Data statistics and transformation revision

Michael J. Watts

http://mike.watts.net.nz

Lecture Outline

- Statistical operations on Data
- Data transformations
  - The objectives of a data transform
  - Linear versus non-linear transformations
  - Transformations for pre-processing of data
  - DFT and FFT Transformations
  - Wavelet Transformations

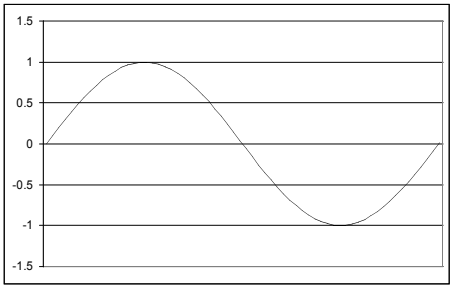
Why Analyse Data?

- Important part of problem solving process
- Can suggest method to use to solve problem
- Answers many important questions about the data set and the problem
- Improves understanding of the problem

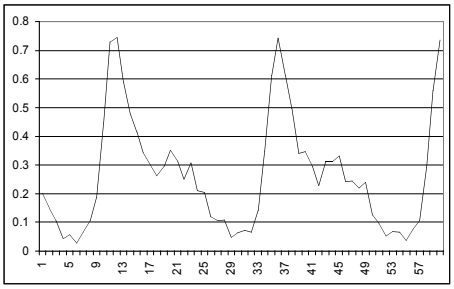
Why Analyse Data?

- what are the statistical parameters of the data?
  - mean, standard deviation, correlation
- what is the nature of the process?
  - periodic, chaotic, random?
  - a random process cannot be predicted at all
  - periodic processes are more easily modelled
  - chaotic processes are a bit harder to model

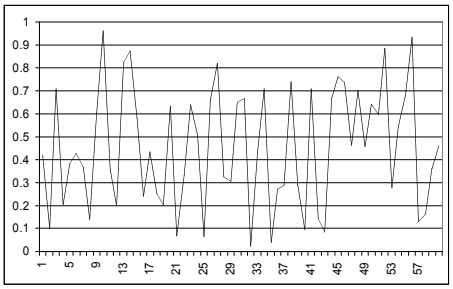
A Periodic Process



A Chaotic Process



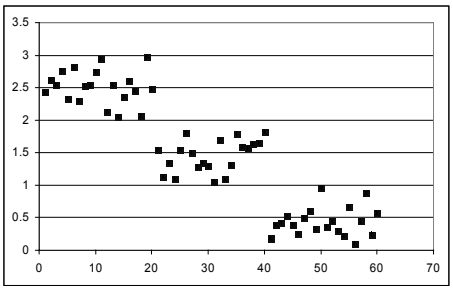
A Random Process



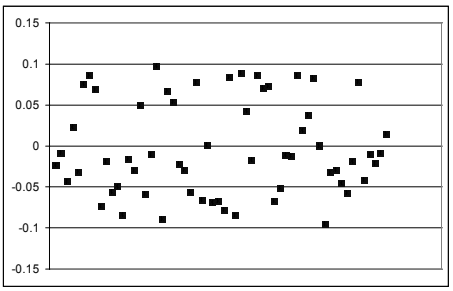
Why Analyse Data?…

- How is the available data distributed?
  - does it naturally cluster together?
  - is it uniformly distributed?
  - does it cover enough of the problem space to be useful?

Clustered Data



Uniform Distribution



Why Analyse Data?

- Is there missing data and how much?
  - is missing data a critical obstacle?
  - can other methods be used to compensate for the gaps?

Why Analyse Data?

- What features can be extracted from the data?
  - reducing the number of variables in the data set
  - can assist with modelling the problem
  - can make correlations / relationships easier to see

Statistical Data Analysis

- Discover repetitiveness in data
- Simple functions
  - mean
  - standard deviation
  - Histogram

Statistical Data Analysis

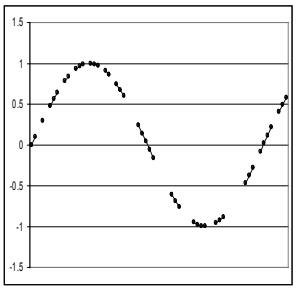
- Arithmetic mean
  - A value that is representative of the population of values
- Standard deviation
  - A measure of how far from the mean values deviate
- Analysis must be appropriate for the data
  - Measurement theory

Correlation

- Correlation
  - Finds linear dependencies between variables
  - Correlation coefficients may change in time for time series data

Regression and Interpolation

- *Regression analysis:*
  - finds a formula which approximates data for a given output variable
- *Interpolation:*
  - fills in gaps in data
  - fit data into curves



PCA

- Principal component analysis (PCA)
  - eliminates redundant variables
  - reduces number of variables in data set
  - makes it easier to model

ICA

- Independent component analysis
  - separates components from a set of unknown independent components
  - Example: the cocktail party problem - separating speakers from a signal taken from cocktail party speech - several people speaking simultaneously

Clustering Methods

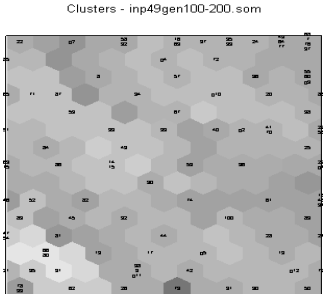
- Assigns each datum into one and only one subset of the data set
- k-means clustering
  - finds k centres in the data space
  - sum of squares of distance between each data point and nearest centre is minimised
  - Distance between cluster centres is maximised

Vector Quantisation

- represents a *n* dimensional space as a *m* dimensional one
- $m < n$
- Preserves distance between examples
  - examples that are close in *n* dimensional space will be close in *m* dimensional space

Example: SOM for vector quantisation of data in Bioinformatics

- SOM !! ---->
- A selected subset of genes expressed in 49 tissue samples (two types of Leukaemia - ALL and AML)



The objectives of a data transformation

- Data rate reduction – meaningful features are extracted from it
- Improving the quality of the information – via noise suppression or image enhancement
- Knowledge discovery and better understanding of the processes and events
- Finding similarities and analogies between processes and events

Linear versus non-linear transformations

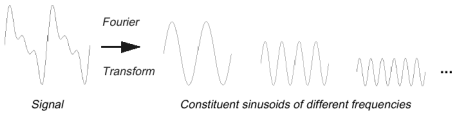
- Linear transformation
  - *F(x)* of a raw data vector *x* such that *F* is a linear function of *x*. E.g. $F(x)=2x+1$
- Non-linear transformations
  - *F(x)* of a raw data vector *x* such that *F* is a non-linear function of *x*. E.g. $F(x)=1(1+e^{-x.c})$
- Other non-linear transformations
  - The logarithmic function, $F(x)=log_{10}x$.

Transformations for pre-processing of data

- Sampling
  - The process of selecting a subset of the available data. Can be applied to continuous time series data such as speech and music.
- Discretisation
  - Representing continuous-valued data with the use of sub-intervals where the real values lie.
- Normalisation
  - Moving the scale of the real data into a predefined scale e.g. [0,1]. Can be linear or non-linear.

DFT and FFT Transformations

- Discrete Fourier Transforms (DFT)
  - A non-linear transformation where the data is represented as a sum of harmonic Fourier series



  - A periodic signal (e.g sin) is characterised by one frequency. Every signal can be represented as a sum of periodic signals with different frequencies.

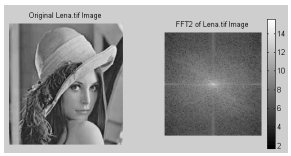
DFT and FFT Transformations

- Fast Fourier Transform (FFT)
  - The fast implementation of a DFT when the number of periodic signals is a power of 2.

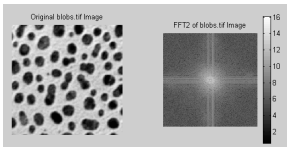
- Applications of FFT Transforms
  - Speech and Image data.
  - Sunspot activity analysis.

DFT and FFT Transformations

- Fast Fourier Transform (FFT) of Images



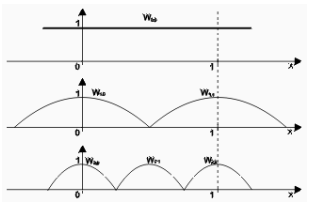
The FFT operator transforms the image from the spatial to the frequency domain



It allows us to analyse the information content of the image

Wavelet Transformations

- Wavelet Transformation
  - A non-linear transformation. Can represent sight changes of the signal within the chosen window from the time scale.
- $W_{a,b}(x) = f(ax - b)$,
  - f = non-linear function
  - a = scaling parameter
  - b = shifting parameter



Summary

- Data analysis is an integral part of the problem solving process
  - can suggest means of solving problem
  - assists in the modelling process
- Statistical techniques, clustering techniques, vector quantisation are all available methods

Summary

- There are many transforms available to apply to datasets, some more appropriate than others.
- Linear and non-linear transformations are simple but effective operations.
- DFT, FFT, and Wavelet transforms are powerful ways of analysing signals.