Cluster Analysis of Protein Expression Data

Zhengyang Fei

CHL5230: Applied Machine Learning for Health Data

Dalla Lana School of Public Health

UNIVERSITY OF TORONTO

April 3, 2025

Contents

Section 1: Introduction

The analysis of high-dimensional biological data, such as protein expression levels, presents significant challenges due to the complexity and scale of the data. One of the key steps in understanding and interpreting such datasets is to identify patterns or groups within the data. Unsupervised learning methods, particularly clustering, are commonly used for this purpose. In this assignment, we explore various clustering techniques to categorize protein expression data from experimental samples.

The following questions guide this exploratory analysis:

1. Are there any distinct clusters in the measurements?

2. Are these clusters associated with any of the different genotype, behaviour and/or treatment?

3. Do we see any particular set of proteins that exhibit a distinct expression pattern (profile) in any or all of these clusters?

Additionally, a significant challenge in this analysis arises from the assumption that the samples are independent. In reality, measurements from multiple samples come from the same mouse, meaning the data points may not be independent. This introduces a potential correlation in the data, which could affect the clustering results. We will discuss the implications of this assumption and explore ways to address the potential correlatedness.

Finally, we will conclude by discussing how unsupervised learning techniques, in particular the ones that will be discussed, can be applied to my personal real-world research problems.

Section 2: Methodology and Results

2.1 Data Processing

This analysis focuses on the exploratory investigation of protein expression data collected from a set of experimental mice. The dataset consists of protein expression measurements taken from 72 mice, including both control and trisomic (Down syndrome) genotypes, with additional experimental factors such as behavior (stimulated to learn vs. not stimulated) and treatment (memantine vs. saline). The protein expression data includes 77 proteins or protein modifications, with each measurement being taken 15 times per sample, resulting in a total of 1080 data points per protein across all samples.

Variable Name	**Type**	**Description**	**Proportion**	**Role**
Genotype	Categorical	Control	0.5278	Feature
		Ts65Dn	0.4722	
Treatment	Categorical	Memantine	0.5278	Feature
		Saline	0.4722	
Behavior	Categorical	C/S	0.4861	Feature
		S/C	0.5139	
Proteins	Numerical	77 protein expression values	Varied mean and SD	Feature
Class	Categorical	c-CS-m	0.1389	Target
		c-CS-s	0.1250	
		c-SC-m	0.1389	
		c-SC-s	0.1250	
		t-CS-m	0.1250	
		t-CS-s	0.0972	
		t-SC-m	0.1250	
		t-SC-s	0.1250	

Table 1: Summary of Data

We addressed missingness in the dataset by setting a threshold for column-level missing values. If a column had more than 15% missing values, we remove that column as instructed. As a result, the following five columns were deleted: ”BAD_N”, ”BCL2_N”, ”H3AcK18_N”, ”EGR1_N”, and ”H3MeK4_N”.

For the remaining data, imputation was employed to fill in missing values. Each protein’s missing values were imputed using the mean value calculated within the same class.

2.2 Clustering Methodologies

Three clustering techniques were employed to the data:

- K-means
- Hierarchical clustering
- Partitioning Around Medoids (PAM)

We applied Principal Component Analysis (PCA) to reduce the dimensionality of the data while retaining as much variance as possible. The cumulative proportion of variance explained plot above illustrates how much of the variance is captured by each successive principal component (PC). From Figure 1, the first few PCs account for the majority of the variance in the dataset. In practice, we often aim to retain at least 95% of the variance, which in this case is achieved with the first 28 PCs.

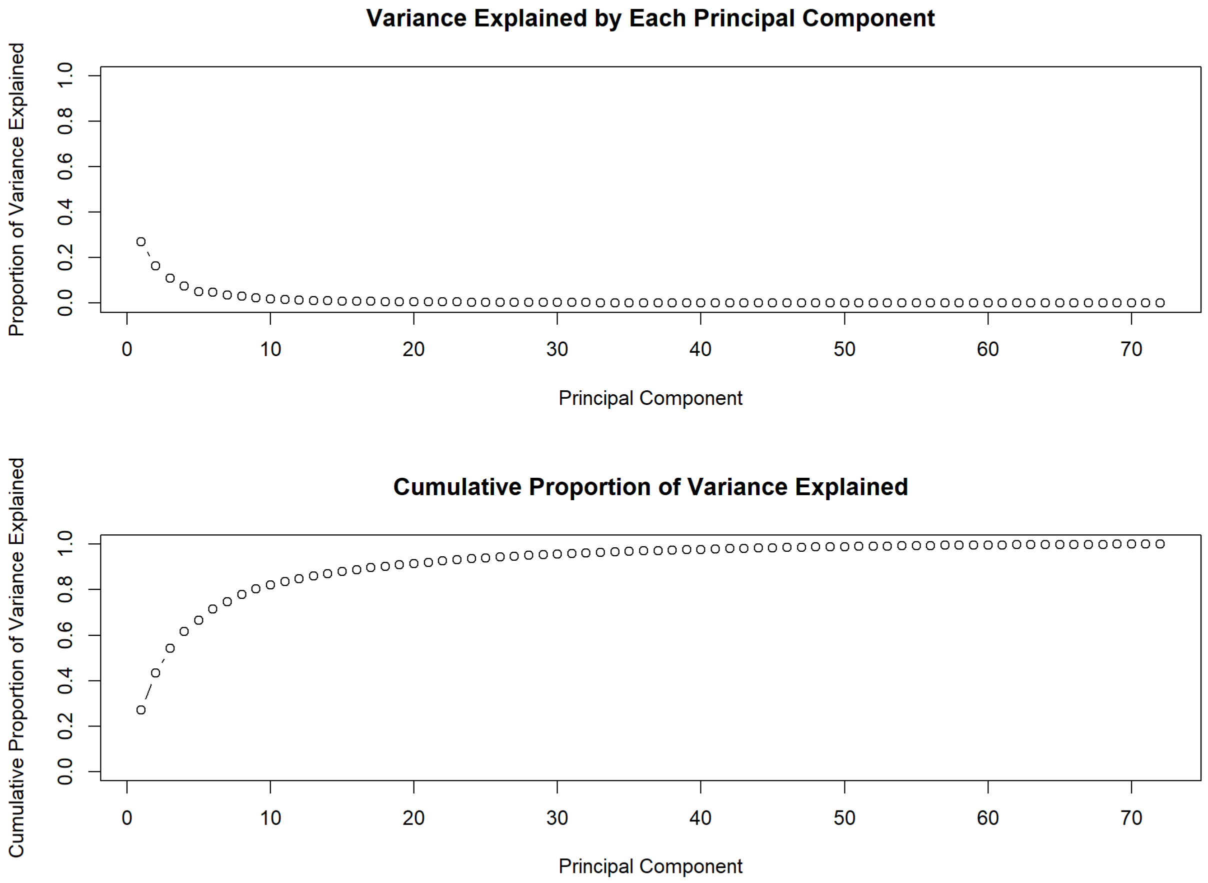

Figure 1: Variance/Cumulative Variance Explained by each Principal Component

Before employing each technique, we applied silhouette plot showing the average silhouette width for different numbers of clusters (from 2 to 10). The silhouette width measures how similar each point is to its own cluster compared to other clusters. The average silhouette width is used to evaluate how well the clustering has been performed.

A higher average silhouette width indicates that the clusters are well-defined and are likely to produce meaningful clusters.

In this case, Figure 2 suggests that exploring clusters with $k = 2$ or $k = 8$ could be beneficial, as these values either exhibit the highest silhouette width or show a sharp increase in width, indicating a more distinct separation between clusters. Another reason exploring $k = 8$ clusters might be insightful is that our dataset contains 8 classes. It would be interesting to investigate the performance of these clustering methods in relation to these predefined classes.

Similar plots were employed for hierarchical clustering and PAM, which showed similar results.

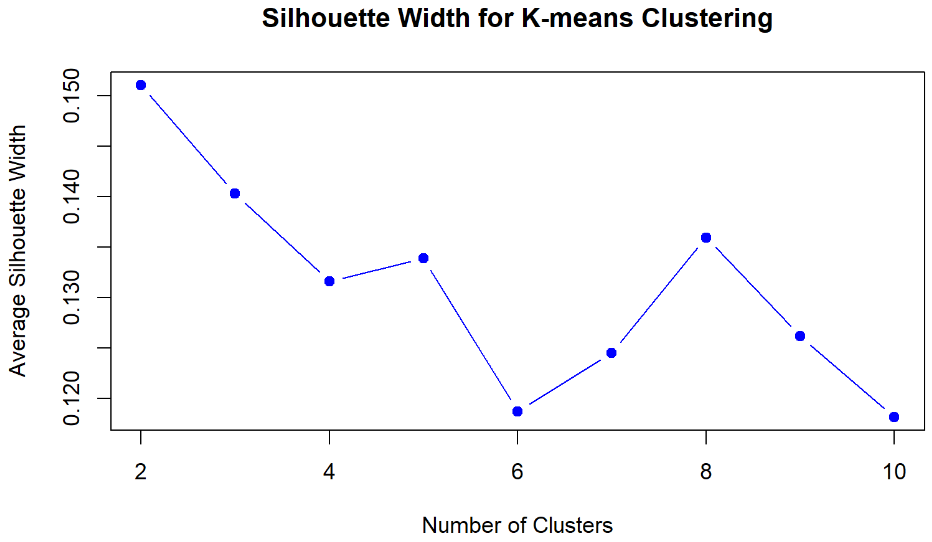

Figure 2: Silhouette Width for K-means Clustering with Cluster Sizes 2 to 10

2.2.1 K-Means Clustering

A visual representation of how the data is clustered into two groups based on the first two PCs can be seen in Figure 3. The two clusters seem to be fairly well-separated with little overlap, suggesting that the first two principal components capture enough of the variance to distinguish the two groups. We do this to visualize the data in a lower-dimensional space while preserving as much variability as possible. This allows us to identify patterns that are otherwise harder to visualize due to high dimensionality.

Figure 4 shows heatmaps illustrating the clustering based on genotype, behavior, and treatment categories. For behavior, the S/C category is predominantly assigned to cluster 2, while C/S is more evenly distributed between cluster 1 and cluster 2, but more-so leaning towards cluster 1. For genotype, the Ts65Dn genotype shows a stronger association with cluster 2, while the control category is mostly assigned to cluster 2 as well. For treatment, saline is mostly in cluster 2, while memantine is also distributed in cluster 2.

This shows that for genotype and treatment, k means with two clusters does not do a very good job at classification.

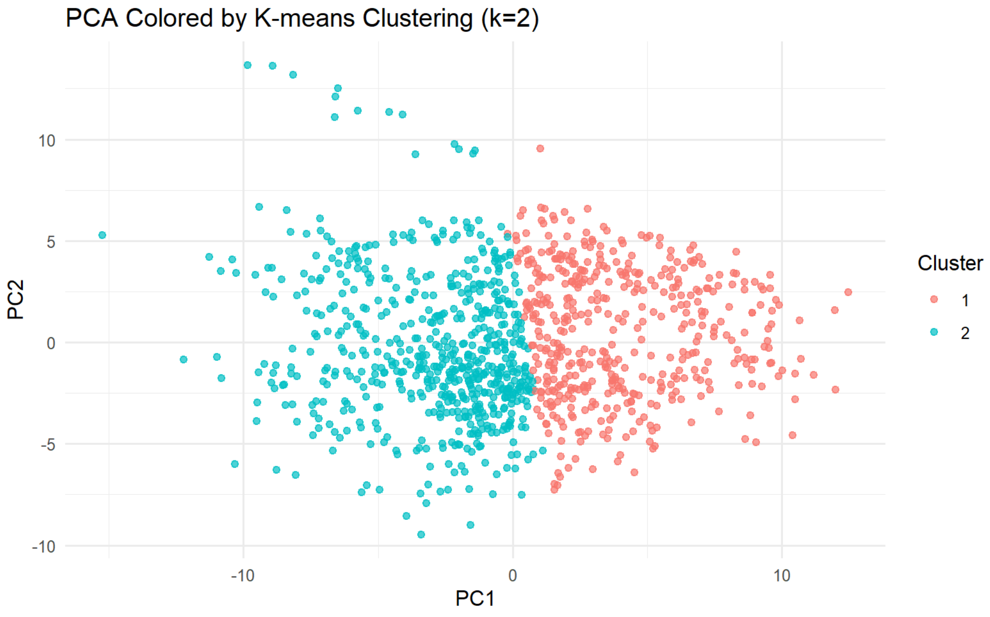

Figure 3: PCA Visualization of K-means Clustering (k = 8)

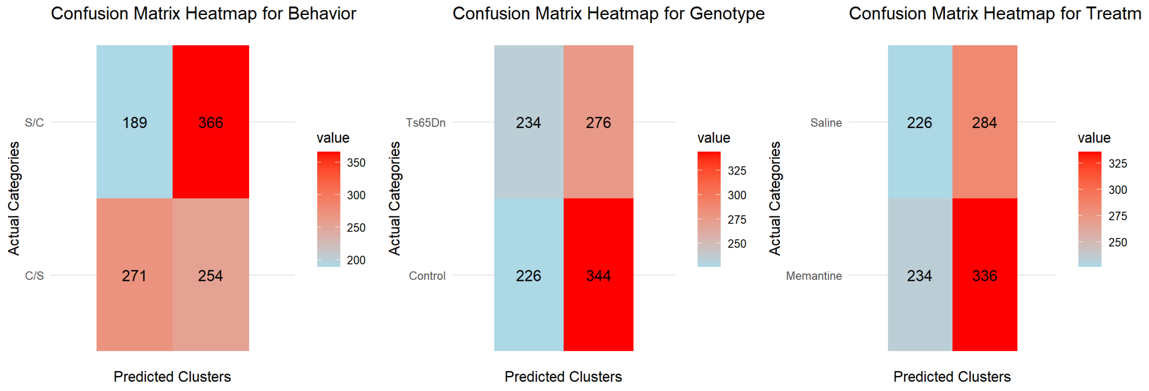

Figure 4: Heatmap of Kmeans Clustering Using Two Clusters

To measure dissimilarity, we can employ Gower's coefficient to display a histogram and a mean. As we see in figure 5, the gower distance has mean 0.1494, indicating similarity between data points.

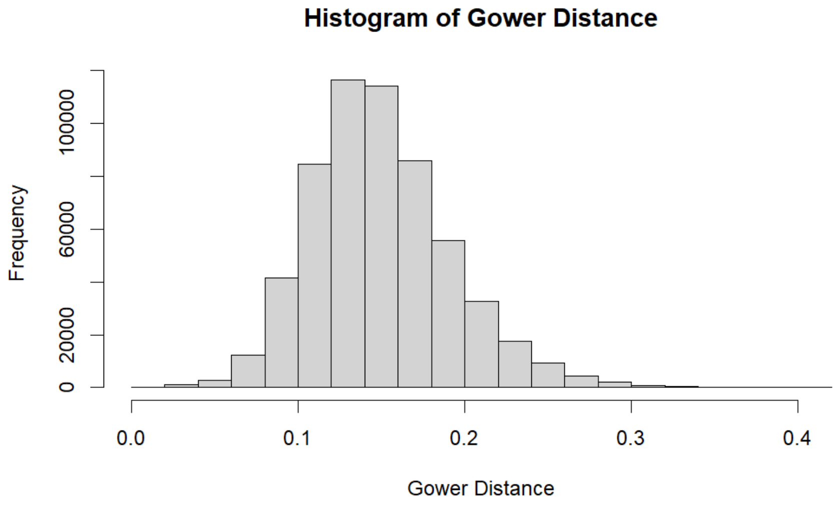

Figure 5: Histogram of Gower's coefficient for two clusters (K-means)

We now repeat the process for eight clusters. Figure 6 reveals some overlap among the clusters, but distinct groups remain observable. Notably, clusters 4 (green), 5 (blue), and 6 (cyan) show significant overlap in the middle of the plot, suggesting that these clusters are not well-separated in the feature space of the first two principal components. In contrast, the remaining clusters are more spread out and better separated.

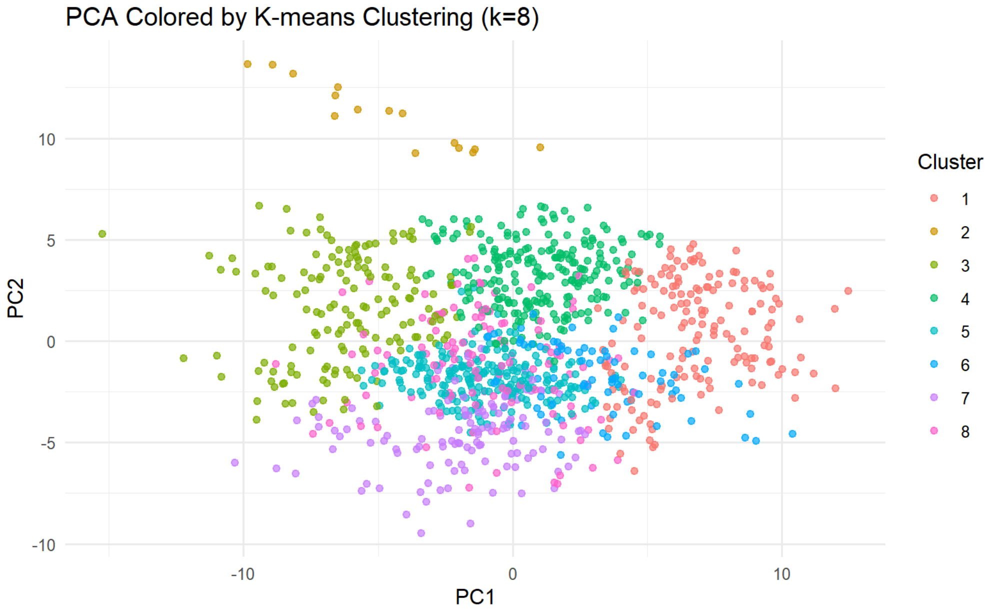

Figure 6: PCA Visualization of K-means Clustering (k = 8)

Figure 7 shows the heatmap of the confusion matrix for genotype, treatment, and behavior categories. For genotype, Ts65Dn is primarily predicted to cluster 1 and cluster 4 while control is primarily in clusters 4 and 5. For treatment, saline is primarily in cluster 4 while memantine is primarily in cluster 4 and 5. For behavior, S/C is primarily in cluster 5 while C/S is primarily in cluster 4.

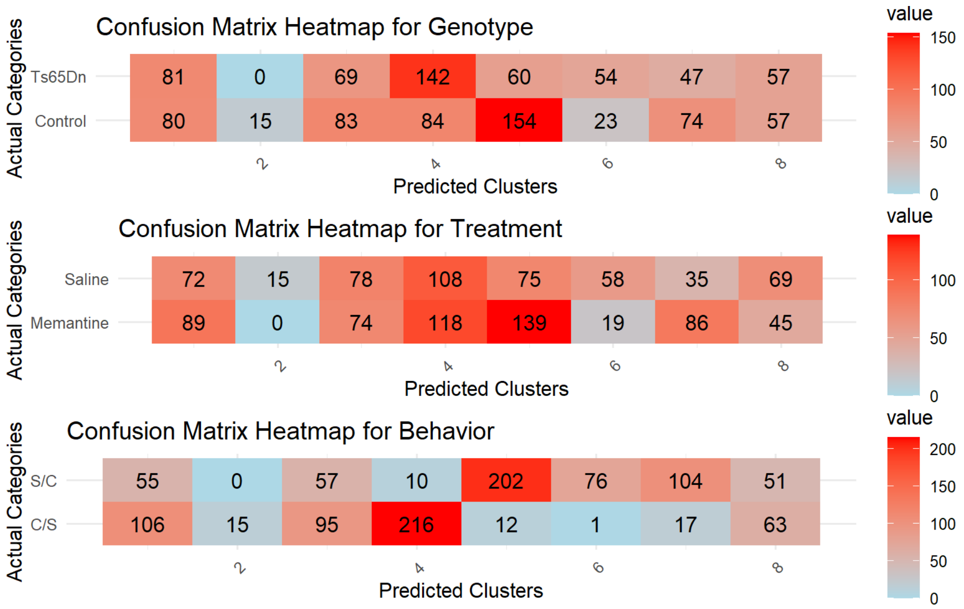

Figure 7: Heatmap of Kmeans Clustering Using Eight Clusters

Using Gower's coefficient, we construct a histogram and a mean. As we see in figure 9, the Gower distance has mean 0.1302, which is slightly lower than the two cluster method using k means.

Using the two-cluster method, we identified the top 10 contributors of protein for the first two principal components in figure 8. For PC1, proteins like MEK_N, BDNF_N, and NR1_N exhibit negative loadings, meaning their expression decreases as PC1 increases. In contrast, for PC2, proteins such as CaNA_N, ITSN1_N, and GSK3B_N show positive loadings, indicating that their expression increases with PC2. However, proteins like Ubiquitin_N, P38_N, and SNCA_N have negative loadings for PC2, meaning their expression decreases as PC2 increases.

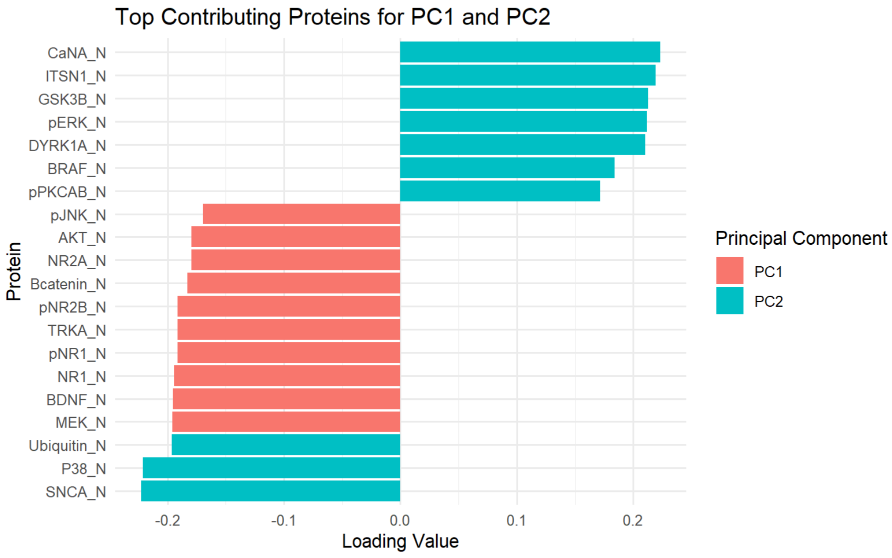

Figure 8: Top Contributing Proteins for PC1 and PC2 using K means with 2 clusters

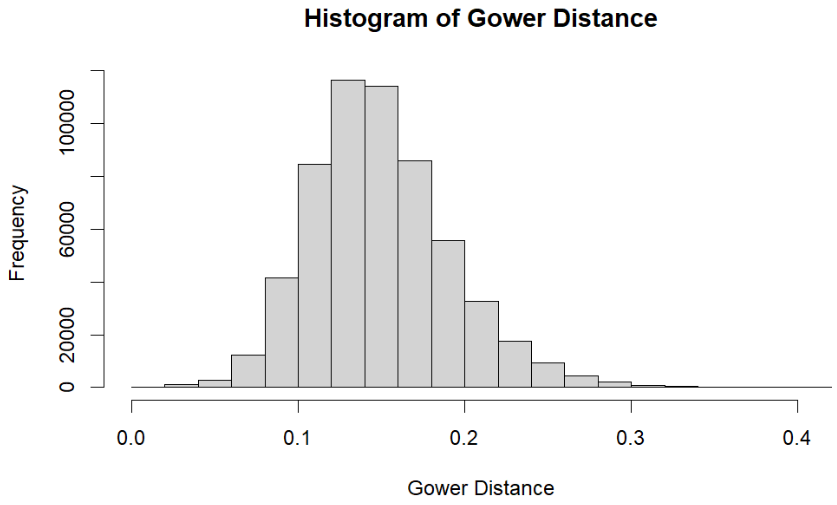

Figure 9: Histogram of Gower's coefficient for eight clusters (K-means)

2.2.2 Hierarchical Clustering

For hierarchical clustering, we will use two dissimilarity metrics (Euclidean and Pearson) as both are appropriate for our analysis. We will begin by using the Euclidean metric.

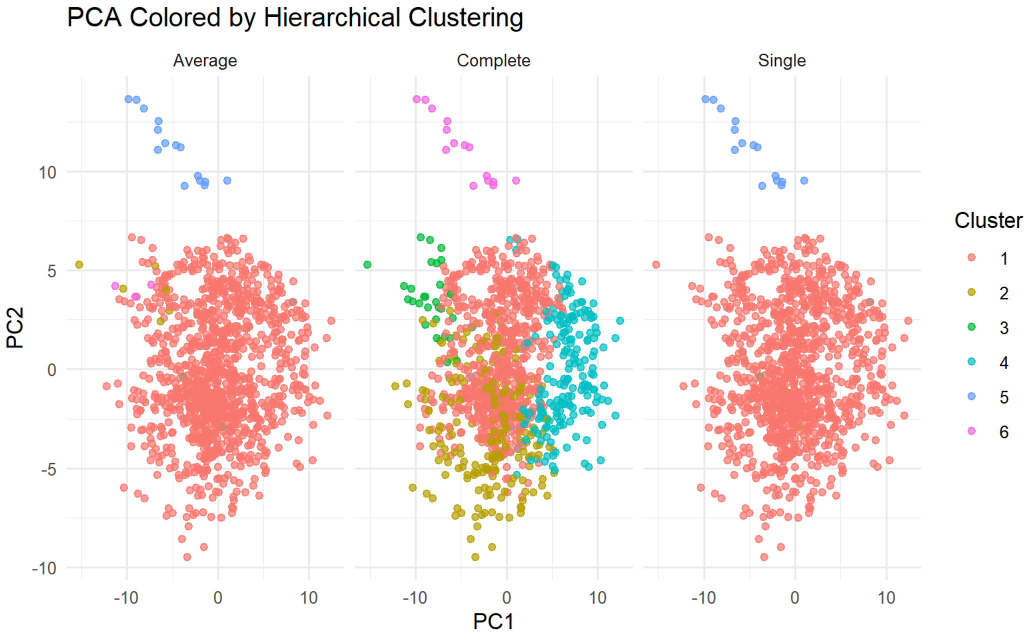

Figure 10: PCA Visualization of Hierarchical Clustering using Euclidean (Average, Complete, and Single Linkage)

Figure 10 gives a visual representation of how the data is clustered into eight groups based on the first two principle components. We choose $k = 8$ as recommended by the hierarchical silhouette plot similar to Figure 2 (included in code). All three linkage methods were used. Looking at the plot, the complete linkage shows more compact clusters with better separation between the clusters compared to the other two methods. Both average and single linkage shows fewer clear boundaries between the clusters.

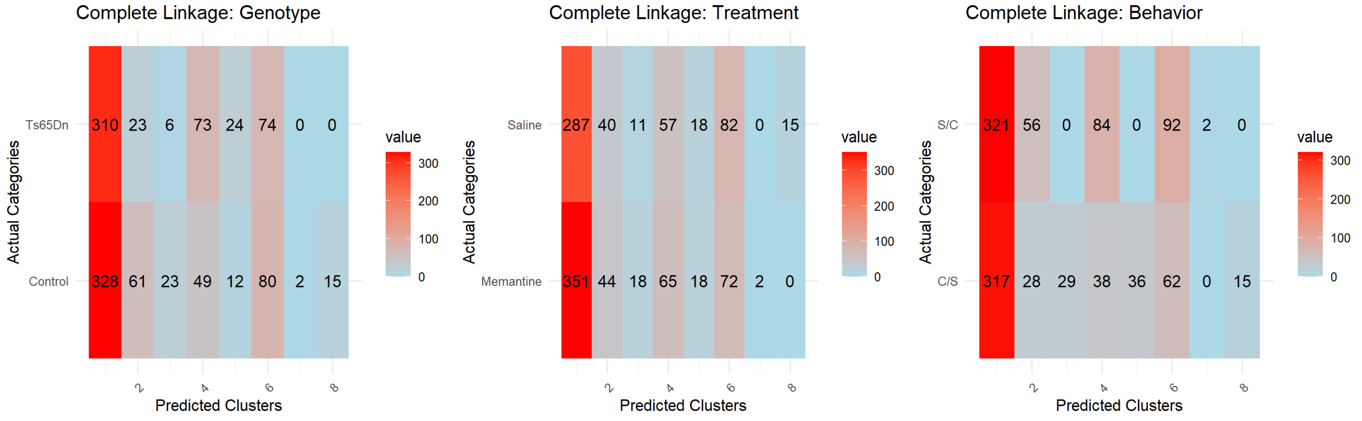

Figure 11: Heatmap for Hierarchical Clustering (Complete Linkage with Euclidean Metric) for Genotype, Treatment, and Behavior

Figure 11 shows the heatmap of the confusion matrix for genotype, treatment, and behavior categories. We only display results for complete linkage as it performs the best out of the three linkages. For genotype, both Ts65Dn and control are primarily predicted to cluster 1. For treatment, both saline and memantine are also primarily predicted to cluster 1. For behavior, both S/C and C/S are predicted to be in cluster 1 as well.

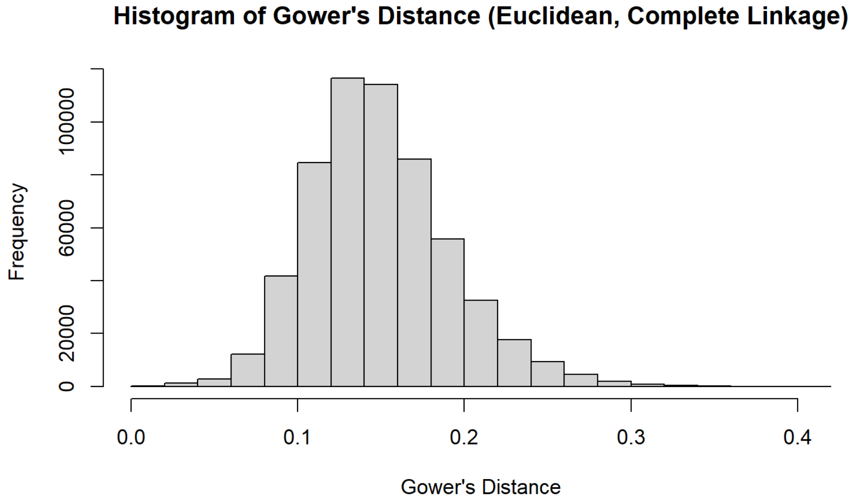

Figure 12: Histogram of Gower's coefficient for eight clusters (Hierarchical Clustering, Euclidean metric, complete linkage)

Using Gower's coefficient, Figure 12 tells us that its mean is 0.1494, showing similarity between data points. Note that we used complete linkage as it gave us the best results in terms of clusters.

We now repeat the clustering process using the Pearson correlation metric. Figure 13 provides a visual representation of how the data is grouped into eight clusters based on the first two principal components. Compared to the Euclidean distance method, the Pearson metric results in noticeably better separation between clusters across all three linkage methods as there is a clearer distinguish between data points of different colors.

Figure 14 shows the heatmap of the confusion matrix for genotype, treatment, and behavior categories. Although, as we mentioned, the Pearson metric improved separation between clusters, we will still only show the results for the complete linkage to be consistent. For genotype, Ts65Dn is evenly predicted into clusters 2, 6 and 8 while control is primarily predicted into cluster 5. For treatment, saline is primarily predicted in cluster 2 while memantine is primarily predicted in cluster 5 and 8. For behavior, S/C is evenly predicted into clusters 5 and 8 while C/S is evenly predicted into clusters 1 and 2.

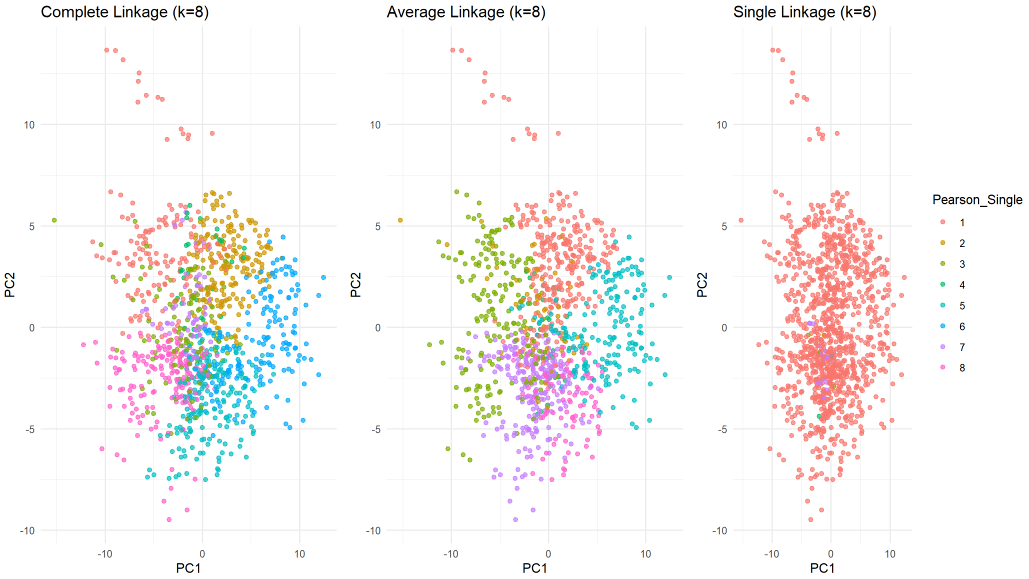

Figure 13: PCA Visualization of Hierarchical Clustering Using Pearson (Average, Complete, and Single Linkage)

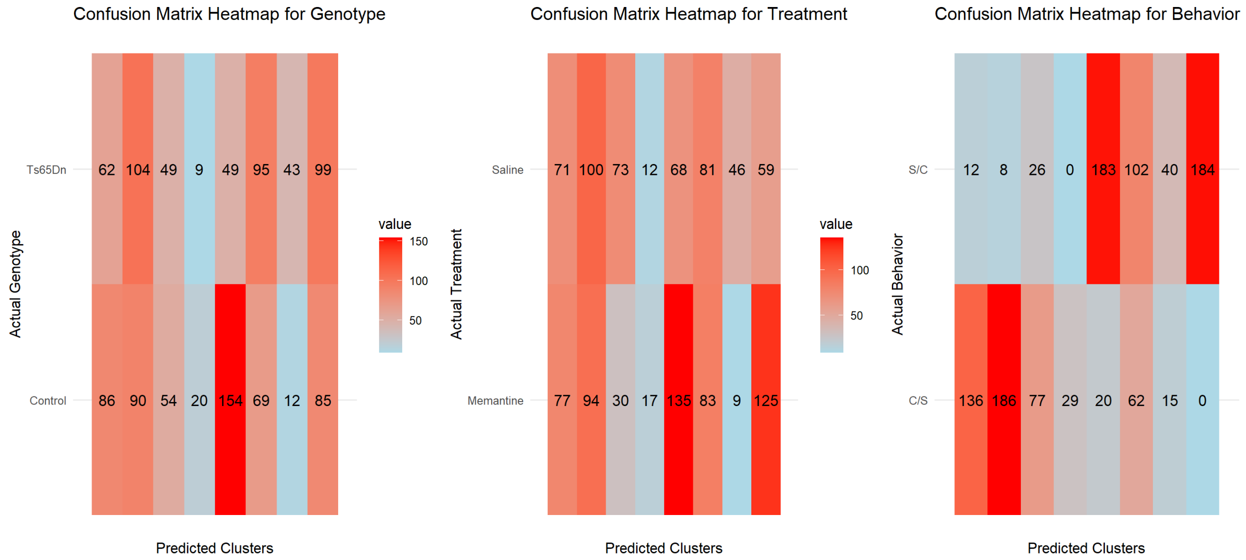

Figure 14: Heatmap for Hierarchical Clustering (Complete Linkage using Pearson metric) for Genotype, Treatment, and Behavior

We use hierarchical clustering using the Pearson metric with complete linkage, we identified the top 10 contributors for the first two principal components in figure 15. For PC1, proteins like MEK_N, BDNF_N, and NR1_N exhibit negative loadings, meaning their expression decreases as PC1 increases. In contrast, for PC2, proteins such as CaNA_N, ITSN1_N, and GSK3B_N show positive loadings, indicating that their expression increases with PC2. However, proteins like Ubiquitin_N, P38_N, and SNCA_N have negative loadings for PC2, meaning their expression decreases as PC2 increases.

Gower's coefficient mean using Pearson metric with complete linkage is 0.1494, which is similar to using Euclidean metric with complete linkage.

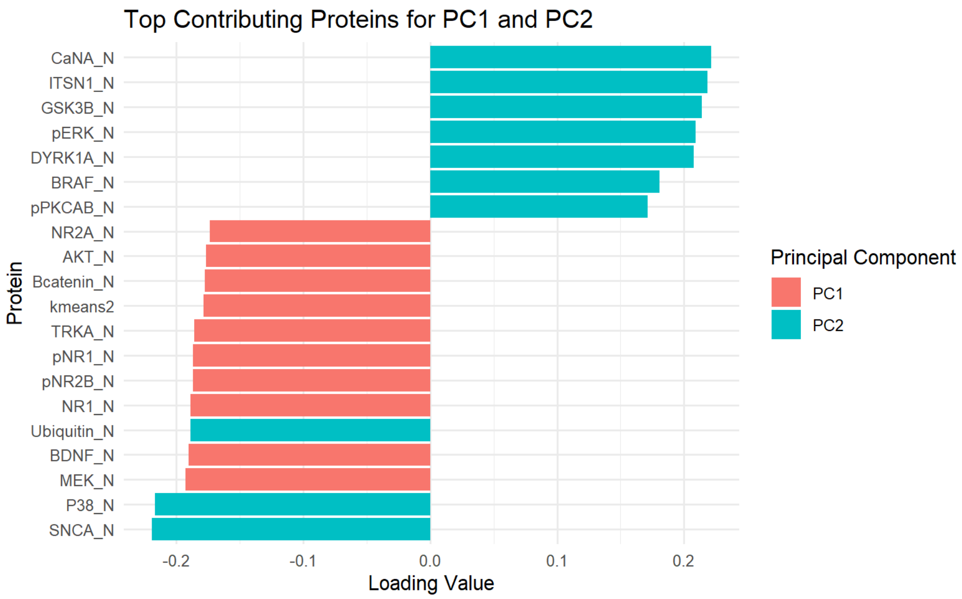

Figure 15: Top Contributing Proteins for PC1 and PC2 using Hierarchical Clustering (Pearson Metric + Complete Linkage)

2.2.3 Partitioning Around Medoids

For PAM, we first note that the silhouette plot shows that the optimal number of clusters is $k = 2$ and $k = 6$. However, I do not think that $k = 6$ clusters is particularly meaningful in our context. So I will use $k = 2$ cluster as it serves as a nice comparison to the K-means method. Additionally, $k = 2$ has the highest silhouette width.

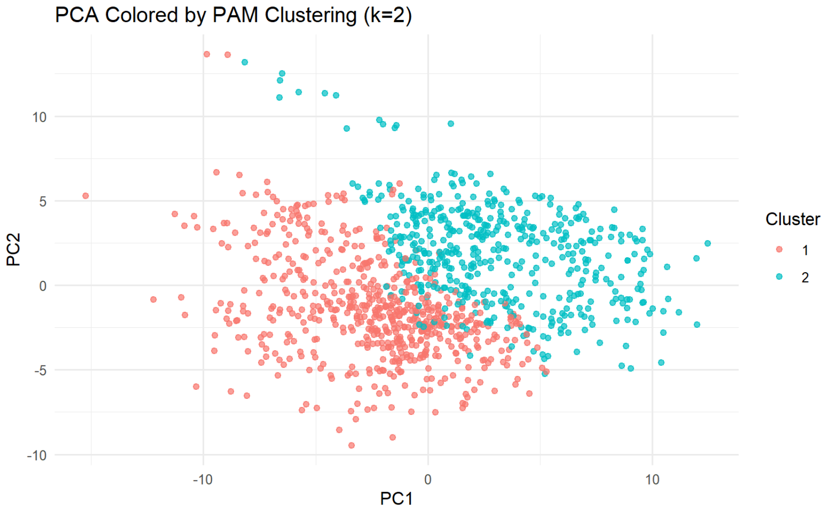

Figure 16: PAM with 2 clusters

Figure 16 shows that the two clusters seem to be farily well-separated with little overlap, suggesting the first two principal components capture enough of the variance to distinguish the two groups.

Figure 17 shows the heatmap of the confusion matrix for genotype, treatment, and behavior categories. Results show that, for genotype Ts65Dn is mainly predicted in cluster

2 while control is mainly predicted in cluster 1. For treatment, saline is evenly predicted in both clusters 1 and 2 while memantine is predicted in cluster 1. For behavior, S/C is mainly predicted in cluster 1 while C/S is mainly predicted in cluster 2.

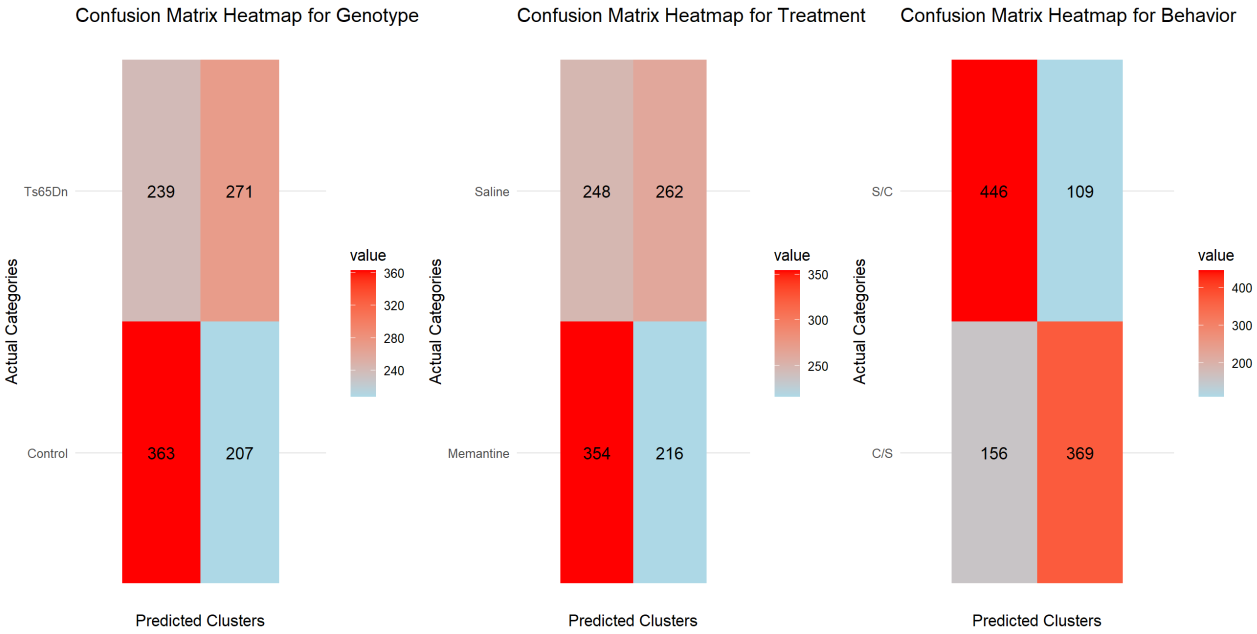

Figure 17: Heatmap for Hierarchical Clustering (Complete Linkage with Euclidean Metric) for Genotype, Treatment, and Behavior

We continue to identify the top 10 contributors for the first two principal components in figure 18. For PC1, proteins like MEK_N, BDNF_N, and NR1_N exhibit negative loadings, meaning their expression decreases as PC1 increases. In contrast, for PC2, proteins such as CaNA_N, ITSN1_N, and GSK3B_N show positive loadings, indicating that their expression increases with PC2. However, proteins like Ubiquitin_N, P38_N, and SNCA_N have negative loadings for PC2, meaning their expression decreases as PC2 increases.

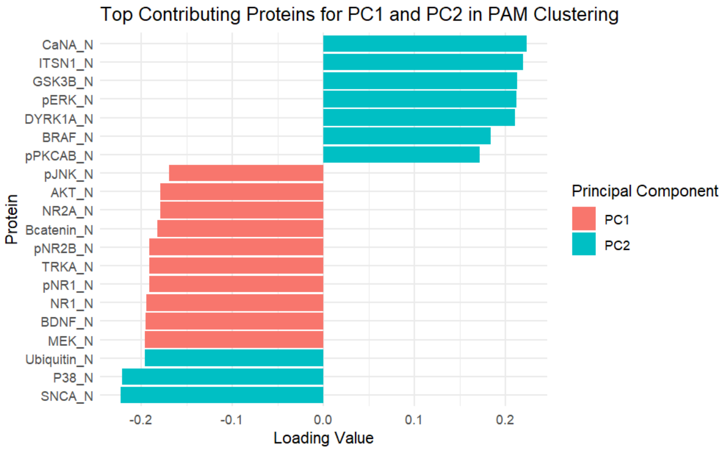

Figure 18: Top Contributing Proteins for PC1 and PC2 using two clusters (PAM)

Gower's coefficient mean is 0.1494, indicating small amount of dissimilarity.

Section 3: Discussion

In this section, we will discuss the three questions which guides our analysis, followed by the bonus question and question 2.

Q1.1: Are there any distinct clusters in the measurements?

Yes, distinct clusters were observed in the measurements. Across all three methods—K-Means, Hierarchical Clustering, and PAM—clustering on the reduced principal component space revealed clear groupings. For example, K-Means with k = 2 and k = 8 showed well-separated clusters in the PCA plots, and PAM with k = 2 showed minimal overlap. Hierarchical clustering with Pearson distance also demonstrated strong visual separation, especially when using complete linkage. Silhouette analysis supported the presence of distinct clusters, particularly for k = 2 and k = 8, reinforcing that natural grouping structures exist in the data. However, the other methods such as hierarchical clustering with Euclidean distance did not show as strong of distinct clusters in comparison.

Q1.2: Are these clusters associated with any of the different genotype, behaviour and/or treatment?

To some extent, yes. While none of the clustering methods perfectly aligned with the known categories, there were observable associations. For instance,

- Genotype: PAM and K-Means with k = 2 showed some distinction between control and Ts65Dn mice, with Ts65Dn more commonly found in cluster 2 while control mice more commonly found in cluster 1.
- Behavior: In K-Means with k = 8, S/C mice were more commonly found in cluster 5 while C/S mice was more commonly found in cluster 4.
- Treatment: In hierarchical clustering using Pearson metric, saline mice were commonly found in cluster 2 while memantine mice were more commonly found in clusters 5 and 8.

Hierarchical clustering using Pearson distance and complete linkage showed the most meaningful alignment with these experimental factors. However, to determine if there is statistical significance, one should proceed to test significance with perhaps a Chi-squared Pearson test.

Q1.3: Do we see any particular set of proteins that exhibit a distinct expression pattern (profile) in any or all of these clusters?

Yes, specific proteins contributed significantly to the clustering patterns:

- The top contributors to the first two principal components were consistently identified across methods, including MEK N, BDNF N, NR1 N (strong negative loadings on PC1), and CaNA N, ITSN1 N, GSK3B N (positive loadings on PC2)
- These proteins demonstrated varying expression levels across clusters, suggesting they may underlie some of the group separation seen in the PCA plots.
- Heatmaps and loading plots highlighted distinct expression profiles, with several proteins consistently separating the clusters along key principal components.

Q2: Think of ways that you could use some of the unsupervised learning methods in problems you are currently facing or have recently faced in your research or work. Explain how and why they can be helpful in investigating the problems and answering the questions you are dealing with.

I am recently working on two projects where I think unsupervised learning methods can be extremely useful.

The first project I am working on involves analyzing biomarkers in systemic juvenile idiopathic arthritis (sJIA). In this project, we are particularly focused on identifying and understanding the role of biomarkers in disease progression and treatment response. One of the challenges in studying sJIA is the complexity of the data, which includes four biomarkers, dozens of clinical features, and genetic data. These datasets are often high-dimensional, making it difficult to interpret patterns and relationships using traditional analysis techniques. Unsupervised learning methods, such as clustering and PCA, are especially useful for addressing these problems. K-means or hierarchical clustering can allow us to group patients based on their biomarker profiles and/or clinical features. I can see this being particularly valuable during exploratory data analysis, as it helps us understand how similar or different these features are across patients. PCA on the other hand, enables us to reduce the dimensionality of the data while retaining the most significant sources of variance. This technique helps identify which biomarkers/features contribute most to the differences between patients.

The second project which is a part of my practicum focuses on analyzing inpatient survey responses to understand the factors affecting patient experience in Ontario hospitals. This project aims to identify key variables that influence how patients perceive the care they receive, which can help improve patient satisfaction, outcomes, and overall healthcare quality. Given the large volume of survey responses and the diversity of factors that contribute to patient experiences, I can imagine that unsupervised learning methods like clustering and PCA can be particularly beneficial. We could use K-means or hierarchical clustering to group patients with similar responses, which allows us to identify different patient profiles or subgroups that share common experiences. PCA can also be applied to reduce the dimensionality of the dataset. By identifying principal components, we can pinpoint the most influential factors affecting patient satisfaction.

Bonus Question: In the data we consider the samples as independent. The reality is that most likely they are not, since multiple measurements are coming from the same mouse. Are there any implications about this? Discuss possible ways if any that you could improve the overall analysis plan by taking into consideration this potential correlatedness in the data. Note: I am not looking for any particular right answer here. You could do some research in order to answer this question.

When the samples are not independent, as in the case of repeated measures on the same individual (mice in this case), the data may exhibit within-subject correlation. This could lead to biased statistical results, especially if the correlated nature of the data is not accounted for in the analysis. For example, when if we take multiple measurements from the same mouse, there may be inherent similarities or shared variability across those measurements, which could inflate the type I error rate (false positives) in traditional analyses that assume independence. In other words, you might find associations or patterns that appear significant but are actually a result of shared variance between the repeated measurements within the same animal, rather than a true effect.

One approach to take could be the use of mix-effect models. In these models, the repeated measurements from the same mouse subject are treated as ”nested” within that subject, and the model includes both fixed effects and random effects. This way, we can model the individual variability between subjects and the repeated measurements within subjects separately, reducing the risk of inflated significance levels.

Another approach could involve using generalized estimating equations (GEE). GEEs provide a method for estimating the parameters of a model while accounting for within-subject correlation, which is especially useful for analyzing correlated data that might not follow a normal distribution. This approach can help provide more accurate standard errors and confidence intervals when repeated measures are involved.

Another method could involve the use of bootstrapping methods or cross-validation techniques that allow for subject-level resampling, rather than treating each measurement as independent, might also provide more robust and reliable estimates.

Section 4: Conclusion

In this exploratory analysis of protein expression data using unsupervised learning, we applied three clustering techniques—K-Means, Hierarchical Clustering, and Partitioning Around Medoids (PAM)—to uncover underlying patterns and groupings within the data.

Hierarchical Clustering using Pearson correlation with complete linkage demonstrated the best overall performance. It provided better cluster separation, more meaningful alignment with known class labels, and clearer biological interpretability. Therefore, for high-dimensional biological data such as protein expression, hierarchical clustering with Pearson distance is recommended as the most effective method in this context.