2. Basic Descriptive Techniques

Haipeng Xing

State University of New York, Stony Brook

Outline

Types of variation

Traditional methods of time-series analysis are mainly concerned with decomposing the variation in a series into components representing trend, seasonal variation and other cyclic changes. Any remaining variation is attributed to 'irregular' fluctuations.

- Trend. This may be loosely defined as 'long-term change in the mean level'. A difficulty with this definition is deciding what is meant by 'long term', hence we must take into account the number of observations available and make a subjective assessment of what is meant by the phrase 'long term'.
- Seasonal variations. Some time series exhibit variations at a fixed period.
- Irregular fluctuations. After trend and cyclic variations have been removed from a set of data, we are left with a series of residuals that may or may not be 'random'.

Examples: Financial time series

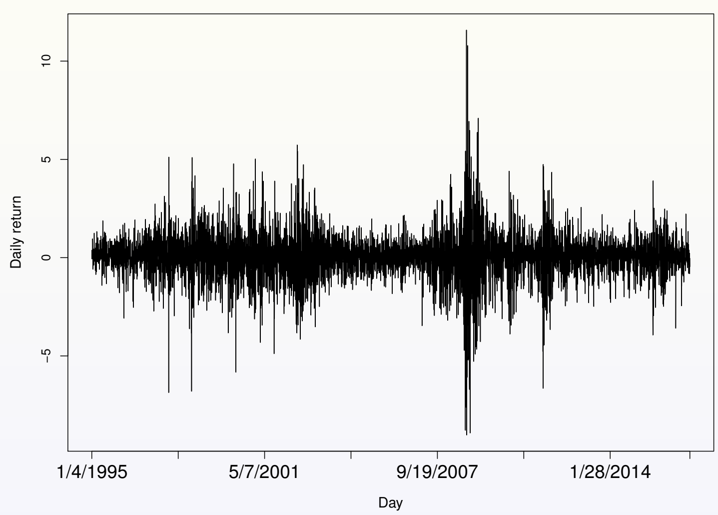


Figure 2.1: Daily returns of the adjusted closing prices of S&P500 index from January 4, 1995 to December 30, 2016.

The mean of the return series seems to be stable with an average return of approximately zero, but the volatility of data changes over time.

Stationary time series

- A time series is said to be **stationary** if there is no systematic change in mean (no trend), if there is no systematic change in variance and if strictly periodic variations have been removed. In other words, the properties of one section of the data are much like those of any other section.
- Strictly speaking, it is very often that time series data violate the stationarity property. However, the phrase is often used for time series data meaning that they exhibit characteristics that suggest a stationary model can sensibly be fitted.
- Much of the probability theory of time series is concerned with stationary time series, and for this reason time series analysis often requires one to transform a non-stationary series into a stationary one so as to use this theory.

The time plot

- The first, and most important, step in any time-series analysis is to plot the observations against time. This graph, called a **time plot**, will show up important features of the series such as trend, seasonality, outliers and discontinuities. The plot is vital, both to describe the data and to help in formulating a sensible model.
- Plotting a time series is not as easy as it sounds. The choice of scales, the size of the intercept and the way that the points are plotted (e.g. as a continuous line or as separate dots or crosses) may substantially affect the way the plot ‘looks’, and so the analyst must exercise care and judgement.

Transformations: Box-Cox transformation

There are several reasons to transform the data.

- To stablize the variance.
- To make the seasonal effect additive
- To make the data normally distributed

Given an observed time series $\{x_t\}$ and a transformation parameter λ, the the Box–Cox transformation of the series is given by

$$y_t = \begin{cases} (x_t^\lambda - 1)/\lambda & \lambda \neq 0 \\ \log x_t & \lambda = 0. \end{cases}$$

The 'best' value of λ can be 'guesstimated', or alternatively estimated by a proper inferential procedure, such as maximum likelihood.

Analysing series with trend but no seasonal variations

- Curve fitting. A traditional method of dealing with non-seasonal data that contain a trend, particularly yearly data, is to fit a simple function of time such as a polynomial curve (linear, quadratic, etc.). For example, a series X_t with global linear trend is given by

  $$X_t = \alpha + \beta t + \epsilon_t,$$

  where α, β are constants and ϵ_t denotes a random error term with zero mean. Then the trend term is $m_t = \alpha + \beta t$.
- Filtering. Another way is to use a **linear filter** to converts one time series, $\{x_t\}$, into another, $\{y_t\}$, by the linear operation

  $$y_t = \sum_{r=-q}^{+s} a_r x_{t+r},$$

  where $\{a_r\}$ is a set of weights with $\sum a_r = 1$.

Analysing series with trend but no seasonal variations

- Filtering with weights whose sum is 1 is often referred to as a moving average. Moving averages are often symmetric with $s = q$ and $a_j = a_{-j}$.
  - The simplest example of a symmetric smoothing filter is the simple moving average, for which $a_r = 1/(2q+1)$ for $r = -q, \ldots, +q$, and the smoothed value of x_t is given by

$$\mathrm{Sm}(x_t) = \frac{1}{2q+1} \sum_{r=-q}^{+q} x_{t+r}. \quad (1)$$

  - Another symmetric example is provided by the case where the $\{a_r\}$ are successive terms in the expansion of $(\frac{1}{2} + \frac{1}{2})^{2q}$. Thus when $q = 1$, the weights are $a_{-1} = a_{+1} = \frac{1}{4}, a_0 = \frac{1}{2}$. As q gets large, the weights approximate to a normal curve.
- Question: How to choose the window size?

Analysing series with trend but no seasonal variations

- Define B as the backward shift opterator such that $Bx_t = x_{t-1}$.

Example 1 (Backward operators)

Compute the series that results from the following operators: (a) $[1 - 2B + 3B^2]x_t$, (b) $(1 - B)(1 - 3B)x_t$, (c) $(1 - B - B^2 + B^3)x_t$ (d) $\frac{1}{1-\alpha B} X_t$ ($|\alpha| < 1$, $\alpha \neq 0$).

- Differencing. A special type of filtering is simply to difference a given time series until it becomes stationary, which is an integral part of the so-called Box-Jenkins procedure.

Analysing series with trend but no seasonal variations

The differencing operator can be written as $\nabla := 1 - B$. Then
$\nabla x_t = (1-B)x_t = x_t - x_{t-1}$,
$\nabla^2 x_t = (1-B)^2 x_t = (1-B)\nabla x_t = x_t - 2x_{t-1} + x_{t-2}$, and
$\nabla^j x_t = (1-B)^j x_t = \nabla^{j-1}\nabla x_t$, $j \geq 3$.

- For non-seasonal data, first-order differencing is usually sufficient to attain apparent stationarity. Here a new series, say $\{y_2, \dots, y_N\}$, is formed from the original observed series, say $\{x_1, \dots, x_N\}$, by $y_t = x_t - x_{t-1} = \nabla x_t$ for $t = 2, 3, \dots, N$.
- Occasionally second-order differencing is required using the operator ∇^2.

Analysing Series that Contain Seasonal Variation

- Three seasonal models in common use
  - Additive seasonality: $X_t = m_t + S_t + \epsilon_t$
  - Multiplicative seasonality: $X_t = m_t S_t + \epsilon_t$
  - Multiplicative seasonality and error: $X_t = m_t S_t \epsilon_t$
- The seasonality indices $\{S_t\}$ are usually assumed that $S_t = S_{t-s}$, where s is the period of the cyclic behavior.
- The indices $\{S_t\}$ are usually normalized to that $\sum_{t=1}^{s} S_t = 0$ in the additive case or $\prod_{t=1}^{s} S_t = 1$ in the multiplicative case.
- The analysis of time series with seasonal variation depends on whether the purpose is to measure the seasonal effect and/or to eliminate seasonality.

Seasonal Effect Estimation and Elimination

Assume that the seasonal part has a period of d (i.e., $S_{t+d} = S_t$ and $\sum_{j=1}^{d} S_j = 0$).

- Moving average method: We first estimate the trend part by a moving average filter running over a complete cycle so that the effect of the seasonality is averaged out. Depending on whether d is odd or even, we perform one of the following two steps for $t = q+1, \ldots, n-q$,
  - If $d = 2q$, $\text{Sm}(x_t) = \frac{1}{d}(\frac{1}{2}x_{t-q} + x_{t-q+1} + \cdots + x_{t+q-1} + \frac{1}{2}x_{t+q})$.
  - If $d = 2q+1$, $\text{Sm}(x_t) = \frac{1}{d}(x_{t-q} + x_{t-q+1} + \cdots + x_{t+q-1} + x_{t+q})$.

  The seasonal effect can then be estimated by calculating $x_t - \text{Sm}(x_t)$ or $x_t/\text{Sm}(x_t)$ for the additive or multiplicative case.
- Seasonal differencing: Use the dth differencing of data $\nabla_d = x_t - x_{t-d}$.

Autocorrelation and the Correlogram

- Sample autocorrelation coefficients measure the correlation between observations at different distances apart.
- Given N observations $x_1, \ldots, x_N$ on a time series, we can find the correlation between observations that are k steps apart, or the autocorrelation coefficient at lag k,

$$r_k = \frac{\sum_{t=1}^{N-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{N}(x_t - \bar{x})^2}$$

  where $\bar{x} = \sum_{t=1}^{N} x_t/N$.
- r_k can be calculated by autocovariance coefficient at lag k,

$$c_k = \frac{1}{N}\sum_{t=1}^{N-k}(x_t - \bar{x})(x_{t+k} - \bar{x})$$

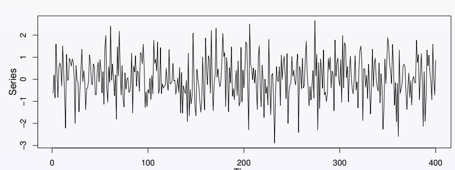
  We then compute $r_k = c_k/c_0$ for $k = 1, \ldots, M$ where $M < N$.

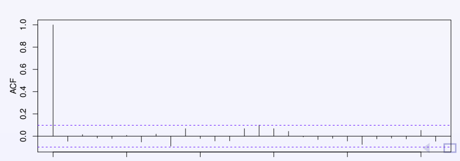
Interpreting the correlogram

- A correlogram is the plot of the sample autocorrelation coefficients r_k against the lag k for $k = 0, 1, \ldots, M$, where M is usually much less than N.
- For example if $N = 200$, then the analyst might look at the first 20 or 30 coefficients.
- Note that r_0 is always unity, but is still worth plotting for comparative purposes. The correlogram may alternatively be called the sample autocorrelation function (ac.f.).

Interpreting the correlogram: Random series

- Random series: A series of independent observations having the same distribution. For large N, we expect to find that $r_k \approx 0$ for all non-zero values of k.
- We can show that, for a random time series, r_k, $k \geq 1$, is approximately N$(0, 1/N)$. Thus, if a time series is random, we can expect 19 out of 20 of the values of r_k to lie between $\pm 1.96/\sqrt{N}$. As a result, it is common practice to regard any values of r_k outside these limits as being 'significant'.





Interpreting the correlogram: Short-term series

- Short-term series: Stationary series that exhibit short-term correlation are characterized by large values of r_k for small lags followed by values of r_k being approximately 0 for larger lags.

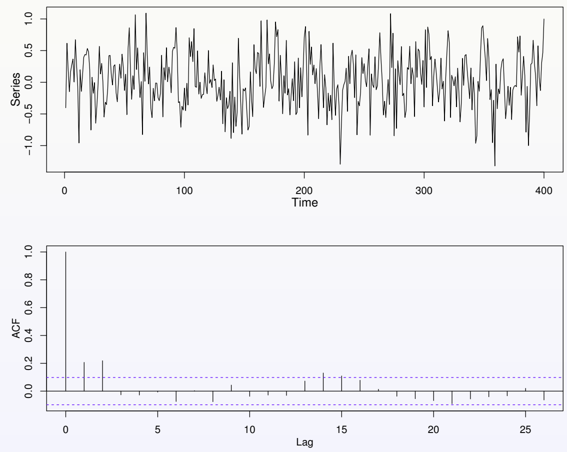


Figure 2.3: A time series showing short-term correlation together with its correlogram.

Interpreting the correlogram: Non-stationary series

- If a time series contains a trend, then the values of r_k will not come down to zero except for very large values of the lag. This is because an observation on one side of the overall mean tends to be followed by a large number of further observations on the same side of the mean because of the trend.
- A typical non-stationary time series together with its correlogram is shown in Figure 2.4. Little can be inferred from a correlogram of this type as the trend dominates all other features.
- In fact the sample ac.f. $\{r_k\}$ is only meaningful for data from a stationary time-series model and so any trend should be removed before calculating $\{r_k\}$.

Interpreting the correlogram: Non-stationary series

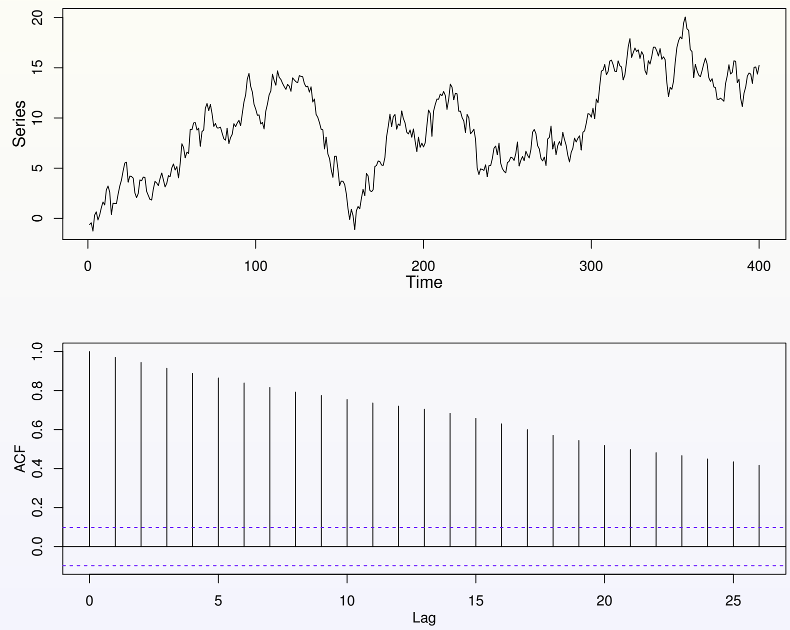
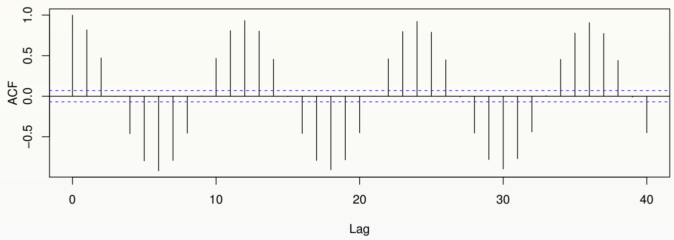


Figure 2.4: A non-stationary time series together with its correlogram.

Interpreting the correlogram: Seasonal series

- If a time series contains seasonal variation, then the correlogram will also exhibit oscillation at the same frequency.
- The top panel of Figure 2.5 shows the correlogram of the monthly air temperature data shown in Figure 1.3. The sinusoidal pattern of the correlogram is clearly evident, but for seasonal data of this type the correlogram provides little extra information, as the seasonal pattern is usually displayed in the time plot of the data.
- If the seasonal variation is removed from seasonal data, then the correlogram may provide useful information.

Interpreting the correlogram: Seasonal series



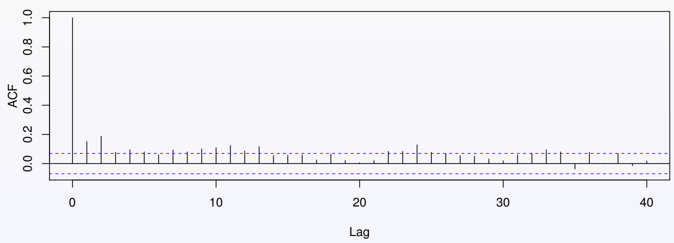


Figure 2.5: The correlograms of monthly observations on air temperature in Anchorage, Alaska for the raw data (top) and for the seasonally adjusted data (bottom).