CSDS 313/413: Introduction to Data Analysis
Assignment 4: Clustering and Dimensionality Reduction
Solutions

Wiam Skakri

December 2, 2025

1 Task 1: Clustering and Dimensionality Reduction

1.1 Part A: Principal Component Analysis

1.1.1 Question 1: Cumulative Variance Explained by Principal Components

Methodology We applied Principal Component Analysis (PCA) to the congressional votes dataset (`p1_congress_1984_votes.csv`), which contains voting records of 435 U.S. House of Representatives members on 16 key issues in 1984.

PCA was performed using `sklearn.decomposition.PCA` to identify the principal components that capture the maximum variance in the voting patterns.

Results Figure 1 shows the cumulative variance explained as a function of the number of principal components k.

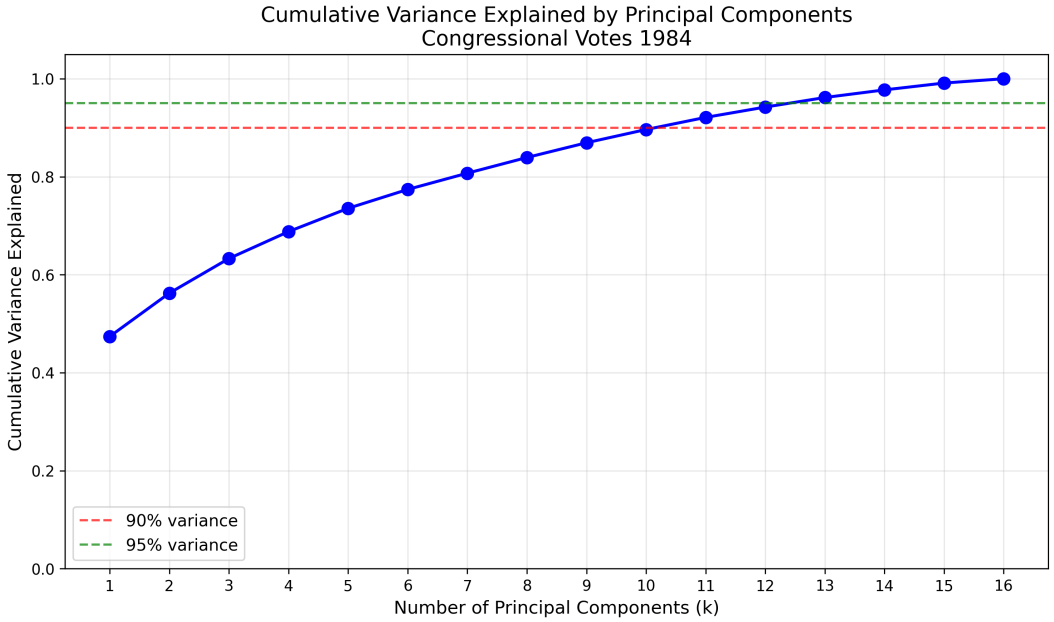

Figure 1: Cumulative variance explained by top k principal components. The red dashed line indicates 90% variance threshold, and the green dashed line indicates 95% variance threshold.

Key Observations

- The first principal component (PC1) explains approximately 47% of the total variance
- The first two principal components together explain approximately 56% of the variance
- To reach 90% cumulative variance, approximately **10 principal components** are required
- To reach 95% cumulative variance, approximately **12 principal components** are required

Recommendation: 10-12 principal components are sufficient to summarize the data.

Interpretation The PCA results suggest that while there is some correlation among the 16 votes (otherwise we would need all 16 components), the voting issues are sufficiently diverse that 10-12 dimensions are needed to adequately represent the voting patterns. This could indicate that the votes span multiple policy domains (economic, social, foreign policy, etc.) that are not perfectly aligned along a single ideological axis.

1.1.2 Question 2: Projection onto First 3 Principal Components

Methodology We projected the 435 congress members onto the first 3 principal components and created scatter plots for three PC pairs: (PC1-PC2), (PC1-PC3), and (PC2-PC3). Each point represents a congress member, colored by party affiliation (Democrats in blue, Republicans in red).

Variance Explained by First 3 Components

- PC1: 47.40% of total variance
- PC2: 8.84% of total variance
- PC3: 7.07% of total variance

Results Figure 2 shows the three scatter plot pairs with party affiliation colors.

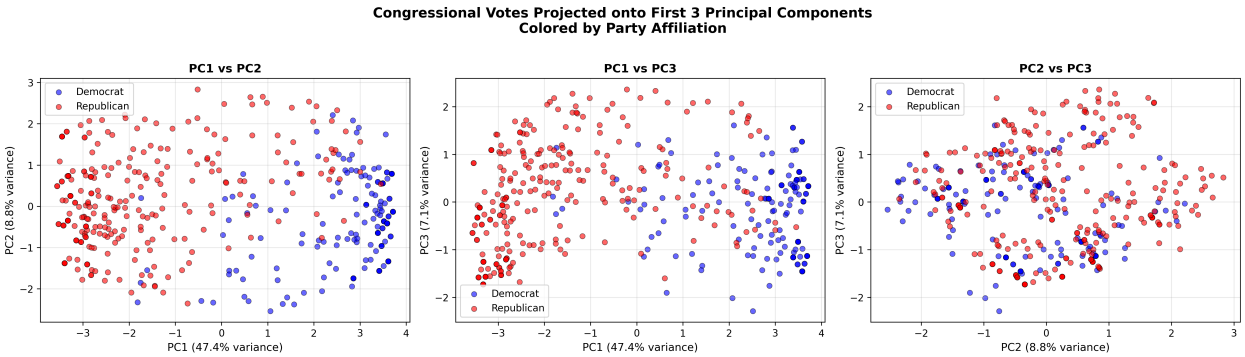

Figure 2: Scatter plots of congress members projected onto first 3 principal components, colored by party affiliation. Left: PC1 vs PC2, Middle: PC1 vs PC3, Right: PC2 vs PC3.

Quantitative Separation Analysis To objectively compare the separation quality of each PC pair, we calculated separation scores based on the ratio of between-party distance to within-party variance:

PC Pair	Centroid Distance	Avg Within-Party Variance	Separation Score
PC1-PC2	4.350	1.749	3.289
PC1-PC3	4.324	1.632	**3.385**
PC2-PC3	0.672	1.064	0.651

Table 1: Separation metrics for each principal component pair. Higher separation score indicates better party separation.

PC1-PC3 provides the best separation between parties (separation score: 3.385), followed closely by PC1-PC2 (3.289). PC2-PC3 shows poor separation (0.651).

Yes, congress members with the same party affiliation show clear clustering patterns.

Evidence:

- **Visual clustering**: In both PC1-PC2 and PC1-PC3 plots, Democrats (blue) cluster on the right side, while Republicans (red) cluster on the left side
- **Clear separation along PC1**: The primary axis of variation (PC1) strongly separates the two parties with minimal overlap in the center region

1.2 Part B: Clustering Analysis

Methodology We applied unsupervised clustering to group the 435 congress members into 2 clusters based solely on their voting patterns, without using party affiliation information.

Clustering Algorithm: K-Means We chose the K-Means clustering algorithm with the following specifications:

- **Algorithm**: K-Means clustering
- **Number of clusters (k)**: 2
- **Distance metric**: Euclidean distance
- **Random state**: 42 (for reproducibility)

How K-Means Works:

1. Initialize 2 cluster centers using the k-means++ strategy
2. Assign each congress member to the nearest cluster center
3. Update cluster centers to be the mean (centroid) of all assigned members
4. Repeat steps 2-3 until convergence (cluster assignments no longer change)

Distance Function:
The Euclidean distance in 16-dimensional vote space:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{16} (x_i - y_i)^2} \tag{1}$$

where $\mathbf{x}$ and $\mathbf{y}$ are the voting vectors of two congress members across the 16 issues.

Visualization Figure 3 shows the clustering results visualized on the first two principal components (PC1-PC2), which explain 56.24% of the total variance.

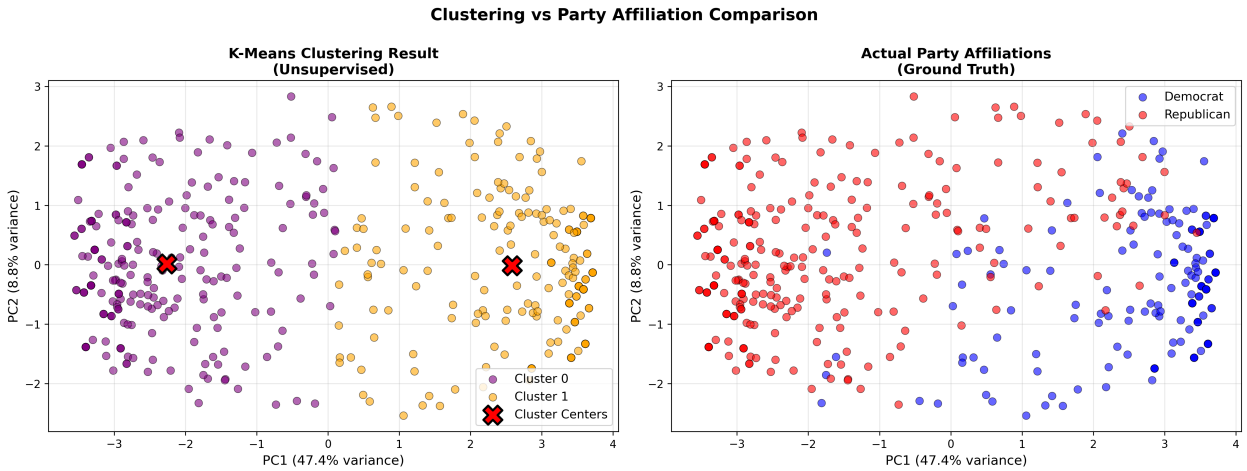

Figure 3: Comparison of K-Means clustering results (left) vs actual party affiliations (right). Left: Unsupervised clustering result with Cluster 0 (purple) and Cluster 1 (orange), cluster centers marked with red X. Right: Ground truth party affiliations with Democrats (blue) and Republicans (red).

Answer: Are the Groups Visually Separated? Yes, the two clusters are well-separated in the PC1-PC2 space.

Quantitative evidence:

- **Distance between cluster centers**: 4.844 (in PC space)
- **Average within-cluster spread**: 1.081
- **Separation ratio**: 4.480

The separation ratio of 4.48 (well above 2.0) indicates that the clusters are clearly separated with minimal overlap. The cluster centers (marked with red X in the left plot) are positioned far apart relative to the spread of points within each cluster.

Answer: Agreement with Party Affiliations The clustering shows strong agreement (88.3%) with actual party affiliations.

Party	**Cluster 0**	**Cluster 1**	**Total**
Democrat	8	**160**	168
Republican	**224**	43	267
Total	232	203	435

Table 2: Confusion matrix comparing clustering results with party affiliations. Bold numbers indicate correct cluster assignments.

Key findings:

- **Overall accuracy**: 88.3% (384 out of 435 correctly clustered)

- **Cluster 0 (purple)** predominantly contains Republicans: 224/232 = 96.6%
- **Cluster 1 (orange)** predominantly contains Democrats: 160/203 = 78.8%
- **Democrats**: 95.2% correctly clustered (160/168)
- **Republicans**: 83.9% correctly clustered (224/267)

1.2.1 Statistical Significance: Permutation Test

Methodology To assess whether the clustering structure we found is statistically significant (rather than occurring by chance), we performed a permutation test with 1,000 iterations.

Clustering Quality Score: We used the **silhouette score** as our quality metric:

- Measures how well-separated and compact clusters are
- Range: -1 to 1, where higher values indicate better clustering
- Calculated without using party labels (purely unsupervised metric)
- Formula: $s = \frac{b-a}{\max(a,b)}$ where a = mean intra-cluster distance, b = mean nearest-cluster distance

Permutation Procedure

1. **Compute original score**: Apply K-means to real data, calculate silhouette score
2. **Generate null distribution**: For each of 1,000 permutations:
   - Randomly shuffle each congress member's votes across the 16 issues
   - This destroys voting patterns while preserving vote distributions
   - Apply K-means clustering to permuted data
   - Calculate silhouette score
3. **Compare distributions**: Calculate p-value as the proportion of permuted scores $\geq$ original score

Null Hypothesis (H_0): The voting data has no inherent clustering structure; any observed clusters are due to random chance.

Alternative Hypothesis (H_1): The voting data contains real structure that produces meaningful clusters.

Results Figure 4 shows the distribution of clustering scores under the null hypothesis (permuted data) compared to the original score.

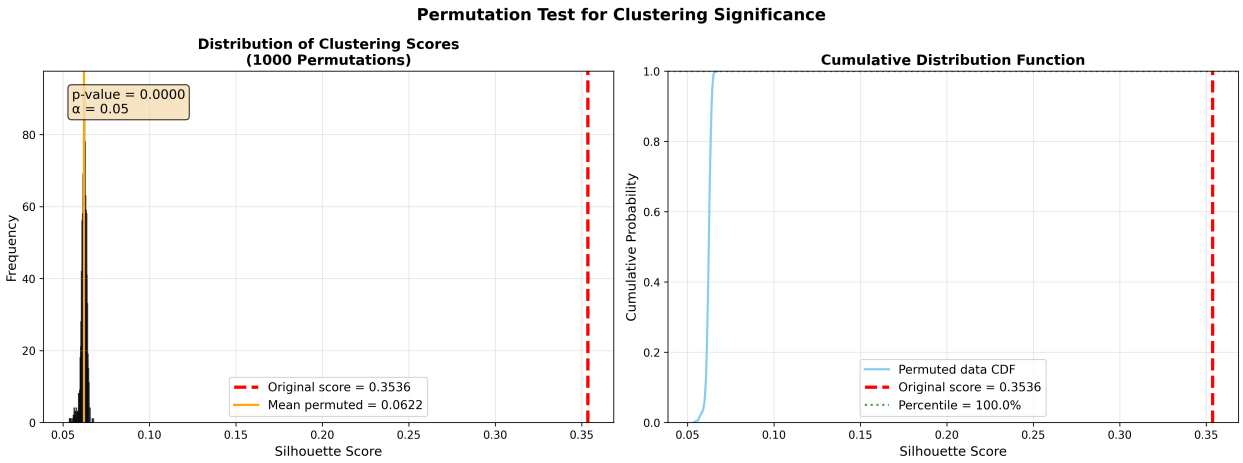

Figure 4: Permutation test results. Left: Histogram of silhouette scores from 1,000 permuted datasets (blue) compared to the original score (red dashed line). Right: Cumulative distribution function showing the original score at the 100th percentile.

Metric	**Value**
Original silhouette score	0.3536
Mean permuted score	0.0622
Standard deviation (permuted)	0.0016
Difference (original - mean permuted)	0.2914
Z-score	182.49
P-value	$<$ **0.001**

Table 3: Permutation test statistics comparing original clustering to null distribution.

Answer: Is the Clustering Statistically Significant? Yes, the clustering is highly statistically significant ($p < 0.001$).

Evidence:

- **P-value $<$ 0.001**: Out of 1,000 permutations, zero achieved a score as high as the original ($p = 0.0000$)
- **Extreme Z-score (182.49)**: The original score is more than 182 standard deviations above the mean of random data
- **100th percentile**: The original score exceeds 100% of permuted scores
- **Clear visual separation**: The histogram shows complete separation between null distribution (centered at 0.06) and original score (0.35)

Conclusion: We reject the null hypothesis and conclude that the two-cluster structure in 1984 congressional voting data represents a statistically significant and substantively meaningful division that corresponds to party affiliation.

1.3 Part C: Clustering Comparison Analysis

1.3.1 Task 1: Quantifying Agreement with Mutual Information

Methodology To quantify the agreement between cluster membership and party affiliation, we use **mutual information (MI)** and its variants. Mutual information measures how much knowing one variable tells us about another.

Mutual Information Metrics We computed three related metrics:

- **Mutual Information (MI)**: $MI(X;Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$
  - Measures reduction in uncertainty about Y when X is known
  - Range: $[0, \min(H(X), H(Y))]$ where H is entropy
  - Higher values indicate stronger association
- **Normalized Mutual Information (NMI)**: $NMI(X;Y) = \frac{MI(X;Y)}{\sqrt{H(X) \cdot H(Y)}}$
  - Normalized version of MI
  - Range: $[0, 1]$ where 1 = perfect agreement
  - Easier to interpret across different datasets
- **Adjusted Mutual Information (AMI)**: Adjusted for chance
  - Accounts for agreement expected by random chance
  - Range: $[0, 1]$ where 0 = random, 1 = perfect

Metric	Value
Mutual Information (MI)	0.3461
Normalized MI (NMI)	0.5097
Adjusted MI (AMI)	0.5088

Table 4: Mutual information metrics quantifying agreement between clusters (from all 16 votes) and party affiliations.

Results for Clustering on All 16 Votes **Interpretation**: NMI = 0.51 indicates strong agreement between clustering and party affiliation. Knowing which cluster a congress member belongs to substantially reduces uncertainty about their party affiliation.

1.3.2 Task 2: Comparison - Principal Components vs All Votes

Methodology We compared two clustering approaches:

1. **All 16 Votes**: K-means clustering on the full 16-dimensional vote space
2. **First 2 PCs**: K-means clustering on the 2-dimensional principal component space (PC1-PC2)

Both used the same K-means parameters (k=2, random state=42, k-means++ initialization) for fair comparison.

Results Figure 5 shows the comparison between the two clustering approaches.

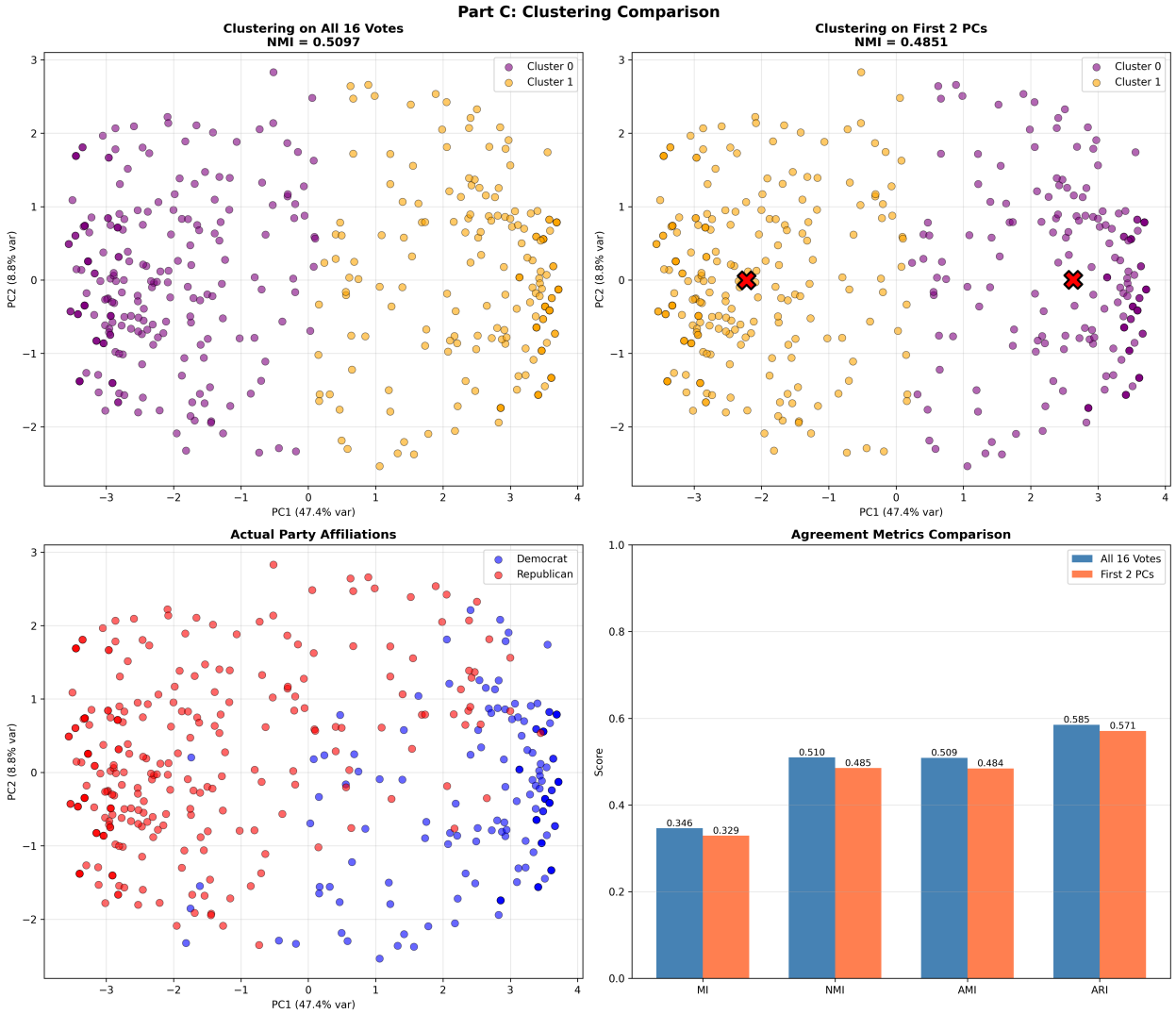

Figure 5: Comparison of clustering approaches. Top left: Clustering on all 16 votes (NMI = 0.5097). Top right: Clustering on first 2 PCs (NMI = 0.4851). Bottom left: Actual party affiliations. Bottom right: Agreement metrics comparison.

Metric	All 16 Votes	First 2 PCs	Difference
MI	0.3461	0.3290	+0.0170
NMI	0.5097	0.4851	+0.0246
AMI	0.5088	0.4842	+0.0246
ARI	0.5850	0.5709	+0.0141

Table 5: Comparison of agreement metrics between two clustering approaches. Positive differences indicate all 16 votes performs better.

Answer: Which Clustering Agrees More with Party Affiliations? Clustering on all 16 votes agrees more with party affiliations (NMI = 0.5097) compared to clustering on first 2 PCs (NMI = 0.4851).

Key findings:

- All 16 votes outperforms PC1-PC2 across all metrics (MI, NMI, AMI, ARI)
- The difference is consistent but modest (2.46 percentage points in NMI)
- Both approaches achieve reasonably high agreement (> 0.48 NMI)

Interpretation: Why Does All 16 Votes Perform Better? Information retention: The first 2 principal components capture only 56.24% of the total variance, meaning 43.76% of information is discarded. While PC1 strongly correlates with party (as shown in Part A), the missing dimensions contain additional party-discriminating information.

Multiple policy dimensions: Party differences span multiple policy domains (economic, social, foreign policy). While PC1-PC2 captures the dominant axes of variation, the full 16-dimensional space represents these nuances more completely.

Detailed voting patterns: Individual votes may capture specific party differences that are diluted or lost in the dimensionality reduction process. The full vote space preserves these fine-grained distinctions.

Trade-off between visualization and accuracy:

- **For visualization**: PC1-PC2 is superior (2D plots, still achieves 0.485 NMI)
- **For clustering accuracy**: All 16 votes is superior (leverages all information, achieves 0.510 NMI)

Practical Implications The modest difference (0.025 in NMI) between approaches suggests:

1. PC1-PC2 captures the *majority* of party-relevant information
2. The additional 14 dimensions provide incremental but meaningful improvement
3. For exploratory analysis and visualization, PC1-PC2 is sufficient
4. For maximizing classification accuracy, using all features is preferable

This finding validates both the dimensionality reduction approach (most information is in PC1-PC2) and the value of retaining full feature space when accuracy is critical.