Reading Report:
A Tutorial on Principal Component Analysis

Boyuan Du (2024151470021)
School of Software, **Sichuan University**
Email: 2024151470021@stu.scu.edu.cn

September 21, 2025

Course / Assignment: Reading report on PCA.
Primary Source: J. Shlens, "A Tutorial on Principal Component Analysis," 2014.

Abstract

PCA is a powerful, fundamental technique for extracting useful information from complex (high-dimensional) data by reducing dimensionality to the data's essential features. The essence of PCA lies in its ability to identify the directions of maximum variance in the data, represented by the principal components. By projecting the data onto these components, PCA effectively reduces redundancy and highlights the most informative aspects of the dataset. The tutorial emphasizes two main approaches to PCA: eigen-decomposition (ED) of the covariance matrix and singular value decomposition (SVD) of the data matrix. Both methods yield the same principal components. However, they share a common limitation: they assume principal components are orthogonal, which restricts them from capturing non-orthogonal structure; Independent Component Analysis (ICA) can address this by seeking statistically independent (not necessarily orthogonal) components. "About the experiment on MNIST . . . "

Keywords: PCA, covariance, eigen-decomposition, SVD, variance.

Contents

1 Foundation & Settings

Principal Component Analysis (PCA) is a technique to extract useful and hidden information and knowledge from complex and large datasets. It is entirely based on linear algebra to achieve the goal of dimensionality reduction, in which vectors and matrices are crucial and fundamental elements. Like the example of a fluctuating spring, PCA can extract the main movement along a single axis x from three different path records at different angles, which are both noisy and redundant. A matrix in PCA can serve multiple roles, such as

$$X = [x^{(1)} \ \cdots \ x^{(n)}] \in \mathbb{R}^{m \times n}, \tag{1}$$

in which each column is a sample and each row reflects a dimension of the dataset,
and **linear transformations** such as stretching, rotation, and projection:

$$C_Y = P C_X P^\top. \tag{2}$$

P projects C_X onto a new basis in which the off-diagonal elements are zero.
In the process of PCA, we measure the dataset along orthonormal directions and find the direction of the largest variance, which is called the **principal component (PC)**. Thus, in linear algebra, we use a set of orthonormal vectors as a basis to formalize how we measure the data.
The essence of PCA is to find a linear transformation P that projects the data X onto a new basis, $Y = PX$, in which the covariance C_Y becomes a diagonal matrix, in order to remove noise and redundancy in the raw data.

2 Covariance, Redundancy, and SNR

2.1 Covariance

Intuitive Definition of Covariance
Covariance represents the relationship between two variables. In PCA, it reflects the degree to which different dimensions of a dataset overlap, which we call *redundancy*. The larger the covariance is, the more two dimensions are correlated and overlapped, which means greater redundancy.

$$C_X = \tfrac{1}{n} X X^T, \qquad (C_X)_{ij} = \tfrac{1}{n} \sum_{\ell=1}^{n} X_{i\ell} X_{j\ell}. \tag{3}$$

2.2 Covariance, Redundancy, and Noise

In contrast to covariance, the diagonal elements of C_X represent the variance of each dimension, which reflects the useful information of the data; we call this direction *signal*. The *noise* is the opposite information in the direction perpendicular to the signal.
So we first divide the covariance matrix into two parts: diagonal elements and off-diagonal elements. The former consists of the variances of the signal direction and the noise direction. The latter consists of the covariances between different dimensions.
Now it is clear that PCA aims to reduce both the irrelevant information (noise) and the repetitive information (redundancy) by finding a new basis to project the data onto, where the covariance matrix becomes diagonal and the variance along each dimension is ordered from large to small.

2.3 SNR & Variance

Before we take steps to decrease noise, we measure the quality of the data by comparing the variance between the signal and the noise. Thus, we define the *Signal-to-Noise Ratio (SNR)* as:

$$\mathrm{SNR} = \frac{\sigma^2_{\text{signal}}}{\sigma^2_{\text{noise}}} \tag{4}$$

where σ^2_{signal} is the variance along the signal direction and σ^2_{noise} is the variance along the noise direction. A higher SNR indicates a cleaner and more reliable measurement, while a lower SNR suggests that the data are more contaminated by noise.
In PCA, we make an effort to increase the SNR by seeking directions where σ^2_{noise} approaches zero.

3 Assumptions & Limits of PCA

I. Linearity
PCA assumes a **linear change of basis**. With a dataset $X \in \mathbb{R}^{m\times n}$ and covariance $C_X = \frac{1}{n}XX^\top$, it seeks a linear transformation P to project X onto a new basis $Y = PX$ where the covariance $C_Y = \frac{1}{n}YY^\top$ becomes diagonal.

$$Y = PX, \quad P \in O(m), \quad P^\top P = I_m. \tag{5}$$

However, this assumption restricts PCA to extracting only linear features from data; Independent Component Analysis (ICA) can further tackle this problem efficiently.

II. Large Variance = Important Structure
PCA assumes that directions with the largest variance correspond to the most important underlying structure, which contains the most valuable information. We assume that the useful information is embedded in direction G, so this assumption is:

$$\forall\, g_{ij} \in G,\ \sigma^2 = \mathbf{g}_{ij} C_X \mathbf{g}_{ij}^\top \to \infty. \tag{6}$$

However, this assumption may not always be correct, and in real cases, it often leads to the wrong direction.

III. The principal components are orthogonal
PCA assumes that the principal components are orthogonal to each other. Now we demonstrate this assumption in the process of PCA: PCA seeks a linear transformation P to project X onto a new basis $Y = PX$, where

$$C_Y = PC_XP^\top \tag{7}$$

becomes diagonal. We assume that

$$P \in O(m), \quad p_i^\top p_j = 0 \ \ (\forall\, i \neq j), \qquad P^\top P = I_m \ \ (\forall\, i = j,\ p_i^\top p_j = 1). \tag{8}$$

This assumption is crucial because it ensures that the principal components are uncorrelated, which simplifies the analysis and interpretation of the data. However, it also limits PCA's ability to capture non-orthogonal features in the data.

4 PCA as a Change of Basis

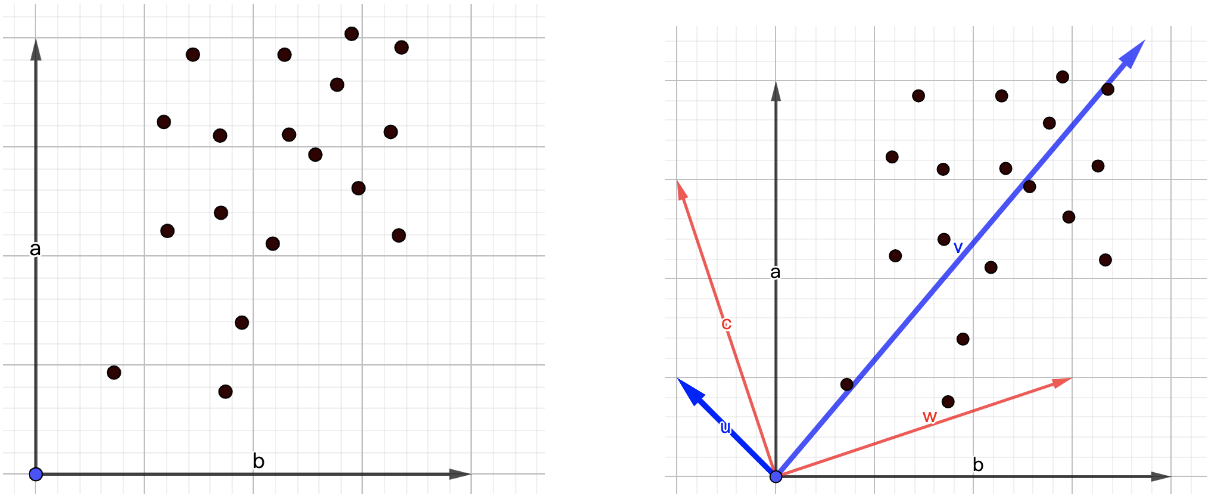
Change-of-Basis Formulation.

$$\mathbf{y} = \mathbf{P}\mathbf{x}, \qquad \mathbf{C}_Y = \mathbf{P}\mathbf{C}_X\mathbf{P}^\top \tag{9}$$

Principal component analysis is the process by which we change the basis of the data matrix; it is the way we change our directions of measurement and evaluation of the data, reducing noise and redundancy to approach an optimal condition. Thus, the principal component information is embedded in the matrix $\mathbf{P}$ that makes $\mathbf{C}_Y$ diagonal. Each row of $\mathbf{P}$ is a principal component, and the variance along that component is given by the corresponding diagonal element of $\mathbf{C}_Y$.

Projection and Re-expression

The way we re-express the data in the new basis is by projecting the data onto each principal component. The projection of a data point $\mathbf{x}$ onto a principal component $\mathbf{p}_i$ is given by the dot product $\langle \mathbf{p}_i, \mathbf{x} \rangle$. This projection gives us the coordinate of the data point in the new basis along that principal component. We can visualize this as a rotation and stretching of the basis vectors.



(a) Raw data with a random basis

(b) After PCA

Figure 1: Comparison of data before and after PCA. Figure (a) shows the raw data in a random basis, where the axes are not aligned with the data's main variance direction, leading to a noisy and redundant representation. Figure (b) illustrates the same data after PCA, where the axes are aligned with the principal components, resulting in a clearer and more structured representation.

Relation to Covariance

PCA is a linear transformation that re-expresses the dataset in a new orthonormal basis so that the covariance matrix becomes diagonal. In this process, the off-diagonal elements, which represent redundancy between variables, are eliminated, and the diagonal elements are the variances along the principal components. In conclusion, PCA is a tool for changing the basis of the data, and the covariance matrix is a measure of how well the data are represented in that basis.

5 Two Powerful Routes to PCA: ED & SVD

ED Route ***(Eigen Decomposition)***

As mentioned before, PCA seeks a linear transformation $\mathbf{P}$ to project $\mathbf{X}$ onto a new basis

$\mathbf{Y} = \mathbf{PX}$, where the covariance $\mathbf{C}_Y = \frac{1}{n}\mathbf{YY}^\top$ becomes diagonal. In linear algebra, we know that a symmetric matrix can be diagonalized by its eigenvectors. Thus, we can find the eigenvectors of the covariance matrix $\mathbf{C}_X = \frac{1}{n}\mathbf{XX}^\top$ to form the projection matrix $\mathbf{P}$. The eigenvectors of $\mathbf{C}_X$ are orthogonal and can be used as the new basis for the data. Intuitively, the eigenvectors represent the directions of maximum variance in the data, and the corresponding eigenvalues represent the amount of variance along those directions. By projecting the data onto the eigenvectors, we can effectively reduce redundancy and highlight the most informative aspects of the dataset.

$$\mathbf{C}_X = \mathbf{EDE}^\top, \qquad \text{PCs} \equiv \text{columns of } \mathbf{E},\ \text{variance} \equiv \text{diag}(\mathbf{D}) \tag{10}$$

SVD Route ***(Singular Value Decomposition)***
Another viable route to PCA is to use the singular value decomposition (SVD) of the data matrix $\mathbf{X}$. The left singular vectors u_i and right singular vectors v_i of $\mathbf{X}$ are the eigenvectors of $\mathbf{XX}^\top$ and $\mathbf{X}^\top\mathbf{X}$, respectively. The singular values σ_i are the square roots of the eigenvalues of $\mathbf{X}^\top\mathbf{X}$. SVD decomposes $\mathbf{X}$ into three matrices: $\mathbf{U}$, $\mathbf{\Sigma}$, and $\mathbf{V}^\top$. The columns of $\mathbf{V}$ are the right singular vectors of $\mathbf{X}$, which correspond to the principal components. The diagonal elements of $\mathbf{\Sigma}$ are the singular values, which are related to the variance along each principal component.

$$\mathbf{X} = \mathbf{U\Sigma V}^\top, \qquad \text{PCs} \equiv \text{columns of } \mathbf{V} \tag{11}$$

Up to now, it is clear and easy to see the pros and cons of the two methods: **ED** is more intuitive and directly related to the covariance matrix, but the realization of **ED** requires a couple of extra conditions on the matrix, namely that it must be square and symmetric. **SVD** is more general and can handle any data matrix, but it is less intuitive, and the decomposition into three matrices can be more computationally intensive.

6 The Viability of PCA in Dimensionality Reduction

PCA, as its name suggests, is a method to find the principal components of the data. But the way we define "*principal*" is not entirely clear. In the context of dimensionality reduction, the degree to which a reduced representation can predict the original data is how we define success; this, in turn, reveals how "principal" the components we use to simplify the data are. Quantitatively, we use the **Mean Squared Error (MSE)** to measure the difference between the original data and the data predicted from the simplified representation.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} \|x^{(i)} - \hat{x}^{(i)}\|^2 \tag{12}$$

To minimize the MSE, we need to maximize the variance along the principal components we choose to keep. The mathematical result is that we should choose the top k eigenvectors of the covariance matrix $\mathbf{C}_X$ or the top k right singular vectors of the data matrix $\mathbf{X}$, as in PCA.

$$\text{MSE} = \sum_{i=k+1}^{m} \lambda_i \tag{13}$$

Thus, PCA is a viable method for dimensionality reduction because it effectively captures the most important features of the data while minimizing the loss of information.

7 Conclusion

PCA is a powerful and fundamental technique for extracting useful information from complex, high-dimensional data by transforming the data into a new basis in which the covariance matrix becomes diagonal. The essence of PCA lies in its ability to identify the directions of maximum variance in the data, which are represented by the principal components. By projecting the data onto these components, PCA effectively reduces redundancy and highlights the most informative aspects of the dataset. ED and SVD are viable and powerful methods for achieving PCA as well as dimensionality reduction. However, PCA has limitations, such as its reliance on the assumption that principal components are orthogonal. In cases where this assumption does not hold, alternative techniques such as Independent Component Analysis (ICA) may be more appropriate.

A Mathematical Notations

Table 1: Alphabetical summary of mathematical notations used in the PCA tutorial

Notation	**Definition**	**Corresponds to**
A, B	Two general matrices used to define or explain other definitions below	tool matrices
a_i, b_i	i-th samples of A and B	scalar observations in the illustrative example
$C_X = \frac{1}{n}XX^\top$	Covariance matrix of X	reveals redundancy in dataset X
$C_Y = \frac{1}{n}YY^\top$	Covariance of $Y = PX$ in the new basis	covariance matrix after change of basis
D	Diagonal matrix of eigenvalues in eigendecomposition	variances along principal components
E	Matrix whose columns are eigenvectors of C_X	principal directions
I	Identity matrix	orthonormal basis of $\mathbb{R}^m$
k	Target dimension for reduction (also a scalar in the SVD example $Xa = kb$)	
m	Number of features (measurement types)	dimensionality of the dataset
n	Number of samples (trials)	size of the dataset
$P = [p_1^\top \cdots p_m^\top]$	Rotation and stretching that transform X into Y	projection matrix
p_i	i-th principal component (row of P)	principal axis
r	Rank of X (or $X^\top X$)	intrinsic dimensionality
$\mathrm{SNR} = \sigma^2_{\mathrm{signal}}/\sigma^2_{\mathrm{noise}}$	Signal-to-noise ratio	measurement quality
$\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$	Diagonal matrix of singular values	covariance in the new basis
σ_i	i-th singular value of X; $\lambda_i = \sigma_i^2/n$ (if $C_X = \frac{1}{n}XX^\top$)	scale of mode i
σ^2	Variance of a scalar variable/sequence	spread/energy
U	Left singular vectors of X	orthonormal basis of the column space
V	Right singular vectors of X	orthonormal basis of the row space
$\hat{u}_i$	i-th left singular vector; $\hat{u}_i = \frac{1}{\sigma_i}X\hat{v}_i$	output direction of mode i
$\hat{v}_i$	i-th eigenvector of $X^\top X$	input direction of mode i
$X \in \mathbb{R}^{m\times n}$	Data matrix	stacked measurements dataset
$x^{(j)}$	j-th sample vector (a column of X)	per-sample measurement
$Y = PX$	Data expressed in PCA coordinates	projections onto PCs
$Z = U^\top X$	Coordinates in the left-singular basis	transformed data
λ_i	i-th eigenvalue of C_X	variance along the i-th PC
δ_{ij}	Kronecker delta ($= 1$ if $i = j$, else 0)	orthogonality indicator
$\lVert \cdot \rVert$	Euclidean norm	vector length
$(\cdot)^\top$	Transpose	matrix transpose
$\cdot$	Dot product	inner product

References

[1] J. Shlens, *A Tutorial on Principal Component Analysis*, 2014.