Unsupervised Learning and Dimensionality Reduction

Brittany N. Tan
Georgia Institute of Technology, brittanytan@gatech.edu

***Abstract* – This project includes analyses on different unsupervised learning and dimensionality reduction algorithms on two different datasets, the spam e-mail dataset which contains 4601 instances and 58 attributes and the diabetes dataset which contains 768 instances and 9 attributes. The algorithms implemented include *k*-Means Clustering, Expectation Maximization, Principal Component Analysis (PCA), Independent Component Analysis (ICA), Randomized Projections, and Feature Selection based on Information Gain.**

K-MEANS CLUSTERING

This algorithm separates the data into *k* number of clusters with randomly picked centers and minimizes the mean squared distance from each data point when clustering [1]. The algorithm was implemented using both datasets with varying amounts of clusters and the within clusters sum of squared errors (SSE) was analyzed. This value represents the sum of the squared differences between the group's mean and each observed data point [2]. For the spambase dataset, the optimal cluster was found using the elbow method. This method is a visual method where an "elbow" is identified [3]. This is the point where the SSE stops decreasing as rapidly and begins to plateau more. This elbow is indicated in the graph shown in Figure 1.

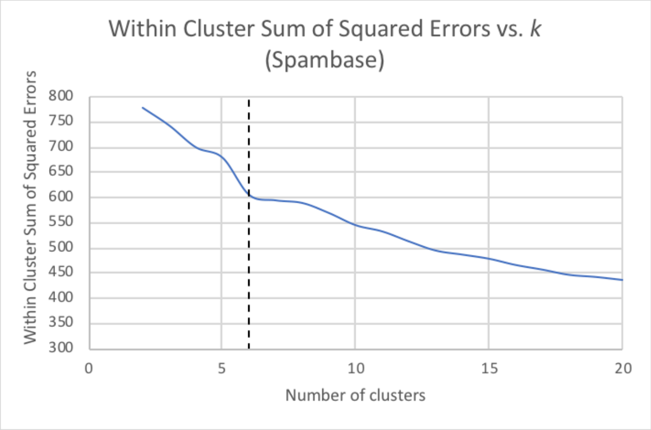


FIGURE 1
SUM OF SQUARED ERRORS VS. NUMBER OF CLUSTERS (SPAMBASE)

There does not exist an unambiguous elbow, however, in the diabetes dataset. For this dataset, our optimal *k* value is chosen using the calculated silhouette width *s(i)*. The *k*-value with the largest *s(i)* is considered to be the optimal [3]. The silhouette width *s(i)* for entity $i \in I$ is defined in (1).

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i)} . \qquad (1)$$

a(i) is the average Euclidean distance between *i* and the other data points in the cluster, and *b(i)* is the minimum of average Euclidean distances between *i* and the other data points. This value is evaluated and indicated in Figure 2.

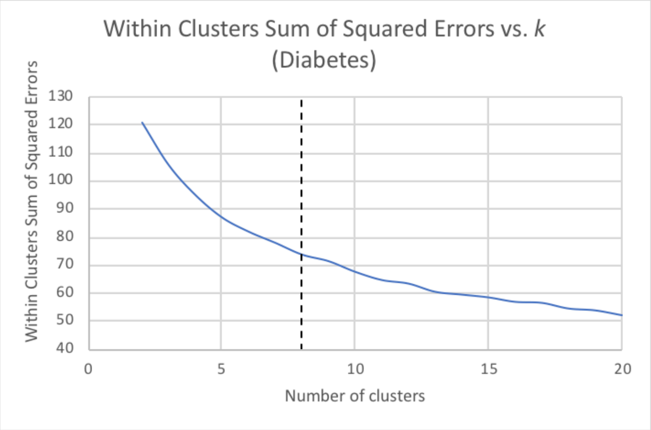


FIGURE 2
SUM OF SQUARED ERRORS VS. NUMBER OF CLUSTERS (DIABETES)

EXPECTATION MAXIMIZATION

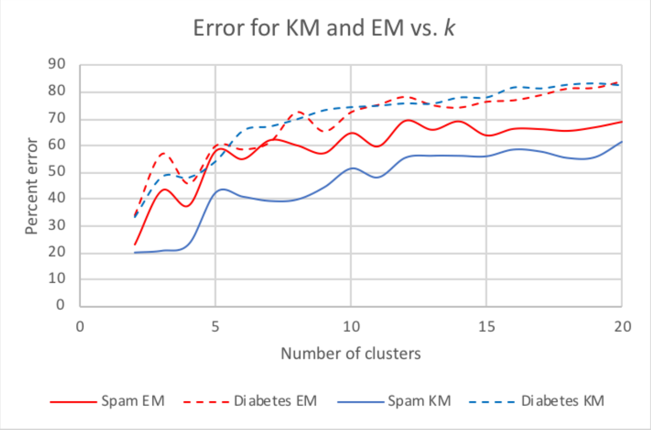


FIGURE 3
ERROR VS. NUMBER OF CLUSTERS FOR K-MEANS AND EXPECTATION MAXIMIZATION

This algorithm is a soft clustering algorithm that alternates between two different probabilistic calculations, expectation and maximization [4]. This was implemented with both datasets, varying the number of clusters between 2 and 20 for each. One observation to note is that this algorithm resulted in a greater percent error than *k*-Means for both datasets. This may be because there may exist outliers or attributes that are not directly pertinent to classifying, and so the domain knowledge was not well-suited for the problems.

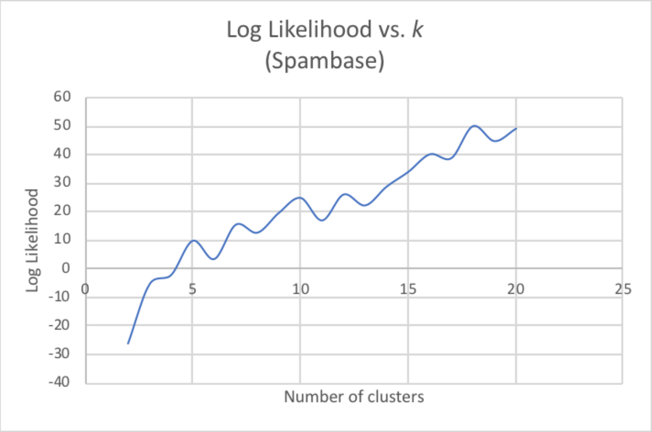


FIGURE 4
LOG LIKELIHOOD VS. NUMBER OF CLUSTERS (SPAMBASE)

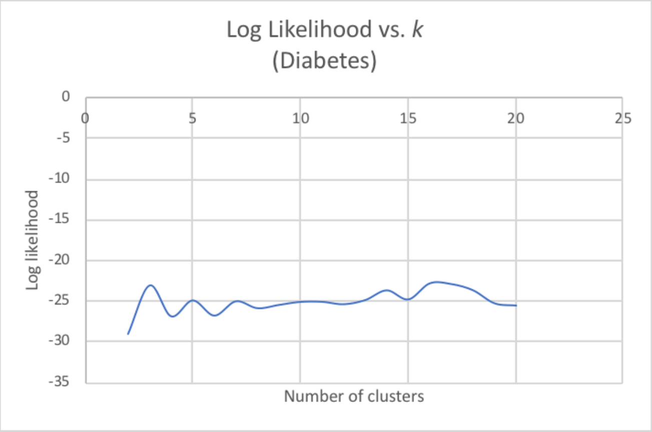


FIGURE 5
LOG LIKELIHOOD VS. NUMBER OF CLUSTERS (DIABETES)

Figures 4 and 5 display the log likelihoods of the data based on number of clusters from 2 to 20. One observation is that the log likelihood experiences an overall increase as the number of clusters increases for the spam e-mail dataset, but remains roughly the same for the diabetes dataset. The behavior for log likelihood for the spam e-mail dataset is as expected. A possible explanation for why this is not the case for the diabetes dataset may be because this problem may not be appropriate for this algorithm. Since expectation maximization is a soft clustering algorithm, it is vulnerable to getting stuck on local optima [5].

Figure 5 and Figure 6 display visual representations of the clusters based on the optimal *k*-values found from *k*-Means clustering in the previous section. The clusters for the spam e-mail dataset are not as evenly distributed as the clusters in the diabetes dataset, which indicates some degree of overfitting. Additionally, the clusters for the diabetes dataset are much more spread out than those in the spam e-mail dataset. This indicates higher variability in the diabetes data.

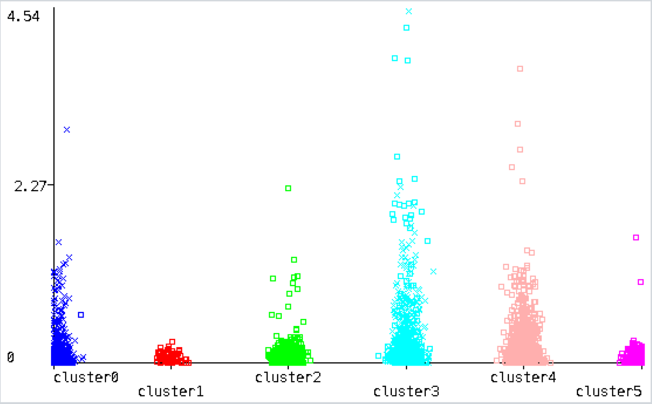


FIGURE 6
VISUALIZATION OF CLUSTERS (SPAMBASE)

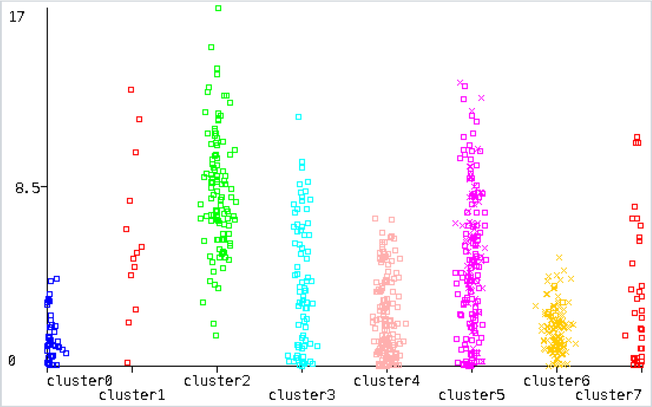


FIGURE 6
VISUALIZATION OF CLUSTERS (DIABETES)

PRINCIPAL COMPONENT ANALYSIS (PCA)

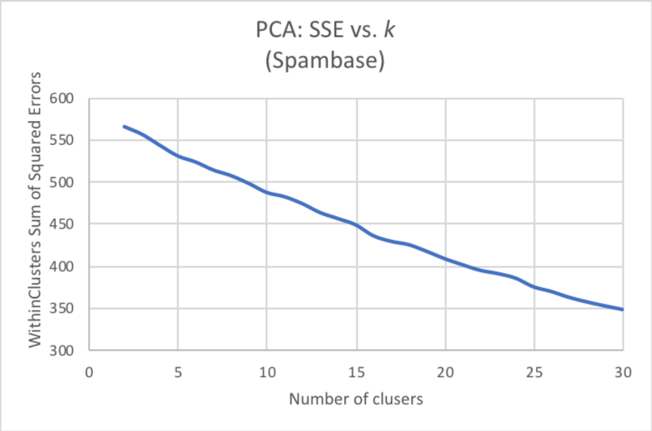


FIGURE 6
PCA: SSE VS. CLUSTERS (SPAMBASE)

Both datasets were filtered using Weka's Principal Components filter with a 0.95 variance and then run using both clustering algorithms discussed before. When applying this filter to the spam e-mail dataset, there no longer exists an unambiguous elbow point in the plot for SSE vs. number of clusters as shown in Figure 7. However, using the silhouette width method, a similar optimum of $k = 5$.

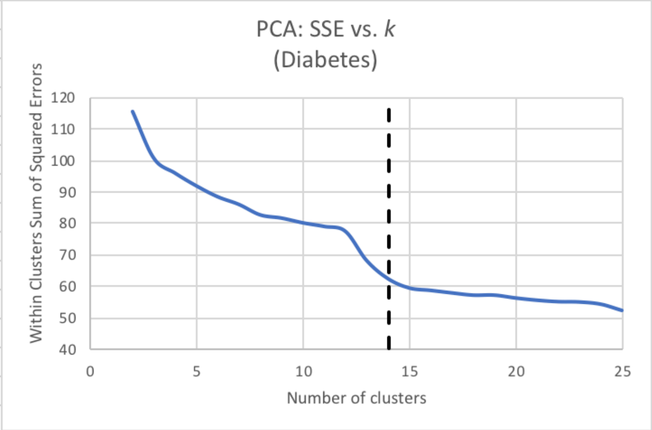


FIGURE 7
PCA: SSE VS. CLUSTERS (DIABETES)

When doing the same to the diabetes dataset, there still exists an elbow in the curve, but it differs a bit from the one found using *k*-Means. This difference in *k*-value is because PCA transforms nominal attributes into binary attributes before filtering the data, therefore adding extraneous attributes [6]. When these attributes are factored, the data ends up being more skewed.

INDEPENDENT COMPONENT ANALYSIS (ICA)

Similarly to PCA, ICA is a feature transformation algorithm, but instead of trying to maximize variance, it tries to maximize independence. Overall, ICA behaves very similarly to PCA with runtimes and accuracies. After running ICA on both datasets, the kurtosis for each independent component was observed. Figure 8 shows kurtoses and accuracies for each attribute of the spambase data after the transformation, and Figure 9 displays the kurtoses for the diabetes dataset.

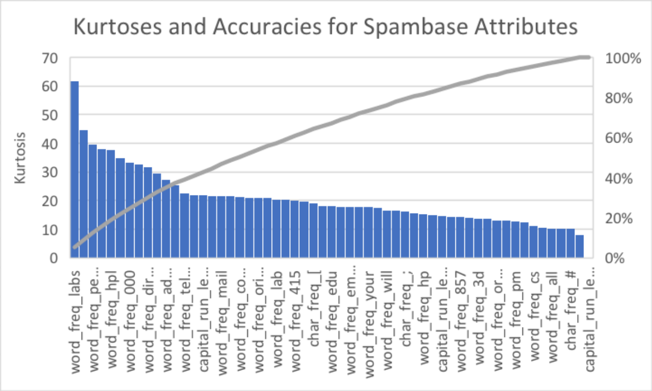


FIGURE 8
KURTOSES AND ACCURACIES FOR SPAMBASE ATTRIBUTES

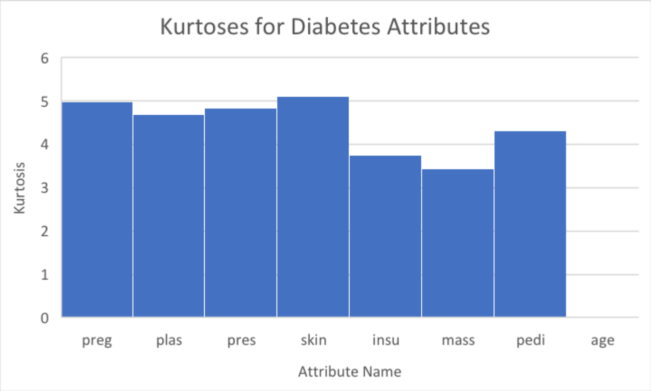


FIGURE 9
KURTOSES FOR DIABETES ATTRIBUTES

Given that the ideal kurtosis of a normal distribution is 3, and all of the kurtoses for the spam e-mail dataset are far from 3, this indicates high levels of mutual independence for this dataset. The kurtoses for diabetes, however, are closer to 3. While these attributes are still mutually independent, they are not as strong and independent as those of spam.

RANDOMIZED PROJECTIONS

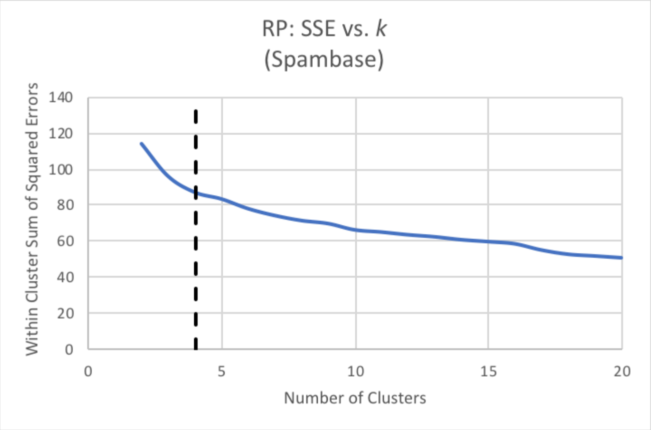


FIGURE 10
RP: SSE VS. CLUSTERS (SPAMBASE)

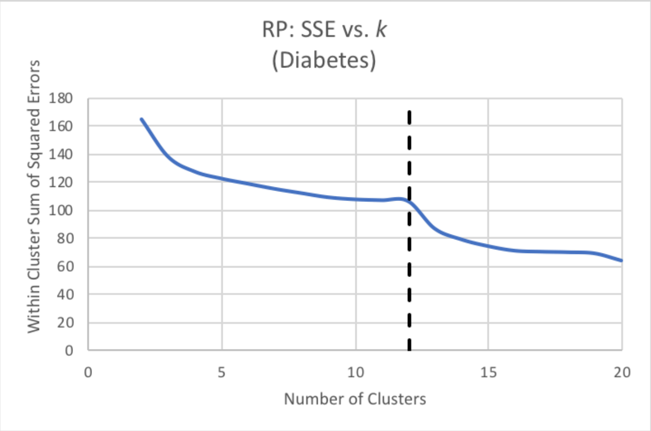


FIGURE 11
RP: SSE VS. CLUSTERS (DIABETES)

This is a dimensionality reduction algorithm that generates a randomized matrix based on a target number of dimensions and uses said matrix to apply the transformation. After applying the transformation to each algorithm, KM and EM were run and compared the same way as before. Figure 10 and Figure 11 display the SSE based on number of clusters. A notable observation to make is that the behaviors of the graphs are very similar to PCA and ICA, and the optimal *k*-values found are also very similar. However, this observation is not necessarily conclusive due to the randomized nature of the algorithm. This randomness is also displayed in Figure 12 with the large spike in log likelihood at 3 clusters. Otherwise, the likelihoods remain fairly consistent.

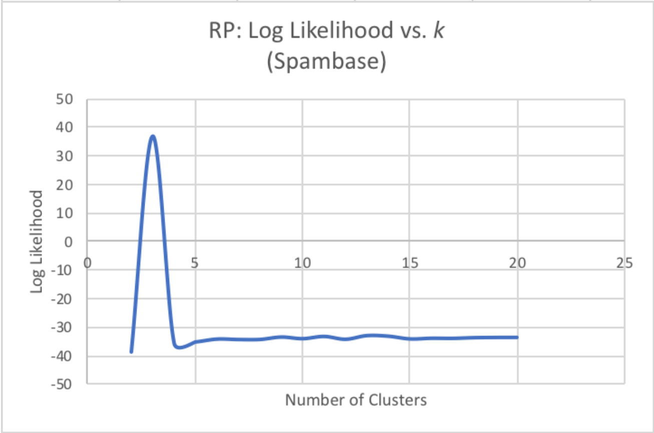


FIGURE 12
LOG LIKELIHOOD VS. NUMBER OF CLUSTERS (SPAMBASE)

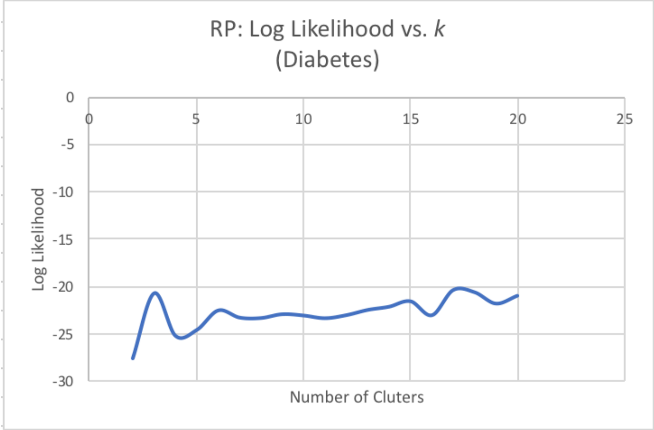


FIGURE 13
LOG LIKELIHOOD VS. NUMBER OF CLUSTERS (DIABETES)

INFORMATION GAIN (IG)

Information Gain is a feature selection algorithm, while PCA, ICA, and RIP are feature transformation algorithms. The information gain is the change in entropy from a prior state to a current state. Overall, as able to be seen in Figure 14, removing the number of attributes does not have any significant effect on performance. This is likely because of the many attributes of the spam e-mail dataset that do not directly pertain to classification. The oscillations in performance may be because as attributes are removed, certain attributes that may be skewing the data are no longer affecting it. It also may simply be due to the fact that all attributes are numerical except for the one nominal class attribute.

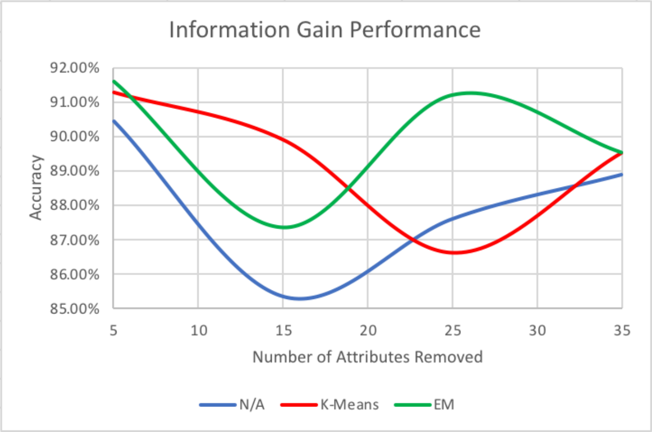
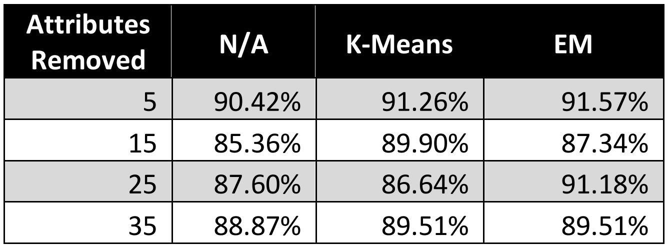


FIGURE 14
INFORMATION GAIN ACCURACIES BASED ON NUMBER OF ATTRIBUTES REMOVED

NEURAL NETWORK LEARNING

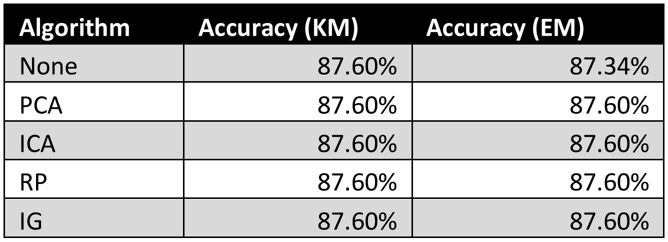
Neural networks were run on the spam e-mail datasets. As shown in Table I of accuracies of neural networks based on different dimensionality reduction algorithms, the accuracies only differ marginally, so there are not any benefits from using the dimensionality reduction. Even with clustering as shown in Table II, the results are not any different.

TABLE I
PERFORMANCE OF NEURAL NETWORKS



Attributes Removed	N/A	K-Means	EM
5	90.42%	91.26%	91.57%
15	85.36%	89.90%	87.34%
25	87.60%	86.64%	91.18%
35	88.87%	89.51%	89.51%

TABLE II
PERFORMANCE OF NEURAL NETWORKS WITH CLUSTERING



Algorithm	Accuracy (KM)	Accuracy (EM)
None	87.60%	87.34%
PCA	87.60%	87.60%
ICA	87.60%	87.60%
RP	87.60%	87.60%
IG	87.60%	87.60%

Although the transformations and selections do not have any significant effect on this dataset, it is not necessarily indicative of the algorithms having little overall effect just because they are not significant on this dataset.

References

[1] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman and A. Wu, "An efficient k-means clustering algorithm: analysis and implementation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881-892, 2002.

[2] G. Milligan and M. Cooper, "An examination of procedures for determining the number of clusters in a data set", *Psychometrika*, vol. 50, no. 2, pp. 159-179, 1985.

[3] T. Kodinariya1 and P. Makwana, "Review on determining number of Cluster in K-Means Clustering", *International Journal of Advance*

[4] T. Moon, "The expectation-maximization algorithm", *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47-60, 1996.*Research in Computer Science and Management Studies*, vol. 1, no. 6, pp. 90-95, 2013.

[5] S. Wold, K. Esbensen and P. Geladi, "Principal component analysis", *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1-3, pp. 37-52, 1987.

[6] H. Oja and K. Nordhausen, "Independent Component Analysis", *Encyclopedia of Environmetrics*, 2013.

Author Information

Brittany N. Tan, Undergraduate Student, College of Computing, Georgia Institute of Technology.