Series Variables

Ella Bingham
Laboratory of Computer and Information Science, HUT
3rd November 1999
Tik-61.181 Special Course in Information Technology I:
Data Preparation for Data Mining
Based on Chapters 9.1–9.5 in Pyle's book
"Data Preparation for Data Mining"

2

Contents

Note: the handouts contain no figures, as the figures are found in the book

1 Properties of series data

In series data, the **ordering** of the data set is important:
data set consists of measurements $x(t)$

- t is called **displacement variable** or **index variable** and it is always monotonic
- t usually denotes time, but may also be distance etc.
- t is recorded or the amount of displacement is defined
- t is recorded numerically
- x(t) is usually numerical
- x(t) might also be nonnumerical: for example, ACBABBCAAB ...
  If some regular (for example, cyclical) pattern is found in the series, that should be preserved

- it is also possible to record several variables $x(t), y(t), \ldots$ using the same index variable t
- random sampling from a data set cannot be done, because random sampling destroys ordering
- when analyzing series data, it is important to extract and preserve the properties of series data — techniques for nonseries data may not be suitable

2 Describing a series using Fourier analysis

By adding together sine and cosine waves of different frequencies, phases and amplitudes, it is possible to create **any periodic** series shape.

- Frequency: how many times a waveform repeats its pattern in a given time
- Phase: where the peaks and valleys of a wave occur in relation to peaks and valleys of other waves
- Amplitude: distance between highest and lowest values of a wave

Nonperiodic waveforms are created assuming that the pattern found in the data set at hand repeats itself in the time periods before and after our observations.

Read more in *Digital Signal Processing*, E. M. Ifeachor and B. W. Jervis, Addison-Wesley, or any other signal processing book

3 Describing a series using spectrum

Frequency spectrum or **power spectrum** shows what frequencies the waveform contains.

- The **height** of each spike in the power spectrum corresponds to the **amplitude** of each component waveform
- The **position** of each spike corresponds to the **frequency** of each component waveform

4 Patterns in the series

In the so called **classical decomposition**, the series is regarded as being built from four separate components: trend, seasonality, cycles and noise.

4.1 Trend

Trend is a noncyclic, monotonically increasing or decreasing component of the waveform.

Power spectrum of a trended waveform does not show any other detail! Trend must be **removed** before analyzing the power spectrum, for example by computing a linear regression $x(t) = at + b$ and replacing the series with the residuals of this regression.

The trend may also be nonlinear (Error in Pyle's figure 9.13!)
In this case, the trend could be removed by nonlinear regression.

Note: what is regarded as a trend over one time period may be a cycle over a different period — danger of misinterpretation!

Detrending nontrended data hides some other important properties of the series, and adds a nonexistent trend in the predictions.
If there is some a priori knowledge that the process does not generate a trend, the data should not be detrended even if it appears to have a trend.

4.2 Cycles

Cycles are fluctuations in the level of the series that have some identifiable repetitive form and structure.
For example: sine and cosine waves, although cycles are not necessarily based on sines and cosines

4.3 Seasonality

Pyle: Regardless of any other trend, cycle or noise influence, certain seasons are inherently different.
For example, consumers spend more in December than in other months regardless of economic conditions etc. This is caused by a phenomenon that is local to the season: Christmas. Although Christmas occurs cyclically, it is not a cyclic event itself.

Reader: But why shouldn't it still be regarded as cyclic?

Pyle: When comparing the effects of a sales campaign in June and another in December, the effect of Christmas must be removed to make a fair comparison of the campaigns.

Reader: In this case, the comparison should always be done keeping in mind the effect of Christmas.

Seasonal changes usually require a domain expert to explain — is there any damage done if they are regarded as cyclic components?

4.4 Noise

Noise is left when trend, cyclic and seasonal components are extracted. It is random (if it was not, it would be characterized as one of the other patterns!)
Noise is named according to its frequency distribution:
white noise contains equal amounts of all frequencies;
blue noise contains high frequencies etc.
Noise sources can sometimes be identified by looking at the power spetrum.

Example: **Random walk** is one kind of noise. (Is it?)
Next point in the series is at distance 0 ... 1 up or down from the current point. Distance and direction are chosen at random.

By coincidence, the plot shows some trend and cyclical pattern which are not genuine. Warning: it is easy to "discover" meaningless patterns and draw wrong conclusions!

Describing an existing waveform and predicting its future shape are entirely different activities!

5 Describing a series using autocorrelation

Correlation r measures how values of one variable change as values of another variable change.

Properties of r:
$-1 \leq r \leq 1$
$r = 1$: two variables x and y are completely linearly related and move in the same direction, $y = ax + b$ where $a > 0$
$r = -1$: x and y are completely linearly related and move in opposite directions, $y = ax + b$ where $a < 0$
$r = 0$: knowing the value of one variable tells nothing about the value of the other variable

Coefficient of determination r^2:
$0 \leq r^2 \leq 1$

Autocorrelation measures how well one part of the series correlates with another part of the series.
The distance between index points is called the **lag**.

In a **correlogram**, autocorrelations of lag one, two, three ... are plotted.
The correlogram is useful when building a model for time series data.
For example, the positions of spikes show the lengths of cycles.

6 Repairing series data problems

6.1 Missing values

Either the feature value x or the index value t in $x(t)$ may be missing.

Feature value is missing: replace missing values using **autoregression** that measures the self-similarity of the waveform across different lags, or **smooth** the missing values.
Both methods enhance a pattern that is discovered elsewhere in the series.
Problem: When data is later modeled, this enhanced pattern is "discovered" and it may predominate in a spectral analysis or a correlogram.
Solution: add noise to the replacement values. The level of noise could be the same as the level of noise found in the present values (if this can be estimated).

When this kind of tricks are done, it would be preferable that the person doing DP also does DM!

Index value is missing: the order of data points is not known. For example, $x(?) = 3435$, $x(??) = 5344$. If no empty slots in the data series can be filled with these numbers, the observations must be rejected.

6.2 Outliers

Questions:
Are they really significant?
What kind of process has created them?
Can they be translated back into the normal range if they are errors?

If no explanation for the outliers is found, replace them using the same methods as with missing feature values.

6.3 Nonuniform displacement

Most tools assume that measurements are taken at uniformly sampled index points: $x(n-1), x(n), x(n+1), \ldots$.
Pyle: Techniques for noise removing work well.
Reader: If I know what the displacements are, I could choose the smallest displacement as a constant displacement and treat data as one having missing values

6.4 Trend

Trend removal discussed earlier!