Spring 2013 Statistics 153 (Time Series) : Lecture One

Aditya Guntuboyina

22 January 2013

A time series is a set of numerical observations, each one being recorded at a specific time. Time series data arise everywhere. The aim of this course is to teach you how to analyze such data.

We will focus on two approaches to time series analysis: (1) Time Domain Approach, and (2) Frequency Domain Approach (also known as the spectral or Fourier analysis of time series). In the Time domain approach, one works directly with the data while in the Frequency Domain approach, one works with the Discrete Fourier Transform of the data.

Very roughly, 60 percent of the course will be on time domain methods and 40 percent will be on frequency domain methods.

1 Time Series Models

The primary objective of time series analysis is to develop mathematical models that provide plausible descriptions for sample time series data.

We assume that the observed data $x_1, \dots, x_n$ is a realization of a sequence of random variables $X_1, \dots, X_n$. The basic strategy of modelling is always to start simple and to build up.

We will study models for time series in this class. Basic Models:

1. White Noise
2. Deterministic trend + white noise
3. Deterministic seasonality + white noise
4. Deterministic trend + deterministic seasonality + white noise
5. Stationary time series models
6. Stationary ARMA models
7. ARIMA models
8. Seasonal ARIMA models
9. Modeling and estimating Spectral Density

1.1 White Noise

$X_1, \ldots, X_n$ are called white noise if they have mean zero, variance σ^2 and are uncorrelated.

An important special case of white noise is Gaussian White Noise where $X_1, \ldots, X_n$ are i.i.d $N(0, \sigma^2)$.

How to check if white noise is a good model for a given dataset? Think in terms of forecasting. For white noise, the given data cannot help in predicting X_{n+1}. The best estimate of X_{n+1} is $E(X_{n+1}) = 0$. In particular, X_1 cannot predict X_2; X_2 cannot predict X_3 and so on. Therefore, the correlation coefficient between $Y = (X_1, \ldots, X_{n-1})$ and $Z = (X_2, \ldots, X_n)$ must be close to zero.

The formula for the correlation between Y and Z is:

$$r := \frac{\sum_{t=1}^{n-1}(X_t - \bar{X}_{(1)})(X_{t+1} - \bar{X}_{(2)})}{\sqrt{\sum_{t=1}^{n-1}(X_t - \bar{X}_{(1)})^2 \sum_{t=1}^{n-1}(X_{t+1} - \bar{X}_{(2)})^2}}$$

where

$$\bar{X}_{(1)} = \frac{\sum_{t=1}^{n-1} X_t}{n-1} \text{ and } \bar{X}_{(2)} = \frac{\sum_{t=1}^{n-1} X_{t+1}}{n-1}.$$

This formula is usually simplified to obtain

$$r_1 = \frac{\sum_{t=1}^{n-1}(X_t - \bar{X})(X_{t+1} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2}$$

where $\bar{X} := \sum_{t=1}^{n} X_t/n$. Note that we are calling this correlation r_1 (note the subscript 1). This quantity r_1 is called the sample autocorrelation coefficient (sample ACF) of $X_1, \ldots, X_n$ at lag one. Lag one because this correlation is between X_t and X_{t+1}.

When $X_1, \ldots, X_n$ are obtained from white noise, r_1 is close to zero, particularly when n is large.

One can similarly define sample autocorrelations at other lags:

$$r_k = \frac{\sum_{t=1}^{n-k}(X_t - \bar{X})(X_{t+k} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2} \qquad \text{for } k = 1, 2, \ldots$$

Here is an important mathematical fact: Under certain additional conditions (which are satisfied for gaussian white noise), if $X_1, \ldots, X_n$ are white noise, then the sample autocorrelations $r_1, r_2, \ldots$ are **independently** distributed according to the normal distribution with mean zero and variance $1/n$.

Note that the variance decreases to zero as n increases and the mean is zero. Thus for large n, the sample autocorrelations should be very close to zero. Also note that the sample autocorrelations for different lags are independent.

Therefore, one way of checking if the white noise model is a good fit to the data is to plot the sample autocorrelations. This plot is known as the **correlogram**. Use the function **acf** in **R** to get the correlogram. The blue bands in the correlogram correspond to levels of $\pm 1.96 n^{-1/2}$.

How to interpret the correlogram? When $X_1, \ldots, X_n$ are white noise, the probability that a fixed r_k lies outside the blue bands equals 0.05. A value of r_k outside the blue bands is **significant** i.e., it gives evidence against pure randomness. However, the overall probability of getting atleast one r_k outside the bands increases with the number of coefficients plotted. For example, if 20 r_ks are plotted, one expects to get one significant value under pure randomness.

Here are a couple rules of thumb for deciding if a correlogram indicates departure from white noise:

- A single r_k just outside the bands may be ignored but two or three values well outside indicate a departure from pure randomness.

- A single significant r_k at a lag which has some physical interpretation such as lag one or a lag corresponding to seasonal variation also indicates evidence of non-randomness.