COMS21202: Symbols, Patterns and Signals
Data Acquisition and Data Characteristics

[based on Dima Damen lecture notes]

Rui Ponte Costa
rui.costa@bristol.ac.uk

University of Bristol, Department of Computer Science
Bristol BS8 1UB, UK

February 3, 2020

Agenda

Today we are going to talk about

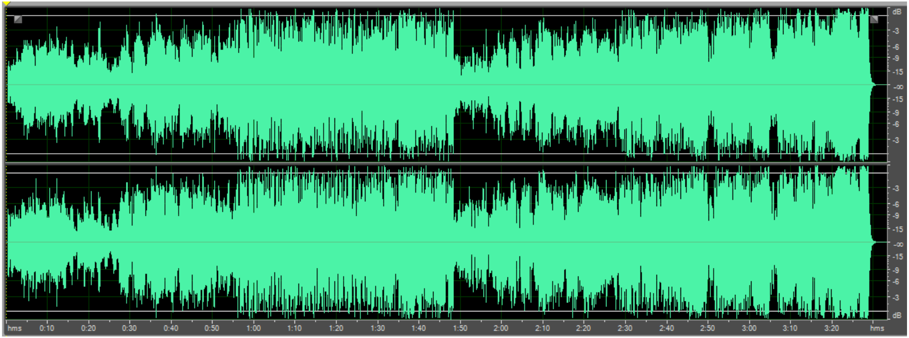
1. Data acquisition
2. Data characteristics: distance measures
3. Data characteristics: summary statistics [reminder]
4. Data normalisation and outliers

Data Acquisition - Analogue to Digital Conversion

Analogue to Digital conversion involves

1. Sampling
2. Quantisation

e.g. Audio Signal - 1D [low zoom]

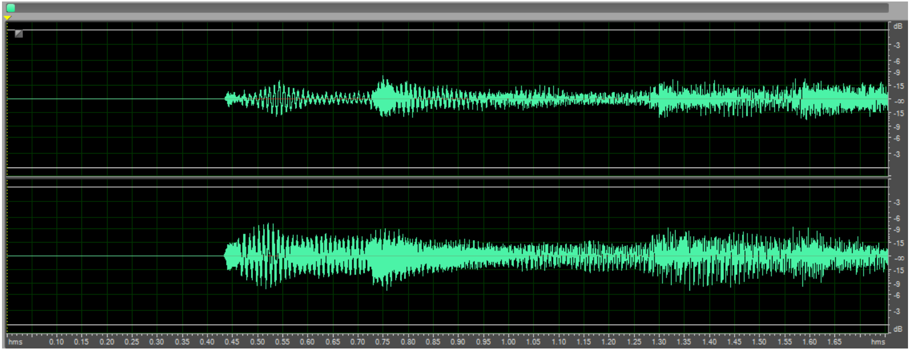


Data Acquisition - Analogue to Digital Conversion

Analogue to Digital conversion involves

1. Sampling
2. Quantisation

e.g. Audio Signal - 1D [medium zoom]

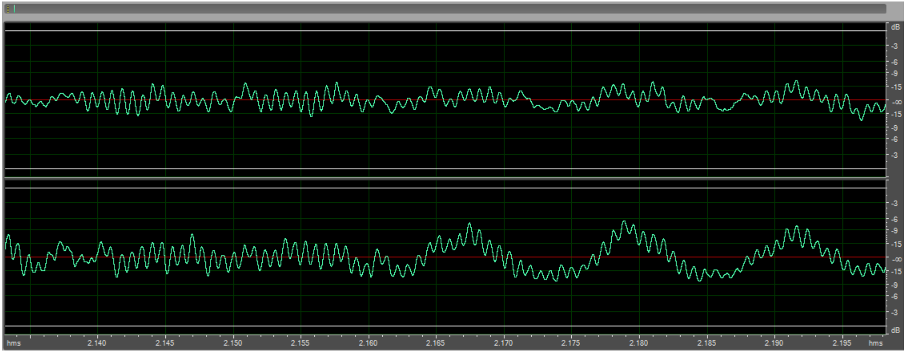


Data Acquisition - Analogue to Digital Conversion

Analogue to Digital conversion involves

1. Sampling
2. Quantisation

e.g. Audio Signal - 1D [high zoom] : How do you represent data digitally?

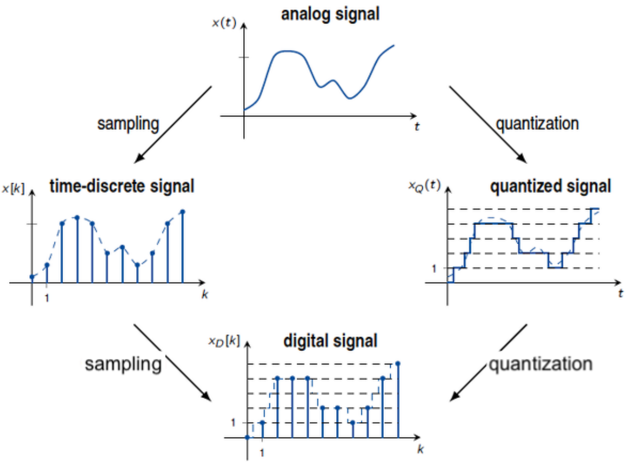


Data Acquisition - Analogue to Digital Conversion

You need to:

1. Sample
2. Quantise

example from dsp-nbsphinx.readthedocs.io (chapter 5.1)



Data Acquisition - Analogue to Digital Conversion

Theorem

Nyquist-Shannon sampling theorem:
If a function $x(t)$ contains no frequencies higher than f_{max} hertz, it is completely determined by giving its ordinates at a series of points spaced $\frac{1}{2f_{max}}$ seconds apart.

In other words,

- Suppose the highest frequency for a given analog signal is f_{max},
- According to the Theorem:
  sampling period, $T_s \leq \frac{1}{2f_{max}}$
  which is equivalent to *sampling rate*, $f_s \geq 2f_{max}$

Data Acquisition - Analogue to Digital Conversion

Examples of sampling and quantisation of standard audio formats

- Speech (e.g. phone call)
  - Sampling: 8 KHz samples
  - Quantisation: 8 bits / sample
- Audio CD
  - Sampling: 44 KHz samples
  - Quantisation: 16 bits / sample
  - Stereo (2 channels)

Note: Higher sampling/quantisation achieves better signal quality, but also larger memory/storage.

Data Acquisition - Analogue to Digital Conversion

Images - Multi-Dimensional

- Sampling: Resolution in digital photography
- Quantisation: Representation of each pixel in the image

- 8 Mega Pixel Camera - 3264x2448 pixels
- Quantisation 8 bits per colour
- Colour images: 3 channels: Red, Green, Blue

- Greyscale images: 1 channel: intensity $= \frac{R+G+B}{3}$

- Binary Images: Black/White 1 bit per pixel

Agenda

Today we are going to talk about

1. Data acquisition
2. **Data characteristics: distance measures**
3. Data characteristics: summary statistics [reminder]
4. Data normalisation and outliers

Distance

- Distance is measure of separation between data.
- Can be defined between single-dimensional data, multi-dimensional data or data sequences.
- Distance is important as it:
  - enables data to be ordered
  - allows numeric calculations
  - enables calculating similarity and dissimilarity
- Without defining a distance measure, almost all statistical and machine learning algorithms will not be able to function.

Distance

A valid distance measure $D(a, b)$ between two components a and b has the following properties

- non-negative: $D(a, b) \geq 0$
- reflexive: $D(a, b) = 0 \iff a = b$
- symmetric: $D(a, b) = D(b, a)$
- satisfies triangular inequality: $D(a, b) + D(b, c) \geq D(a, c)$

Distance (Numerical)

Distances between numerical data points in Euclidean space $\mathbb{R}^n$, for a point $x = (x_1, x_2, .., x_n)$ and a point $y = (y_1, y_2, .., y_n)$, the Minkowski distance of order p (p-norm distance) is defined as:

$$D(x, y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

Distance (Numerical)

Distances between numerical data points in Euclidean space $\mathbb{R}^n$, for a point $x = (x_1, x_2, .., x_n)$ and a point $y = (y_1, y_2, .., y_n)$, the Minkowski distance of order p (p-norm distance) is defined as:

$$D(x, y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

- $p = 1$
- 1-norm distance (L_1)
- Also known as *Manhattan Distance*
- $$D(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Rui Ponte Costa
rui.costa@bristol.ac.uk
COMS21202: Data Acquisition

Distance (Numerical)

Distances between numerical data points in Euclidean space $\mathbb{R}^n$, for a point $x = (x_1, x_2, .., x_n)$ and a point $y = (y_1, y_2, .., y_n)$, the Minkowski distance of order p (p-norm distance) is defined as:

$$D(x, y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

- $p = 2$
- 2-norm distance (L_2)
- Also known as *Euclidean Distance*
- Can be expressed in vector form

$$\begin{aligned} D(x, y) &= \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \\ &= \|\mathbf{x} - \mathbf{y}\| \\ &= \sqrt{(\mathbf{x} - \mathbf{y})^T (\mathbf{x} - \mathbf{y})} \end{aligned} \tag{1}$$

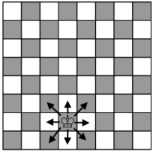
Distance (Numerical)

Distances between numerical data points in Euclidean space $\mathbb{R}^n$, for a point $x = (x_1, x_2, .., x_n)$ and a point $y = (y_1, y_2, .., y_n)$, the Minkowski distance of order p (p-norm distance) is defined as:

$$D(x,y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

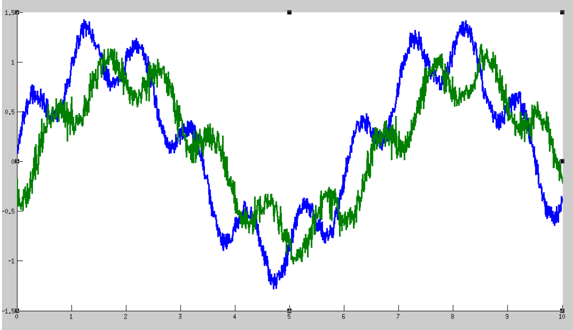
- $p = \infty$
- ∞-norm distance (L_∞)
- Also known as *Chebyshev distance*

$$\begin{aligned} D(x,y) &= \lim_{p \to \infty} \sum_{i=1}^{n} (|x_i - y_i|^p)^{\frac{1}{p}} \\ &= max(|x_1 - y_1|, |x_2 - y_2|, .., |x_n - y_n|) \end{aligned}$$

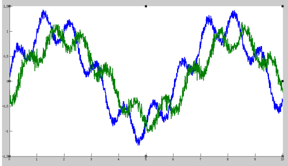


Distance (Numerical Time Series)

- Time Series: successive measurements made over a time interval
- Assume you recorded an audio signal of two people saying the same word *w*



Distance (Numerical Time Series)



P-Norm distances can only

- Compare time series of the same length
- very sensitive to signal transformations:
  - shifting
  - uniform amplitude scaling
  - non-uniform amplitude scaling
  - uniform time scaling

Distance (Numerical Time Series)

Adv. distance: Dynamic Time Warping (Berndt and Clifford, 1994)

- Replaces Euclidean one-to-one comparison with many-to-one
- Recognises similar shapes even in the presence of shifting and/or scaling

- Dynamic Time Warping (DTW) can be defined recursively as
  For two time series $\mathbf{X} = (x_0, .., x_n)$ and $\mathbf{Y} = (y_0, .., y_m)$

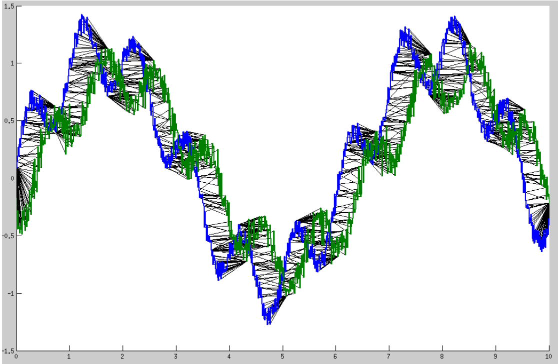
$$DTW(\mathbf{X}, \mathbf{Y}) = D(x_0, y_0) + min\{DTW(\mathbf{X}, REST(\mathbf{Y})), DTW(REST(\mathbf{X}), \mathbf{Y}), DTW(REST(\mathbf{X}), REST(\mathbf{Y}))\}$$

where $REST(\mathbf{X}) = (x_1, .., x_n)$

Distance (Numerical Time Series)

Adv. distance: Dynamic Time Warping (Berndt and Clifford, 1994)

- Can be used for aligning sequences



Distance (Symbolic)

- Distance is not always between numerical data
- Distance between symbolic data is less well-defined, but gaining interest (e.g. text data)
- Distance in text could be:
  - syntactic
  - semantic

Distance (Symbolic)

Syntactic - e.g. Hamming Distance

- Defined over symbolic data of *the same* length
- Measures the number of substitutions required to change one string/number into another
- e.g.
  - $\begin{array}{ccccccc} B & r & i & s & t & o & l \\ B & u & r & t & t & o & n \end{array}$ D('Bristol', 'Burtton') = 4
  - $\begin{array}{cccc} 5 & 2 & 4 & 3 \\ 6 & 2 & 1 & 3 \end{array}$ D(5243, 6213) = 2
  - 1011101
1001001 D(1011101, 1001001) = 2
- For binary strings, hamming distance equals L_1

Distance (Symbolic)

Syntactic - e.g. Edit Distance

- Defined on text data of any length
- Measures the *minimum* number of 'operations' required to transform one sequence of characters into another
- 'Operations' can be: insertion, substitution, deletion

- e.g. D('fish', 'first') = 2
- 'fish' $\xrightarrow{\textit{insertion}}$ 'firsh' $\xrightarrow{\textit{substitution}}$ 'first'

- used in spelling correction, DNA string comparisons

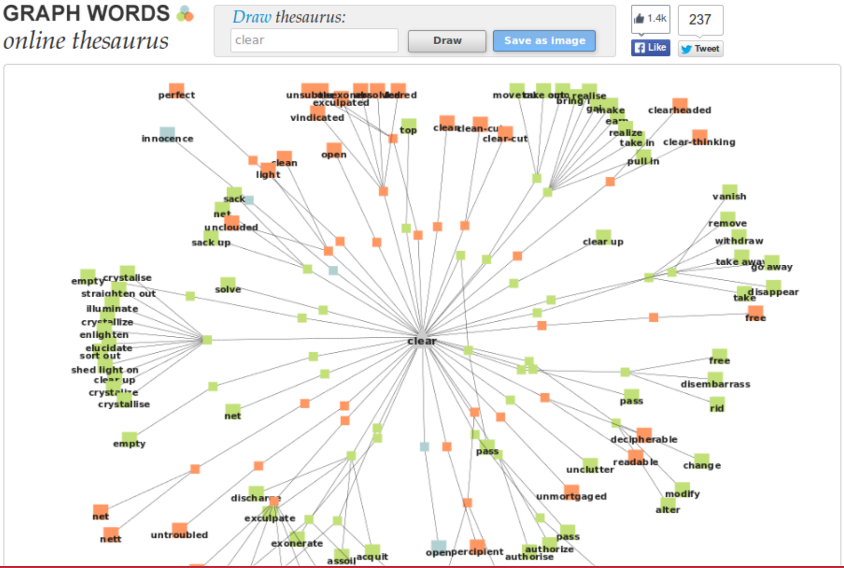
Distance (Symbolic)

Semantic - e.g. WUP Relatedness Measure

- Built on top of a hierarchy of word semantics
- Most commonly used is WordNet (Princeton) `http://wordnet.princeton.edu/`
- WordNet contains more than 117,000 synsets (synset: set of one or more synonyms that are interchangeable in some context)

Distance (Symbolic)

Semantic - e.g. WUP Relatedness



Rui Ponte Costa
rui.costa@bristol.ac.uk

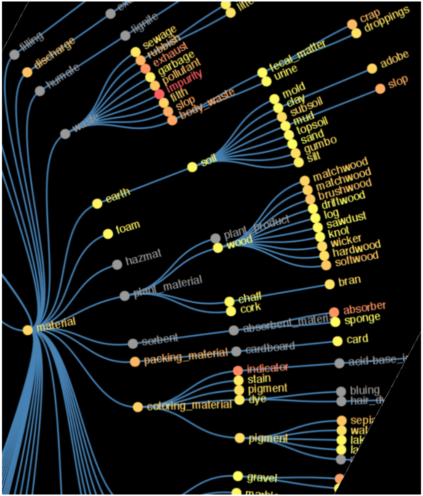
Distance (Symbolic)

Semantic - e.g. WUP Measure

- In WordNet, directed relationships are defined between synsets
  - hyponymy (is-a relationship) e.g. furniture $\rightarrow$ bed
  - meronymy (part-of relationship) e.g. chair $\rightarrow$ seat
  - troponymy [for verb hierarchies] (specific manner) e.g. communicate $\rightarrow$ talk $\rightarrow$ whisper
  - antonymy (strong contract) e.g. wet $\leftrightarrow$ dry

Distance (Symbolic)

Semantic - e.g. hyponymy

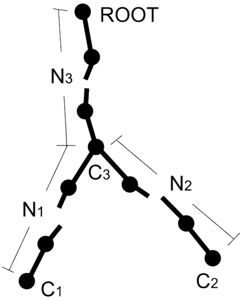


Rui Ponte Costa
rui.costa@bristol.ac.uk

Distance (Symbolic)

Semantic - e.g. WUP Measure

- WUP Measure - Wu and Palmer Distance (1994)
- WUP finds the path length to the root node from the least common subsumer (LCS) of the two concepts, which is the most specific concept they share as an ancestor. This value is scaled by the sum of the path lengths from the individual concepts to the root.



$$WUP(C_1, C_2) = \frac{2 * N_3}{N_1 + N_2 + 2 * N_3}$$

- WUP, along with other relatedness measures can be calculated via Java API for WordNet Searching (JAWS)
- or online: `http://ws4jdemo.appspot.com/`

Distance (Symbolic)

Semantic - e.g. WUP Measure

- **HOWEVER** WUP is a similarity measure, not a distance measure
- It is effectively the inverse of a distance measure, taking higher values for similar data points.
- WUP(w1, w1) = 1
- Similarity measures can be converted to distance measures, depending on the values they take:

$$D_{WUP} = 1 - WUP$$

Distance - Conclusion

- Once you define a distance measure on your data, you can perform numeric operations
- Different distance measures will enable you to use the same data for various goals

Questions?

- What is the Edit Distance for [Menti.com]:

  D('horse', 'rose') = 2

  'horse' $\xrightarrow{\textit{substitution}}$ 'rorse' $\xrightarrow{\textit{deletion}}$ 'rose'

  D('AGGCTATCACCTGACC', 'TGGCCTATCACCTGAC') = 3

  'AGGCTATCACCTGACC' $\xrightarrow{\textit{del}}$ 'AGGCTATCACCTGAC' $\xrightarrow{\textit{sub}}$ 'TGGCTATCACCTGAC' $\xrightarrow{\textit{ins}}$ 'TGGCCTATCACCTGAC'

- Which are numeric distances? [Menti.com]

Agenda

Today we are going to talk about

1. Data acquisition
2. Data characteristics: distance measures
3. **Data characteristics: summary statistics [reminder]**
4. Data normalisation and outliers

Mean and Variance (Reminder)

For one-dimensional data $\{x_1, .., x_n\}$,
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 $\{\mathbf{x}_1, .., \mathbf{x}_n\}$ where $\mathbf{x}_i$ is an m-dimensional vector,
Mean: calculated independently for each dimension

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

Variance can still be computed along each dimension

Covariance Matrix: spread and correlation

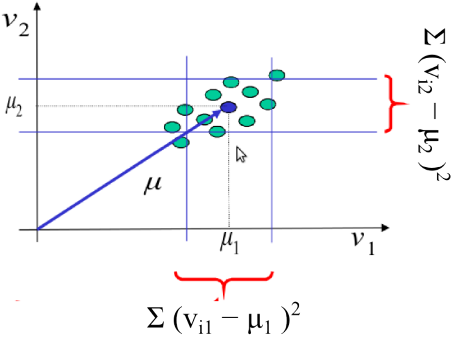
$$\begin{aligned}\Sigma &= \frac{1}{N-1}\sum_i (\mathbf{x}_i - \boldsymbol{\mu})^2 \\ &= \frac{1}{N-1}\sum_i (\mathbf{x}_i - \boldsymbol{\mu})^T(\mathbf{x}_i - \boldsymbol{\mu})\end{aligned}$$

WARNING: Σ is the capital letter of σ, not the summation sign!

Covariance Matrix

In two dimensions,

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



Rui Ponte Costa
rui.costa@bristol.ac.uk

Covariance Matrix

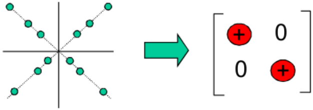
In two dimensions,

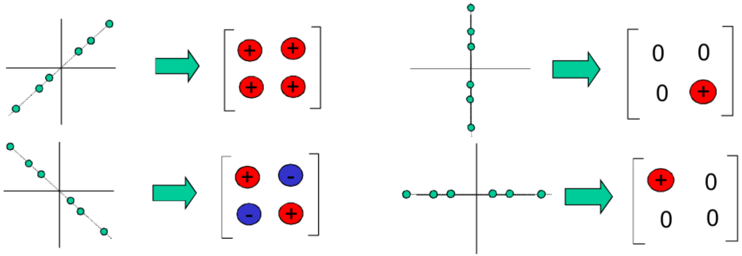
$$\Sigma = \frac{1}{N-1} \sum_i \begin{bmatrix} (v_{i1} - \mu_1)^2 & (v_{i1} - \mu_1)(v_{i2} - \mu_2) \\ (v_{i1} - \mu_1)(v_{i2} - \mu_2) & (v_{i2} - \mu_2)^2 \end{bmatrix}$$

- In addition to the variances along each dimension, the covariance matrix measures the correlation between components
- A positive covariance between two components means a proportional relationship between the variables.
- A negative covariance value indicates and inverse proportional relationship.

Covariance Matrix

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





Rui Ponte Costa
rui.costa@bristol.ac.uk

Covariance and correlation: Demo

geogebra.org/m/wrSFAnkh

Covariance Matrix

In three dimensions,

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

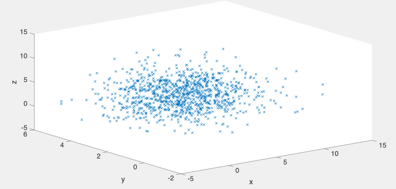
Covariance matrix is always

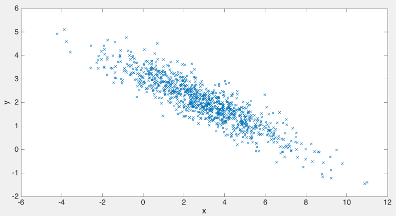
- square and symmetric
- variances on the diagonal
- covariance between each pair of dimensions is included in non-diagonal elements

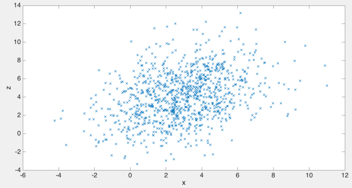
Covariance Matrix - e.g.

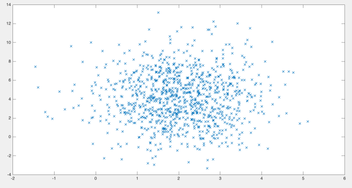
For the covariance matrix,

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





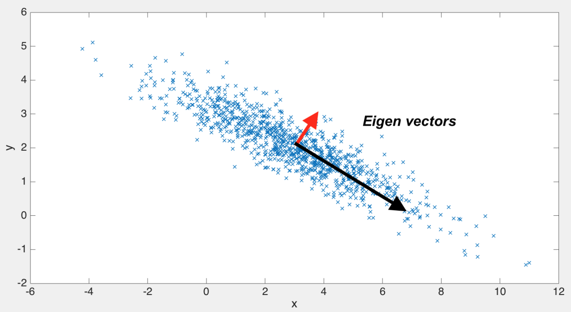




Rui Ponte Costa
rui.costa@bristol.ac.uk

Covariance Matrix: Eigen analysis

- Eigenvectors and eigenvalues define **principal axes** and spread of points along directions
- Commonly used to reduce data dimensionality (e.g. Principal component analysis [PCA])



Covariance Matrix: Eigen analysis

Definition

For a square matrix A,
if there exists a non-zero column vector v where

$$Av = \lambda v$$

then,
$v \rightarrow$ eigenvector of matrix A
$\lambda \rightarrow$ is eigenvalue of matrix A

e.g.

$$A = \begin{bmatrix} 0 & -1 \\ 2 & 3 \end{bmatrix}, v_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Covariance Matrix: Eigen analysis

- To calculate eigenvectors of a square matrix, solve $|A - \lambda\mathbf{I}| = 0$ where
  - $\mathbf{I}$ is the identity matrix
  - |A| is the determinant of the matrix
- For 2×2 matrices, two eigenvalues are found λ_1, λ_2

e.g.

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

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

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

Covariance Matrix: Eigen analysis

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

For $\lambda_1 = 1$

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

- This simplified to:

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

- Its set $v_{12} = 1$ then we get the eigenvector[1]:

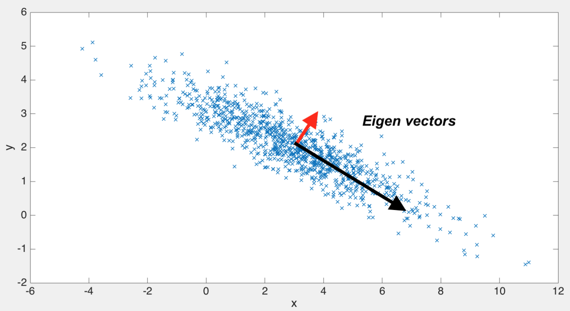
$$\begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \tag{4}$$

- Verify that this is indeed a valid eigenvector by calculating $Av = \lambda v$

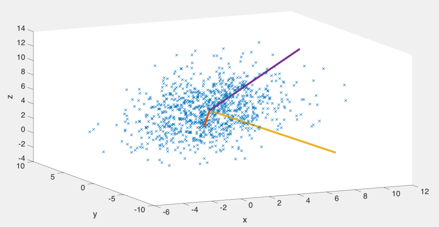
[1] Note that there many eigenvectors that work for a particular eigenvalue, but they all have the same direction. We could consider the eigenvector with $\|v_1\| = 1$, $v_1 = (\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}})$.

Covariance Matrix: Eigen analysis

- **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: another example



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

Mean vs. Median

- An alternative to arithmetic mean is the **median value**
- But median is difficult to work with
- e.g. median of two sets cannot be defined in terms of the individual medians

Note - Sample Variance vs. Variance

Given sample $\{x_1, x_2, .., x_N\}$

$$\mu \approx \bar{x} = \frac{1}{N}\sum_i x_i \tag{5}$$

$$\sigma^2 \approx s^2 = \frac{1}{N-1}\sum_i (x_i - \bar{x})^2 \tag{6}$$

- These are only estimates of the 'true' mean and variance
- $N - 1$ gives unbiased estimate of the variance [2]
- As $N \to \infty$
  - $\bar{x} \to \mu$
  - $s^2 \to \sigma^2$

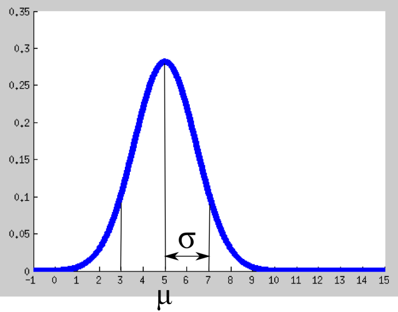
[2] This means that this variance estimator is equal to the true variance when $N \to \infty$. More information about this can be found here.

Rui Ponte Costa
rui.costa@bristol.ac.uk

Normal Distribution (Reminder)

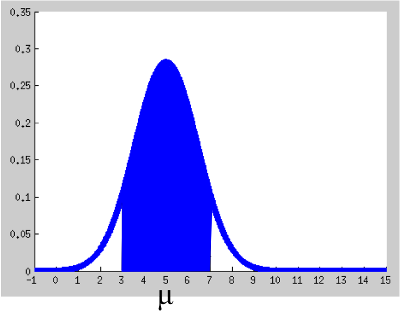
For a normal distribution $\mathcal{N}(\mu, \sigma^2)$ in one dimension, the probability density function (pdf) can be calculated as

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{7}$$



Normal Distribution (Reminder)

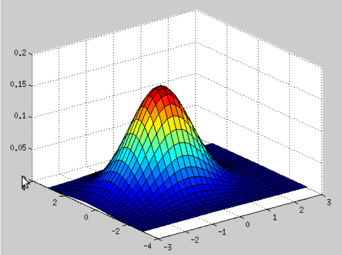
- 68% of the sample *should* lies within one standard deviation of the mean
- 95% of that area lies within two standard deviations of the mean
- 99.9% of that area lies within three standard deviations of the mean

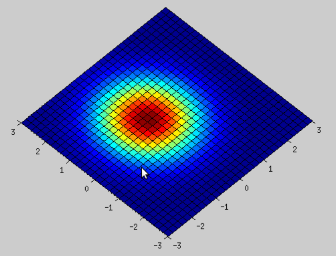


Normal Distribution - Multi-dimensional

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

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^M |\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})} \tag{8}$$





WARNING: **Σ is the capital letter of σ, not the summation sign!**

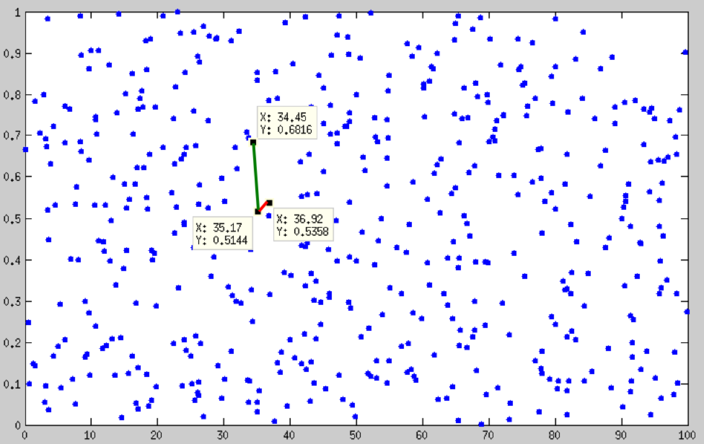
Agenda

Today we are going to talk about

1. Data acquisition
2. Data characteristics: distance measures
3. Data characteristics: summary statistics [reminder]
4. **Data normalisation and outliers**

Data Characteristic - Data Normalisation

- Multi-dimensional data may need to be normalised before distance is calculated [3]



[3] note the difference in magnitude between the two dimensions below!

Data Characteristic - Data Normalisation

- Multi-dimensional data may need to be normalised before distance is calculated.
- Methods for normalisation:
  1. Rescaling

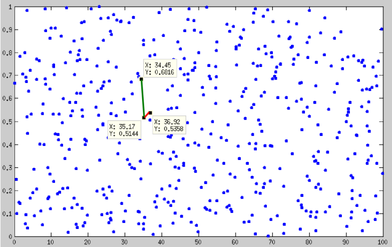
$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

  2. Standardisation (also known as z-score)

$$x' = \frac{x - \mu}{\sigma}$$

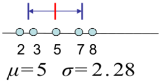
  3. Scaling to unit length

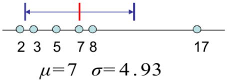
$$x' = \frac{x}{\|x\|}$$

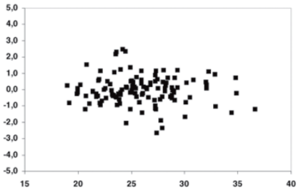


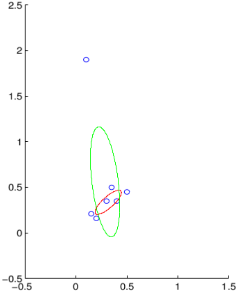
Data Characteristic - Outliers

- 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 other points
  - usually due to fault in measurement
  - not always easy to remove









Problem class tomorrow!

- Problem Class Tomorrow (Thur 1-2): Data Acquisition
- Prepare your answers in advance [problem sheet on github SPS page]

Further Reading

- Fundamentals of Multimedia
  Li and Drew (2004)
  - Section 6.1 Digitization of Sound
- Applied Multivariate Statistical Analysis
  Hardle and Simar (2003)
  - Section 1.2
  - Section 1.4
  - Section 3.1
  - Section 3.2
- Linear Algebra and its applications
  Lay (2012)
  - Section 6.5
  - Section 6.6
- Advances in Data Mining Knowledge Discovery and applications
  Karahoca (Ed.) (2012)
  - Chapter 3. Similarity Measures and Dimensionality Reduction Techniques for Time Series Data Mining