1 Differencing for Trend Elimination

In the last class, we studied trend models: \( X_t = m_t + Z_t \) where \( m_t \) is a deterministic trend function and \( \{Z_t\} \) is white noise.

The residuals obtained after fitting the trend function \( m_t \) in the model \( X_t = m_t + Z_t \) are studied to see if they are white noise or have some dependence structure that can be exploited for prediction.

Suppose that the goal is just to produce such detrended residuals. Differencing is a simple technique which produces such de-trended residuals.

One just looks at \( Y_t = X_t - X_{t-1}, t = 2, \ldots, n \). If the trend \( m_t \) in \( X_t = m_t + Z_t \) is linear, then this operation simply removes it because if \( m_t = \alpha t + b \), then \( m_t - m_{t-1} = \alpha \) so that \( Y_t = \alpha + Z_t - Z_{t-1} \).

Suppose that the first differenced series \( Y_t \) appears like white noise. What then would be a reasonable forecast for the original series: \( X_{n+1} \)? Because \( Y_t \) is like white noise, we forecast \( Y_{n+1} \) by the sample mean \( \bar{Y} := (Y_2 + \cdots + Y_n)/(n-1) \). But since \( Y_{n+1} = X_{n+1} - X_n \), this results in the forecast \( X_n + \bar{Y} \) for \( X_{n+1} \).

Sometimes, even after differencing, one can notice a trend in the data. In that case, just difference again. It is useful to follow the notation \( \nabla \) for differencing:
\[
\nabla X_t = X_t - X_{t-1} \quad \text{for} \ t = 2, \ldots, n
\]
and second differencing corresponds to
\[
\nabla^2 X_t = \nabla(\nabla X_t) = \nabla X_t - \nabla X_{t-1} = X_t - 2X_{t-1} + X_{t-2} \quad \text{for} \ t = 3, \ldots, n.
\]

It can be shown that quadratic trends simply disappear with the operation \( \nabla^2 \). Suppose the data \( \nabla^2 X_t \) appear like white noise, how would you obtain a forecast for \( X_{n+1} \)?

Differencing is a quick and easy way to produce detrended residuals and is a key component in the ARIMA forecasting models (later). A problem however is that it does not result in any estimate for the trend function \( m_t \).

2 Random Walk with Drift Model

Consider the following model for \( X_t \):
\[
R_t = \delta + R_{t-1} + W_t
\]
for \( t = 1, 2, \ldots \), with initial condition \( R_0 = 0 \) and \( W_t \) being white noise. This model can also be written as:

\[
R_t = \delta t + \sum_{j=1}^{t} W_j
\]

When \( \delta = 0 \), this model is called the Random Walk model. This is used often to model trend. This would be a stochastic model for trend as opposed to the previous ones which are deterministic models.

Consider the model \( X_t = m_t + Z_t \) where \( Z_t \) is white noise and \( m_t \) is a random walk with drift:

\[
m_t = \delta + m_{t-1} + W_t
\]

This is an example of a Dynamic Linear Model (DLM). \( W_t \) is called evolution error and \( Z_t \) is called observational error.

The differenced series for \( X_t \) is:

\[
\nabla X_t = X_t - X_{t-1} = m_t - m_{t-1} + Z_t - Z_{t-1} = \delta + W_t + Z_t - Z_{t-1}.
\]

Therefore, \( \nabla X_t \) is a detrended series. Ways for modelling detrended series through stationary models will be studied later.

3 Models for Seasonality

Many time series datasets exhibit seasonality. Simplest way to model this is: \( X_t = s_t + Z_t \) where \( s_t \) is a periodic function of a known period \( d \) i.e., \( s_{t+d} = s_t \) for all \( t \). Such a function \( s \) models seasonality. These models are appropriate, for example, to monthly, quarterly or weekly data sets that have a seasonal pattern to them.

This model, however, will not be applicable for datasets having both trend and seasonality which is the more realistic situation. These will be studied a little later.

Just like the trend case, there are three different approaches to dealing with seasonality: fitting parametric functions, smoothing and differencing.

3.0.1 Fitting a parametric seasonality function

The simplest periodic functions of period \( d \) are: \( a \cos(2\pi ft/d) \) and \( a \sin(2\pi ft/d) \). Here \( f \) is a positive integer. The quantity \( a \) is called Amplitude and \( f/d \) is called frequency and its inverse, \( d/f \) is called period. The higher \( f \) is, the more rapid the oscillations in the function are.

More generally,

\[
s_t = a_0 + \sum_{f=1}^{k} (a_j \cos(2\pi ft/d) + b_j \sin(2\pi ft/d)) \tag{1}
\]

is a periodic function. Choose a value of \( k \) (not too large) and fit this to the data.

For \( d = 12 \), there is no need to consider values of \( k \) that are more than 6. With \( k = 6 \), every periodic function with period 12 can be written in the form (1). More on this when we study the frequency domain analysis of time series.

3.0.2 Smoothing

Because of periodicity, the function \( s_t \) only depends on the \( d \) values \( s_1, s_2, \ldots, s_d \). Clearly \( s_1 \) can be estimated by the average of \( X_1, X_{1+d}, X_{1+2d}, \ldots \). For example, for monthly data, this corresponds to
estimating the mean term for January by averaging all January observations. Thus

\[ \hat{s}_i := \text{average of } X_i, X_{i+d}, X_{i+2d}, \ldots \]

Note that here, we are fitting 12 parameters (one each for \(s_1, \ldots, s_d\)) from \(n\) observations. If \(n\) is not that big, fitting 12 parameters might lead to overfitting.

### 3.0.3 Differencing

How can we obtain residuals adjusted for seasonality from the data without explicitly fitting a seasonality function? Recall that a function \(s\) is a periodic function of period \(d\) if \(s_{t+d} = s_t\) for all \(t\). The model that we have in mind here is:

\[ X_t = s_t + Z_t. \]

Clearly \(X_t - X_{t-d} = s_t - s_{t-d} + Z_t - Z_{t-d} = Z_t - Z_{t-d}\). Therefore, the lag-\(d\) differenced data \(X_t - X_{t-d}\) do not display any seasonality. This method of producing deseasonalized residuals is called **Seasonal Differencing**.

### 4 Data Transformations

Suppose that the time series data set has a trend and that the variability increases along with the trend function. An example is the UKgas dataset in R. In such a situation, transform the data using the logarithm or a square root so that the resulting data look reasonably homoscedastic (having the same variance throughout).

Why log or square root? It helps to know a little bit about variance stabilizing transformations. Suppose \(X\) is a random variable having mean \(m\). A very heuristic calculation gives an approximate answer for the variance of a function \(f(X)\) of the random variable \(X\)? Expand \(f(X)\) in its Taylor series up to first order around \(m\):

\[ f(X) \approx f(m) + f'(m)(X - m) \]

As a result,

\[ \text{var}(f(X)) \approx \text{var}(f(m) + f'(m)(X - m)) = (f'(m))^2\text{var}(X). \]

Thus if

1. \(\text{var}(X) = Cm\) and \(f(x) = \sqrt{x}\), we would get \(\text{var}(X) \approx C/4\).
2. \(\text{var}(X) = Cm^2\) and \(f(x) = \log x\), we would get \(\text{var}(X) \approx C\).

The key is to note that in both the above cases, the approximate variance of \(f(X)\) does not depend on \(m\) anymore.

The above rough calculation suggests the following insight into time series data analysis. A model of the form \(X_t = m_t + W_t\) where \(m_t\) is a deterministic function and \(W_t\) is purely random or stationary (next week) assumes that the variance of \(X_t\) does not vary with \(t\). Suppose however that the time plot of the data shows that the variance of \(X_t\) increases with its mean \(m_t\), say \(\text{var}(X_t) \propto m_t\). Then the rough calculation suggests that \(\text{var}(\sqrt{X_t})\) should be approximately constant (does not depend on \(t\)) and hence the model \(m_t + W_t\) should be fit to the transformed data \(\sqrt{X_t}\) instead of the original data \(X_t\). Similarly, if \(\text{var}(X_t) \propto m_t^2\), then \(\text{var}(\log X_t)\) should be approximately constant.

Thus, if the data show increased variability with a trend, then apply a transformation such as log or square root depending on whether the variability in the resulting data set is constant across time.
By the way, count data are usually modelled via Poisson random variables and the variance of a Poisson equals its mean. So one typically works with square roots while dealing with count (Poisson) data.

If one uses the model $X_t = m_t + W_t$ with a non-deterministic (stochastic) trend function $m_t$, this automatically allows for $X_t$ to have a variance that changes with $t$. In that case, we may not need to use transformations on the data. These models can be seen as special cases of State Space Models that we will briefly look at later.

**Box-Cox transformations**: The square-root and the logarithm are special cases of the Box-Cox Transformations given by:

$$Y_t = \begin{cases} 
\frac{X_t^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\
\log X_t & \text{if } \lambda = 0.
\end{cases}$$

(2)

Square root essentially corresponds to $\lambda = 1/2$. 

\[ \]