STAT8008: Times Series & MV Analysis

Autoregressive Integrated Moving Average (ARIMA)

Justin McGuinness

Cork Institute of Technology

What is ARIMA?

- The assumption of this approach for forecasting is that the common variation in the series can be divided into three components:

1. Autoregressive (AR)
2. Integrated (I) or Difference
3. Moving Average (MA)

- An ARIMA model can have one, two, or all three of these components.
- In addition, these components can operate at both the seasonal and nonseasonal level. For example, sales this month may be related to sales last month (nonseasonal) as well as sales one year ago (seasonal).
- There are many different types of ARIMA models but the general form of the model is ARIMA(*p*,*d*,*q*)(P,D,Q).

ARIMA(p,d,q)(P,D,Q)

1. *p* refers to the order of the nonseasonal autoregressive process incorporated into the ARIMA model (and *P* is the order of the seasonal autoregressive process).

2. *d* refers to the order of integration or differencing (and *D* is the order of the seasonal integration or differencing).

3. *q* refers to the order of the moving average process incorporated in the model (and *Q* is the order of the seasonal moving average process).

Autoregressive

- In regression analysis we assume that one variable influences another variable.

- For example, advertising spending affects sales of a product.

- Alternatively in regression it might also be the case that **lagged values of one variable are a good predictor of another variable.**

- For example, the amount of advertising two periods ago is a good indicator of what sales are now.

Autoregressive

- In an analogous manner to regression, **ARIMA models use lagged values to predict the dependent variable**.

- The term ***autoregressive*** implies that a variable based on the actual dependent variable is used in some way to predict the dependent variable.

- More specifically, the autoregressive component of an ARIMA model uses the *lagged* values of the dependent variable as predictors of the current value of the dependent variable.

- For example, a good predictor of sales could be the number of sales in the previous time period. The lagged values of the dependent variable account for part of the model's fit.

Autoregressive

- There can be different ‘orders of autoregression’.

- The **order of autoregression** refers to:

  The time difference between the dependent variable and the lagged dependent variable, which is being used as a predictor.

Autoregressive

- If the dependent variable is influenced by the dependent variable lagged one time period, then this is an autoregressive model of order one and is sometimes called an AR(1) process.
- The AR(1) component of the ARIMA model is saying that the value of the dependent variable in the previous period (t–1) is a good indicator and predictor of what the dependent variable will be now (t).
- If on the other hand, a good predictor of a dependent variable is the dependent variable two periods previous, then the autoregressive process is of order two, an AR(2) process. This pattern continues for higher order processes.

Integrated

- The “I” part of the ARIMA model refers to Integrated. This term relates to whether a time series requires **differencing in order to become stationary**.

- The idea of stationarity is very important in ARIMA modeling.

- The dependent variable in an ARIMA model should be **stationary to meet the assumptions** of this technique.

- The Integration component of ARIMA is typically associated with **removing trend from the series**, which would violate the constant mean component of stationarity.

Integrated: Stationarity

In time series analysis the term stationarity is often used to describe **how a particular time series variable changes over time**.

Stationarity has **three components**:

- First, the series has a **constant mean**, which implies that there is no tendency for the mean of the series to increase or decrease over time.
- Second, the **variance of the series is assumed constant** over time. So, for example, if the magnitude of the seasonal swings increase over time, then a series is not stationary.
- Finally, any **autocorrelation pattern is assumed constant** throughout the series. For example, if there is an AR(2) pattern in the series, it is assumed to be present throughout the entire series.

Integrated: Stationarity

- Any **violation of stationarity creates estimation problems** for ARIMA models.

- It is often the case in time series analysis that the mean of a variable increases or decreases over time.

- For example, if your aim is to study how car ownership has changed over time, we know that ownership has increased greatly. Therefore, the mean of the series will also have increased over time. If the mean of the series changes over time then the variable is non-stationary.

- In order to make a series containing trend stationary, you can **create a new series that is the difference of the original series**.

Integrated

- A first order difference creates a value for the new series which is the difference between the series value in the current period minus the series value in the previous period.

- Often the differenced series will have a stationary mean. If the differenced series does not have a stationary mean then it might be necessary to take first differences of the differenced series. This transformation is known as second order differencing as the original series has now been differenced twice.

- The number of times a series needs to be differenced is known as the order of integration. Differencing can be performed at the seasonal (current time period value minus the value from one season ago) or non-seasonal (current time period minus the value from the previous time period).

Integrated: In Short

- **Differencing can remove trend** from a series in order to create the stationary series that forms the basis of ARIMA, and **integration later builds the trend back** into the series when predictions (forecasts) are produced.
- If a series is stationary then there is no need to difference the series and the order of integration is zero, thus the dependent variable should be left in its original values. In this case, ARIMA models will be of the form ARIMA (*p,0,q*) where *p* is the order of the autoregressive process and *q* is the order of the moving average process in the model.
- If a series is non-stationary then usually first differencing the series will make it stationary. If first differencing makes the series stationary then the order of integration is one and ARIMA models will be in the form of ARIMA (*p,1,q*).
- Very occasionally it might be necessary to difference a series twice to make it stationary in which case the order of integration is two. It is however nearly always the case that first differences will make a non-stationary series stationary.

Moving Average

- The autoregressive component of an ARIMA model uses lagged values of the dependent variable as independent variables.

- In contrast to this, the moving average component of the model uses lagged values of the **model error** as independent variables. It can be successful at extracting autocorrelated patterns in the series that would have otherwise been included in the model error.

Moving Average

- The **order of moving average refers to the lag length between the error and the dependent variable**. For example, if the dependent variable is influenced by the model's error lagged one period then this is a moving average process of order one and is sometimes called an MA(1) process. The MA(1) component of the ARIMA model is saying that the model's error in the previous period is related to what the dependent variable will be now.

- Consider a time series model with a tendency for a positive error to be followed two periods later by an increase in the series value. A moving average of order two would use the positive error two periods previous in order to better predict this pattern. Techniques such as regression can not extract this variation as part of its model fit and as a result the model error in regression would be autocorrelated.

ARIMA Equation

- Let Y be the dependent series transformed to stationarity (i.e. the differenced series of order d). Then the general form of the model is:

$$Y_t = f_1 Y_{t-1} + f_2 Y_{t-2} + \cdots + f_p Y_{t-p} + \epsilon_t - b_1 \epsilon_{t-1} - b_2 \epsilon_{t-2} - \cdots - b_q \epsilon_{t-q}$$

- Y is predicted by its own past values along with current and past errors.
- The challenge with any dependent series is identifying the order of differencing and seasonal differencing, the order of autoregressive parameters, both nonseasonal and seasonal, and the order of moving average parameters, both nonseasonal and seasonal.

Identifying the type of ARIMA model

- There are **many different ARIMA models** which could be fit to a particular time series of interest. The type of ARIMA model depends upon the selected orders of autoregression (*p*), integration (*d*), and moving average (*q*).

- It is often the case that identifying the *p*, *d*, and *q* combination to give the "best-fit or forecasting" ARIMA model is a process of trial and error. In many ways the identification stage is by far the most subjective in the entire ARIMA modeling process. The Expert Modeler will, if you wish, do the identification for you (but you should not always trust it blindly!)

Identifying the type of ARIMA model

- The identification stage involves using exploratory techniques (sequence charts and autocorrelation and partial autocorrelation plots) in order to determine the most likely combination of *p, d,* and *q* that will give the closest fit to the historic data.

- The **order of integration (*d*) can usually be identified by looking at sequence charts** for the dependent variable (or the dependent variable after differencing).

- **Autocorrelation and partial autocorrelation plots of the dependent variable are used to suggest plausible values for *p* and *q***, the orders of autoregression and moving average.

Identifying the Order of Integration

- Recall that a series is stationary when there is no tendency for the **mean of a variable to increase or decrease over time**, the variance is homogeneous, and the autocorrelation pattern is constant throughout the series (although it is difficult to assess this last requirement).

- Here we focus on problems with stationary violations due to mean shifts.

- If the series is stationary then the order of integration (*d*) is zero and there is no need to difference the series. The type of ARIMA model will be an ARIMA(*p,0,q*).

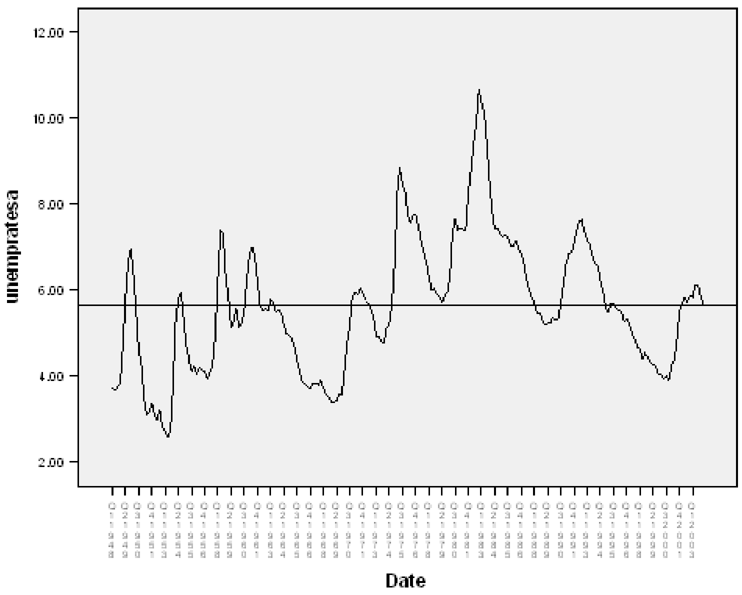
Identifying the Order of Integration

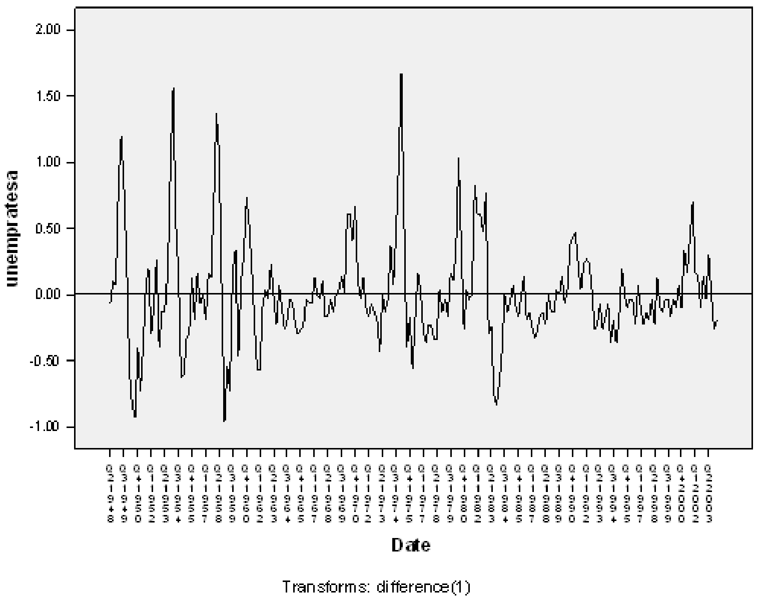
- If the series is non-stationary due to trend, then the next stage is to run a sequence chart of the first differenced values of the dependent variable. Usually, first differencing the series will make it stationary. If the sequence chart of the first differenced variable shows that the first differences are stationary, then the order of integration is one. The type of ARIMA model will be an ARIMA(*p,1,q*).

- Finally, occasionally it will be necessary to second difference the series in order to make it stationary. If this is the case the type of ARIMA model is an ARIMA(*p,2,q*).

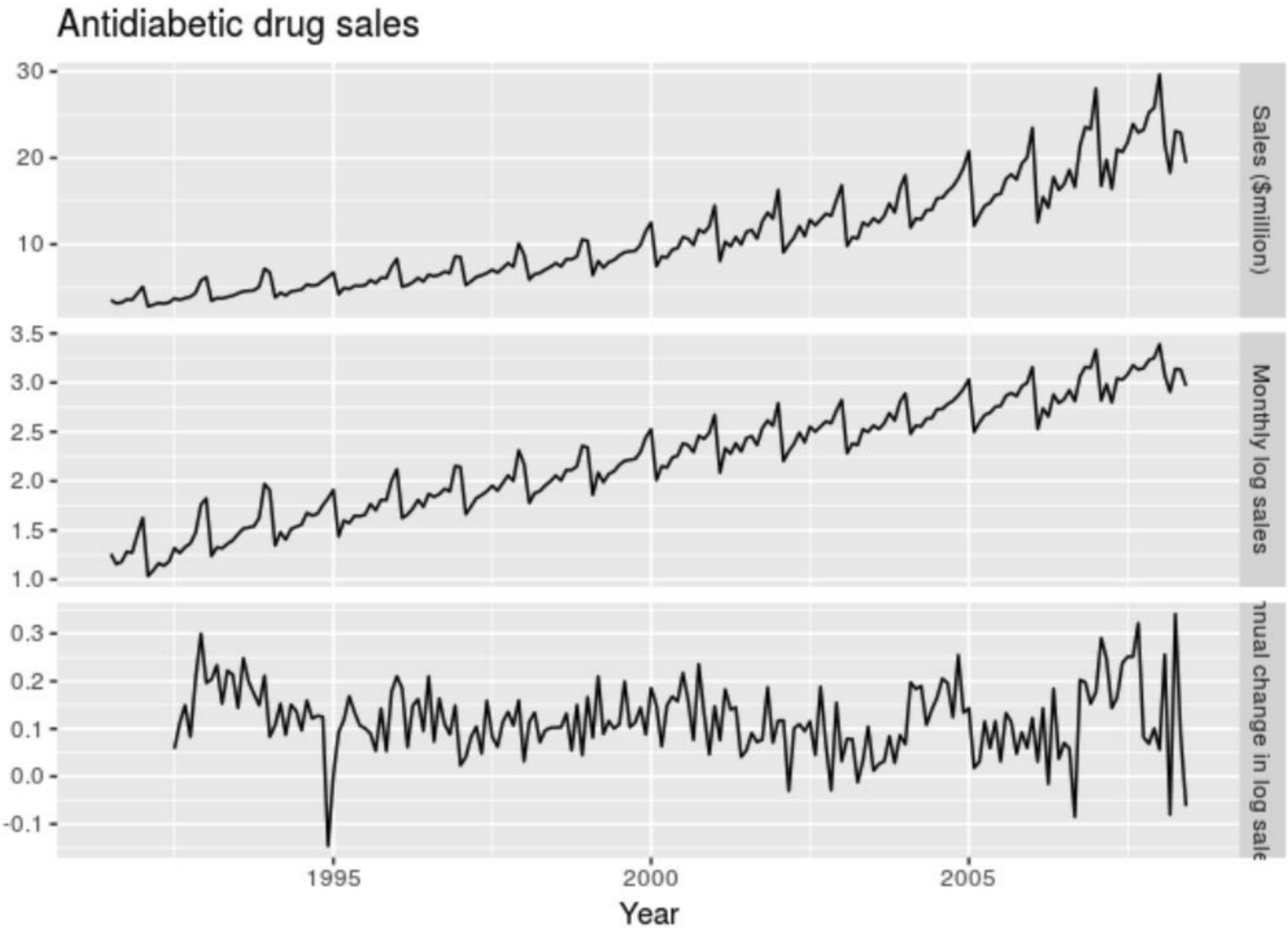
Identifying the Order of Integration

Zero order difference versus first order difference





Identifying the Order of Integration



Identifying the Autoregressive (*p*) and Moving Average (*q*) Orders

- The exploratory process for identifying the orders of autoregression and moving average is the subjective part of model identification.

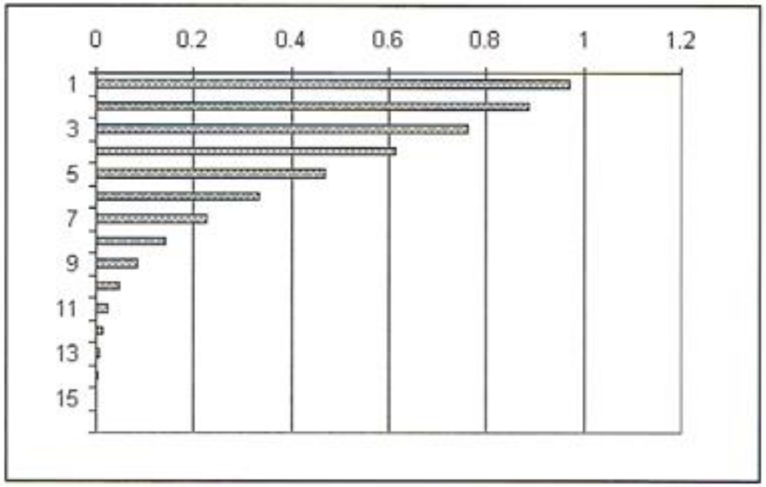
- Identification of possible autoregressive (*p*) and moving average (*q*) orders requires examination of the autocorrelation and partial autocorrelation functions for the dependent variable

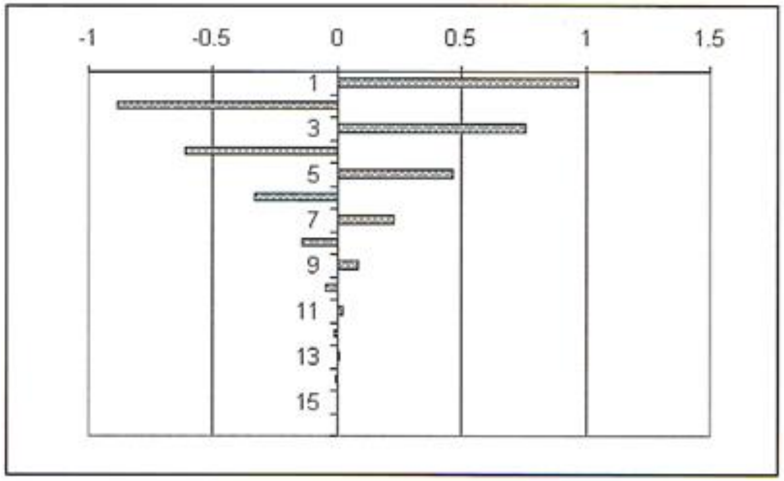
Identifying an Autoregressive Process

- ARIMA models which have an autoregressive component but no moving average component - i.e., ARIMA(p,d,0) models, have **exponentially or sine wave declining autocorrelation functions**.

- The plots assume any necessary differencing has already been performed

Autocorrelation Function (ACF) Plots

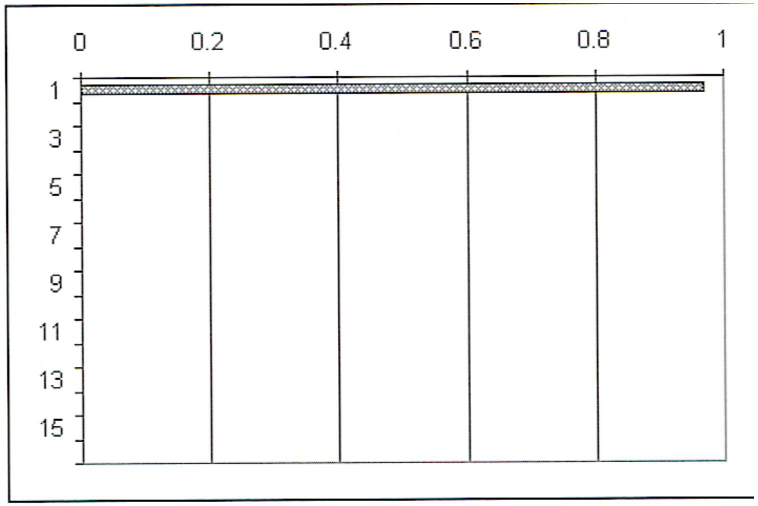


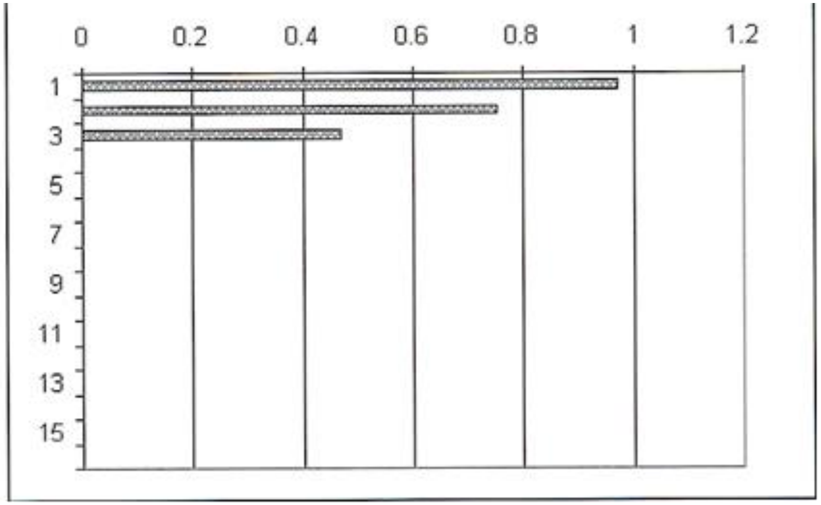


Identifying an Autoregressive Process

- A pure autoregressive process with no moving average component is also characterised by significant bars (spikes) on the low lags of the partial autocorrelation functions followed by a sudden stop in the significance of the bars for all subsequent lags. **The number of significant bars (spikes) indicates the order of the autoregressive process.**
- If the ARIMA model followed a first order autoregressive process—an ARIMA(1,d,0)—the partial autocorrelation function will have one spike followed by a sudden decline to zero for all subsequent lagged bars.
- Similarly, if the ARIMA model followed a third order autoregressive process, an ARIMA(3,d,0), the partial autocorrelation function will have three spikes followed by a sudden decline to zero for all subsequent lagged bars.

Partial Autocorrelation Function (PACF) Plots

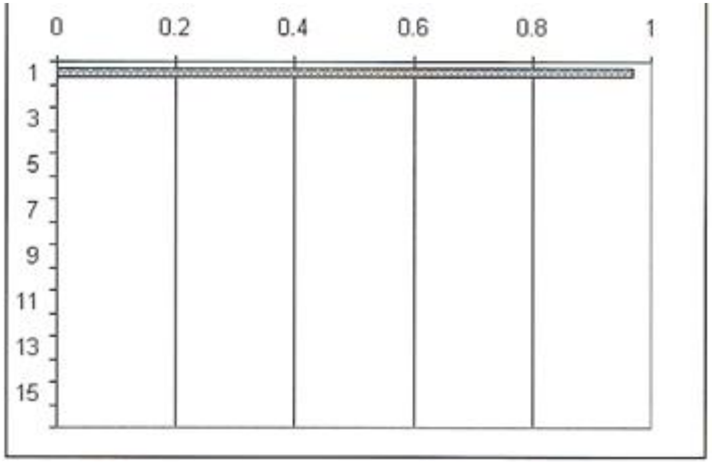




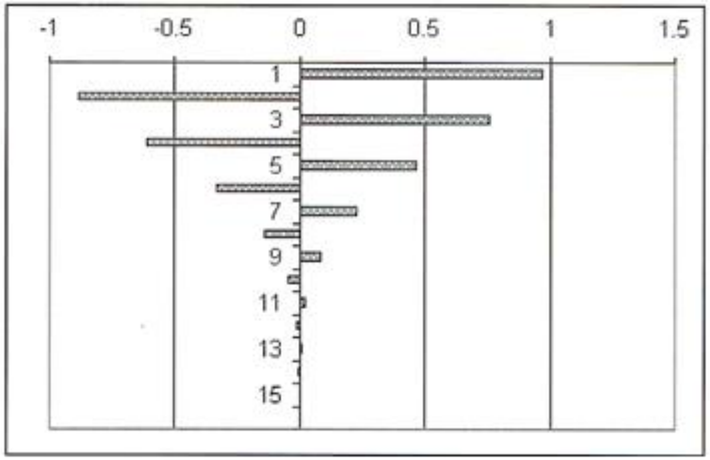
Identifying a Moving Average Process

- In contrast to the autogressive process, a **moving average process has the spikes in the autocorrelation function and the declining exponential or sine wave in the partial autocorrelation function**.
- The figures show the respective plots for a moving average process of order one - i.e., an ARIMA(0,d,1).
- In the autocorrelation plot, the number of spikes indicates the order of the moving average process. In this case, there is one spike for the autocorrelation function along with an exponentially declining or sine wave declining partial autocorrelation plot.

ACF Plot



PACF Plot



Identifying an ARIMA Model in Practice

- **Parameter estimates**. After estimation, all parameter estimates should be **statistically significant**. You might remove non-significant parameter estimates from the model and estimate a simpler model. You can tentatively add parameters to the model, but the added parameter estimates should be statistically significant.
- **Residual ACF and PACF.** The residual ACF and PACF correlations should be small and nonsignificant. By chance, you will occasionally observe significant autocorrelations associated with a random series, but you should pay special attention to significant autocorrelations at the first few lags as well as seasonal lags.
- **Box-Ljung statistic.** The Box-Ljung statistic is a statistic associated with residual autocorrelations up to and including a given lag. The Box-Ljung statistic tests the null hypothesis that the autocorrelations from lag 1 to the given lag are collectively associated with a white noise process.

Identifying an ARIMA Model in Practice

- **Model fit statistics.** Fit statistics such as the mean absolute error are measured in the same metric as the dependent series. Small values of these statistics are associated with better-fitting models. But, take care that such better fit is not attained through adding unnecessary parameters to the model.

- **Information criteria.** There are a number of information criteria with names such as the Akaike Information Criterion (AIC), the Schwartz Bayesian Information Criterion (SBC or BIC), the normalized Bayesian Information Criterion (normalized BIC). These criteria combine the sum of squared residuals and the number of parameter estimates in various fashions to produce a criterion value. The rule for any of these criteria is to select the model with the minimum value.