Measurement analysis - II

Lecture slides for S-38.3183

21.3.2007

Mika Ilvesmäki

Contents

- Dependence statistics
  - cross-correlation
  - autocorrelation
- Time series analysis
  - stability
- Self-similarity
  - Hurst parameter

Goals of this lecture

- After this lecture you should know
  - What different correlation statistics there are
  - What different correlation statistics mean
  - And what things must be considered when evaluating different correlation statistics
  - Preliminary time series analysis
  - What self-similarity means and why it exists in the network
  - How self-similarity is evaluated
    - And how to calculate the Hurst parameter, in three different ways (requires reading ”Chapter 2” also)

Dependence statistics

- Cross-correlation
  - Calculated between two series
  - May be evaluated with a delay
    - Results in correlation series
- Autocorrelation
  - Calculated within the series
  - Correlation series indicates dependence or periodicity (or lack thereof)

Correlation

- If two phenomena covary
  - They do it in a positive or negative sense
    - Or not at all
  - Covariation is always perceived (through measurements)
- Correlation does not imply causality!

Cross-correlation

- A standard method of estimating the degree to which two (different) series are (linearly) correlated
  - aka dot product,
- Normalized correlation coefficient that equals unity indicates perfect match
  - But gives no explanation why there is a perfect match.

HELSINKI UNIVERSITY OF TECHNOLOGY

Mika Ilvesmäki, D.Sc. (Tech.)

Determining cross correlation

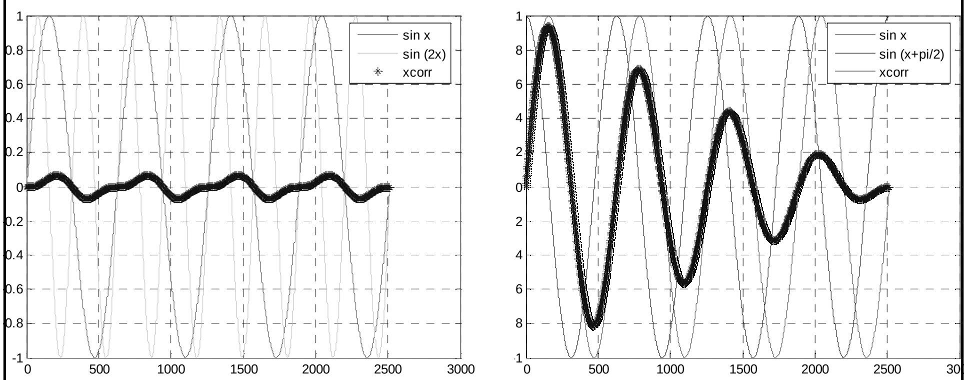
- Definition includes delay d (lag k…)
  - If sample index outside the series a) ignore b) assume zero c) wrap around (preferred)
  - Delay may be significantly less than series length N to test for short delay correlation only
  - If you find correlation with certain d it is an indication of a correlating phenomena with a time delay.
  - When two random processes (x and y) are statistically independent then the R_{xy} and R_{yx} are equal.
  - Hint: Always plot the original signals together with cross-correlation with varying lag (d)

HELSINKI UNIVERSITY OF TECHNOLOGY

Mika Ilvesmäki, D.Sc. (Tech.)

Properties of cross correlation

- High correlation likely indicates periodicity
- Correlation does not indicate any physical relation and correlation is indicated only based on the samples

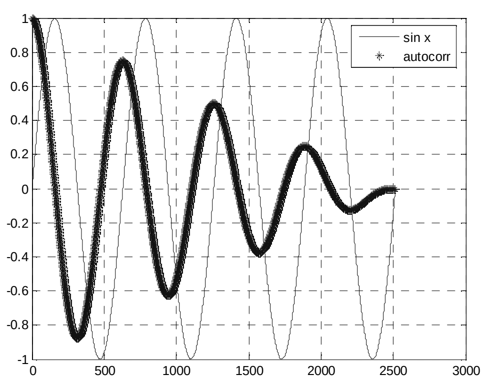


Auto-correlation

- Observations on the signal should be equally spaced (in time or in space)
- Correlation between values of the same variable at different times (lagged signal)
  - A high correlation is likely to indicate a periodicity in the signal of the corresponding time duration.
  - The autocorrelation of a periodic function is, itself, periodic with the very same period.
- Auto-correlation with zero lag will always results in unity (perfect match)
  - Usually, as lag increases the auto-correlation value will decrease
- Used to detect non-randomness in data
- Auto-correlation with varying lag
  - Indicates the persistence (memory) of the proces

Properties of auto-correlation

- Properties of auto-correlation
  - How quickly random signal or processes change with respect to the time function
  - Whether process has a periodic component and what the expected frequency might be
  - The autocorrelation of a white noise signal will have a strong peak at d = 0 and will be close to 0 for all other d.
    - This shows that a sampled instance of a white noise signal is not statistically correlated to a sample instance of the same white noise signal at another time.



Time (or space) series

- Measured events occur in time (or in space)
    - Collect the timestamp (location) of the event in evenly spaced timeslots.
    - Repeat, and you have yourself a timeseries
- Useful for determining the amount of data on a link
    - Arrived data or packets/Time window
- Useful for detecting the start and end of a phenomena

Purpose of Time series analysis

- Time series analysis aims to:
    - identify the nature of the phenomenon represented by the observations
    - predict future values of the time series.
- Time series analysis ables us to extrapolate the identified pattern to predict future events
    - This does not depend up on our understanding of the underlying phenomena and/or the validity of our interpretation (theory) of the phenomenon

Mika Ilvesmäki, D.Sc. (Tech.)

Procedure for time series analysis

- It is assumed that the data consist of a systematic pattern and random noise which usually makes the pattern difficult to identify.
- Is time series stable?
  - First question: Is it (the distribution) heavy tailed?
  - Process in three steps
    - Graph the series (x-axis time, y-axis event)
      - Periodicity, outliers, determine also basic statistics
      - Do histogram of the series
        - Lose temporal structure, gain info on symmetry
      - Do the converging variance test
        - Plot S^2_n for the first n observations as a function of n. If data has finite variance, the sample variance should converge to a finite value.

Mika Ilvesmäki, D.Sc. (Tech.)

Self-similarity

- Self-similar phenomenon looks the same when viewed at different scales of a dimension
  - Time: µs, ms,s,min,h,a etc.
  - Space: µm, mm,cm,m,km etc.
- Typically self-similarity of a phenomena means that there are non-negligible correlations between the event counts in far apart spaced observations (time, space)

Definition of self-similarity

- Self-similarity of a time series:
  - when aggregated…
    - (leading to a shorter time series in which each point is the sum of multiple original points)
  - the new series has the same autocorrelation function as the original…
  - and the series is *distributionally* self-similar.

HELSINKI UNIVERSITY OF TECHNOLOGY Mika Ilvesmäki, D.Sc. (Tech.)

Self-similarity

- Long-range dependence
  - A process with long-range dependence has an autocorrelation function $r(k) \sim k^{-\beta}$ as the lag $k \to \infty$ and $0<\beta<1$
  - Therefore the r(k) of such process decays hyperbolically
    - Poisson traffic decays exponentially
    - Hyperbolic decay is much slower than exponential decay
    - Since $\beta<1$, the sum of autocorrelation values approaches inifinity
- The parameter that is usually (for historic reasons) used to indicate the speed of decay of the series' autocorrelation function is the *Hurst* parameter
  - $H=1-\beta/2$ and therefore $½<H<1$. As H approaches unity, the degree of self-similarity increases.
  - Simplified: To test self-similarity of a series: Is H significantly different from ½?

Hurst parameter

- There are several theoretically sound estimators for Hurst parameter
- However, they may disagree when applied to same data
- Differing views on how to preprocess data
  - At least aim to
    - remove mean,
    - trends,
    - best polynomial fit (of high order, like 10)

HELSINKI UNIVERSITY OF TECHNOLOGY Mika Ilvesmäki, D.Sc. (Tech.)

Hurst parameter: Variance-time

- Variance-time relation
  - Calculate the variance of series as you take more and more of the series into the calculation
  - Plot variance-time relation
    - Log-log plot
    - A straight line with slope $-\beta>-1$ indicates self-similarity
- Estimation is made in time-domain

Hurst parameter: R/S-method

- R/S: Rescaled Range
  - Relies on rescaled range (R/S) statistic growing like a power law with H as a function of number of points *n* plotted.
  - The plot of R/S versus *n* on log-log has slope which estimates H
  - Process:
    - Divide a timeseries into K non-overlapping blocks, blocks vary from 1…n
    - Compute R/S(n), the rescaled adjusted range for all n
      - R is the range of the data in the block n, S is the sample variance of the data in the same block.
      - The R/S values plotted against n should have n^H relation.
      - In log-log space the slope of the R/S vs. n –line is H
- Estimation is made in time-domain

Hurst parameter: Periodograms

- Fact: Spectral density of self-similar processes obeys power law near the origin
  - The slope of the power spectrum of the series as frequency approaches zero (and is near origin)
    - The periodogram slope (in a log-log plot) is a straight line with slope 1-2H close to the origin (10% of the lowest frequencies)
- Estimation made in frequency domain

Other methods for estimating H

- Analysing wavelets
  - Generalized Fourier-transform
- Whittle estimator focuses on making observations near zero frequency
- Both of these methods are in the frequency domain
  - And all of these are dealt with in advanced courses ☺

Meaning of self-similarity

- A high value of Hurst parameter often increases delays and packet loss in a network.
- If buffer provisioning is done using the assumption of Poisson traffic then the network will be underprovisioned.
- The Hurst parameter is a dominant characteristic for a number of packet traffic engineering problems.
- The origins of LRD are uncertain but the most likely cause seems to be the aggregation of file transfer processes (ftp, p2p).

All is not as it seems…

- Trends and periodicities or other corrupting noise may be mistaken for LRD.
  - All techniques to find H are somewhat vulnerable to addition of short-range dependent data.
- A researcher (and a student ☺) relying on a single measure of the Hurst parameter is likely to draw false conclusions.

Applications for self-similarity studies

- Main idea is to statistically analyze traffic process
  - Build traffic models for simulators
  - Be able to analytically handle traffic
- Is Poisson model enough?
  - Recent studies show that using Poisson-modeled traffic significantly overestimates network performance
  - Self-similar models perform better
  - Multi-fractal models are even better
    - Multi-fractals dealt with in advanced courses
- However,
  - Self-similarity analysis is at the moment just ”interesting”
  - Practical applications are few and far between (in networking)

Explanations for self-similar behavior

- Open loop –models (edge oriented)
  - Connections arrive at random
    - Files have size, network has rate
  - Heavy-tailed distribution of file sizes causes LRD
    - Are filesizes really heavytailed?
- Closed loop –models (network oriented)
  - 90% traffic is closed loop (TCP)
    - Transmission of future packets depends up on the faith of the previous packets -> correlation independent of file size
- Mixed models
  - Protocol functionality is layered (TCP->IP->Ethernet)
  - Different layers act on different timescales -> multiple timescales (and self-similarity)

Measurement analysis summary

- Correlation
  - Cross- and auto
  - Significance
  - Intrepretation
- Basics of timeseries analysis
- Self-similarity
  - Methods of how to determine
    - R/S, Variance-time, Periodograms
    - Causes, consequences