High Dimensional Statistics IV MATH4287 - Mini Project Report

Does dimensionality reduction improve clustering performance in high-dimensional gene expression data?

qvns53@durham.ac.uk, Durham University

2025-12-19

Test

1 Introduction

Modern biological datasets often exhibit a "large p, small n" structure, where the number of measured variables far exceeds the number of observations, this is known as high-dimensional data. A canonical example is microarray gene expression data, where thousands of genes are measured across a limited number of samples. In such settings, classical clustering methods based on Euclidean distances are known to perform poorly due to the curse of dimensionality.

A common strategy for addressing these issues is to apply dimensionality reduction prior to clustering. Principal Component Analysis (PCA) is frequently used to project the data onto a lower-dimensional subspace that preserves maximal variance, after which standard clustering methods may be applied.

However, it has been noted that directions of maximal variance do not necessarily correspond to directions that separate clusters. As a result, dimensionality reduction may fail to improve, or may even degrade, clustering performance depending on the data structure.

This mini-project investigates whether dimensionality reduction improves clustering performance on a real high-dimensional dataset. In particular, we study k-means and hierarchical clustering applied to the Human Tumor Microarray dataset (available in the practical material from this course), comparing clustering results obtained on the raw data, PCA-reduced data, and sparse PCA-reduced data.

[1] [2] [3] [4] [5] [6]

2 Objectives & Questions

The objectives of this project are:

- To evaluate whether dimensionality reduction improves clustering performance in a high-dimensional, low-sample-size setting.
- To compare standard PCA and sparse PCA as preprocessing steps for clustering.
- To assess the impact of dimensionality reduction on both clustering accuracy and interpretability.

The primary research question is:

- Does dimensionality reduction improve clustering performance, as measured by the Adjusted Rand Index, on the Human Tumor Microarray dataset?

Secondary questions include:

- Does sparse PCA offer advantages over standard PCA for clustering?
- How does dimensionality reduction affect different hierarchical linkage methods?

3 Literature Review

In high dimensions, Euclidean distances tend to concentrate, reducing their discriminative power. As discussed in the lecture notes, this phenomenon undermines distance-based clustering methods such as k-means and hierarchical clustering.

Although PCA is often applied prior to clustering, theoretical and empirical results suggest that high-variance directions may not align with cluster-separating directions.

Sparse PCA extends classical PCA by imposing sparsity on the loading vectors, improving interpretability by identifying a small subset of influential variables. While sparse PCA can enhance interpretability, its effect on clustering performance is less clear.

4 Methods

4.1 Dataset & Exploratory Analysis

The Human Tumor Microarray dataset consists of n = 64 samples with p = 6830 gene expression measurements. True tumor labels are available and are used only for external evaluation of clustering results.

4.2 Clustering without Dimensionality Reduction

The number of clusters was fixed at k = 5 across all experiments to allow fair comparison between methods.

4.3 Dimensionality Reduction

Rather than fixing the number of principal components directly, variance-explained thresholds were used to determine the reduced dimension.

Sparse PCA was implemented using the spca function, which imposes an L1 penalty on the loading vectors, encouraging sparse gene selection.

4.4 Evaluation Metrics

Clustering performance was evaluated using the Adjusted Rand Index (ARI), which measures agreement between the clustering assignment and the true labels while correcting for chance.

5 Results & Findings

5.1 Baseline clustering

Applying k-means directly to the raw high-dimensional data yielded an ARI of approximately 0.66, which was the highest performance observed across all methods.

5.2 PCA-based clustering

Applying PCA with 90% or 95% variance retention did not meaningfully change clustering performance relative to the raw data.

At the 75% variance threshold, the preferred clustering structure changed, but no improvement in ARI was observed.

5.3 Sparse-PCA based clustering

Sparse PCA did not improve k-means clustering performance as measured by ARI.

However, sparse PCA produced loading vectors involving substantially fewer genes, facilitating interpretation of the dominant components.

5.4 Hierarchical clustering

The impact of dimensionality reduction on hierarchical clustering depended on the choice of linkage method, with sparse PCA affecting average and complete linkage differently.

6 Conclusions

- Dimensionality reduction did not improve clustering performance on the Human Tumor Microarray dataset.
- Raw k-means clustering achieved the highest Adjusted Rand Index.
- PCA preserved clustering structure but did not enhance it, even at aggressive variance thresholds.
- Sparse PCA improved interpretability by identifying a small subset of genes but did not improve clustering accuracy.
- The effectiveness of dimensionality reduction for clustering is highly data-dependent and cannot be assumed a priori.

7 References

[1] K. Y. Yeung and W. L. Ruzzo, "An empirical study on Principal Component Analysis for clustering gene expression data," 2001.

[2] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, Dec. 1985, doi: 10.1007/BF01908075.

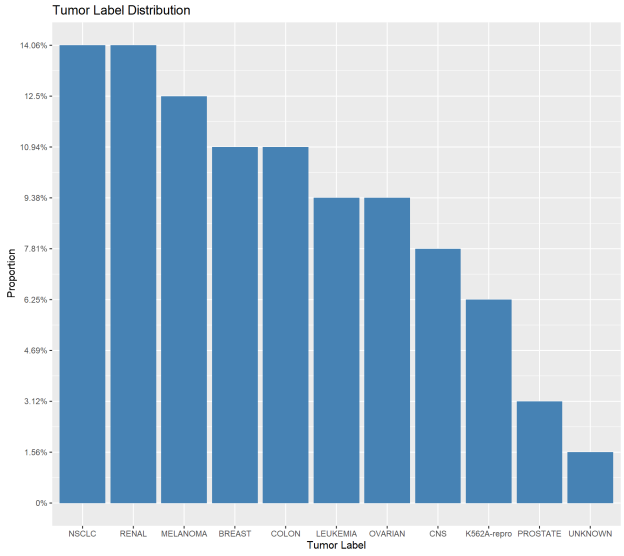
[3] C. Mukherjee and J. Zhang, "Compressibility: Power of PCA in Clustering Problems Beyond Dimensionality Reduction," *ResearchGate*. 2022. Accessed: Dec. 16, 2025. [Online]. Available: https://www.researchgate.net/publication/360186068_Compressibility_Power_of_PCA_in_Clustering_Problems_Beyond_Dimensionality_Reduction

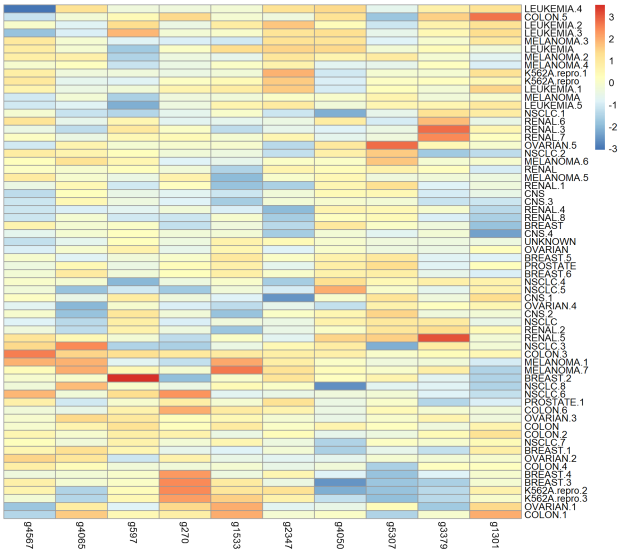
[4] C. Ding and X. He, "K -means clustering via principal component analysis," in *Twenty-first international conference on Machine learning - ICML '04*, Banff, Alberta, Canada: ACM Press, 2004, p. 29. doi: 10.1145/1015330.1015408.

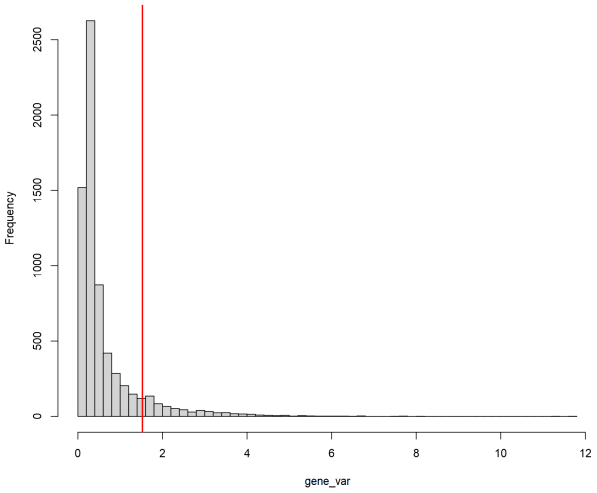
[5] W. M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, Dec. 1971, doi: 10.1080/01621459.1971.10482356.

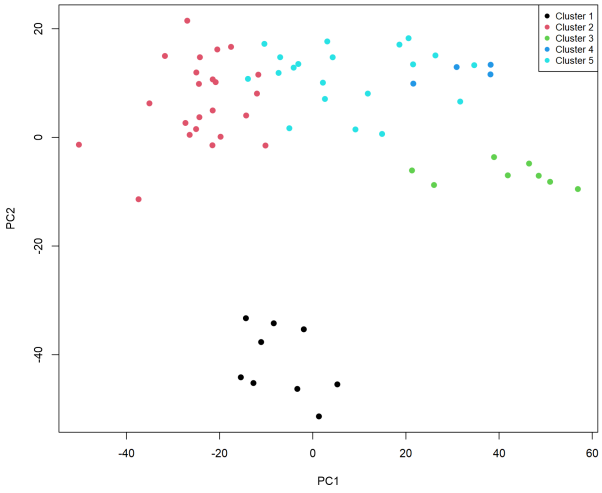
[6] W.-C. Chang, "On Using Principal Components before Separating a Mixture of Two Multivariate Normal Distributions," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 32, no. 3, pp. 267–275, 1983, doi: 10.2307/2347949.

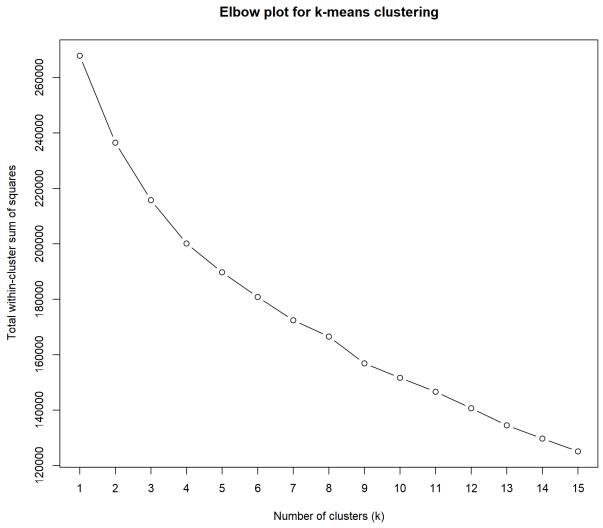
8 Appendix

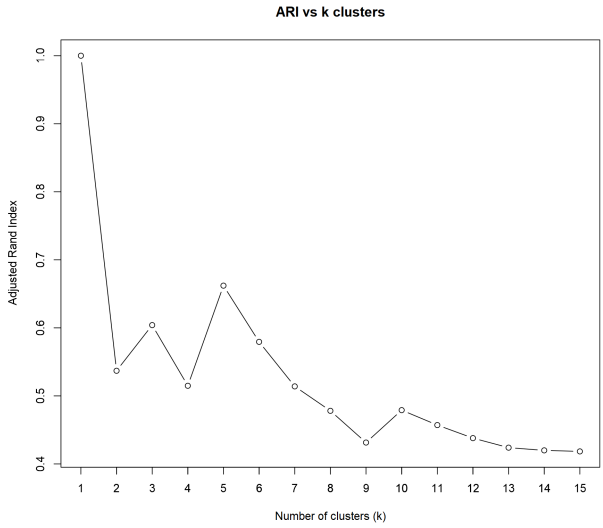

Tumor Label Distribution
Proportion
14.06%
12.5%
10.94%
9.38%
7.81%
6.25%
4.69%
3.12%
1.56%
0%
NSCLC
RENAL
MELANOMA
BREAST
COLON
LEUKEMIA
OVARIAN
CNS
K562A-repro
PROSTATE
UNKNOWN
Tumor Label

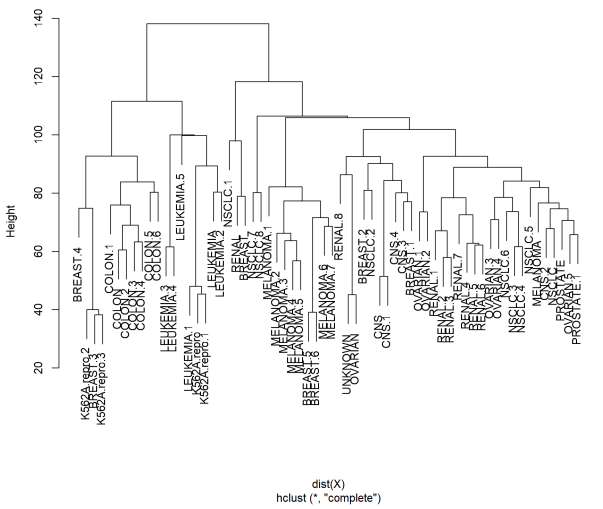

LEUKEMIA.4
COLON.5
LEUKEMIA.2
LEUKEMIA.3
MELANOMA.3
LEUKEMIA
MELANOMA.2
MELANOMA.4
K562A.repro.1
K562A.repro
LEUKEMIA.1
MELANOMA
LEUKEMIA.5
NSCLC.1
RENAL.6
RENAL.3
RENAL.7
OVARIAN.5
NSCLC.2
MELANOMA.6
RENAL
MELANOMA.5
RENAL.1
CNS
CNS.3
RENAL.4
RENAL.8
BREAST
CNS.4
UNKNOWN
OVARIAN
BREAST.5
PROSTATE
BREAST.6
NSCLC.4
NSCLC.5
CNS.1
OVARIAN.4
CNS.2
NSCLC
RENAL.2
RENAL.5
NSCLC.3
COLON.3
MELANOMA.1
MELANOMA.7
BREAST.2
NSCLC.6
NSCLC.8
PROSTATE.1
COLON.6
OVARIAN.3
COLON
COLON.2
NSCLC.7
BREAST.1
OVARIAN.2
COLON.4
BREAST.4
BREAST.3
K562A.repro.2
K562A.repro.3
OVARIAN.1
COLON.1
3
2
1
0
-1
-2
-3
g4567
g4085
g597
g270
g1533
g2347
g4050
g5307
g3379
g1301

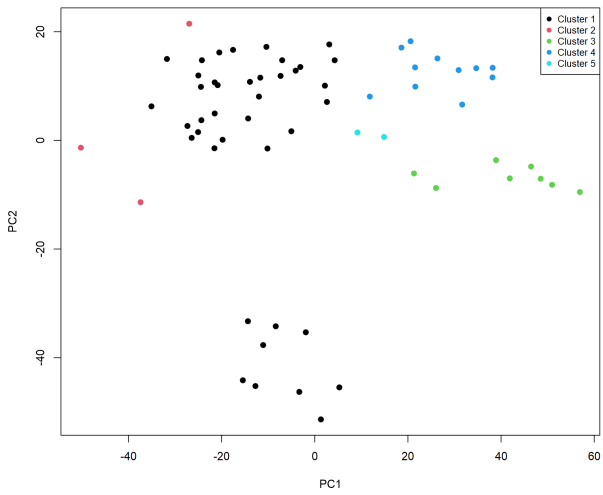

Gene-wise variance distribution
Frequency
2500
2000
1500
1000
500
0
0
2
4
6
8
10
12
gene_var

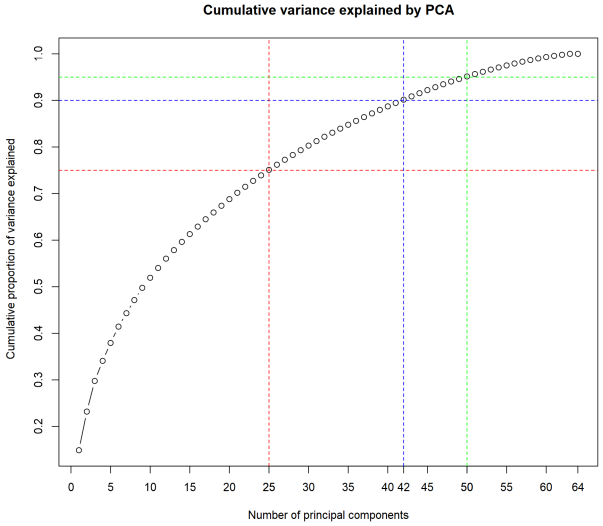

K-means clusters (first 2 PCs, raw clustering)
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
PC2
20
0
-20
-40
-40
-20
0
20
40
60
PC1

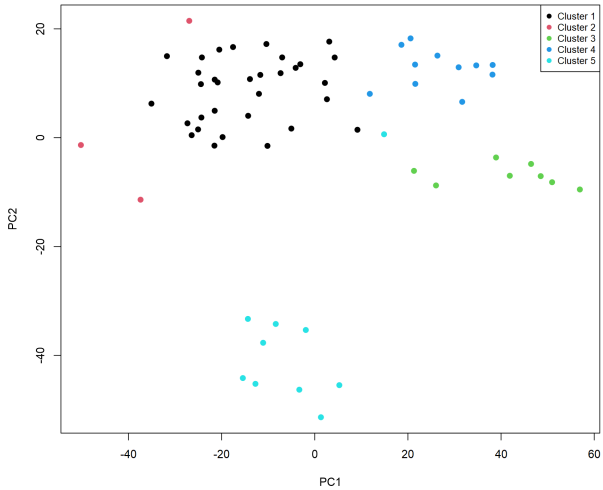

Elbow plot for k-means clustering
Total within-cluster sum of squares
Number of clusters (k)

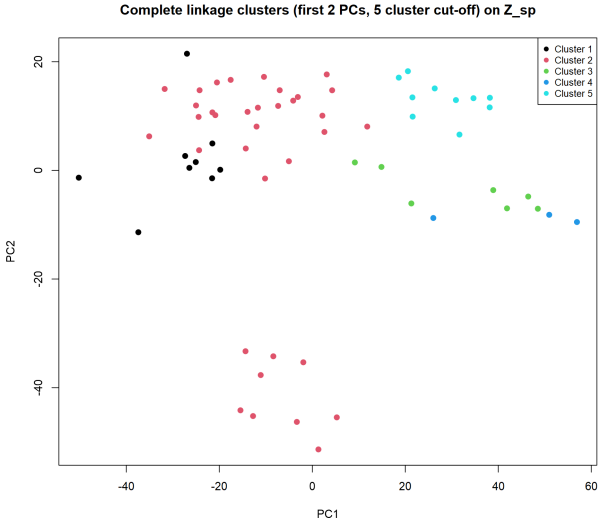

ARI vs k clusters
Adjusted Rand Index
Number of clusters (k)

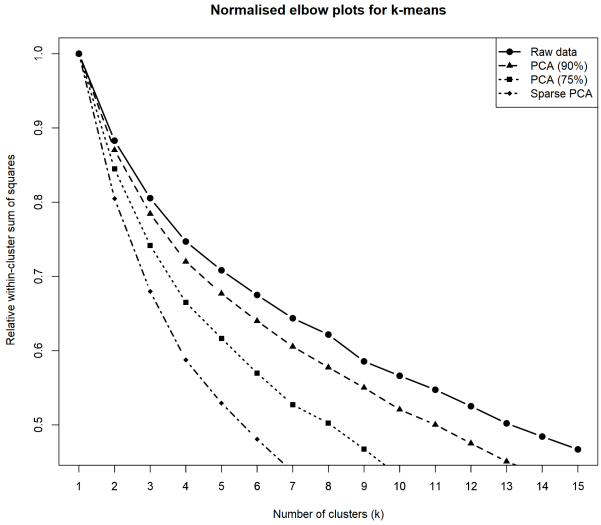

Complete
Height
dist(X)
hclust (*, "complete")

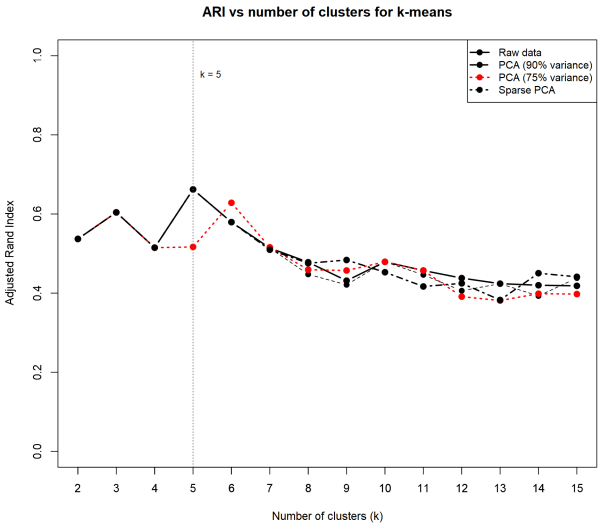

Complete linkage clusters (first 2 PCs, 5 clusters cut-off)
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
PC2
PC1

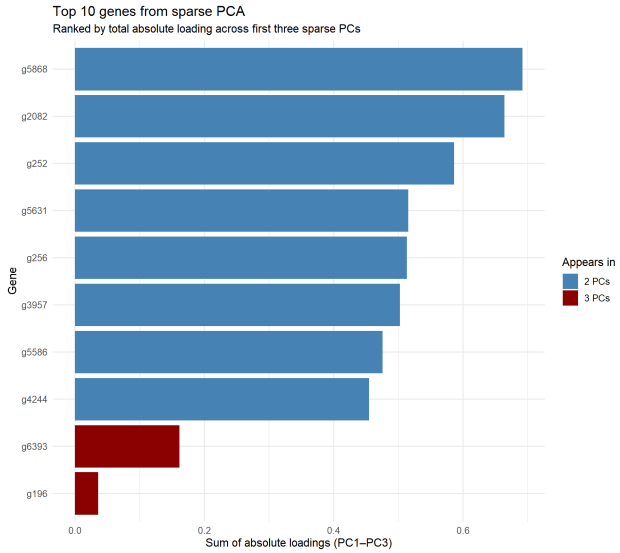

Cumulative variance explained by PCA
Cumulative proportion of variance explained
Number of principal components


Complete linkage clusters (first 2 PCs, 5 cluster cut-off) on Z
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
PC2
PC1


Complete linkage clusters (first 2 PCs, 5 cluster cut-off) on Z_sp
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
PC1
PC2


Normalised elbow plots for k-means
Raw data
PCA (90%)
PCA (75%)
Sparse PCA
Relative within-cluster sum of squares
Number of clusters (k)


ARI vs number of clusters for k-means
Raw data
PCA (90% variance)
PCA (75% variance)
Sparse PCA
k = 5
Adjusted Rand Index
Number of clusters (k)


Top 10 genes from sparse PCA
Ranked by total absolute loading across first three sparse PCs
g5868
g2082
g252
g5631
g256
g3957
g5586
g4244
g6393
g196
Gene
Appears in
2 PCs
3 PCs
Sum of absolute loadings (PC1–PC3)