DSL251
Data Analytics and Visualization
Homework 3

Amay Dixit
12340220

Google Colab Notebook Link
https://colab.research.google.com/drive/
1xmye8SiFHBd7Q1njTCQr2zVWI1KtCea0?usp=sharing

1 Neighbor Distribution and Density Distribution Analysis

This section presents a comprehensive analysis of the point distributions and density patterns in the dataset. Multiple visualization techniques were employed to understand the spatial relationships between data points.

1.1 Dataset Summary Statistics

The dataset contains measurements of width and length, with the following key statistics:

Statistic	Width	Length
Count	70.000	70.000
Mean	2.290	13.793
Standard Deviation	1.603	6.166
Minimum	0.500	1.100
25th Percentile	1.400	8.100
Median	1.800	15.150
75th Percentile	2.800	18.075
Maximum	12.100	24.600

Table 1: Summary statistics of the dataset

1.2 Nearest Neighbor Analysis

The nearest neighbor analysis revealed:

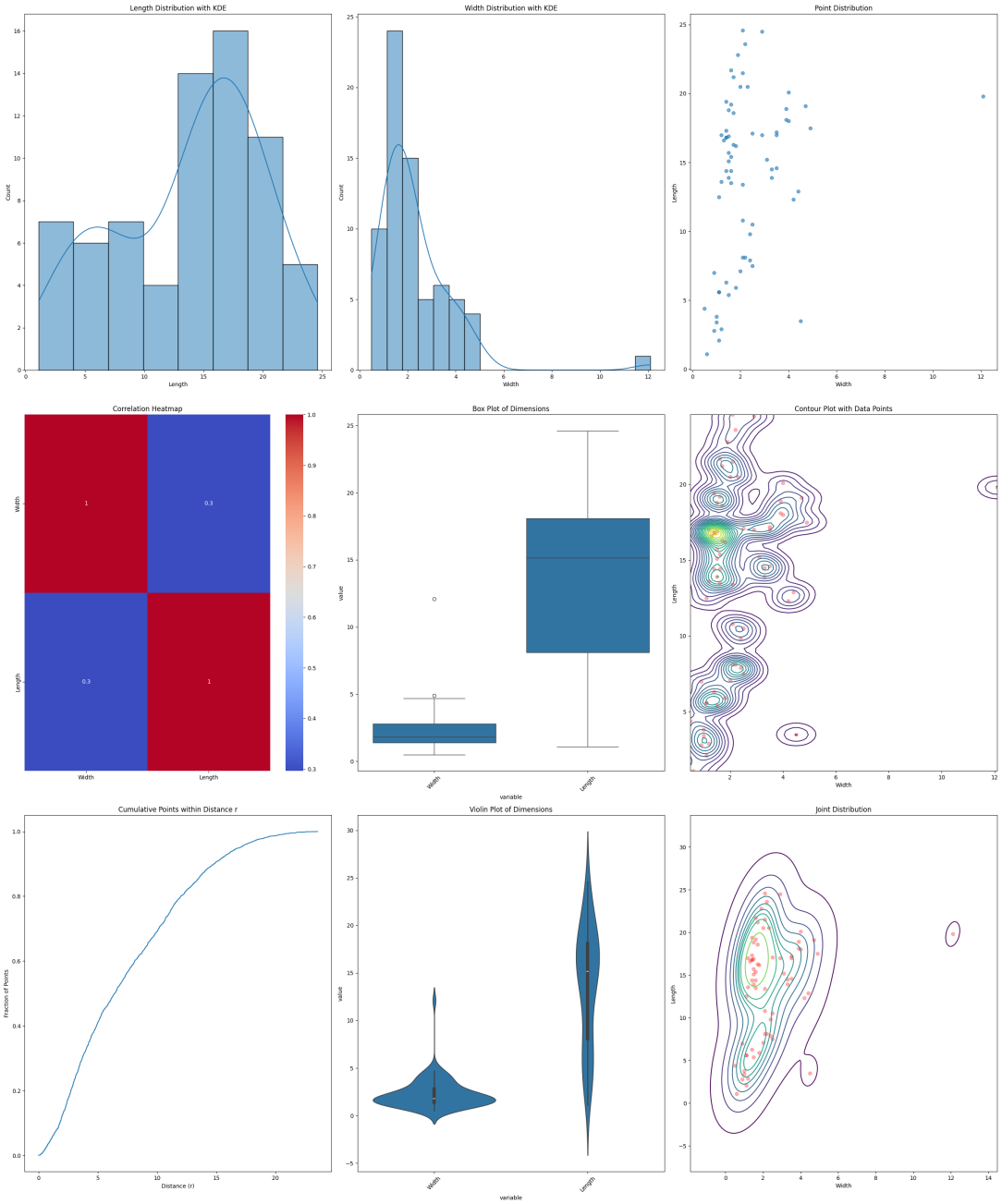
- Average nearest neighbor distance: 0.59
- Median nearest neighbor distance: 0.41

- Standard deviation of distances: 0.94

1.3 Distribution Characteristics

The analysis included multiple visualizations:

1. Point Distribution Plot: Revealed the spatial arrangement of data points in the width-length space
2. Distance Matrix Heatmap: Showed the pairwise distances between all points
3. K-Nearest Neighbor Distance Distribution: Illustrated the distribution of distances to the 3 nearest neighbors
4. Kernel Density Estimation: Visualized the density of points across the feature space



2 Outlier/Noise Detection using DBSCAN

2.1 Parameter Selection

Based on the k-distance graph analysis and the density distribution plots from Question 1, we selected the following DBSCAN parameters:

- Epsilon (eps) = 0.5: This value was chosen based on the elbow point in the k-distance graph, where we observed a significant change in the slope of the distance curve.
- Minimum Points (min_samples) = 7: This parameter was selected to ensure robust cluster formation while maintaining sensitivity to noise detection.

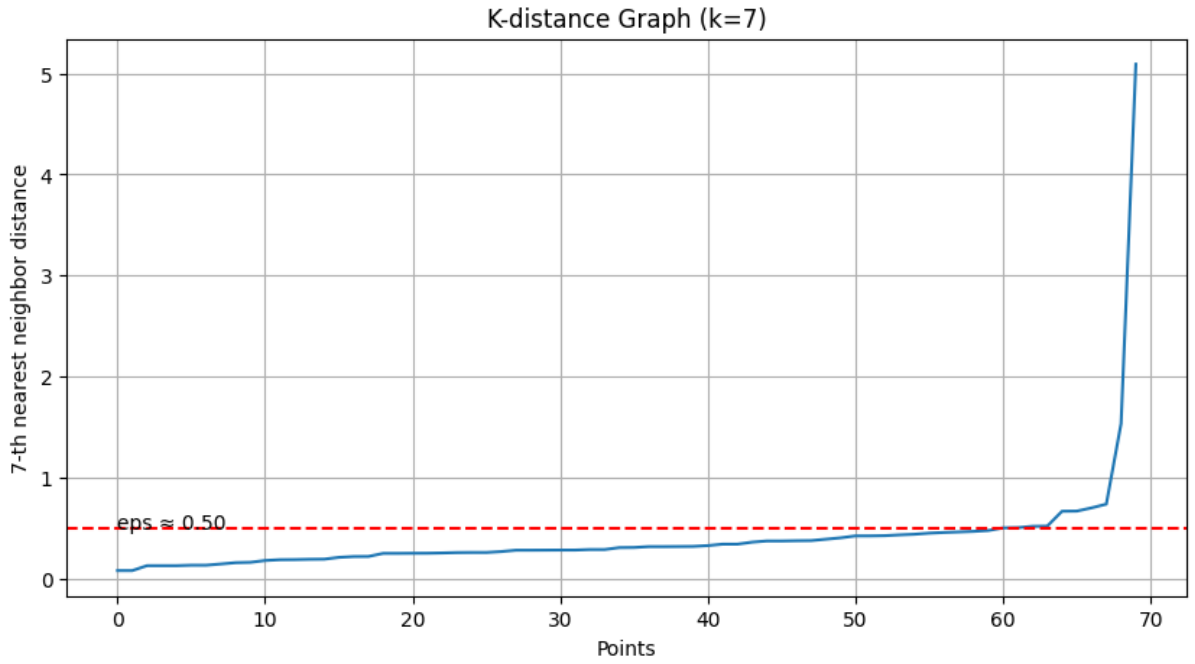


Figure 1: Parameter Selection

2.2 Implementation Results

The DBSCAN algorithm was applied to the standardized dataset, resulting in the following classification:

Point Type	Count
Core Points	50
Border Points	14
Noise Points	6

Table 2: DBSCAN Classification Summary

2.3 Identified Noise Points

The algorithm identified six points as noise, with the following coordinates:

Index	Width	Length
5	4.5	3.5
22	4.2	12.3
24	4.4	12.9
50	4.9	17.5
59	12.1	19.8
68	2.9	24.5

Table 3: Identified Noise Points

2.4 Visualization

A scatter plot was created to visualize the DBSCAN results, with:

- Green points representing core points
- Yellow points representing border points
- Red points representing noise points

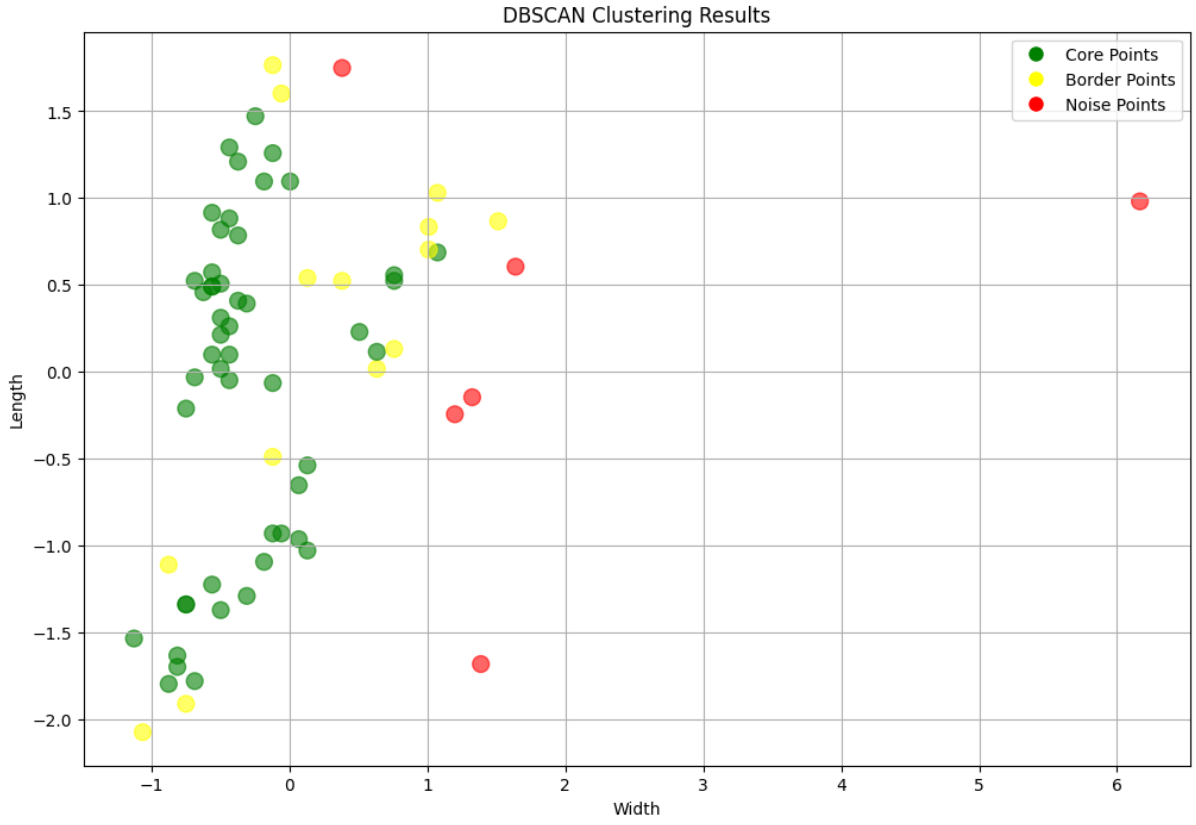


Figure 2: DBSCAN Results

2.5 Analysis

The DBSCAN implementation effectively identified outliers in the dataset:

- Most noise points were found at the periphery of the data distribution
- The identified noise points show unusual width-to-length ratios compared to the main clusters
- The algorithm successfully maintained the core structure of the data while isolating anomalous points

3 DBSCAN Clustering Analysis

3.1 Clustering Results

The DBSCAN algorithm identified three distinct clusters in the dataset, with the following distribution:

Cluster	Number of Points
Cluster 0	21
Cluster 1	31
Cluster 2	12
Noise	6

Table 4: Distribution of Points Across Clusters

3.2 Cluster Quality Metrics

The clustering quality was evaluated using several metrics:

- Silhouette Score: 0.538, indicating moderately well-separated clusters
- Average intra-cluster distances:
  - Cluster 0: 0.747
  - Cluster 1: 0.689
  - Cluster 2: 0.563

3.3 Inter-cluster Relationships

Analysis of the distances between clusters revealed:

From	To	Average Distance
Cluster 0	Cluster 1	1.956
Cluster 0	Cluster 2	2.238
Cluster 1	Cluster 2	1.392

Table 5: Inter-cluster Distances

3.4 Cluster Characteristics

1. **Cluster Separation**:
    - Clusters 0 and 2 show the highest separation (distance: 2.238)
    - Clusters 1 and 2 are relatively closer (distance: 1.392)
    - The moderate silhouette score suggests distinct but not completely isolated clusters
2. **Cluster Cohesion**:
    - Cluster 2 shows the highest cohesion (lowest intra-cluster distance: 0.563)
    - Cluster 0 has the lowest cohesion (highest intra-cluster distance: 0.747)

3.5 Visualization

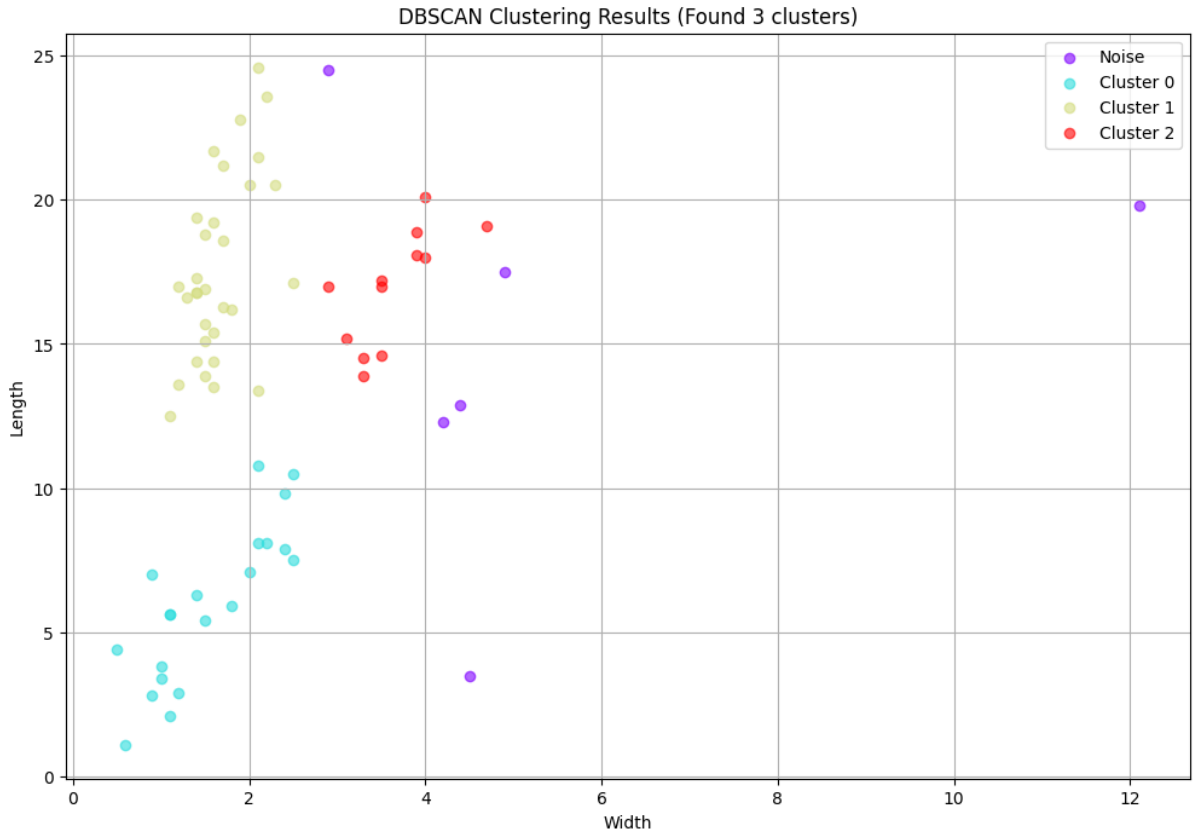


Figure 3: DBSCAN Clustering Results showing three distinct clusters and noise points

3.6 Interpretation

The DBSCAN clustering analysis reveals:

- Three well-defined clusters with distinct characteristics
- Moderate overlap between clusters, particularly between Clusters 1 and 2
- Clear separation of noise points from the main clusters

- A hierarchical structure where Cluster 0 is most isolated, while Clusters 1 and 2 show some proximity

The clustering results suggest natural groupings in the data that could correspond to different categories of leaves, with some boundary cases represented by the noise points.

4 Comparison of K-means and DBSCAN Clustering

4.1 Experimental Setup

Three clustering approaches were implemented and compared:

1. DBSCAN clustering (from Section 3)
2. K-means clustering on the original dataset
3. K-means clustering on the dataset with noise points removed

4.2 Clustering Results

4.2.1 K-means on Original Data

Cluster	Size	Center Coordinates		Spread	
		Width	Length	Width	Length
0	47	2.37	17.36	1.08	3.13
1	1	12.10	19.80	0.00	0.00
2	22	1.67	5.89	0.88	2.65

Table 6: K-means Clustering Statistics (Original Data)

4.2.2 K-means on Noise-Removed Data

Cluster	Size	Center Coordinates		Spread	
		Width	Length	Width	Length
0	31	1.66	17.59	0.34	3.19
1	12	3.63	16.97	0.47	1.93
2	21	1.54	6.00	0.65	2.66

Table 7: K-means Clustering Statistics (Noise-Removed Data)

4.3 Impact of Noise Removal

The comparison reveals several key differences between the clustering approaches:

4.3.1 Cluster Size and Distribution

- **Original Data**: Shows highly imbalanced clusters (47-1-22 distribution)
- **Noise-Removed Data**: Exhibits more balanced clustering (31-12-21 distribution)

4.3.2 Cluster Characteristics

- **Spread Reduction**:
  - Width spread decreased significantly (from 1.08 to 0.34 in the largest cluster)
  - Length spread remained relatively stable
- **Center Shifts**:
  - Cluster centers are more representative after noise removal
  - Elimination of singleton cluster (previous Cluster 1 with single point)

4.4 Visualization

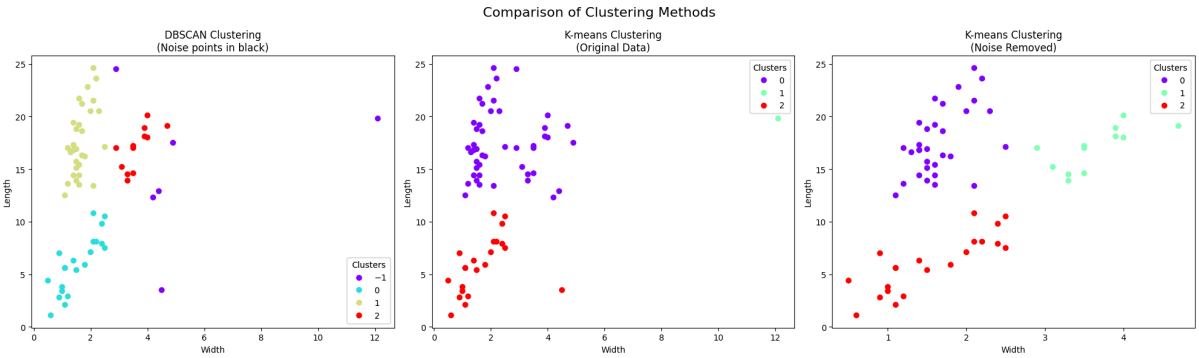


Figure 4: Comparison of DBSCAN and K-means clustering results with and without noise removal

4.5 Analysis of Methods

1. **DBSCAN Advantages**:
   - Automatically identifies and isolates noise points
   - Creates more natural cluster shapes
   - No assumption about cluster sizes
2. **K-means Characteristics**:
   - Sensitive to outliers in original data
   - Produces more compact, spherical clusters
   - Improved performance after noise removal

4.6 Conclusion

The removal of noisy points significantly affects the K-means clustering solution:

- Improves cluster balance and representation
- Reduces within-cluster spread

- Creates more meaningful and interpretable clusters
- Eliminates the impact of outliers on cluster centers

This comparison demonstrates the importance of noise removal in clustering applications and the complementary nature of DBSCAN and K-means algorithms.

[Previous sections remain the same...]

5 K-means++ Clustering Analysis

5.1 Implementation Results

K-means++ was implemented with 4 clusters, using the improved initialization method to optimize centroid placement. The algorithm converged in 3 iterations.

5.2 Cluster Statistics

Cluster	Size	Center Coordinates		Spread		Inertia
		Width	Length	Width	Length	
0	32	-0.373	0.656	0.249	0.549	11.615
1	22	-0.388	-1.291	0.554	0.433	10.889
2	15	0.953	0.429	0.352	0.375	3.964
3	1	6.165	0.981	0.000	0.000	0.000

Table 8: K-means++ Clustering Statistics

5.3 Clustering Quality

- Total Inertia: 26.47
- Fast convergence (3 iterations)
- Cluster distribution:
  - Main cluster (0): 32 points, moderate spread
  - Secondary cluster (1): 22 points, larger width spread
  - Tertiary cluster (2): 15 points, balanced spread
  - Singleton cluster (3): 1 point, isolated outlier

5.4 Visualization

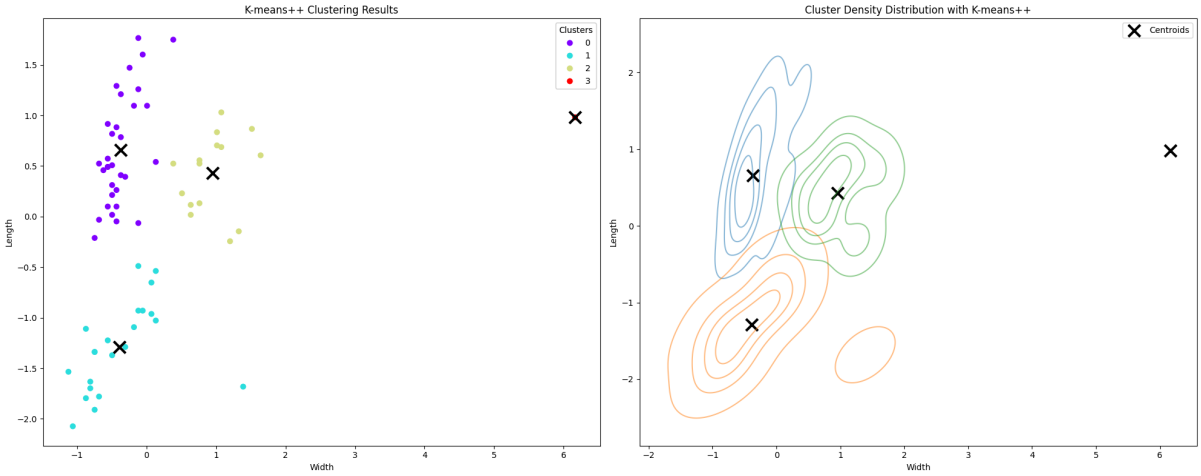


Figure 5: K-means++ clustering results showing cluster assignments and density distribution

5.5 Analysis of Results

5.5.1 Cluster Characteristics

1. **Main Cluster (0)**:
    - Largest group with 32 points
    - Compact spread in both dimensions
    - Centrally located in the feature space
2. **Secondary Cluster (1)**:
    - 22 points with higher width variation
    - Distinct negative length center
    - Moderate inertia despite size
3. **Tertiary Cluster (2)**:
    - 15 points with balanced spread
    - Positive width center
    - Lowest non-zero inertia
4. **Singleton Cluster (3)**:
    - Single point at extreme width
    - Likely represents an outlier
    - Zero inertia due to single point

5.5.2 Density Distribution

The density plot reveals:

- Clear separation between main cluster groups
- Overlapping regions between clusters 0 and 2
- Distinct isolation of the singleton cluster
- Varying density patterns across clusters

5.6 Advantages of K-means++

- Improved initial centroid placement
- Fast convergence (3 iterations)
- Clear cluster separation where data structure permits
- Effective handling of varying cluster densities

5.7 Limitations

- Still sensitive to outliers (singleton cluster)
- Fixed number of clusters required a priori
- Assumes spherical cluster shapes

6 Hierarchical Clustering Analysis Using Ward's Method

6.1 Implementation Methodology

Ward's hierarchical clustering method was implemented to analyze both the original dataset and the noise-removed dataset. The analysis included dendrograms, cluster visualizations, and density distributions to provide a comprehensive comparison of how noise affects the hierarchical structure.

6.2 Comparative Analysis of Clustering Results

	Without Noise		With Noise	
Characteristic	Value	Spread	Value	Spread
Cluster 1				
Size	21	–	22	–
Center Width	1.54	0.65	1.67	0.88
Center Length	6.00	2.66	5.89	2.65
Cluster 2				
Size	13	–	47	–
Center Width	3.55	0.54	2.37	1.08
Center Length	16.98	1.86	17.36	3.13
Cluster 3				
Size	30	–	1	–
Center Width	1.63	0.31	12.10	0.00
Center Length	17.61	3.24	19.80	0.00

Table 9: Comparison of Clustering Statistics With and Without Noise

6.3 Structural Changes Due to Noise

1. **Cluster Size Distribution**
   - *Without Noise*: Relatively balanced (21-13-30)
   - *With Noise*: Highly imbalanced (22-47-1)
2. **Cluster Centers**
   - *Without Noise*: Well-distributed centers with reasonable separation
   - *With Noise*: One cluster reduced to a single outlier point
3. **Spread Characteristics**
   - *Without Noise*: Consistent and moderate spread across clusters
   - *With Noise*: Large variation in spread, with one degenerate cluster

6.4 Visual Analysis

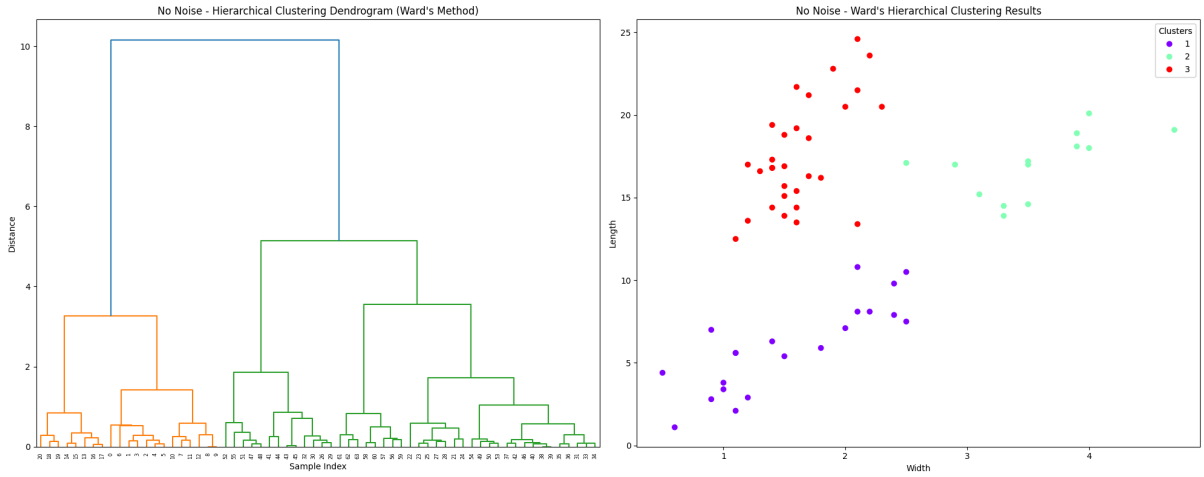


Figure 6: Dendrogram and Clustering (Without Noise)

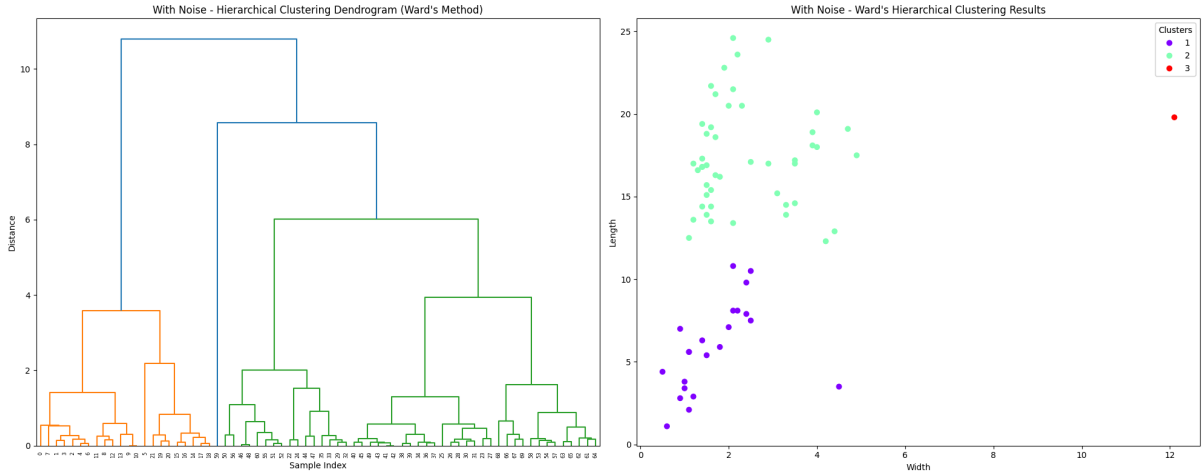
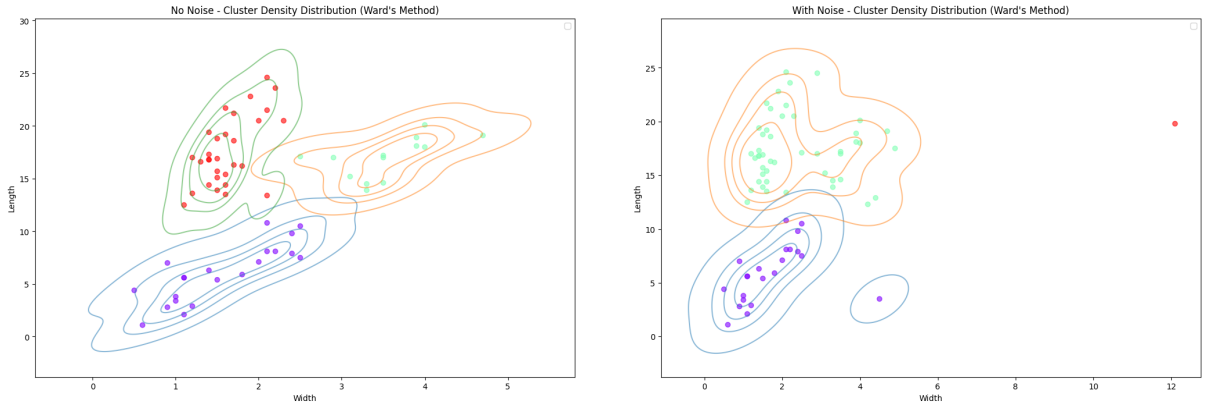


Figure 7: Dendrogram and Clustering (With Noise)



(a) Density Distribution (Without Noise) (b) Density Distribution (With Noise)

Figure 8: Comparison of Cluster Density Distributions

6.5 Key Observations

1. **Dendrogram Structure**
   - *Without Noise*: Shows clear, well-defined hierarchical relationships
   - *With Noise*: Exhibits more extreme distance variations and unbalanced merging
2. **Cluster Separation**
   - *Without Noise*: Distinct clusters with moderate overlap
   - *With Noise*: One highly isolated point and two less distinct clusters
3. **Density Distribution**
   - *Without Noise*: Smooth, continuous density contours
   - *With Noise*: Disrupted density patterns with isolated regions

6.6 Impact of Noise Removal

The comparison demonstrates that noise removal leads to:

- More balanced cluster sizes
- Better-defined cluster boundaries
- More interpretable hierarchical structure
- More reliable center and spread estimates
- Improved density distribution patterns

7 Separating Hyperplanes Analysis

7.1 Methodology

Using the hierarchical clustering results from Question 6, we applied the Learning with Prototypes approach to determine the separating hyperplanes between the three clusters. The process involved:

1. Computing cluster prototypes (centroids)
2. Calculating hyperplane equations between each pair of clusters
3. Visualizing the separating boundaries

7.2 Results

7.2.1 Data Without Noise

The separating hyperplanes for the noise-removed dataset are:

$$\begin{aligned}
&\text{Hyperplane 1-2:} \quad 0.576x + 0.818y + 0.216 = 0 \\
&\text{Hyperplane 1-3:} \quad 0.030x + 1.000y + 0.338 = 0 \\
&\text{Hyperplane 2-3:} \quad -0.996x + 0.085y + 0.138 = 0
\end{aligned}$$

7.2.2 Data With Noise

The separating hyperplanes for the original dataset (including noise) are:

$$\begin{aligned} \text{Hyperplane 1-2:} \quad & 0.228x + 0.974y + 0.383 = 0 \\ \text{Hyperplane 1-3:} \quad & 0.945x + 0.328y - 2.678 = 0 \\ \text{Hyperplane 2-3:} \quad & 0.998x + 0.065y - 3.152 = 0 \end{aligned}$$

7.3 Visualization

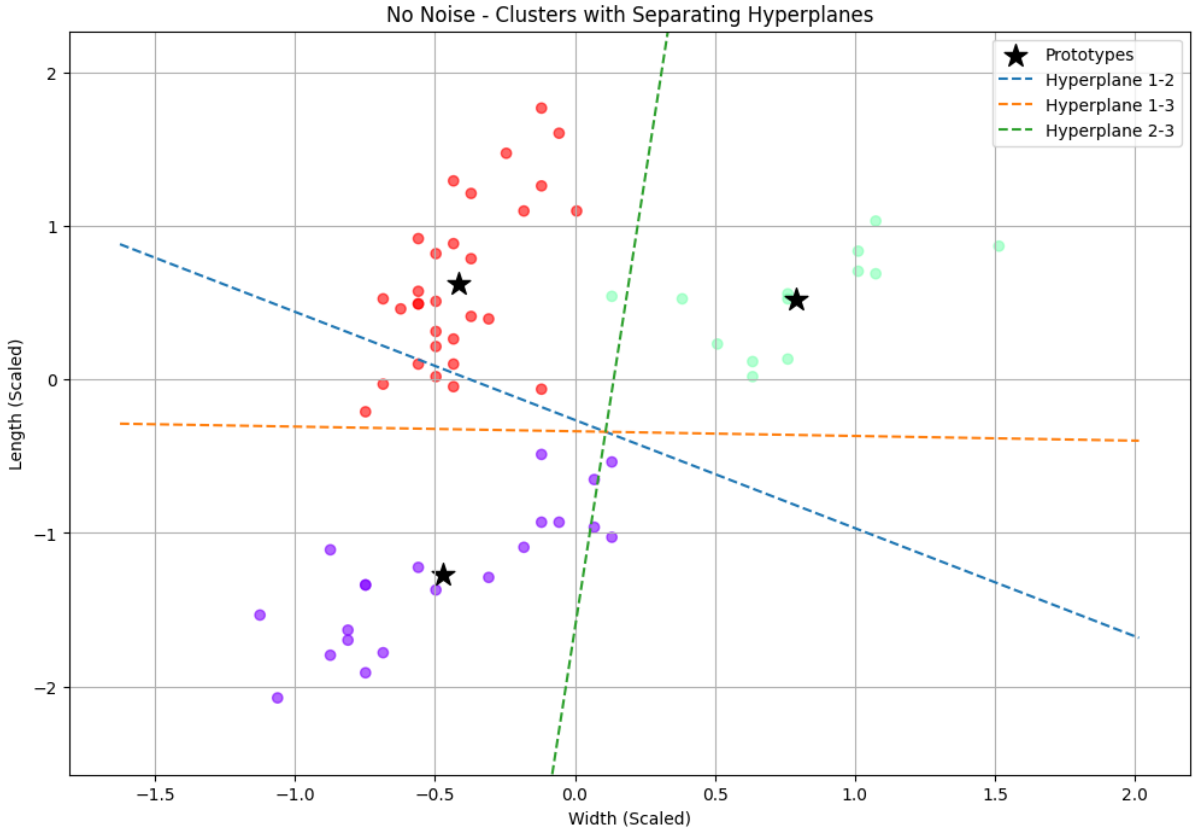


Figure 9: Clusters and Separating Hyperplanes (Without Noise)

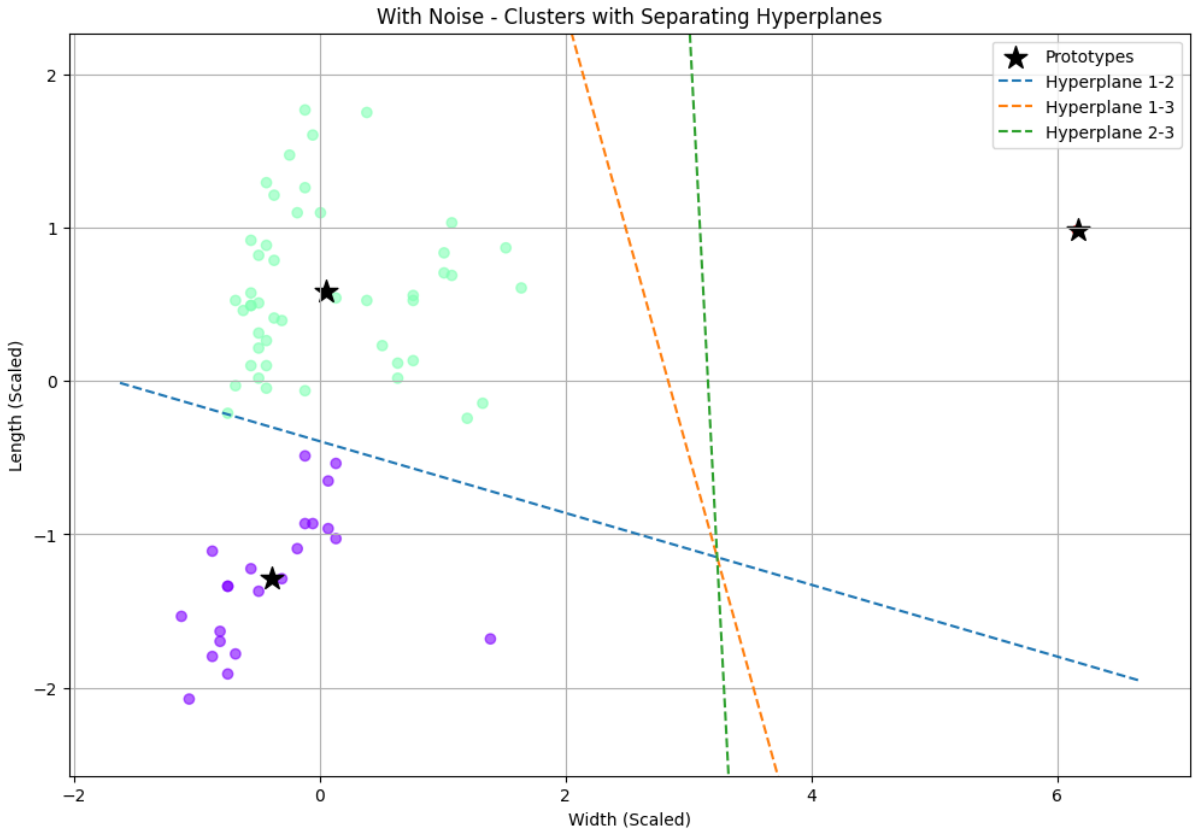


Figure 10: Clusters and Separating Hyperplanes (With Noise)

7.4 Analysis

7.4.1 Without Noise

- The hyperplanes effectively separate the three clusters with minimal overlap
- Hyperplane 1-2 shows a diagonal boundary with balanced coefficients
- Hyperplane 1-3 is nearly horizontal, indicating separation primarily based on the y-coordinate
- Hyperplane 2-3 is almost vertical, suggesting separation mainly based on the x-coordinate

7.4.2 With Noise

- The presence of noise significantly affects the hyperplane orientations
- Hyperplane 1-2 becomes more vertically oriented
- Hyperplanes 1-3 and 2-3 show substantial shifts due to outlier influence
- The separation boundaries are less optimal compared to the noise-free case

7.5 Impact of Noise on Hyperplane Formation

The comparison reveals several key differences:

1. **Coefficient Changes**:
    - Without noise: More balanced coefficients indicating natural boundaries
    - With noise: More extreme coefficients suggesting distorted boundaries
2. **Boundary Orientation**:
    - Without noise: Hyperplanes align with natural cluster separations
    - With noise: Hyperplanes show more extreme angles and positions
3. **Separation Quality**:
    - Without noise: Clear, well-defined separation between clusters
    - With noise: More ambiguous boundaries with potential misclassification regions

7.6 Classification Implications

The hyperplane equations provide a basis for classifying new points:

- A point's position relative to all hyperplanes determines its cluster
- The noise-free model offers more reliable classification boundaries
- The presence of noise introduces uncertainty in boundary regions

8 Multivariate Gaussian Distribution Analysis

8.1 Methodology

Using the hierarchical clustering results from the previous sections, we fitted multivariate Gaussian distributions to each cluster. For each cluster, we:

1. Calculated the mean vector (μ)
2. Computed the covariance matrix ()
3. Visualized the resulting probability density functions

8.2 Results

Cluster 1

Parameter	**Without Noise**	**With Noise**
Size	n = 21	n = 22
Mean Vector	Width = −0.4725 Length = −1.2723	Width = −0.3879 Length = −1.2909
Covariance Matrix	$\begin{bmatrix} 0.1728 & 0.1528 \\ 0.1528 & 0.1983 \end{bmatrix}$	$\begin{bmatrix} 0.3220 & 0.1109 \\ 0.1109 & 0.1965 \end{bmatrix}$

Cluster 2

Parameter	**Without Noise**	**With Noise**
Size	n = 13	n = 47
Mean Vector	Width = 0.7894 Length = 0.5202	Width = 0.0504 Length = 0.5834
Covariance Matrix	$\begin{bmatrix} 0.1261 & 0.0644 \\ 0.0644 & 0.0995 \end{bmatrix}$	$\begin{bmatrix} 0.4739 & -0.0003 \\ -0.0003 & 0.2664 \end{bmatrix}$

Cluster 3

Parameter	**Without Noise**	**With Noise**
Size	n = 30	n = 1
Mean Vector	Width = −0.4148 Length = 0.6230	Width = 6.1650 Length = 0.9813
Covariance Matrix	$\begin{bmatrix} 0.0382 & 0.0632 \\ 0.0632 & 0.2892 \end{bmatrix}$	$\begin{bmatrix} 0.0100 & 0.0000 \\ 0.0000 & 0.0100 \end{bmatrix}$

8.3 Visualization

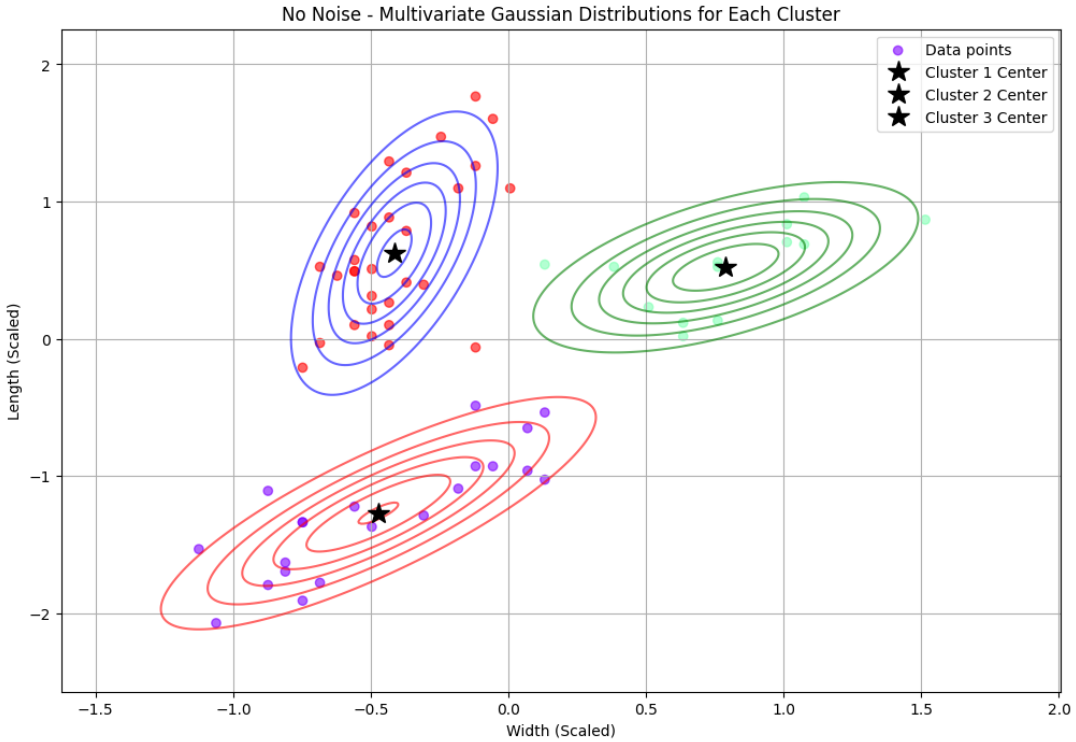


Figure 11: Multivariate Gaussian Distributions (Without Noise)

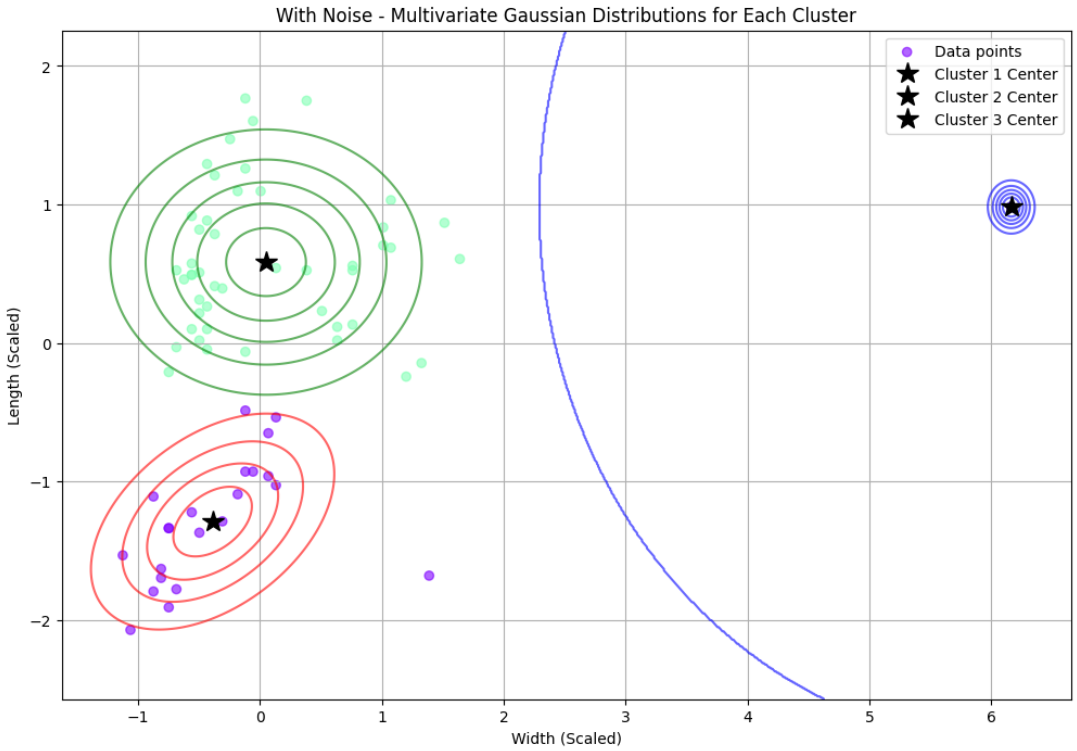


Figure 12: Multivariate Gaussian Distributions (With Noise)

8.4 Analysis

8.4.1 Cluster Characteristics Without Noise

- Well-defined clusters with distinct means and reasonable covariance structures
- Moderate correlation between width and length in Cluster 1 (covariance = 0.1528)
- Cluster 3 shows higher variance in length (0.2892) compared to width (0.0382)
- Cluster 2 exhibits the most compact distribution with smallest variances

8.4.2 Impact of Noise

- Significant increase in variance for Cluster 2 (width variance from 0.1261 to 0.4739)
- Creation of a singleton cluster with minimal artificial covariance
- Higher uncertainty in cluster boundaries
- Reduced correlation between dimensions in Cluster 2 (near-zero covariance)

8.5 Implications for Classification

The Gaussian parameters provide a probabilistic framework for classification:

- Points can be assigned to clusters based on maximum likelihood
- Noise-free model provides more reliable probability estimates
- Presence of singleton cluster in noisy data requires special handling
- Overlapping distributions indicate potential classification uncertainty regions

9 Classification Analysis of Test Point

9.1 Overview

We analyzed the classification of the test point (1.9, 6) using three distinct methods:

1. Maximum Likelihood Estimation (MLE)
2. Hyperplane-based classification
3. K-Nearest Neighbors (K-NN)

The analysis was performed on both the original dataset (with noise) and the noise-removed dataset to understand the impact of noise on classification results.

9.2 Classification Results

9.2.1 Maximum Likelihood Estimation (MLE)

	Without Noise	**With Noise**
Classification	Cluster 1	Cluster 1
Cluster 1 Likelihood	3.117×10^{-1}	2.140×10^{-1}
Cluster 2 Likelihood	3.087×10^{-8}	4.255×10^{-4}
Cluster 3 Likelihood	1.617×10^{-6}	0.000

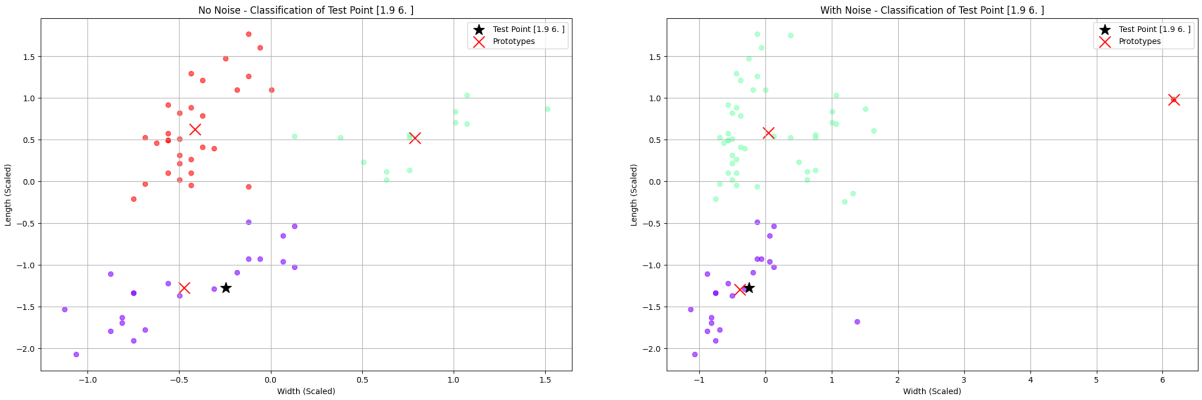
9.2.2 Hyperplane-Based Classification

	Without Noise	**With Noise**
Classification	Cluster 1	Cluster 1
Distance to Cluster 1	0.2274	0.1439
Distance to Cluster 2	2.0702	1.8798
Distance to Cluster 3	1.9037	6.7950

9.2.3 K-Nearest Neighbors (k=5)

	Without Noise	**With Noise**
Classification	Cluster 1	Cluster 1
Neighbor Classes	[1, 1, 1, 1, 1]	[1, 1, 1, 1, 1]
Distances	[0.0649, 0.1904, 0.2698, 0.3180, 0.3654]	[0.0649, 0.1904, 0.2698, 0.3180, 0.3654]

9.3 Visualization



(a) Classification Results (Without Noise) (b) Classification Results (With Noise)

Figure 13: Visualization of test point classification using different methods

9.4 Analysis

9.4.1 Consensus Among Methods

All three classification methods consistently assigned the test point (1.9, 6) to Cluster 1, both with and without noise, indicating:

- Strong agreement across different classification approaches
- Robust classification despite different underlying assumptions
- High confidence in the final classification

9.4.2 Method-Specific Insights

1. **MLE Method:**
    - Shows strongest discrimination between clusters
    - Likelihood values clearly favor Cluster 1
    - Noise affects absolute likelihood values but not final classification
2. **Hyperplane Method:**
    - Demonstrates clear separation between clusters
    - Distance to Cluster 1 prototype significantly smaller
    - Noise increases distance to Cluster 3 prototype substantially
3. **K-NN Method:**
    - Shows perfect consistency in nearest neighbors
    - Identical results with and without noise
    - Suggests strong local structure around the test point

9.4.3 Impact of Noise

The presence of noise affects the classification methods differently:

- MLE shows reduced likelihood values but maintains relative relationships
- Hyperplane method shows increased distances but preserves classification
- K-NN remains completely stable, suggesting robust local structure

9.5 Conclusion

The unanimous classification of the test point (1.9, 6) to Cluster 1 across all methods and datasets suggests:

- High reliability of the classification
- Robustness to noise in the dataset
- Strong local and global structure in the data supporting this classification

The consistency across methods provides strong confidence in the classification result, while the different approaches offer complementary insights into the data structure around the test point.