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

Outline

- Principal Component Analysis
- Maximum Likelihood Estimation

Principal Component Analysis

PCA is

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

Basic concepts needed

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

Data analysis

How two variables are related

Both covariance and correlation indicate how closely two variables 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.

- > 0 both variables increase or decrease together
- < 0 while one variable increases the other decreases
- $= 0$ variables are linearly independent of each other

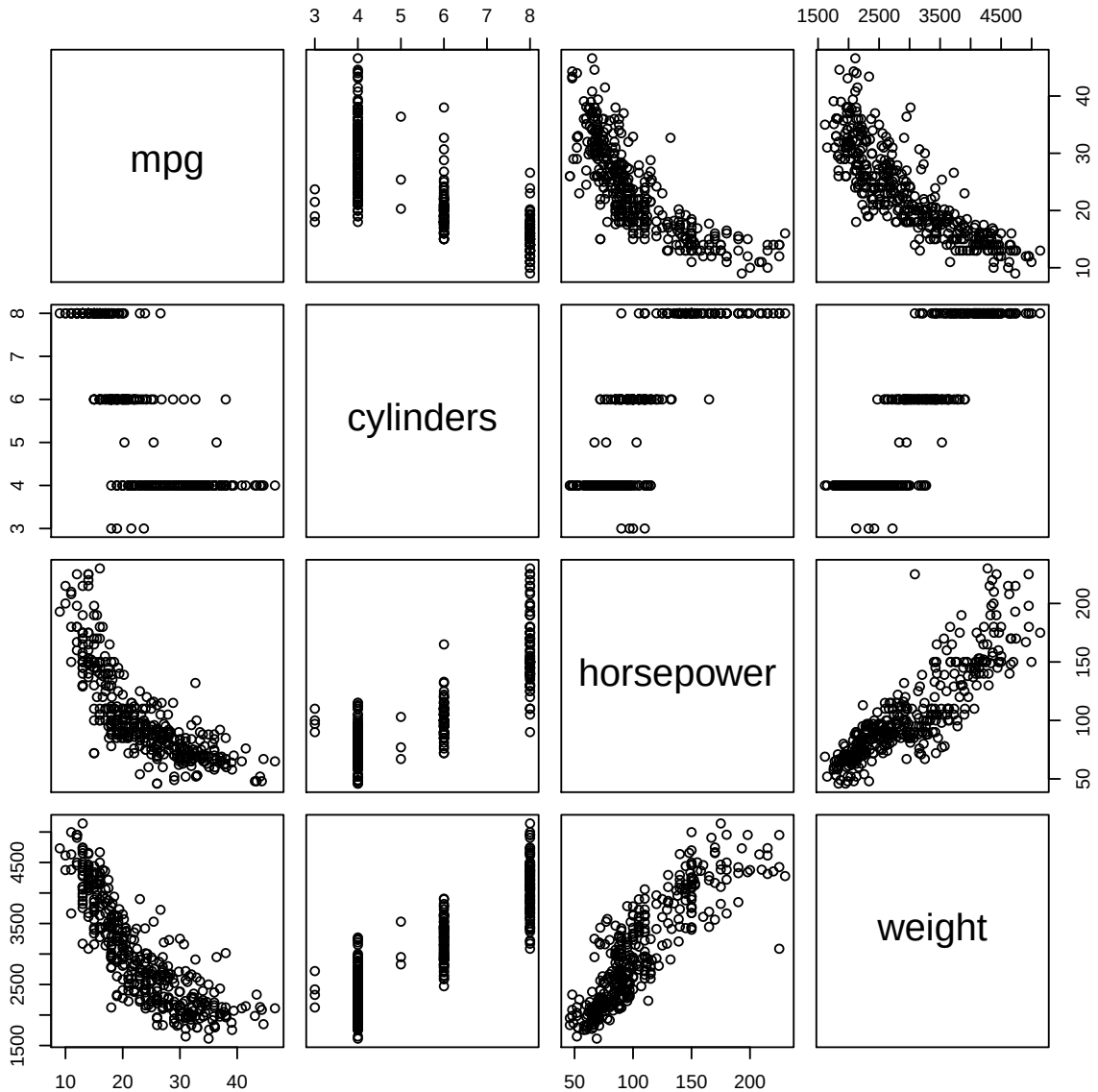
Data analysis

How two variables are related

Therefore normalize the covariance $\rightarrow$ **Pearson correlation** coefficient

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

Auto data set



Data analysis

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

$$\mathbf{C}(\mathrm{A}_1, \ldots, \mathrm{A}_m) = \begin{pmatrix} \mathrm{var}(\mathrm{A}_1) & \mathrm{cov}(\mathrm{A}_1, \mathrm{A}_2) & \ldots & \mathrm{cov}(\mathrm{A}_1, \mathrm{A}_m) \\ \mathrm{cov}(\mathrm{A}_2, \mathrm{A}_1) & \mathrm{var}(\mathrm{A}_2) & \ldots & \mathrm{cov}(\mathrm{A}_2, \mathrm{A}_m) \\ \ldots & \ldots & \ldots & \ldots \\ \mathrm{cov}(\mathrm{A}_m, \mathrm{A}_1) & \mathrm{cov}(\mathrm{A}_m, \mathrm{A}_2) & \ldots & \mathrm{var}(\mathrm{A}_m) \end{pmatrix}$$

- diagonal - variance of the features $\mathrm{var}(\mathrm{A}_i)$
- symmetrical about the diagonal $\mathrm{cov}(\mathrm{A}_i, \mathrm{A}_j) = \mathrm{cov}(\mathrm{A}_j, \mathrm{A}_i)$

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
```

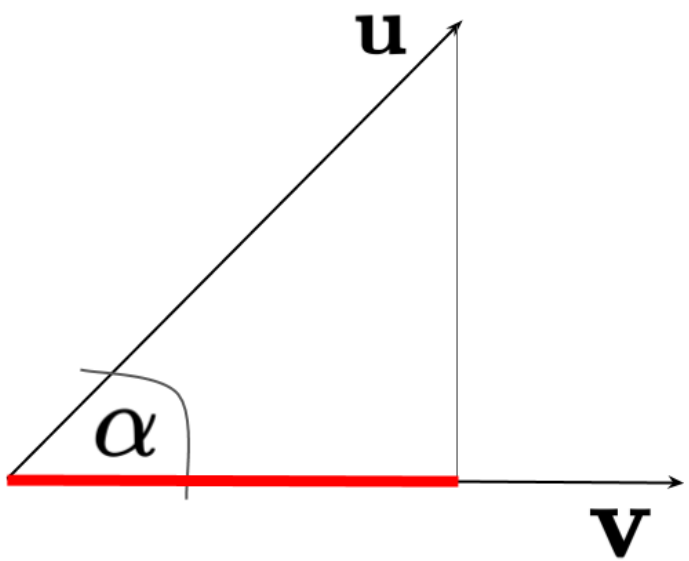
Linear algebra

1. **A** is a linear transformation. **Eigenvector** of **A** is a vector **u** for which exists **eigenvalue** λ so that $\mathbf{A} \cdot \mathbf{u} = \lambda \mathbf{u}$
   - eigenvector **u** does not change its direction under the transformation **A**
   - $\lambda \mathbf{u}$ scales a vector **u** by λ; it changes its length, not its direction
2. The covariance matrix of **X** is an $m \times m$ symmetric matrix $\mathbf{C}(\mathbf{X}) = \frac{1}{n-1}\mathbf{X}\mathbf{X}^\top$
3. Any symmetric matrix $m \times m$ **A** 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 $i \neq j$
4. **A** is a symmetric $m \times m$ matrix and **E** is an $m \times m$ matrix whose i-th column is the i-th eigenvector of **A**. The eigenvectors are ordered in terms of decreasing values of their associated eigenvalues. Then there is a diagonal matrix **D** such that $\mathbf{A} = \mathbf{E} \cdot \mathbf{D} \cdot \mathbf{E}^\top$
5. If the rows of **E** are orthogonal, then $\mathbf{E}^{-1} = \mathbf{E}^\top$

Linear algebra

Dot product

- $\mathbf{u} = \langle u_1, \ldots, u_m \rangle$, $\mathbf{v} = \langle v_1, \ldots, v_m \rangle$
- algebraic definition $\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + \cdots + u_m v_m$
- geometric definition $\mathbf{u} \cdot \mathbf{v} = ||\mathbf{u}|| \cdot ||\mathbf{v}|| \cdot \cos \alpha$
- $\mathbf{u}$ and $\mathbf{v}$ are orthogonal iff $\mathbf{u} \cdot \mathbf{v} = 0$



$$||\mathbf{v}|| = 1 \rightarrow \mathbf{u} \cdot \mathbf{v} = ||\mathbf{u}|| \cdot \cos \alpha$$

Linear algebra

- A set of vectors $\mathbf{x}_i \in \mathcal{R}^m$ is linearly independent if no vector is a linear combination of other vectors.

Basis of $\mathcal{R}^m$ is a set vectors $\mathbf{u}_1, \ldots, \mathbf{u}_m$

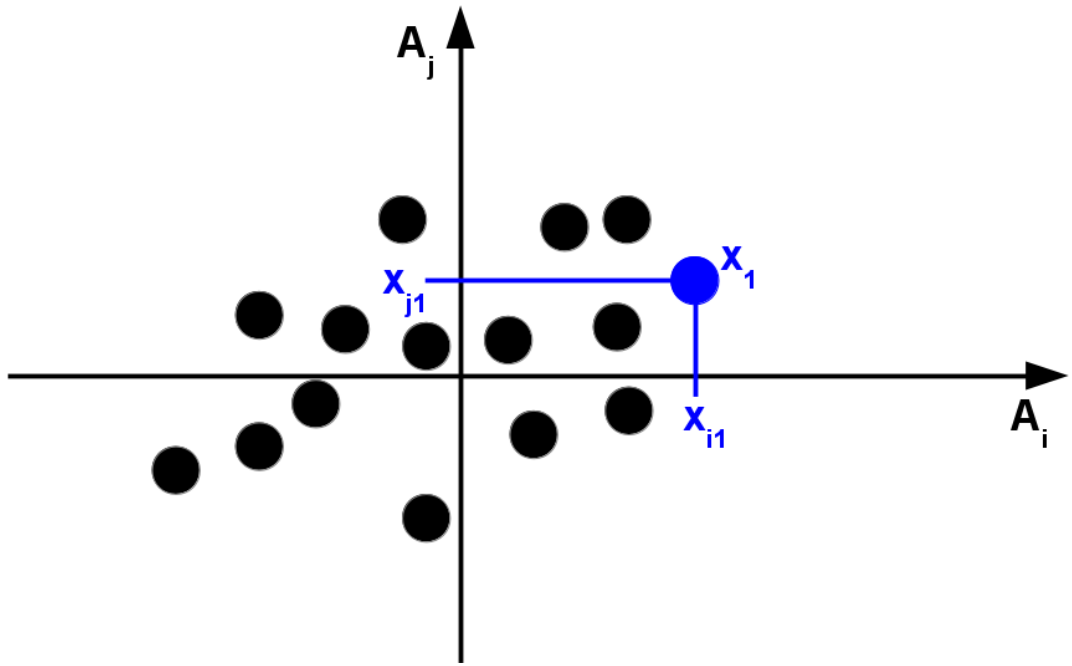
- linearly independent
- $\mathbf{u}_i \cdot \mathbf{u}_j = 0,\ i, j = 1, \ldots, m,\ i \neq j$
- any $\mathbf{u} \in \mathcal{R}^m$: $\mathbf{u} = c_1\mathbf{u}_1 + \cdots + c_m\mathbf{u}_m$
- for example, the standard basis of the *3*-dimensional Euclidean space $\mathcal{R}^3$ consists of $\mathbf{x} = \langle 1, 0, 0 \rangle, \mathbf{y} = \langle 0, 1, 0 \rangle, \mathbf{z} = \langle 0, 0, 1 \rangle$. It is an example of orthonormal basis, so called *naive* basis **I**

Principal Component Analysis

$Data = \{\mathbf{x}_i, \mathbf{x}_i = \langle x_{1i}, \ldots, x_{mi} \rangle\}$, $|Data| = n$

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

(i.e. examples in columns)



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

$\mathbf{C}(\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_\mathrm{m})$

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

- high correlation $\sim$ high redundancy
- 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 $\mathbf{Z}$; $\mathbf{Z}$ is a new representation of $\mathbf{X}$

PCA
Heading for P

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

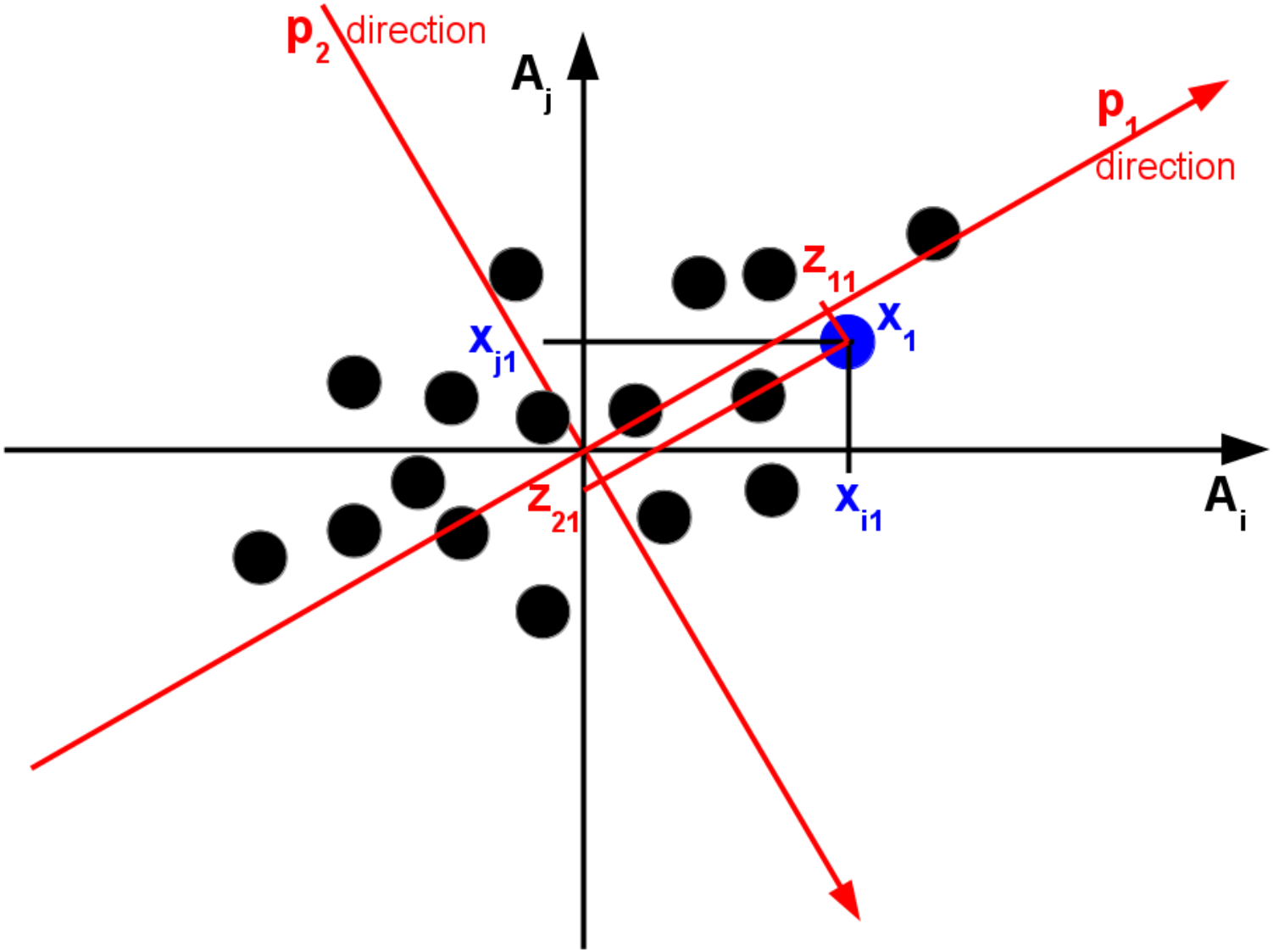
- **principal components** of $\mathbf{X}$ are the vectors $\mathbf{p}_i = \langle p_{1i}, \ldots, p_{mi} \rangle$
- **principal component loadings** of $\mathbf{p}_i$ are the elements $p_{i1}, \ldots, p_{im}$

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: Find a set of directions on which to project the data such that

- the variance of each projection is maximized
- the projections are uncorrelated (random variables X, Y are said to be uncorrelated if their $\mathrm{cov}(X, Y) = 0$)

PCA
Heading for P

Let's compute the variance of a random variable obtained by projecting $\mathbf{X}$ onto a direction represented by the vector $\mathbf{p}$ ($\mu = E[\mathbf{X}]$):

$$\sigma^2 = E[(\mathbf{p}^\top \mathbf{X} - E[\mathbf{p}^\top \mathbf{X}])^2] = \mathbf{p}^\top E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^\top]\mathbf{p} = \mathbf{p}^\top \mathbf{C}(\mathbf{X})\mathbf{p}$$

We use the method of Lagrange multipliers:

Maximize

$$\mathbf{p}^\top \mathbf{C}(\mathbf{X})\mathbf{p}$$

subject to

$$\mathbf{p}^\top \mathbf{p} = 1$$

PCA
Heading for P

Lagrangian function

$$\mathcal{L}(\mathbf{p}, \lambda) = \mathbf{p}^\top \mathbf{C}(\mathbf{X})\mathbf{p} - \lambda \mathbf{p}^\top \mathbf{p}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{p}} = 0 \Rightarrow \mathbf{C}(\mathbf{X})\mathbf{p} = \lambda \mathbf{p}$$

Our problem comes down to seeking eigenvalues and eigenvectors of $\mathbf{C}(\mathbf{X})$.
In the general case, $\mathbf{C}(\mathbf{X})$ has m distinct eigenvectors and eigenvalues.

Which one is the solution we seek?

$$\sigma^2 = \mathbf{p}^\top \mathbf{C}(\mathbf{X})\mathbf{p} = \mathbf{p}^\top \lambda \mathbf{p} = \lambda \mathbf{p}^\top \mathbf{p} = \lambda$$

The variance is maximized if we choose the unit eigenvector that corresponds to the largest eigenvalue of $\mathbf{C}(\mathbf{X})$. Denote these as $\mathbf{p}_1$, λ_1.

Usually we cannot represent the data sufficiently good with just one projection. Thus, we need to find the procedure for computing the next projection directions $\mathbf{p}_2$, λ_2, $\mathbf{p}_3$, λ_3, ...

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

Derivation of PCA

1. preprocessing *Data*
   mean normalization to get centered data $\rightarrow \mathbf{X}$
2. $\mathbf{C}(\mathbf{X}) = \mathbf{A} = \frac{1}{n-1}\mathbf{X}\mathbf{X}^\top$
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 $\mathbf{C}(\mathbf{Z})$ 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

$$\mathbf{C}(\mathbf{P}^\top \cdot \mathbf{X}) \overset{\text{see p.10.4}}{=} \frac{1}{n-1}(\mathbf{P}^\top \cdot \mathbf{X}) \cdot (\mathbf{P}^\top \cdot \mathbf{X})^\top =$$

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

$$\overset{\text{see p.10.4}}{=} \frac{1}{n-1}\mathbf{P}^\top \cdot (\mathbf{P} \cdot \mathbf{D} \cdot \mathbf{P}^\top) \cdot \mathbf{P} \overset{\text{see p.10.5}}{=} \frac{1}{n-1}\mathbf{P}^\top \cdot (\mathbf{P}^\top)^{-1}\mathbf{D} \cdot \mathbf{P}^\top \cdot (\mathbf{P}^\top)^{-1} = \frac{1}{n-1}\mathbf{D}$$

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)

How much of the information in a given data set is lost by projecting the instances onto the first few principal components?

In other words, how much of the variance in the data is not contained in the first few principal components?

- 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 $\mathbf{p}_k$: $\frac{1}{n} \sum_{i=1}^{n} z_{ki}^2$
- $\mathrm{PVE}(\mathbf{p}_k) = \frac{\sum_{i=1}^{n} z_{ki}^2}{\sum_{i=1}^{m} \sum_{i=1}^{n} x_{ij}^2}$
- $\mathrm{PVE}(\mathbf{p}_1, \ldots, \mathbf{p}_M) = \sum_{i=1}^{M} \mathrm{PVE}(\mathbf{p}_i)$, $M \leq m$

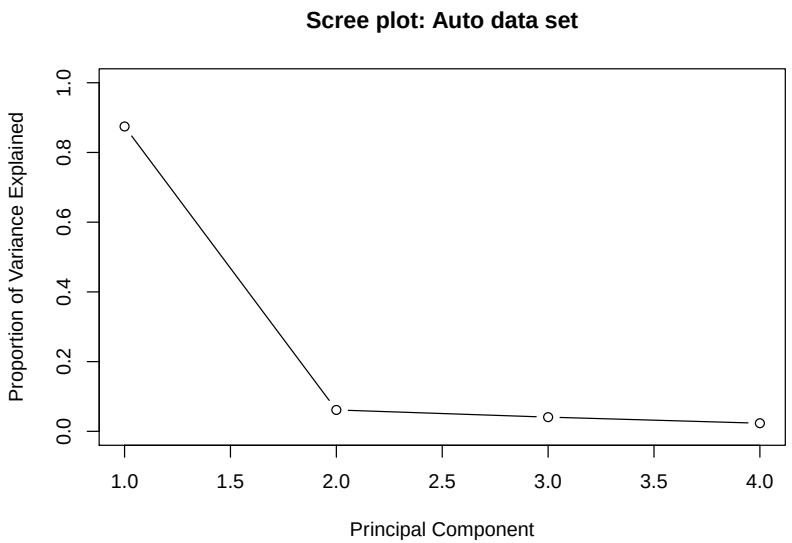
PCA
Auto data set

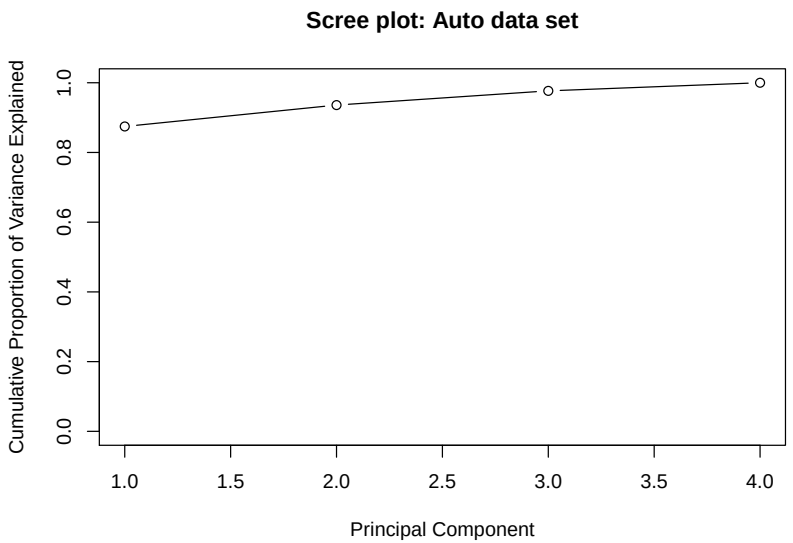
```
> a <- Auto[c("mpg", "cylinders", "horsepower", "weight")]
> pca.a <- prcomp(a, scale = TRUE)
> summary(pca.a)

# 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





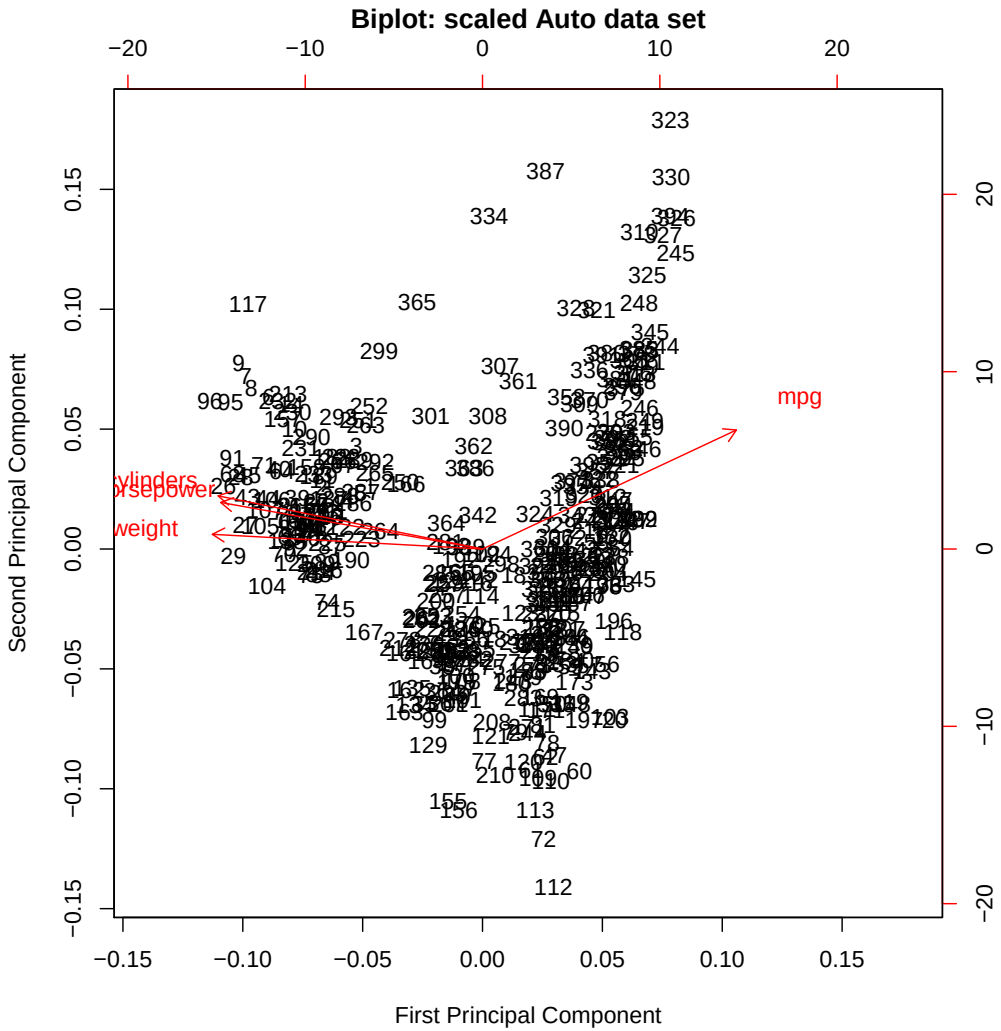
PCA
Auto data set

```
> 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.

Biplot for the Auto data set is showing

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



The biplot for the Auto data set is showing

- the scores of each example (i.e., car) on the first two principal components with axes on the bottom and left
  – 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 top and right
  – see the red arrows
  – their length corresponds to the variabliaty of the original features

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.

Probability vs. likelihood

Task: Predict the outcome of each of 10 coin tosses

probability

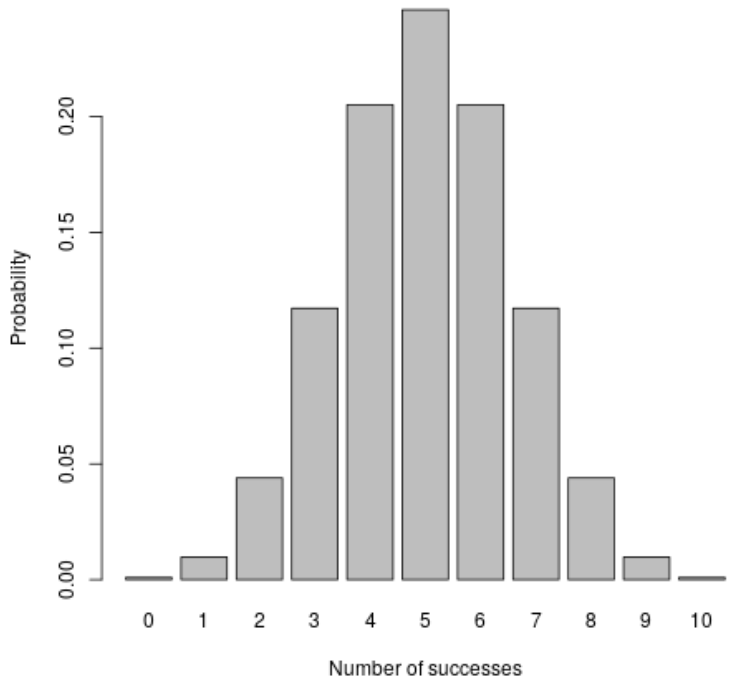
$\Pr(X = k|n = 10, p = 0.8)$

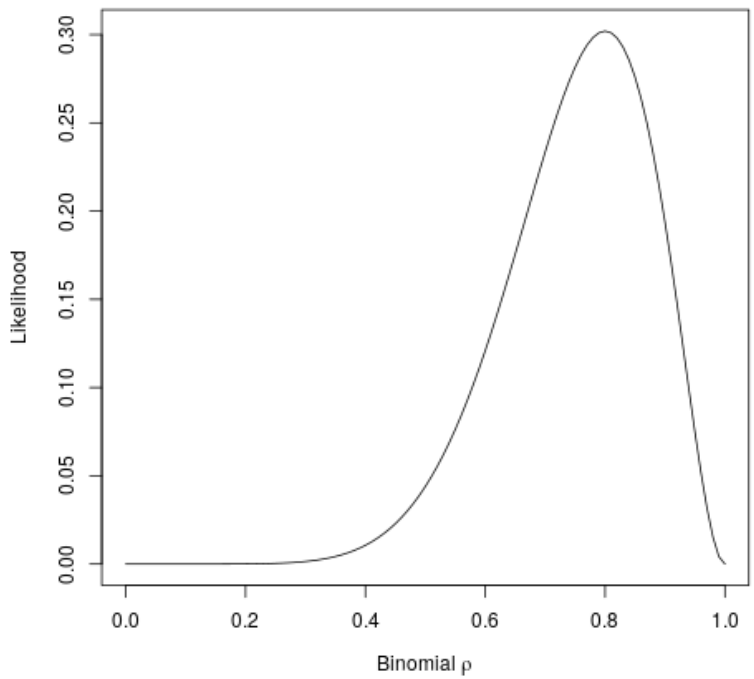
$\Pr(\text{data}|\theta)$

likelihood

$\mathcal{L}(p|X = 8)$

$\mathcal{L}(\theta|\text{data})$





Probability vs. likelihood

The binomial distribution is the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability p, $X \sim Bin(n, p)$.

Task: Predict the outcome of each of 10 coin tosses

probability

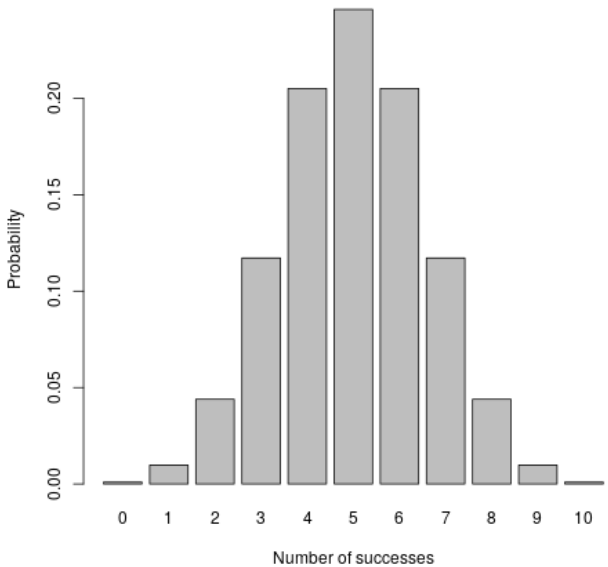
$\Pr(X = k|n = 10, p = 0.8)$

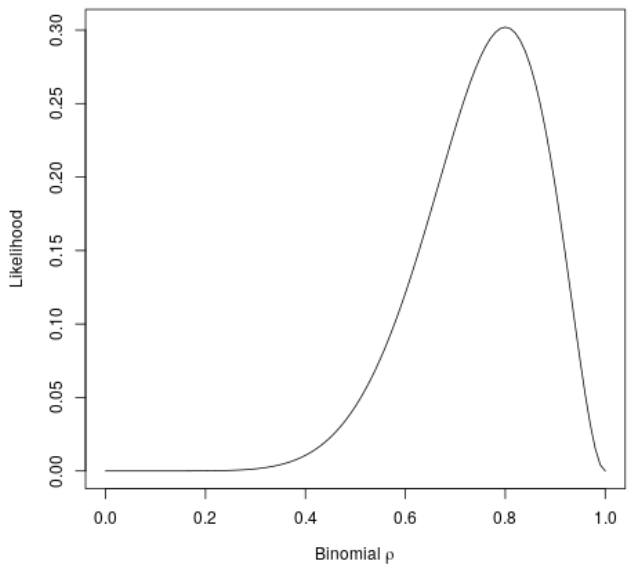
$\Pr(\text{data}|\theta)$

likelihood

$\mathcal{L}(p|X = 8)$

$\mathcal{L}(\theta|\text{data})$



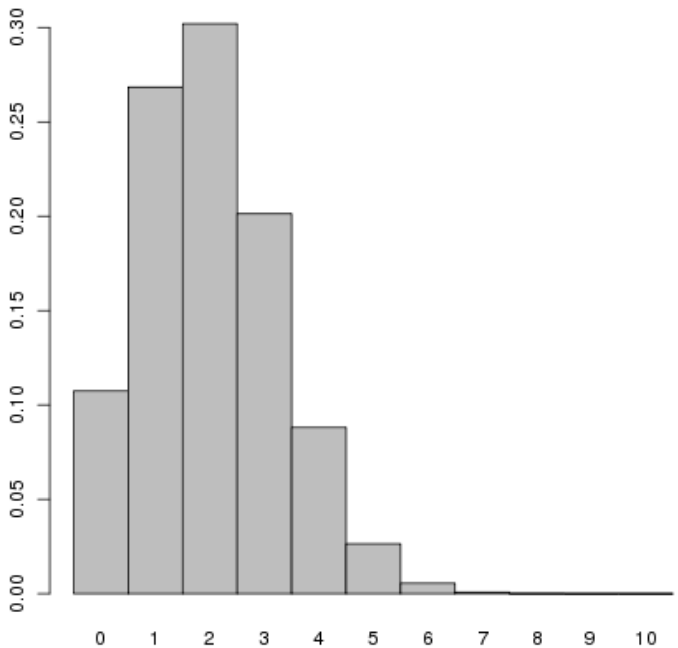


Binomial distribution

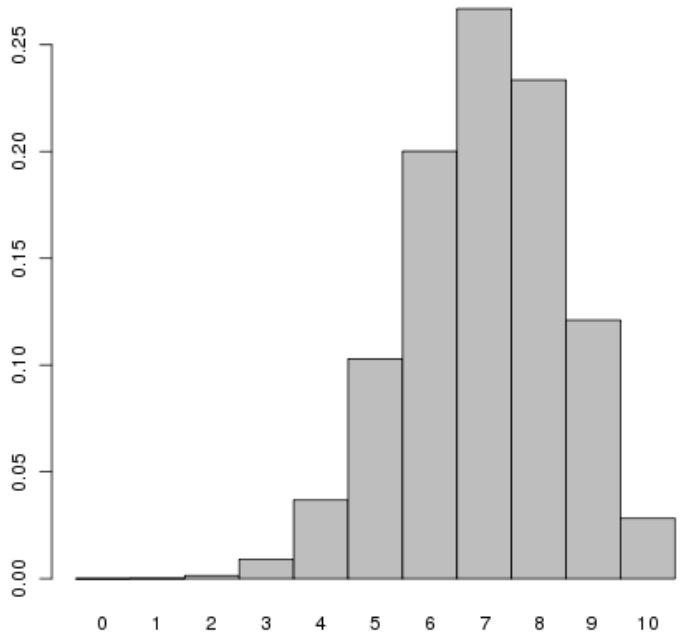
Probabilistic mass function

$$\Pr(X = k|n, p) = f(k; n, p) = \frac{n!}{k!(n-k)!} p^k (1-p)^{(n-k)}$$

$p = 0.2$



$p = 0.7$



Maximum likelihood estimation

- sample $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$

Assumption: $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent and identically distributed with an unknown probability density function $f(\mathbf{X}; \Theta)$

- Θ is a vector of parameters of the probability distribution $\Theta = \langle \theta_1, \ldots, \theta_m \rangle$
- joint density function $f(\mathbf{x}_1, \ldots, \mathbf{x}_n; \Theta) \overset{i.i.d.}{=} \prod_{i=1}^{n} f(\mathbf{x}_i; \Theta)$

We determine what value of Θ would make the data $\mathbf{X}$ most likely.

Maximum likelihood estimation

MLE is a method for estimating parameters from data.

Goal: identify the population that is most likely to have generated the sample.

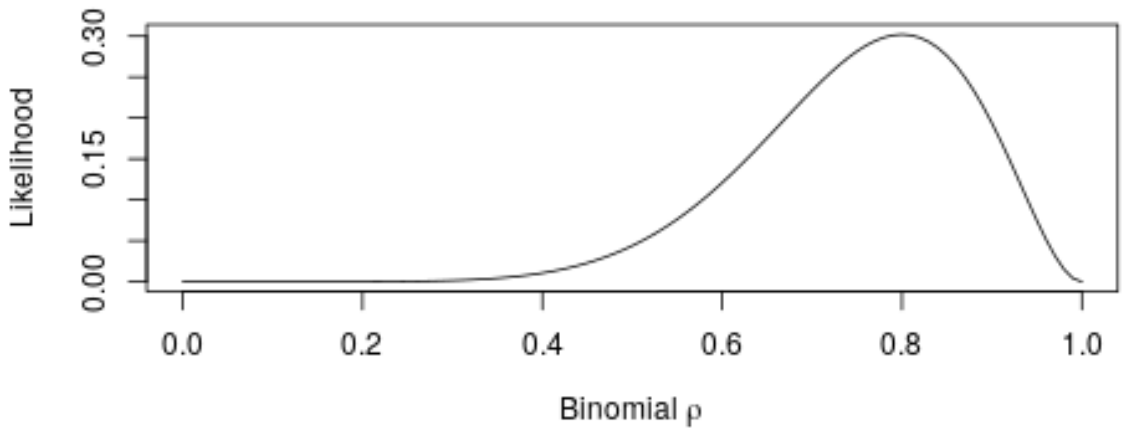
Likelihood function

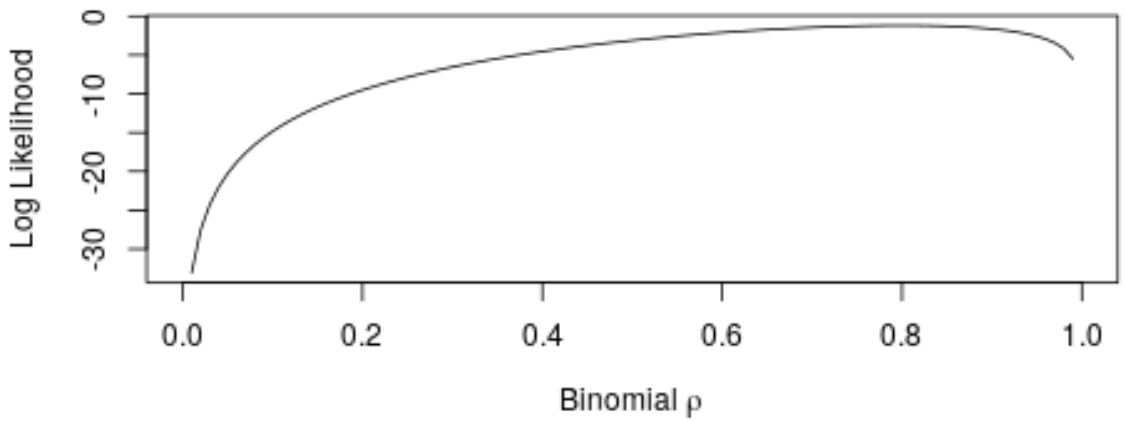
$$\mathcal{L}(\Theta|\mathbf{x}_1, \ldots, \mathbf{x}_n) \stackrel{df}{=} \prod_{i=1}^{n} f(\mathbf{x}_i; \Theta) \tag{1}$$

Log-likelihood function

$$\log \mathcal{L}(\Theta|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n} \log f(\mathbf{x}_i; \Theta) \tag{2}$$

Maximum likelihood estimation





Maximum likelihood estimation

Maximum likelihood estimate of Θ

$$\Theta^{\star}_{MLE} = \operatorname{argmax}_{\Theta} \log \mathcal{L}(\Theta|\mathbf{x}_1, \ldots, \mathbf{x}_n) \tag{3}$$

Maximum likelihood estimation

MLE analytically

- Likelihood equation: $\frac{\partial \log \mathcal{L}(\Theta|X)}{\partial \theta_i} = 0$ at θ_i for all $i = 1, \ldots, m$
- Maximum, not minimum: $\frac{\partial^2 \mathcal{L}(\Theta|\mathbf{x})}{\partial \theta_i^2} < 0$

Numerically

- Use an optimization algorithm (for ex. Gradient Descent)

Maximum likelihood estimation
Binomial distribution

Estimate the probability p that a coin lands head using the result of n coin tosses, k of which resulted in heads. $\Theta = \langle p \rangle$

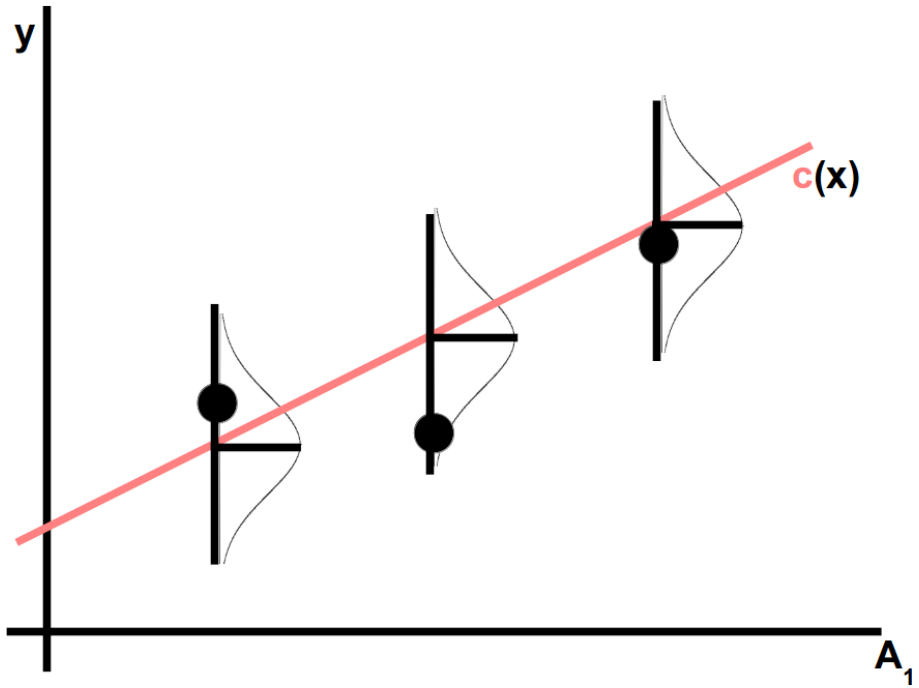
- $f(k; n, p) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$
- $\mathcal{L}(p|n, k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$
- $\log \mathcal{L}(p|n, k) = \log \frac{n!}{k!(n-k)!} + k \log p + (n-k) log(1-p)$
- $\frac{\partial \log \mathcal{L}(p|n,k)}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p} = 0$
- $\hat{p}_{MLE} = \frac{k}{n}$

Maximum likelihood estimation
Linear regression

Learn parameter estimates $\hat{\Theta}^\star$ from $Data = \{\langle \mathbf{x}_i, y_i \rangle, y_i \in \mathcal{R}, i = 1, ..., n\}$ using MLE.

Assumption: At each value of A_1, the output value y is subject to random error ϵ that is normally distributed $N(0, \sigma^2)$

$y_i = \Theta^\top \mathbf{x}_i + \epsilon_i$



Maximum likelihood estimation
Linear regression

- $\epsilon_i = y_i - \Theta^\top \mathbf{x}_i \sim N(0, \sigma^2)$
- probability density function of the Normal distribution

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mathcal{L}(\mu, \sigma|\epsilon) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(\epsilon_i-\mu)^2}{2\sigma^2}}$$

$$\mathcal{L}(\Theta, \sigma|\mathbf{X}, \mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(y_i-\Theta^\top \mathbf{x}_i)^2}{2\sigma^2}}$$

Maximum likelihood estimation
Linear regression

$$\log \mathcal{L}(\Theta, \sigma | \mathbf{X}, \mathbf{y}) = \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(y_i - \Theta^\top \mathbf{x}_i)^2}{2\sigma^2}$$

$$\operatorname{argmax}_\Theta \log \mathcal{L}(\Theta, \sigma | \mathbf{X}, \mathbf{y}) = \operatorname{argmax}_\Theta \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(y_i - \Theta^\top \mathbf{x}_i)^2$$

$$\operatorname{argmax}_{\boldsymbol{\Theta}} \log \mathcal{L}(\Theta, \sigma | \mathbf{X}, \mathbf{y}) = \operatorname{argmin}_\Theta \sum_{i=1}^{n} (y_i - \Theta^\top \mathbf{x}_i)^2$$

The minimum least square estimates are equivalent to the maximum likelihood estimates under the assumption that Y is generated by adding random noise to the true target values characterized by the Normal distribution $N(0, \sigma^2)$.

Maximum likelihood estimation
Logistic regression

Logistic regression models conditional probability using sigmoid function.

$$f(\mathbf{x}) = \frac{1}{1 + e^{-\Theta^T \mathbf{x}}} = \Pr(y = 1|\mathbf{x})$$

Learn parameter estimates $\hat{\Theta}^\star$ from $Data = \{\langle \mathbf{x}_i, y_i \rangle, y_i \in \{0, 1\}, i = 1, ..., n\}$ using MLE.

Maximum likelihood estimation
Logistic regression

$$f(\mathbf{x}; \Theta) = \Pr(y = 1|\mathbf{x})$$

$$\prod_{i=1}^{n} \Pr(y = y_i|\mathbf{x}_i) = \prod_{i=1}^{n} f(\mathbf{x}_i; \Theta)^{y_i}(1 - f(\mathbf{x}_i; \Theta))^{1-y_i}$$

$$\mathcal{L}(\Theta|\mathbf{X}, \mathbf{y}) = \prod_{i=1}^{n} f(\mathbf{x}_i; \Theta)^{y_i}(1 - f(\mathbf{x}_i; \Theta))^{1-y_i}$$

$$\log \mathcal{L}(\Theta|\mathbf{X}, \mathbf{y}) = \sum_{i=1}^{n} y_i \log f(\mathbf{x}_i; \Theta) + (1 - y_i) \log(1 - f(\mathbf{x}_i; \Theta))$$

$$\hat{\Theta}^{\star}_{MLE} = \operatorname{argmax}_{\Theta} \sum_{i=1}^{n} y_i \log f(\mathbf{x}_i; \Theta) + (1 - y_i) \log(1 - f(\mathbf{x}_i; \Theta))$$

Maximum likelihood estimation
Naïve Bayes classifier

$$\hat{y}^\star = argmax_{y_k \in Y} \Pr(y_k) \prod_{j=1}^{m} \Pr(x_j | y_k)$$

Maximum likelihood estimation
Naïve Bayes classifier

Categorical feature $\mathtt{A}_j$

Theorem

The Maximum likelihood estimates for NB take the form

- $\Pr(y) = \frac{c_y}{n}$ *where* $c_y = \sum_{i=1}^{n} \delta(y_i, y)$
- $\Pr(x|y) = \frac{c_{j_{x|y}}}{c_y}$ *where* $c_{j_{x|y}} = \sum_{i=1}^{n} \delta(y_i, y)\delta(\mathbf{x}_{i_j}, x)$

Maximum likelihood estimation
Naïve Bayes classifier

Continuous feature $\mathtt{A}_j$

Typical assumption, each continuous feature has a Gaussian distribution.

Theorem

The ML estimates for NB take the form

- $\overline{\mu_k} = \frac{\sum_{i=1}^{n} x_i^j \delta(y_i = y_k)}{\sum_{j=1}^{n} \delta(Y^j = y_k)}$
- $\overline{\sigma_k}^2 = \frac{\sum_{i=1}^{j} (x_i^j - \overline{\mu_k})^2 \delta(y_i = y_k)}{\sum_j \delta(Y^j = y_k)}$

$$\Pr(x|y_k) = \frac{1}{\sqrt{2\pi\overline{\sigma_k}^2}} e^{\frac{-(x-\overline{\mu_k})^2}{2\overline{\sigma_k}^2}}$$