# Introduction to Machine Learning
## NPFL 054

http://ufal.mff.cuni.cz/course/npfl054

Barbora Hladká
hladka@ufal.mff.cuni.cz

Martin Holub
holub@ufal.mff.cuni.cz

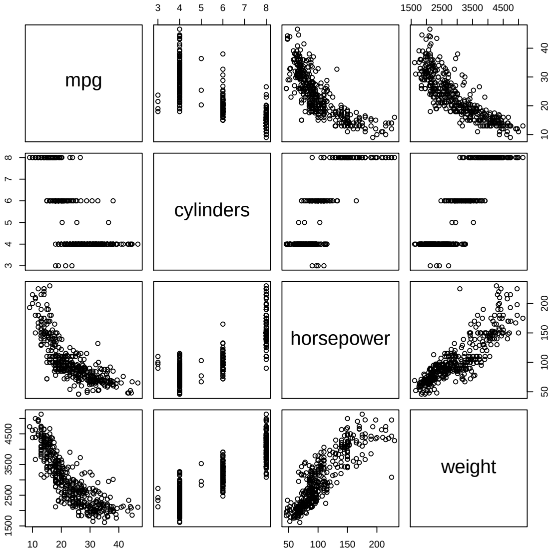
Charles University,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

# Lecture #12

**Principal Component Analysis** is

- a tool to analyze the data
- a tool to do dimensionality reduction

# Auto data set



# Basic concepts needed

- data analysis
  measures of center and spread, covariance and correlation
- linear algebra
  eigenvectors, eigenvalues, dot product, basis

# Data analysis

**How two features are related**

Both covariance and correlation indicate how closely two features relationship follows a straight line.

**Covariance** $\mathrm{cov}(X, Y)$ is a measure of the joint variability of two random variables $X$ and $Y$

$$\mathrm{cov}(X, Y) = E[(X - EX)(Y - EY)]$$

The magnitude of the covariance is not easy to interpret because it is not normalized and hence depends on the magnitudes of the variables. Therefore normalize the covariance $\rightarrow$ **Pearson correlation** coefficient

$$-1 \leq \rho_{X,Y} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y} \leq +1$$

# Data analysis

**Covariance matrix** of features $\mathrm{A}_1, \dots, \mathrm{A}_m$

$$\mathrm{COV}(\mathrm{A}_1, \dots, \mathrm{A}_m) = \begin{pmatrix} \mathrm{var}(\mathrm{A}_1) & \mathrm{cov}(\mathrm{A}_1, \mathrm{A}_2) & \dots & \mathrm{cov}(\mathrm{A}_1, \mathrm{A}_m) \\ \mathrm{cov}(\mathrm{A}_2, \mathrm{A}_1) & \mathrm{var}(\mathrm{A}_2) & \dots & \mathrm{cov}(\mathrm{A}_2, \mathrm{A}_m) \\ \dots & \dots & \dots & \dots \\ \mathrm{cov}(\mathrm{A}_m, \mathrm{A}_1) & \mathrm{cov}(\mathrm{A}_m, \mathrm{A}_2) & \dots & \mathrm{var}(\mathrm{A}_m) \end{pmatrix}$$

# Data analysis
# Auto data set

```
> cov(Auto[c("mpg", "cylinders", "horsepower", "weight")])

#                    mpg   cylinders  horsepower     weight
# mpg           60.91814  -10.352928  -233.85793  -5517.441
# cylinders    -10.35293    2.909696    55.34824   1300.424
# horsepower  -233.85793   55.348244  1481.56939  28265.620
# weight     -5517.44070 1300.424363 28265.62023 721484.709

> cor(Auto[c("mpg", "cylinders", "horsepower", "weight")])

#                   mpg  cylinders horsepower     weight
# mpg         1.0000000 -0.7776175 -0.7784268 -0.8322442
# cylinders  -0.7776175  1.0000000  0.8429834  0.8975273
# horsepower -0.7784268  0.8429834  1.0000000  0.8645377
# weight     -0.8322442  0.8975273  0.8645377  1.0000000
```

# Linear algebra

**Eigenvector $\mathbf{u}$, eigenvalue $\lambda$:** $\mathbf{A} \cdot \mathbf{u} = \lambda \mathbf{u}$

- $\mathbf{u}$ does not change its direction under the transformation
- $\lambda \mathbf{u}$ scales a vector $\mathbf{u}$ by $\lambda$; it changes its length, not its direction

1. The covariance matrix of $\mathbf{X}$ is an $m \times m$ symmetric matrix given by $\frac{1}{n-1}\mathbf{X}\mathbf{X}^\top$
2. Any symmetric matrix $m \times m$ has a set of orthonormal eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ associated with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_m$
   - for any $i$, $\mathbf{A} \cdot \mathbf{v}_i = \lambda_i \mathbf{v}_i$
   - $||\mathbf{v}_i|| = 1$
   - $\mathbf{v}_i \cdot \mathbf{v}_j = 0$ if $i \neq j$
3. $\mathbf{A}$ is a symmetric $m \times m$ matrix and $\mathbf{E}$ is an $m \times m$ matrix whose $i$-th column is the $i$-th eigenvector of $\mathbf{A}$. The eigenvectors are ordered in terms of decreasing values of their associated eigenvalues. Then there is a diagonal matrix $\mathbf{D}$ such that $\mathbf{A} = \mathbf{E} \cdot \mathbf{D} \cdot \mathbf{E}^\top$
4. If the rows of $\mathbf{E}$ are orthogonal, then $\mathbf{E}^{-1} = \mathbf{E}^\top$

# Linear algebra

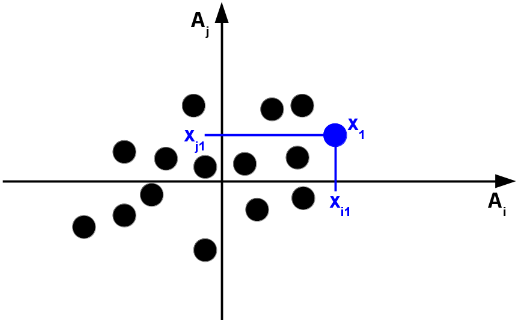
**Basis** of $\mathcal{R}^m$ is a set of linearly independent vectors $\mathbf{u}_1, \dots, \mathbf{u}_m$

- none of them is a linear combination of other vectors
- $\mathbf{u}_i \cdot \mathbf{u}_j = 0$, $i, j = 1, \dots m$, $i \neq j$
- any $\mathbf{u} = c_1\mathbf{u}_1 + \cdots + c_m\mathbf{u}_m$
- for example, the standard basis of the *3*-dimensional Euclidean space $\mathcal{R}^3$ consists of $\mathbf{x} = \langle 1, 0, 0\rangle, \mathbf{y} = \langle 0, 1, 0\rangle, \mathbf{z} = \langle 0, 0, 1\rangle$. It is an example of orthonormal basis, so called *naive* basis $\mathbf{I}$

# Principal Component Analysis

Representation of $Data = \{\mathbf{x}_i, \mathbf{x}_i = \langle x_{1i}, \dots, x_{mi} \rangle\}$, $|Data| = n$ for PCA

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1n} \\ x_{21} & \dots & x_{2n} \\ \dots & \dots & \dots \\ x_{m1} & \dots & x_{mn} \end{pmatrix}$$



# PCA

**Which features to keep?**

- features that change a lot, i.e. high variance
- features that do not depend on others, i.e. low covariance

**Which features to ignore?**

- features with some noise, i.e. low variance

# PCA principles

1. high correlation $\sim$ high redundancy
2. the most important feature has the largest variance

# PCA

- **Question**

  Is there any other representation of $\mathbf{X}$ to extract the most important features?

- **Answer**

  Use another basis

$$\mathbf{P}^\top \cdot \mathbf{X} = \mathbf{Z}$$

  where $\mathbf{P}$ transforms $\mathbf{X}$ into $\mathbf{Z}$

# PCA
## Heading for P

$$\mathbf{P} = \begin{pmatrix} \mathbf{p}_{11} & \dots & \dots & \mathbf{p}_{1m} \\ \mathbf{p}_{21} & \dots & \dots & \mathbf{p}_{2m} \\ \dots & \dots & \dots & \dots \\ \mathbf{p}_{m1} & \dots & \dots & \mathbf{p}_{mm} \end{pmatrix}$$

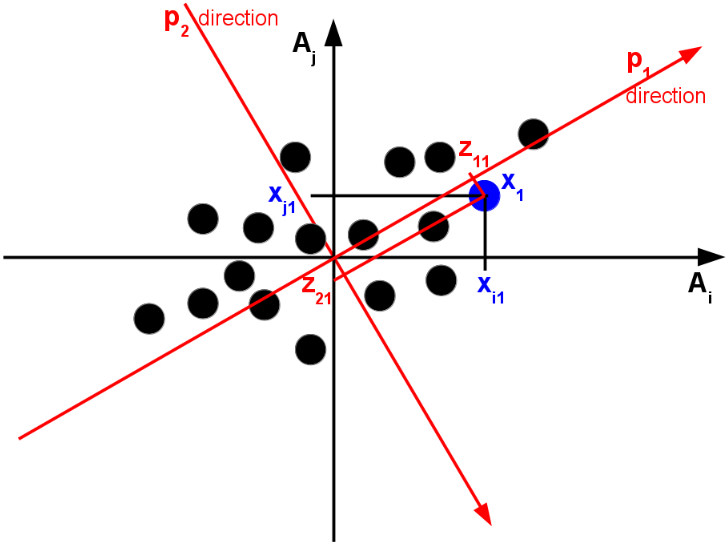
- **principal components** of $\mathbf{X}$ are the vectors $\mathbf{p}_i = \langle p_{1i}, \dots, p_{mi} \rangle$
- **principal component loadings** of $\mathbf{p}_i$ are the elements $p_{i1}, \dots, p_{im}$

# PCA
## Heading for P

$$\mathbf{Z} = \begin{pmatrix} \mathbf{p}_1 \cdot \mathbf{x}_1 & \dots & \dots & \mathbf{p}_1 \cdot \mathbf{x}_n \\ \mathbf{p}_2 \cdot \mathbf{x}_1 & \dots & \dots & \mathbf{p}_2 \cdot \mathbf{x}_n \\ \dots & \dots & \dots & \dots \\ \mathbf{p}_m \cdot \mathbf{x}_1 & \dots & \dots & \mathbf{p}_m \cdot \mathbf{x}_n \end{pmatrix}$$

*i*-**principal component scores** of $n$ instances are $\mathbf{p}_i \cdot \mathbf{x}_1, \mathbf{p}_i \cdot \mathbf{x}_2, \dots, \mathbf{p}_i \cdot \mathbf{x}_n$

# PCA
## Heading for P



# PCA
## Heading for P

- What is a good choice of **P**?
- What features we would like **Z** to exhibit?

**Goal:** **Z** is a new representation of **X**

The new features are linear combinations of the original features whose weights are given by **P**.

The covariance matrix of **Z** is diagonal and the entries on the diagonal are in descending order, i.e. the covariance of any pair of distinct features is zero, and the variance of each of our new features is listed along the diagonal.

# PCA
## Heading for P

- principal components are new basis vectors to represent $\mathbf{x}_j$, $j = 1, \ldots, n$
- $\mathbf{p}_i \cdot \mathbf{x}_j$ is a projection of $\mathbf{x}_j$ on $\mathbf{p}_i$
- changing the basis does not change data, it changes their representation

Covariance matrix $\mathrm{cov}(\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_\mathrm{m})$

- on the diagonal, large values correspond to interesting structure
- off the diagonal, large values correspond to high redundancy

# Derivation of PCA

1. preprocessing *Data*
   mean normalization to get centered data $\rightarrow \mathbf{X}$

2. $\mathrm{cov}(\mathbf{X}) = \mathbf{A} = \frac{1}{n-1}\mathbf{X}\mathbf{X}^\top$

3. Compute eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$ and eigenvalues $\lambda_1, \ldots, \lambda_m$ of $\mathbf{A}$

4. Take the eigenvectors, order them by eigenvalues, i.e. by significance, highest to lowest: $\mathbf{p}_1, \ldots, \mathbf{p}_m, \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$

5. The eigenvectors $\mathbf{p}_1, \ldots, \mathbf{p}_m$ become columns of $\mathbf{P}$

$$\boldsymbol{p}_i = \begin{pmatrix} p_{1i} \\ \ldots \\ p_{mi} \end{pmatrix}$$

# Properties of PCA

$$\mathbf{P}^\top \cdot \mathbf{X} = \mathbf{Z}$$

$$\mathbf{Z} = \begin{pmatrix} \mathbf{p}_1 \cdot \mathbf{x}_1 & \dots & \dots & \mathbf{p}_1 \cdot \mathbf{x}_n \\ \mathbf{p}_2 \cdot \mathbf{x}_1 & \dots & \dots & \mathbf{p}_2 \cdot \mathbf{x}_n \\ \dots & \dots & \dots & \dots \\ \mathbf{p}_m \cdot \mathbf{x}_1 & \dots & \dots & \mathbf{p}_m \cdot \mathbf{x}_n \end{pmatrix}$$

- The $i$-th diagonal value of $\mathrm{cov}(\mathbf{Z})$ is the variance of $\mathbf{X}$ along $\mathbf{p_i}$.
- We calculate a rotation of the original coordinate system such that all non-diagonal elements of the new covariance matrix become zero.
- The principal components define the basis of the new coordinate axes and the eigenvalues correspond to the diagonal elements of the new covariance matrix.
- So the eigenvalues, by definition, define the variance along the corresponding principal components.

# Properties of PCA

$$cov(\mathbf{P}^\top \cdot \mathbf{X}) \overset{\text{see p.49.1}}{=} \frac{1}{n-1}(\mathbf{P}^\top \cdot \mathbf{X}) \cdot (\mathbf{P}^\top \cdot \mathbf{X})^\top =$$

$$\frac{1}{n-1}\mathbf{P}^\top \cdot \mathbf{X} \cdot \mathbf{X}^\top \cdot \mathbf{P} \overset{\text{let } \mathbf{A}=\mathbf{X}\cdot\mathbf{X}^\top}{=} \frac{1}{n-1}\mathbf{P}^\top \cdot \mathbf{A} \cdot \mathbf{P} =$$

$$\overset{\text{see p.49.3}}{=} \frac{1}{n-1}\mathbf{P}^\top \cdot (\mathbf{P} \cdot \mathbf{D} \cdot \mathbf{P}^\top) \cdot \mathbf{P} \overset{\text{see p.49.4}}{=} \frac{1}{n-1}\mathbf{P}^\top \cdot (\mathbf{P}^\top)^{-1} \mathbf{D} \cdot \mathbf{P}^\top \cdot (\mathbf{P}^\top)^{-1} = \frac{1}{n-1}\mathbf{D}$$

# Properties of PCA

**A geometric interpretation for the first principal component $\mathbf{p}_1$**

It defines a direction in feature space along which the data vary the most. If we project the $n$ instances $\mathbf{x}_1, \ldots, \mathbf{x}_n$ onto this direction, the projected values are the principal component scores $z_{11}, \ldots, z_{n1}$ themselves.

# Proportion of Variance Explained (PVE)

The fraction of variance explained by a $k$-th principal component $\mathrm{PVE}(p_k)$ is the ratio between the variance of that principal component and the total variance.

- total variance in $\mathbf{X}$: $\sum_{j=1}^{m} \mathrm{var}(A_j) = \sum_{i=1}^{m} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$
  (assuming feature normalization)
- variance expressed by $\mathbf{p}_k$: $\frac{1}{n} \sum_{i=1}^{n} z_{ki}^2$
- $\mathrm{PVE}(\mathbf{p}_k) = \frac{\sum_{i=1}^{n} z_{ki}^2}{\sum_{i=1}^{m} \sum_{i=1}^{n} x_{ij}^2}$
- $\mathrm{PVE}(\mathbf{p}_1, \ldots, \mathbf{p}_M) = \sum_{i=1}^{M} \mathrm{PVE}(\mathbf{p}_i)$, $M \leq m$

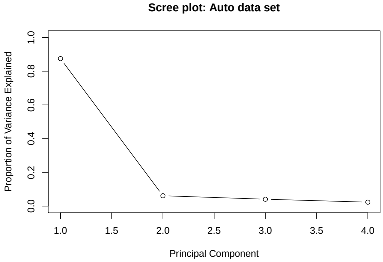
# PCA
# Auto data set

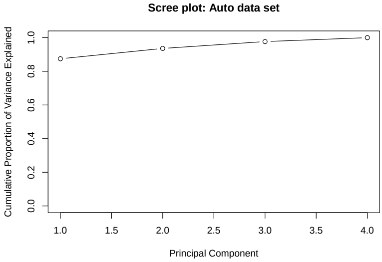
```
> a <- Auto[c("mpg", "cylinders", "horsepower", "weight")]
> pca.a <- prcomp(a, scale = TRUE)
> summary(pca.a)

# Importance of components:
#                          Comp.1    Comp.2   Comp.3   Comp.4
Standard deviation     1.8704 0.49540 0.40390 0.30518
Proportion of Variance 0.8746 0.06135 0.04078 0.02328
Cumulative Proportion  0.8746 0.93593 0.97672 1.00000
```

# PCA
# Auto data set

## Scree plot



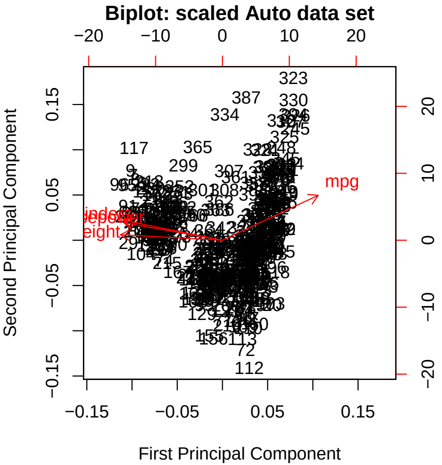
# PCA
## Auto data set

```
> pca.a$rotation
                  PC1       PC2         PC3        PC4
mpg         0.4833271 0.8550485 -0.02994982  0.1854453
cylinders  -0.5033993 0.3818233 -0.55748381 -0.5385276
horsepower -0.4984381 0.3346173  0.79129092 -0.1159714
weight     -0.5143380 0.1055192 -0.24934614  0.8137252
```

- PC1 places approximately equal weight on `cylinders`, `horsepower`, `weight` with much higher weight on `mpg`.
- PC2 places most of its weight on `mpg` and less weight on the other three features.

# PCA
# Auto data set

A biplot displays both the PC scores and the PC loadings.



# PCA
## Auto data set

**The biplot for the Auto data set is showing**

- the scores of each example (i.e., cars) on the first two principal components with axes on the top and right
  – see the id cars in black
- the loading of each feature (i.e., `mpg`, `weight`, `cylinders`, `horsepower`) on the first two principal components with axes on the bottom and left
  – see the red arrows

# PCA

In general, a $m \times n$ matrix $\mathbf{X}$ has $\min(n-1, m)$ distinct principal components.

- **Question**
  How many principal components are needed?
- **Answer**
  There is no single answer to this question. Study scree plots.

# Summary of Examination Requirements

- Principal Component Analysis
  data nalaysis, derivation, scree plot, biplot