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 and fundamental technique for extracting useful information from complex(high-dimensional) data by lowering dimension onto basic features of the data. 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. 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 same results with the help of eigenvector. However they has a common weakness that they rely on the assumption that principal components are orthogonal which limits the two techniques to extract non-vertical principal component, but Independent Component Analysis can handle this problem efficiently. "about the experiment of MNIST..."

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

Contents

1 Foundation & Settings

Principal Component Analysis(PCA) is a technique to extract useful and hiden information and knowlegde form confuse and big datasets. It completely bases on linear algebra to achieve the goal of dimension reduction of data, in which, the vector and matrix is crucial and foundamental elements. Like the example of a flunctuate spring, PCA can extract a main movement in a single axis x from three different path record in different angle which is both noisy and redundant. A matrix in PCA can serve multiple roles such as a

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

In which each column is a sample and each row reflect a dimension of datasets,

and **linear transformation** as stretch or rotation and projection:

$$C_Y = PC_XP^\top \tag{2}$$

P projects C_X onto a new basis where the off-diagonal elements equal to zero.
In the process of PCA, we measure the dataset from different directions orthonormally and find the direction with largest variance which is called **principal component(PC)**. So in linear algebra, we use a set of orthonormal vertors as basis, to reveal the the way we measure the data. The essence of PCA is to find a linear transformation P to project the data X onto a new basis $Y = PX$ where the covariance C_Y becomes a diagnoal matrix in order to remove the noise and redundancy in raw data material.

2 Covariance, Redundancy, and SNR

2.1 Covariance

Intuitive Definition of Covariance
Covariance represents the relationship between two variables. In PCA, it reflects the the degree of different dimensions of datasets overlapped which we called *Redundancy*. The larger the covariance is , the more two dimensions are correlated and overlapped which means bigger redundancy.

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

2.2 Covariance, Redundancy, and Noise

In contrast of 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*. And the emphnoise is the opposite infomation in the direction vertical to the signal. So we first divide the covariance matrix into two parts: diagonal elements and off-diagonal elements. The former consists of the variance of signal direction and the noise direction. The latter consists of the covariance between different dimensions. Now it is clear that PCA aims to reduce both the irrelavant 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 the noise, we measure the quality of data by compare the variance between signal and noise. So logically, 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 is more contaminated by noise. In PCA, we make effort to increase the SNR by seeking directions where σ^2_{noise} closes in on zero.

3 Assumptions & Limits of PCA

I. Linearity

PCA assumes a **linear change of basis**. With 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 only extract linear features from data, but Independent Component Analysis(ICA) can further tackle this problem efficiently.

II. Large Variance = Important Structure

PCA assumes that directions with 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, this often leads to a 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

$$C_Y = PC_X P^\top \tag{7}$$

becomes diagonal. And we assume that:

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

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 the ability of PCA to capture non-orthogonal features in the data.

4 PCA as a Change of Basis

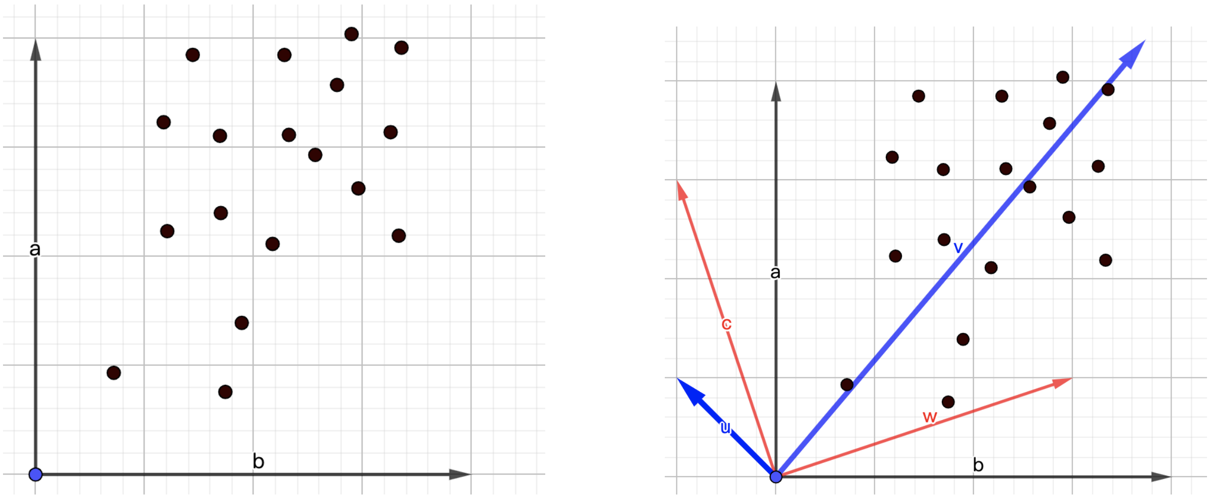
Change-of-Basis Formulation.

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

Principal components is the process where we change the basis of the data matrix; it is the very way we change our directions of measurement and evaluation on the data, reducing the noise and redundancy to approach the best condition. So the principal component is embedded in the matrix $\mathbf{P}$ that makes $\mathbf{C}_Y$ diagonal. Every 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 x onto a principal component p_i is produced 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. And we visualize it as the rotation and stretch of the basis vectors.



(a) Raw data with a random basis

(b) After PCA

Figure 1: Comparison of data before and after PCA.

Figure 2: 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) illustrate 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 which represent redundancy between variables are eliminated, and the diagonal elements is the variances along the principal components. In conclusion, PCA is a tool to change the basis of data, and the covariance matrix is a measure of how well the data is represented in that basis.

5 Two powerful route to PCA: ED & SVD

ED Route ***Eigen decomposition***

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

$\mathbf{Y} = \mathbf{P}\mathbf{X}$ where the covariance $\mathbf{C}_Y = \frac{1}{n}\mathbf{Y}\mathbf{Y}^\top$ becomes diagonal. In linear algebra, we know that a symmetric matrix can be diagonalized by its eigenvectors. So we can find the eigenvectors of the covariance matrix $\mathbf{C}_X = \frac{1}{n}\mathbf{X}\mathbf{X}^\top$ 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.

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

SVD Route *Singular Value Decomposition*
Another viable way 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 $\mathbf{X}\mathbf{X}^\top$ 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 matrix: $\mathbf{U}$, $\mathbf{\Sigma}$, 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.

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

Till now, it's 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** requries a couple of extra conditions of data matrix that it must be square and symmetric. **SVD** is more general and can deal with any data matrix, but it is less intuitive and the features of decomposing data into three matirx requires a surge of computation.

6 The Viability of PCA in Dimensional Reduction

PCA as it's name suggests, is a method to find the principal components of the data. But the way we define "*principal*" is not so clear. In the context of dimensional reduction, the degree to which a reduced representation can predict the original data is a how we define success and that 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 predict the data from the simplified representation.

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

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

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

So PCA is a viable method for dimensional 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 and high dimensional data by transform the data into a new basis where 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. And the way ED and SVD proved to be a viable and powerful method to achieve PCA as well as dimensional reduction. However, PCA has some limitations, such as its reliance on the assumption that principal components are orthogonal. In cases where these assumptions do not hold, substitute techniques such as Independent Component Analysis (ICA) may be more appropriate.

A Mathmatical Notations

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

Notation	Definition	Corresponds to
A, B	Two general matrix used to define or explain other definitions below	tool matrix
a_i, b_i	i-th samples of A and B	scalar observations in demonstrating example
$C_X = \frac{1}{n}XX^\top$	Covariance matrix of X	reveals the redundancy of dataset X
$C_Y = \frac{1}{n}YY^\top$	Covariance of $Y = PX$ under a 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 in $\mathbb{R}^m$
k	Target dimension for reduction($Xa = kb$in the explaination of SVD)	
m	Number of features (measurement types)	dimensions of dataset
n	Number of samples (trials)	scale of training set
$P = [p_1^\top \cdots p_m^\top]$	rotation and stretch to transforms$X$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
$\mathrm{SNR} = \sigma^2_{\text{signal}}/\sigma^2_{\text{noise}}$	Signal-to-noise ratio	measurement quality
$\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$	Diagonal matrix of singular values	covarience in 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 column space
V	Right singular vectors of X	orthonormal basis of 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}	element U($= 1$ 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.