

# Machine Learning

## Lecture 18: Principal Component Analysis (PCA)

Kia Nazarpour

# Context

Digital technologies, machine learning and AI are revolutionising the fields of medicine, research and public health.

1. How can we make sense from this data?
2. Is it all useful data?
3. How can we make the data smaller, without losing much information?



1 PB = 1,000 TB

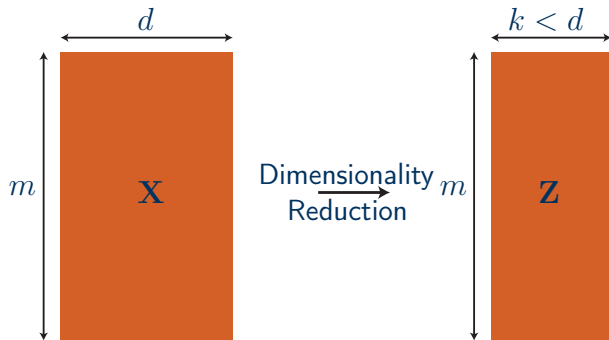
# Learning Outcomes

1. Learn about the key motivation behind the use of the PCA method
2. Understand the geometrical explanation of the PCA method
3. Explain steps in one of the derivations of the PCA method
4. Apply the PCA method on a real dataset

## References:

1. James *et al.*, *An Introduction to Statistical Learning*, Springer, 2013. (Sections 6.3, 6.7, and 10.2)
2. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2008. (Section 12.1)

# Dimensionality Reduction



# Applications and Considerations

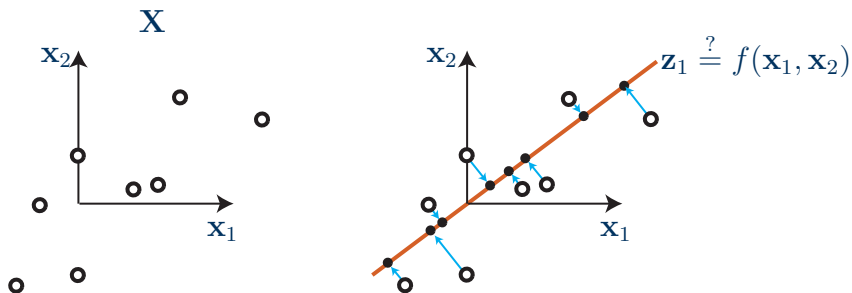
## Applications of the PCA method (and many other dimensionality reduction methods)

1. Visualisation
2. Exploration
3. Compression

## Key considerations:

1. Reducing the number of columns ( $d \rightarrow k$ ) by deletion is not meaningful.
2. Columns of  $\mathbf{Z}$  are uncorrelated, *i.e.* minimal redundancy.
3. It is OK to make our variables *less interpretable!*

# Principal Component Analysis



## Notes

1. We are interested in finding projections of data points that are as similar to the original data points as possible, but which have a significantly lower intrinsic dimensionality.
2. Without loss of generality, we assume that the mean of data is zero.

# Principal Component Analysis

$$\mathbf{X}_{m \times d} \xrightarrow{PCA} \mathbf{Z}_{m \times k} \quad (k < d)$$

$$\mathbf{Z}_{m \times k} = \mathbf{X}_{m \times d} \mathbf{U}_{d \times k}$$

$$\mathbf{Z} = [\mathbf{z}_1 \ \mathbf{z}_2 \ \cdots \ \mathbf{z}_k] \quad \mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_d] \quad \mathbf{U} = [\mathbf{u}_1 \ \mathbf{u}_2 \ \cdots \ \mathbf{u}_k]$$

$$\mathbf{z}_1 = \mathbf{X} \mathbf{u}_1$$

## Remarks

1. Principal components are a sequence of projections of the data, mutually uncorrelated and ordered in variance.
2. The columns  $\mathbf{u}_{1 \dots k}$  of  $\mathbf{U}$  are orthonormal, so that  $\mathbf{u}_i^T \mathbf{u}_j = 0$  if and only if  $i \neq j$  and  $\mathbf{u}_i^T \mathbf{u}_i = 1$ .

# Key Different Perspectives to PCA

## Three key approaches to PCA:

1. **Maximum variance formulation** (Hotelling 1933)
2. Minimum error formulation (Pearson 1901)
3. Probabilistic formulation (Tipping & Bishop 1997)



## Maximum Variance Formulation

$$\mathbf{X}_{m \times d} \xrightarrow{PCA} \mathbf{Z}_{m \times k} \quad (k < d)$$

$$\mathbf{Z}_{m \times k} = \mathbf{X}_{m \times d} \mathbf{U}_{d \times k}$$

$$\mathbf{Z} = [\mathbf{z}_1 \ \mathbf{z}_2 \ \cdots \ \mathbf{z}_k] \quad \mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_d] \quad \mathbf{U} = [\mathbf{u}_1 \ \mathbf{u}_2 \ \cdots \ \mathbf{u}_k]$$

$$\max_{\mathbf{u}_1} \text{Var}[\mathbf{z}_1] = \max_{\mathbf{u}_1} \text{Var}[\mathbf{X}\mathbf{u}_1]$$

## Maximum Variance Formulation

$$\mathbf{X}_{m \times d} \xrightarrow{PCA} \mathbf{Z}_{m \times k} \quad (k < d)$$

$$\mathbf{Z}_{m \times k} = \mathbf{X}_{m \times d} \mathbf{U}_{d \times k}$$

$$\mathbf{Z} = [\mathbf{z}_1 \ \mathbf{z}_2 \ \cdots \ \mathbf{z}_k] \quad \mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_d] \quad \mathbf{U} = [\mathbf{u}_1 \ \mathbf{u}_2 \ \cdots \ \mathbf{u}_k]$$

$$\max_{\mathbf{u}_1} \text{Var}[\mathbf{z}_1] = \max_{\mathbf{u}_1} \text{Var}[\mathbf{X}\mathbf{u}_1]$$

$$\max_{\mathbf{u}_1} \mathbf{z}_1^T \mathbf{z}_1 = \max_{\mathbf{u}_1} \mathbf{u}_1^T \mathbf{X}^T \mathbf{X} \mathbf{u}_1$$

$$= \max_{\mathbf{u}_1} \mathbf{u}_1^T \Sigma_{\mathbf{X}} \mathbf{u}_1 \quad \Sigma_{\mathbf{X}} = \mathbf{X}^T \mathbf{X}: (\mathbf{N} \times \text{covariance of } \mathbf{X})$$

## Maximum Variance Formulation

$$\mathbf{X}_{m \times d} \xrightarrow{PCA} \mathbf{Z}_{m \times k} \quad (k < d)$$

$$\mathbf{Z}_{m \times k} = \mathbf{X}_{m \times d} \mathbf{U}_{d \times k}$$

$$\mathbf{Z} = [\mathbf{z}_1 \ \mathbf{z}_2 \ \cdots \ \mathbf{z}_k] \quad \mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_d] \quad \mathbf{U} = [\mathbf{u}_1 \ \mathbf{u}_2 \ \cdots \ \mathbf{u}_k]$$

$$\max_{\mathbf{u}_1} \text{Var}[\mathbf{z}_1] = \max_{\mathbf{u}_1} \text{Var}[\mathbf{X}\mathbf{u}_1]$$

$$\max_{\mathbf{u}_1} \mathbf{z}_1^T \mathbf{z}_1 = \max_{\mathbf{u}_1} \mathbf{u}_1^T \mathbf{X}^T \mathbf{X} \mathbf{u}_1$$

$$= \max_{\mathbf{u}_1} \mathbf{u}_1^T \Sigma_{\mathbf{X}} \mathbf{u}_1 \quad \Sigma_{\mathbf{X}} = \mathbf{X}^T \mathbf{X}: (\mathbf{N} \times \text{covariance of } \mathbf{X})$$

$$= \max_{\mathbf{u}_1} \mathbf{u}_1^T \Sigma_{\mathbf{X}} \mathbf{u}_1 \quad \text{s.t.} \quad \|\mathbf{u}_1\| = \mathbf{u}_1^T \mathbf{u}_1 = 1$$

## Maximum Variance Formulation

Using the Lagrange multipliers method:

$$L(\mathbf{u}_1, \lambda_1) = \mathbf{u}_1^T \Sigma_{\mathbf{X}} \mathbf{u}_1 - \lambda_1 (\mathbf{u}_1^T \mathbf{u}_1 - 1)$$

$$\frac{\partial L}{\partial \mathbf{u}_1} = 2\Sigma_{\mathbf{X}} \mathbf{u}_1 - 2\lambda_1 \mathbf{u}_1 = 0$$

## Maximum Variance Formulation

Using the Lagrange multipliers method:

$$L(\mathbf{u}_1, \lambda_1) = \mathbf{u}_1^T \Sigma_{\mathbf{X}} \mathbf{u}_1 - \lambda_1 (\mathbf{u}_1^T \mathbf{u}_1 - 1)$$

$$\frac{\partial L}{\partial \mathbf{u}_1} = 2\Sigma_{\mathbf{X}} \mathbf{u}_1 - 2\lambda_1 \mathbf{u}_1 = 0$$

$$\Sigma_{\mathbf{X}} \mathbf{u}_1 = \lambda_1 \mathbf{u}_1 \quad \rightsquigarrow \quad \lambda_1 \text{ and } \mathbf{u}_1 \text{ are an eigenvalue-eigenvector pair of } \Sigma_{\mathbf{X}}$$

## Maximum Variance Formulation

Using the Lagrange multipliers method:

$$L(\mathbf{u}_1, \lambda_1) = \mathbf{u}_1^T \Sigma_{\mathbf{X}} \mathbf{u}_1 - \lambda_1 (\mathbf{u}_1^T \mathbf{u}_1 - 1)$$

$$\frac{\partial L}{\partial \mathbf{u}_1} = 2\Sigma_{\mathbf{X}} \mathbf{u}_1 - 2\lambda_1 \mathbf{u}_1 = 0$$

$\Sigma_{\mathbf{X}} \mathbf{u}_1 = \lambda_1 \mathbf{u}_1 \quad \rightsquigarrow \quad \lambda_1$  and  $\mathbf{u}_1$  are an eigenvalue-eigenvector pair of  $\Sigma_{\mathbf{X}}$

$$\text{Var}[\mathbf{z}_1] = \mathbf{u}_1^T \Sigma_{\mathbf{X}} \mathbf{u}_1 = \mathbf{u}_1^T \lambda_1 \mathbf{u}_1 = \lambda_1 \underbrace{\mathbf{u}_1^T \mathbf{u}_1}_1 = \lambda_1$$

## Maximum Variance Formulation

Using Lagrange multipliers:

$$L(\mathbf{u}_1, \lambda_1) = \mathbf{u}_1^T \Sigma_{\mathbf{X}} \mathbf{u}_1 - \lambda_1 (\mathbf{u}_1^T \mathbf{u}_1 - 1)$$

$$\frac{\partial L}{\partial \mathbf{u}_1} = 2\Sigma_{\mathbf{X}} \mathbf{u}_1 - 2\lambda_1 \mathbf{u}_1 = 0$$

$\Sigma_{\mathbf{X}} \mathbf{u}_1 = \lambda_1 \mathbf{u}_1 \quad \rightsquigarrow \quad \lambda_1$  and  $\mathbf{u}_1$  are an eigenvalue-eigenvector pair of  $\Sigma_{\mathbf{X}}$

$$\text{Var}[\mathbf{z}_1] = \mathbf{u}_1^T \Sigma_{\mathbf{X}} \mathbf{u}_1 = \mathbf{u}_1^T \lambda_1 \mathbf{u}_1 = \lambda_1 \underbrace{\mathbf{u}_1^T \mathbf{u}_1}_1 = \lambda_1$$

For  $\Sigma_{\mathbf{X}}$  there are  $d$  eigenvalue-eigenvector pairs:

$$\begin{array}{ccccccc} e_1 & > & e_2 & > & e_3 & > & \cdots & > & e_d \\ \mathbf{v}_1 & & \mathbf{v}_2 & & \mathbf{v}_3 & & \cdots & & \mathbf{v}_d \end{array}$$

## Maximum Variance Formulation

$$\mathbf{X}_{m \times d} \xrightarrow{PCA} \mathbf{Z}_{m \times k} \quad (k < d)$$

$$\mathbf{Z}_{m \times k} = \mathbf{X}_{m \times d} \mathbf{U}_{d \times k}$$

$$\mathbf{Z} = [\mathbf{z}_1 \ \mathbf{z}_2 \ \cdots \ \mathbf{z}_k] \quad \mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_d] \quad \mathbf{U} = [\mathbf{u}_1 \ \mathbf{u}_2 \ \cdots \ \mathbf{u}_k]$$

The first principal direction  $\mathbf{u}_1$  must be the eigenvector of  $\Sigma_{\mathbf{X}}$  that corresponds to largest eigenvalue ( $e_1$ ).

$$\mathbf{z}_1 = \mathbf{X}\mathbf{u}_1 \quad \rightarrow \quad \mathbf{z}_1 = \mathbf{X}\mathbf{v}_1$$



## Maximum Variance Formulation

$$\mathbf{X}_{m \times d} \xrightarrow{PCA} \mathbf{Z}_{m \times k} \quad (k < d)$$

$$\mathbf{Z}_{m \times k} = \mathbf{X}_{m \times d} \mathbf{U}_{d \times k}$$

$$\mathbf{Z} = [\mathbf{z}_1 \ \mathbf{z}_2 \ \cdots \ \mathbf{z}_k] \quad \mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_d] \quad \mathbf{U} = [\mathbf{u}_1 \ \mathbf{u}_2 \ \cdots \ \mathbf{u}_k]$$

The first principal direction  $\mathbf{u}_1$  must be the eigenvector of  $\Sigma_{\mathbf{X}}$  that corresponds to largest eigenvalue ( $e_1$ ).

$$\mathbf{z}_1 = \mathbf{X}\mathbf{u}_1 \rightarrow \mathbf{z}_1 = \mathbf{X}\mathbf{v}_1$$

What about other principal components?  $\mathbf{z}_{2 \dots k} = \mathbf{X}\mathbf{u}_{2 \dots k} \stackrel{?}{=} \mathbf{X}\mathbf{v}_{2 \dots k}$

Each new principal direction  $\mathbf{u}_i$  should:

- maximise  $\text{Var}[\mathbf{z}_i]$ ;
- be orthogonal to all other  $\mathbf{u}_j$ ; extracting something new from  $\mathbf{X}$ .

## Maximum Variance Formulation

$$\mathbf{X}_{m \times d} \xrightarrow{PCA} \mathbf{Z}_{m \times k} \quad (k < d)$$

$$\mathbf{Z}_{m \times k} = \mathbf{X}_{m \times d} \mathbf{U}_{d \times k}$$

$$\mathbf{Z} = [\mathbf{z}_1 \ \mathbf{z}_2 \ \cdots \ \mathbf{z}_k] \quad \mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_d] \quad \mathbf{U} = [\mathbf{u}_1 \ \mathbf{u}_2 \ \cdots \ \mathbf{u}_k]$$

For  $\mathbf{z}_2$ :

$$\mathbf{z}_2 = \mathbf{X} \mathbf{u}_2$$

$$\max_{\mathbf{u}_2} \text{Var}[\mathbf{z}_2] = \max_{\mathbf{u}_2} \mathbf{u}_2^T \Sigma_{\mathbf{X}} \mathbf{u}_2$$

$$\text{s.t.} \quad \|\mathbf{u}_2\| = 1 \quad \& \quad \mathbf{u}_2^T \mathbf{u}_1 = 0$$

$$\mathbf{z}_2 = \mathbf{X} \mathbf{u}_2 \quad \rightarrow \quad \mathbf{z}_2 = \mathbf{X} \mathbf{v}_2$$

Because  $\mathbf{u}_2$  must be the eigenvector of  $\Sigma_{\mathbf{X}}$  that corresponds to second largest eigenvalue ( $e_2$ ).

## Summary - Maximum Variance Formulation

$$\mathbf{X}_{m \times d} \xrightarrow{PCA} \mathbf{Z}_{m \times k} \quad (k < d)$$

$$\mathbf{Z}_{m \times k} = \mathbf{X}_{m \times d} \mathbf{U}_{d \times k} = \mathbf{X}_{m \times d} \mathbf{V}_{d \times k}$$

$$\mathbf{Z} = [\mathbf{z}_1 \ \mathbf{z}_2 \ \cdots \ \mathbf{z}_k] \quad \mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_d] \quad \mathbf{V} = [\mathbf{v}_1 \ \mathbf{v}_2 \ \cdots \ \mathbf{v}_k]$$

where columns of  $\mathbf{V}_{d \times k}$  are the eigenvectors of  $\Sigma_{\mathbf{X}} = \mathbf{X}^T \mathbf{X}$ .

# An example - Public Health in Scotland

Source: Scottish Public Health Observatory (ScotPHO)

Region: All 32 Councils in Scotland

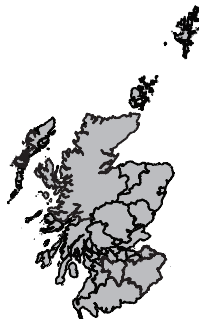
Year: 2019

Data: Six indicators were extracted

- 1) Active travel to school
- 2) Alcohol-related hospital admissions
- 3) Drug-related deaths
- 4) Attempted murder & serious assault
- 5) Drug crimes recorded
- 6) Smoking quit attempts

Labels: Employment deprivation level

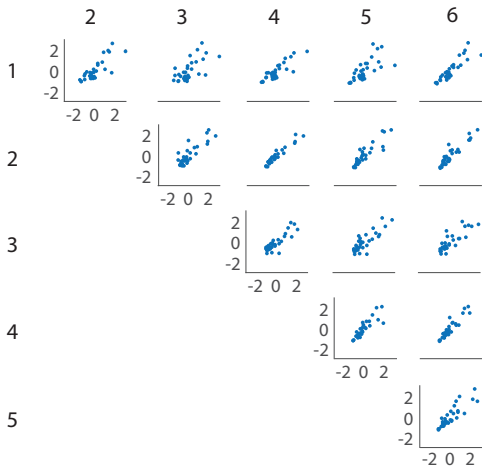
Low v.s. High



# An example - Public Health in Scotland

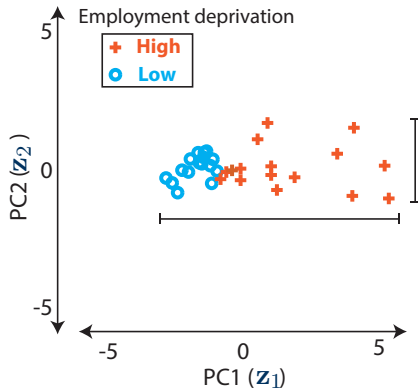
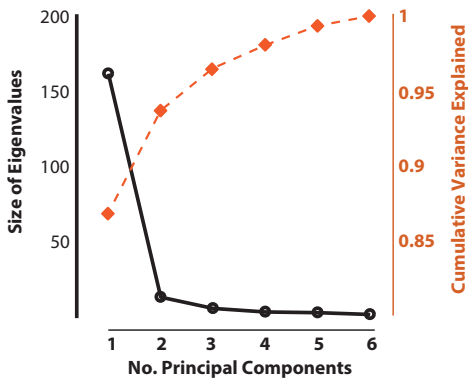
## Data Exploration

- 1) Active travel to school
- 2) Alcohol-related hospital admissions
- 3) Drug-related deaths
- 4) Attempted murder & serious assault
- 5) Drug crimes recorded
- 6) Smoking quit attempts



# An example - Public Health in Scotland

PCA Results:



Cumulative variance explained =  $\frac{\sum_{i=1}^k e_i}{\sum_{i=1}^d e_i}$   
where  $e_i$  is the  $i^{\text{th}}$  eigenvalue

## PCA - Bad Applications

1. Doing PCA to avoid overfitting is a bad idea. Instead use regularisation.
2. Doing PCA to for dimensionality reduction before classification is also a bad idea. Instead use a method called, linear discriminant analysis (LDA).

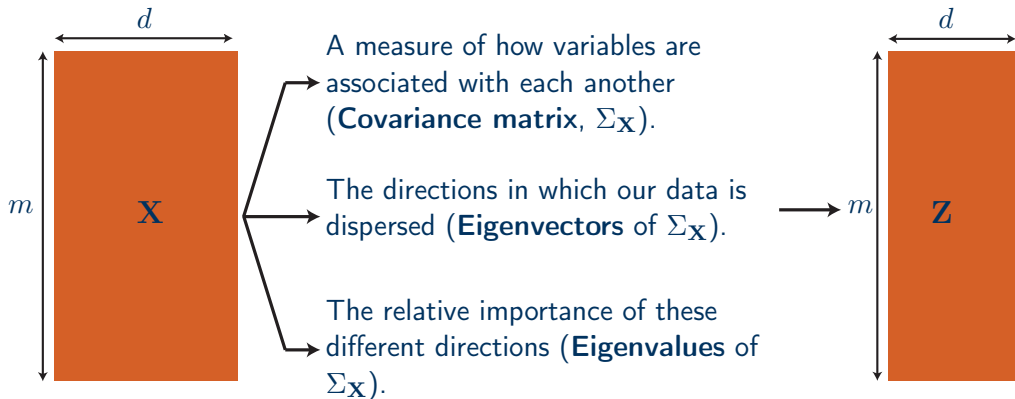
## PCA Implementation

There are three (potentially four) implementations for the PCA methods. For the centred design matrix  $\mathbf{X}_{m \times d}$  with the covariance matrix  $\Sigma_{\mathbf{X}} = \frac{1}{m} \mathbf{X}^T \mathbf{X}$

1. Eigenvector decomposition of  $\Sigma_{\mathbf{X}}$  - computational cost  $\mathcal{O}(d^3)$
2. Singular value decomposition of  $\Sigma_{\mathbf{X}}$  - computational cost  $\mathcal{O}(d^3)$
3. Singular value decomposition of  $\mathbf{X}$  - computational cost  $\mathcal{O}(md^2)$ 
  - Prove it as practice.
  - Start with the singular value decomposition of  $\mathbf{X}$ , that is  $\mathbf{X} = \mathbf{U} \Sigma_{\mathbf{X}} \mathbf{V}^T$
4. Eigenvector decomposition of Gram matrix  $K = \mathbf{X} \mathbf{X}^T$  - computational cost  $\mathcal{O}(d^3)$



## PCA - Summary



PCA linearly combines our variables and allows us to drop projections that are less informative.