AUTOREGRESSIVE INTEGRATED MOVING AVERAGE MODELS

STA 3050: Time Series and Forecasting

In the previous sections of this unit, you have learnt different time series models such as autoregressive (AR), moving average (MA), and autoregressive moving average (ARMA). These models assume that the time series is stationary. However, in the real world, most time series variables are nonstationary.

In general, trends and periodicity exist in many time series data. Hence, the AR, MA, and ARMA models do not apply to nonstationary time series, so there is a need to remove these effects before applying such models. Therefore, if the input time series is nonstationary, we first have to transform the series from nonstationary into stationary and then apply models such as the AR, MA, and ARMA.

To transform a nonstationary time series to stationary, we may use differencing, as discussed, once, twice, or three times, and so on until the series is at least approximately stationary. As AR and MA processes are described by the order, in a similar way, the differencing process is also described by the order of differencing, such as 1, 2, 3, etc. Therefore, to describe a model for nonstationary time series, the elements make up a triple (p,d,q) instead of two (p,q), which defines the type of model applied where the degree of the differencing is represented by the d parameter. Combining the differencing of a nonstationary time series with the ARMA model provides a powerful family of models that can be applied in a wide range of situations. The model is described as an autoregressive moving average (ARMA) model.

In this form, the letter "I" in ARIMA refers to the fact that the time series data has been initially differenced, and when the modelling is completed, the results then have to be summed or integrated to produce the final estimations and forecasts. Box and Jenkins played a significant role in the development of this extended variant of the model; therefore, ARIMA models are also referred to as Box-Jenkins models.

The ARIMA model is discussed below:

Autoregressive Integrated Moving Average (ARIMA) Model

The ARIMA model is a combination of differencing with autoregressive and moving average models. We can express the ARIMA model as follows:

$$y'_t = \delta + \phi_1 y'_{t-1} + \phi_2 y'_{t-2} + \cdots + \phi_p y'_{t-p} - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q} + \epsilon_t$$

where y'_t is the differenced series, which may have been differenced more than once, and p and q are the orders of the autoregressive and moving average parts.

For the first difference, we can write the ARIMA model as:

$$y_t - y_{t-1} = \delta + \phi_1 (y_{t-1} - y_{t-2}) + \cdots + \phi_p (y_{t-p} - y_{t-p-1}) - \theta_1 \epsilon_{t-1} - \cdots - \theta_q \epsilon_{t-q} + \epsilon_t$$

Based on different values of p, q, and d, the ARIMA model has different types, which are discussed as follows:

Various Forms of ARIMA Models

The ARIMA model has various forms for different values of the parameters p, d, and q of the model. We discuss some standard forms as follows:

Model Name	Form	Use
ARIMA(0,0,0)	White noise $y_t = \epsilon_t$	The errors are uncorrelated across time.
ARIMA(1,0,0)	First-order autoregressive model $y_t = \mu + \phi_1 y_{t-1} + \epsilon_t$	The series is stationary and autocorrelated with its previous value.
ARIMA(0,1,0)	Random walk $y_t = y_{t-1} + \epsilon_t$	The series is not stationary.
ARIMA(1,1,0)	Differenced first-order autoregressive model $y_t - y_{t-1} = \mu + \phi_1(y_{t-1} - y_{t-2}) + \epsilon_t$	The time series is not stationary and autocorrelated with its previous values.
ARIMA(0,1,1)	Simple exponential smoothing model $y_t = \theta_1 \epsilon_{t-1} + \epsilon_t$	The time series is not stationary, and the errors are correlated across time.
ARIMA(0,1,1) with constant	Simple exponential smoothing with growth $y_t = \mu + \theta_1 \epsilon_{t-1} + \epsilon_t$	The time series is not stationary, and the errors are correlated across time.
ARIMA(1,1,1)	Damped-trend linear exponential smoothing model $y_t - y_{t-1} = \mu + \phi_1(y_{t-1} - y_{t-2}) + \theta_1 \epsilon_{t-1} + \epsilon_t$	The series is not stationary, and the series has an upward trend.

After understanding various time series models, we now discuss an important topic: how to select a suitable time series model for real-life time series data.

Time Series Model Selection

I hope you understand the various time series models. Broadly, you can divide all time series models into two categories: the models used for stationary time series, such as AR, MA, and ARMA, and the models used for nonstationary time series, such as ARIMA.

When you deal with real-time series data, the first question that may arise in your mind is how you know which time series model is most suitable for a particular time series data. Don't worry about that.

Here we describe the methodology for the same in steps so that you can easily identify/select the model and its order for the given time series data. It has the following steps:

Step 1: Check for Stationarity

Since there are two types of models used for stationary and nonstationary time series, first of all, we plot the time series data and check whether the time series data is stationary or nonstationary as you have learned.

Step 2: Determine the Model Type

If the time series is stationary, we have to decide which model out of AR, MA, and ARMA is suitable for our time series data. To distinguish among them, we calculate the autocorrelation function (ACF) and the partial autocorrelation function (PACF) as discussed . After that, we plot the ACF and PACF versus the lag, that is, correlogram as discussed , and try to identify the pattern of both.

The ACF plot is most useful for identifying the AR model and PACF plot for the order of the AR model, whereas the PACF plot is most useful for identifying the MA model and ACF plot for the order of the MA model. We now try to understand how to distinguish between AR, MA, and ARMA models as follows:

Case I (AR model)

In the plot of ACF versus the lag (correlogram), if you see a gradual diminish in amount or exponential decay, this indicates that the values of the time series are serially correlated, and the series can be modelled through an AR model. For determining the order of an AR model, we use a plot of PACF versus the lag.

If the PACF output cuts off, which means the PACF is almost zero at lag p+1, then it indicates that the AR model is of order p. We can also calculate PACF by increasing the order one by one, and as soon as this lies within the range of $\pm\frac{2}{\sqrt{n}}$ (where n is the size of the time series), we should stop and take the order as the last significant PACF as the order of the AR model.

Case II (MA model)

In the plot of PACF versus the lag, if you see a gradual diminish in amount or exponential decay, this indicates that the series can be modelled through an MA model, and if the ACF output cuts off, meaning the ACF is almost zero at lag q+1, then it indicates that the MA model is of order q.

Case III (ARMA model)

If the autocorrelation function (ACF) as well as the partial autocorrelation function (PACF) plots show a gradual diminish in amount (exponential decay) or damped sinusoid pattern, this indicates that the series can be modelled through an ARMA model, but it makes the identification of the order of the ARMA (p, q) model relatively more difficult.

For that, extended ACF, generalized sample PACS, etc. are used, which are beyond the scope of this course. For more detail, you can consult “Time Series Analysis: Forecasting and Control,” 4th Edition, written by Box, Jenkins, and Reinsel.

Step 3: Differencing Nonstationary Time Series

If the time series is nonstationary, we obtain the first, second, etc. differences of the time series as discussed until it becomes stationary and ensure that trend and seasonal components are removed and find d.

Suppose after the second difference the series becomes stationary; then d is 2. Generally, one or two-stage differencing is sufficient. The differenced series will be shorter (as you have observed) than the source series. An ARMA model is then fitted to the resulting time series.

Since ARIMA models have three parameters, there are many variations to the possible models that could be fitted. We should choose the ARIMA models as simple as possible, i.e., contain as few terms as possible (small values of p and q). For more detail, you can consult "Time Series Analysis: Forecasting and Control," 4th Edition, written by Box, Jenkins, and Reinsel.

Step 4: Parameter Estimation

After identifying the model, we estimate the parameters of the model using the method of moments, maximum likelihood estimation, least squares methods, etc. The method of moments is the simplest of these.

In this method, we equate the sample autocorrelation functions to the corresponding population autocorrelation functions, which are the functions of the parameters of the model and solve these equations for the parameters of the model. However, this method is not a very efficient method of estimation of parameters.

For moving average processes, usually the maximum likelihood method is used, which gives more efficient estimates when n is large. We shall not discuss this anymore here, and if someone is interested in this, they may refer to "Time Series Analysis: Forecasting and Control," 4th Edition, written by Box, Jenkins, and Reinsel.

Step 5: Diagnostic Checking

After fitting the best model, we give a diagnostic check to the residuals to examine whether the fitted model is adequate or not. It helps us to ensure no more information is left for extraction and check the goodness of fit.

For the residual analysis, we plot the ACF and PACF of the residual and check whether there is a pattern or not. For the adequate model, there should be no structure in ACF and PACF of the residual and should not differ significantly from zero for all lags greater than one.

For the goodness of fit, we use Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC). We have not discussed all the above aspects in detail here, but if someone is interested, they should consult "Time Series Analysis: Forecasting and Control," 4th Edition, written by Box, Jenkins, and Reinsel.

After understanding the procedure of selection of a time series model, let us take an example.

Example 4:

The temperature (in °C) in a particular area on different days collected by the meteorological department is as follows:

Day	Temperature	Day	Temperature
1	27	9	28
2	29	10	30
3	31	11	30
4	27	12	26
5	28	13	30
6	30	14	31
7	32	15	27
8	29		

1. Examine which model (AR or MA) is suitable for this data.
2. Find the order and estimate the parameter of the selected model.
3. Write the model.

Solution:

First, we check whether the given time series is stationary or nonstationary. For that, we plot the time series data by taking days on the X-axis and temperature on the Y-axis. We get the time series plot as shown in Fig.1.

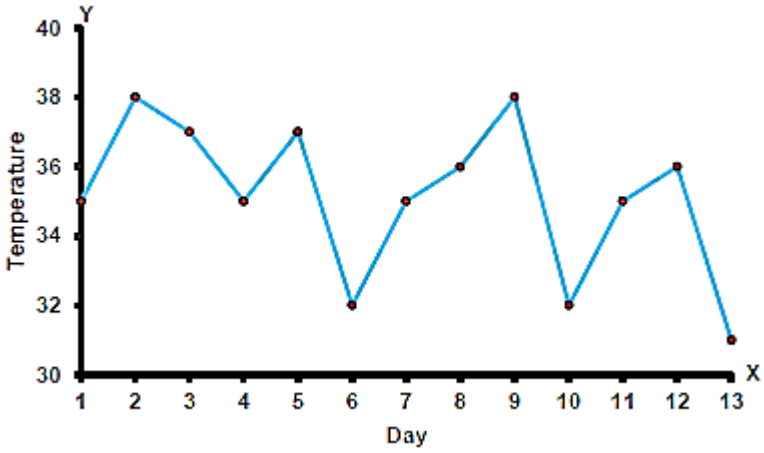

Fig.1: Time series plots of the temperature data.

Fig.1 shows that there is no consistent trend (upward or downward) over the entire period. The series appears to slowly wander up and down. Also, the variance is constant.

Almost by definition, there is no seasonality or trend. So we can say that this time series is stationary.

To examine the model and its order, we have to compute the sample autocorrelation (ACF) and partial autocorrelation (PACF) as we have discussed . Therefore, for the sake of time, we just write them here:

$r_1 = 0.835,\ r_2 = 0.676,\ r_3 = 0.469,\ r_4 = 0.280$ $\qquad$ $\hat{\phi}_{11} = r_1 = 0.835,\ \hat{\phi}_{22} = 0.088$

Since the autocorrelation function (ACF) gradually diminishes (decreases) in amount, it indicates that the series can be modelled through an AR model. The PACF output is almost zero at lag 2, indicating that the AR model is of order 1. Hence, we may conclude that the AR(1) model is suitable for this data. Therefore, the model is:

$$y_t = \delta + \phi_1 y_{t-1} + \epsilon_t$$

We now estimate the parameters (δ and ϕ_1) of the model using the method of moments. In this method, we equate the sample autocorrelation functions to the population autocorrelation functions, which are the function of the parameters of the model, and solve these as below:

$$r_1 = \rho_1$$

$$0.835 = \phi_1 \rightarrow \phi_1 = 0.835$$

For estimating the parameter δ, first, we find the mean of the given data and then we use the relationship:

$$\text{Mean} = \frac{\delta}{1 - \phi_1}$$

$$\text{Mean} = \frac{1}{15}\sum_{i=1}^{15} y_i = \frac{435}{15} = 29$$

Therefore,

$$\delta = \text{Mean} \times (1 - \phi_1) = 29 \times (1 - 0.835) = 4.785$$

Therefore, the suitable model for the temperature data is:

$$y_t = 4.785 + 0.835 y_{t-1} + \epsilon_t$$

EXERCISE 4

A researcher wants to develop an autoregressive model for the data on COVID-19 patients in a particular city. For that, he collected the data for 100 days and calculated the autocorrelation function as follows:

$$r_1 = 0.73,\ r_2 = 0.39,\ r_3 = 0.07$$

By calculating the sample PACF, estimate the order of the autoregressive model to be fitted.