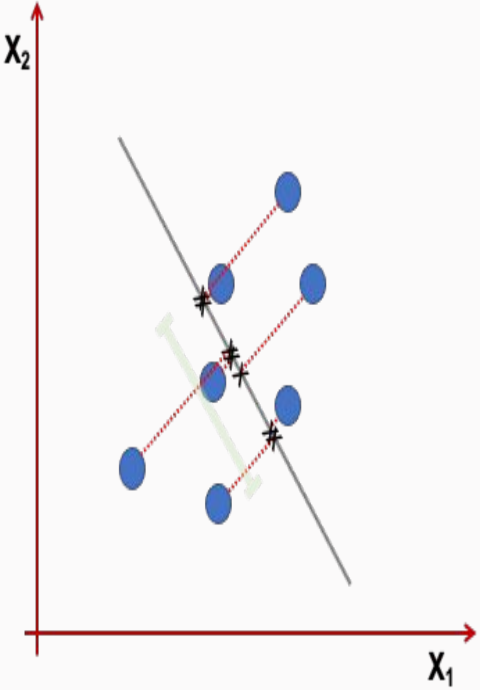
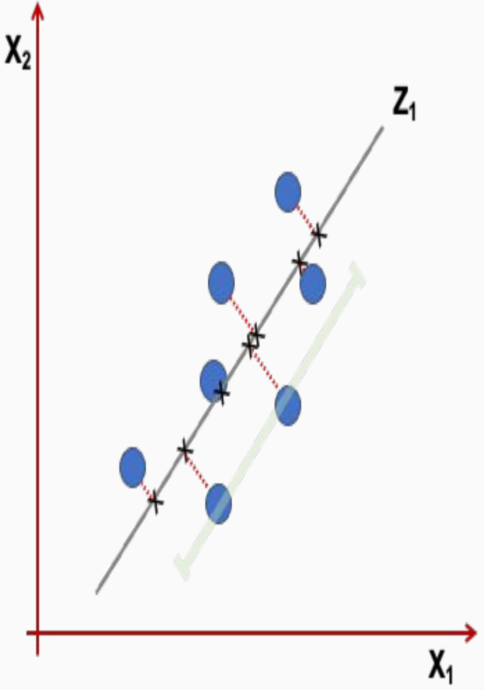
Principal Component Analysis (PCA)

Principal Component Analysis is a dimensionality reduction technique

Goal is to find directions, known as principal components, that maximize variance





Algorithm & Runtime

Original n x m data matrix A, where each column is a variable and each row is an observation (data point)

1. *Center data: Each entry is subtracted by the variable's mean. Result in centered matrix X.* $O(nm)$

2. *Compute covariance matrix* $C = \frac{1}{n} X^\top X. O(nm^2)$

3. *Perform eigendecomposition on matrix* $C = Q \Lambda Q^\top . O(m^3)$

4. *k principal components = k eigenvectors that correspond to the largest k eigenvalues*

Runtime = $O(nm + nm^2 + m^3) \leq O(2nm^2 + m^3) = O(nm^2 + m^3)$

Correctness

Why the eigenvector with the largest eigenvalue is the direction that captures the most variance?

After applying Lagrange multiplier:

$$v^\top C v + \alpha(1 - v^\top v)$$

Taking derivative and set to 0:

$$Cv = \alpha v$$

Alternative method: Singular Value Decomposition

A better method: without the need of building a covariance matrix.

$$X = U\Sigma V^\top$$

Linearly independent vectors only, $r = \text{rank}(X)$. Compacted SVD:

$$X = U_r \Sigma_r V_r^\top$$

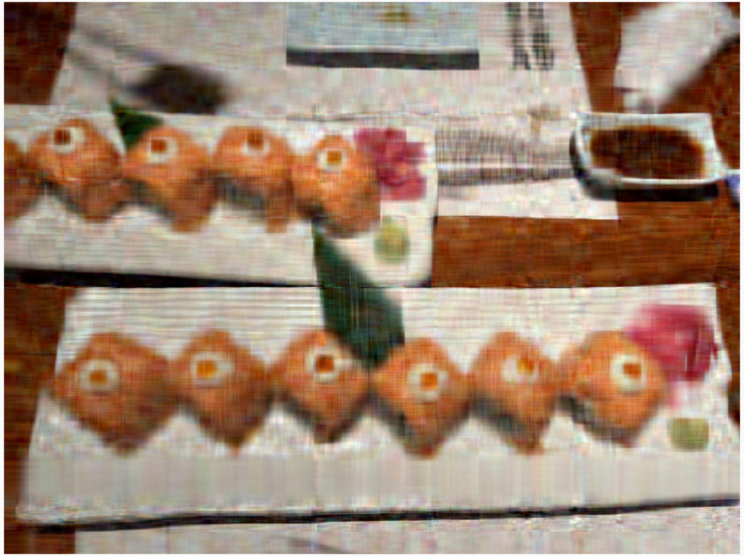
$$C = V_r(\frac{1}{n}\Sigma_r^2)V_r^\top \qquad C = Q\Lambda Q^\top$$

$$V_r = Q$$

$$(\frac{1}{n}\Sigma_r^2) = \Lambda$$

Demonstration

Original
#variables = 4032



$k = 25$

Demonstration

$k = 200$

$k = 300$

Extension

1. Can see diminishing returns in the variance captured as k increases. Find the optimal number of principal components

2. Explore variants of PCA. When the relationship is non-linear (e.g circular), then finding a line doesn't make sense: Kernel PCA

References

1. Shlens, “A Tutorial on Principal Component Analysis,” 1404.1100, Apr. 2014. [Online]. Available: https://arxiv.org/pdf/1404.1100
2. T. Roughgarden and G. Valiant, “CS168: The Modern Algorithmic Toolbox Lecture #7: Understanding and Using Principal Component Analysis (PCA),” 2024. [Online]. Available: https://web.stanford.edu/class/cs168/l/l7.pdf?
3. C. Bishop, Pattern Recognition and Machine Learning. 2006, Chapter 12.
4. E. Fetaya, J. Lucas, E. Andrews, and University of Toronto, “CSC 411 Lecture 12: Principle Components Analysis.” [Online]. Available: https: //www.cs.toronto.edu/~jlucas/teaching/csc411/lectures/lec12_handout.pdf?
5. Y. Qiu, “Large-Scale Eigenvalue Decomposition and SVD with RSpectra,” Jul. 18, 2024. https://cran.r-project.org/web/packages/RSpectra/vignettes/introduction.html
6. NeuralNine, “Image compression using PCA in Python,” YouTube. Jun.08, 2022. [Online]. Available: https://www.youtube.com/watch?v=3aUshxvxGhY
7. Dr. Trefor Bazett, “Lagrange Multipliers — Geometric Meaning & Full example,” YouTube. Nov. 27, 2019. [Online]. Available: https://www.youtube.com/watch?v=8mjcnxGMwFo
8. Caltech, “8.6 David Thompson (Part 6): Nonlinear Dimensionality Reduction: KPCA,” YouTube. May 25, 2018. [Online]. Available: https://www.youtube.com/watch?v=HbDHohXPLnU
9. Libretexts, “20.4: Sparse principal component analysis,” Biology LibreTexts, Mar. 17, 2021. Available: https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.)/20%3A_Networks_I-_Inference_Structure_Spectral_Methods/20.04%3A_Sparse_Principal_Component_Analysis