BIOST 527 Final Project Report

Yueqi Xu

2023-05-20

1 Data Processing

The original data set contains 12,000 observations and 64 variables, with no missing values. Upon an initial inspection, it appears that the data in each dimension follows an approximately normal distribution

The clustering methods I plan to use are mean-shift clustering and hierarchical clustering. Mean-shift clustering is based on the kernel density estimate (KDE), and it is well-known that the performance of kernel density estimators degrades as the dimension increases, caused by the sparseness of the data in high-dimensional spaces (Hyrien 2016). Therefore, a proper dimension reduction method is desired.

I initially attempted principal component analysis (PCA), which is a commonly used linear dimension reduction method. The scree plot of PCA is presented in Figure 1, illustrating the variance explained by each principal component (PC). However, the plot indicates a linear decrease in variance without a distinct elbow point, making it challenging to determine the optimal number of PCs to retain. Hence, PCA may not be the most suitable dimension reduction method for this dataset.

The next dimension reduction method I tried was t-distributed stochastic neighbor embedding (t-SNE), which is a non-linear dimension reduction technique. A 2D Visualization of dimension-reduced data with t-SNE is shown in Figure 2.

Upon examining the visualizations, it is evident that there are several distinct clusters in the t-SNE projected data. The transformed data from t-SNE appear to be suitable for clustering analysis. Therefore, I decided to utilize the t-SNE projected data for mean shift clustering. However, upon further investigation, I discovered that it may be inappropriate to directly use the t-SNE transformed data for clustering purposes. Section 2.1 provides more details about the t-SNE transformation and its validity, highlighting any potential limitations or considerations.

Hierarchical clustering, on the other hand, employs a measure of distance or similarity to create clusters and is extensively used to organize high-dimensional objects (Sean et al. 2013). Thus no data processing was done for hierarchical clustering, as it tends to perform well with high-dimensional data.

2 Method Description

This section provides a more detailed description of the clustering methods employed in this study.

2.1 Mean Shift Clustering

Mean shift clustering is a non-parametric density-based clustering algorithm. It is a flexible clustering algorithm that does not require a pre-specified number of clusters. It is particularly effective in identifying clusters of arbitrary shapes and sizes. The main idea behind mean shift clustering is to iteratively shift data points towards the nearest local maxima in the estimated density curve, typically estimated by kernel

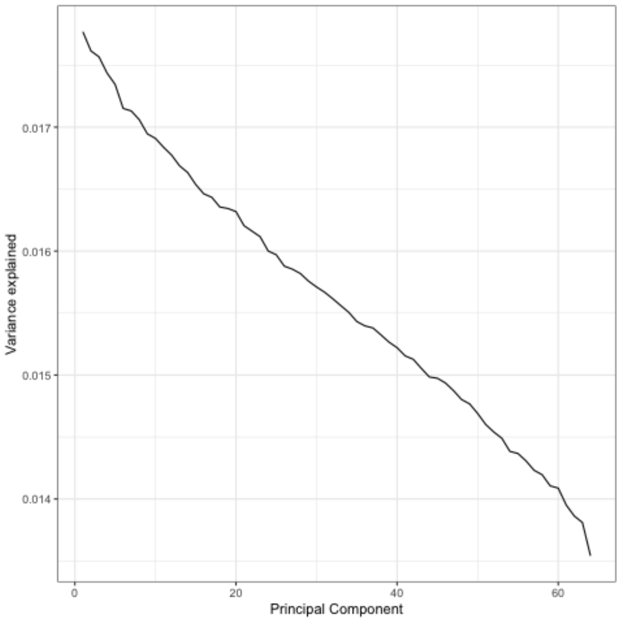

Figure 1: Scree plot of PCA

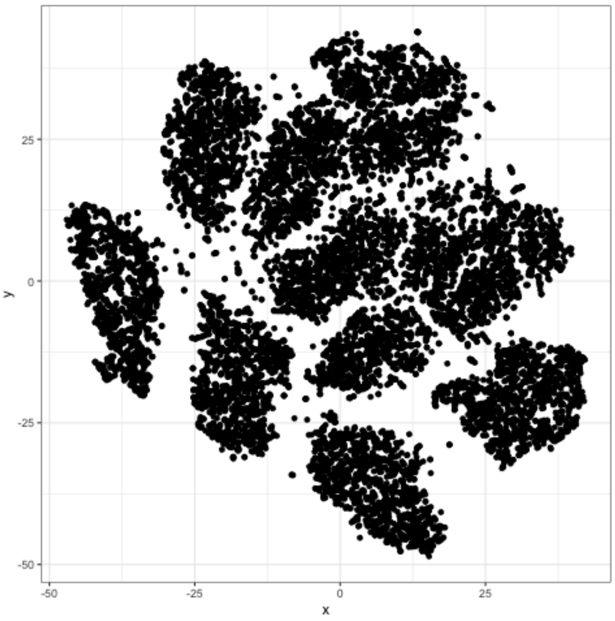

Figure 2: 2D Visualization of t-SNE projected data

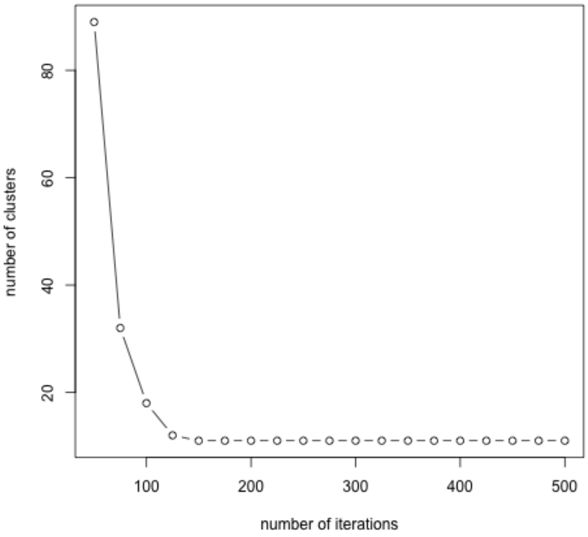

Figure 3: Number of clusters by number of iterations

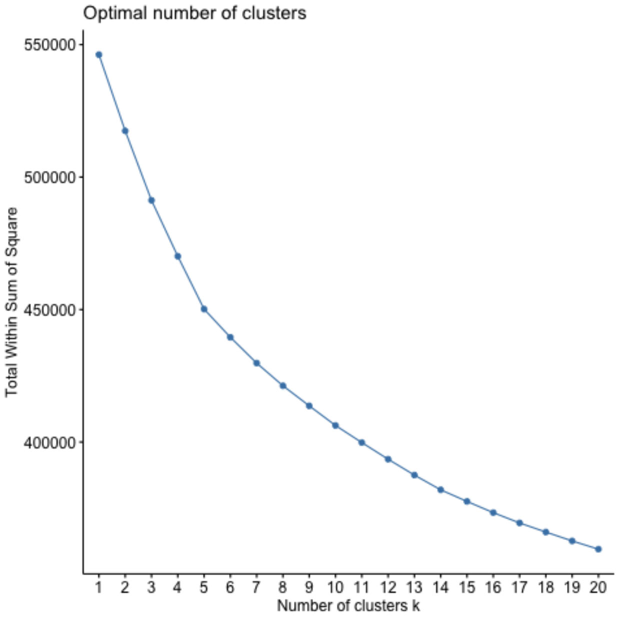

Figure 4: WSS by number of clusters

density estimator. In order to do so, a kernel function needs to be defined, and a bandwidth need to be specified. Commonly used kernel functions include the Gaussian kernel, uniform kernel, triangular kernel, and rectangular kernel.

As mentioned in Section 1, the performance of kernel density estimators tends to degrade as the dimension increases. Therefore, for mean shift clustering, the data was projected into a 2-dimensional space using t-SNE. t-SNE is a technique commonly used to visualize high-dimensional data by giving each data point a location in a two or three-dimensional map (Van 2008). t-SNE is capable of capturing much of the local structure of the high-dimensional data very well, while also revealing global structure such as the presence of clusters at several scales, which I thought may be beneficial in terms of finding clusters.

further research conducted during the writing of this report revealed that t-SNE does not preserve the distance relationships between observations or the density information (Erich, 2022). Consequently, running distance- or density-based algorithms directly on the t-SNE-transformed data may result in information loss and potentially yield poor clustering results. Therefore, it may be more appropriate to use t-SNE solely for visualization purposes and employ other dimension-reduction methods for data processing.

2.2 Hierarchical Clustering

Hierarchical clustering is a non-parametric distance/similarity-based clustering algorithm. There are two types of hierarchical clustering algorithms: agglomerative and divisive. The agglomerative approach starts with treating each observation as a single cluster, and merging clusters with the lowest dissimilarity; the divisive approach starts with treating all observations as one big cluster and then splitting clusters with the greatest dissimilarity. Either way, the algorithm will result in a tree-like diagram called a dendrogram, with the y-axis representing the dissimilarity between clusters at each merging or splitting step.

Various dissimilarity measures can be employed in hierarchical clustering, including single linkage, complete linkage, centroid linkage, and Ward's method (based on least squares).

One advantage of hierarchical clustering is that it does not require the pre-specification of the number of clusters, unlike some other algorithms such as K-means. This makes it an efficient approach, as the algorithm only needs to be run once, and the number of clusters can be determined later based on specific needs or requirements using the produced dendrogram.

3 Clustering Strategy and Parameter Choices.

3.1 Mean Shift Clustering

The function `Rtsne` from package `Rtsne` was used to generate t-SNE projected data. All parameters were set as the default value. I chose to convert the data into 2 dimensions because visualizations in 2-dimensions are more clear and easier for human brains to process than in 3-dimension. However, when performing the t-SNE, I did not realize the initialization was randomized so no random seed was set for the output I turned in. Therefore, the outputs shown in this report may be a little different from the one I submitted, but the difference should not be extreme as this algorithm is stabled on random seeds. Please check section 4.2 for more details. For the outputs in this report, all random seed was set a 527, unless mentioned otherwise.

A mean-shift algorithm was employed, with a Gaussian kernel for KDE since the data is smooth and continuous across all dimensions. The bandwidth was chosen using the `Hpi.diag()` function in package `ks`. It is the multivariate plug-in selector of Wand & Jones (1994), which is shown to be more stable and reliable than some previously proposed multivariate bandwidth selection procedures such as least squares cross-validation and over-smoothing, and exhibits good theoretical and practical performance. The selected bandwidths are `H = c(4.32, 4.85)`.

With the selected bandwidth, a plot showing the number of clusters vs. the number of iterations is shown in Figure 3. The number of clusters converges to 11 at `iterations = 150` and remains unchanged thereafter. Therefore, the number of iterations is set to 150 to allow enough time for the clusters to converge while the runtime is still acceptable. The resulting clusters are visualized in Figure 6.

In contrast, the clusters generated from the mean shift clustering algorithm with bandwidth selected by cross-validation are presented in Figure 5. With the same number of iterations, the bandwidth selected by the plug-in procedure result in fewer clusters (11 vs. 237) with clearer boundaries.

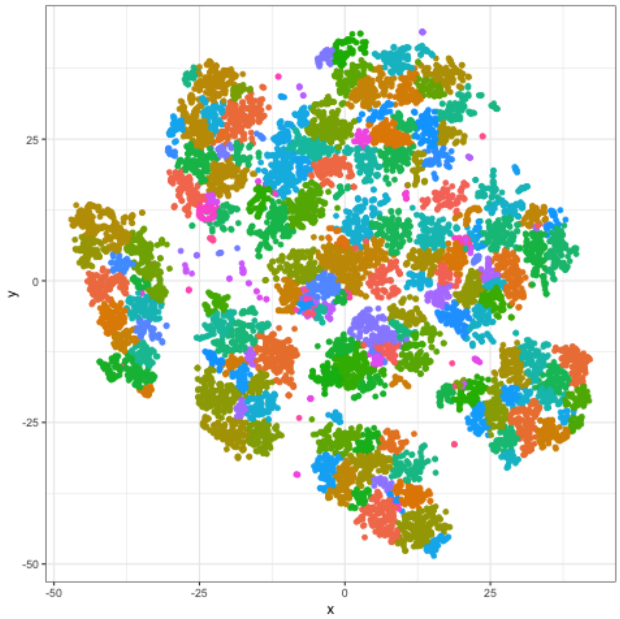

Figure 5: Mean shift clustering result with dimension reduced data and CV bandwidth; 150 iterations

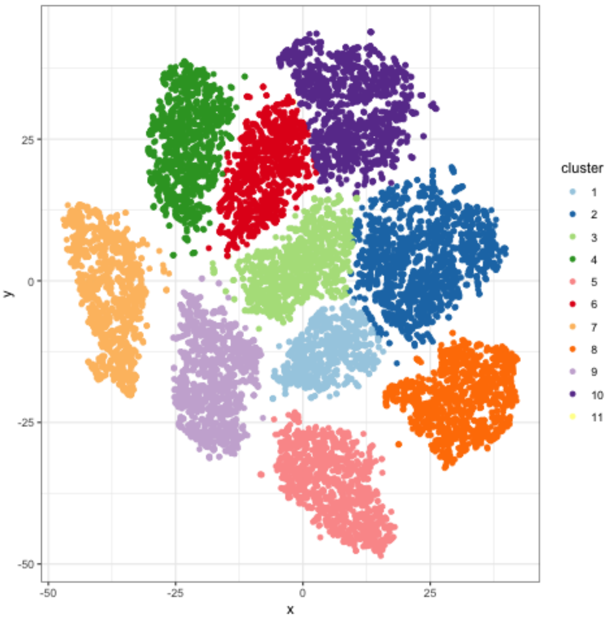

Figure 6: Mean shift clustering result with dimension reduced data and plug-in bandwidth; 150 iterations

3.2 Hierarchical Clustering

For the second method, I would like to employ a method that is feasible and efficient in clustering high dimensional data, in order to assess the similarity of clusters generated on the original data and on the t-SNE transformed data. Therefore, a hierarchical clustering was employed, with the original dataset without performing dimension reduction.

Various studies, including Ferreira & Hitchcock (2009) and Doğan (2013) have shown that Ward's method is usually the most robust method among the common linkage choices of Hierarchical clustering. Thus in this study, the Ward method is used, with an agglomerative approach. The clustering was performed using the `hclust` function in the `stats` package.

The number of clusters is determined according to both the within-cluster sum of squared (WSS) score and visual assessment of the 2D projected visualization of the original data (Figure 2). A line plot of the WSS score versus the number of clusters k is shown in Figure 4. There is not an obvious elbow point on this plot; however, we can still observe that the decrease in total WSS is getting slower as the number of clusters increases: the total WSS decreased by approximately 150,000 from $k = 1$ to $k = 10$, and decreased less than 50,000 from $k = 11$ to $k = 20$. And from Figure 2, I observed 10 major clusters. Thus I decided to cut the dendrogram where the number of clusters is equal to 10, and a 2D visualization of the resulting clusters is presented in Figure 7.

In contrast, an agglomerative hierarchical clustering with the Ward method was performed on the t-SNE transformed data, the number of clusters was set at 10, and the resulting cluster assignments in displayed in Figure 8.

4 Evaluation

The clustering results are evaluated with three metrics: visual assessment, stability analysis, and internal evaluation. At the end of this section, the clustering results obtained from the two methods are also cross-compared.

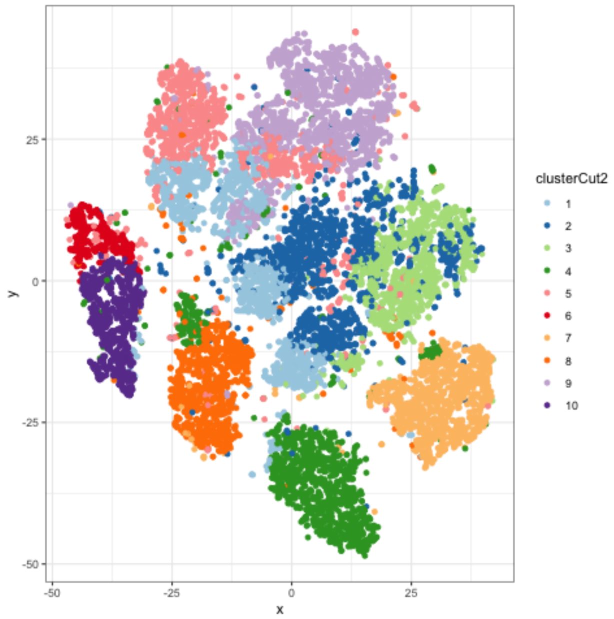

Figure 7: Hierarchical clustering on original data

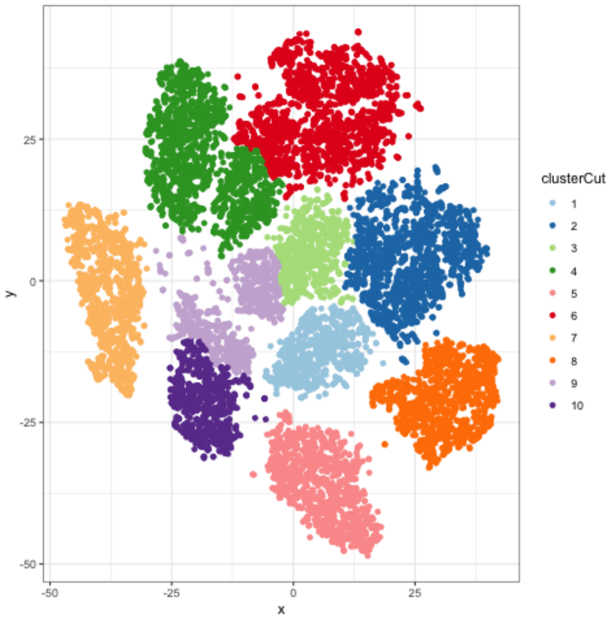

Figure 8: Hierarchical clustering on t-SNE output

4.1 Visual Accessment

Visualization of clusters generated by mean shift clustering and hierarchical clustering are presented in Figure 6 and Figure 7, respectively. The clusters assigned by the mean shift algorithm have much clearer boundaries than the clusters generated by hierarchical clustering on the t-SNE projected plot. However, we need to be careful here that the clusters of the mean shift were generated based on the t-SNE projected data, while the clusters of Hierarchical clustering were created based on the original data. Therefore, better separation on the t-SNE projected plot does not necessarily imply better clusterings.

4.2 Stability Analysis

To assess the stability of the clustering results, we conducted a stability analysis for both mean shift clustering and hierarchical clustering.

For mean shift clustering, I evaluated the impact of different initialization of t-SNE dimension reduction (controlled by random seed) and different bandwidths on the clustering results. A table of rand indices subject to different changes is presented in Table 1.

Change	`set.seed(533)`	`set.seed(579)`	`H = c(4.5, 5)`	`H = c(4, 4.5)`
Rand Index	0.972	0.979	0.9997	0.994

Table 1: Rand index of mean shift clusterings

Rand index is a measure of similarity between clustering assignments, ranging between 0 to 1, with higher values indicating more similar clustering results. In the table above, we can see that when keeping the bandwidth the same and changing the random seed of t-SNE transformation to 533 and 579, the new cluster assignment obtains a Rand index of 0.971 and 0.978, respectively, compared to the clustering result in 3.1. In addition, when keeping the random seed the same at 527 and making small changes to the bandwidths, the resulting cluster assignments are almost the same as what we obtained in section 3.1, which indicates that the method we used in section 3.1 is stable.

For hierarchical clustering performed on original data, I employed bootstrapping to evaluate stability. $k = 20$ bootstrap samples of size 2,000 were sampled from the original data with replacement, with random seed equals 527. For each bootstrap sample, hierarchical clustering with least squares linkage was performed, and the number of clusters was set at 10. Then the Rand index of similarity of the resulting cluster assignment and the cluster assignments in section 3.2 of the corresponding observations was computed. The average Rand index of the 20 bootstrap samples is 0.890.

For the hierarchical clustering performed on t-SNE transformed data, the same approach was taken, and the average Rand index of the 20 bootstrap samples is 0.977.

In contrast, $k = 20$ bootstrap samples of size 2,000 were taken from the t-SNE transformed data with replacement with the same random seed. For each bootstrap sample, mean shift clustering with bandwidth `H = c(4.32, 4.85)` was performed. Then the Rand index of similarity of the resulting cluster assignment and the cluster assignments in section 3.1 of the corresponding observations was computed. The average Rand index of the 20 bootstrap samples is 0.988.

In summary, when resampling the data, the mean shift algorithm shows higher stability than the hierarchical algorithm. In addition, the mean shift algorithm also shows stability under a change of random seed for t-SNE transformation and a small change of bandwidth.

4.3 Internal Evaluation

Table 2 shows some internal evaluation metrics, where

- WCSS is the within clusters sum of squares.
- Avg Silhouette is the average silhouette width, with values ranging from -1 to 1, where higher values indicate better-defined and well-separated clusters.
- Avg Between is the average distance between clusters.
- CH Index is the Calinski-Harabasz Index, where higher values indicate better-defined and well-separated clusters.

All metrics are evaluated on the original data set.

Metric	Mean Shift	Hierarchical (original)	Hierarchical (t-SNE)
WCSS	406075.8	408968.1	409187.1
Avg Silhouette	0.0331	0.0374	0.0570
Avg Between	9.586	9.580	9.583
CH Index	413.7	447.1	446.1

Table 2: Internal evaluation metrics

Regarding the evaluation metrics, the clustering methods employed have approximately the same overall performance. The only metric that has a fixed range and hence is interpretable is the Silhouette score. For both clustering methods used, the Silhouette scores are close to zero, indicating that there might exist overlapping clusters – the clusters are not very well-defined.

4.4 Compare Clustering Results

Overall, among the employed clustering methods, mean shift clustering with t-SNE reduced data exhibits the highest stability, hierarchical clustering with original data demonstrates the lowest stability, and hierarchical clustering with t-SNE reduced data falls in the middle in terms of stability. Regarding the quality of clusters, all the employed methods perform similarly.

In terms of similarity, the rand index of similarity of the clusters generated by hierarchical clustering with original and t-SNE transformed data is 0.913. On the other hand, the rand index of similarity of the clusters generated by hierarchical clustering and mean shift clustering on t-SNE reduced data is 0.988. The cluster assignments created by hierarchical clustering with original and t-SNE transformed data do not differ significantly, and when performing hierarchical clustering and mean shift clustering on the same dataset (t-SNE transformed data), they yield highly similar clustering results.

References

1. Hyrien O, Baran A. Fast Nonparametric Density-Based Clustering of Large Data Sets Using a Stochastic Approximation Mean-Shift Algorithm. J Comput Graph Stat. 2016; 25(3):899-916. doi: 10.1080/10618600.2015.1051625. Epub 2016 Aug 5. PMID: 28479847; PMCID: PMC5417725.

2. Sean Gilpin, Buyue Qian, and Ian Davidson. 2013. Efficient hierarchical clustering of large high dimensional datasets. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management (CIKM '13). Association for Computing Machinery, New York, NY, USA, 1371–1380. https://doi.org/10.1145/2505515.2505527

3. Carreira-Perpinán, M.A., 2015. A review of mean-shift algorithms for clustering. arXiv preprint arXiv:1503.00687.

4. Erich Schubert (https://stats.stackexchange.com/users/18215/erich-schubert), Clustering on the output of t-SNE, URL (version: 2022-05-17): https://stats.stackexchange.com/q/264647

5. van der Maaten and Hinton, 2008. Visualizing data using t-SNE. J. Mach. Learn. Res., 9 (2008), pp. 2579-2605

6. Wand, M.P. & Jones, M.C. (1994) Multivariate plug-in bandwidth selection. Computational Statistics, 9, 97-116.

7. Laura Ferreira & David B. Hitchcock (2009) A Comparison of Hierarchical Methods for Clustering Functional Data, Communications in Statistics - Simulation and Computation, 38:9, 1925-1949, DOI: 10.1080/03610910903168603

8. Saraçli, S., Doğan, N. & Doğan, İ. Comparison of hierarchical cluster analysis methods by cophenetic correlation. J Inequal Appl 2013, 203 (2013). https://doi.org/10.1186/1029-242X-2013-203