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

2025-12-17

1 Introduction

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

2 Objectives & Questions

3 Literature Review

4 Methods

5 Results & Findings

6 Conclusion

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.

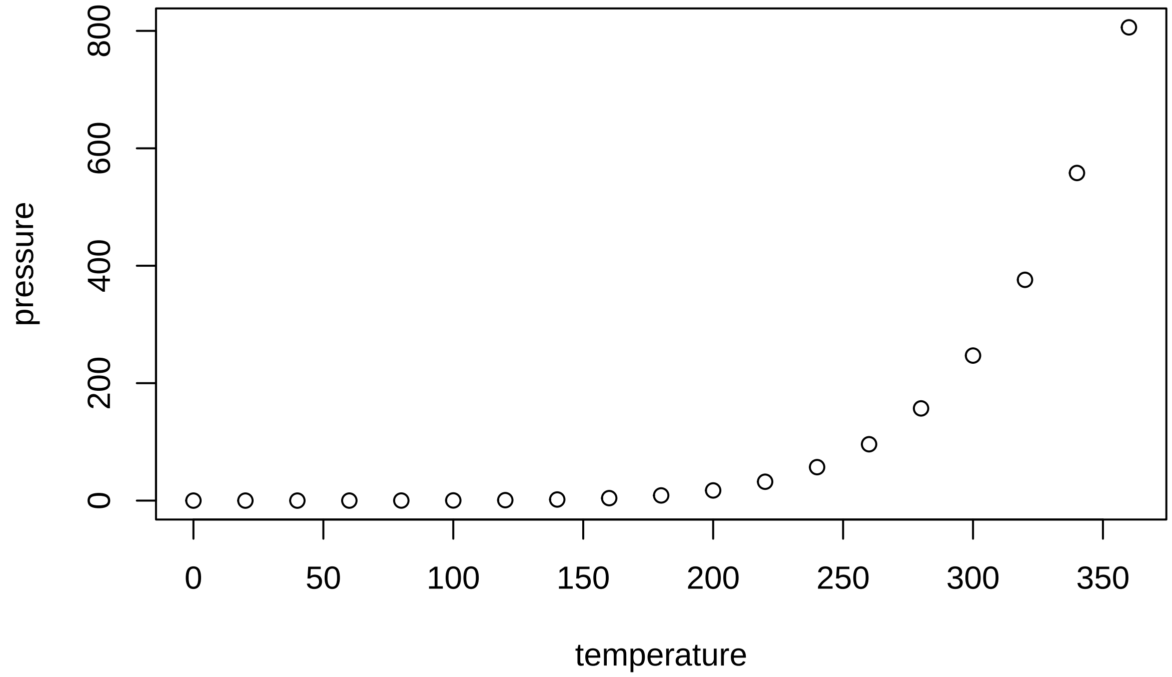
[5] W. M. Rand, “Objective Criteria for the Evaluation of Clustering Methods,” *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, Dec. 1971, doi: 10.1080/01621459.1971.10482356.

[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

```r
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```


pressure
temperature