CS7641 A3: Unsupervised Learning and Dimensionality Reduction

Scott Schmidl
sschmidl3@gatech.edu

I. Introduction

KMeans Clustering (KM), Gaussian Mixture Model (GMM), Principal Component Analysis (PCA), Independent Component Analysis (ICA), Sparse Randomized Projection (SRP) and Hessian Locally Linear Embedding (HLLE) Manifold Learning represent an array of unsupervised learning techniques in the field of machine learning. Each of these methods plays a crucial role in uncovering patterns, reducing dimensions, and extracting meaningful information from complex datasets.

In this paper, I delve into the applications and comparative analyses of these key methods over two different sets of data. I begin with KMeans clustering and Gaussian Mixture Model before exploring PCA, ICA and SRP, and HLLE. Finally, I combine clustering and dimensionality reduction with neural networks to evaluate the differences in performance of these techniques. I aim to provide insight into their respective strengths, weaknesses, and applications.

Datasets

For the first dataset, I use cardiovascular disease (CVD) factors to predict whether a person has CVD or not given their height, weight, blood pressure, and age. The metric chosen to represent the performance of the model for the first dataset is recall, which helps identify true positives and false negatives.

For the second dataset, I use a nutrition facts to predict food groups given their level of protein, carbohydrate, cholesterol, water, and sodium. The metric chosen to represent the model performance for dataset two is balanced accuracy (BA) due to the data being imbalanced.

II. Step 1

My hypothesis regarding clustering algorithms is that they will effectively partition the data into distinct groups. Given the inherent structure within these two datasets, Gaussian Mixture Model and KMeans will successfully identify latent patterns present in the data.

GMM assigns each data point to each cluster with a certain probability, using the EM algorithm to maximize the likelihood of the data. KMeans partitions the data into K clusters by iteratively updating the positions of K centroids to minimize the sum of squared distances within the cluster.

The Calinski-Harabasz score which calculates the proportion of dispersion between clusters to dispersion within clusters. BIC which assesses the complexity of a model by imposing penalties on models with increased parameters. Inertia which quantifies the total squared distances from each sample to its designated centroid within the cluster. AMIS and ARS which measure the agreement between the clustering labels and the ground-truth labels. Finally, the cluster center scatter plots show the centroids of each cluster in the feature space.

Cardiovascular Disease

A. Expectation Maximization

In figure 1(a) BIC is analyzed, values decrease as the number of clusters increases, indicating that the model fit improves with more clusters. The slight zigzags observed suggest fluctuations in the model complexity or overfitting at different numbers of clusters. The significant decrease in BIC from 2 to 12 clusters suggests that adding more clusters improves the fit of the model substantially. Hence, 12 clusters indicates the best balance between model complexity and fit. The selection of 12 clusters implies that the data exhibits a complex structure that requires a relatively large number of clusters to capture effectively.

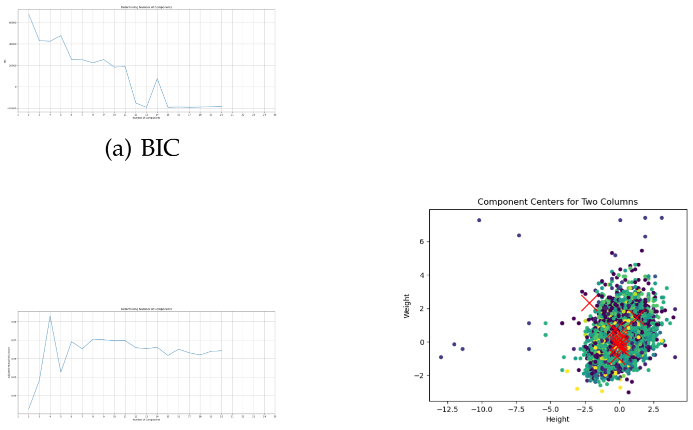
In figure 1(b), AMIS was analyzed to help compare the 12 clusters to the ground truth labels, of which there are two. The AMIS increases from 0.03 to 0.08 as the number of clusters increases from 2 to 4, suggesting improved agreement between clustering and results. The increase in AMIS indicates that additional clusters capture more nuances or variations in the data, leading to better clustering performance. The AMIS dropping to 0.05 at 5 clusters suggests that adding another cluster beyond 4 does not significantly improve the agreement with the ground truth labels. It indicates that the KMeans may be overfitting. The zigzags around 0.65 from 5 to 20 clusters suggests instability in the clustering performance.

In figure 1(c), I plot the component centers. With 12 clusters, GMM can most likely be seen to overfit, capturing noise or minor variations as separate clusters. GMM is struggling to assign data points to appropriate clusters, leading to overlapping cluster centers. To help reduce these issues and better represent my data I should explore a significant reduction in the number of clusters to help capture the essential structure of the data more effectively or explore different preprocessing steps.

GMM has struggled to produce well-defined and meaningful clusters. The overlap observed in the cluster centers suggests that the clusters identified by GMM do not accurately capture the underlying structure of the data. It would be beneficial to explore alternative clustering algorithms, other preprocessing techniques, perform better tuning to address these issues and obtain more robust clustering results. A value of 12 clusters was chosen as optimal for this model, as explained above.

B. KMeans Algorithm

In figure 2(a), the Calinski score is analyzed. The decrease in the Calinski score as the number of clusters increases suggests that the clusters become less well defined or separated as more clusters are added. This decrease in score indicates that the



(a) BIC

(b) AMIS

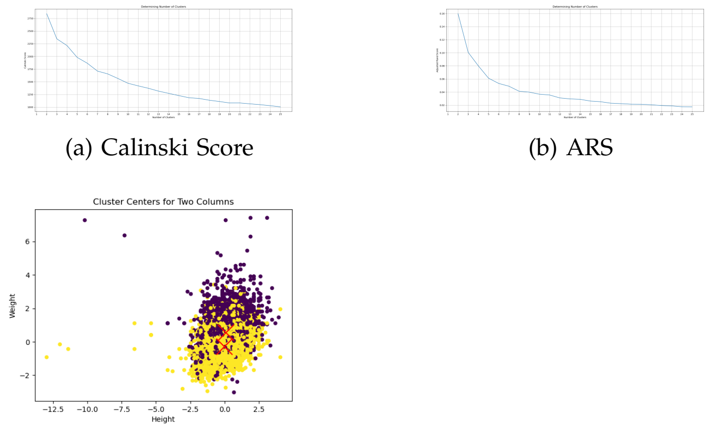
(c) Cluster Centers

Fig. 1: CVD Gaussian Mixture Method

clustering algorithm is overfitting the data or capturing noise as separate clusters, leading to less cohesive clusters.

In figure 2(b), the Adjusted Rand score is analyzed. The decrease in the Adjusted Rand Score as the number of clusters increases suggests that the clustering results become less consistent with the true labels, which are two. This decrease in score indicates that the clusters identified by KMeans may not align well with the underlying structure or patterns in the data as the number of clusters increases.

In figure 2(c), I plot the cluster centers. The overlap in cluster centers observed when plotting two clusters suggests that these clusters are not well-separated in the feature space. This overlap indicates that the centroids of these clusters are close together, making it difficult to distinguish them based solely on their centroids. It suggests that the clusters may be similar in terms of their feature distributions or that KMeans may not have effectively partitioned the data into distinct groups.



(a) Calinski Score

(b) ARS

(c) Cluster Centers

Fig. 2: CVD KMeans Algorithm

It clear to see that KMeans struggles to produce well-defined and meaningful clusters beyond a certain two clusters. The overlap observed in the cluster centers suggests that the clusters identified by KMeans may not accurately capture the underlying structure of the data. It would be beneficial to explore alternative clustering algorithms, other preprocessing techniques, own better tuning to address these issues and obtain more robust clustering results. Additionally, further investigation of the data's characteristics after clustering may provide insight into the appropriate number of clusters. A value of 2 clusters was chosen as optimal for this model, as explained above.

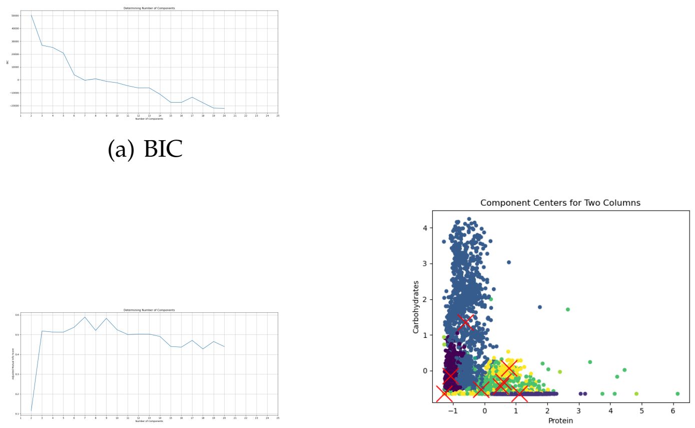
Nutrition Facts

C. Expectation Maximization

In figure 3(a) BIC was analyzed. The significant decrease from 2 to 8 clusters suggests that adding more clusters significantly improves the fit of the model. The decrease that starts to plateau after 8 clusters, indicates diminishing returns in model improvement beyond this point.

In figure 3(b) AMIS was analyzed. The increase from 2 to 7 clusters suggests that adding more clusters initially improves the agreement with true labels; despite there only being five true labels. The drop in AMIS after 7 clusters suggests that adding more clusters beyond this point leads to a decrease in agreement with true labels, indicating overfitting or capturing noise as separate clusters.

In figure 3(c) cluster centers is plotted. The analysis of cluster centers at 8 clusters reveals some separation between clusters but also some overlap, indicating that the clusters are not perfectly distinct. The overlap suggests that some clusters may have similar centroids or that the data points within these clusters have overlapping feature distributions.



(a) BIC

(b) AMIS

(c) Cluster Centers

Fig. 3: NF Gaussian Mixture Method

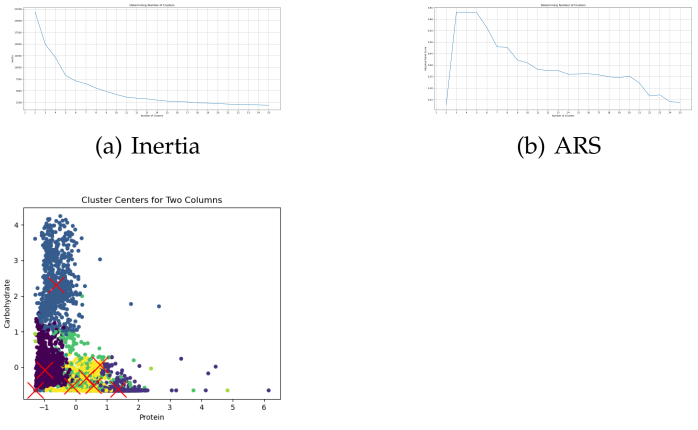
These results suggest that GMM would benefit from further exploration of clustering parameters, or alternative clustering algorithms may be more beneficial to improve performance. A value of 8 clusters was chosen as optimal for this model, as explained above.

D. KMeans Algorithm

In figure 4(a), Inertia was analyzed. The significant decrease in inertia from 2 to 8 clusters suggests that adding more clusters improves the compactness of the clusters and reduces the variance within the cluster. Beyond 8 clusters, the decrease in inertia slows down, indicating diminishing returns in cluster compactness.

In figure 4(b), ARS was analyzed. The increase in ARS from 2 to 3 clusters suggests that adding more clusters initially improves the agreement with true labels. However, the drop in ARS after 5 clusters suggests that adding more clusters beyond this point leads to a decrease in agreement with true labels, indicating overfitting or capturing noise as separate clusters.

In figure 4(c), I plotted the cluster centers. The analysis of cluster centers at 8 clusters shows some separation between clusters but also some overlap. The overlap suggests that some clusters may have similar centroids or that the data points within these clusters have overlapping feature distributions.



(a) Inertia

(b) ARS

(c) Cluster Centers

Fig. 4: NF KMeans Algorithm

These results suggest that KMeans would benefit from further exploration of clustering parameters or alternative clustering algorithms may be more beneficial to improve performance. A value of 8 clusters was chosen as optimal for this model, as explained above.

III. Step 2

My hypothesis with regard to dimensionality reduction techniques is that they will effectively capture the underlying structure of the data. This hypothesis posits that these techniques will efficiently extract and preserve the most informative features from the datasets while discarding redundant or noisy attributes.

PCA performs eigenvalue decomposition on the data and projects it to a lower dimension preserving maximum variance. ICA is a technique that aims to separate a multivariate signal into additive, statistically independent components by assuming that the observed data are independent and non-Gaussian. SRP is a technique that uses random projection matrices to project data to a lower dimensional space. Hessian LLE is a nonlinear-dimensionality reduction technique that embeds data into a lower-dimensional space by approximating the local geometry by fitting a quadratic surface.

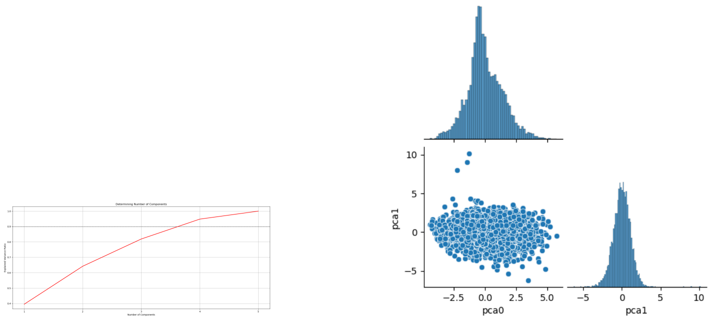
The explained variance ratio indicates the proportion of variance in the original data explained by each principal component. Kurtosis is a measure of the ”tailedness” of the probability distribution of a dataset, where high kurtosis indicates non-Gaussianity in the projected data. The reconstruction error measures the difference between the original data and the data reconstructed. A pair plot of two components visualizes the data in the reduced space which helps in understanding the relationships between data points in the reduced space.

Cardiovascular Disease

A. Principal Component Analysis

In figure 5(a), explained variance ratio was analyzed. The rise in explained variance ratio from 0.4 to 0.82 at 3 components and from 0.82 to 0.94 at 4 components indicates that adding more components captures a larger portion of the variance in the data. This suggests that the first few principal components explain a significant amount of the variance in the dataset, with additional components capturing smaller amounts of variance.

In figure 5(b), a pair plot of the first two components was used to evaluate components. The pair plot of components one and two shows that both components are Gaussian distributed, indicating that they capture similar patterns or structures in the data. However, the pair plot also shows that, when plotted together, the data points form a horizontal blob, suggesting that the two components are not correlated.



(a) Explained Variance Ratio

(b) Component Pair Plot

Fig. 5: CVD PCA Algorithm

A value of 3 components was chosen as optimal for this model, as explained above.

B. Independent Component Analysis

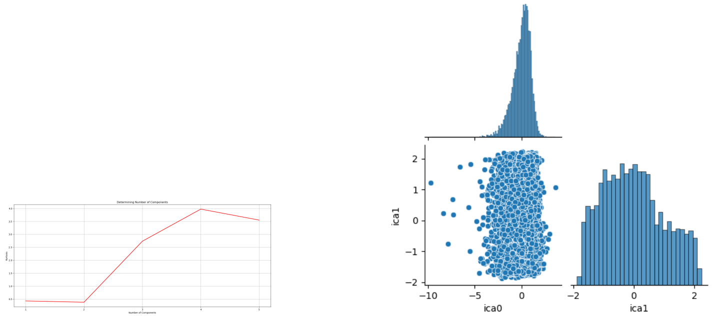
In figure 6(a), kurtosis was analyzed. The steady value in kurtosis from 1 to 2 components suggest some Gaussian structure in the data. A rise from 2 to 4 components, suggest a non-Gaussian structure, while a drop at 5 components, suggests a slight change in the non-Gaussian structure captured by the independent components. The initial rise in kurtosis indicates an increase in non-Gaussianity as more components are added, while the subsequent drop suggests a decrease in non-Gaussian structure beyond four components.

In figure 6(b), a pair plot of the first two components was used to evaluate components. The pair plot of two components shows that one component is Gaussian, while the other has a bit of a right tail, indicating some non-Gaussian structure. However, when plotted together, the data points form a vertical blob, suggesting no correlation and some independence.

A value of 4 components was chosen as optimal for this model, as explained above.

1) Sparse Random Projections: In figure 7(a), reconstruction error was analyzed. The decrease in reconstruction error from 0.9 to 0.6 from 1 to 3 components, from 0.6 to 0.2 at 4 components, and from 0.2 to 0 at 5 components indicates that adding more components improves the quality of the dimensionality reduction. Lower reconstruction errors suggest that the lower-dimensional representation effectively captures the essential information from the original data.

In figure 7(b), a pair plot was used to analyze the results. The pair plot of two components shows that both components

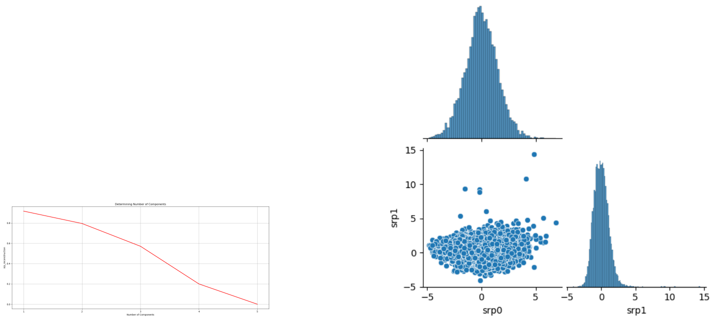


(a) Kurtosis

(b) Component Pair Plot

Fig. 6: CVD ICA Algorithm

are Gaussian distributed, indicating that they capture similar patterns or structures in the data. However, when plotted together, the data points form a horizontal blob, suggesting that the two components are not correlated.



(a) Reconstruction Error

(b) Component Pair Plot

Fig. 7: CVD SRP Algorithm

A value of 4 components was chosen as optimal for this model, as explained above.

2) Hessian Locally Linear Embedding: In figure 8(a), reconstruction error was analyzed. The decrease in reconstruction error from -0.25 to -2.5 from 2 to 4 components and from -2.5 to -4.25 from 4 to 5 components indicates that adding more components improves the quality of the reduction of dimensionality. Lower reconstruction errors suggest that the lower-dimensional representation better preserves the local relationships in the original data.

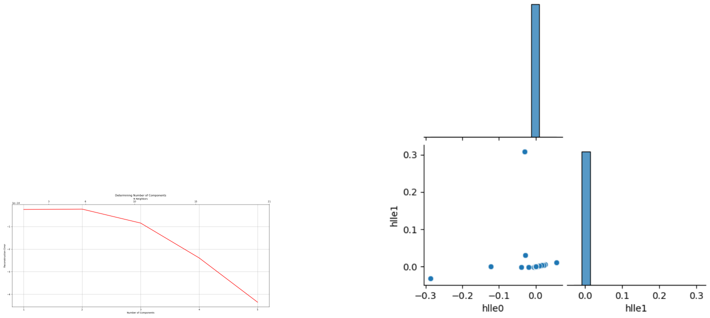
In figure 8(b), a pair plot was used to evaluate the components. The pair plot of two components shows that both components appear as straight vertical bars at 0.0, suggesting that they do not capture meaningful variation in the data at this scale, but that zooming in on the axis might show more information. When plotted together, the data points show no discernible structure, indicating there is no correlation between components.

A value of 4 components was chosen as optimal for this model, as explained above.

Nutrition Facts

C. Principal Component Analysis

In figure 9(a), explained variance ratio was analyzed. The rise in explained variance ratio from 0.3 to 0.82 at 3 components and from 0.82 to 0.98 at 4 components indicates that adding more components captures a larger portion of the variance in the data. This suggests that the first few principal components explain a significant amount of the variance in the dataset, with additional components capturing smaller amounts of variance.

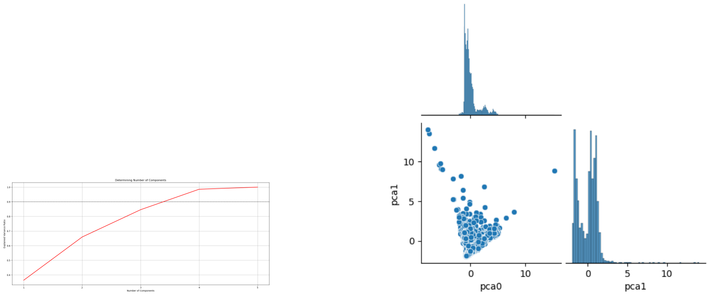


(a) Reconstruction Error

(b) Component Pair Plot

Fig. 8: CVD HLLE Algorithm

In figure 9(b), a pair plot of the first two components was used to evaluate components. The pair plot of components one and two shows that both components have a distribution that is non-Gaussian, indicating that they don't capture similar patterns or structures in the data. However, the pair plot also shows that, when plotted together, the data points seem to have a negative correlation to zero, and then positive correlation to 15. This suggests some nonlinear relationship between the two components.



(a) Explained Variance

(b) Component Pair Plot

Fig. 9: NF PCA Algorithm

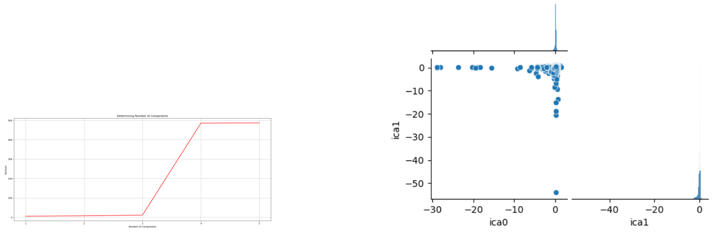
A value of 3 components was chosen as optimal for this model, as explained above.

D. Independent Component Analysis

In figure 10(a), kurtosis was analyzed. The steady value in kurtosis from 1 to 3 components suggests some Gaussian structure in the data. A rise from 3 to 4 components, suggests a non-Gaussian structure, while steadiness from 4 to 5 components, suggests no change in the non-Gaussian structure captured by the independent components.

In figure 10(b), a pair plot of the first two components was used to evaluate components. The pair plot of two components shows that both components migth have a Gaussian distribution. However, when plotted together, the data points form a ”cornered” shape, suggesting no correlation and some independence.

A value of 4 components was chosen as optimal for this model, as explained above.

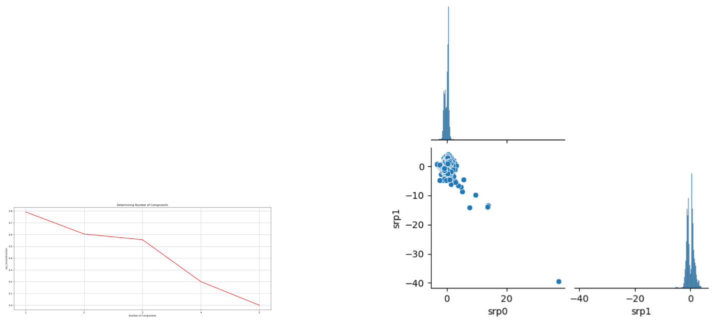


(a) Kurtosis
(b) Component Pair Plot

Fig. 10: NF ICA Algorithm

1) Sparse Random Projections: In figure 11(a), reconstruction error was analyzed. The decrease in reconstruction error from 0.8 to 0.55 from 1 to 3 components, from 0.55 to 0.2 at 4 components, and from 0.2 to 0 at 5 components indicates that adding more components improves the quality of the dimensionality reduction. Lower reconstruction errors suggest that the lower-dimensional representation effectively captures the essential information from the original data.

In figure 11(b), a pair plot was used to analyze the results. The pair plot of two components shows that both components have a Gaussian-like distribution, but with two humps. This indicates that they capture somewhat similar patterns or structures in the data. However, when plotted together, the data points show a negative correlation, suggesting some more investigation is required.



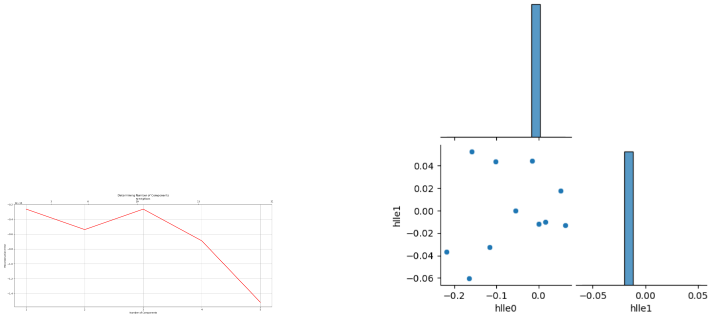
(a) Reconstruction Error
(b) Component Pair Plot

Fig. 11: NF SRP Algorithm

A value of 4 components was chosen as optimal for this model, as explained above.

2) Hessian Locally Linear Embedding: In figure 12(a), reconstruction error was analyzed. The decrease in reconstruction error from -0.25 to -0.55 from 1 to 2 components and from -0.25 to -1.5 from 3 to 5 components indicates that adding more components improves the quality of the reduction of dimensionality. The increase in error from 2 - 3 components is concern and should be investigated.

In figure 12(b), a pair plot was used to evaluate the components. The pair plot of two components shows that both components appear as straight vertical bars at 0.0, suggesting that they do not capture meaningful variation in the data at this scale, but that zooming in on the axis might show more information. When plotted together, the data points show no discernible structure, indicating there is no correlation between components.



(a) Reconstruction Error
(b) Component Pair Plot

Fig. 12: NF HLLE Algorithm

A value of 4 components was chosen as optimal for this model, as explained above.

IV. Step 3

My hypothesis is that clustering after dimension reduction process will reveal more meaningful clusters. This hypothesis suggests that by reducing the dimensionality of the data, the clustering algorithms will operate in a more condensed feature space forcing clearer boundaries.

Cardiovascular Disease

A. Sparse Random Projection and KMeans

In figure 13(a), the Calinski score is analyzed. The decrease in the Calinski score as the number of clusters increases suggests that the clusters become less well defined or separated as more clusters are added. This decrease in score indicates that the clustering algorithm is overfitting the data or capturing noise as separate clusters, leading to less cohesive clusters. This lead me to choosing two clusters.

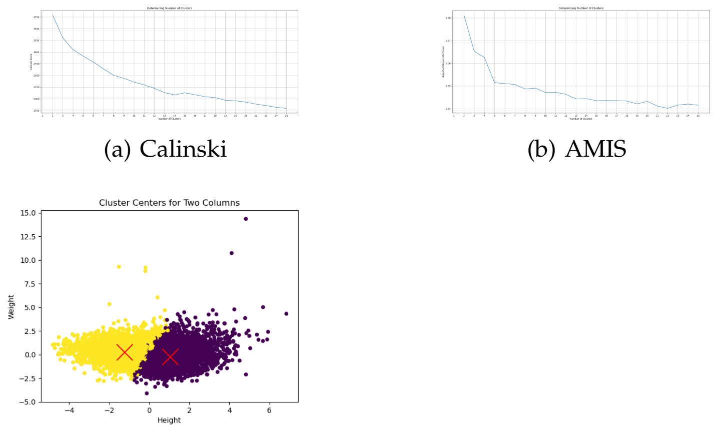
In figure 13(b), AMIS was analyzed to help compare the 2 clusters to the ground truth labels, of which there are two. The AMIS continue to decrease as the number of clusters increase, suggesting reduced agreement between clustering and results. While the max score here is two, the value is still much lower than 1, suggesting that even though two clusters were optimal from Calinski, the clusters did not agree much with the ground truth.

In figure 13(c), I plot the cluster centers. After dimensiona reduction and finding two clusters, this scatter plot looks remarkable. The model appears to have found a good split between the data with only minor overlap at the center. It suggests that KMeans may have effectively partitioned the data into distinct groups, despite there difference from the ground truth.

While the clusters look well-defined, it's worth more investigation to determine in what ways these clusters differ from the ground truth.

B. Hessian Locally Linear Embedding and KMeans

In figure 14(a), Inertia was analyzed. The significant decrease in inertia from 2 to 4 clusters suggests that adding more clusters improves the compactness of the clusters and reduces the variance within the cluster. Beyond 4 clusters, the decrease in inertia slows down, indicating diminishing returns in cluster compactness. This lead me to choosing four clusters.



(a) Calinski

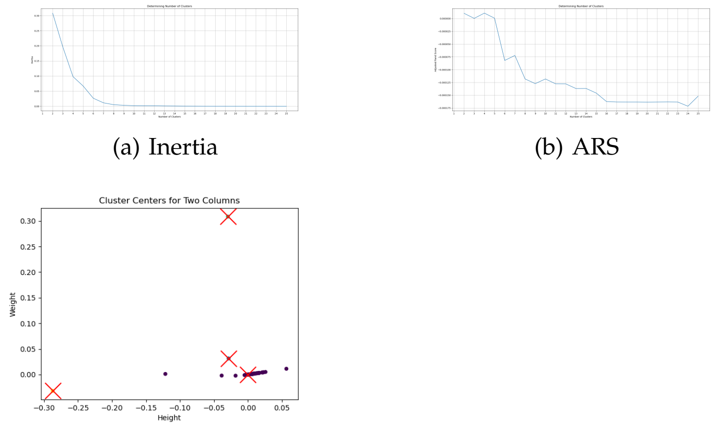
(b) AMIS

(c) Cluster Centers

Fig. 13: CVD KMeans and SRP

In figure 14(b), ARS was analyzed. The zigzag and decreasing pattern suggest that with more clusters there model grows farther away from the ground truth. The peaks are at 2 and 4, with 4 appearing to be slightly higher. After 4, there is marked drop in agreement.

In figure 14(c), I plot the cluster centers. These clusters are far apart and don't appear to be well-defined. Using HLLE doesn't appear to be the best model as it's not capturing meaningful variation in the data.



(a) Inertia

(b) ARS

(c) Cluster Centers

Fig. 14: CVD KMeans and HLLE

While four clusters came out of this analysis, it's worth exploring HLLE some more to address some issues seen in the scatter plot. It could also be the case that there are no nonlinear relationships in the data.

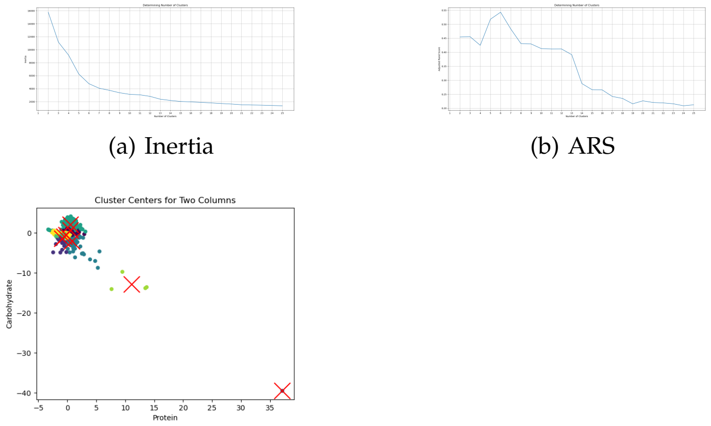
Nutrition Facts

C. Sparse Random Projection and KMeans

In figure 15(a), Inertia was analyzed. The significant decrease in inertia from 2 to 8 clusters suggests that adding more clusters improves the compactness of the clusters and reduces the variance within the cluster. Beyond 8 clusters, the decrease in inertia slows down, indicating diminishing returns in cluster compactness. This lead me to choose 8 clusters as optimal.

In figure 15(b), ARS was analyzed. The increase in ARS from 4 to 6 clusters suggests that adding more clusters here improves the agreement with true labels. However, the drop in ARS after 6 clusters suggests that adding more clusters beyond this point leads to a decrease in agreement with true labels, indicating overfitting or capturing noise as separate clusters. The max value here is 0.55, which while better than the CVD dataset still shows room for improvement as a score of 1 is the best agreement.

In figure 15(c), I plot the cluster centers. The overlap in cluster centers observed when plotting 8 clusters suggests that these clusters are not well-separated in the feature space. This overlap indicates that the centroids of these clusters are close together, making it difficult to distinguish them based solely on their centroids. It suggests that the clusters may be similar in terms of their feature distributions or that KMeans may not have effectively partitioned the data into distinct groups.



(a) Inertia

(b) ARS

(c) Cluster Centers

Fig. 15: NF KMeans and SRP

While 8 clusters came out of this analysis, further investigation into the affects of SRP and KMeans should be explored.

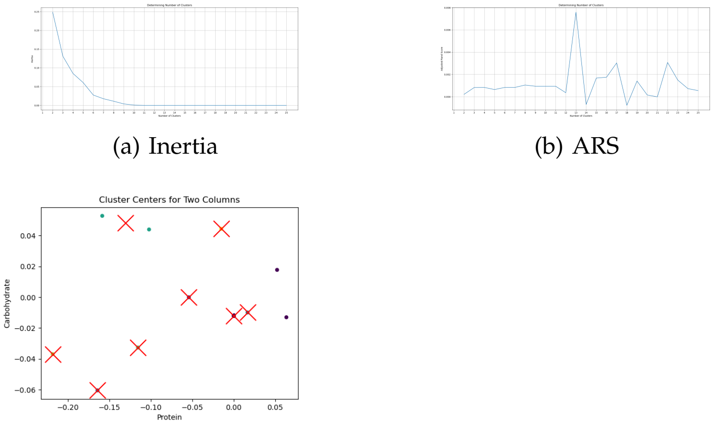
D. Hessian Locally Linear Embedding and KMeans

In figure 16(a), Inertia was analyzed. The significant decrease in inertia from 2 to 6 clusters suggests that adding more clusters improves the compactness of the clusters and reduces the variance within the cluster. Beyond 6 clusters, the decrease in inertia slows down, indicating diminishing returns in cluster compactness. While I initially choose 8 as the optimal number of clusters, analyzing again, I believe 6 would have been a better optimal number. If given more time, 6 clusters would be explored.

In figure 16(b), ARS was analyzed. The increase in ARS from 12 to 13 clusters with a score ranging from 0.001 to 0.007 suggests that 13 more clusters here improves the agreement with true labels. However, the score here is quite poor as a score of 1.0 is the best agreement. This indicates that Rand score, while the best of scores between those that measure agreeness to ground truth, is not a good measure at face value. This does, however, suggest that there is a lot of room to explore the affects of HLLE and KMeans.

In figure 16(c), I plot the cluster centers. These clusters are far apart and don't appear to be well-defined. Using HLLE doesn't appear to be the best model as it's not capturing meaningful variation in the data.

While eight clusters came out of this analysis, it's worth exploring HLLE some more to address some issues seen in the



(a) Inertia

(b) ARS

(c) Cluster Centers

Fig. 16: NF KMeans and HLLE

scatter plot. It could also be the case that there are no nonlinear relationships in the data.

V. Step 4

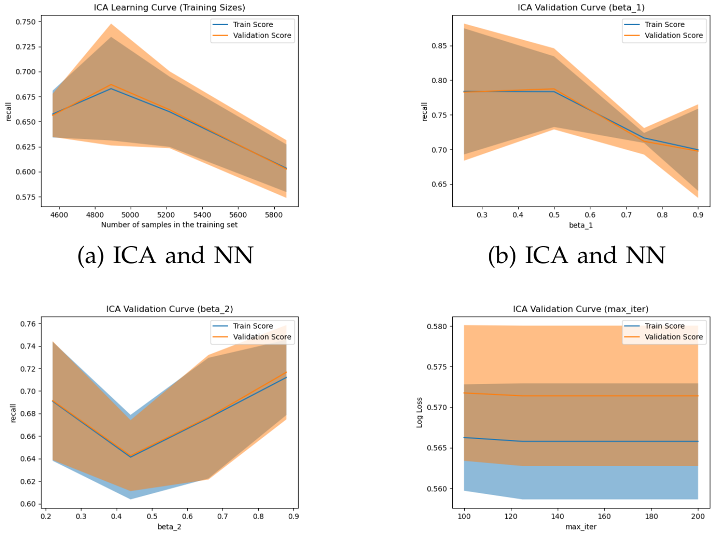
My hypothesis is that performing dimension reduction before running a neural network is that the results will be enhanced. Using a compact feature space, neural networks are expected to exhibit improved generalization, robustness, and interpretability.

Cardiovascular Disease

asghafha;kjgha;kjgha;s

A. Independent Component Analysis and Neural Network

afsdfasfas



(a) ICA and NN

(b) ICA and NN

(c) ICA and NN

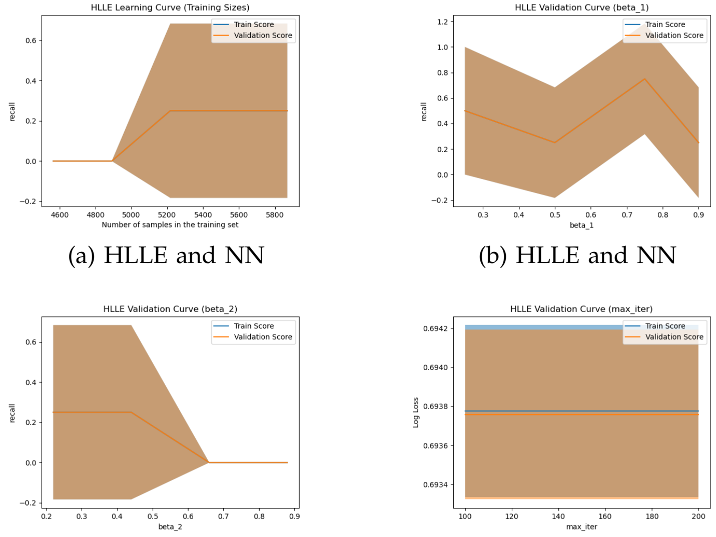
(d) ICA and NN

Fig. 17: ICA and NN

asfasf

B. Hessian Locally Linear Embedding and Neural Network

blahblahblahbalhbakd;jad
adsfsdfas



(a) HLLE and NN

(b) HLLE and NN

(c) HLLE and NN

(d) HLLE and NN

Fig. 18: HLLE and NN

VI. Step 5

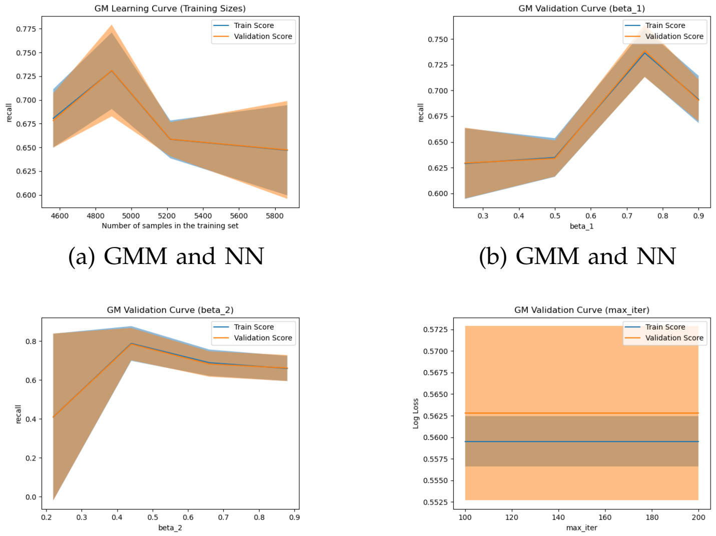
My hypothesis is that using clustering before running a neural network will enhance my results. This hypothesis posits that incorporating clustering results into the neural network's feature space will augment its ability to capture and leverage the inherent structure of the data.

Cardiovascular Disease

asfsdf

A. Gaussian Mixture Method and Neural Network

adsfas



(a) GMM and NN

(b) GMM and NN

(c) GMM and NN

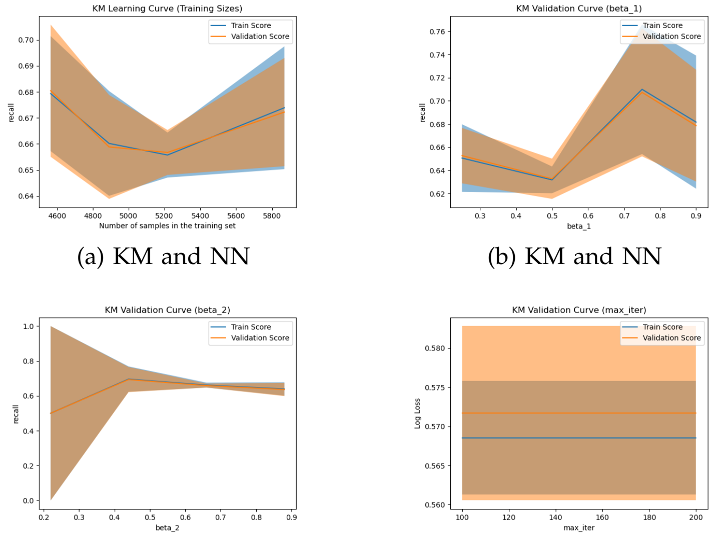
(d) GMM and NN

Fig. 19: GMM and NN

dafss

B. KMeans Clustering and Neural Network

asfsdaf
asdfsdf



(a) KM and NN

(b) KM and NN

(c) KM and NN

(d) KM and NN

Fig. 20: KM and NN

VII. Conclusion

In conclusion, the methods of KM, GMM, PCA, ICA, SRP, and HLLE stand as indispensable tools in any machine learning toolkit. From organizing unlabeled data points into meaningful clusters to uncovering the latent structure of high-dimensional datasets, each technique offers unique capabilities and insights into the underlying patterns within data.

As I navigated the complexities of my two datasets characterized by mild dimensionality, noise, and nonlinear relationships, the importance of these methods cannot be overstated. Whether it be for exploratory data analysis, feature extraction, or dimensionality reduction, understanding these methods empowers machine learning practitioners to extract insights and drive decision-making.

VIII. Resources

[1] *API Reference*. Python. https://www.python.org/.
[2] *API Reference*. Pandas. https://pandas.pydata.org/.
[3] *API Reference*. Jupyter. https://jupyter.org/.
[4] *API Reference*. IPython. https://ipython.org/.
[5] *API Reference*. MatPlotLib. https://matplotlib.org/stable/.
[6] *API Reference*. NumPy. https://numpy.org/.
[7] *API Reference*. Seaborn. https://seaborn.pydata.org/index.html.
[8] *API Reference*. OpenPyXl. https://openpyxl.readthedocs.io/en/stable/tutorial.html.
[9] *API Reference*. Scikit-Learn. https://scikit-learn.org.
[10] Nakamura, K. (2023). *ML LaTeX Template*.
[11] *Data Source* Nutrition Facts Database Tools and Spreadsheet. https://tools.myfooddata.com/nutrition-facts-database-spreadsheet.php.
[12] *Data Source* Cardiovascular Disease. https://www.kaggle.com/datasets/colewelkins/cardiovascular-disease.