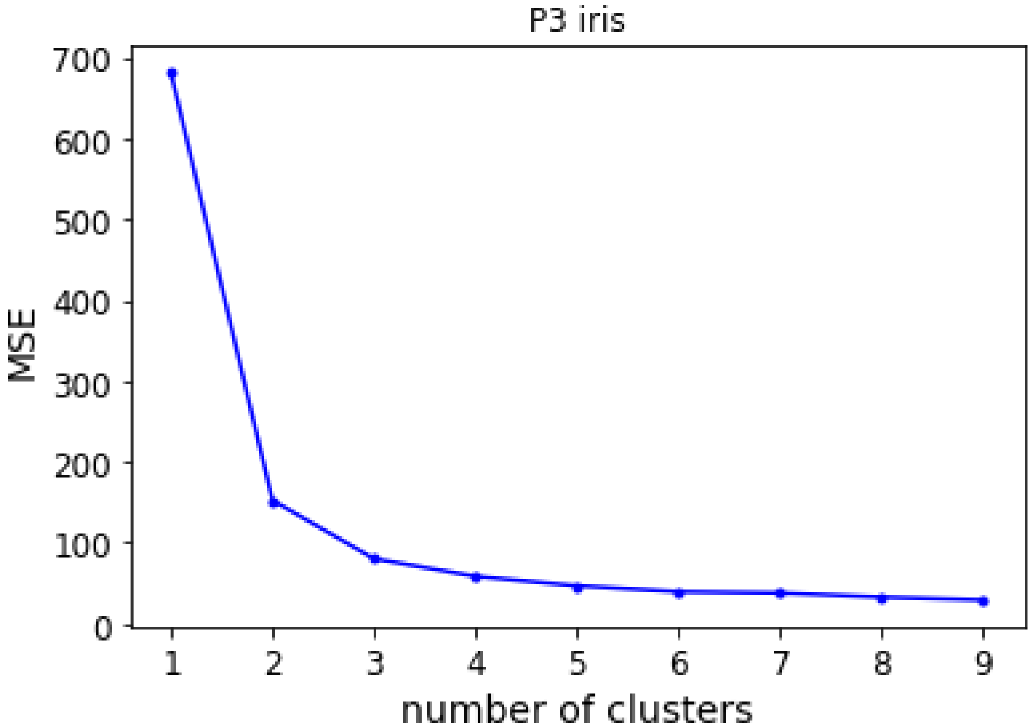
Since there is a minimum value for the SSE (its at least 0), and each step decreases the SSE, we can conclude that the kmeans algorithm converge to that when you set the number of clusters to be the number of points, in that case the SSE will be 0.

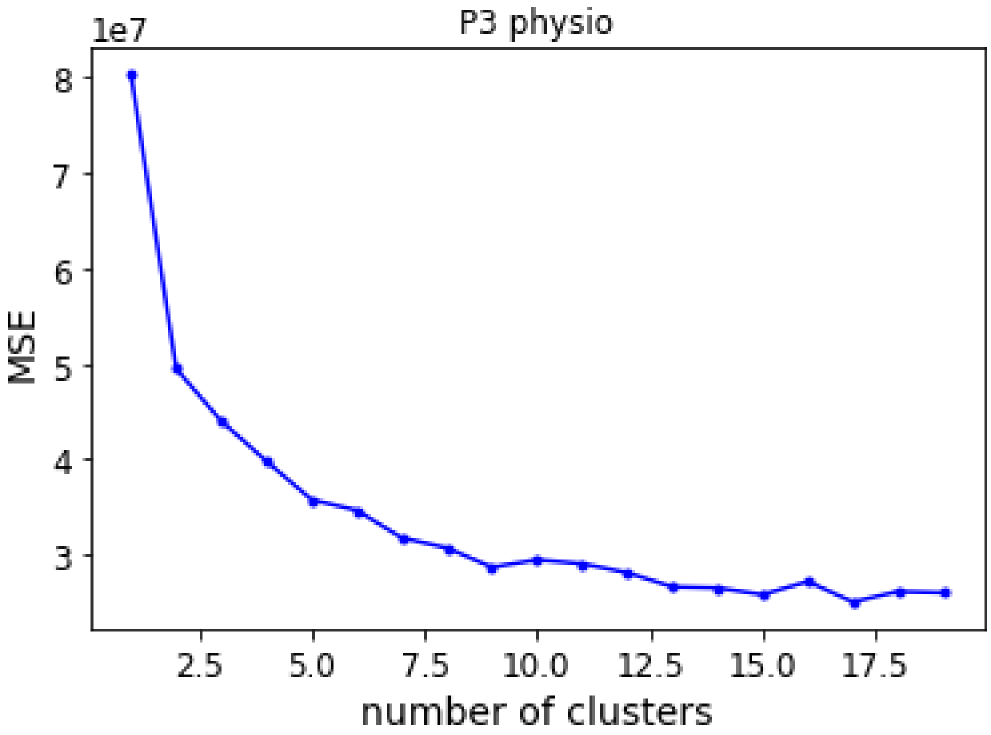
Answer to Problem 2

(a) the hold-out MSE is 1.08470454253
(b) it is a reasonable value
(c) the hold-out AUC = 0.63
(d) the reason which I didn't get reach the expected value is because I didn't do the transform after get the best K(number of features). Instead I pipe the feature_selection with KNN together. In that case I didn't get the best number of features (k=2) and best number of neighbors (n=16). Plus, I use the predict method insteand of predict_proba method, which is not accurate for prediction in the problem.

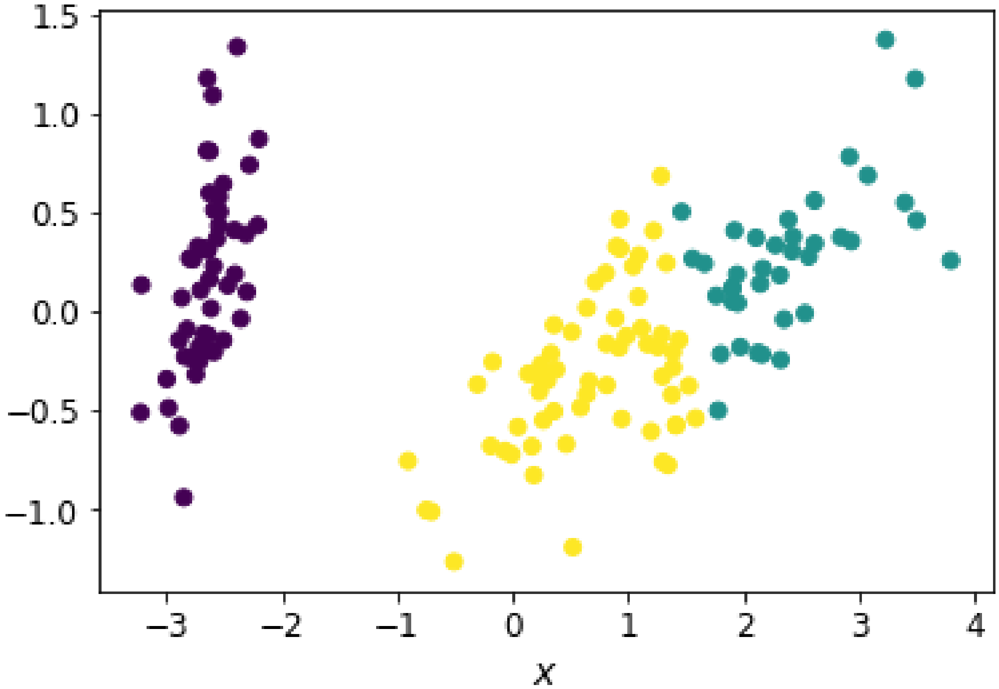
Answer to Problem 3

(a)

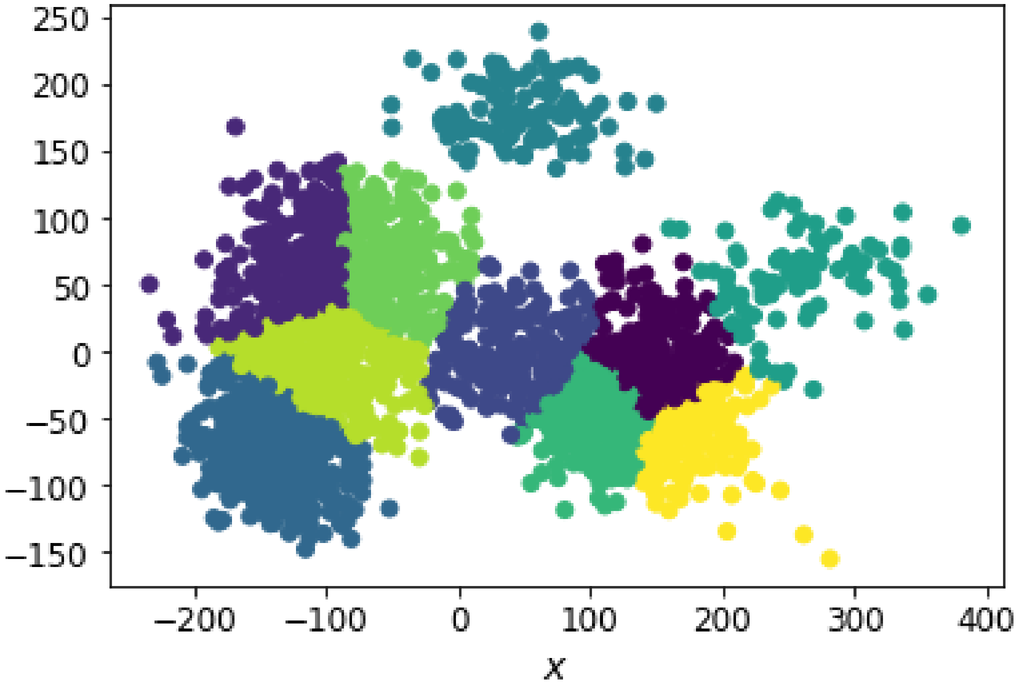




(b)
so for iris data, we can see that the MSE dramatically decreases when k = 3, so we choose k = 3

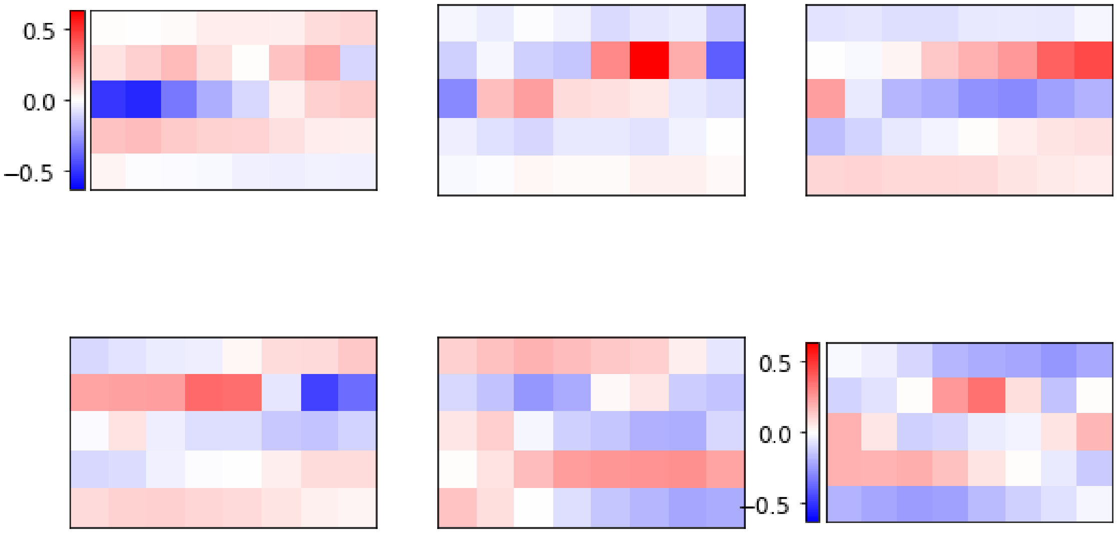


and for physio data, we can see that the MSE dramatically decreases when k = 10, so we choose k = 10

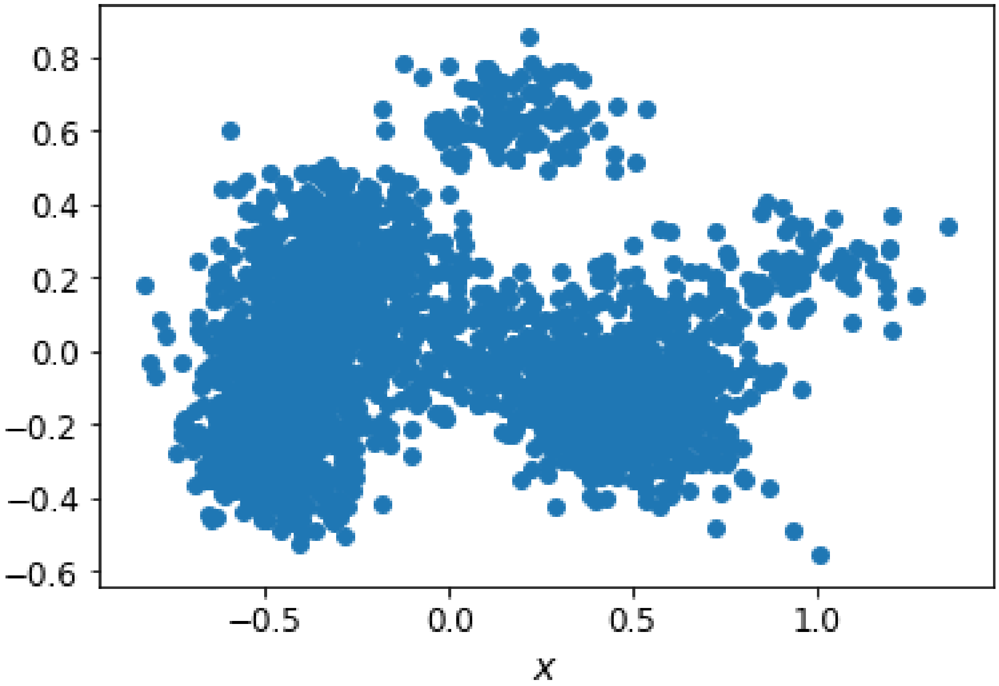


Answer to Problem 4

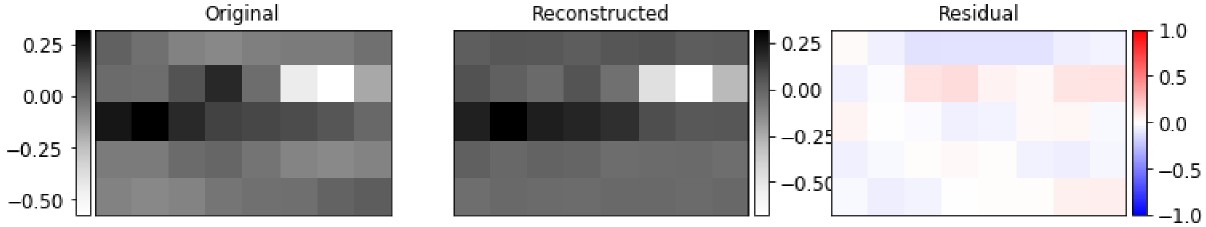
(a)
visualization of the components:



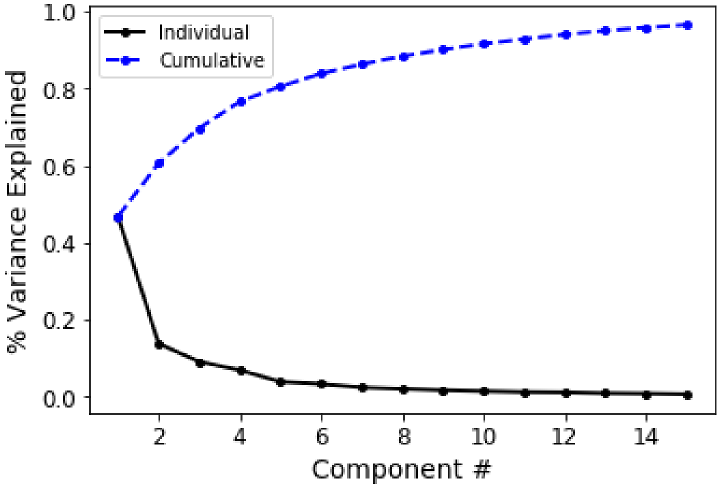
(b)
visualization after embedding it in the top 2 principle components:



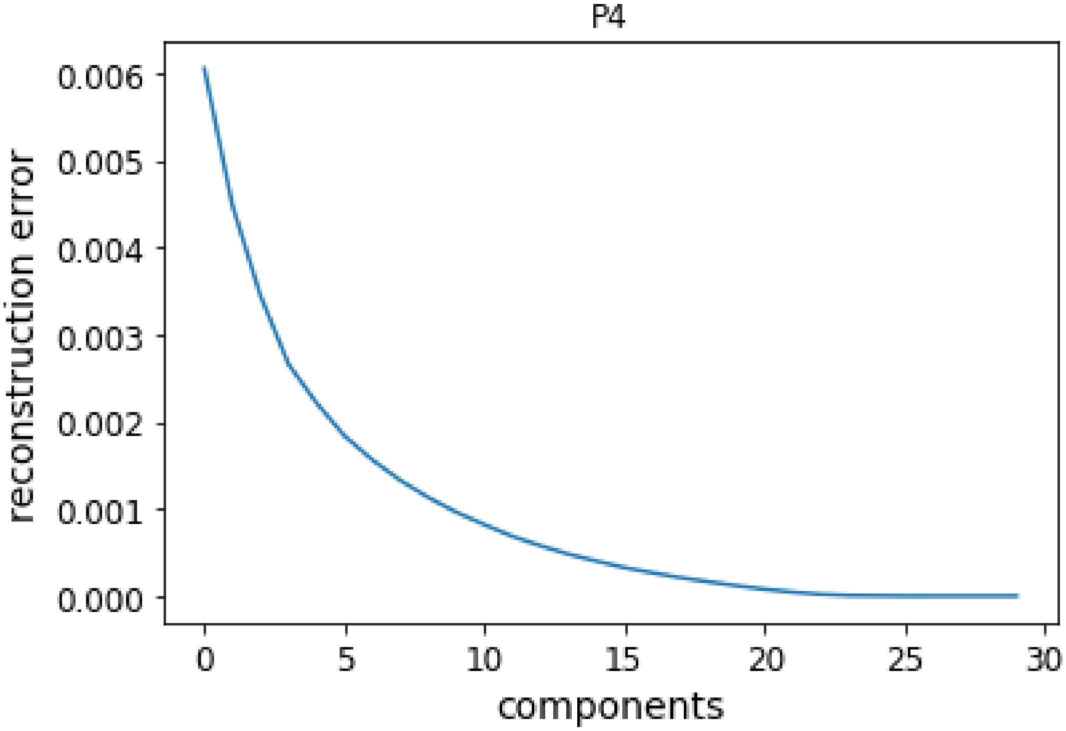
for one single image (one row of the dataset), the reconstruction plot is:



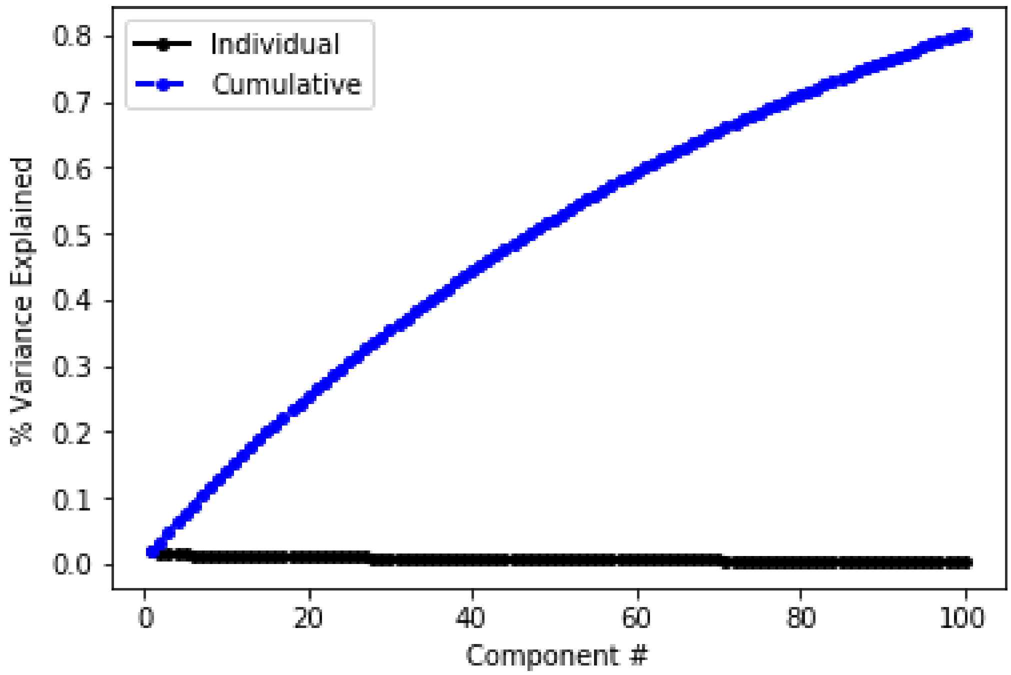
and the variance plot:



from the accumulated variance plot we can see, to maintain 80% variance, we are going to pick the number of components to be 6.
we can get the reconstruction errors as a function of the number of components:



Answer to Problem 5



(i)To choose the number of components, from the variance plot, we can see that to maintain 80% variance, we would choose the number of components toe be 70.

(ii) using the model_selection on the number of components, gamma, and C, we can get the the best result to be 0.75 scored by roc_auc, with settins as: n_components = 10, gamma = 0.0001, C=2.21