Discrete Stochastic Modelling

Time Series Analysis

1. MINITAB

We shall be using MINITAB for statistical analysis, as the standard version of Excel lacks the appropriate statistical capability. MINITAB is a statistical package intended for educational use and is able to handle the kind of procedures we have in mind. These notes were written using MINITAB 10.5; more advanced formats may require a change in the notes.

When starting MINITAB, notice that it uses a number of windows. The most important are the Session window, where commands and text output appear, and the Data window, which shows data in a spreadsheet format. Each graphic also appears in its own window. Text and graphics can be pasted into a Word document. Remember to use a fixed-width font like Courier New for text.

2. Data

The data files for this session can all be downloaded from the web page http://www.staff.city.ac.uk/r.j.gerrard/courses/3ts/datasets.html, except the DAX data, which are at http://www.staff.city.ac.uk/r.j.gerrard/courses/2dsm/ . I shall try to put a MINITAB worksheet on the T: drive to make life easier, but I may not have time. If not, you can just use the web page.

When downloading files in text format, as most of them are, it is best to save them first in your local file space (all are fairly small files), then to load them into MINITAB using "Open Worksheet…".

3. Plots

Start by opening the tree-ring data, which represent the widths of the rings on an Australian pencil pine from 1028 until 1975 (data collected by Lamarche, Ogden, Campbell and Dunwiddie, 1977). From the Graph menu choose Time Series Plot. The first time you use this, no variable is selected for plotting, so double-click the "treering" entry in the list of data columns, then press OK, and you will get a time series plot of the data.

You can't tell very much from the TS plot in this case – it is used for detecting periodic variation or trend – so the next thing to do is to look at the sample ACF and sample PACF, both of which are available on the Stat | Time Series menu. Each of these has options, but choose the default setting to start with.

You will see that the sample ACF decreases more or less geometrically, and that the sample PACF is not significantly different from 0 for $k > 3$. The issue is slightly obscured by the huge number of lags for which the sample (P)ACF is calculated: feel free to produce the plot again with no more than 20 lags (set the appropriate option).

4. ARIMA model fitting

The SACF/SPACF seem to suggest that an ARMA(3,0) is a suitable model. On the Stat | Time Series menu you will find the ARIMA.. item; this invites you to specify which data set is to be fitted and asks you for p, d and q: use 3, 0 and 0. In addition, tick the box which instructs MINITAB to store the residuals for later use. Store the fitted values too.

You see that MINITAB presents in the session window a set of iterative estimates of the parameter values, followed by the final estimates and t-ratios, then such diagnostics as sum of squares of the residuals and the Ljung-Box χ^2 statistic.

If you observe the Data window you find that the residuals have been stored under the name RESI1, the fitted values as FITS1. Plot residuals against fitted values (Graph | Plot, then fill in whichever series you want for the y-value and the x-value), then produce a histogram of the residuals (Graph | Histogram) to check for normality. It is also possible to get a normal probability plot (Graph | Normal plot).

5. Stationarity

Look now at the DAX data set, which depicts the closing prices of the DAX index every day for years. The SACF of the DAX data is close to 1 for many lags, indicating that differencing is a sensible procedure. So you need to look at the first difference of the DAX data: Stat | Time Series has the procedure, called Difference. You will need a lag-1 difference.

Now look at a time series plot of the first difference data. Does it look stationary? It looks to me as through the residuals are larger when the fitted values are larger; thus a data transformation is in order: logarithmic is traditional, but you might like to experiment with something like square root and see what happens.

Calc | Functions is the menu required here. Choose the data set to alter, the function required and the column in which to store the transformed values. The transformed data will still need differencing, but the plot of the differenced data is much more acceptable and the SACF/SPACF look much better. Decide which ARMA model is appropriate for the differenced data and fit it as before. Look at the residuals and decide whether the model fits the data. (Note that, instead of differencing the data and fitting ARMA, it would have been possible to fit ARIMA directly.)

6. Seasonal variation

A data set with a built-in seasonal component needs to have an appropriate model.

Look at the hotel occupancy data. Analysis of both trend and seasonal component can be carried out by the routine Stat | Time Series | Decomposition. You will need to specify the period of the seasonal variation: 12 for monthly data, 4 for quarterly, etc. Make sure you save the residuals, then try to fit an ARMA model to them.

Instead of doing the seasonal decomposition followed by fitting an ARMA model, it is possible to fit a seasonal ARIMA model (SARIMA) directly. A SARIMA$(p, d, q)\times(P, D, Q)$ process normally has too many parameters to make it attractive: in addition to the standard ARIMA parameters it has P seasonal AR parameters, Q seasonal MA parameters and is seasonally differenced D times. I recommend setting $P = Q = 0, D = 1$.

7. Akaike

When fitting a variety of models to the same data, some objective criterion is required for determining best fit. You may well find that there is more than one model which leaves residuals whose SACF and SPACF look right and which appear normally distributed.

MINITAB does not implement Akaike's Information Criterion directly, but it does print the residual sum of squares associated with each model, so that $n\log(\mathrm{RSS})+2(p+q)$ can be calculated for each fitted model.

Work on the Colorado river data, fitting a variety of models which seem appropriate and recording the residual sum of squares and the value of $n\log(\mathrm{RSS})+2(p+q)$.