A Tutorial on Principal Components Analysis

Research Review by William Steimel

Source

- A Tutorial on Principal Components Analysis
  - Johnathon Shlens – Google Research Mountain View, CA 94043 - 2014

Table of Contents

Abstract

- PCA is very popular in mainstream data analysis but is often considered a black-box that is widely used and poorly understood.
- This paper's aim is to "dispel the magic behind this black box" by building both a solid intuition on how PCA works and the mathematics behind PCA

I. Introduction

- Principal Components Analysis (PCA) is a standard tool in modern data analysis used from fields like neuroscience to computer graphics
  - Non-parametric, simple method for reducing complex data to lower dimensions to find hidden structure
- The goal of this paper is to educate readers about the usage and mathematics behind Principal Components Analysis

II. Motivation: Toy Example

- Lets pretend we are an experimenter trying to understand some unknown phenomenon by measuring quantities like (spectra, voltage, velocities. Etc.)
- This paper uses an example of trying to understand the motion of a physicist’s ideal spring as pictured in figure 1.
  - A ball of mass m is attached to a massless, frictionless spring
  - The ball is released a distance away from equilibrium and because the spring is ideal the ball oscillates along the x-axis indefinitely.
- This is a common problem in physics in which motion along the x direction is solved by an explicit function of time.
  - In reality, the system can be explained by a single variable x
- However, we do not know any of this so we decide to measure the balls position in 3 dimensional space
  - We do not know our true x, y,and z, axes in regard to this system so we chose three camera positions $\underset{a}{\rightarrow}$,$\underset{b}{\rightarrow}$, and $\underset{c}{\rightarrow}$ at arbitrary angles with respect to the system.
  - In the real world, experimenters do not often know which measurements best reflect the system and sometimes record more dimensions than needed.
- The question is how do we get from this dataset to a simple equation that represents x.
  - The goal of this tutorial is to understand how to systematically extract x using PCA
  - Our aim is to reduce dimensions to only the valuable dimensions

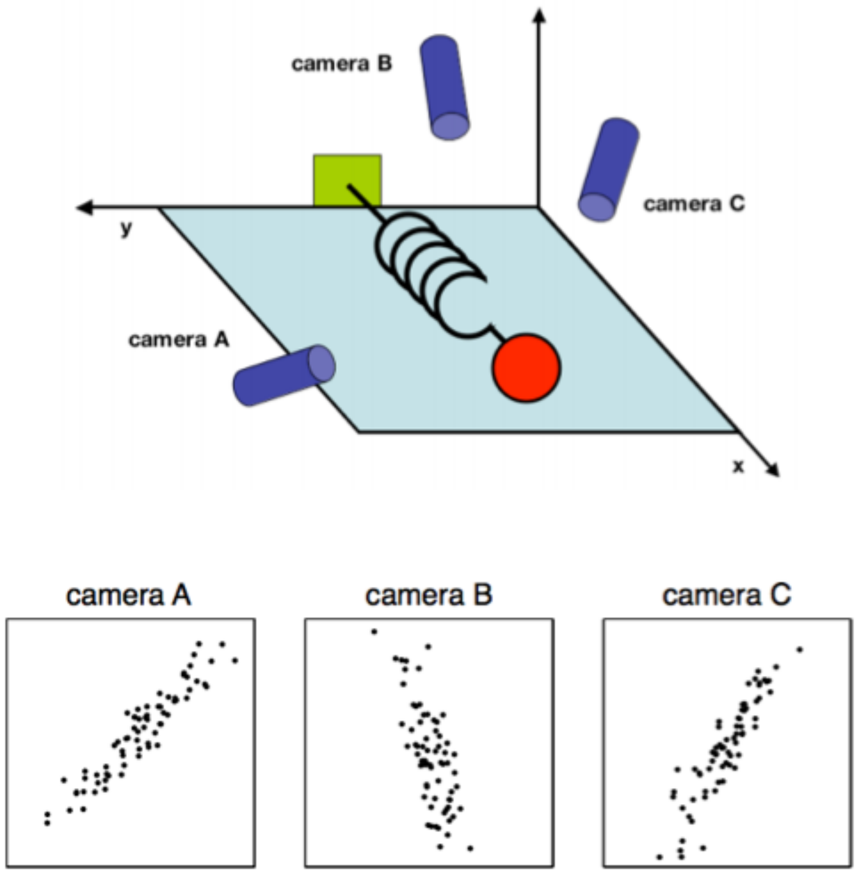

FIG. 1 A toy example. The position of a ball attached to an oscillating spring is recorded using three cameras A, B and C. The position of the ball tracked by each camera is depicted in each panel below.

III. Framework: Change Of Basis

- "The goal of principal component analysis is to identify the most meaningful basis to re-express a data set."
- The goal is this basis will filter out the noise and reveal hidden structure
  - Applied to the previous example, the goal of PCA is to determine that the unit basis vector along the spring or x-axis is the important dimension
- This technique allows an experimenter to determine which dimensions are important, redundant, or noise.

A. A Naïve Basis

- We must first define our data:
- Camera A records corresponding ball position at a point in time (x_A, y_A)
  - Each camera (A, B, C) contributes a 2-dimensional projection of the balls position.
- One sample trial can be represented as a 6 dimensional column vector.

$$\vec{X} = \begin{bmatrix} x_A \\ y_A \\ x_B \\ y_B \\ x_C \\ y_C \end{bmatrix}$$

- If we record this for 10 mins at 120 hertz than will have 72000 of these vectors.
  - 10 x 60 x 120

A. A Naïve Basis

- Each sample $\underset{x}{\rightarrow}$ is an m-dimensional vector where m is number of measurement types.
  - Every sample in the vector lies in an m-dimensional vector space spanned by some orthonormal basis.
- What is the orthonormal basis?
  - The naïve choice - The naïve basis reflects the method we gathered the data.
  - How do we express this naïve basis in linear algebra?
    - In the 2-dimensional case - 2x2 Identity Matrix $\{(1,0),(0,1)\}$
  - We can extend this to the m -dimensional case by constructing an $m\ x\ m$ identity matrix where each row is an orthonormal basis vector b_i with m components.

$$\mathbf{B} = \begin{bmatrix} \mathbf{b_1} \\ \mathbf{b_2} \\ \vdots \\ \mathbf{b_m} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \mathbf{I}$$

- All of our data has been recorded in this basis and can be expressed as a linear combination of $\{\boldsymbol{b}_i\}$

B. Change of Basis

- We must ask the question, “Is there another basis, which is a linear combination of the original basis, that best re-expresses our data set ?”
- PCA makes one strong assumption of linearity
  - With this assumption PCA is limited to re-expressing the data as linear combinations of its basis vectors
- Let $\boldsymbol{X}$ be the original data set, where each column is a single sample of our data set.
  - In the spring example, X is an $m\ x\ n$ matrix where m = 6 and n = 72000
- Let y be another $m\ x\ n$ matrix related by a linear transformation P
  - $\boldsymbol{X}$ is the original recorded dataset and $\boldsymbol{Y}$ is the new representation of the dataset as represented by the below formula:
    - $\boldsymbol{PX} = \boldsymbol{Y}$ (1)

B. Change of Basis

- In this section we follow the below definitions:
  - P_i are the rows of $\boldsymbol{P}$
  - X_i are the columns of $\boldsymbol{X}$
  - Y_i are the columns of $\boldsymbol{Y}$
- The previous equation $\boldsymbol{PX} = \boldsymbol{Y}$ represents a change in basis:
- There are a number of interpretations of this:
  - $\boldsymbol{P}$ is a matrix that transforms $\boldsymbol{X}$ into $\boldsymbol{Y}$
  - Geometrically, $\boldsymbol{P}$ is a rotation and stretch which transforms $\boldsymbol{X}$ into $\boldsymbol{Y}$
  - The rows of $\boldsymbol{P}$, $\{\boldsymbol{p_1}, \ldots, \boldsymbol{p_m}\}$ are a set of new basis vectors for expressing the columns of $\boldsymbol{X}$ (as seen by writing out the dot products of $\boldsymbol{PX}$)
- Each coefficient of $\boldsymbol{y_i}$ is a dot-product of $\boldsymbol{x_i}$ corresponding with row $\boldsymbol{P}$
  - This is the form of an equation where $\boldsymbol{y_i}$ is a a projection on to the basis of $\{\boldsymbol{p_1}, \ldots, \boldsymbol{p_m}\}$
    - The rows of $\boldsymbol{P}$ represent a new set of basis vectors for representing columns of $\boldsymbol{X}$

$$\mathbf{PX} = \begin{bmatrix} \mathbf{p_1} \\ \vdots \\ \mathbf{p_m} \end{bmatrix} \begin{bmatrix} \mathbf{x_1} & \cdots & \mathbf{x_n} \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} \mathbf{p_1 \cdot x_1} & \cdots & \mathbf{p_1 \cdot x_n} \\ \vdots & \ddots & \vdots \\ \mathbf{p_m \cdot x_1} & \cdots & \mathbf{p_m \cdot x_n} \end{bmatrix}$$

Dot Products of PX

$$\mathbf{y}_i = \begin{bmatrix} \mathbf{p_1 \cdot x_i} \\ \vdots \\ \mathbf{p_m \cdot x_i} \end{bmatrix}$$

Form of each column of Y

C. Questions Remaining

- The row vectors $\{\boldsymbol{p_1}, \ldots, \boldsymbol{p_m}\}$ in this transformation will become the principal components of $\boldsymbol{X}$.
- There are some questions that need to be answered.
  - “What is the best way to re-express **X**?”
  - “What is a good choice of basis **P**?”
- Assumptions other than linearity are needed in determining what features we would like $\boldsymbol{Y}$ to have.

IV. Variance and the Goal

- This section will answer the question of “What does best express the data mean?”

A. Noise and Rotation

- Measurement noise must be low or no information about the signal can be extracted. (regardless of analysis technique used)
- How to measure noise?
  - There is no absolute measure and it is usually represented relative to signal strength.
- Signal-to-Noise ratio (SNR) (ratio of variances) $= \dfrac{\sigma^2 signal}{\sigma^2 noise}$
  - High SNR indicates high precision >> 1
  - low SNR indicates high noise.
- Typically, the points of interest are along the directions with largest variance and highest SNR
  - In this case, the camera vectors $(x_A,\ y_A)$ do not represent the directions of highest variance.
- To maximize the variance we need to find right the rotation of a naïve basis
  - We need to rotate the naïve basis to lie parallel to the best fit line (direction indicated by the $\sigma^2 signal$) which would reveal the direction of the motion of the spring.

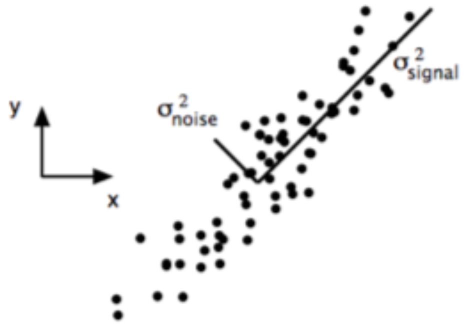

FIG. 2 Simulated data of (x, y) for camera A. The signal and noise variances σ^2_{signal} and σ^2_{noise} are graphically represented by the two lines subtending the cloud of data. Note that the largest direction of variance does not lie along the basis of the recording (x_A, y_A) but rather along the best-fit line.

B. Redundancy

- Figure 3. details another factor called redundancy in our data which heavily impacts the spring example.
  - The left plot shows no apparent relationship and is considered uncorrelated.
  - The plot on the right is considered on the other extreme and has highly correlated (redundant) recordings.
- There are two possible reasons for this redundancy in this example including:
  - Cameras A and B being very close
  - A plot where one axis is in meters and another is in inches (not on the same unit scale)
- Did we really need to record two variables in the case of high redundancy as one variable would express the data more effectively. (2 → 1 dimensions)
  - This is the central idea behind dimensionality reduction.

FIG. 3 A spectrum of possible redundancies in data from the two separate measurements r_1 and r_2. The two measurements on the left are uncorrelated because one can not predict one from the other. Conversely, the two measurements on the right are highly correlated indicating highly redundant measurements.

C. Covariance Matrix

- With 2 variables it is easy to identify redundant cases by drawing the best fitting line but what about higher dimensions?
  - Co-variance!
- Covariance measures the degree of linear relationship between two variables.
  - Large positive indicates positively correlated data.
  - Large negative indicates negatively correlated data.
  - Magnitude measures the degree of redundancy

C. Covariance Matrix

- Pre-requisites for Co-variance Matrix -
  - Consider two sets of measurements with zero means:
    - $A = \{a_1, a_2, \dots, a_n\}, \quad B = \{b_1, b_2, \dots, b_n\}$
  - The individual variance of A and B can be defined as:

$$\sigma_A^2 = \frac{1}{n}\sum_i a_i^2, \; \sigma_B^2 = \frac{1}{n}\sum_i b_i^2$$

  - The co-variance between A and B can be generalized as:

$$\sigma_{AB}^2 = \frac{1}{n}\sum_i a_i b_i$$

  - We can convert A and B into row vectors to express co-variance as a a dot product matrix computation.
    - $\boldsymbol{a} = [a_1 a_2 \dots a_n], \quad \mathbf{b} = [b_1 b_2 \dots b_n]$
    - $\sigma_{AB}^2 = \frac{1}{n} ab^T$
  - We can then rename the row vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ as $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ and define a new $m\ x\ n$ matrix $\boldsymbol{X}$ depending on the amount of row vectors (variables) $\boldsymbol{x_3}, \dots, \boldsymbol{x_m}$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_1} \\ \vdots \\ \mathbf{x_m} \end{bmatrix}$$

    - Each row of X corresponds to all measurements of a particular type
    - Each column of X corresponds to a set of measurements from one observation

C. Covariance Matrix

- Definition of Co-variance Matrix
  - $\boldsymbol{C_X} = \frac{\boldsymbol{1}}{\boldsymbol{n}}\boldsymbol{XX^T}$
  - The $ijth$ element of $\boldsymbol{C_X}$ is the dot product between the vector of the ith measurement type with the vector of the jth measurement type.
- Properties of Co-variance Matrix $\boldsymbol{C_X}$
  - $\boldsymbol{C_X}$ is square symmetric $m\ x\ m$ matrix
  - The diagonal terms of $\boldsymbol{C_X}$ are the variance of measurement types
  - Off diagonal terms of $\boldsymbol{C_X}$ are co-variance between measurement types.
- $\boldsymbol{C_X}$ tells us the covariance between all possible pairs of measurements.
  - The values reflect the noise and redundancy in our measurements.
    - Diagonal Terms - Large values (Interesting Structure) - Variance Terms
    - Off-Diagonal Terms – Large Magnitudes (High Redundancy) – Co-variance Terms

D. Diagonalize the Covariance Matrix

- As from the previous sections, we can see clearly our goal is to minimize redundancy (covariance) and maximize the signal (variance)
- What would an optimized covariance matrix $\boldsymbol{C_Y}$ look like ?
  - Off-diagonal terms in $\boldsymbol{C_Y}$ should be 0 meaning $\boldsymbol{C_Y}$ is a diagonal matrix and **Y** is decorrelated
  - Each dimension in **Y** should be ordered by variance.
    - With ordering we can judge the importance of each principal direction (component)
- There are many methods for diagonalizing matrices but PCA assumes that all basis vectors $\{\boldsymbol{p_1}, \dots, \boldsymbol{p_m}\}$ are orthonormal.
- In the simple 2-d example from Figure 2, **P** is an orthonormal matrix and acts as a rotation to align a basis with the axis of maximal variance
- In multiple dimensions this can be performed with a Simple Algorithm:
  - Select normalized direction in m-dimensional space in which variance $\boldsymbol{X}$ is maximized- Save as vector $\boldsymbol{p_1}$
  - Find another direction which variance is maximized but restrict search to all directions orthonormal to all previous directions. Save this vector as $\boldsymbol{p_i}$
  - Repeat until m vectors are selected.
  - The ordered set of variances $\boldsymbol{p}'s$ are called the principal components.

E. Summary of PCA Assumptions

- Summary of the previously presented assumptions:
  - Linearity
    - Change of Basis is a core element of PCA
  - Large Variances have important structure
    - Principal components with larger associated variance represent interesting structure while those with smaller associated variance represent noise (sometimes)
  - The Principal Components are orthogonal
    - This assumption makes PCA soluble with linear algebra decomposition techniques

V. Solving PCA Using Eigenvector Decomposition

- This paper talks about two methods for deriving PCA through eigenvectors of covariance and Singular Value Decomposition.
- Typically the steps for computing PCA using Eigenvector Decomposition of dataset **X** $(m\ x\ n$ matrix) includes:
  - Step 1 - Subtracting off the mean of each measurement type
  - Step 2 - Calculate the Co-Variance Matrix $\boldsymbol{C_X}$
    - $\boldsymbol{C_X = \frac{1}{n} X X^T}$
  - Step 4 - Calculate the eigenvectors and eigenvalues of the covariance matrix C_x
    - The Principal components of **X** are the eigenvectors of $\boldsymbol{C_X = \frac{1}{n} X X^T}$
  - Step 5 – Choosing the components and forming a feature vector (Hyperparameter – number of components)
    - $\boldsymbol{P}$ - Eigenvectors with higher associated eigenvalues are considered the components with highest explained variance.
  - Step 5 – Transform to new dataset
    - $\boldsymbol{PX = Y}$

VII. Discussion

- PCA has many widespread applications as it can reveal simple structure in complex data using analytical solutions from linear algebra
- The measurement of variance is useful as it allows for comparison of importance of each dimension
  - The goal is for a small number of components (smaller than the dataset features) to represent a great deal of information "characterization" from the dataset
    - This is the main goal of dimensionality reduction methods
- Any scientist or researcher one should ask and be aware of when PCA may or can fail.
    - One of the beautiful aspects of PCA is that it is a non-parametric method without hyperparameters
    - Data can be plugged in with an answer outputted.
    - This characteristic of PCA being "agnostic" to data source, can also be considered a weakness as seen from the Ferris wheel example.

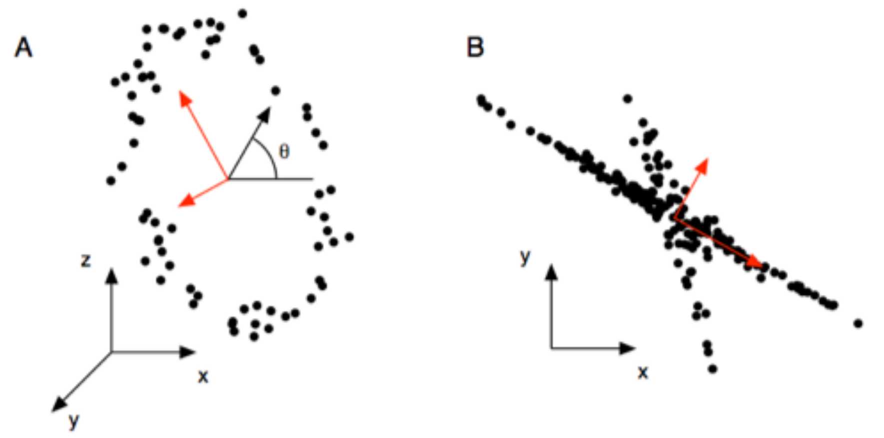

FIG. 6 Example of when PCA fails (red lines). (a) Tracking a person on a ferris wheel (black dots). All dynamics can be described by the phase of the wheel θ, a non-linear combination of the naive basis. (b) In this example data set, non-Gaussian distributed data and non-orthogonal axes causes PCA to fail. The axes with the largest variance do not correspond to the appropriate answer.