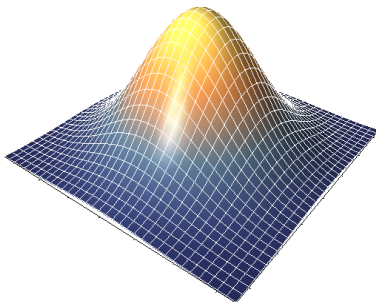


COMS20011 – Data-Driven Computer Science



February 2021

Majid Mirmehdi, R Ponte Costa, D Damen, A Calway

This lecture



Analog Signal



Digital Signal

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

Mean and Variance

For one-dimensional data $\mathbf{v} = \{v_1, \dots, v_n\}$,

Mean: [average]

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

Variance: [spread]

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

Standard Deviation:

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

Mean and Covariance

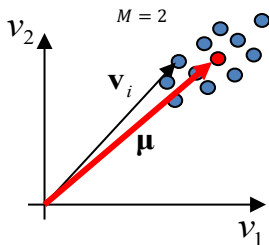
For multi-dimensional data:

e.g. M dimensions with $\{\mathbf{v}_1, \dots, \mathbf{v}_N\}$, 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:

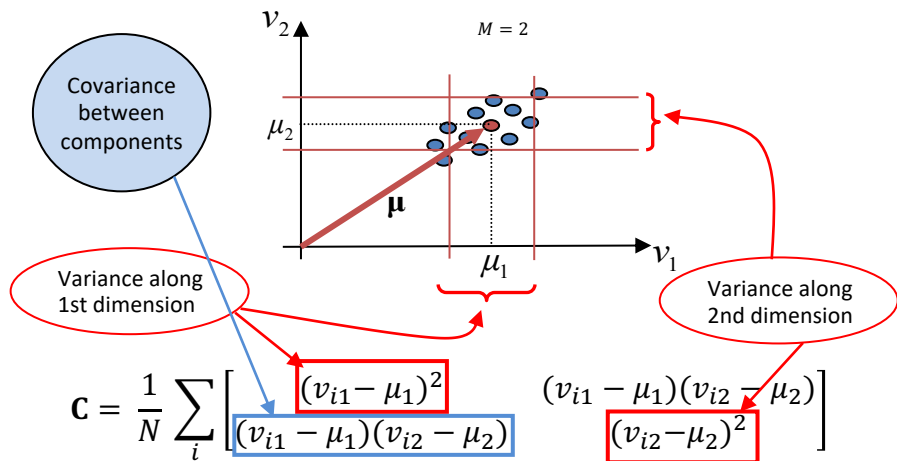
Gives both spread and
correlation

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

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

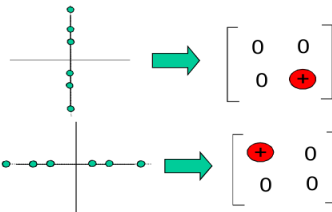
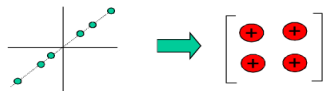
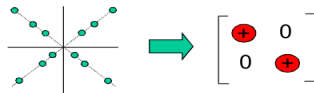
$$\mathbf{C} = \frac{1}{N} \sum_i \begin{bmatrix} (v_{i1} - \mu_1)^2 & (v_{i1} - \mu_1)(v_{i2} - \mu_2) \\ (v_{i1} - \mu_1)(v_{i2} - \mu_2) & (v_{i2} - \mu_2)^2 \end{bmatrix}$$

Mean and Covariance



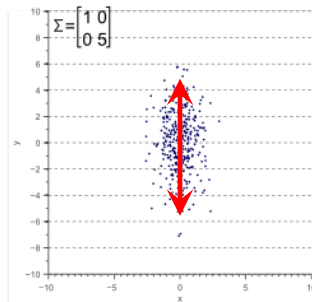
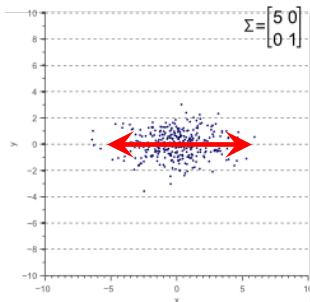
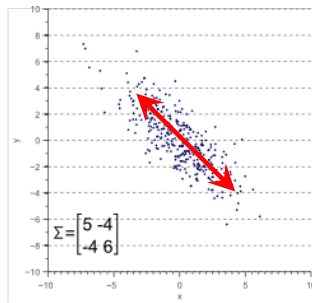
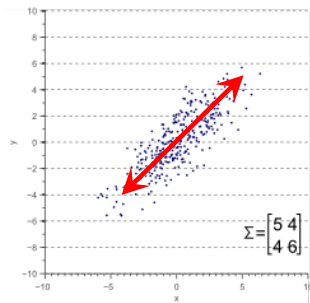
Covariance Matrix

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



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,

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

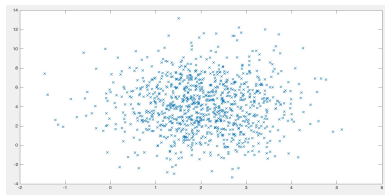
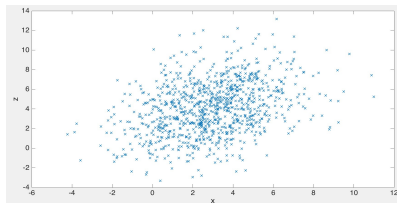
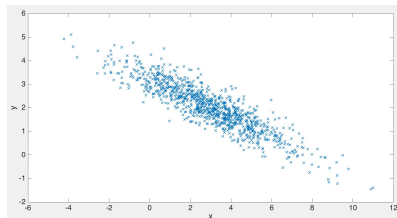
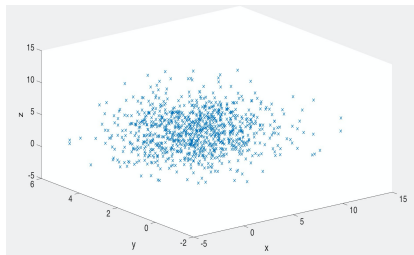
A Covariance matrix is always:

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

Covariance Matrix example

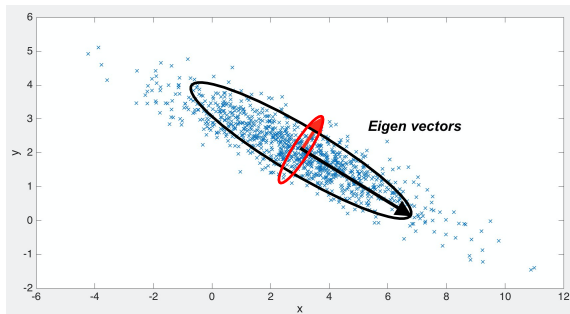
For the covariance matrix,

$$\mathbf{c} = \begin{bmatrix} 5 & -2 & 2 \\ -2 & 1 & 0 \\ 2 & 0 & 7 \end{bmatrix}$$



Covariance Matrix: Eigen analysis

- Eigenvectors and eigenvalues define **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



Covariance Matrix: Eigen analysis

Definition

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

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

then,

$\mathbf{v} \rightarrow$ eigenvector of matrix \mathbf{C}

$\lambda \rightarrow$ eigenvalue of matrix \mathbf{C}

e.g.

$$\mathbf{C} = \begin{bmatrix} 0 & -1 \\ 2 & 3 \end{bmatrix}, \mathbf{v}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \lambda_1 = 1$$

Covariance Matrix: Eigen analysis

- ▶ To calculate eigenvectors of a square matrix, e.g. a covariance matrix, then solve

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

where

- ▶ \mathbf{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

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

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

$$\lambda_1 = 1, \lambda_2 = 2$$

Covariance Matrix: Eigen analysis

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

For $\lambda_1 = 1$

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

- ▶ This simplifies to:

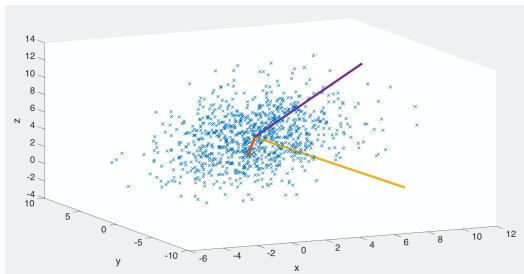
$$\begin{bmatrix} -v_{12} \\ 2v_{11} + 3v_{12} \end{bmatrix} = \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} \quad (3)$$

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

$$\begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad (4)$$

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

Covariance Matrix: another example



➤ Eigenvalues $\rightarrow \lambda_1 = 0.08 \quad \lambda_2 = 4.52 \quad \lambda_3 = 8.40$

➤ Eigenvectors $\rightarrow v_1 = \begin{bmatrix} -0.42 \\ -0.90 \\ 0.12 \end{bmatrix} \quad v_2 = \begin{bmatrix} 0.71 \\ -0.40 \\ -0.57 \end{bmatrix} \quad v_3 = \begin{bmatrix} 0.57 \\ -0.15 \\ 0.81 \end{bmatrix}$

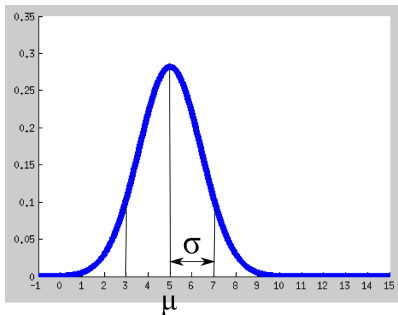
➤ Principal/Major axis is v_3 (corresponding to the largest eigenvalue)

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:

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

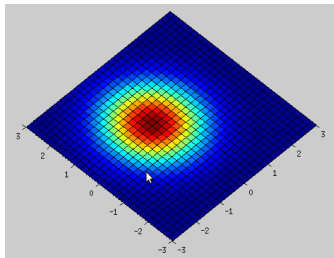
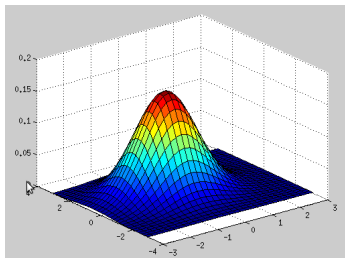
68% of data within 1σ of μ
92% within 2σ of μ
99% within 3σ of μ



Normal Distribution - Multi-dimensional (reminder)

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

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



WARNING: Σ is the capital letter of σ , not the summation sign!
So here Σ is the covariance matrix.

This lecture



Analog Signal

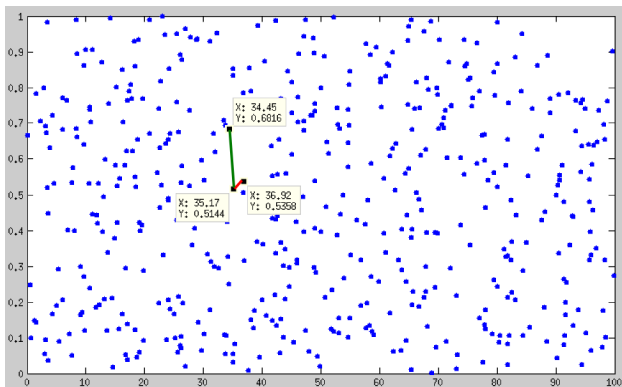


Digital Signal

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

Data Characteristic - Data Normalisation

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



Data Characteristic - Data Normalisation

► Methods for normalisation:

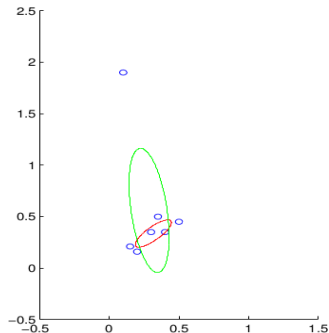
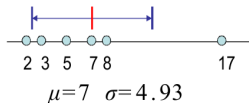
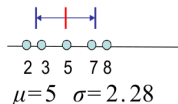
1. Rescaling
$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

2. Standardisation (also known as z-score)
$$x' = \frac{x - \mu}{\sigma}$$

3. Scaling to unit length
$$x' = \frac{x}{\|x\|}$$

Data Characteristic - Outliers

- Mean, variance and covariance can provide concise description of 'average' and 'spread', but not when outliers are present in the data
- **outliers**: small number of points with values significantly different from that of the other points
- usually due to fault in measurement and not always easy to remove



Concluding initial part of COMS20011



Analog Signal



Digital Signal

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