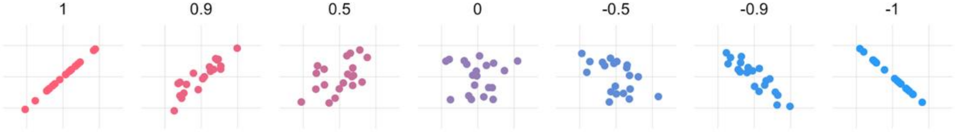
COMS20017 – Algorithms & Data



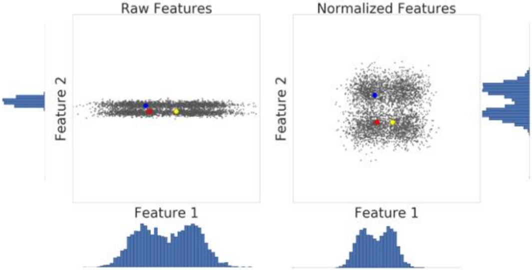
September 2025

Majid Mirmehdi

with some slides from Rui Ponte Costa & Dima Damen

Lecture MM-03

This lecture



- Data acquisition
- Data characteristics: distance measures
- **Data characteristics: summary statistics**
- **Data normalisation and outliers**

Mean and Variance

For one-dimensional data $\mathbf{x} = \{x_1, .., x_n\}$,

Mean: [average]

$$\mu = \frac{1}{N} \sum_i x_i$$

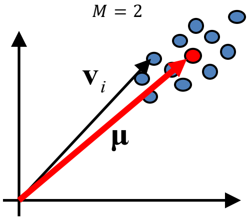
Variance: [spread]

$$\sigma^2 = \frac{1}{N-1} \sum_i (x_i - \mu)^2$$

Standard Deviation:

$$\sigma = \sqrt{\frac{1}{N-1} \sum_i (x_i - \mu)^2}$$

Mean and Covariance



For multi-dimensional data:

e.g. M dimensions with $\{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$, i.e there are N vectors/datapoints where each vector has M elements.

Mean vector:

Computed independently for each dimension

$$\boldsymbol{\mu} = \frac{1}{N} \sum_i \mathbf{v}_i$$

Covariance:

Gives spread and how variables change together

$$\mathbf{C} = \frac{1}{N-1} \sum_i (\mathbf{v}_i - \boldsymbol{\mu})^2$$

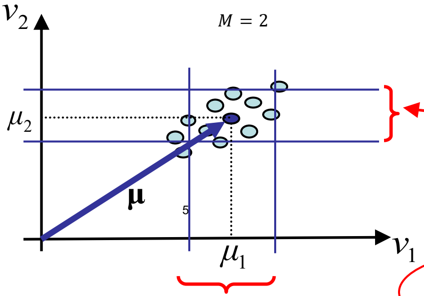
$$\mathbf{C} = \frac{1}{N-1} \sum_i (\mathbf{v}_i - \boldsymbol{\mu})^{\mathrm{T}} (\mathbf{v}_i - \boldsymbol{\mu})$$

$$\mathbf{C} = \frac{1}{N-1} \sum_i \begin{bmatrix} (v_{i1} - \mu_1)^2 & (v_{i1} - \mu_1)(v_{i2} - \mu_2) \\ (v_{i1} - \mu_1)(v_{i2} - \mu_2) & (v_{i2} - \mu_2)^2 \end{bmatrix}$$

N when the population mean is known, N-1 when not!

Mean and Covariance

$$\boldsymbol{\mu} = \frac{1}{N}\sum_i \mathbf{v}_i$$



Covariance between components

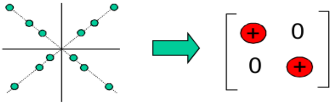
Variance along 1st dimension

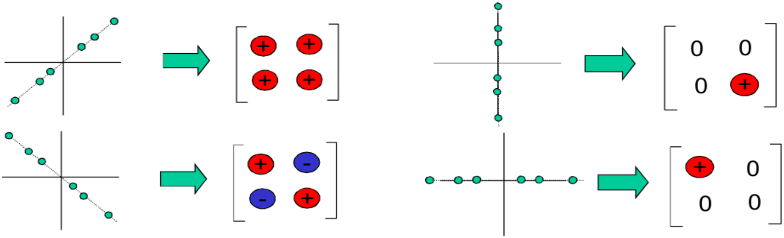
Variance along 2nd dimension

$$\mathbf{C} = \frac{1}{N-1}\sum_i \begin{bmatrix} \boxed{(v_{i1} - \mu_1)^2} & (v_{i1} - \mu_1)(v_{i2} - \mu_2) \\ \boxed{(v_{i1} - \mu_1)(v_{i2} - \mu_2)} & \boxed{(v_{i2} - \mu_2)^2} \end{bmatrix}$$

Covariance Matrix

$$\mathbf{C} = \frac{1}{N}\sum_i \begin{bmatrix} (v_{i1} - \mu_1)^2 & (v_{i1} - \mu_1)(v_{i2} - \mu_2) \\ (v_{i1} - \mu_1)(v_{i2} - \mu_2) & (v_{i2} - \mu_2)^2 \end{bmatrix}$$

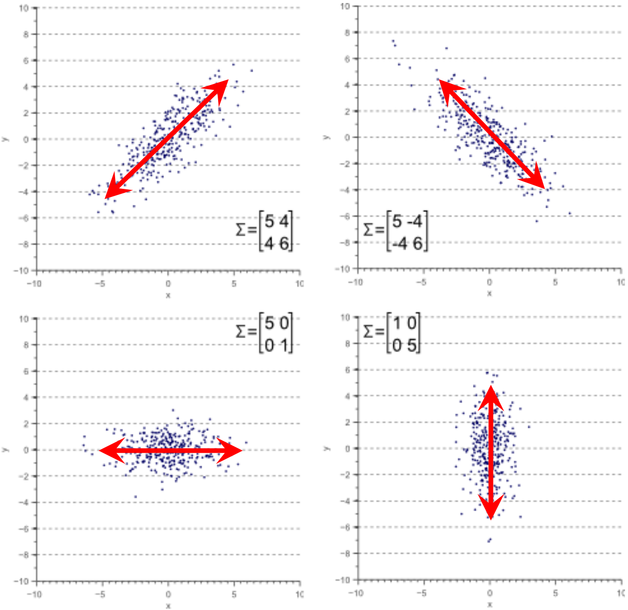




Spread and Covariance

- The shape of the data is defined by the covariance matrix.

- Diagonal spread is captured by the covariance, while axis-aligned spread is captured by the variance.



Covariance Matrix

In three dimensions,

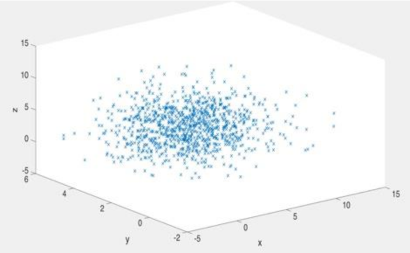
$$\mathbf{C} = \frac{1}{N}\sum_i \begin{bmatrix} (v_{i1}-\mu_1)^2 & (v_{i1}-\mu_1)(v_{i2}-\mu_2) & (v_{i1}-\mu_1)(v_{i3}-\mu_3) \\ (v_{i1}-\mu_1)(v_{i2}-\mu_2) & (v_{i2}-\mu_2)^2 & (v_{i2}-\mu_2)(v_{i3}-\mu_3) \\ (v_{i1}-\mu_1)(v_{i3}-\mu_3) & (v_{i2}-\mu_2)(v_{i3}-\mu_3) & (v_{i3}-\mu_3)^2 \end{bmatrix}$$

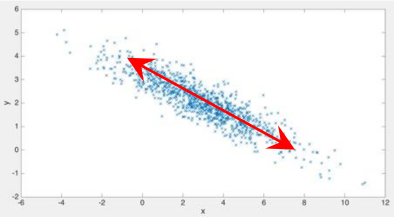
A Covariance matrix is always:

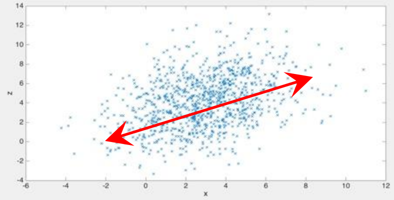
- square
- symmetric
- variances on the diagonal
- covariance between each pair of dimensions in non-diagonal elements

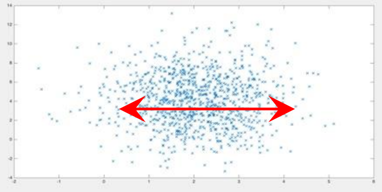
Covariance Matrix example

For the covariance matrix, $$\mathbf{c} = \begin{bmatrix} 5 & -2 & 2 \\ -2 & 1 & 0 \\ 2 & 0 & 7 \end{bmatrix}$$









Correlation

- We are often interested in the degree of similarity between two sequences in terms of their variation (independent of their absolute values)

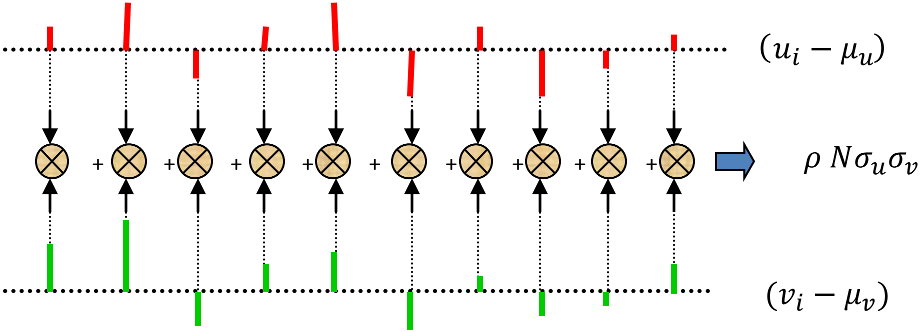
- Correlation of two sequences u and v of length N

$$\rho = \frac{1}{N\sigma_u\sigma_v} \sum_{i=0}^{N-1} (u_i - \mu_u)\,(v_i - \mu_v)$$

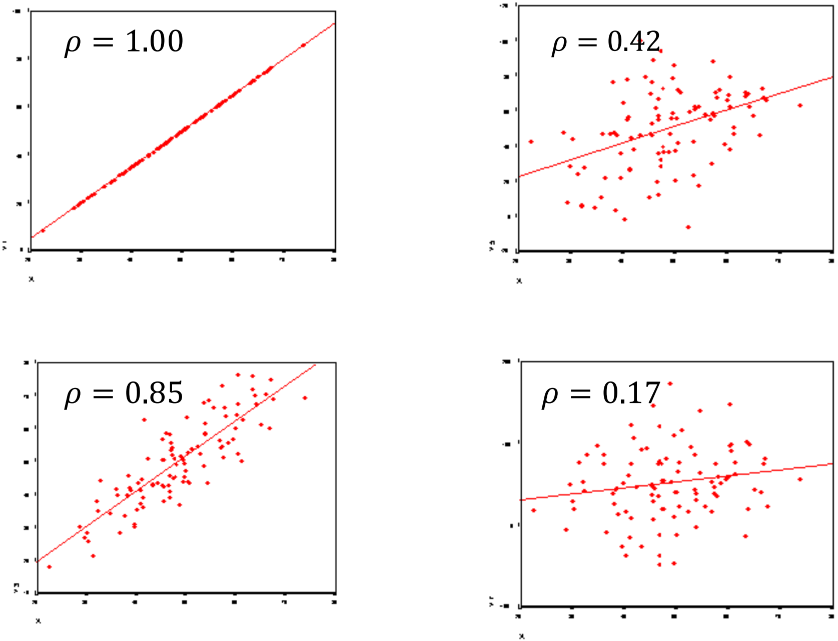
Division by variance product $\sigma_u\sigma_v$ normalises measure to be independent of absolute value and is unitless → captures similarity in variation (or *structure*).

Example: Correlating Coefficient for Two Sequences

$$\rho = \frac{1}{N\sigma_u\sigma_v}\sum_{i=0}^{N-1}(u_i - \mu_u)\,(v_i - \mu_v)$$



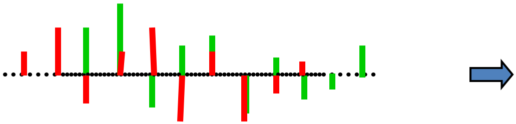
Correlation Example

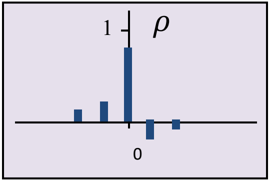


Correlation – shifting similarity

- Sometimes sequences are similar after applying a shift.
- Can be measured with **cross-correlation**

$$\rho = \frac{1}{N\sigma_u\sigma_v}\sum_{i=0}^{N-1}(u_i - \mu_u)(v_{i-j} - \mu_v)$$





'most likely shift' given by position j of maximum value in cross-correlation

Self Study: Correlation and Covariance

MEANING AND DEFINITION

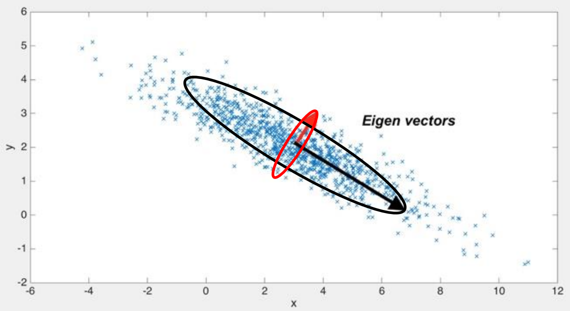
CORRELATION	COVARIANCE
There is said to be correlation between two, when change in one results in change in another.	Covariance talks about the direction of the relationship between the two variables (positive or negative)

CORRELATION	COVARIANCE
Measures the strength of the variables under comparison	Measures the extent of change in one with regards to change in another.
Correlation is a scaled down version of covariance.	Covariance is considered as a part of correlation.
Value here lies between -1 and +1.	Value here lies between -infinity to +infinity
Correlation is a unit-free measure	Covariance value is the product of the units of the variables.
There would be no change in correlation due to scale.	Any change in scale affects covariance.

https://efinancemanagement.com/investment-decisions/correlation-vs-covariance

Eigen analysis

- Eigenvectors and eigenvalues define the principal axes and spread of points along directions
- **Major axis** - eigenvector corresponding to larger eigenvalue (i.e. larger variance)
- **Minor axis** - eigenvector corresponding to smaller eigenvalue (i.e. smaller variance)
- These can be represented using major and minor axes of ellipses



Eigen analysis

Definition

For a square matrix $\mathbf{C}$,
if there exists a non-zero column vector $\mathbf{v}$ where

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

then,

$v \rightarrow$ eigenvector of matrix C
$\lambda \rightarrow$ eigenvalue of matrix C

e.g. $\mathbf{C} = \begin{bmatrix} 0 & -1 \\ 2 & 3 \end{bmatrix}, v_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \lambda_1 = 1$

Eigen analysis

- To calculate eigenvectors of a square matrix, solve $\mathbf{C}v = \lambda v$

$$\longrightarrow \quad |\mathbf{C} - \lambda \boldsymbol{I}| = 0$$

where

- $\boldsymbol{I}$ is the identity matrix
- $|\mathbf{C}|$ is the determinant of the matrix

For 2×2 matrices, there are two eigenvalues λ_1, λ_2

$$\mathbf{C} - \lambda \boldsymbol{I} = \begin{bmatrix} 0 & -1 \\ 2 & 3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = \begin{bmatrix} -\lambda & -1 \\ 2 & 3-\lambda \end{bmatrix}$$

$$|\mathbf{C} - \lambda \boldsymbol{I}| = \lambda^2 - 3\lambda + 2 = (\lambda - 1)(\lambda - 2)$$

$$\lambda_1 = 1 \quad and \quad \lambda_2 = 2$$

Eigen analysis

- After the eigenvalues are found, the eigenvectors can be calculated

For $\lambda_1 = 1$

$$\begin{bmatrix} 0 & -1 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} \tag{2}$$

- This simplifies to:

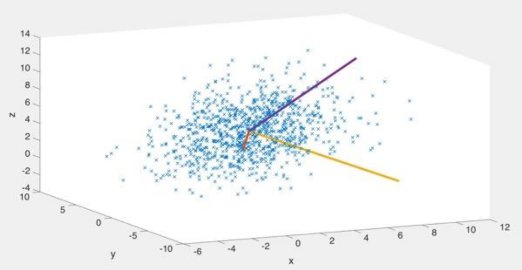
$$\begin{bmatrix} -v_{12} \\ 2v_{11} + 3v_{12} \end{bmatrix} = \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} \tag{3}$$

- If we set $v_{12} = 1$, then we get the eigenvector:

$$\begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \tag{4}$$

- Verify that this is indeed a valid eigenvector by calculating $\mathbf{C}v = \lambda v$

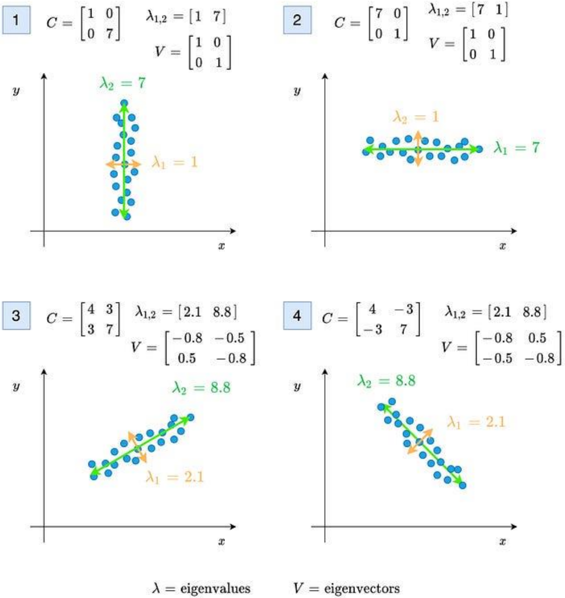
3D example



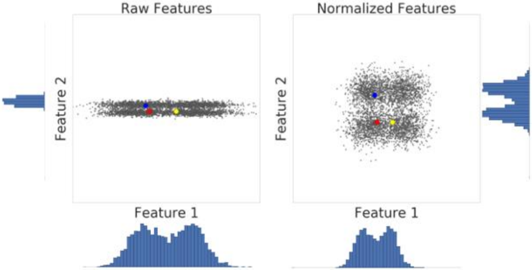
- Eigenvalues → $\lambda_1 = 0.08$ $\lambda_2 = 4.52$ $\lambda_3 = 8.40$
- Eigenvectors → $v_1 = \begin{bmatrix} -0.42 \\ -0.90 \\ 0.12 \end{bmatrix}$ $v_2 = \begin{bmatrix} 0.71 \\ -0.40 \\ -0.57 \end{bmatrix}$ $v_3 = \begin{bmatrix} 0.57 \\ -0.15 \\ 0.81 \end{bmatrix}$
- Principal/Major axis is v_3 (corresponding to the largest eigenvalue)

Eigenvalue and Variance

If the covariance matrix of the data is a diagonal matrix, i.e the covariances are zero, the variances are equal to the eigenvalues.



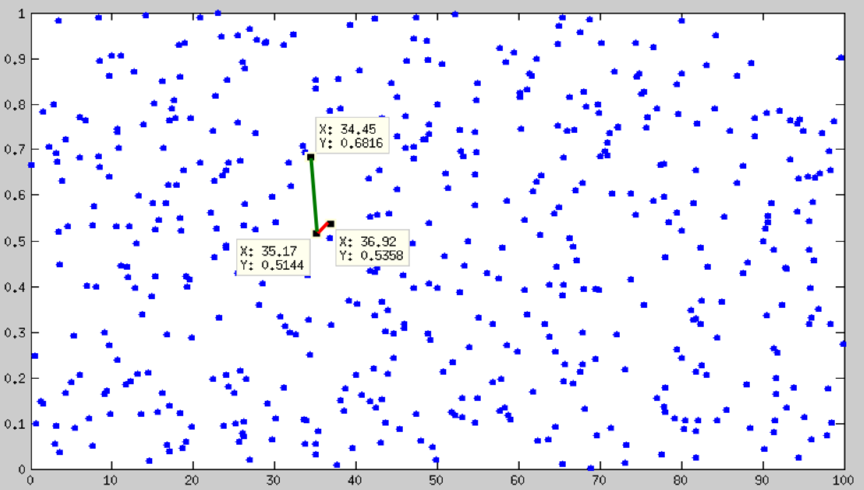
Next



- Data acquisition
- Data characteristics: distance measures
- Data characteristics: summary statistics
- **Data normalisation and outliers**

Data Characteristic - Data Normalisation

- Note the difference in magnitude between the two dimensions below!
- Data may need to be normalised before distance is calculated



Data Characteristic - Data Normalisation

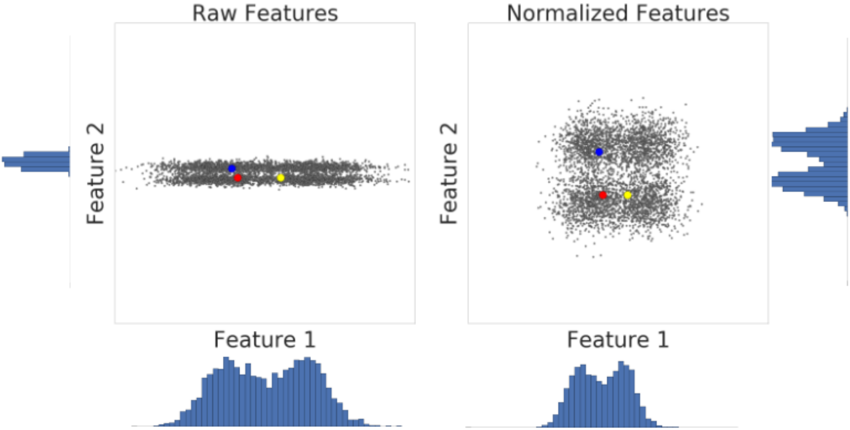
- Methods for normalisation:

1. Rescaling
$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$
rescales the range of features to [0, 1]

2. Standardisation (also known as *z*-score)
$$x' = \frac{x - \mu}{\sigma}$$
makes the values of each feature in the data have zero-mean and unit-variance

3. Scaling to unit length
$$x' = \frac{x}{||\mathbf{x}||}$$
scales components of feature vector so that the complete vector has length one

Normalisation Example



Brief return to Distance Measures

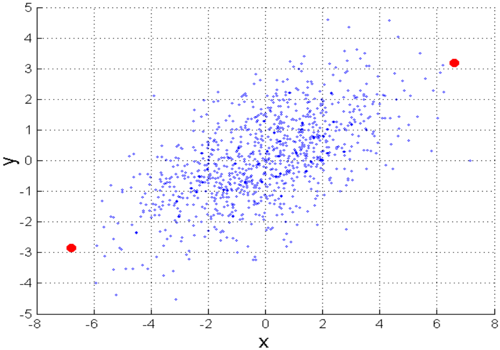
Mahalanobis Distance is a measure of distance between a data vector and a set of data, or a variation that measures the distance between two vectors from the same dataset:

$$mahalanobis(a,b) = \sqrt{(a-b)^T \Sigma^{-1} (a-b)}$$

$$\text{where } \operatorname{cov}(X,Y) = \Sigma = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

Warning: Σ is the covariance matrix of the input data D

For red points, the Euclidean distance is 14.7, and the Mahalanobis distance is 6.



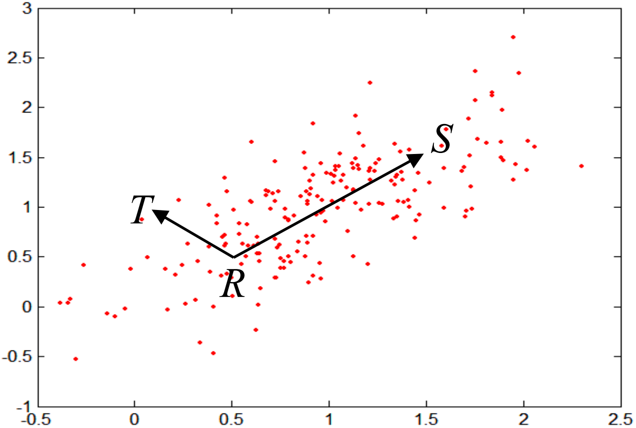
Brief return to Distance Measures

Mahalanobis Distance example:

Given $R = (0.5, 0.5)$, $S = (1.5, 1.5)$, $T = (0.0, 1.0)$, find the Mahalanobis distance RS and RT.

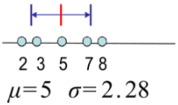
$$\sum = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

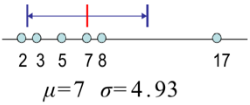
$$RS = 2$$
$$RT = \sqrt{5}$$

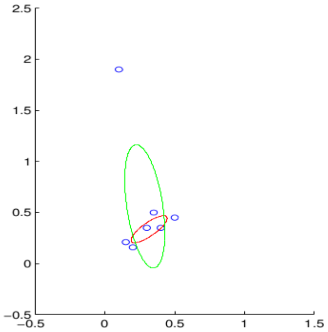


Data Characteristics - Outliers

- Mean, variance and covariance can provide concise description of ‘average’ and ‘spread’, but not when outliers are present in the data
- **outliers**: An *outlier* is an observation that lies an abnormal distance from other values in a random sample from a population.
- usually due to fault in measurement and not always easy to remove





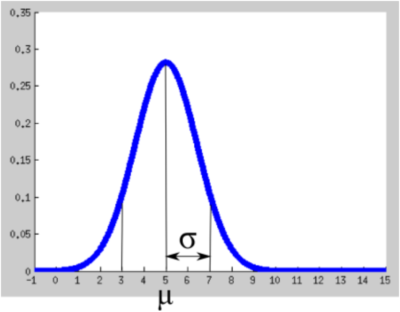


Normal or Gaussian Distribution (Reminder)

For a normal distribution $N(\mu, \sigma^2)$ in one dimension, the probability density function (pdf) can be calculated as:

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

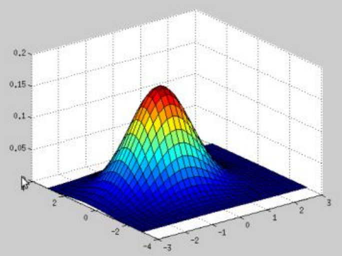
68% of data within 1σ of μ
92% within 2σ of μ
99% within 3σ of μ

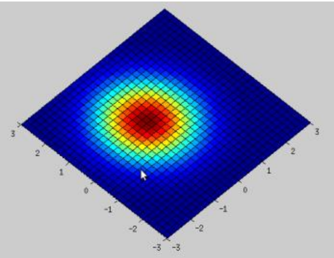


Normal Distribution - Multi-dimensional (reminder)

For multi-dimensional normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the probability density function (pdf) can be calculated as

$$p(\mathbf{x}) = \frac{1}{2\pi\|\boldsymbol{\Sigma}\|^{1/2}} \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))$$





WARNING: Σ **here is the capital letter of** σ **and represents the covariance matrix. It is NOT the summation sign!**

Some of the topics next in COMS20017

- Least Squares and Regression
- Clustering data
- Classification of data
- The Fourier transform
- Principal Components Analysis
- Convolutions