ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

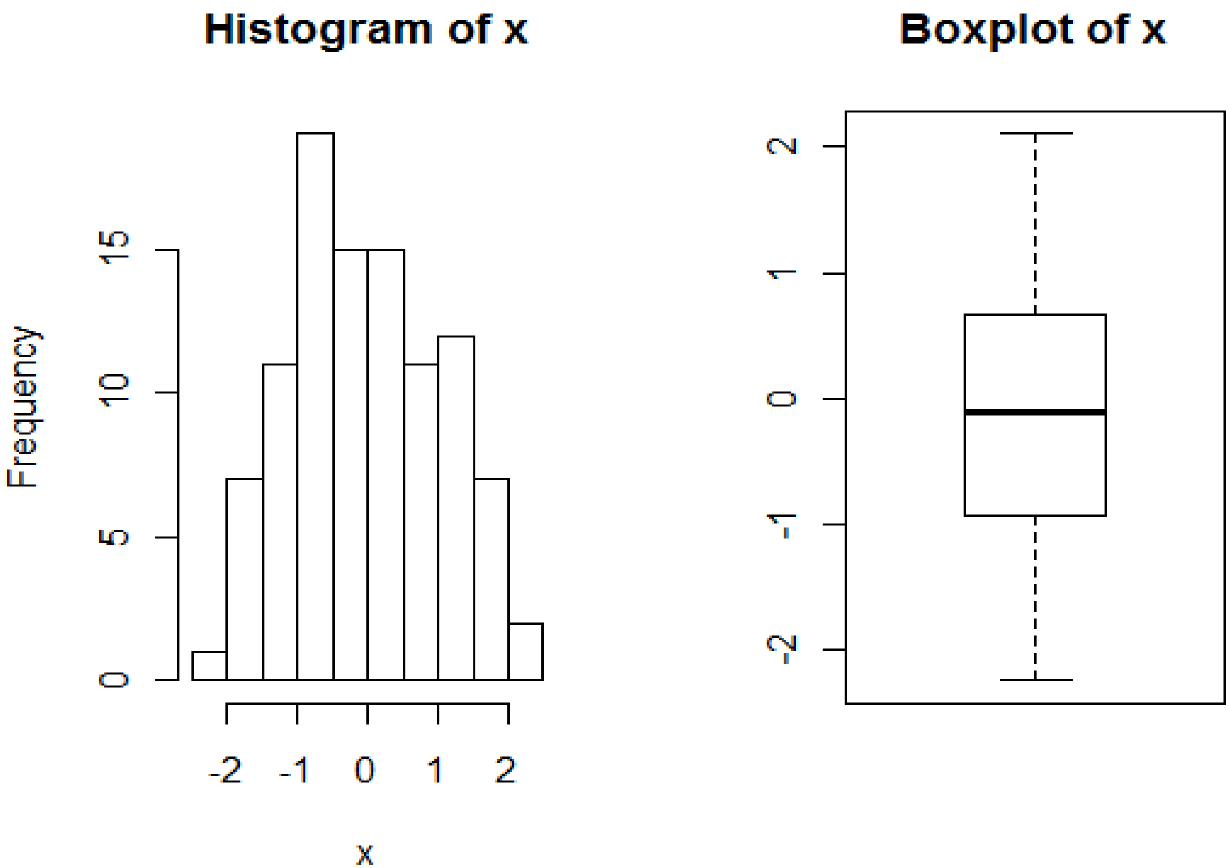
Visualization 1

Applied Multivariate Statistics – Spring 2013

Goals

- Covariance, Correlation (true / sample version)
- Test for zero correlation: Fisher's z-Transformation
- Scatterplot / Scatterplotmatrix
- Covariance matrix / Correlation matrix
- Multivariate Normal Distribution
- Mahalanobis distance

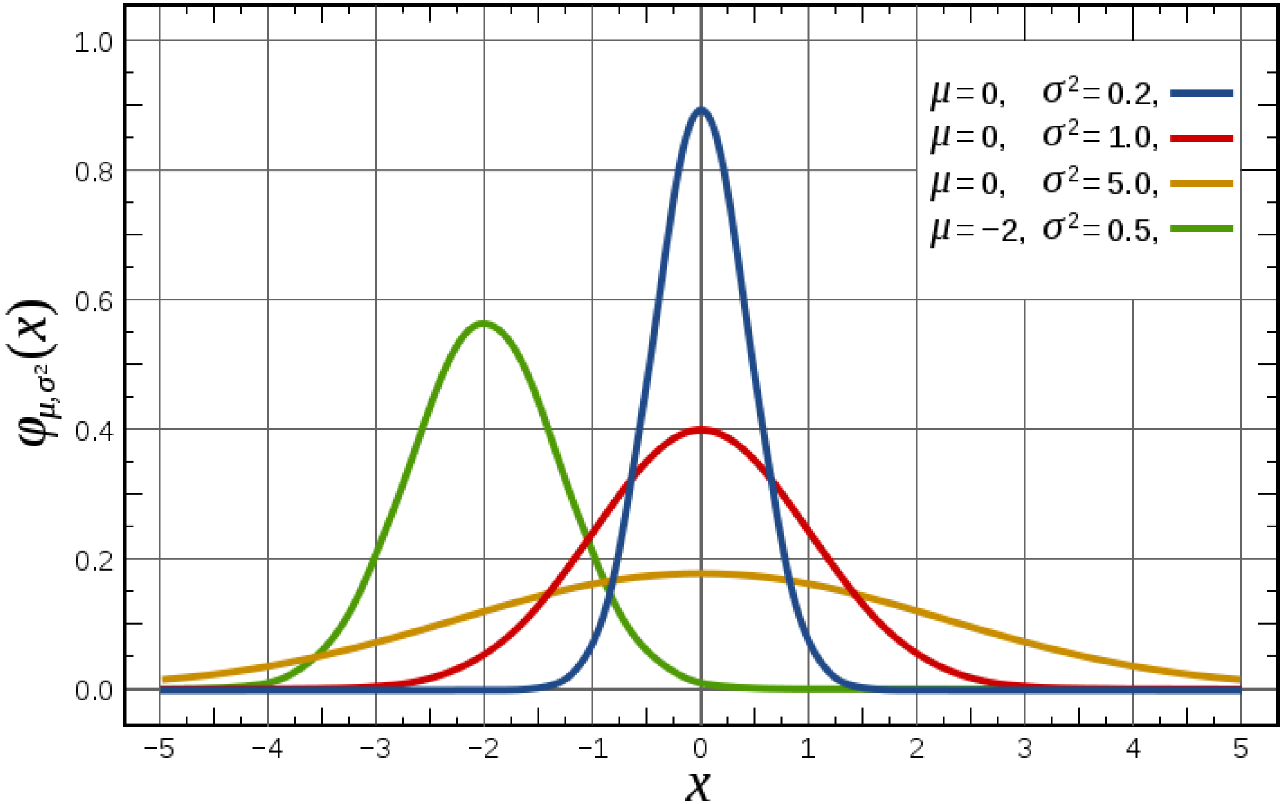
Visualization in 1d



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Normal distribution in 1d: Most common model choice

$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right)$$



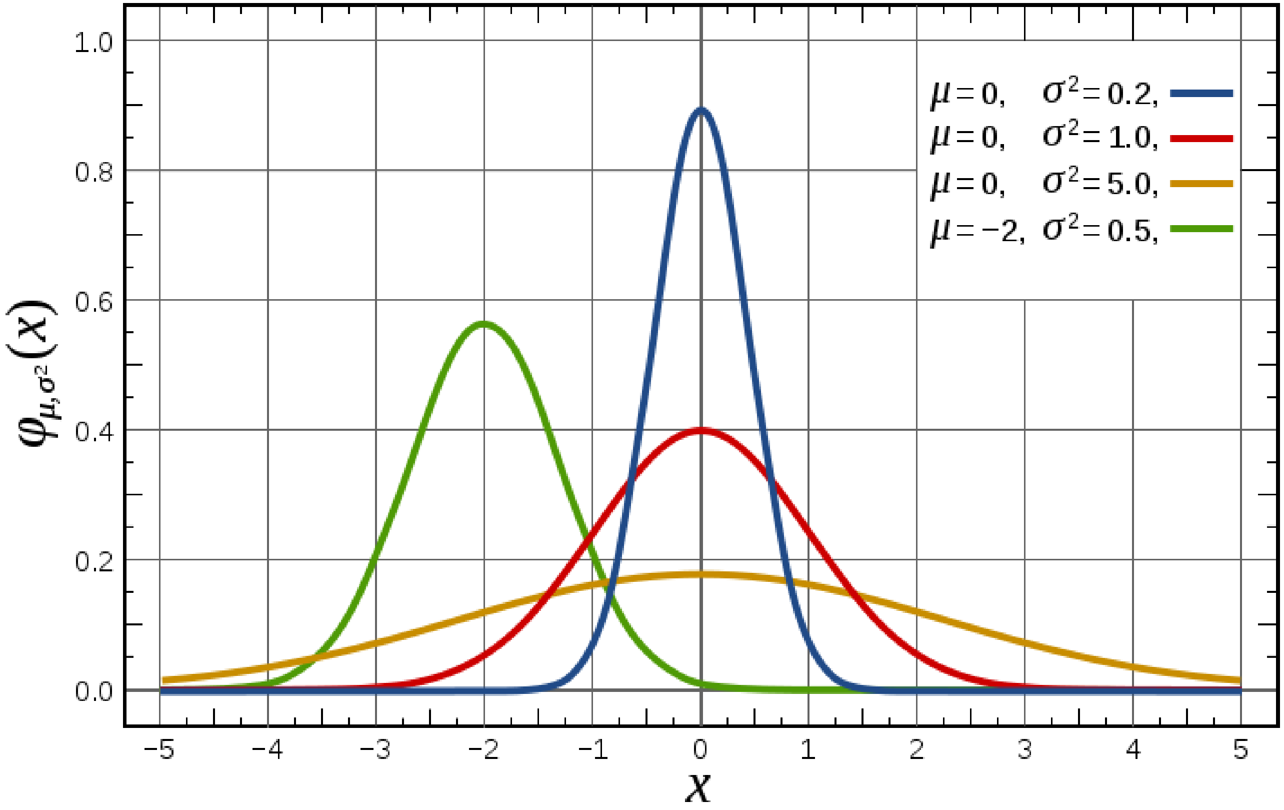
Squared Mahalanobis Distance

=

Sq. Distance from mean in standard deviations

Normal distribution in 1d: Most common model choice

$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \cdot \boxed{\frac{(x-\mu)^2}{\sigma^2}}\right)$$



ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

One variable: Expected value and variance

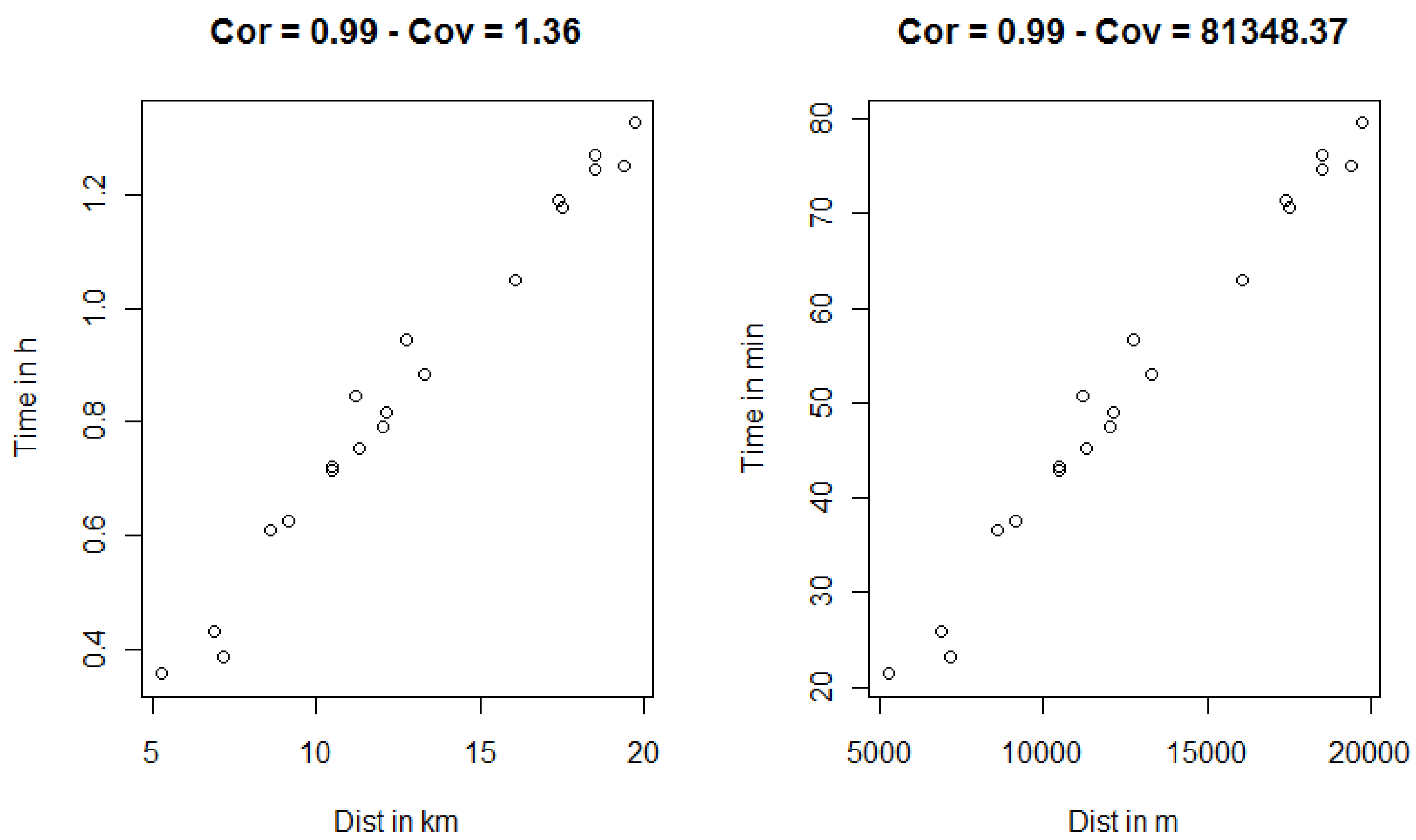
- Expected value: $\mu = E(X) = \int xf(x)dx$
  Estimate: Mean $\hat{\mu} = \overline{x} = \frac{1}{n}\sum x_i$

- Variance:
  $\sigma_X^2 = Var(X) = E\left(\left(X - E(X)\right)^2\right) = \int \left(x - E(X)\right)^2 dx$
  Estimate: Sample Variance
  $$\widehat{\sigma_X^2} = \frac{1}{n-1}\sum(x_i - \overline{x})^2$$
- Standard deviation: $\sigma_X = \sqrt{Var(X)}$
  Estimate: Square root of Sample Variance

Two variables: Covariance and Correlation

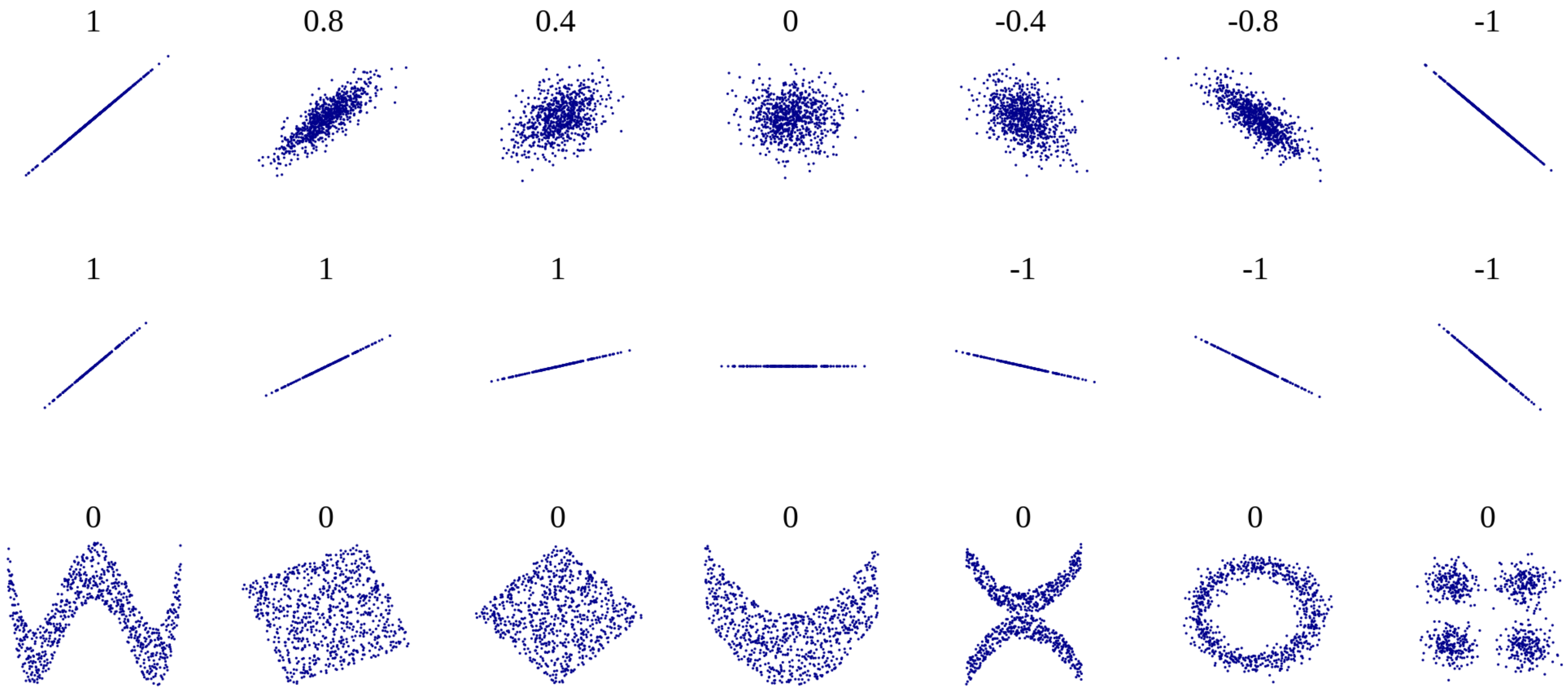
- Covariance: $Cov(X,Y) = E[(X - E[X])(Y - E[Y])] \quad \in [-\infty; \infty]$
- Correlation: $Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \quad \in [-1; 1]$
- Sample covariance: $\widehat{Cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$
- Sample correlation: $r_{xy} = \widehat{Cor}(x,y) = \frac{\widehat{Cov}(x,y)}{\hat{\sigma}_x \hat{\sigma}_y}$
- Correlation is invariant to changes in units, covariance is not (e.g. kilo/gram, meter/kilometer, etc.)

Scatterplot: Correlation is scale invariant



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Intuition and pitfalls for correlation
Correlation = LINEAR relation



Source: Wikipedia

Test for zero correlation: Fisher's z-Test

- X, Y (bivariate) normal distributed with **true correlation** ρ
- Take n samples
- Compute **sample correlation r**

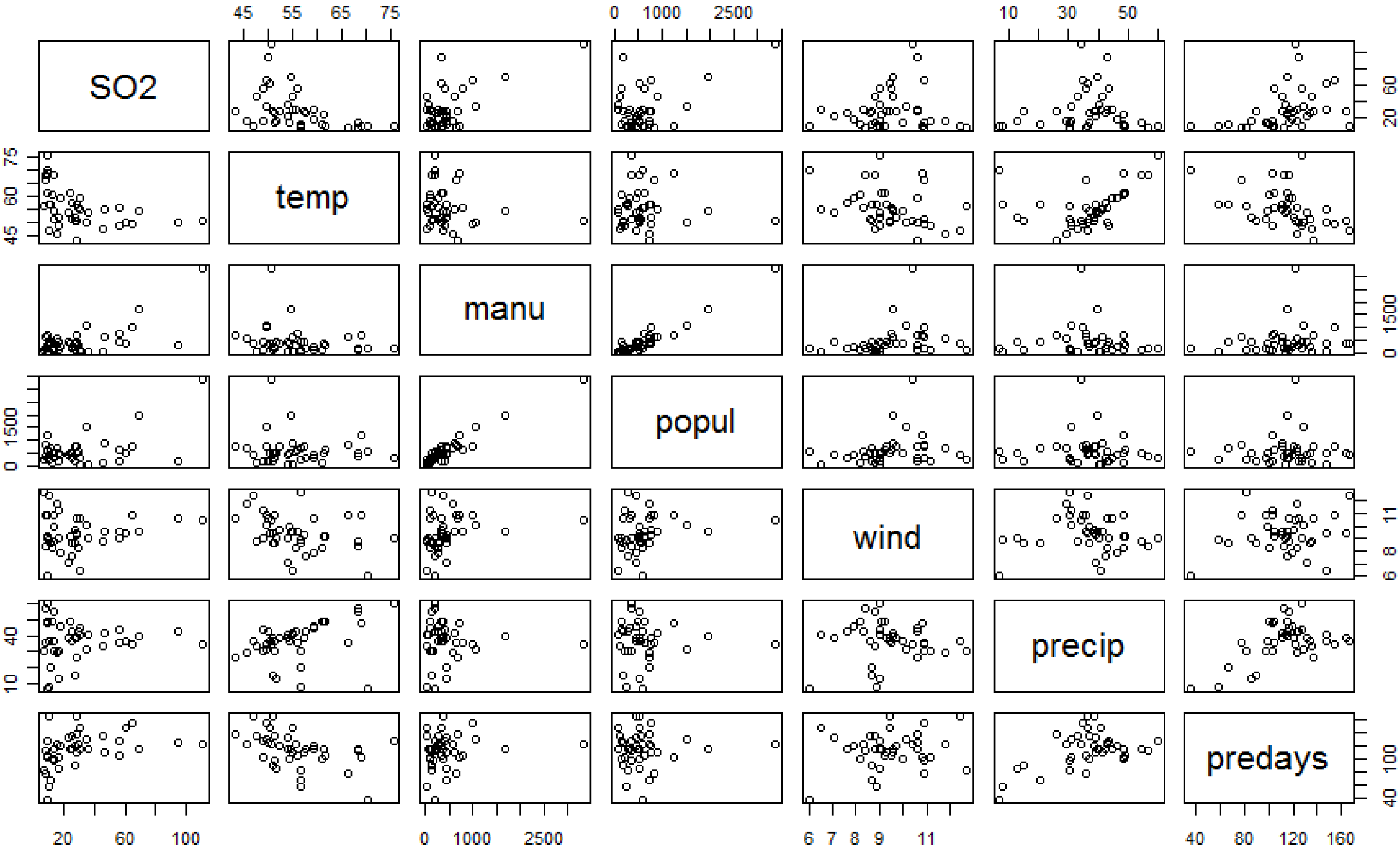
Compute $z = \frac{1}{2} \log\left(\frac{1+r}{1-r}\right)$

Compute $\xi = \frac{1}{2} \log\left(\frac{1+\rho}{1-\rho}\right)$

- For large n: $\sqrt{n-1}(z-\xi) \sim N(0,1)$

- Use cor.test() in R to test and get confidence intervals

Many dimensions: Scatterplot matrix



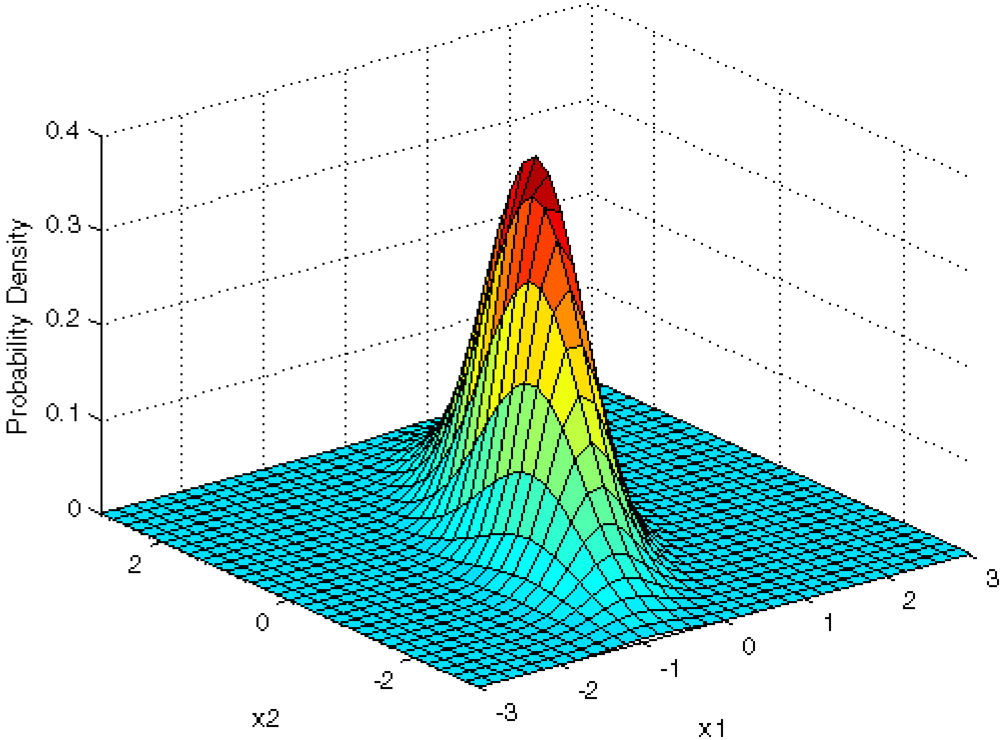
Covariance matrix / correlation matrix: Table of pairwise values

- True covariance matrix: $\Sigma_{ij} = Cov(X_i, X_j)$
- True correlation matrix: $C_{ij} = Cor(X_i, X_j)$

- Sample covariance matrix: $S_{ij} = \widehat{Cov}(x_i, x_j)$
  Diagonal: Variances
- Sample correlation matrix: $R_{ij} = \widehat{Cor}(x_i, x_j)$
  Diagonal: 1

Multivariate Normal Distribution: Most common model choice

$$f(x;\mu,\Sigma) = \frac{1}{(2\pi)^{(p/2)}|\Sigma|^{(1/2)}} \exp\left(-\frac{1}{2}\cdot(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$



Multivariate Normal Distribution: Funny facts

If $X_1, \ldots, X_p$ multivariate normal, then

- every linear combination $Y = a_1 X_1 + \ldots + a_p X_p$ is normally distributed
- every projection on a subspace is multivariate normally distributed

If margins are normally distributed, then it is NOT GUARANTEED that the underlying distribution is multivariate normal

(i.e., “multivariate” is stronger than just normal margins)

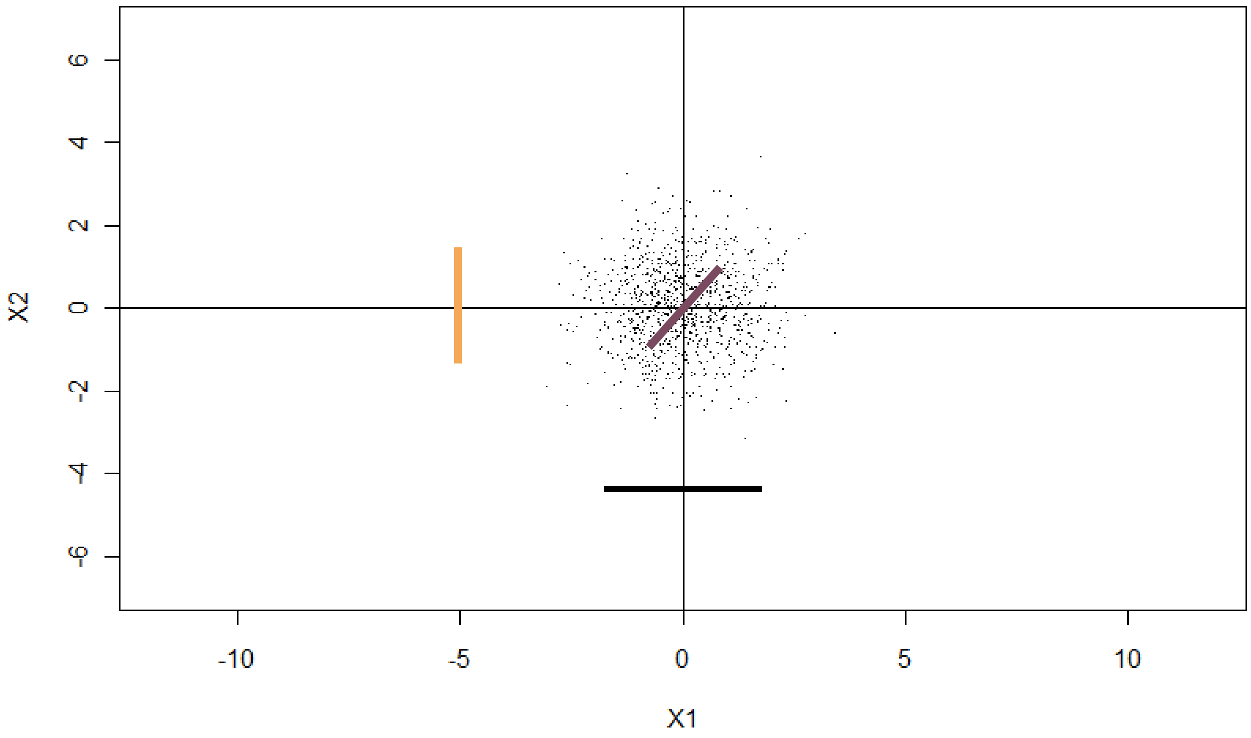
Multivariate Normal Distribution: Example 1
1000 random samples

Variance along X1

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Covariance btw. X1 and X2

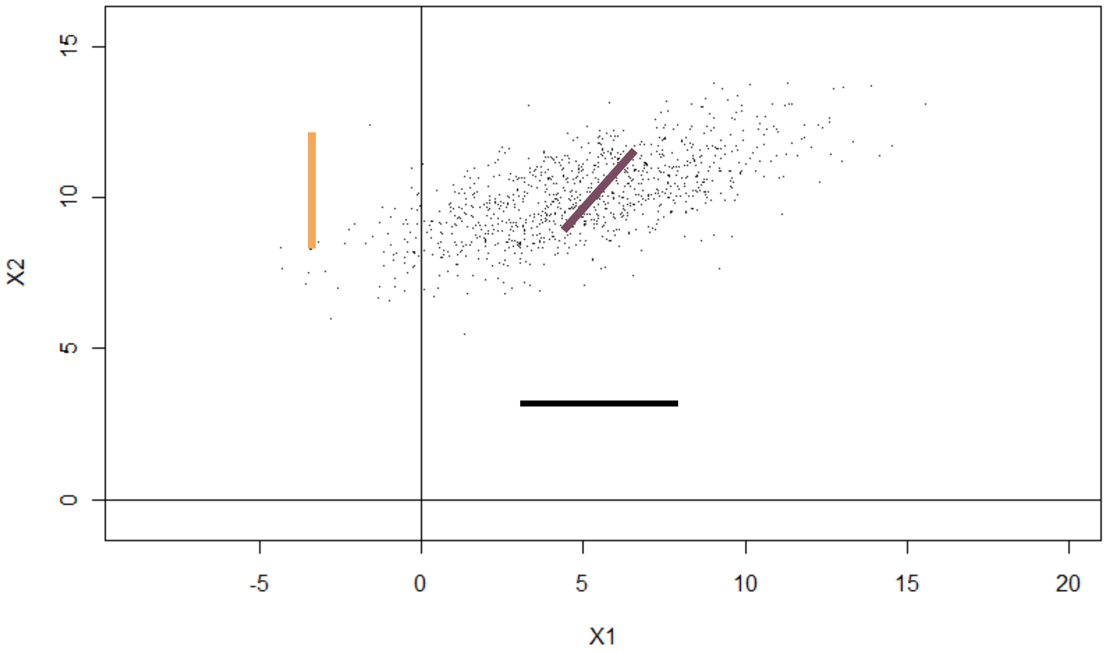
Variance along X2



Variance along X1

Multivariate Normal Distribution: Example 2
1000 random samples

$$\mu = \begin{pmatrix} 5 \\ 10 \end{pmatrix}, \; \Sigma = \begin{pmatrix} 10 & 3 \\ 3 & 2 \end{pmatrix}$$



Covariance btw. X1 and X2

Variance along X2

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Sq. Mahalanobis Distance $MD^2(x)$

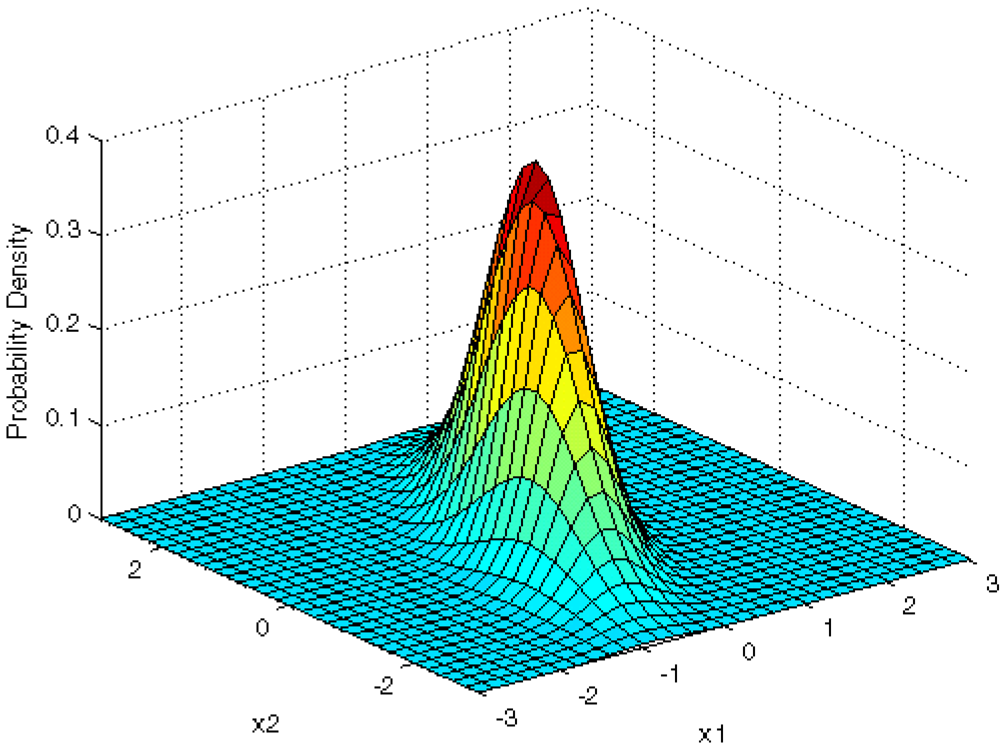
=

Sq. distance from mean in standard deviations

IN DIRECTION OF X

Multivariate Normal Distribution: Most common model choice (p dimensions)

$$f(x;\mu,\Sigma) = \frac{1}{(2\pi)^{(p/2)}|\Sigma|^{(1/2)}} \exp\left(-\frac{1}{2}\cdot \boxed{(x-\mu)^T\Sigma^{-1}(x-\mu)}\right)$$

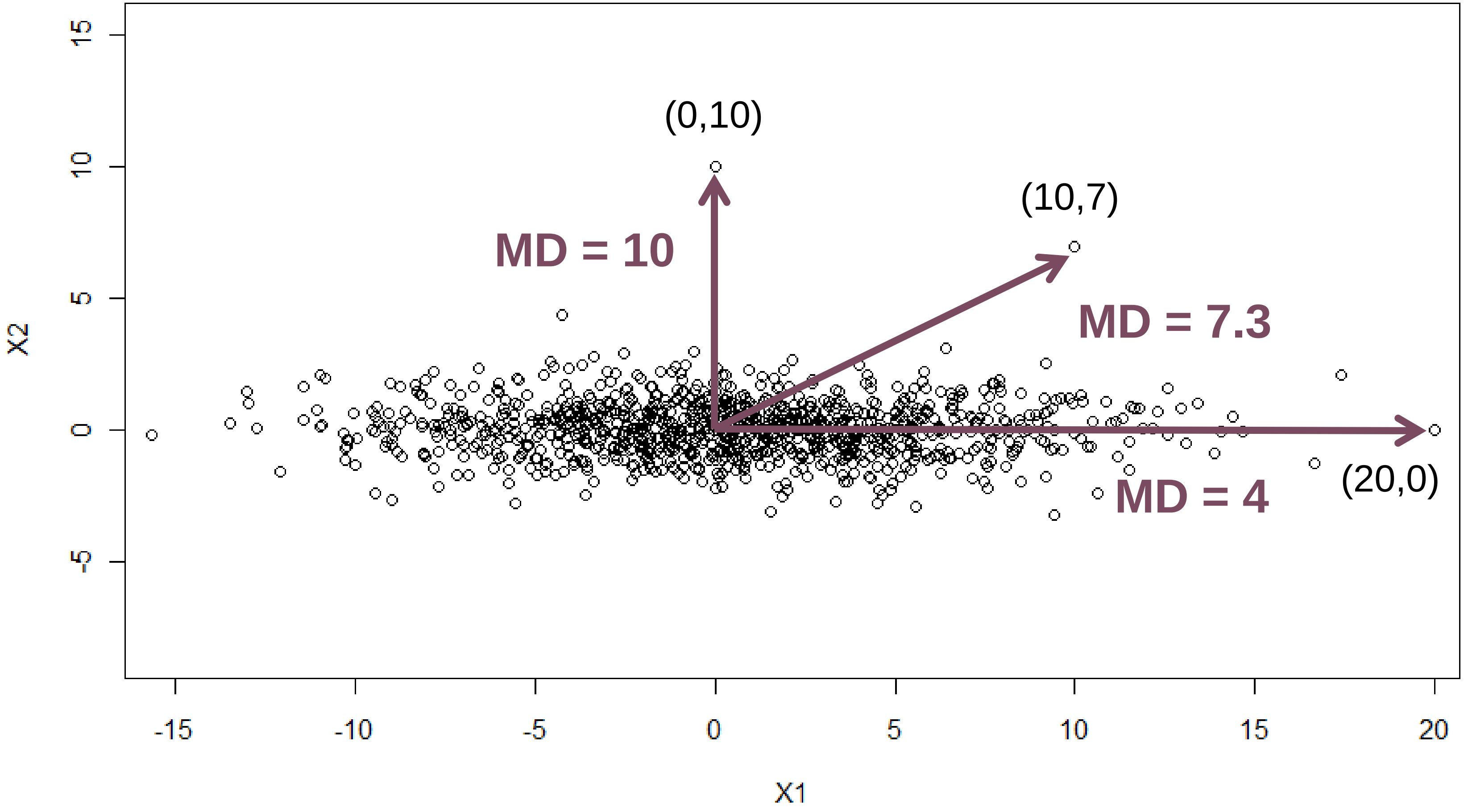


ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

Mahalanobis distance: Example

$$\Sigma = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix}$$



Concepts to know

- Covariance, Correlation (true / sample version)
- Test for zero correlation: Fisher's z-Transformation
- Scatterplot / Scatterplotmatrix
- Covariance matrix / Correlation matrix
- Multivariate Normal Distribution
- Mahalanobis distance

R commands to know

- read.csv, head, str, dim
- colMeans, cov, cor
- mvrnorm, t, solve, %*%