**An Analysis of Unsupervised Learning Algorithms**

Assignment 3

Clement Lu   
*CS 7641 Machine Learning*  
*Department of Computer Science, Georgia Tech University*Atlanta, GA  
clu328@gatech.edu

*Abstract*—Unsupervised learning is a subset of machine learning in which inferences are made on the data without the use of labels. In this space, clustering and dimensionality reduction are commonly used. In this paper, the performance of two clustering algorithms, k-means and expectation-maximization (with Gaussian mixture models), are compared across the Credit Card Default and Free Music Archive datasets before and after applying dimensionality reduction. For dimensionality reduction, principle component analysis, independent component analysis, random projections, and linear discriminant analysis, are applied and compared across the same datasets. Finally, for the Free Music Archive dataset, the performance of a neural net was compared for both dimensionally reduced and clustering-augmented versions of the dataset.

# Introduction

As implied by the name, *unsupervised learning* is a form of machine learning that finds relationships in unlabeled data, i.e. training without supervision. Two common applications of unsupervised learning are *clustering* and *dimensionality reduction.* In clustering, the objective is to group the data into subsets or “clusters” using either a distance heuristic or a measure of relatedness [1]. Dimensionality reduction refers to the technique of transforming data into a lower dimension while attempting to retain as much of the original structure as possible [2].

# Data

## Default of Credit Card Clients

Default of Credit Card Clients (DCCC) [3], consists of data for credit card clients in Taiwan from April 2005 to September 2005. Every client is marked either as *1* for having defaulted on their next payment (in October 2005) which will be referred to as *default*, or *0* for not having defaulted on their next payment which will be referred to as *non-default*. DCCC contains a mix of both continuous and categorical features which should prove interesting to observe how clustering algorithms, especially ones that use Euclidean distance measures, handle this data. After one hot encoding, DCCC contains 91 total features.

## Free Music Archive

*Free Music Archive* (FMA) [4], contains audio files, metadata, and extracted features for 106,574 songs. The extracted features contain commons features such as *mel frequency cepstral coefficient* (MFCC) that are extracted by *librosa*, a Python package for music and audio analysis, and other characteristics provided by *Echonest* (now *Spotify*). FMA contains 518 continuous features. FMA is interesting for clustering as it may reveal new genre groupings besides the ones provided by human labelers. Additionally, the high dimensionality of FMA makes it an interesting dataset for applying dimensionality reduction as there may be significant performance gains after removing irrelevant features and reducing the complexity of the dataset.

# Clustering Algorithms

## K-Means

K-means is an iterative clustering algorithm that seeks to group data into *k* clusters in such a way that minimizes the average dissimilarity within each cluster [1]. k-Means performs *hard clustering,* meaning that each object is assigned specifically to one and only one cluster*.* Initially, *k* points are chosen within the dataset and objects are assigned to each cluster based on distance. Typically, a form of Euclidean distance, e.g. squared or weighted Euclidean distance, is used as the distance measure. Once all points have been assigned, a set of means for each cluster is computed and set as the new cluster centers. The assignment of points and computation of cluster centers is repeated until the assignment of points do not change between iterations. This algorithm, assuming squared Euclidean distance, can be represented more succinctly as the following equation where is the cluster assignment and :

(1)

Because of the use of Euclidean distance as a measure of dissimilarity, k-Means is usually only well-suited for datasets that exclusively contain quantitative features as Euclidean distances are not easily measured or representative of dissimilarity for categorical features.

## Expectation-Maximization

Expectation-Maximization (EM) is an iterative algorithm that seeks to approximate the maximum likelihood function. While EM itself is not explicitly a clustering algorithm, it can be used with a Gaussian mixture model (GMM) where each cluster is a Gaussian density with parameters where is the mean and is the standard deviation of the distribution. EM can be used with GMM to perform *soft clustering* where data points are not assigned to only one cluster, but instead, the probability of belonging to each cluster is computed. By choosing the number of clusters or components within a GMM, one assumes the underlying distribution of the data; this allows hidden or “latent” variables to be discovered when such assumptions are utilized [5]. EM consists of two main steps as implied by the name: expectation and maximization.

(2)

The expectation step computes, for missing data , parameters , and observed data , the likelihood function as shown in Equation 2. The maximization step then seeks to find such that it maximizes [5]. Both the expectation and maximization steps are repeated until the log-likelihood does not change significantly between iterations.

# Methodology

## Data Pre-Processing

Both k-Means and GMM require numerical, continuous features since k-Means uses Euclidean distance and GMM uses continuous distributions. In DCCC, all categorical features were one-hot encoded to satisfy these requirements; however, it should be noted that this increases the dimensionality of the dataset and may decrease the performance of the clustering. Both k-Means and GMM are also sensitive to scale and both DCCC’s and FMA’s features were normalized between 0 and 1.

## K-Means

For choosing the optimal number of clusters *k*, the elbow method was used. This method simply involves plotting the scores of some evaluation metric for increasing numbers of *k* and identifying if at any value an “elbow” is observed. The location of the elbow is usually signified by a sharp decline in the gradient of the curve, indicating diminishing returns in increasing *k* values after that point. The *Kneedle* algorithm, an automatic knee (or elbow) detection algorithm, was used in determining the elbow point [6]. A smooth curve that does not show a clear elbow, however, may indicate that the metric used in the elbow analysis is not the right one and/or the dataset is not attuned to clustering. In this analysis, the elbow method was utilized prior to clustering both DCCC and FMA with both the distortion score and the silhouette score. The distortion score measures the sum of squared distances from each point to its assigned cluster centroid [7]. As such, a lower distortion score generally indicates better clustering.

(3)

The silhouette score is the mean of the silhouette coefficientsof all the data points. The silhouette coefficient is calculated by Equation 3 where is the mean distance between and all other data points in the same cluster and is the smallest mean distance between and all data points in any other cluster except its own [8]. In general, a higher silhouette score indicates better clustering.

## EM+GMM

Like k-Means, the elbow method was used for selecting the number of clusters; however, instead of calculating distortion score and silhouette score, the *Akaike information criterion* (AIC) and the *Bayesian information criterion* (BIC) were used.

(4)

(5)

Equation 4 shows how AIC is calculated where *k* is the number of clusters and *L* is the maximum value of the likelihood function. Similarly, Equation 5 shows how BIC is calculated with the addition of *N* which is the number of samples in the dataset. A naïve approach to *k* selection for EM+GMM would be to simply choose the model with the highest likelihood; however, this could lead to overfitting [9]. Both AIC and BIC address this by penalizing the use of more clusters while still attempting to maximize the likelihood. A lower AIC and BIC score generally indicate a better clustering.

## Performance Metrics

To evaluate clustering performance, homogeneity, adjusted rand index (ARI), and silhouette scores are computed for all clusterings. Given the ground truth, homogeneity measures the diversity of classes within the clusters. ARI is an adjusted-for-chance metric that scores the similarities between two clusterings by comparing clustering assignments between all pairs of points in the ground truth and the clustering prediction. The silhouette score has been previously described and is shown in Equation 3.

## Principle Component Analysis

Principle component analysis (PCA) is technique that linearly combines features into new components such that each is orthogonal to the preceding component, i.e. linearly uncorrelated [10]. PCA components are ordered by variance and the first principle component will always account for the most variability in the data. A scree plot was utilized for determining the number of principle components to keep. After plotting the cumulative explained variance of the first *k* principle components, a threshold of retaining 90% of the overall explained variation was used as the cut-off point in choosing the number of principle components.

## Independent Component Analysis

Independent component analysis (ICA) is a technique that can be used to separate a multivariate dataset into statistically independent components [11]. ICA seeks to maximize non-Gaussianity to find independent features; therefore, kurtosis is used for determining an optimal number of independent components to use. Kurtosis is the fourth moment of a Gaussian distribution and measures the distribution’s similarity to a Gaussian distribution [12]. Using the Fisher definition of kurtosis, a kurtosis of 0 represents a Gaussian distribution, whereas any positive or negative number represents a non-Gaussian. As the absolute value of the kurtosis increases, the distribution’s likeness to a Gaussian distribution decreases. Therefore, the number of components with the highest average absolute value kurtosis should represent the set of components with the most independent features.

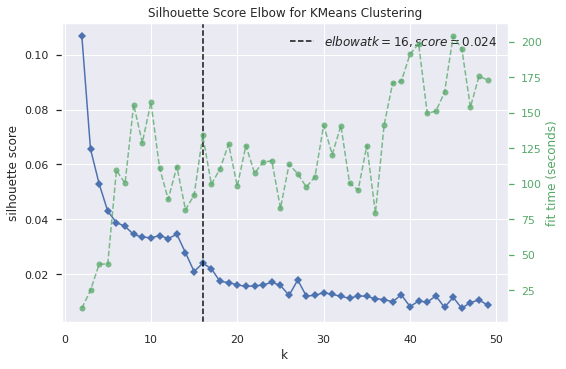
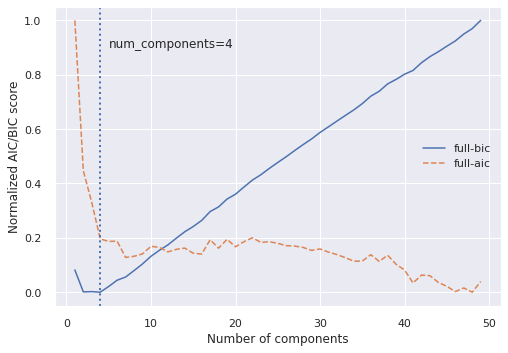
 

Fig. 2. FMA elbow plots. Left: (a). Distortion score for k-Means. Center: (b). Silhouette score for k-Means.

Right: (c). AIC/BIC scores for EM+GMM

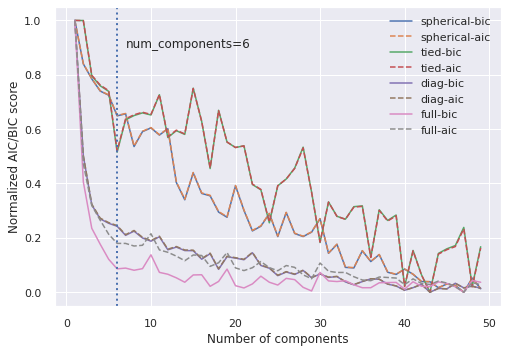
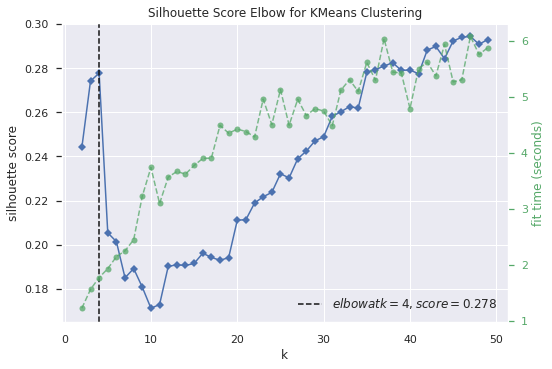
 

Fig. 1. DCCC elbow plots. Left: (a) Distortion score for k-Means. Center: (b). Silhouette score for k-Means.

Right: (c). AIC/BIC scores for EM+GMM.

## Random Projection

Based on the Johnson-Lindenstrauss lemma, random projection seeks to map a set of high-dimensional points to a lower dimension subspace “in such a way that relative distances between data points are nearly preserved” [13]. This lower dimension subspace is chosen at random based on some distribution. In this experiment, Gaussian random projection is used. The reconstruction error for the random projection is calculated and plotted for each component. The lowest number of components to achieve less than or equal to 1.0 mean squared error is chosen. To achieve consistent results, the random projection is run 10 times and the projection that returns the lowest number of components based on the criteria mentioned above is used.

## Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a technique like PCA that seeks to find linear combinations of features that account for the highest variability in the data. However, LDA is a supervised method that seeks to separate the data based on its classes unlike PCA which is unsupervised. The number of components used in LDA can be selected by plotting the cumulative explained variance versus the number of components selected. For consistency, a threshold of 90% is also used here for selecting the optimal number of components.

# Experiments and Results

## Clustering

### K-Means

Plotting the distortion scores and silhouette scores for k-Means clustering on DCCC, as illustrated in Fig. 1(a) and Fig. 1(b), show that an optimal *k* is either 12 or 4. However, the distortion score plot does not display a clear elbow as the silhouette plot does so 4 is chosen as *k* for clustering. For FMA, the distortion and silhouette scores are plotted in Fig. 2(a) and Fig. 2(b), respectively. The elbow points are selected as 15 and 16; however, neither display a strong elbow point which may be indicative of poor clustering ability for FMA. As such, the point with the highest silhouette score at 2 is chosen for FMA clustering. Of note, the silhouette scores for FMA are all relatively low with the majority below 0.04.

### EM+GMM

Plotting the AIC and BIC scores for EM+GMM fit on DCCC in Fig. 1(c) shows an optimal elbow at 6. The lowest scores are attributed to using a “full” covariance type in which each component has its own general covariance matrix. Fig. 2(c) shows the AIC and BIC scores for EM+GMM fit on FMA where a distinct elbow is observed at 4 components for both AIC and BIC. The scores have only been computed with a “full” covariance type as the large dimensionality of FMA made fitting EM+GMM on all covariance type for all number components impractical. Interestingly, the elbow for the BIC scores is also its global minimum, with the scores monotonically increasing past 4 components as opposed to AIC which continues to decrease. This displays the stronger regularization factor in BIC that penalizes the use of more components.

## Dimensionality Reduction

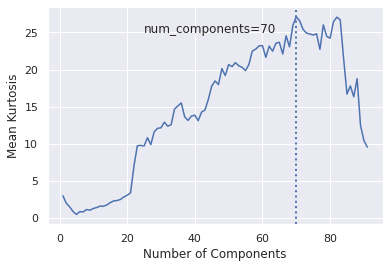


Fig. 7. Mean kurtosis for DCCC ICA for 1 to 91 components

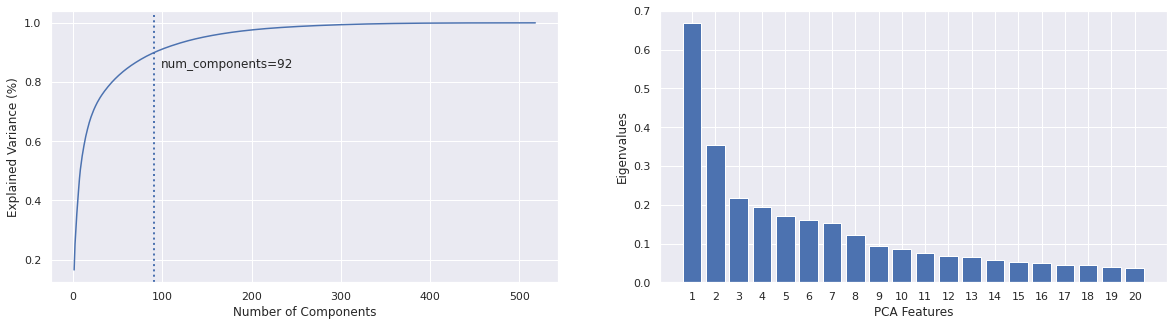


Fig. 6. FMA PCA scree plot for top 20 principle components

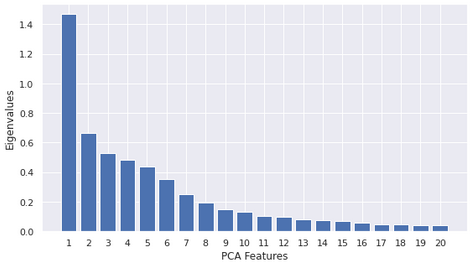


Fig. 4. DCCC PCA scree plot for top 20 principle components

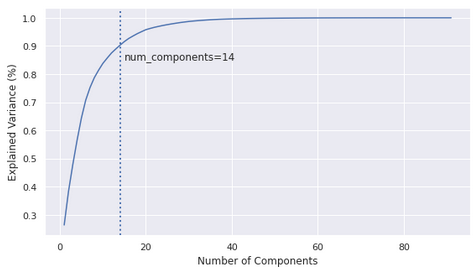


Fig. 3. DCCC PCA cumulative explained variance for 1 to 91 components

### PCA

Fig. 3 and Fig. 5 show the cumulative explained variances of DCCC and FMA, respectively, by the number of principle components. Using 90% explained variance as the threshold, 14 principle components are kept for DCCC PCA dimensionality reduction and 91 principle components are kept for FMA PCA dimensionality reduction. Additionally, as shown in Fig. 4, the scree plot for DCCC PCA shows that the first principle component accounts for a significant amount of variance relative to the other principle components. PC1 has an eigenvalue of over 1.4, while PC2 has a value of just over 0.6. FMA displays similar behavior in Fig. 6 where the scree plot shows that FMA’s first two principle components have eigenvalues of 0.68 and 0.35, respectively, where the following principle components have eigenvalues of around 0.2 and below.

### ICA

Fig. 7 and Fig. 8 show the mean absolute value kurtosis of the resulting features after performing ICA on DCCC and FMA, respectively, on a varying number of components. The most kurtotic responses occur at 70 components selected for DCCC and 171 components selected for FMA.

### Random Projection

Fig. 9 and Fig. 10 show the MSE of 10 random projections of DCCC data and FMA data, respectively, from 1 to 50 components. For this experiment, an MSE of less than or equal to 1.0 is chosen as a threshold for the number of components to use. This threshold is first achieved at 5 components for DCCC and at 76 components for FMA. There is a relatively large amount of variance in MSE between runs for both DCCC and FMA using less than 10 components. As more components are used, the MSE seems to converge between runs for random projection.

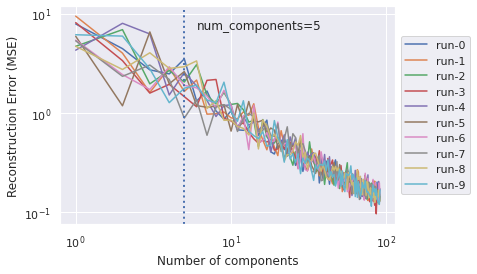
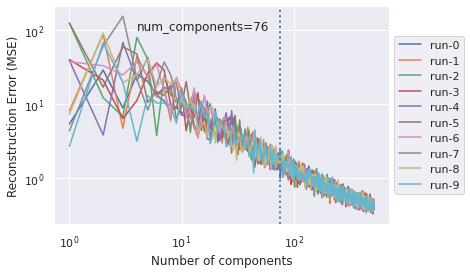


Fig. 9. DCCC random projection reconstruction error for 1 to 91 components



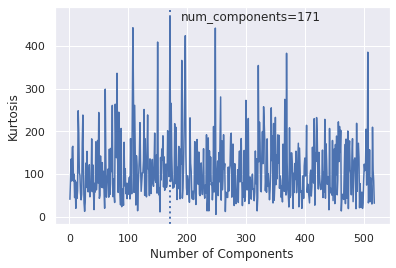


Fig. 8 Mean kurtosis for FMA ICA for 1 to 518 components

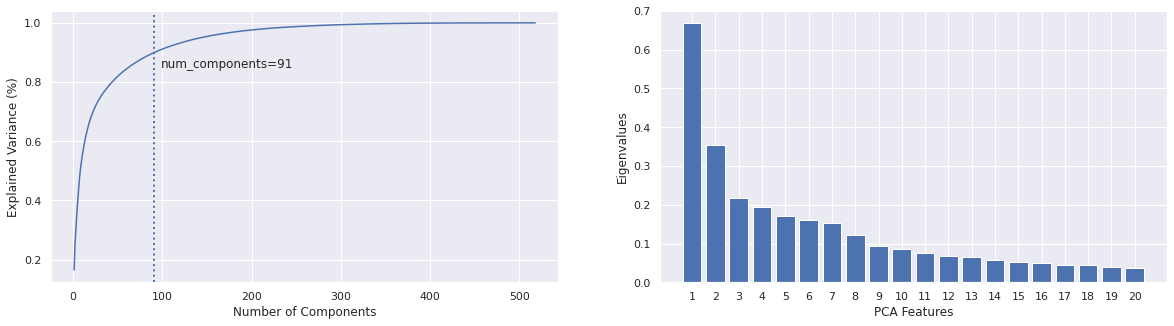


Fig. 5. FMA PCA cumulative explained variance for 1 to 518 components

Fig. 10. FMA random projection reconstruction error for 1 to 518 components

### Linear Discriminant Analysis

Since LDA can at most utilize components, where is the number of classes, there is no need to select the number the number of components for binary classification datasets like DCCC where only 1 component can be selected.

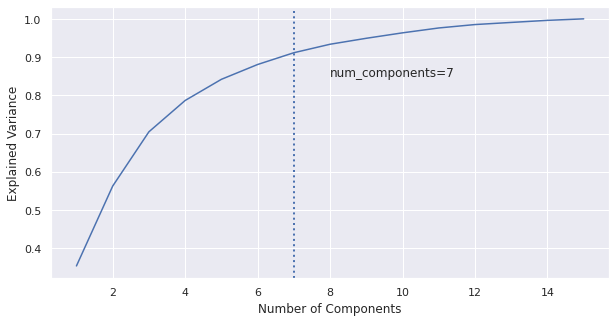
****

Fig. 11. FMA LDA cumulative explained variance per number of components

For FMA, plotting the cumulative explained variance from 1 to 15 components (FMA has 16 classes), as illustrated in Fig. 11, shows that 7 components can explain over 90% of the total variance.

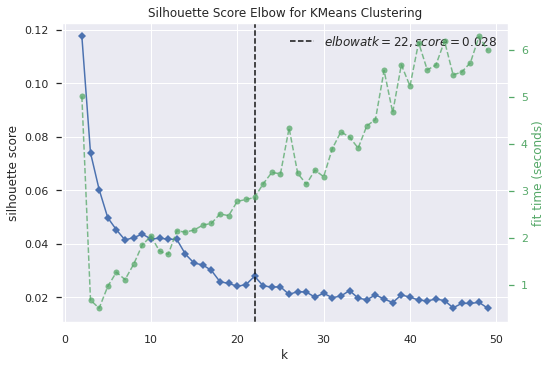
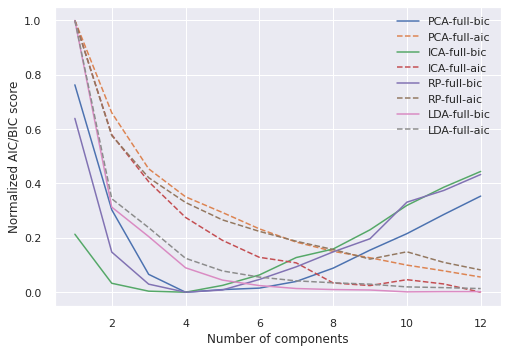
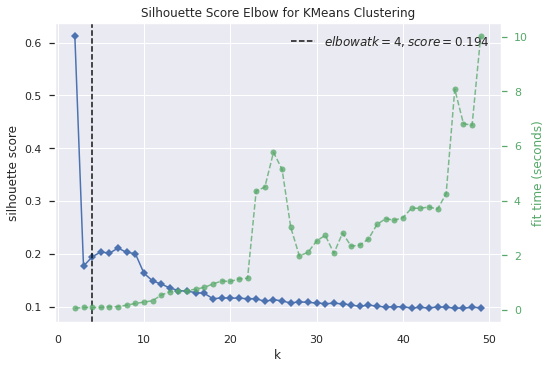
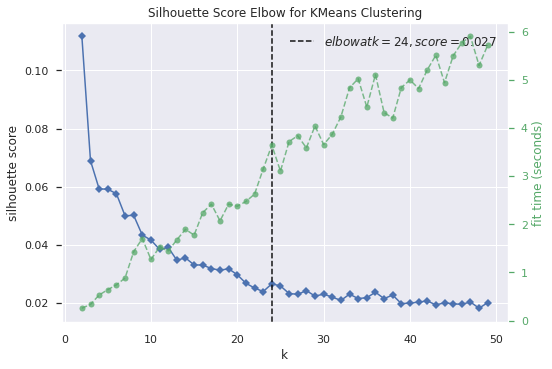
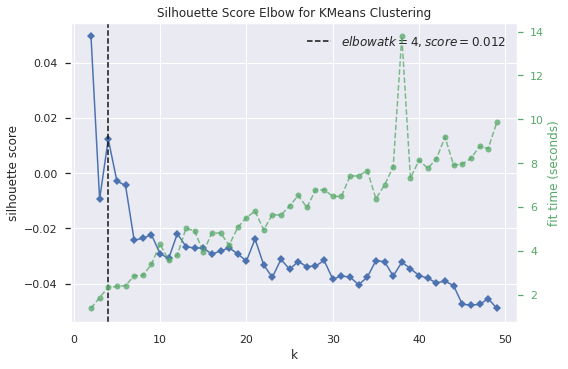
 

Fig. 13. FMA elbow plots. Top Left: (a) Silhouette scores for k-Means with PCA. Top Center: (b). Silhouette scores for k-Means with ICA.

Top Right: (c). Silhouette scores for k-Means with RP. Bottom Left: (e). Silhouette scores for k-Means with LDA.

Bottom Right: (f). AIC/BIC scores for EM+GMM for all dimensionally reduced data.

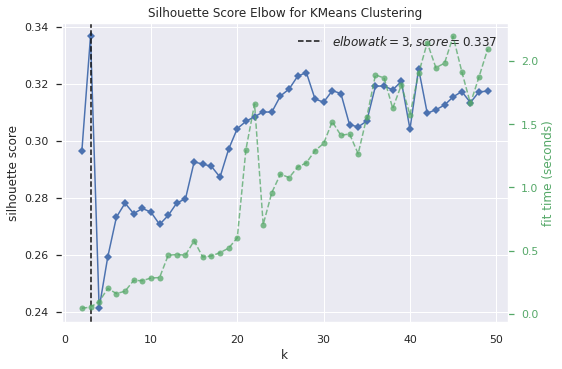
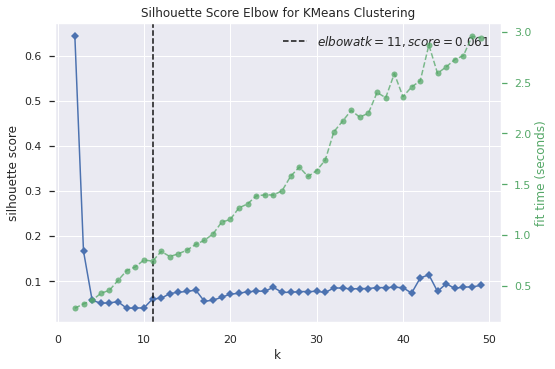
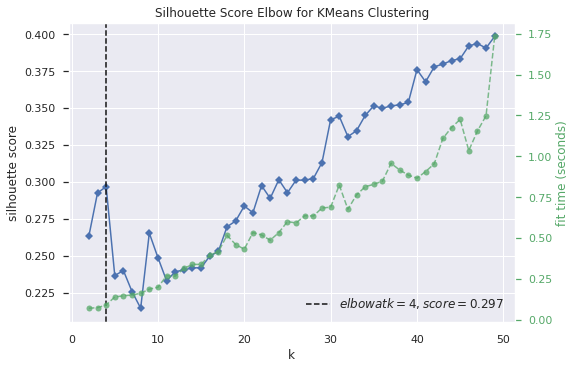
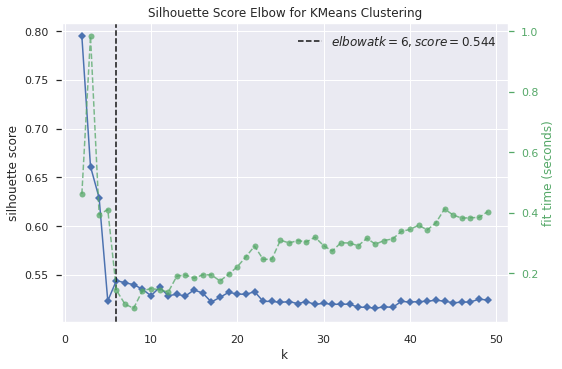
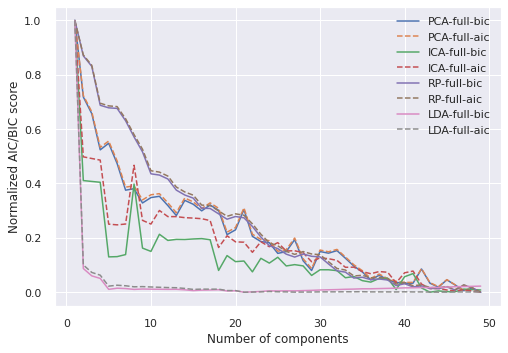
  

Fig. 12. DCCC elbow plots. Top Left: (a) Silhouette scores for k-Means with PCA. Top Center: (b). Silhouette scores for k-Means with ICA.

Top Right: (c). Silhouette scores for k-Means with RP. Bottom Left: (e). Silhouette scores for k-Means with LDA.

Bottom Right: (f). AIC/BIC scores for EM+GMM for all dimensionally reduced data.

## Clustering with Dimensionality Reduction

### K-Means

Fig. 12(a), (b), (c), and (e) show the silhouette elbow plots for DCCC k-Means clustering on PCA reduced data, ICA reduced data, RP reduced data, and LDA reduced data, respectively. The elbow points for PCA and RP are observed at 4 and 3 clusters, respectively. However, for ICA and LDA, the silhouette scores generally decrease as more clusters are used; therefore, 2 is selected for the optimal number of clusters for ICA and LDA as the highest silhouette score occurs at that points for both. Fig. 13(a), (b), (c), and (e) show the silhouette elbow plots for FMA k-Means clustering on PCA, ICA, RP, and LDA reduced data, respectively. Like the original dataset, dimensionally reduced FMA also displays poor clusterability with low initial and generally decreasing silhouette scores as the number of clusters grow. The highest mean silhouette score occurs at 2 for all dimensionally reduced datasets for FMA. The distortion score curves for both DCCC and FMA for all dimensionally reduced datasets lacked a clear elbow point and have been omitted.

### EM+GMM

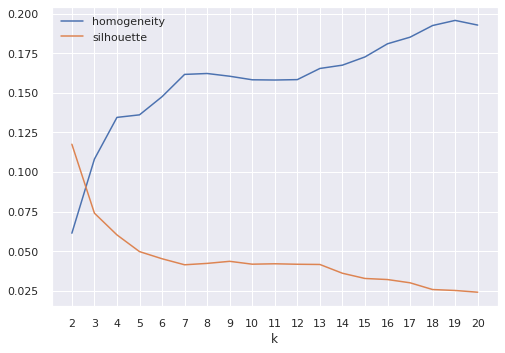


Fig. 14. Homogeneity and silhouette scores for *k* 2 to 20 for FMA PCA k-Means clustering

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Accuracy** | **Macro** **F1** | **Weighted F1** | **Macro Precision** | **Weighted Precision** | **Macro Recall** | **Weighted Recall** |
| FMA-baseline | 0.6548 | 0.4576 | 0.6381 | 0.5332 | 0.6333 | 0.4415 | 0.6548 |
| FMA-PCA | 0.6484 | 0.4698 | 0.6307 | 0.5008 | 0.6286 | 0.4617 | 0.6484 |
| FMA-ICA | 0.6600 | **0.4883** | 0.6445 | **0.5464** | 0.6401 | **0.4677** | 0.6600 |
| FMA-RP | 0.6006 | 0.3724 | 0.5724 | 0.4390 | 0.5651 | 0.3803 | 0.6006 |
| FMA-LDA | **0.6664** | 0.3943 | 0.6412 | 0.4350 | 0.6326 | 0.3986 | **0.6664** |
| FMA+k-Means | 0.6400 | 0.3778 | 0.6102 | 0.4513 | 0.6275 | 0.3752 | 0.6400 |
| FMA+EM+GMM | 0.6644 | 0.4404 | **0.6520** | 0.4578 | **0.6495** | 0.4369 | 0.6644 |

Table 2. ANN classification results on FMA reduced and clustering-augmented data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Clusters** | **Fit Time (s)** | **Homogeneity** | **ARI** | **Silhouette** |
| **DCCC** | | | | | |
| kMeans-baseline | 4 | 4.26 | 0.096 | 0.080 | 0.278 |
| kMeans-PCA | 4 | 0.42 | 0.097 | 0.080 | 0.297 |
| kMeans-ICA | 2 | 0.32 | 0.000 | -0.001 | 0.643 |
| kMeans-RP | 3 | 0.06 | 0.026 | 0.035 | 0.337 |
| kMeans-LDA | 2 | **0.02** | **0.134** | **0.295** | **0.795** |
| EM+GMM-baseline | 6 | 4.60 | 0.032 | 0.018 | 0.099 |
| EM+GMM-PCA | 7 | 0.75 | 0.095 | 0.049 | 0.122 |
| EM+GMM-ICA | 5 | 3.82 | 0.030 | 0.050 | 0.095 |
| EM+GMM-RP | 10 | 1.45 | 0.085 | -0.009 | 0.061 |
| EM+GMM-LDA | 2 | 0.08 | 0.131 | 0.264 | 0.700 |
| **FMA** | | | | | |
| kMeans-baseline | 2 | 7.48 | 0.061 | 0.075 | 0.107 |
| kMeans-PCA | 2 | 0.59 | 0.061 | 0.075 | 0.118 |
| kMeans-ICA | 2 | 1.41 | 0.070 | 0.091 | 0.049 |
| kMeans-RP | 2 | 0.27 | 0.062 | 0.079 | 0.112 |
| kMeans-LDA | 2 | **0.07** | 0.041 | 0.017 | **0.612** |
| EM+GMM-baseline | 6 | 124.79 | 0.081 | 0.060 | 0.043 |
| EM+GMM-PCA | 4 | 6.95 | **0.126** | **0.154** | 0.013 |
| EM+GMM-ICA | 4 | 8.07 | 0.095 | 0.095 | -0.011 |
| EM+GMM-RP | 4 | 10.92 | 0.123 | 0.146 | -0.003 |
| EM+GMM-LDA | 2 | 0.17 | 0.098 | 0.121 | 0.260 |

Table 1. Clustering performance for DCCC and FMA

Fig. 12(f) shows the AIC and BIC scores for DCCC EM+GMM on all dimensionally reduced data. The elbow points for PCA, ICA, RP, and LDA are observed at 7, 5, 10, and 2 clusters, respectively. Fig. 13(f) shows the AIC and BIC scores for FMA EM+GMM on all dimensionally reduced data. The elbow points for PCA, ICA, RP, and LDA are observed at 4, 4, 4, and 2, respectively.

## ANN Learning on Dimensionally Reduced Data

A neural network learner was trained for PCA, ICA, RP, and LDA reduced FMA data. A grid search was run for each learner to obtain the parameters that resulted in the highest macro F1 score. Table 2 shows the results for each learner.

## ANN Learning on Clustering Augmented Data

Since the number of clusters selected for FMA differs drastically from the total number of classes in the original dataset, clustering labels have been used to augment FMA data rather than replacing it altogether. This could potentially work as a filtering step by including a feature that pre-groups data points that have some sort of similarity (as determined by clustering). A neural network learner was trained using both k-Means and EM+GMM augmented FMA data. Table 2 shows the results for each learner.

# Discussion

## Clustering and Dimensionality Reduction

Based on the results in Table 1, it does not appear that either DCCC or FMA are clustering-friendly – at least in accordance with the ground truth labels and evaluating homogeneity. For FMA especially, this is evident in the fact that the clustering method with the highest homogeneity score, *EM+GMM-PCA* (0.126), has one of the lowest silhouette scores (0.013). Likewise, *kMeans-LDA*, has the highest silhouette score for FMA clustering (0.612), but has one of the lowest homogeneity scores (0.041). Intuitively, it makes sense that homogeneity increases with *k* as there at least needs to be 16 clusters to have homogenous clusters when there are 16 ground truth labels. Fig. 14 displays this inverse relationship between homogeneity and silhouette scores which suggests there does not exist a such that homogeneity and the silhouette score are maximized. This suggest that the relationship between a song’s features and its genre may not be linear as k-Means produces linear decision boundaries. This does not mean that this is a poor clustering, but rather another labeling has been discovered. In many cases, where unsupervised learning is most appropriate, the ground truth is not known. The silhouette score, which is an unsupervised metric, seems to indicate that a more conclusive grouping occurs at . Using the top 5 features weight from PC1 and PC2, as shown in Fig. 15, gives an idea of the features that account for the most variance in the principle components: *chroma* and *spectral roll-off*. Note that this does not imply that these features are necessarily the most important, but rather that they are the most important for PC1 and PC2. However, this does provide insight into potentially what the two new clusters represent and what k-Means is considering in its clustering.

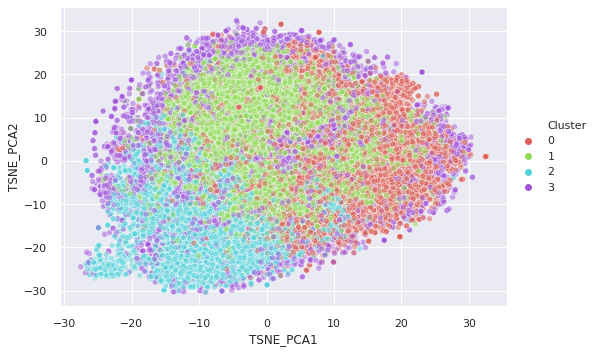
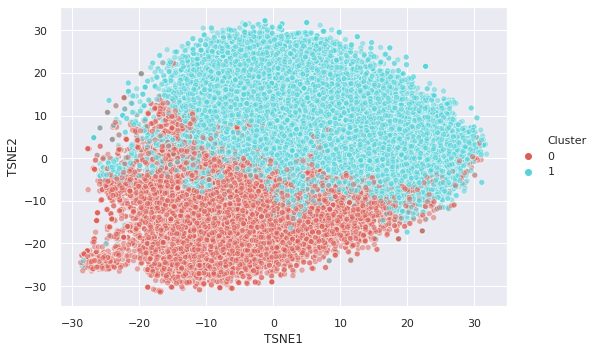


Fig. 17. Left: TSNE projection on k-Means clustered FMA data, Right: TSNE projection on EM+GMM clustered, FMA-PCA reduced data

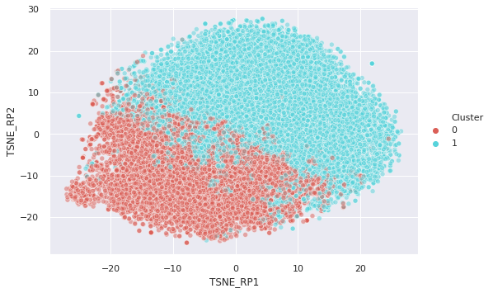
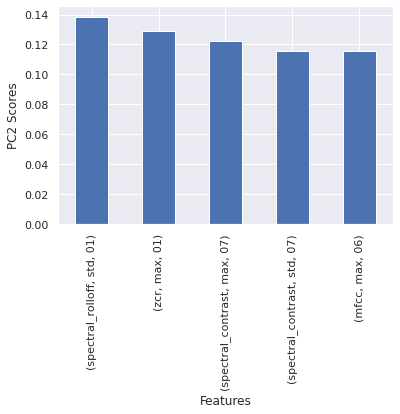
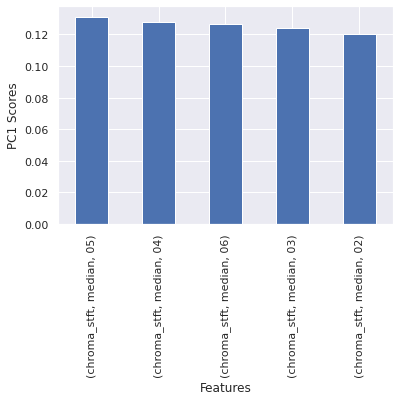
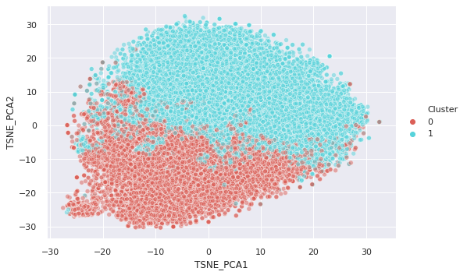
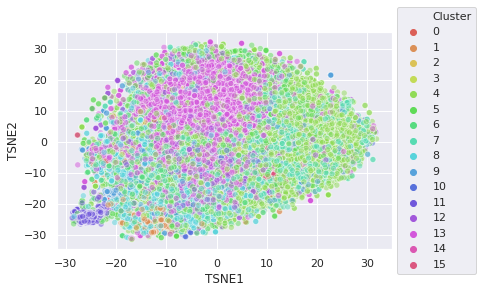


Fig. 18. TSNE projection on k-Means clustered RP-reduced data

Fig. 15. Top 5 contributing features for PC1 and PC2 for FMA

Fig. 16. Left: TSNE projection on FMA data, Right: TSNE projection on k-Means clustered, FMA-PCA reduced data

Observing the TSNE visualizations of the original FMA data and the FMA-PCA reduced data as shown in Fig. 16 shows that PCA has retained the structure of the original data quite well; however, the clustering shows that by maximizing clustering performance, as measured by the silhouette score, seems to ignore the groupings present in the original dataset provided by the ground truth labels. Interestingly, the TSNE projection for k-Means clustering on FMA PCA-reduced (Fig. 16 right) and non-reduced data (Fig. 17 left) appear almost identical and lends to the intuition that both k-Means and PCA actually seek to maximize the same objective function [14]. This is further evident in that all performance metrics for FMA k-Means clustering on the original dataset versus the PCA-reduced dataset are nearly identical except for fit times, which are considerably lower for PCA by an order of magnitude, due to the reduced dimensions. Even faster, however, is the use of random projections with k-Means clustering. While PCA will maintain the best projections, RP is able to preserve distances nearly as well with significantly less computational overhead [15]. Fig. 18 shows that RP projects a space that is almost identical to that of the original FMA dataset and the PCA-reduced counterpart. As such, the resulting k-Means clustering is almost identical as is evident by the similar homogeneity, ARI, and silhouette scores achieved by performing k-Means on all three.

In terms of homogeneity, EM+GMM on FMA PCA-reduced data achieves the best score of 0.126. The right plot in Fig. 17 visualizes this clustering and it is evident that the EM+GMM clusters somewhat resemble the larger genre groupings in the ground truth (Fig. 16 left). This also supports the claim that the relationship between a song’s features and its genre is non-linear as GMMs are capable of modeling non-linear relationships.

## Neural Network Learning

When evaluating the macro F1 score, the neural network trained on FMA PCA and ICA-reduced data *outperforms* the baseline trained on the original dataset with FMA-ICA achieving the highest macro F1 score of 0.4883. While this may seem counterintuitive since information is lost when dimensionality reduction is performed, representing the data in less dimensions may actually remove irrelevant features and reduce overfitting. In addition, PCA and ICA may also serve to reduce noise that may also lead to overfitting on the neural network learner. ICA itself lends well to FMA, a collection of audio features and measure, since it is commonly applied to blind source audio separation problems [16] and has been successfully applied for music genre identification [17]. Presented with features such as *Mel Frequency Cepstral Coefficients* (MFCCs) and *chroma* that are present in FMA, ICA is able to identify the intendent signals represent specific instruments or tones that may only be used in specific genres of music [18]. Curiously, as LDA is a supervised dimensionality reduction technique, the neural network trained on LDA-reduced data performed relatively poorly with a low macro F1 score of 0.3943. This may be a case of overfitting, given the highly unbalanced nature of FMA with two genres accounting for over half of the samples, as the observed accuracy from LDA is the highest of all the learners while it achieves one of the lowest macro F1 scores.

# Conclusion

In this experiment, the performance of two clustering methods, k-Means and expectation-maximization with Gaussian mixture models were explored. Several methods, including the elbow method, silhouette score, and Bayesian information criterion, were shown to help determine the number of clusters to select. Additionally, four dimensionality reduction techniques, principle component analysis, independent component analysis, random projection, and linear discriminant analysis were explored. Techniques for selecting a critical number of components to keep for each dimensionality reduction problem were demonstrated. Finally, it was shown that dimensionality reduction can be applied to clustering and supervised learning problems in way to increase both fit times and clustering/classification performance.

# References

1. <https://link.springer.com/content/pdf/10.1007%2F978-0-387-84858-7_14.pdf>
2. <https://members.loria.fr/moberger/Enseignement/AVR/Exposes/TR_Dimensiereductie.pdf>
3. Archive.ics.uci.edu. (2020). *UCI Machine Learning Repository: default of credit card clients Data Set*. [online] Available at: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients [Accessed 10 Feb. 2020].
4. Defferrard, M., Benzi, K., Vandergheynst, P. and Bresson, X. (2020). *FMA: A Dataset For Music Analysis*. [online] arXiv.org. Available at: https://arxiv.org/abs/1612.01840v2.
5. <https://www.ics.uci.edu/~smyth/courses/cs274/notes/EMnotes.pdf>
6. https://raghavan.usc.edu//papers/kneedle-simplex11.pdf
7. <https://www.scikit-yb.org/en/latest/api/cluster/elbow.html>
8. <https://en.wikipedia.org/wiki/Silhouette_(clustering)>
9. <https://arxiv.org/pdf/1708.00605.pdf>
10. <https://en.wikipedia.org/wiki/Principal_component_analysis>
11. https://en.wikipedia.org/wiki/Independent\_component\_analysis
12. <http://labs.seas.wustl.edu/bme/raman/Lectures/Lecture14_ICA.pdf>
13. W. B. Johnson and J. Lindenstrauss, “Extensions of lipschitz mappings into a hilbert space,”Contemporary mathematics, vol. 26, no. 189-206, p. 1, 1984.
14. <https://dl.acm.org/doi/10.1145/1015330.1015408>
15. <http://users.ics.aalto.fi/ella/publications/randproj_kdd.pdf>
16. <https://arxiv.org/abs/1907.03432>
17. <https://link.springer.com/article/10.1007/s10844-017-0464-5>
18. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.112.2709&rep=rep1&type=pdf