

# Introduction to Machine Learning

## Principal Component Analysis

Varun Chandola

April 20, 2020

### Outline

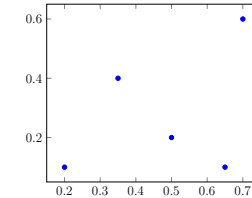
### Contents

1	Principal Components Analysis	1
1.1	Introduction to PCA	1
1.2	Principle of Maximal Variance	2
1.3	Defining Principal Components	3
1.4	Dimensionality Reduction Using PCA	4
1.5	PCA Algorithm	4
1.6	Recovering Original Data	4
1.7	Eigen Faces	5

## 1 Principal Components Analysis

### 1.1 Introduction to PCA

- Consider the following data points
- *Embed* these points in 1 dimension
- What is the best way?
  - **Along the direction of the maximum variance**
  - Why?



### 1.2 Principle of Maximal Variance

- Least loss of information
- Best capture the “spread”
- What is the direction of maximal variance?
- Given any direction ( $\hat{\mathbf{u}}$ ), the projection of  $\mathbf{x}$  on  $\hat{\mathbf{u}}$  is given by:

$$\mathbf{x}_i^\top \hat{\mathbf{u}}$$

- Direction of maximal variance can be obtained by maximizing

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i^\top \hat{\mathbf{u}})^2 &= \frac{1}{N} \sum_{i=1}^N \hat{\mathbf{u}}^\top \mathbf{x}_i \mathbf{x}_i^\top \hat{\mathbf{u}} \\ &= \hat{\mathbf{u}}^\top \left( \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^\top \right) \hat{\mathbf{u}} \end{aligned}$$

### Finding Direction of Maximal Variance

- Find:

$$\max_{\hat{\mathbf{u}}: \hat{\mathbf{u}}^\top \hat{\mathbf{u}} = 1} \hat{\mathbf{u}}^\top \mathbf{S} \hat{\mathbf{u}}$$

where:

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^\top$$

- $\mathbf{S}$  is the sample (empirical) covariance matrix of the mean-centered data

The solution to the above constrained optimization problem may be obtained using the Lagrange multipliers method. We maximize the following w.r.t.  $\hat{\mathbf{u}}$ :

$$(\hat{\mathbf{u}}^\top \mathbf{S} \hat{\mathbf{u}}) - \lambda(\hat{\mathbf{u}}^\top \hat{\mathbf{u}} - 1)$$

to get:

$$\begin{aligned} \frac{d}{d\hat{\mathbf{u}}}(\hat{\mathbf{u}}^\top \mathbf{S} \hat{\mathbf{u}}) - \lambda(\hat{\mathbf{u}}^\top \hat{\mathbf{u}} - 1) &= 0 \\ \mathbf{S} \hat{\mathbf{u}} - \lambda \hat{\mathbf{u}} &= 0 \\ \mathbf{S} \hat{\mathbf{u}} &= \lambda \hat{\mathbf{u}} \end{aligned}$$

Obviously, the solution to the above equation is an eigen vector of the matrix  $\mathbf{S}$ . But which  $\mathbf{S}$ ? Note that for the optimal solution:

$$\hat{\mathbf{u}}^\top \mathbf{S} \hat{\mathbf{u}} = (\hat{\mathbf{u}}^\top \lambda \hat{\mathbf{u}}) = \lambda$$

Thus we should choose the largest possible  $\lambda$  which means that the first solution is the eigen vector of  $\mathbf{S}$  with largest eigen value.

### 1.3 Defining Principal Components

- First PC: Eigen-vector of the (sample) covariance matrix with largest eigen-value
- Second PC?
- Eigen-vector with next largest value
- Variance of each PC is given by  $\lambda_i$
- Variance captured by first  $L$  PC ( $1 \leq L \leq D$ )

$$\frac{\sum_{i=1}^L \lambda_i}{\sum_{i=1}^D \lambda_i} \times 100$$

- What are eigen vectors and values?

$$\mathbf{A} \mathbf{v} = \lambda \mathbf{v}$$

$\mathbf{v}$  is eigen vector and  $\lambda$  is eigen-value for the **square matrix**  $\mathbf{A}$

- Geometric interpretation?

For the second PC we need to optimize the variance with a second constraint that the solution should be orthogonal to the first PC. It is easy to show that this solution will be the PC with second largest eigen value, and so on.

### 1.4 Dimensionality Reduction Using PCA

- Consider first  $L$  eigen values and eigen vectors
- Let  $\mathbf{W}$  denote the  $D \times L$  matrix with first  $L$  eigen vectors in the columns (sorted by  $\lambda$ 's)
- PC score matrix

$$\mathbf{Z} = \mathbf{XW}$$

- Each input vector ( $D \times 1$ ) is replaced by a shorter  $L \times 1$  vector

### 1.5 PCA Algorithm

1. *Center*  $\mathbf{X}$

$$\mathbf{X} = \mathbf{X} - \hat{\boldsymbol{\mu}}$$

2. Compute sample covariance matrix:

$$\mathbf{S} = \frac{1}{N-1} \mathbf{X}^\top \mathbf{X}$$

3. Find eigen vectors and eigen values for  $\mathbf{S}$
4.  $\mathbf{W}$  consists of first  $L$  eigen vectors as columns

- Ordered by decreasing eigen-values
- $\mathbf{W}$  is  $D \times L$

5. Let  $\mathbf{Z} = \mathbf{XW}$

6. Each row in  $\mathbf{Z}$  (or  $\mathbf{z}_i^\top$ ) is the lower dimensional embedding of  $\mathbf{x}_i$

### 1.6 Recovering Original Data

- Using  $\mathbf{W}$  and  $\mathbf{z}_i$

$$\hat{\mathbf{x}}_i = \mathbf{W} \mathbf{z}_i$$

- **Average Reconstruction Error**

$$J(\mathbf{W}, \mathbf{Z}) = \frac{1}{N} \sum_{i=1}^N \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2$$

**Theorem 1** (Classical PCA Theorem). *Among all possible orthonormal sets of  $L$  basis vectors, PCA gives the solution which has the minimum reconstruction error.*

- Optimal “embedding” in  $L$  dimensional space is given by  $z_i = \mathbf{W}^\top \mathbf{x}_i$

## 1.7 Eigen Faces

**EigenFaces** [1]

- **Input:** A set of images (of faces)
- **Task:** Identify if a new image is a face or not.

## References

## References

- [1] M. Turk and A. Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on*, pages 586–591, Jun 1991.