Principal Component Analysis

Christopher Ting

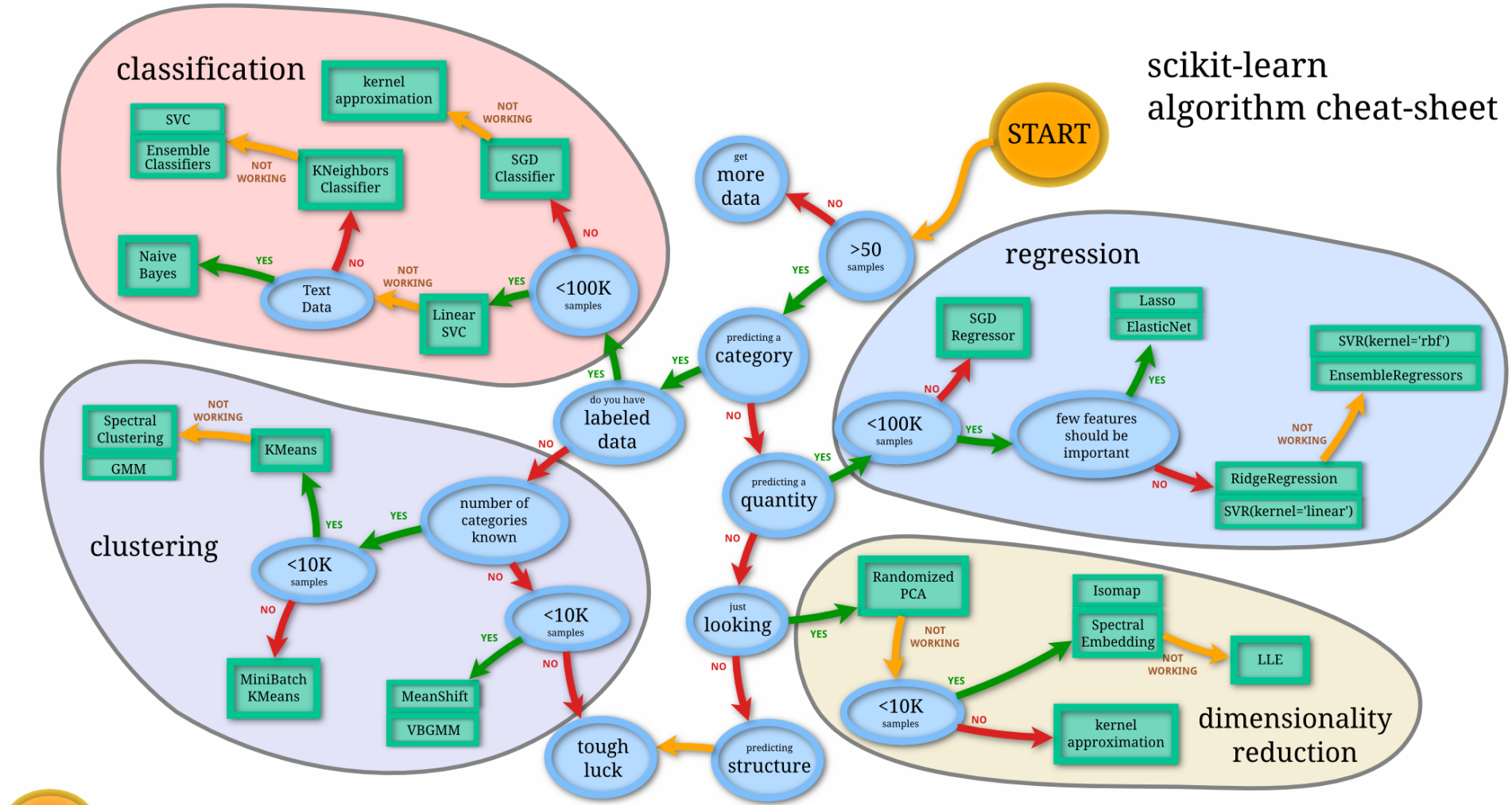
http://cting.x10host.com/CUHKSZ/CUHKSZ.html

http://www.mysmu.edu/faculty/christophert/
✉: christophert@smu.edu.sg

June 27, 2018

Broad Lesson Plan

Machine Learning with Python



Introduction

- Principal component analysis (PCA) is a technique that is useful for the **compression** and **classification** of data.
- The intent is to **reduce the dimensionality** of a data set (sample) by finding a new set of variables, smaller than the original set of variables, that nonetheless retains most of the sample's information.
- PCA is an application of **linear algebra**.
- PCA is basically rotating and shifting the axes to one that presents important information.
- PCA has a lot of applications across different fields.

Definition of Vector Differentiation

- Let $\boldsymbol{x}$ be a column k-vector. Consider the function

$$g(\boldsymbol{x}) = g(x_1, x_2, \ldots, x_k) : \Re^k \longrightarrow \Re.$$

- Vector derivative of a scaler $g(\boldsymbol{x})$ is

$$\frac{\partial}{\partial \boldsymbol{x}} g(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1} g(\boldsymbol{x}) \\ \frac{\partial}{\partial x_2} g(\boldsymbol{x}) \\ \vdots \\ \frac{\partial}{\partial x_k} g(\boldsymbol{x}) \end{bmatrix}. \tag{1}$$

and

$$\frac{\partial}{\partial \boldsymbol{x}^\intercal} g(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1} g(\boldsymbol{x}) & \frac{\partial}{\partial x_2} g(\boldsymbol{x}) & \cdots & \frac{\partial}{\partial x_k} g(\boldsymbol{x}) \end{bmatrix} \tag{2}$$

Definition of Vector Differentiation (Cont'd)

- Vector derivative of a vector function, e.g. $\begin{bmatrix} f(\boldsymbol{x}) \\ g(\boldsymbol{x}) \end{bmatrix}$ is to treat each element of the vector as a scalar and apply the rules (1) and (2) for vector derivative of a scalar.

$$\frac{\partial}{\partial \boldsymbol{x}} \begin{bmatrix} f(\boldsymbol{x}) \\ g(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{x}) \\ \frac{\partial}{\partial \boldsymbol{x}} g(\boldsymbol{x}) \end{bmatrix} \tag{3}$$

and

$$\frac{\partial}{\partial \boldsymbol{x}^\intercal} \begin{bmatrix} f(\boldsymbol{x}) \\ g(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{x}^\intercal} f(\boldsymbol{x}) & \frac{\partial}{\partial \boldsymbol{x}^\intercal} g(\boldsymbol{x}) \end{bmatrix} \tag{4}$$

- Likewise for a matrix of k-valued functions

Basic Properties

- For constant vector $\boldsymbol{a}$ and matrix $\boldsymbol{A}$,

$$\frac{\partial}{\partial \boldsymbol{x}}\left(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{x}\right) = \frac{\partial}{\partial \boldsymbol{x}}\left(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{a}\right) = \boldsymbol{a}, \qquad \frac{\partial}{\partial \boldsymbol{x}^{\mathsf{T}}}\left(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{x}\right) = \boldsymbol{a}^{\mathsf{T}}$$

$$\frac{\partial}{\partial \boldsymbol{x}^{\mathsf{T}}}\left(\boldsymbol{A}\boldsymbol{x}\right) = \boldsymbol{A}$$

$$\frac{\partial}{\partial \boldsymbol{x}}\left(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x}\right) = \left(\boldsymbol{A} + \boldsymbol{A}^{\mathsf{T}}\right)\boldsymbol{x}$$

$$\frac{\partial^2}{\partial \boldsymbol{x}\partial \boldsymbol{x}^{\mathsf{T}}}\left(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x}\right) = \boldsymbol{A} + \boldsymbol{A}^{\mathsf{T}}$$

Vector of Random Variables

- Let $\boldsymbol{x}$ be a vector of m random variables (m-variate) with mean vector $\boldsymbol{\mu} = \mathbb{E}\left(\boldsymbol{x}\right)$ and covariance matrix $\boldsymbol{\Sigma} = \mathbb{C}\left(\boldsymbol{x}\right)$.
- Example: $m = 2$

$$\boldsymbol{x} = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

  - $\sigma_{ii} = \sigma_i^2$ is the variance of r_i, $i = 1, 2$.
  - The covariance of r_1 with r_2 is $\sigma_{12} = \sigma_{21}$.
- Consider a linear combination of m random variables:

$$c = a_1 x_1 + a_2 x_2 + \cdots + a_m x_m$$

- Variance of c,

$$\mathbb{V}(c) = \mathbb{V}\left(\boldsymbol{a}^\intercal \boldsymbol{x}\right) = \boldsymbol{a}^\intercal \mathbb{C}\left(\boldsymbol{x}\right)\boldsymbol{a} = \boldsymbol{a}^\intercal \boldsymbol{\Sigma} \boldsymbol{a}$$

Example: $n = 2$ (Bivariate)

- Note that $\mathbb{V}\left(c\right) = \mathbb{C}\left(c, c\right)$
- Hence,

$$\begin{aligned}
\mathbb{C}\left(a_1 r_1 + a_2 r_2,\, a_1 r_1 + a_2 r_2\right) &= a_1^2\,\mathbb{C}\left(r_1, r_1\right) + a_1 a_2\,\mathbb{C}\left(r_1, r_2\right) \\
&\quad + a_2 a_1\,\mathbb{C}\left(r_2, r_1\right) + a_2^2\,\mathbb{C}\left(r_2, r_2\right) \\
&= a_1^2\,\mathbb{V}(r_1) + 2 a_1 a_2\,\mathbb{C}\left(r_1, r_2\right) + a_2^2\,\mathbb{V}(r_2) \\
&= a_1\sigma_{11} + a_1 a_2 \sigma_{12} + a_2 a_1 \sigma_{21} + a_2^2 \sigma_{22} \\
&= \begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \\
&= \boldsymbol{a}^{\intercal} \boldsymbol{\Sigma} \boldsymbol{a}
\end{aligned}$$

Maximization with Constraints

- Problem formulation

$$\max_{\boldsymbol{a}} \boldsymbol{a}^\intercal \boldsymbol{\Sigma} \boldsymbol{a} \qquad \text{subject to } \boldsymbol{a}^\intercal \boldsymbol{a} = 1.$$

- Use the Lagrange multiplier technique. Consider instead a cost function f

$$f(\boldsymbol{a}, \lambda; \boldsymbol{x}) = \boldsymbol{a}^\intercal \boldsymbol{\Sigma} \boldsymbol{a} + \lambda\left(1 - \boldsymbol{a}^\intercal \boldsymbol{a}\right)$$

- The first-order condition is

$$\frac{\partial f}{\partial \boldsymbol{a}} = 2\boldsymbol{\Sigma}\boldsymbol{a} - 2\lambda\boldsymbol{a} = \boldsymbol{0}$$
$$\frac{\partial f}{\partial \lambda} = 1 - \boldsymbol{a}^\intercal \boldsymbol{a} = 0$$

- We obtain the characteristic equation of eigenvalue problem:

$$\boldsymbol{\Sigma}\boldsymbol{a} = \lambda\boldsymbol{a}$$

First Principal Component

- With respect to the eigenvector $\boldsymbol{a}$,

$$\mathbb{V}\left(c\right) = \boldsymbol{a}^{\intercal}\boldsymbol{\Sigma}\boldsymbol{a} = \boldsymbol{a}^{\intercal}\left(\lambda\boldsymbol{a}\right) = \lambda\boldsymbol{a}^{\intercal}\boldsymbol{a} = \lambda$$

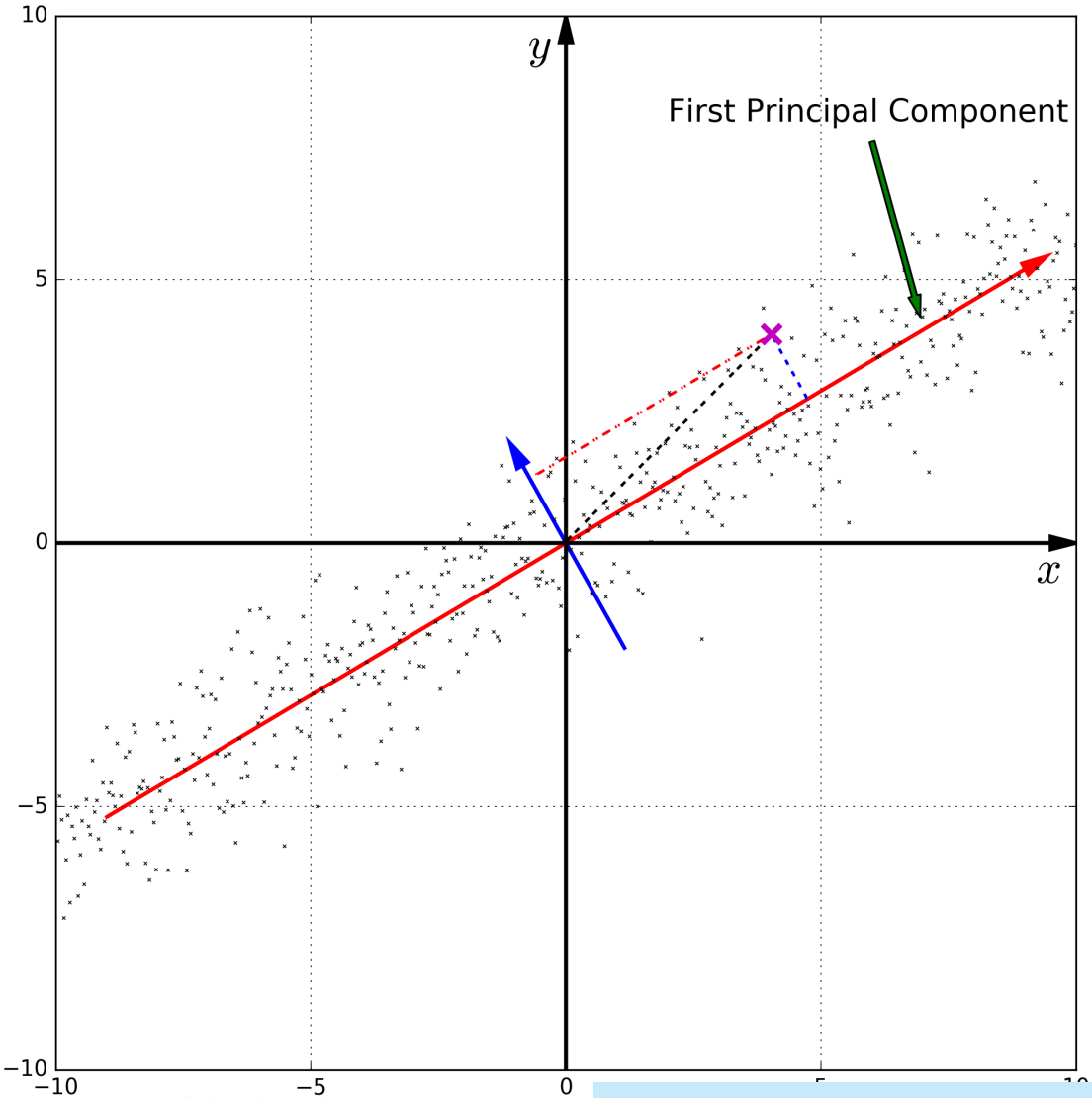
- Therefore, $\mathbb{V}(c)$ is maximized if λ is the largest eigenvalue.
- Correspondingly, c is the first principal component for the first eigenvector $\boldsymbol{a}$.
- In general, suppose there are m random variables, eigenvectors $\boldsymbol{a}_i$ of eigenvalues λ_i, $i = 1, 2, \ldots, m$, we have,

$$\mathbb{V}\left(\boldsymbol{a}_i^{\intercal}\boldsymbol{x}\right) = \lambda_i \tag{5}$$

PCA: Coordinate Transformation

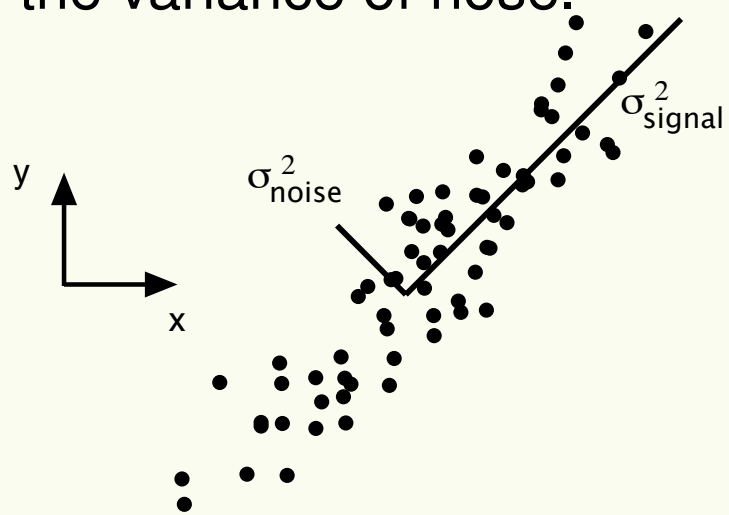
- Variance along the direction of first principal component (u axis) is largest.
- The axis perpendicular to the u axis is second principal component (v axis).
- Since the radius r (length of dashed black lines) should not change,

$$r^2 = x^2 + y^2 = u^2 + v^2$$

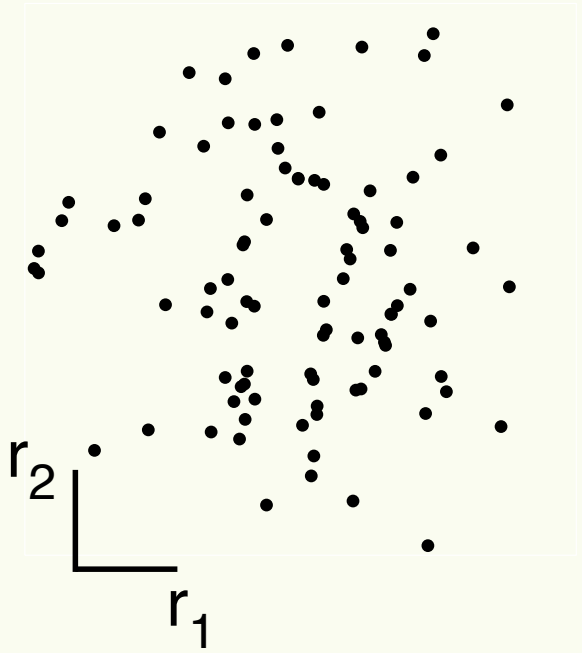


Insight!

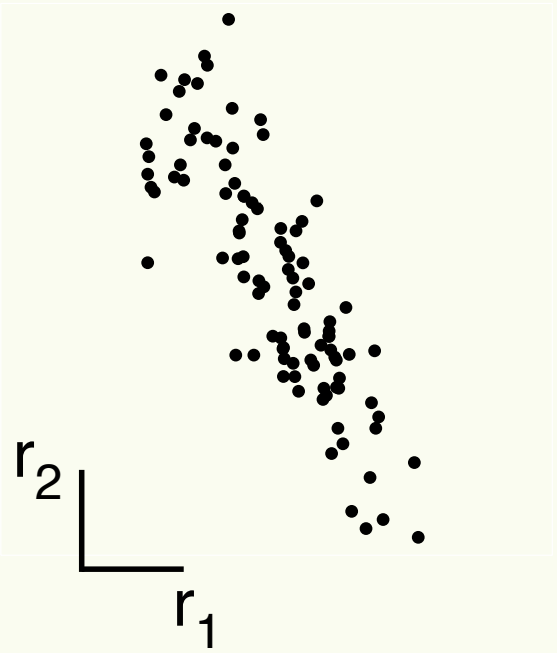
- In standard asset pricing since Markowitz, variance (or standard deviation) is a measure of risk (more appropriately uncertainty).
- In PCA, variance σ_u^2 is a measure of information (signal). The larger dispersion of data along the u axis suggests that more bits are needed to describe it.
- In the two-dimension case, the variance σ_v^2 of dispersion in the direction of the u axis can be considered as the variance of nose.
- **Signal-to-noise ratio** SNR $= \dfrac{\sigma_u^2}{\sigma_v^2}$

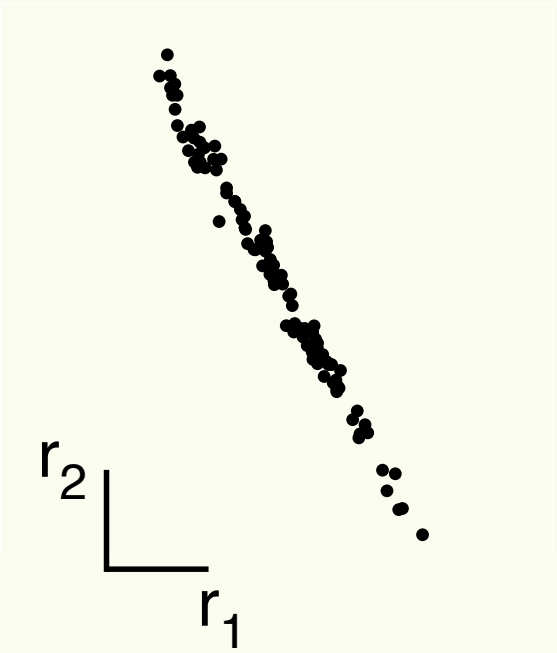


Redundancy



low redundancy



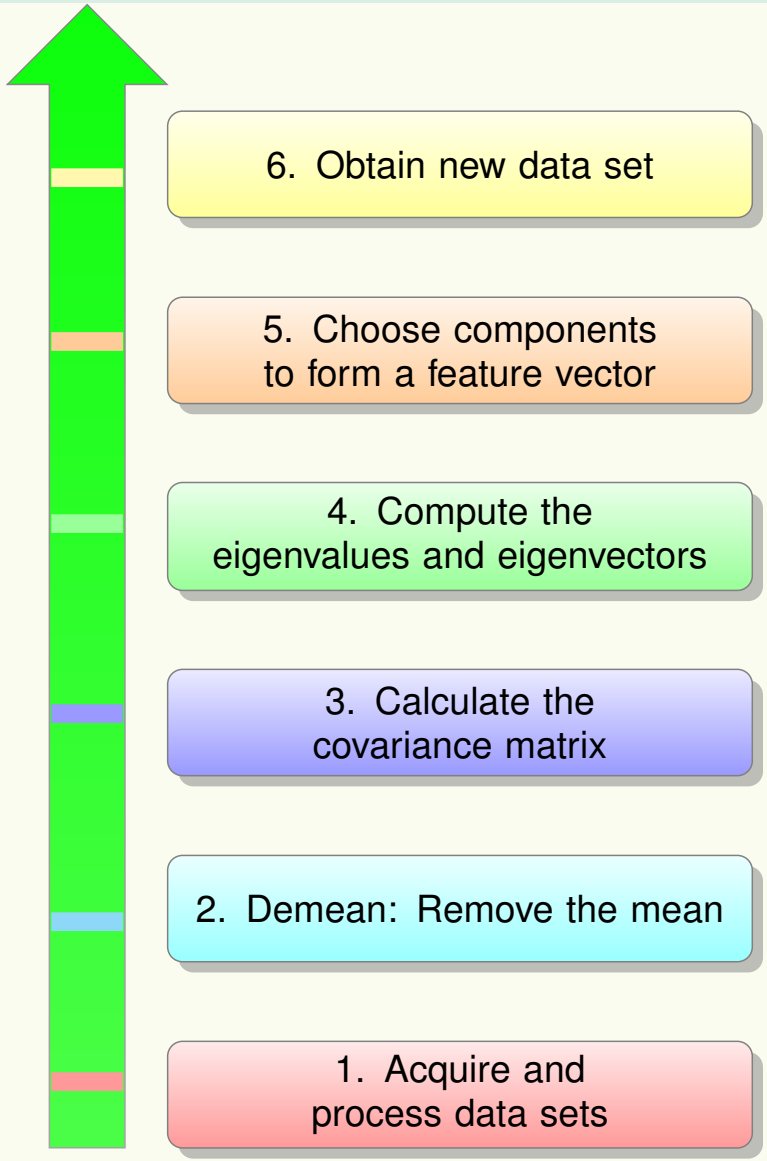


high redundancy

➠ Central idea: correlation, given r_1, you know quite well about r_2 and vice versa

➠ Implication: **Dimension reduction**!

PCA Algorithm



Demo Python Codes

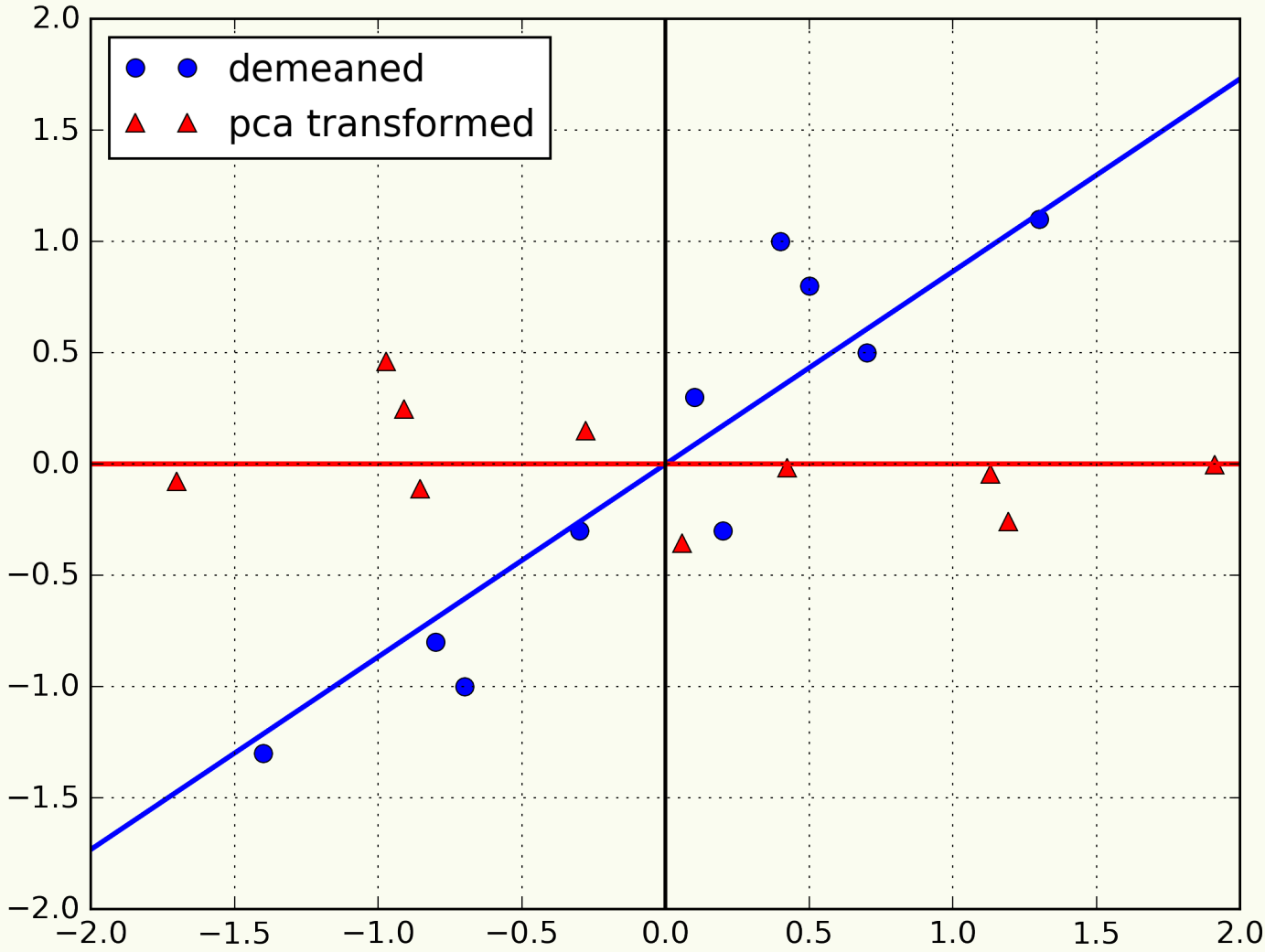
```
from __future__ import  print_function, division
import numpy as np
import matplotlib.pyplot as plt

# Step 1: Collect data
year = range(2007,2017)
x_original = [2.5,0.4,2.2,1.9,3.1,2.3,2.0,1.0,1.5,1.1]   # Annual return of Stock X in %
y_original = [2.4,0.6,2.9,2.2,3.0,2.7,1.6,1.1,1.6,0.9]   # Annual return of Stock Y in %

# Step 2: Demean
mx = np.mean(x_original)
my = np.mean(y_original)

x = np.asarray(x_original - mx)
y = np.asarray(y_original - my)

X = np.matrix([x, y])

# Sep 3: Compute the variance-covariance matrix
Sigma = np.cov(X, ddof=1)

# Step 4: Compute the eigenvalues and eigenvectors
eigenvalues, eigenvectors = np.linalg.eig(Sigma)

# Step 5: Choose components to form features
features2 = eigenvectors
```

Demo Python Codes (Cont'd)

```
# Step 6: Get the data back after transformation
fdata2 = features2.dot(X)
fx = np.asarray(fdata2[0])
fy = np.asarray(fdata2[1])
fx = fx[0]
fy = fy[0]

# Visual presentation of results
plt.plot(x, y, 'bo',
markersize=7, label="demeaned")
plt.plot(fx, -fy, 'r^',
markersize=7, label="pca transformed")

slope = Sigma[0,1]/np.var(y, ddof=1)
lx = np.linspace(-2, 2, 100)
ly = lx * slope
plt.plot(lx, ly, '-b', linewidth=2)

ly2 = np.zeros(100, dtype=float)
plt.plot(lx, ly2, '-r', linewidth=2)
plt.plot(ly2, lx, '-k', linewidth=1.5)
plt.grid()
plt.legend(loc="upper left")
plt.show()
```

Results

- Blue line is the principal component, obtained by OLS regression.
- Data (in red) obtained back from PCA have no slope.
- PCA eigenvectors are invariant with respect to sign change.



Remarks

- Linearity sets up the problem as a change of basis.
- Large variances have important structure.
- The principal components are orthogonal.

Basis

- A coordinate system can be expressed as m vectors that are **orthogonal** to each other.
- m-dimensional coordinate system: $m \times m$ identity matrix

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{b}_1^\intercal \\ \boldsymbol{b}_2^\intercal \\ \vdots \\ \boldsymbol{b}_m^\intercal \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} =: \boldsymbol{I}$$

where each row is an **orthonormal basis vector** $\boldsymbol{b}_i^\intercal$ with m components.

- Examples:
  - A portfolio or index with m component stocks
  - Yield curve with m anchor maturities

Change of Basis

- The goal of principal component analysis is to identify the most meaningful **basis** to re-express a data set.

- Is there another basis, which is a linear combination of the original basis, that best re-expresses our data set?

- Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be $m \times n$ matrices related by a linear transformation $\boldsymbol{P}$.

$$\boldsymbol{PX} = \boldsymbol{Y}$$

- $\boldsymbol{X}$ is the original recorded data set and $\boldsymbol{Y}$ is a **re-representation** of that data set.

Covariance Matrix from Data Matrix

- Data matrix $\boldsymbol{X}$: n sets of observations. Each set has m observations for m variables of interest
- Sample variance-covariance matrix

$$\boldsymbol{C}_X = \frac{1}{n-1}\boldsymbol{X}\boldsymbol{X}^\intercal$$

- $\boldsymbol{C}_X$ is a square symmetric $m \times m$ matrix.
- The diagonal terms of $\boldsymbol{C}_X$ are the variances of particular measurement types.
- The off-diagonal terms of $\boldsymbol{C}_X$ are the covariances between measurement types.

Covariance Matrix in PCA

- The diagonal terms, by assumption, correspond to principal structure, especially for the largest ones.
- For the off-diagonal terms, large magnitudes correspond to high redundancy.
- Our goals are to
  1. minimize redundancy, measured by the magnitude of the covariance
  2. maximize the signal, measured by the variance.
- What should be the properties of C_Y for PCA?

Diagonalization of Covariance Matrix

- All off-diagonal terms in C_Y should be zero.
- Thus, C_Y must be a diagonal matrix. Or, said another way, Y is **decorrelated**.
- Each successive dimension in Y should be rank-ordered according to variance.

How PCA Works

- $\boldsymbol{P}$ acts as a **generalized rotation** to align a basis with the axis of maximal variance. In multiple dimensions this could be performed by a simple algorithm.
- Select a normalized direction in m-dimensional space along which the variance in $\boldsymbol{X}$ is maximized. Save this vector as $\boldsymbol{p}_1$.
- Repeat this procedure until m vectors are selected.
- The resulting ordered set of $\boldsymbol{p}$'s are the **principal components**.

PCA Eigenvector Decomposition

- Find some orthonormal matrix $\boldsymbol{P}$ in $\boldsymbol{Y} = \boldsymbol{PX}$ such that $\boldsymbol{C}_Y$ is a diagonal matrix.
- The m-dimensional rows of $\boldsymbol{P}$ are the principal components of $\boldsymbol{X}$.
- Relation between $\boldsymbol{C}_Y$ and $\boldsymbol{C}_X$:

$$\begin{aligned}\boldsymbol{C}_Y &= \frac{1}{n-1}\boldsymbol{YY}^\intercal = \frac{1}{n-1}\boldsymbol{PX}(\boldsymbol{PX})^\intercal \\ &= \frac{1}{n-1}\boldsymbol{PXX}^\intercal\boldsymbol{P}^\intercal = \boldsymbol{P}\left(\frac{1}{n-1}\boldsymbol{XX}^\intercal\right)\boldsymbol{P}^\intercal \\ &= \boldsymbol{PC}_X\boldsymbol{P}^\intercal.\end{aligned}$$

Orthogonal

Definition 1

A $m \times n$ matrix $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_n \end{bmatrix}$ is said to be orthogonal

$$\boldsymbol{a}_i^\intercal \boldsymbol{a}_j = \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{otherwise} \end{cases},$$

where each $\boldsymbol{a}_i$, $i = 1, 2, \ldots, n$ is a column vector of m rows.

Theorem 1

The inverse of an orthogonal matrix is its transpose.

Proof

Note that the $\left(\boldsymbol{A}^\intercal \boldsymbol{A}\right)_{ij} = \boldsymbol{a}_i^\intercal \boldsymbol{a}_j = \delta_{ij}$, where δ_{ij} is Kronecker's delta function. Since $\boldsymbol{A}^\intercal \boldsymbol{A} = \boldsymbol{I}$, it follows that $\boldsymbol{A}^{-1} = \boldsymbol{A}^\intercal$.

Symmetric Matrix

Definition 2

A $m \times m$ square matrix $\boldsymbol{A}$ is said to be symmetric if $A_{ij} = A_{ji}$, i.e., row index and column index are interchangeable: $\boldsymbol{A}^\intercal = \boldsymbol{A}$.

Theorem 2

For any $m \times n$ matrix $\boldsymbol{A}$, $\boldsymbol{A}^\intercal \boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^\intercal$ are symmetric.

Proof

$$\left(\boldsymbol{A}\boldsymbol{A}^\intercal\right)^\intercal = \boldsymbol{A}^{\intercal\intercal}\boldsymbol{A}^\intercal = \boldsymbol{A}\boldsymbol{A}^\intercal$$

$$\left(\boldsymbol{A}^\intercal\boldsymbol{A}\right)^\intercal = \boldsymbol{A}^\intercal\boldsymbol{A}^{\intercal\intercal} = \boldsymbol{A}^\intercal\boldsymbol{A}$$

Symmetric and Orthogonally Diagonalizable

Definition 3

A matrix $\boldsymbol{A}$ is said to be diagonalizable if there exists some $\boldsymbol{E}$ such that $\boldsymbol{A} = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^{\intercal}$, where $\boldsymbol{D}$ is a diagonal matrix and $\boldsymbol{E}$ is some special matrix that diagonalizes $\boldsymbol{A}$. Additionally, if $\boldsymbol{E}$ is orthogonal, then $\boldsymbol{A}$ is said to be orthogonally diagonalizable.

Theorem 3

A matrix is symmetric if it is orthogonally diagonalizable.

Proof

Suppose $\boldsymbol{A}$ is orthogonally diagonalizable. Let us compute $\boldsymbol{A}^{\intercal}$.

$$\boldsymbol{A}^{\intercal} = \left(\boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^{\intercal}\right)^{\intercal} = \boldsymbol{E}^{\intercal\intercal}\boldsymbol{D}^{\intercal}\boldsymbol{E}^{\intercal} = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^{\intercal} = \boldsymbol{A}.$$

Hence, if $\boldsymbol{A}$ is orthogonally diagonalizable, it must also be symmetric.

Orthonormal Eigenvectors

Let $\boldsymbol{A}$ be a square $n \times n$ symmetric matrix with associated eigenvectors $\{\boldsymbol{e}_i\}_{i=1}^n$. Let $\boldsymbol{E} = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \cdots & \boldsymbol{e}_n \end{bmatrix}$, and $D_{ij} = \lambda_i \delta_{ij}$

Theorem 4

A symmetric matrix $\boldsymbol{A}$ is diagonalized by a matrix of its orthonormal eigenvectors.

Proof (Part 1)

This theorem asserts that there exists a diagonal matrix $\boldsymbol{D}$ such that $\boldsymbol{A} = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^\intercal$.

Let $\boldsymbol{A}$ be any matrix, not necessarily symmetric, and let it have independent eigenvectors $\boldsymbol{e}_i$ (i.e. no degeneracy).

$$\boldsymbol{A}\boldsymbol{E} = \begin{bmatrix} \boldsymbol{A}\boldsymbol{e}_1 & \boldsymbol{A}\boldsymbol{e}_2 & \cdots & \boldsymbol{A}\boldsymbol{e}_n \end{bmatrix} = \begin{bmatrix} \lambda_1\boldsymbol{e}_1 & \lambda_2\boldsymbol{e}_2 & \cdots & \lambda_n\boldsymbol{e}_n \end{bmatrix} = \boldsymbol{E}\boldsymbol{D}.$$

Since $\boldsymbol{A}\boldsymbol{E} = \boldsymbol{E}\boldsymbol{D}$, it follows that $\boldsymbol{A} = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^{-1}$.

Orthonormal Eigenvectors (Cont'd)

Proof (Part 2)

We will show that a symmetric matrix $\boldsymbol{A}$ always has orthogonal eigenvectors. For any i and j such that $i \neq j$,

$$\lambda_i \boldsymbol{e}_i^\intercal \boldsymbol{e}_j = (\boldsymbol{A}\boldsymbol{e}_i)^\intercal \mathbf{e_j} = \boldsymbol{e}_i^\intercal \boldsymbol{A}^\intercal \boldsymbol{e}_j = \boldsymbol{e}_i^\intercal \boldsymbol{A}\boldsymbol{e}_j = \boldsymbol{e}_i^\intercal (\lambda_j \boldsymbol{e}_j)$$

$$\implies \lambda_i \boldsymbol{e}_i^\intercal \boldsymbol{e}_j = \lambda_j \boldsymbol{e}_i^\intercal \boldsymbol{e}_j$$

$$\implies (\lambda_i - \lambda_j)\boldsymbol{e}_i^\intercal \boldsymbol{e}_j = 0$$

Since we have assumed that the eigenvalues are unique, it must be that $\boldsymbol{e}_i^\intercal \boldsymbol{e}_j = 0$. Therefore, the eigenvectors of a symmetric matrix are orthogonal. By theorem 1 that $\boldsymbol{E}^\intercal = \boldsymbol{E}^{-1}$ and we can rewrite the final result

$$\boldsymbol{A} = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^\intercal$$

Orthogonal Basis

Theorem 5

For any $m \times n$ matrix $\boldsymbol{X}$, the symmetric matrix $\boldsymbol{X}^\intercal \boldsymbol{X}$ has a set of orthonormal eigenvectors $\{\hat{\boldsymbol{v}}_1, \hat{\boldsymbol{v}}_2, \ldots, \hat{\boldsymbol{v}}_n\}$ and a set of associated eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$. The set of vectors $\{\boldsymbol{X}\hat{\boldsymbol{v}}_1, \boldsymbol{X}\hat{\boldsymbol{v}}_2, \ldots, \boldsymbol{X}\hat{\boldsymbol{v}}_n\}$ then form an orthogonal basis, where each vector $\boldsymbol{X}\hat{\boldsymbol{v}}_i$ is of length $\sqrt{\lambda_i}$.

Proof

$$\begin{aligned}(\boldsymbol{X}\hat{\boldsymbol{v}}_i)^\intercal(\boldsymbol{X}\hat{\boldsymbol{v}}_j) &= \hat{\boldsymbol{v}}_i^\intercal \boldsymbol{X}^\intercal \boldsymbol{X}\hat{\boldsymbol{v}}_j = \hat{\boldsymbol{v}}_i^\intercal(\lambda_j\hat{\boldsymbol{v}}_j) = \lambda_j\hat{\boldsymbol{v}}_i^\intercal\hat{\boldsymbol{v}}_j \\ &= \lambda_j\delta_{ij}\end{aligned}$$

The squared length of each vector is

$$\|\boldsymbol{X}\hat{\boldsymbol{v}}_i\|^2 = (\boldsymbol{X}\hat{\boldsymbol{v}}_i)^\intercal(\boldsymbol{X}\hat{\boldsymbol{v}}_i) = \lambda_i.$$

Application to PCA

- For a symmetric matrix $\boldsymbol{A}$, Theorem 4 provides $\boldsymbol{A} = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^\intercal$, where $\boldsymbol{D}$ is a diagonal matrix and $\boldsymbol{E}$ is a matrix of eigenvectors of $\boldsymbol{A}$ arranged as columns.
- We select the matrix $\boldsymbol{P}$ to be a matrix where each row $\boldsymbol{p}_i$ is an eigenvector of $\frac{1}{n-1}\boldsymbol{X}\boldsymbol{X}^\intercal$, i.e., $\boldsymbol{P} = \boldsymbol{E}^\intercal$.
- With this relation and Theorem 1 ($\boldsymbol{P}^{-1} = \boldsymbol{P}^\intercal$),

$$\begin{aligned}\boldsymbol{C}_Y &= \boldsymbol{P}\boldsymbol{C}_X\boldsymbol{P}^\intercal = \boldsymbol{P}\left(\boldsymbol{E}^\intercal\boldsymbol{D}\boldsymbol{E}\right)\boldsymbol{P}^\intercal = \boldsymbol{P}\left(\boldsymbol{P}^\intercal\boldsymbol{D}\boldsymbol{P}\right)\boldsymbol{P}^\intercal \\ &= (\boldsymbol{P}\boldsymbol{P}^\intercal)\boldsymbol{D}\left(\boldsymbol{P}\boldsymbol{P}^\intercal\right) = \left(\boldsymbol{P}\boldsymbol{P}^{-1}\right)\boldsymbol{D}\left(\boldsymbol{P}\boldsymbol{P}^{-1}\right) \\ &= \boldsymbol{D}\end{aligned}$$

- It is evident that the choice of $\boldsymbol{P}$ diagonalizes $\boldsymbol{C}_X$, which is the goal for PCA.

Summary

- The principal components of $\boldsymbol{X}$ are the eigenvectors of $\boldsymbol{C}_X = \frac{1}{n-1}\boldsymbol{X}\boldsymbol{X}^\intercal$.
- The i-th diagonal value of $\boldsymbol{C}_Y$ is the variance of $\boldsymbol{X}$ along $\boldsymbol{p}_i$.
- Note that $\boldsymbol{C}_X$ is symmetric, and the same analysis works for $\boldsymbol{C}_X = \boldsymbol{C}_X^\intercal = \frac{1}{n-1}\boldsymbol{X}^\intercal\boldsymbol{X}$.

Framework for Singular Vector Decomposition

- Let $\boldsymbol{X}$ be an arbitrary $n \times m$ matrix and $\boldsymbol{X}^\intercal \boldsymbol{X}$ be a rank r, square, and symmetric $m \times m$ matrix.
- $\{\hat{\boldsymbol{v}}_i\}_{i=1}^r$ is the set of orthonormal $m \times 1$ eigenvectors with associated eigenvalues $\{\lambda_i\}_{i=1}^n$ for the symmetric matrix $\boldsymbol{X}^\intercal \boldsymbol{X}$

$$\boldsymbol{X}^\intercal \boldsymbol{X} \hat{\boldsymbol{v}}_i = \lambda_i \hat{\boldsymbol{v}}_i$$

- $\sigma_i := \sqrt{\lambda_i}$ is positive and real. It is called the **singular values**.
- $\{\hat{\boldsymbol{u}}_i\}_{i=1}^r$ is the set of $n \times 1$ vectors defined by

$$\hat{\boldsymbol{u}}_i := \frac{1}{\sigma_i} \boldsymbol{X} \hat{\boldsymbol{v}}_i$$

- By Theorem 5, we have $\hat{\boldsymbol{u}}_i^\intercal \hat{\boldsymbol{u}}_j = \delta_{ij}$ and $\|\boldsymbol{X} \hat{v}_i\| = \sigma_i$

Singular Value Decomposition

- The scalar version of singular value decomposition is just a restatement of the definition

$$\boldsymbol{X}\hat{\boldsymbol{v}}_i = \sigma_i \hat{\boldsymbol{u}}_i$$

- Now, eigenvectors $\{\hat{\boldsymbol{v}}_i\}_{i=1}^r$ and vectors $\{\hat{\boldsymbol{u}}_i\}_{i=1}^r$ are both orthonormal sets or bases in r-dimensional space.
- For accompanying orthogonal matrices,

$$\boldsymbol{V} = [\hat{\boldsymbol{v}}_{\tilde{1}} \ \hat{\boldsymbol{v}}_{\tilde{2}} \ \cdots \ \hat{\boldsymbol{v}}_{\tilde{m}}] \qquad \boldsymbol{U} = [\hat{\boldsymbol{u}}_{\tilde{1}} \ \hat{\boldsymbol{u}}_{\tilde{2}} \ \cdots \ \hat{\boldsymbol{u}}_{\tilde{n}}]$$

  where we have appended an additional $(m-r)$ and $(n-r)$ orthonormal vectors to "fill up" the matrices for $\boldsymbol{V}$ and $\boldsymbol{U}$ respectively (i.e. to deal with degeneracy issues).

- **Singular Value Decomposition**: $\boldsymbol{XV} = \boldsymbol{U\Sigma}$

Final Form of SVD

- Because $\boldsymbol{V}$ is orthogonal, we can multiply both sides by $\boldsymbol{V}^{-1} = \boldsymbol{V}^{\intercal}$ to arrive at the final form of the decomposition.

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\intercal}$$

- Any arbitrary matrix $\boldsymbol{X}$ can be decomposed to an orthogonal matrix, a diagonal matrix and another orthogonal matrix.

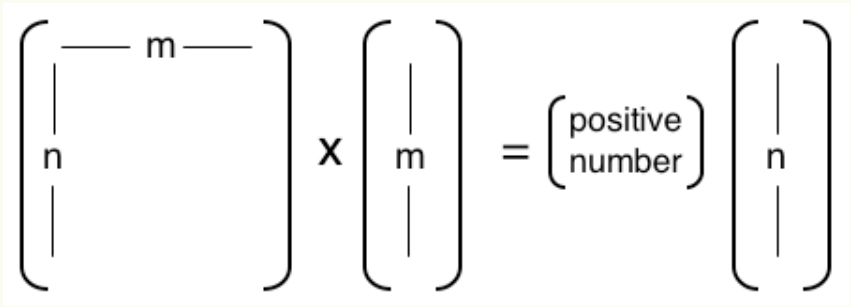
- Any matrix performs a rotation, a stretch, and another rotation!

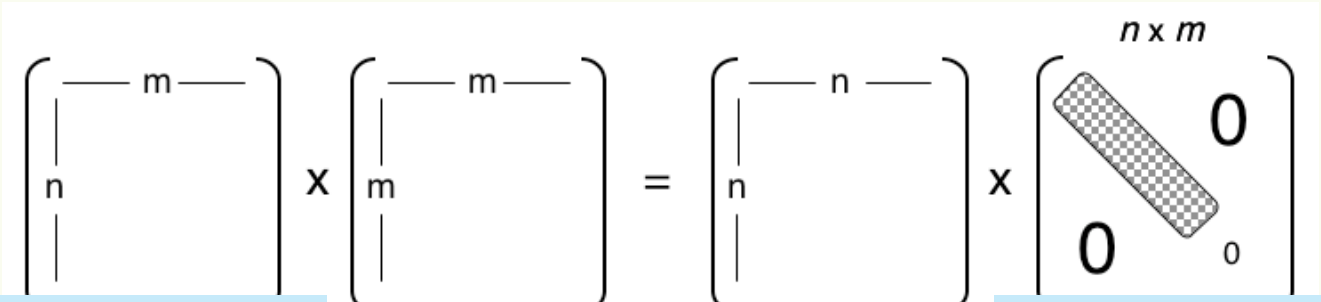
Singular Value Decomposition Illustrated

A new diagonal matrix $\Sigma :=$

$$\begin{bmatrix} \sigma_{\tilde{1}} & & & & \\ & \ddots & & 0 & \\ & & \sigma_{\tilde{r}} & & \\ & & & 0 & \\ & 0 & & & \ddots \\ & & & & & 0 \end{bmatrix}$$

where $\sigma_{\tilde{1}} \geq \sigma_{\tilde{2}} \geq \ldots \geq \sigma_{\tilde{r}}$ are the rank-ordered set of singular values.





Interpreting SVD

- Let $Z := \Sigma V^\intercal$. Then

$$X = U\Sigma V^\intercal \implies U^\intercal X = \Sigma V^\intercal \implies U^\intercal X = Z$$

- $U^\intercal$ performs a change of basis from X to Z in the column space.
- On the other hand, in the row space, define $W := U^\intercal \Sigma$, and

$$(XV)^\intercal = \left(\Sigma U\right)^\intercal \implies V^\intercal X^\intercal = V^\intercal \Sigma \implies V^\intercal X^\intercal = W.$$

SVD and PCA

- Given a $m \times n$ data matrix $\boldsymbol{X}$, define a new $n \times m$ matrix $\boldsymbol{Y}$ as

$$\boldsymbol{Y} := \frac{1}{\sqrt{n-1}} \boldsymbol{X}^\intercal,$$

  where each column of $\boldsymbol{Y}$ has zero mean.

- Properties

$$\boldsymbol{Y}^\intercal \boldsymbol{Y} = \left(\frac{1}{\sqrt{n-1}} \boldsymbol{X}^\intercal \right)^\intercal \left(\frac{1}{\sqrt{n-1}} \boldsymbol{X}^\intercal \right) = \frac{1}{n-1} \boldsymbol{X} \boldsymbol{X}^\intercal = \boldsymbol{C}_X$$

- If we calculate the SVD of $\boldsymbol{Y}$, the columns of matrix $\boldsymbol{V}$ contain the eigenvectors of $\boldsymbol{Y}^\intercal \boldsymbol{Y} = \boldsymbol{C}_X$. Therefore, the columns of $\boldsymbol{V}$ are the principal components of $\boldsymbol{X}$.

- Finding the principal components amounts to finding an orthonormal basis that spans the column space of $\boldsymbol{X}$.

Quick Algorithm of PCA by SVD

1. Organize data as an $m \times n$ matrix, where m is the number of measurement types and n is the number of samples.
2. Subtract off the mean for each measurement type.
3. Calculate the SVD or the eigenvectors of the covariance.

Daily Treasury Yield Curve Rates

- Commonly referred to as "**Constant Maturity Treasury**" rates
- The yield curve, which relates the yield on a security to its time to maturity, is based on the closing market bid yields on actively traded Treasury securities in the over-the-counter market.
- These market yields are calculated from composites of quotations obtained by the Federal Reserve Bank of New York.
- The Treasury yield curve is estimated daily using a cubic spline model. Inputs to the model are primarily bid-side yields for on-the-run Treasury securities.
- Sample period: From 9 Feb, 2006, all maturities from 1 month to 30 years have data.

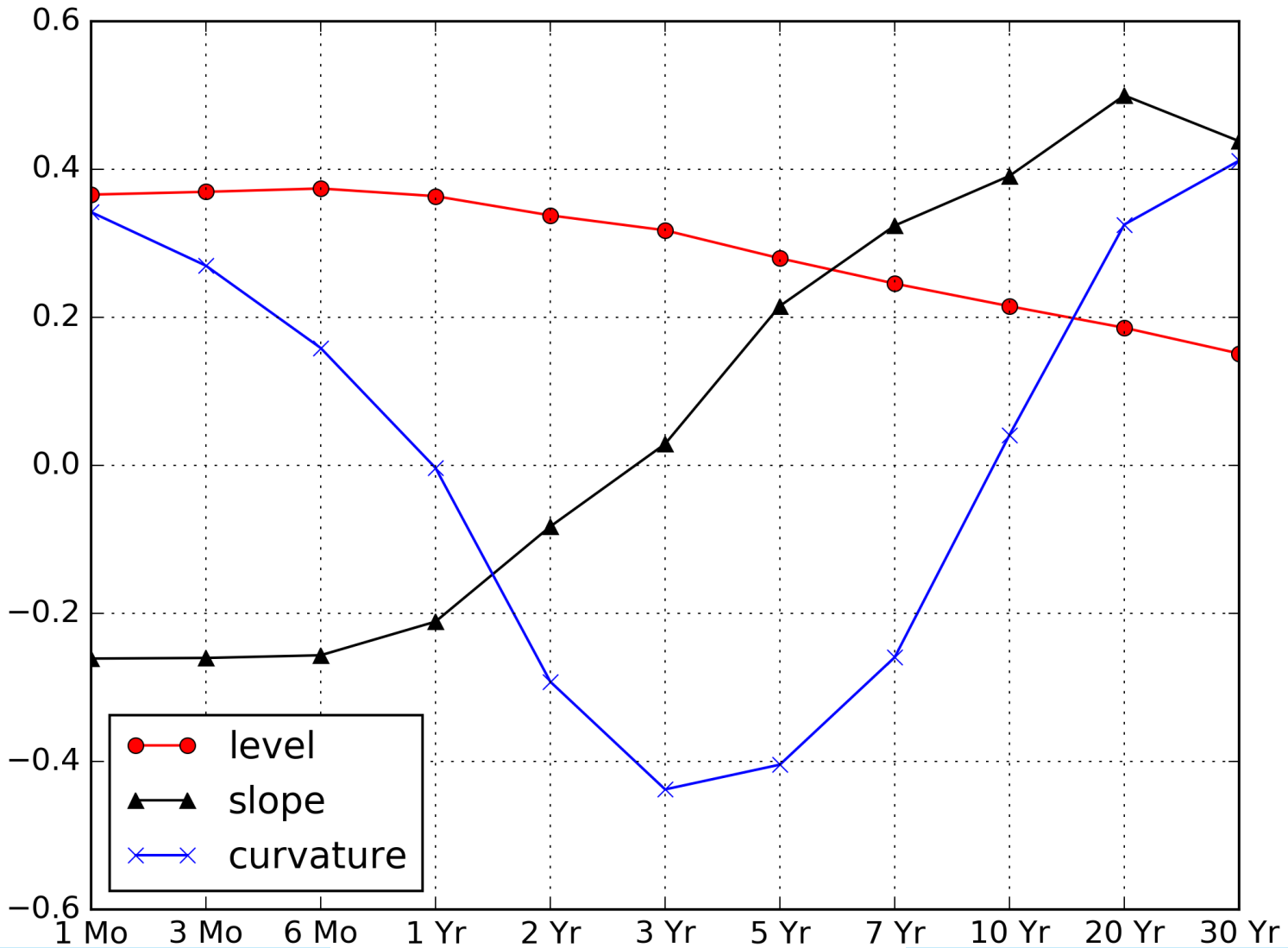
PCA of Yield Curve

- **Yield Curve Level**
  The first principal component can explain 92.49% of the variation in the data set.

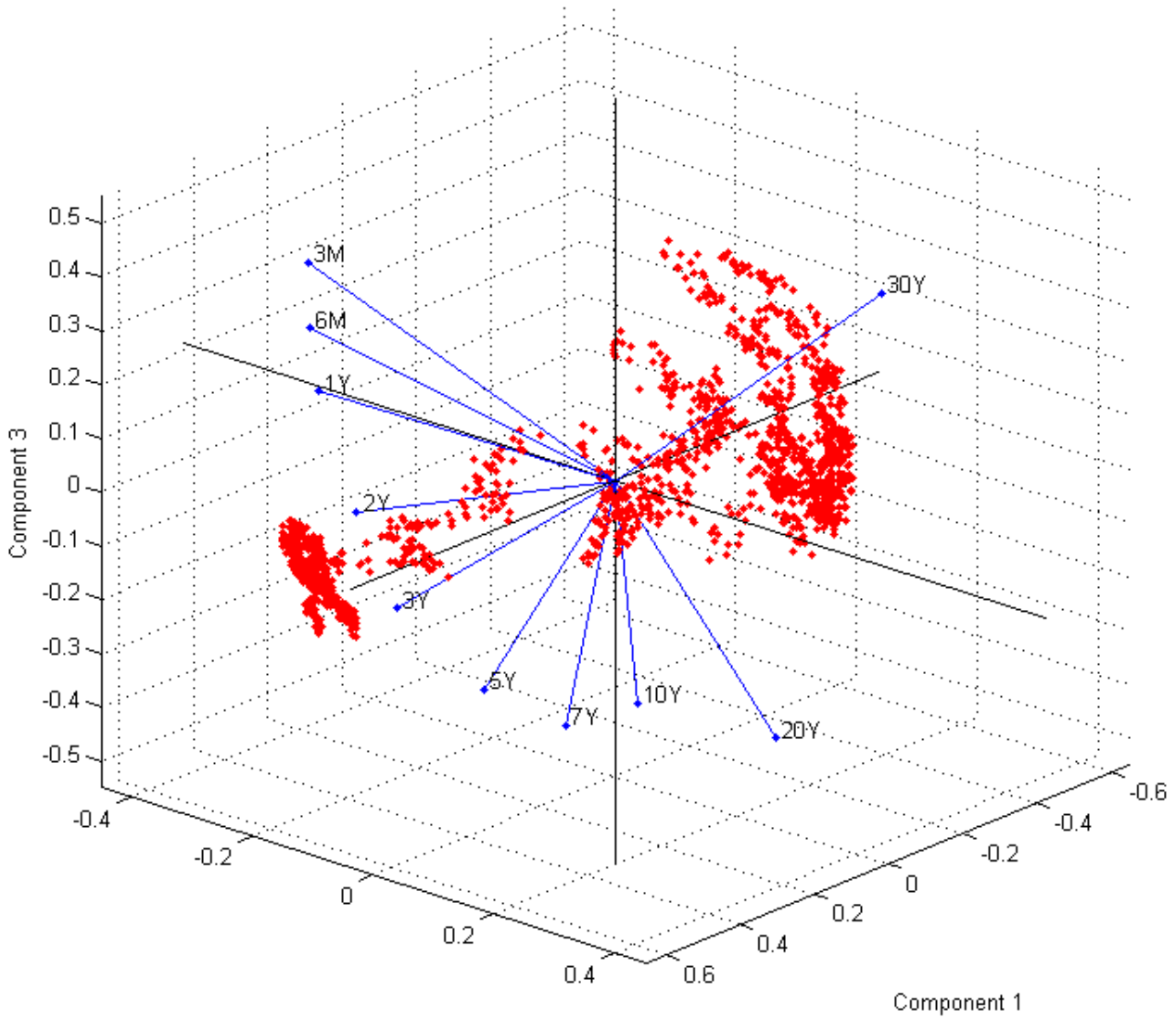
- **Yield Curve Slope**
  The first and second components can explain 99.07% of the variation.

- **Yield Curve Curvature**
  The first three components explain 99.79% of the variation.

Three Principal Components of Yield Curve



Mapping of Principal Components



Takeaways

- PCA is a linear algorithm for finding the dimensions that capture most information in data.
- Information in the PCA framework is measured by variance.

$$\mathbb{V}\left(\boldsymbol{e}_i^\intercal \boldsymbol{x}\right) = \lambda_i, \qquad i = 1, 2, \ldots, m.$$

- The eigenvectors in PCA are orthonormal, and form an uncorrelated basis.
- PCA show that U.S. Treasury Yield Curve can be explained by three “factors”: level, slope, and curvature.

Additional Reading

1 **A Tutorial on Principal Component Analysis**, Jonathon Shlens, Google Research, Working Paper (2014)