Singular Value Decomposition in Principal Component Analysis

Aidan Olson

December 2024

Abstract

Principal Component Analysis (PCA) is a critical tool for dimensionality reduction and data analysis, enabling the transformation of high-dimensional datasets into lower-dimensional representations while preserving their essential structure. This paper outlines the mathematical foundations of PCA, emphasizing its reliance on the Singular Value Decomposition (SVD) for efficient computation. Through practical examples, including genomic data for ovarian cancer prediction, student alcohol consumption patterns, and health metrics related to diabetes, we illustrate PCA's versatility and effectiveness in uncovering patterns and reducing complexity. Additionally, we explore the online PCA algorithm, which extends PCA's applicability to dynamic, growing datasets. These examples demonstrate PCA's ability to simplify complex datasets.

Introduction

Principal Component Analysis (PCA) is a foundational technique in statistical data analysis, widely employed for dimensionality reduction and data simplification. As described by Abdi and Williams [2010], PCA extracts the most significant information from a dataset, condenses its size while preserving essential features, and simplifies its structure. By computing new variables, called principal components (PCs), as linear combinations of the original variables, PCA identifies dimensions of maximum variance. These principal components are ordered by their ability to explain variance, with the first component representing the "direction" of highest variance.

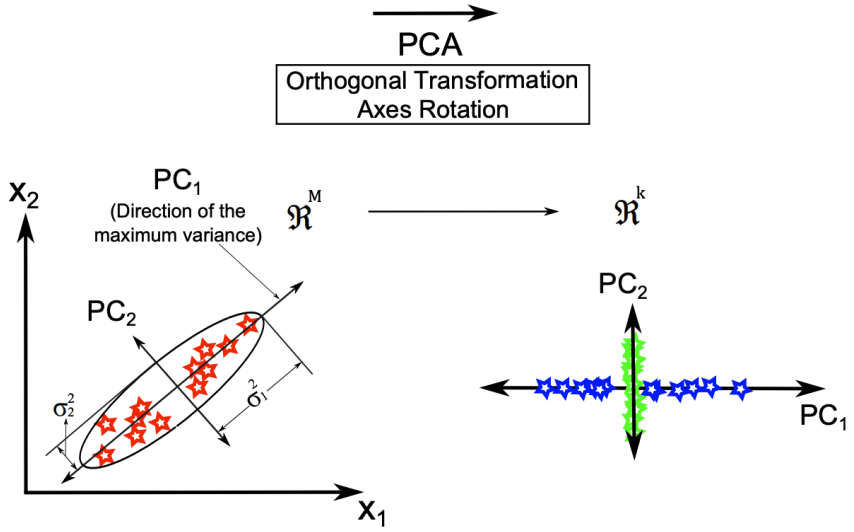


Figure 1: 2-Dimensional Visualization of PCA (*Tharwat [2016] Figure 1*). *The transformation of a 2D dataset into its mean-centered, rotated principal components.*

The primary motivation for PCA lies in its ability to transform high-dimensional data into a more manageable and computationally efficient format without significant loss of information. In today's data-driven

world, where massive datasets are increasingly common, PCA serves as a powerful tool to analyze, interpret, and visualize complex data structures. Furthermore, the orthogonality of the principal components ensures efficient storage and processing, as each component is independent of the others. This efficiency is particularly valuable in applications requiring high computational performance or data compression.

In this paper, we will outline the statistical derivation of PCA and demonstrate how the Singular Value Decomposition (SVD) provides an expedited and robust method for its computation. Using examples from Brunton and Kutz [2022] and publicly available datasets, we will illustrate the practical implementation and advantages of PCA. Finally, we will explore iterative approaches to PCA, such as the online PCA algorithm, which enables dynamic updates to the principal components as new data becomes available.

Mathematical Framework of PCA

In this section, we discuss the mathematical foundation of PCA, focusing on the statistical derivation of principal components from the covariance matrix of mean-centered data. Additionally, we emphasize how the SVD provides an expedited approach to conducting a principal component analysis.

Mean-Centered Data Matrix

Let $A \in \mathbb{R}^{n \times m}$ represent a collection of n observations and m features:

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,m} \end{bmatrix}.$$

To center the data, we compute the mean vector $\mu \in \mathbb{R}^n$:

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}, \quad \text{where} \quad \mu_i = \frac{1}{m} \sum_{j=1}^{m} a_{i,j}.$$

The mean-centered data matrix B is then given by:

$$B = A - \mu[1, \ldots, 1].$$

If μ is the zero vector, the dataset A is already mean-centered, and this step can be skipped.

Covariance Matrix

The covariance matrix C of the mean-centered data B is defined as:

$$C = \frac{1}{n} B^T B.$$

The covariance matrix encapsulates relationships between variables, revealing whether they tend to vary together or inversely. Each entry in C is a symmetric matrix that represents the pairwise covariance between two variables. High-magnitude entries indicate strong correlations, while low-magnitude entries suggest weak correlations. Positive covariance reflects direct correlation, while negative covariance implies inverse correlation. The main diagonal entries of C correspond to the variance of each variable. Importantly, the covariance matrix serves as the foundation for identifying the principal components through its eigendecomposition.

Eigen System of Covariance Matrix

The eigendecomposition of C reveals the principal components and their associated variances:

$$C = V \Lambda V^T,$$

where V is an orthogonal matrix whose columns are the eigenvectors of C, and Λ is a diagonal matrix whose entries are the eigenvalues of C. The eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are ordered such that:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n.$$

Each eigenvalue represents the variance captured by its corresponding eigenvector. If two or more eigenvalues are equal, their corresponding eigenvectors span a subspace of equal variance.

The principal components P can be computed as:

$$P = BV.$$

The first principal component P_1 captures the largest variance, while subsequent components capture progressively smaller variances. A rapid decay in eigenvalues indicates that most of the dataset's variability is concentrated in the first few components, enabling an accurate low-dimensional representation.

Principal Components and the SVD

The Singular Value Decomposition (SVD) of B offers an efficient alternative for computing principal components:

$$B = U\Sigma V^T,$$

where U contains the left singular vectors, Σ is a diagonal matrix of singular values equivalent to the matrix Λ, and V contains the right singular vectors. The principal components can be extracted directly as:

$$P_B = U\Sigma \quad \text{or} \quad P_B = BV.$$

SVD is particularly valuable for its numerical stability and ability to handle matrices of any shape. Its decomposition provides a geometric interpretation of the matrix as a sequence of rotations, scalings, and projections. By leveraging the SVD, PCA becomes a fast and reliable method for dimensionality reduction.

To recover the principal components of the original dataset A, we add back the mean vector μ:

$$P_A = U\Sigma + \mu.$$

In this section, we outlined the mathematical foundation of principal component analysis, starting from the mean-centering of data, through the computation of the covariance matrix, to the extraction of principal components using eigen decomposition and singular value decomposition. These steps reveal how PCA transforms a dataset into a lower-dimensional space while retaining its most significant features. The mathematical simplicity and computational efficiency of PCA make it a valuable tool for analyzing large, high-dimensional datasets. In the next section, we will apply PCA to a real-world dataset, breaking it down into principal components and visualizing the results to demonstrate its practical utility. This example will further illustrate how PCA enables us to uncover patterns and insights in complex data structures.

Example 1: Ovarian Cancer

As outlined in Chapter 1.5 of Brunton and Kutz [2022], MATLAB's integrated dataset on ovarian cancer patients provides an excellent example of PCA's utility. In this section, we perform a PCA on this dataset and demonstrate how it can be leveraged to predict a patient's susceptibility to ovarian cancer.

We start with a mean-centered data matrix $B \in \mathbb{R}^{216\times 4000}$, representing 4000 numerical gene expressions for 216 patients. Since B is already mean-centered, its economic singular value decomposition can be computed directly:

$$B = U\Sigma V^T.$$

The matrix $V \in \mathbb{R}^{4000\times 216}$ is the loading matrix whose columns are orthogonal "eigen-genomes." By taking the inner product of each patient's genome with the first three eigen-genomes, we can visualize the first three principal components in three-dimensional space. This is equivalent to computing the principal component matrix P as:

$$P = BV.$$

We then retain only the first three columns of P. The three-dimensional PCA space can be visualized by plotting each patient's (PC_1, PC_2, PC_3) values.

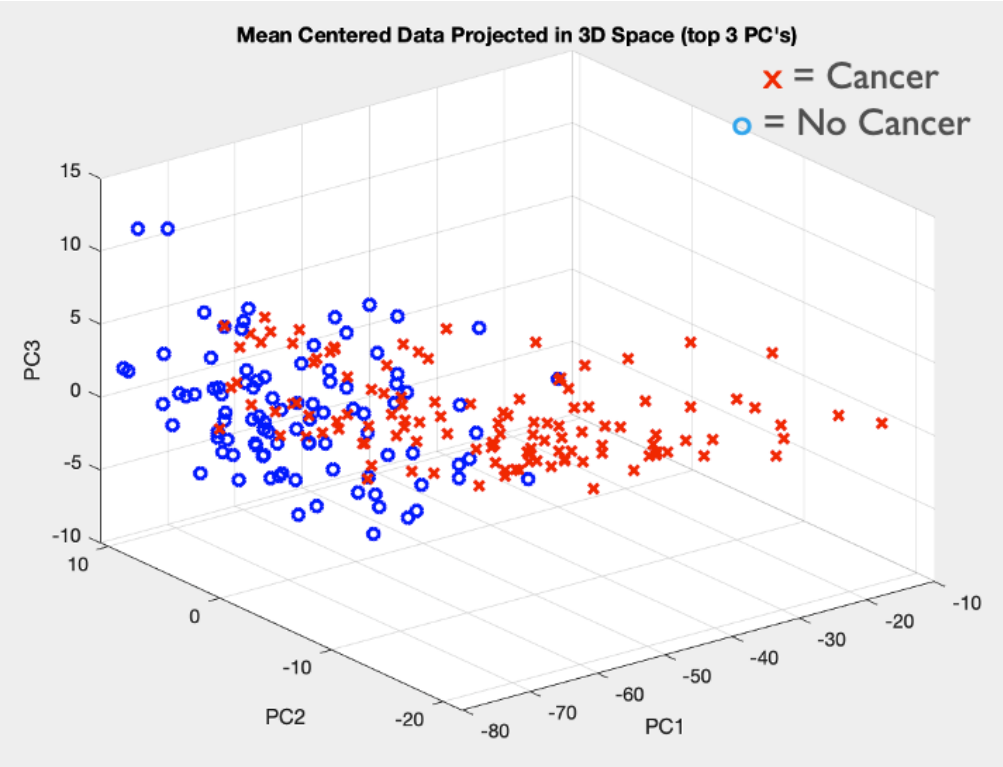


Figure 2: 3-D Visualization of Ovarian Cancer PCA. *Each point reflects a patient's PC_1, PC_2, and PC_3 values. Patients diagnosed with ovarian cancer are marked with a red $\boldsymbol{x}$, while normal patients are marked with a blue circle.*

The visualization reveals distinct clusters for cancer and normal patients, suggesting that PCA effectively separates these groups in lower-dimensional space. To confirm that PCA retains significant information from the original dataset, we examine the singular values and variance captured by the PCA system. Figure 3 shows:

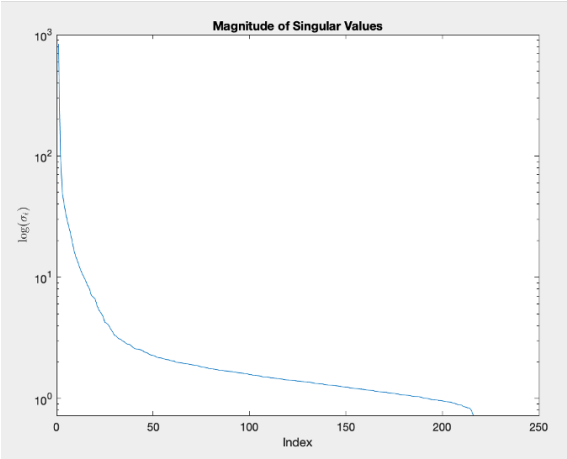
1. A log plot displaying the decay in singular value magnitudes.
2. The percentage of variance captured as the number of principal components increases.

The rapid decay of singular values indicates that the majority of variance is captured by the first few principal components. Specifically, the first three components capture 62% of the variance, with PC_1 alone accounting for 52.6%. This confirms the effectiveness of the dimensionality reduction achieved through PCA.

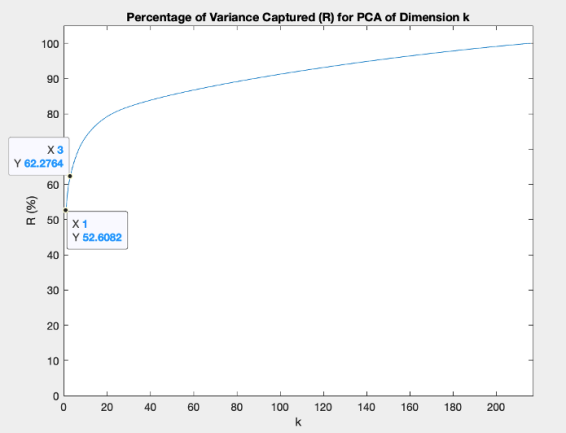
Example 2: Student Alcohol Consumption

In this example, we analyze a publicly available dataset containing demographic, academic, and lifestyle information about students, including their self-reported alcohol consumption. The goal is to determine whether the first few principal components can model the dataset effectively and reveal patterns, such as clustering students based on alcohol consumption.

Applying PCA to this dataset, we generate three plots: a 3D PCA visualization, a singular value decay plot, and a variance capture plot. Figure 4 presents these results:



(a) Singular Value Decay



(b) Variance Captured

Figure 3: Singular Value Decay and Variance Capture Plots.

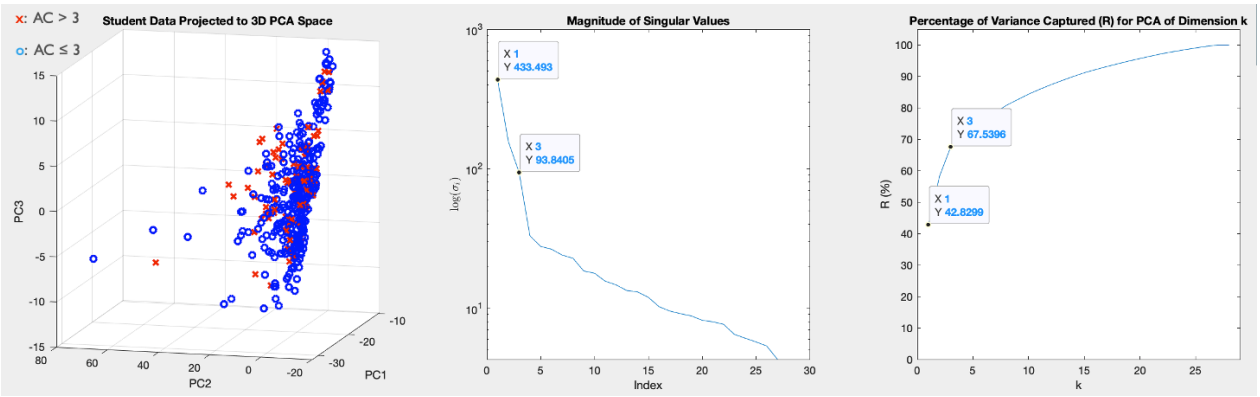


Figure 4: Principal Component, Singular Value Magnitude, and Variance Capture Plots for Student Alcohol Consumption. *Students rating their alcohol consumption above 3 are marked with a red* $\boldsymbol{x}$*, while those rating it 3 or below are marked with a blue circle.*

Although the 3D PCA visualization does not reveal clear clustering based on alcohol consumption, the singular value decay plot confirms that the first three components capture 67% of the variance. This indicates that a small number of "eigen-students" can effectively represent the dataset.

Example 3: Diabetes

The third example involves a dataset on patient health metrics related to diabetes, including glucose levels, blood pressure, and body mass index. The goal is to identify whether PCA can effectively model health profiles and cluster patients by diabetic status.

As with the previous examples, PCA is applied, and the results are visualized in Figure 5:

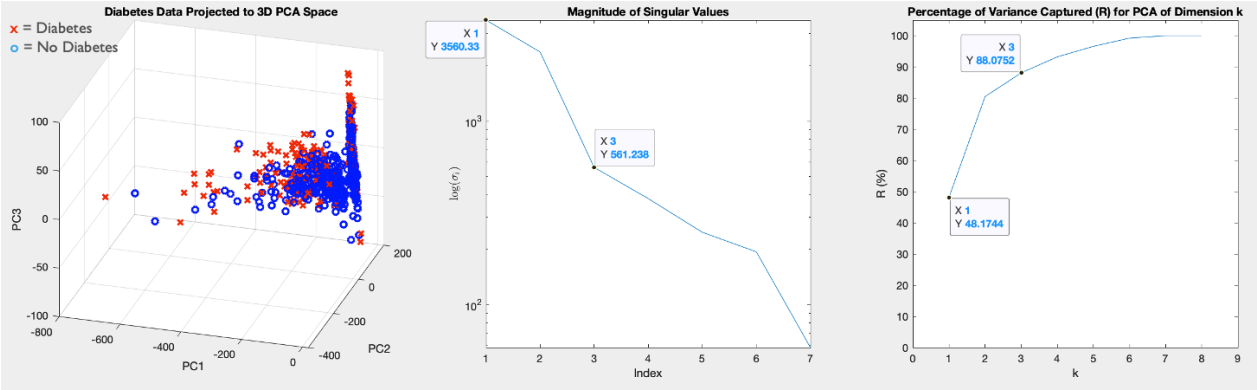


Figure 5: Principal Component, Singular Value Magnitude, and Variance Capture Plots for Diabetes Patients. *Patients with diabetes are marked with a red* $\boldsymbol{x}$*, while those without diabetes are marked with a blue circle.*

While no clear clusters are observed in the 3D PCA plot, the first three components capture 88% of the variance. This suggests that PCA effectively reduces dimensionality while preserving critical information, though the dataset's small size may limit the insights derived from clustering.

Online PCA

An interesting adaptation of the PCA is the so called online PCA as rigorously outlined in Boutsidis et al. [2014]. The online PCA is a sort of iterative PCA that progressively appends principal components to a PCA space rather than simultaneously. The online PCA is applicable when data is entering a database in real time and maintaining an updated principal component basis is important. There are also cases where a data matrix is too big to be stored. When one knows that most data will be ignored in the PCA construction, it saves space and time to use the online PCA.

The online PCA can be constructed by iteratively adding dimensions to the SVD. Similarly to before, we start with a mean-centered data matrix B. When the $t+1^{th}$ row ($b_{t+1} \in \mathbb{R}^{1\times n}$) is added to B, we must ensure that it is mean-centered by subtracting μ_{t+1}. Only then can we update the SVD as follows:

$$B_{t+1} = \begin{bmatrix} B \\ b_{t+1} \end{bmatrix} = U\Sigma V^T + \begin{bmatrix} \mathbf{0} \\ b_{t+1} \end{bmatrix},$$

where $U\Sigma V^T$ is the current SVD of B, b_{t+1} is the newly added row, and $\mathbf{0}$ is the zero matrix of size $t \times n$.

$$p = b_{t+1}V,$$

where V is the $n \times t$ matrix of right singular vectors of B. To properly orthoganlize V_{t+1}, we must calculate the residual r, which is the portion of b_{t+1} orthogonal to the current basis:

$$r = b_{t+1} - pV^T.$$

If $r = 0$, then b_{t+1} already lies in the span of V and can be discarded. Otherwise, append U with u_{t+1}:

$$u_{t+1} = \frac{r}{\|r\|}, \quad U = \begin{bmatrix} u_1 \mid u_2 \mid \cdots \mid u_{t+1} \end{bmatrix}.$$

The columns of U are orthogonal vectors. The new column u_{t+1} is the normalized component of our new vector b_{t+1} that is orthogonal to the other vectors in U. Thus, we maintain the unitary property of the matrix U. With this new addition to the principal component basis, adjusting the SVD matrices accordingly provides the PCA space for the $t+1^{th}$ iteration of the online PCA:

$$\Sigma_{t+1} = \begin{bmatrix} \Sigma & 0 \\ 0 & \|r\| \end{bmatrix}, \quad V_{t+1} = \begin{bmatrix} V & 0 \\ 0 & 1 \end{bmatrix},$$

$$\text{and,} \quad B_{\text{t+1}} = U_{\text{t+1}}\Sigma_{\text{t+1}}V_{\text{t+1}}^T.$$

Similarly to before, we can extract the principal components from the SVD by either multiplying $U\Sigma$ or BV where B is the data matrix that has been appended to include the new data vector b_{t+1}. As expressed earlier, the online PCA is useful for growing datasets. It leverages the SVD's ability to append itself in a computationally inexpensive and efficient manner by reusing previous information.

Conclusion

This paper presented the theoretical foundation and practical applications of PCA, emphasizing its role in dimensionality reduction and data interpretation. By leveraging the SVD, PCA becomes a robust and efficient tool for analyzing high-dimensional datasets. Examples from genomics, student surveys, and health metrics demonstrated its versatility, while the online PCA highlighted its adaptability to growing datasets. PCA remains an indispensable method for uncovering patterns and insights in complex data structures.

References

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

Christos Boutsidis, Dan Garber, Zohar Karnin, and Edo Liberty. *Online principal components analysis*. SIAM, 2014.

Steven L Brunton and J Nathan Kutz. *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press, 2022.

Alaa Tharwat. Principal component analysis-a tutorial. *International Journal of Applied Pattern Recognition*, 3(3):197–240, 2016.