**Review of the Cover-type Dataset**

In assignment one, the cover-type dataset was used in a supervised learning setting. Given a planar view of a 32x32 meters square of forestland. There were 54 features gathered from each of the 15120 instances. The instances are then classified into 1 of 7 different ‘cover types.’ Cover types are classified by whatever foliage covers a specific area. In the previous assignment, the data was run against 5 supervised algorithms. It performed with a considerably low rate of error when run through a neural network and a boosted decision tree. It is important to note that the data was mixed with discrete values, binary values, and continuous values. This cover-type date set was chosen for this analysis due to its interesting array of feature types and classifications. This is opposed to the single-feature type datasets with fewer labels like the Connect 4 and Wilting data sets that were used in the past assignments. It remains to be proven if this data set can stand the test of unsupervised learning and feature selection. It is also important to note that before any of the algorithms for this analysis were run, this dataset was scaled equally to help balance all of the features.

**The Olivetti Face Dataset**

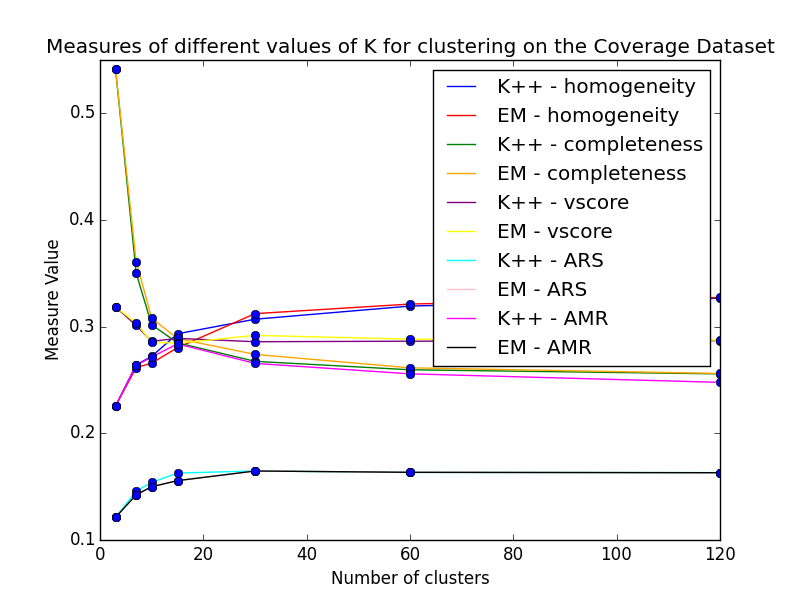
This classic dataset was created by and is credited to AT&T Laboratories Cambridge. It was described on the original website[[1]](#footnote-1) as follows:

*There are ten different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement).*

The Olivetti Face Dataset is used in this unsupervised analysis due to the assumption that it would perform well with the clustering and feature transformation algorithms. The dataset contains data of 400 different pictures, 10 for each subject. The dataset also contains a discrete target class that pertains to the person that the image shows. Each label contains 10 images or instances within the data, which yields 400 instances within the data. Each feature refers to a gray scale intensity pixel mapping from 0 to 255 in a 64x64 image producing 4096 features. This dataset was chosen because of its powers to demonstrate transformation reduction algorithms.

**Clustering Algorithms on the Coverage Dataset**

There were 7 iterations of differently sized max clusters that the analysis observed. Five different measurements were used against the supervised clusterings. Because the clustering labels are unidentified, similarity measurements must be used to make sense of the results. The 5 measurements are stated and briefly explained.

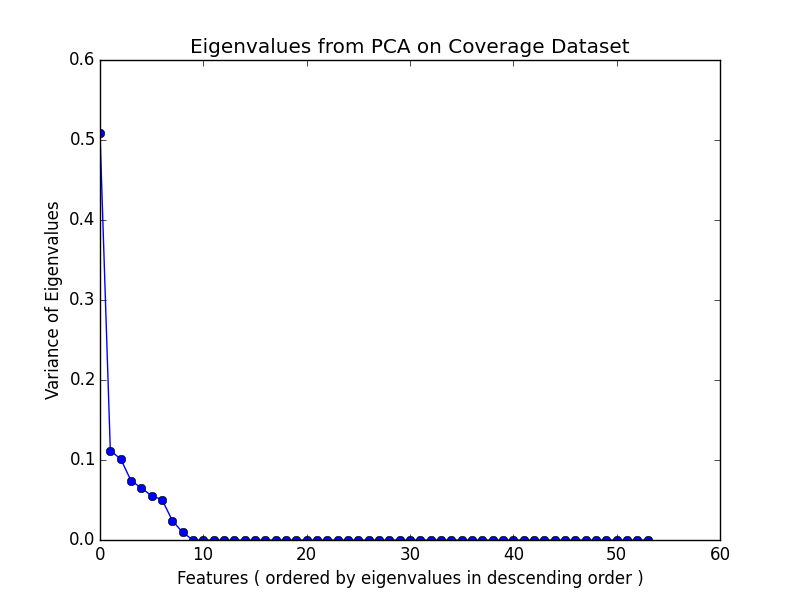
**Adjusted Rand Index (ARI):** No assumption of the predicted labels are made against the ground truth labels besides having a similar output structure. **Adjusted Mutual Information (AMI):** No assumption of the predicted labels are made against the ground truth labels besides having a agreeable output structure. This is based off the entropy of the two sets. **Homogeneity:** This shows how well the clusters contain a single label. **Completeness: This shows how well the same labels are contained in the same clusters. V-measure: This is similar to the harmonic mean and is like NMI (mutual information), but it is normalized over the labels’ entropies to find the best K value for the clustering algorithms. The measurements were scrutinized against each other. Most of the measurements seem to make an intersection when k = 15. Interestingly enough, when k = 10, the algorithms return slightly better results than when k = 7 (the size of the unique labels). As clusters go down, the data appears to have a higher completeness rate. This result seems obvious because more clusters would spread the labels apart. Moreover, homogeneity should increase as clusters increase but does not appear to increase as quickly as completeness decreases. There appears to be more mutual data within the clusters when k increases.**

**PCA on the coverage Database**

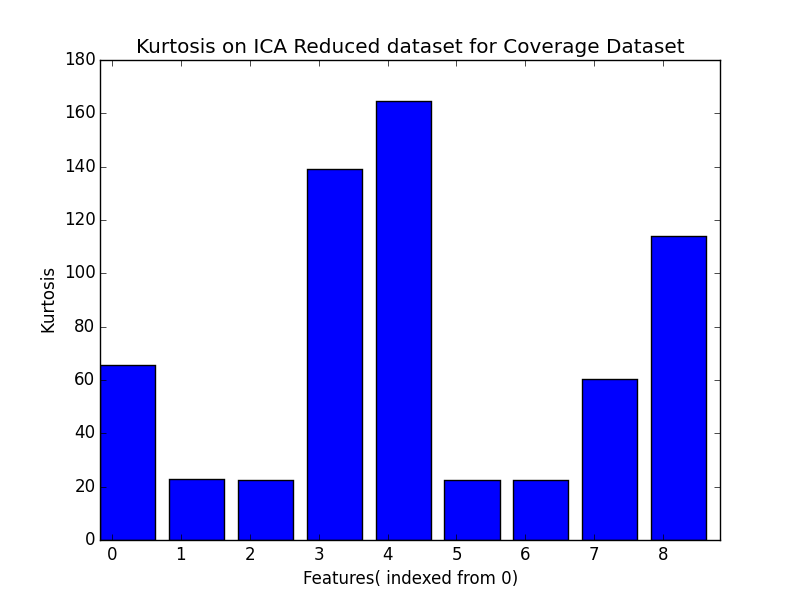
PCA time homo compl v-meas ARI AMI

K-means 1.02s 0.272 0.301 0.286 0.154 0.272

EM PCA 0.03s 0.266 0.308 0.285 0.150 0.265



According to the figure, the 54 features from the dataset only contained 9 features with any bit of variance. Thus, PCA helped conclude how many features should be extracted by the algorithms (this includes the datasets for ICA and RCA as well). The features that had nearly 0 variance were the binary features because they contained a lot of 0’s for their values, and this did not add variance to the dataset.

 **ICA on the Coverage Dataset**

ICA time homo compl v-meas ARI AMI clusters

k-means 0.73s 0.081 0.223 0.119 0.014 0.080 10

EM 0.03s 0.003 0.148 0.006 0.000 0.003 3

**ICA performed the worst out of the 4 transformation algorithms. If the kurtosis is looked at, it appears that the data either does not express any independent variables or struggles to find the distribution. The ICA reduced data appears to contain very leptokurtic distributions. It is important to note that on EM, ICA could only make 3 clusters. This is due to how the ICA algorithm computed and transformed the data. The ICA algorithm had failed to categorize 9 independent variables for most the instances within the dataset. It appears that instances that don’t fit the algorithms process of extracting independent variables simply become arrays of 0s. Changing the values of independent variables did not improve the results.**

time homo compl v-meas ARI AMI

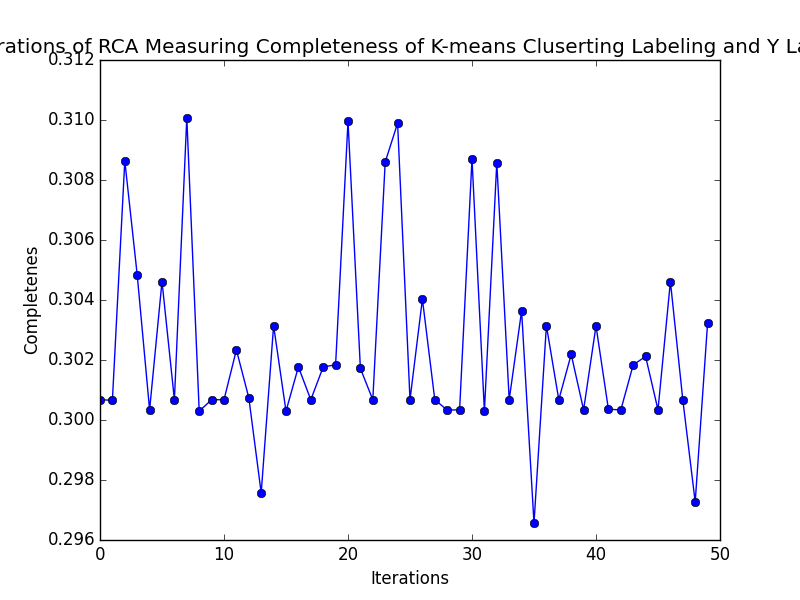
k-mean 1.07s 0.273 0.300 0.286 0.154 0.272

EM 0.03s 0.265 0.310 0.286 0.153 0.265

**RCA on the Coverage Dataset**

RCA was able to replicate PCA performance. RCA was iterated 50 times, and the completeness was computed to find the best set of coordinates.

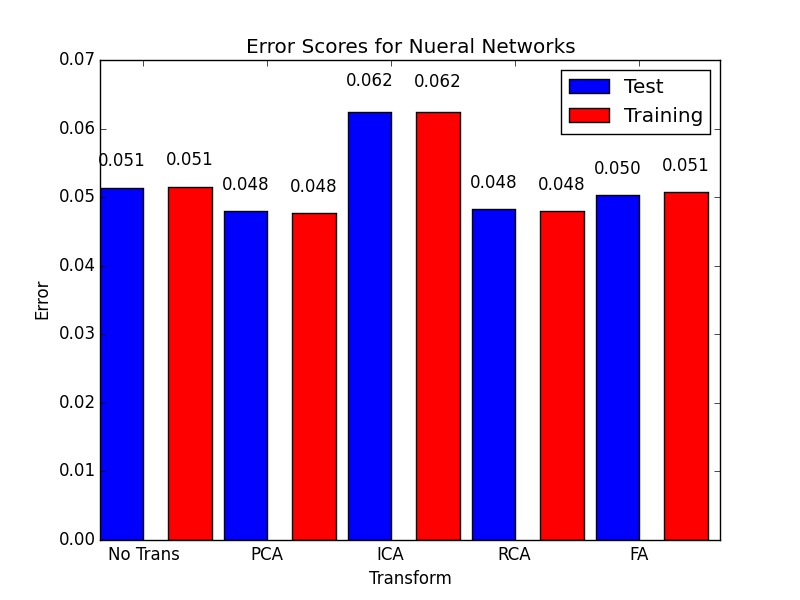
time homo compl v-meas ARI AMI

k-means 0.89s 0.278 0.392 0.325 0.141 0.277

EM 0.02s 0.267 0.531 0.355 0.126 0.266

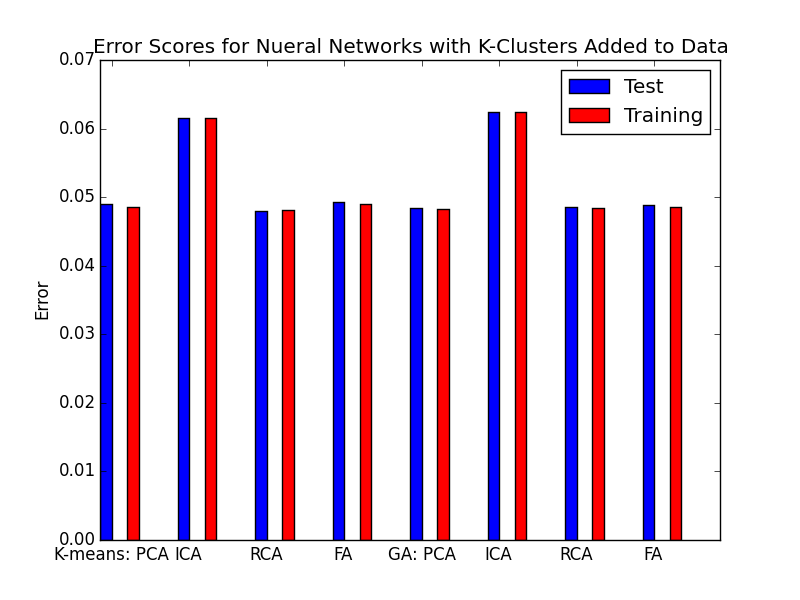
**FA on the Coverage Dataset**

Rather than computing individual variances and projecting them in a different space to maximize them, factor analysis takes the covariance of all the features and tries to find factors that relate to the data. Those factors will then be used instead of the features. FA on EM seemed to perform very well. Creating these ‘feature’ variables seems to work well with the coverage database because once the data is scaled it retains similar variances. Specifically, 3 features were found and were used as the new data for this transformation. By adding more factors, the measurements returned worse results, therefore it appears there are 3 factors that the cover type data set expresses. Clustering with EM gave only 8 values for k. This shows that the 3 factors express 8 different variations.



The results show that all of the transformed data performed faster than the base data, but the ICA data was the only set that did not perform better results. The Neural Network had 5 hidden nodes and a 7 node output layer. Ten epochs were run on the network, and then 50% of the data was allocated to the testing data. The testing error was computed and returned.

Clocking times: no transform: 46.9745259285 PCA: 43.8967571259 ICA : 44.7598409653 RCA: 44.3045909405 FA:43.112457037



Although no significant gain was found by including the cluster as a feature in the Neural Network, all algorithms lowered their error. Therefore, the clustering did help the classification problem slightly. To identify this case, given that the data has been reduced to more than 75% its original size and not much gain was found from reduction, it appears that the cover dataset does not really suffer from a curse of dimensionality. To further this, when the clustering algorithms are run against the new datasets with the previous clusters included, we see that most of the data is recomputed back into the same clusters.

algorithm time homo compl v-meas ARI AMI

kM PCA 1.15s 1.000 0.912 0.954 0.955 0.912

EM PCA 0.04s 1.000 0.897 0.946 0.923 0.896

k- ICA 0.90s 1.000 0.894 0.944 0.951 0.894

EM ICA 0.04s 1.000 0.899 0.947 0.983 0.899

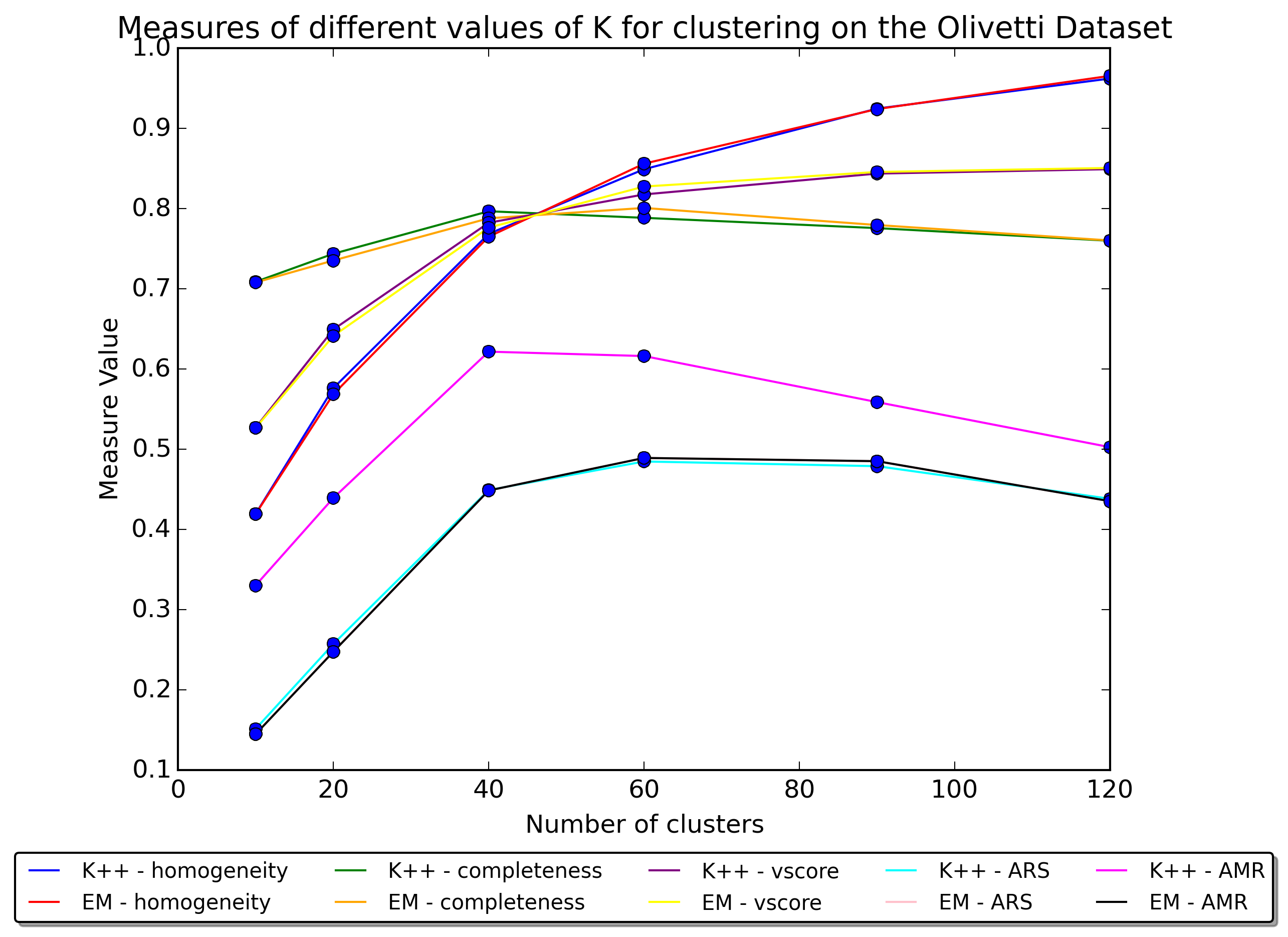
k- RCA 1.28s 0.999 0.915 0.955 0.938 0.914

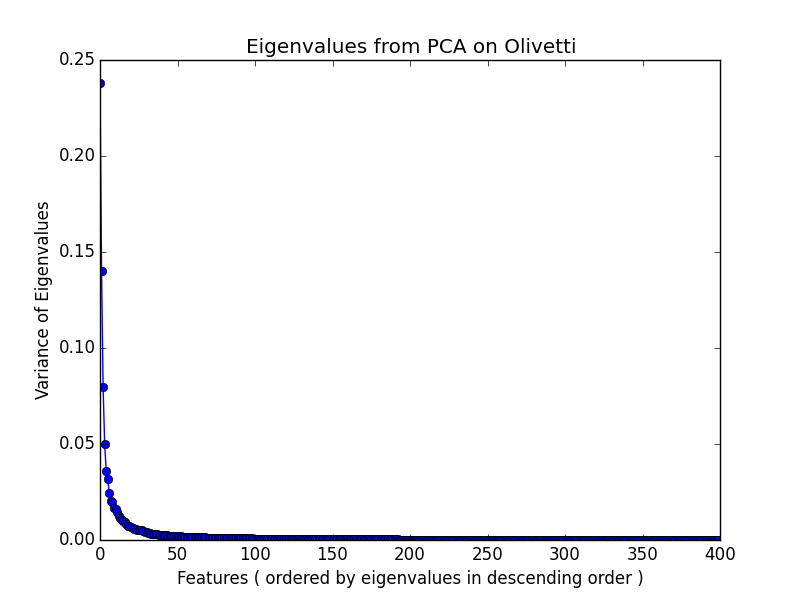
EM RCA 0.04s 1.000 0.877 0.935 0.937 0.877

k- FA 1.06s 1.000 0.903 0.949 0.972 0.902

EM FA 0.02s 1.000 0.866 0.928 0.897 0.866

The homogeneity is nearly perfectly fitted back. Therefore, the algorithms had no problem putting each label into a single cluster.

**On the Olivetta Datasebase**

The measurements in the Olivetta Dataset intersect at around 50 clusters (yet this is assumed as 50 clusters were not run during the analysis). The ARI and AMI values make a strong comeback at 40, which seems to make since given that there are truly 40 distinct faces within the data.

**Olivetta Dataset on PCA**

time homo compl v-meas ARI AMI

k- PCA 0.65s 0.832 0.765 0.797 0.444 0.571

EM PCA 0.01s 0.887 0.823 0.853 0.571 0.678

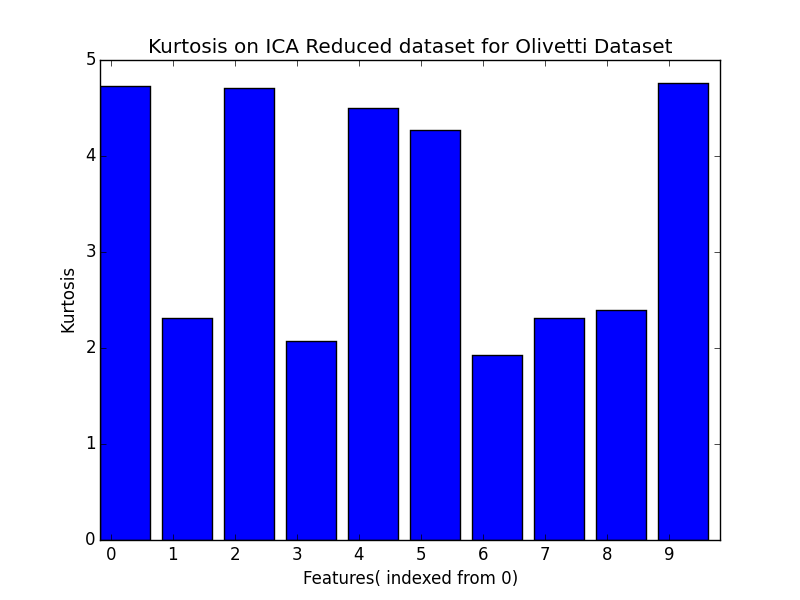
Since there are fewer instances (images) than features (pixels), the eigenvalues are of each instance rather than each feature. PCA seems to perform well with this dataset, but perhaps it would do better if more than 4096 instances were gathered. This number of instances would allow the most variable pixels to be extracted rather than the most variable images. The eigenvalues begin to level out around 40, which is interestingly the number of classifications. Since there are only 400 instances, the most important 10 features seem to carry most of the variance and so were collected for PCA, ICA, and RCA.

**Olivetta Dataset on ICA**

time homo compl v-meas ARI AMI

k- ICA 0.66s 0.841 0.779 0.809 0.469 0.598

EM ICA 0.01s 0.853 0.810 0.831 0.490 0.660

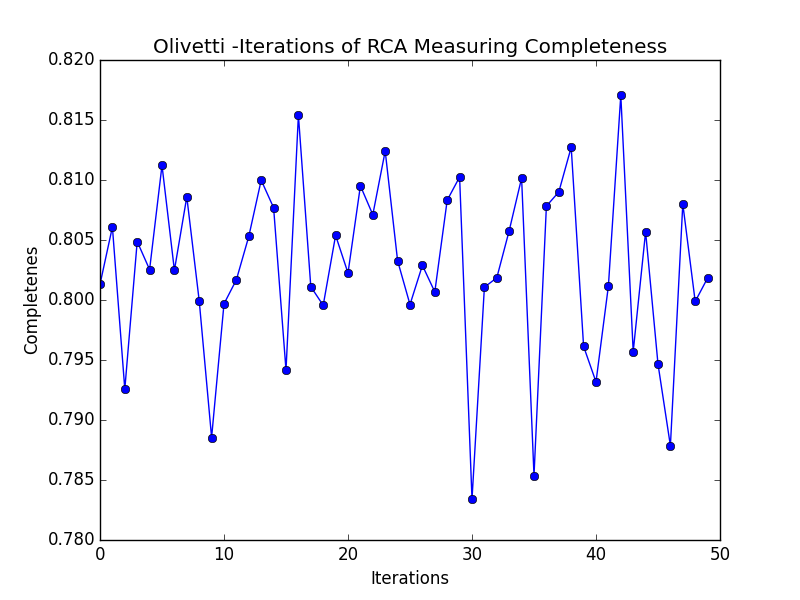


Since the kurtosis of a normal distribution is 3, it is evident that the ICA dataset lingers around a normal distribution. In practice, however, most independently constructed distributions are not going to be normal distributions. Compared to the last dataset’s findings, this dataset seems to have much better results on ICA.

**Olivetta Dataset on RCA**

time homo compl v-meas ARI AMI

k- RCA 0.65s 0.821 0.760 0.789 0.424 0.564

****EM RCA 0.01s 0.877 0.808 0.841 0.544 0.651

RCA once again gives similar results to that of PCA, but not as close as the last dataset. This is probably due to there being many more features for the algorithm to randomly step through.

**Olivetti Data on FA**

time homo compl v-meas ARI AMI

k- FA 0.94s 0.890 0.832 0.860 0.587 0.697

EM FA 0.02s 0.904 0.852 0.877 0.638 0.733

FA had difficulties with this dataset when the number of selected factors was set to 10. Forty clusters were used in this analysis and gave better results. Therefore, it can be deduced that there are many covariant factors that need to be considered when analyzing this data set as opposed to a few.

**Conclusions**

In general, it appears that unsupervised learning can express ways of redefining data that can perform better in supervised learning settings. Despite the results from the measurements, the cover data set went through the process of feature transformation and performed with better results on its supervised classifications. It appears that even with small correlations to information gain, feature transformation and even k clustering features can help performance. K clustering can even go beyond the scope of the labels of the data, which should explain what the data truly represents. For example, the cover data set showed better results with a higher K than the labeling. Perhaps there is noise in the gathering of this data, and its labels should be relabeled, or other features need to be measured for better results. ICA had a very difficult time in categorizing the covertype dataset, but did quite well on the Olevitti dataset. Therefor, it is evident that some datasets are better represent removable independent variables. It seems that when the dataset has a wide variance in its features kurosis values that ICA has a harder time extracting these independent variables. Overall, on both datasets RCA seemed to perform just as well as PCA, which shows that randomized iterations of orthogonal projections on the data appear to give good results overall, but iterations need to be done to find the best results. FA shows that covariance of the features can create interesting factors that can be used to express the original dataset. Overall, FA performed very well on both dataset due to the datasets show similar covariance wthin its featuers this is especially more apparent in the covertype dataset. Overall EM gave better results than K mean on these datasets. Therefore, both datasets are better fit when white-noise is represented within the model.

1. Website has been taken offline but sklearn refers to this website for additional information http://www.cs.nyu.edu/~roweis/ [↑](#footnote-ref-1)