Georgia Institute of Technology

CS 7641: Machine Learning

Assignment 3 – Unsupervised Learning and Dimensionality Reduction

Implement two clustering algorithms and four dimensionality reduction algorithms

Name: Tianyu Yang
GTid: 903645962

Date: 2020/10/31

1. Abstract

In this unsupervised learning and dimensionality reduction project, two clustering algorithms (k-means clustering and expectation maximization) and four dimensionality reduction algorithms (PCA, ICA, Randomized Projections and Factor Analysis) will be implemented. The datasets I used is from the assignment 1 which is the NBA games and LOL games. I will apply the dimensionality reduction algorithms to these datasets and rerun my neural network learner on the newly projected data. Besides, I will apply the clustering algorithms to these datasets and rerun my neural network learner again. Finally, I will evaluate each model's performance and make a comparison.

All codes are on the Github and the link is https://github.com/simonyang0701/CS7641-Machine-Learning-Unsupervised-Learning-and-Dimensionality-Reduction.git.

2. Introduction

In this project, I will cover two clustering algorithms including k-means clustering and expectation maximization. I will evaluate their performance and plot some useful figures. Besides, I will also implement four dimensionality reduction algorithms, including PCA, ICA, Randomized Projections and Factor Analysis. I will also reproduce the clustering experiment and apply it to the same datasets I used in the assignment 1. I will also apply clustering algorithms to the result I mentioned before and make a comparison on the newly projected data.

The datasets I used are from the first assignment which are NBA games and LOL games. For me, as a big fan of NBA games and LOL games, I searched in Kaggle and downloaded NBA games and LOL games matches data in detail. That is because I know that some advanced data will influence the result of each single games. These two datasets are both classification dataset. For the NBA games datasets, it contains 965 rows of data and five features to classify. For the LOL games datasets, it also contains 965 rows of data so that it can be compared with the NBA games in the same count. It has six features. These data are unlabeled datasets.

The next part is to demonstrate the result of six algorithms. I will show how the parameters are selected and analyze the result. The last part is the conclusion part. In this part, I will describe how the data looks like in the new spaces I created through these algorithms and answered the questions from the description of this assignment. Last but not least, I will make a comparison of these algorithms and give my summary of conclusion.

3. Methodology

3.1 K-means clustering

K-means algorithm originated from a vector quantization method in signal processing, and now it is more popular in the field of data mining as a cluster analysis method. The purpose of k-means clustering is to divide n points (which can be an observation or an instance of the sample) into k clusters, so that each point belongs to the nearest mean (this That is, the cluster corresponding to the cluster center) is used as the clustering standard. This problem will boil down to a problem of dividing the data space into Voronoi cells.

3.2 Expectation Maximization

The maximum expectation algorithm is used in statistics to find the maximum likelihood estimation of parameters in a probability model that depends on unobservable hidden variables.

In statistical calculations, the maximum expectation (EM) algorithm is an algorithm for finding the maximum likelihood estimation or maximum a posteriori estimation of parameters in a probability model, where the probability model depends on unobservable hidden variables. The maximum expectation algorithm is often used in the field of data clustering in machine learning and computer vision.

3.3 PCA

PCA (Principal Component Analysis) is a common data analysis method, often used for dimensionality reduction of high-dimensional data and can be used to extract the main feature components of the data.

The mathematical derivation of PCA can be carried out from the two aspects of maximum separability and nearest reconstruction. The optimization condition of the former is that the variance is the largest after division, and the optimization condition of the latter is that the distance from the point to the division plane is the smallest. To prove it.

3.4 ICA

ICA (Independent Component Analysis) was first applied to Blind Source Separation (BBS). Originated from the "cocktail party problem", it is described as follows: In a noisy cocktail party, many people are talking at the same time, and there may be background music, but the human ear can hear each other's words accurately and clearly. This phenomenon in which you can choose the sounds you are interested in from the mixed sounds and ignore other sounds is called the "cocktail party effect."

Independent component analysis is a data-driven signal processing method developed from blind source separation technology. It is an analysis method based on high-order statistical characteristics. It uses statistical principles to perform calculations and separates data or signals into statistically independent linear combinations of non-Gaussian signal sources through linear transformation.

3.5 Randomized Projections

Random projection is a technique used to reduce the dimensionality of a set of points in Euclidean space. Compared with other methods, the random projection method is known for its power, simplicity and low error rate. According to the experimental results, random projection can maintain the distance very well, but there are few empirical results. They have been applied to many natural language tasks under the name of random index.

3.6 Factor Analysis (Additional dimensionality reduction algorithm)

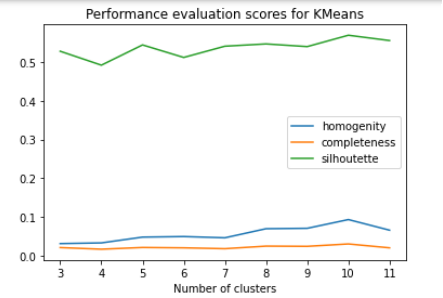
Factor analysis is a statistical method used to describe the observed variability between related variables based on a potentially small number of unobserved variables called factors. For example, it is possible that changes in six observed variables mainly reflect changes in two unobserved variables. The factor analysis search responds to such joint changes in unobserved latent variables. The observed variables are modeled as a linear combination of latent factors plus an "error" term. Factor analysis aims to find independent latent variables.

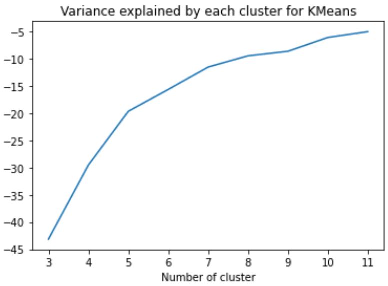
4. Result

4.1 K-means clustering

I used the packages in scikit-learn and cluster_function. After importing and cleaning the datasets, use the train_test_split function to split the datasets into training sets and testing's sets. In order to doing K-means clustering, I preprocessing data standardize the data between 0 and 1. The number of class for the K-means model is 2 for NBA datasets and class of number is 7 for LOL datasets. The component list is from 3 to 11. After that, I plot the figure of performance evaluation scores, variance explained by each cluster. For both datasets, the number of clusters k is selected based on a primarily method: See the output figure to indicate that the improvement of performance and variance with the increasing k. The second method for selecting the value of k is the V metric (entropy-based method) or the information criterion method (the trade-off between goodness and model fit and complexity)

<u>For NBA games datasets:</u>

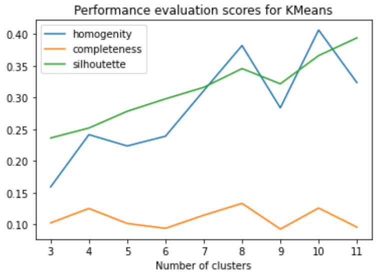


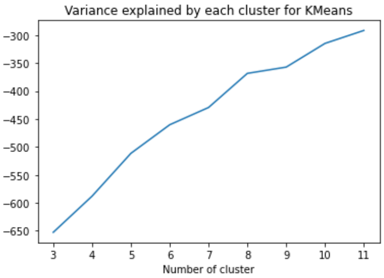


The training accuracy for K-means is K=2:56.60621761658031

The Testing accuracy for K-means is K=2:55.95854922279793

<u>For LOL games datasets:</u>

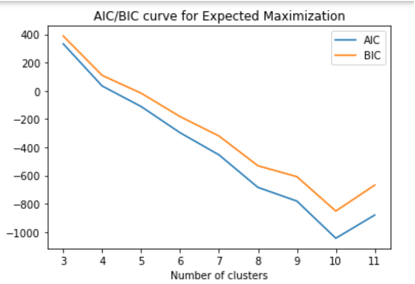


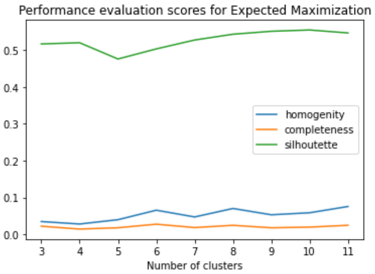


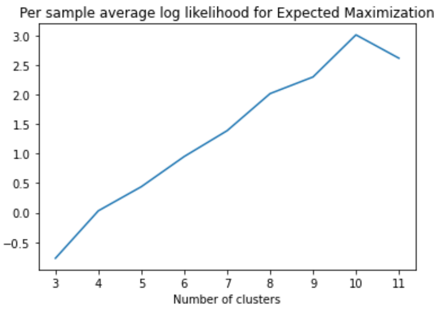
4.2 Expectation Maximization

Similar to K-means clustering method, the number of class for the K-means model is 2 for NBA datasets and class of number is 7 for LOL datasets. The component list is from 3 to 11.

For NBA games datasets:



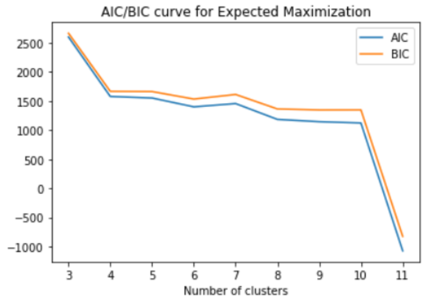


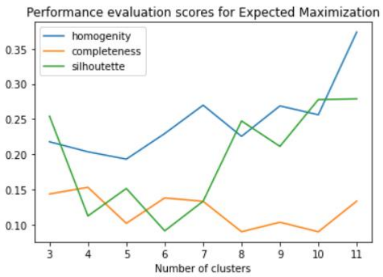


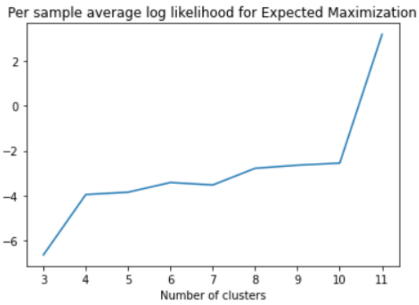
The training accuracy for K-means is K=2:59.067357512953365

The Testing accuracy for K-means is K=2:56.476683937823836

For LOL games datasets:



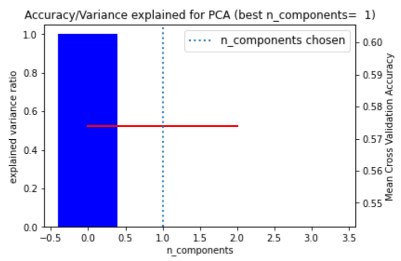
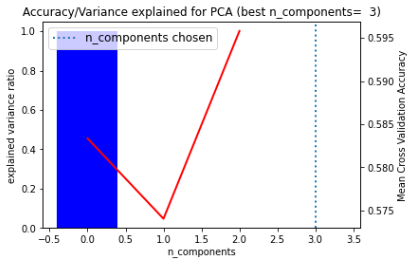
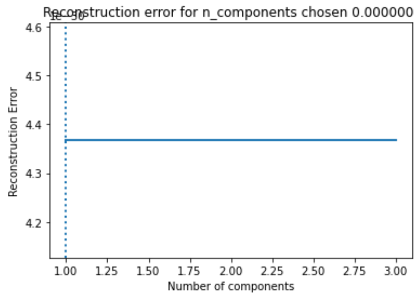
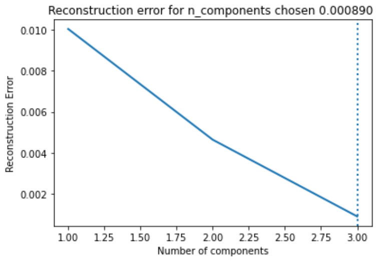
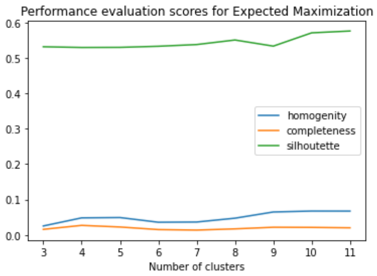
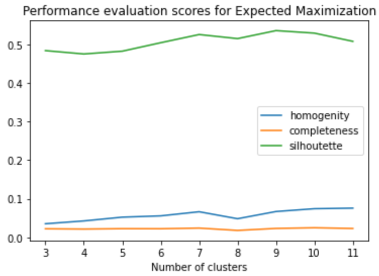
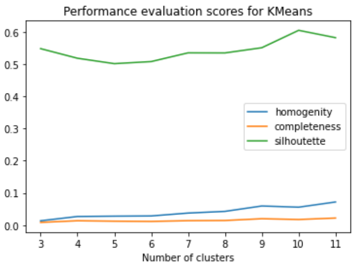
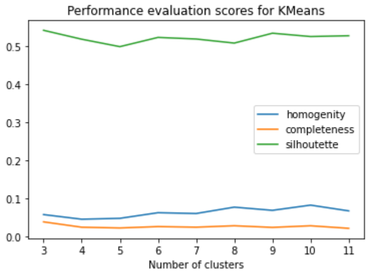




4.3 PCA

As for PCA dimensionality reduction, I firstly build up the neural networks model and then fit the sample datasets. And then plot the spectrum of the PCA. After checking the accuracy for taking all combinations of components (The count should be size of the datasets), I plotted the accuracies and the best components. As an example of NBA datasets, the chosen components should be 1. The accuracy is about 0.575. After that, I plot the relationship between the number of components and the reconstruction errors to judge the evaluation performance of the dimensionality reduction. I re-ran my neural network algorithms and get the performance of the evaluation. After that, I applied this on the clustering model and plot the same figure as what I do before in clustering algorithms implementation, which is helpful to make a comparison for the performance. Firstly, select the optimal dimensions for the PCA algorithms. After that applying k-means clustering and EM clustering to the datasets, and then seed location to show some evaluation scores and variations in the result. Lastly, determine the new optimal value of k clusters for the datasets.

The results are shown as below:

	NBA games	LOL games
Accur		
Re- error		
Eva-em		
Eva-km		
Acur-em-tra	60.880829015544045	59.067357512953365
Acur-em-tes	58.549222797927456	59.067357512953365

Acur-km-tra	58.2901554404145	59.067357512953365
Acur-km-tes	57.51295336787565	59.067357512953365

Accur: Accuracy for the algorithms.

Re-error: Reduction error for n components chosen

Eva-em: Performance Evaluation scores for Expected Maximization

Eva-km: Performance Evaluation scores for K-means

Acur-em-tra: Training accuracy for Expected Maximization

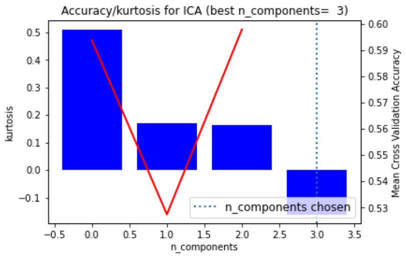
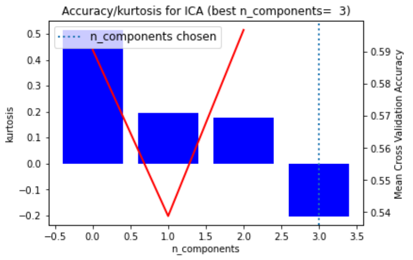
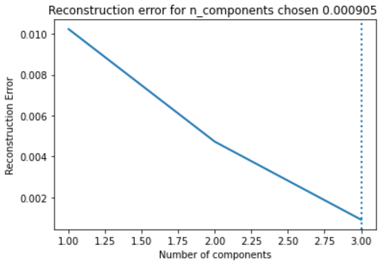
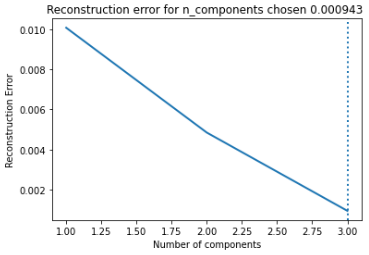
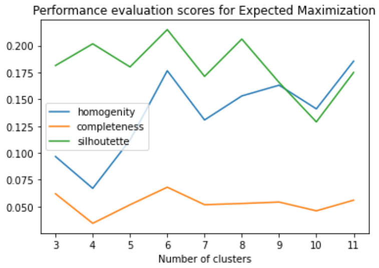
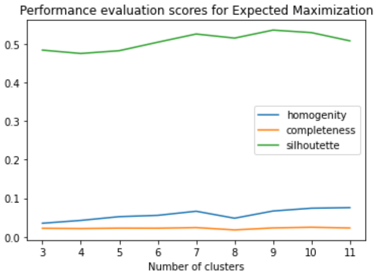
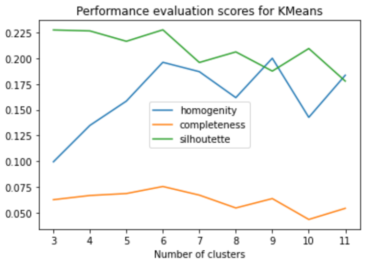
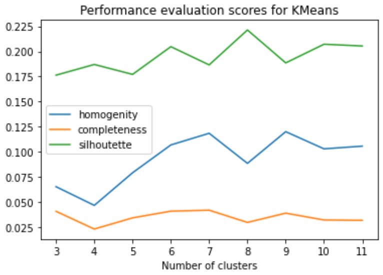
Acur-em-tes: Testing accuracy for Expected Maximization

Acur-km-tra: Training accuracy for K-means

Acur-km-tes: Testing accuracy for K-means

4.4 ICA

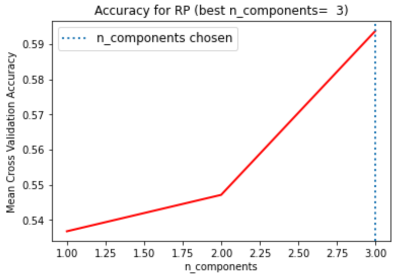
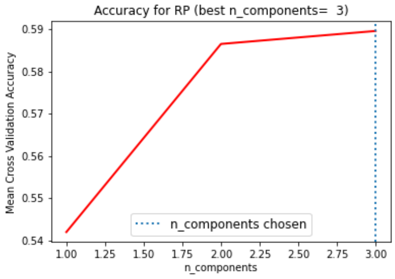
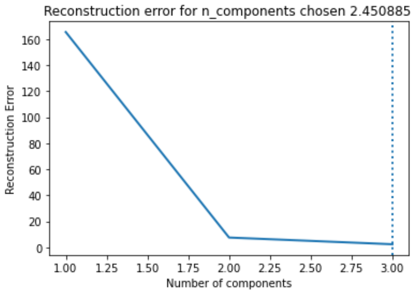
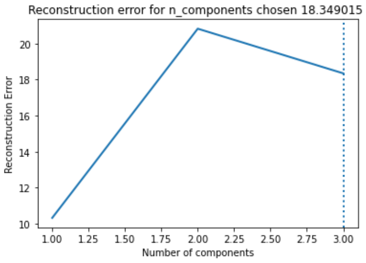
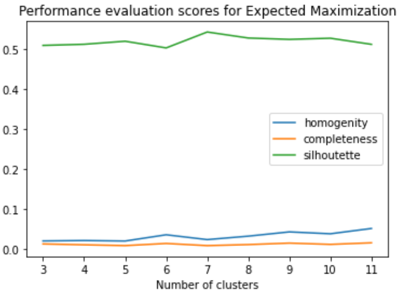
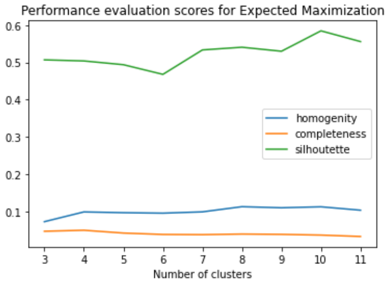
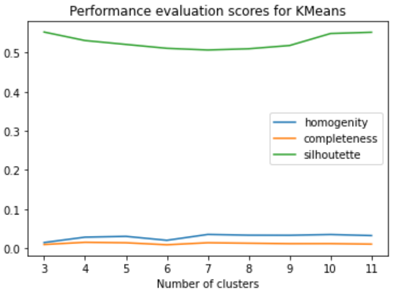
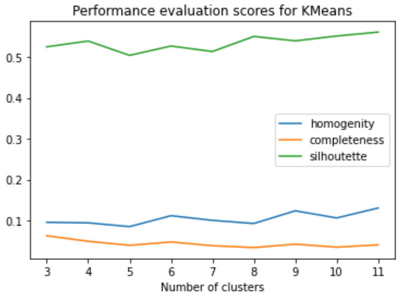
Similar to the method above, the results are shown as below:

	NBA games	LOL games
Accur		
Re- error		
Eva-em		
Eva-km		
Acur-em-tra	61.398963730569946	52.2020725388601

Acur-em-tes	65.80310880829016	50.259067357512954
Acur-km-tra	60.10362694300518	55.181347150259064
Acur-km-tes	68.9119170984456	54.92227979274611

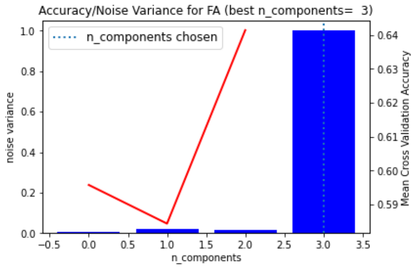
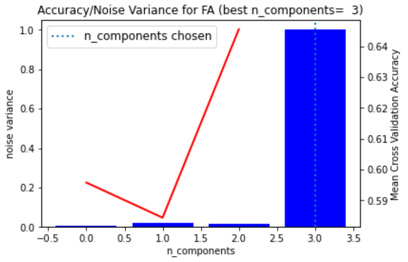
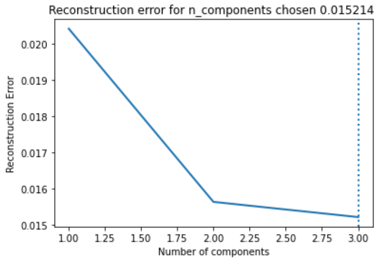
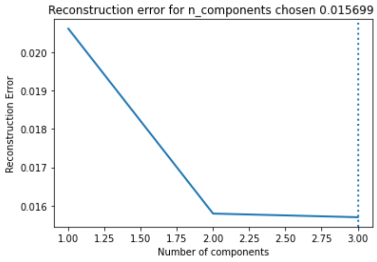
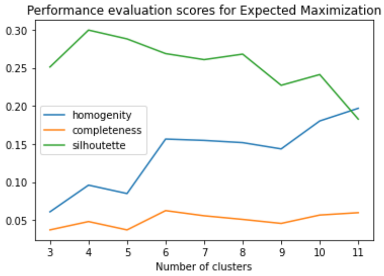
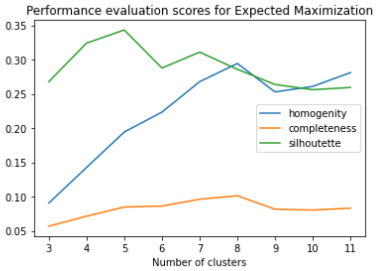
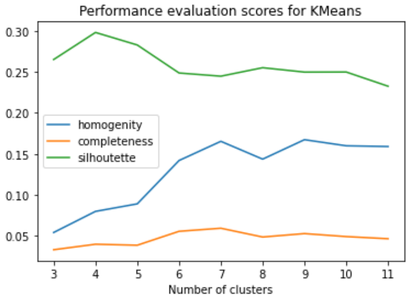
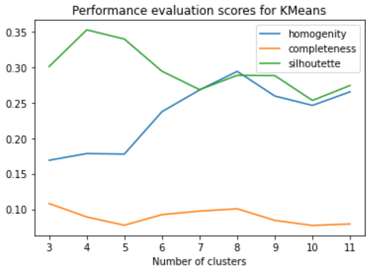
4.5 Randomized Projections

Similar to the method above, the results are shown as below:

	NBA games	LOL games
Accur		
Re- error		
Eva-em		
Eva-km		
Acur-em-tra	59.97409326424871	47.92746113989637
Acur-em-tes	59.97409326424871	51.813471502590666
Acur-km-tra	59.97409326424871	56.865284974093264
Acur-km-tes	59.97409326424871	63.212435233160626

4.6 Factor analysis

Similar to the method above, the results are shown as below:

	NBA games	LOL games
Accur		
Re- error		
Eva-em		
Eva-km		
Acur-em-tra	59.067357512953365	58.41968911917098
Acur-em-tes	60.10362694300518	58.549222797927456
Acur-km-tra	59.196891191709845	59.4559585492228
Acur-km-tes	60.10362694300518	59.067357512953365

5. Conclusion

5.1 Clustering algorithms comparison

As a result, for K-means model, comparing to homogeneity, completeness and silhouette, silhouette has the best performance evaluation scores for both two datasets. For NBA datasets, when number of clusters is increasing, the variance increased and converge to -5. For LOL datasets, this value converges to -275. As for EM model, both AIC and BIC decrease when clusters are increasing for NBA datasets. But for LOL datasets, they are increasing. As for the performance, silhouette is higher for NBA games and homogeneity is higher for LOL games. Per sample average log likelihood is increasing when clusters are increasing for NBA datasets. However, for LOL games, when clusters are 10, it reaches the peak. For these two datasets, we can conclude that the number of clusters, the seed location can influence the performance of the algorithms. Besides, K-means has better training and testing accuracy than EM algorithm.

5.2 Dimensionality reduction algorithms comparison

According to the tables above, comparing with four dimensionality reduction algorithms, we can conclude that for the best components of NBA datasets, PCA is 1 and others is 3. As for the reconstruction errors, PCA is constant but others are decreasing with the increasement of numbers of components. The output curves for AIC/BIC are similar for PCA, ICA. But randomized projections are different from the others. As for the performance of four algorithms, silhouette has the highest scores for EM and K-means after dimensionality reduction. For the per sample average log likelihood of EM, randomized projections and Factor analysis is similar but ICA is completely opposite. For PCA, it reaches the peak when the number of clusters is 8. Last but not least, variance of KM is increasing for all four algorithms. These figures are helpful for me to choose the parameters and numbers of clusters k in the unsupervised learning process.

What's more, comparing the training and testing accuracy from previous result, we can find the accuracy of training and testing obviously increases after using the dimensionality reduction method. The clusters I selected after the dimensionality reduction methods are different the what I used before. The reason is that from the pre-processing for dimensionality reduction, I get the better parameters from the figures so that I can adjusted the parameter in the K-means and EM models.

As a result, clustering organizes data in an unsupervised manner and is a useful preprocessing technique. Dimensionality reduction helps eliminate useless functions, thereby improving clustering and learner accuracy. As described in the algorithm implementation, clustering is applied without isolating tags, which creates a positive bias for including clusters as features. The analysis is still applicable, but the absolute results are biased.

5.3 Summary

From this assignment, I learned how to implement the clustering algorithms and dimensionality reduction algorithms on the datasets. By evaluating the performance of each methods, I can find the best approach to each of the datasets to solve the problem. With the comparison of the previous model and the model after dimensionality reduction, I found that reduce the dimension is an effective way to improve the performance of unsupervised learning. From the course materials, I learn the theoretical knowledge about these four algorithms and by experiencing the algorithms on my own project, I have a deep understanding on the unsupervised learning and dimensionality reduction methods.