Foundations of Data Science - CS F320

A REPORT
ON

Multivariate Time Series Analytics

By

Uday Mittal (2019B4A70662P)
Anubhav Srivastava (2018A8PS0030P)

Introduction

A time series is a collection of observations made sequentially in time. Time series observations can be done in two different ways. They are

1) Univariate Time Series Analytics - It consists of taking in values by a single variable at periodic instances of time. A basic example of it is noting down the temperature of a particular place at certain instances of time, by using a thermometer, by neglecting the contributions of humidity, wind speed, etc. on the changing weather.
2) Multivariate Time Series Analytics (MVTS) - It consists of taking in values by different variables at periodic instances of time. Extending the example taken above, if along with measuring temperature, the device takes the humidity, wind speed, etc. into consideration and then draws the estimate for the whether of the surroundings, then it come under MVTS.

Applications

1) Monthly Sales Data: Say, if we are running an organization use time-series forecasting to forecast store sales on data from the organization.
   The training data looks like this:

id	date	store_nbr	family	sales	onpromotion
0	01-01-2013	1	AUTOMOTIVE	0	0
1	01-01-2013	1	BABY CARE	0	0
2	01-01-2013	1	BEAUTY	0	0
3	01-01-2013	1	BEVERAGES	0	0
4	01-01-2013	1	BOOKS	0	0
5	01-01-2013	1	BREAD/BAKERY	0	0
6	01-01-2013	1	CELEBRATION	0	0
7	01-01-2013	1	CLEANING	0	0
8	01-01-2013	1	DAIRY	0	0
9	01-01-2013	1	DELI	0	0
10	01-01-2013	1	EGGS	0	0
11	01-01-2013	1	FROZEN FOODS	0	0
12	01-01-2013	1	GROCERY I	0	0
13	01-01-2013	1	GROCERY II	0	0
14	01-01-2013	1	HARDWARE	0	0
15	01-01-2013	1	HOME AND KITCHEN I	0	0

   a) The training data, comprising time series of features store_nbr, family, and onpromotion as well as the target sales.
   b) store_nbr identifies the store at which the products are sold.
   c) family identifies the type of product sold.

d) sales gives the total sales for a product family at a particular store at a given date. Fractional values are possible since products can be sold in fractional units (1.5 kg of cheese, for instance, as opposed to 1 bag of chips).
e) onpromotion gives the total number of items in a product family that were being promoted at a store at a given date.

2) Daily Climate data: We can use the data from the Indian meteorological department to train and predict the model on weather forecasting for Indian climate. The 4 parameters used here are: mean temperature, humidity, wind speed, mean pressure.

date	meantemp	humidity	wind_speed	meanpressure
01-01-2017	15.91304348	85.86956522	2.743478261	59
02-01-2017	18.5	77.22222222	2.894444444	1018.278
03-01-2017	17.11111111	81.88888889	4.016666667	1018.333
04-01-2017	18.7	70.05	4.545	1015.7
05-01-2017	18.38888889	74.94444444	3.3	1014.333
06-01-2017	19.31818182	79.31818182	8.681818182	1011.773
07-01-2017	14.70833333	95.83333333	10.04166667	1011.375
08-01-2017	15.68421053	83.52631579	1.95	1015.55
09-01-2017	14.57142857	80.80952381	6.542857143	1015.952
10-01-2017	12.11111111	71.94444444	9.361111111	1016.889
11-01-2017	11	72.11111111	9.772222222	1016.778
12-01-2017	11.78947368	74.57894737	6.626315789	1016.368
13-01-2017	13.23529412	67.05882353	6.435294118	1017.529
14-01-2017	13.2	74.28	5.276	1018.84
15-01-2017	16.43478261	72.56521739	3.630434783	1018.13
16-01-2017	14.65	78.45	10.38	1017.15
17-01-2017	11.72222222	84.44444444	8.038888889	1018.389

3) Forecasting property sales using property sales data for the 2007-2019 period for one specific region. The data contains sales prices for houses and units with 1,2,3,4,5 bedrooms. These are the cross-dependent variables. The data is

shown below:

datesold	postcode	price	propertyType	bedrooms
07-02-2007 00:00	2607	525000	house	4
27-02-2007 00:00	2906	290000	house	3
07-03-2007 00:00	2905	328000	house	3
09-03-2007 00:00	2905	380000	house	4
21-03-2007 00:00	2906	310000	house	3
04-04-2007 00:00	2905	465000	house	4
24-04-2007 00:00	2607	399000	house	3
30-04-2007 00:00	2606	1530000	house	4
24-05-2007 00:00	2902	359000	house	3
25-05-2007 00:00	2906	320000	house	3
26-06-2007 00:00	2902	385000	house	3
27-06-2007 00:00	2906	305000	house	3
27-06-2007 00:00	2612	850000	house	4
28-06-2007 00:00	2904	765000	house	4

The data can be summarized as:

a) date of sale
b) price
c) property type: unit or house
d) number of bedrooms: 1,2,3,4,5 - these are the multi-variables that also influence the pricing.
e) 4 digit postcode - place of property can influence the pricing.

4) Stock Market or in general Financial Markets: They are markets where stocks, shares, equities, etc. of companies are traded, under a regulatory authority, for example, in India we have Securities and Exchange Board of India (SEBI). The stock market is very unpredictable and events like cyclones, or floods cxan affect it. Whole world witnessed it very clearly during March and April 2020, when the world markets crashed due to the spread of the Coronavirus. In such cases the price trends of certain stocks are predicted using some parameters, such as the valuation of the company, the investment company attracts, the loans which the company has taken and others. A popular time series model named ARIMA (Auto Regressive Integrated Moving Average) model is used to predict the linear time series data.

Date	Open	High	Low	Close
Jul-05	13	14	11.25	12.46
Aug-05	12.58	14.88	12.55	13.42
Sep-05	13.48	14.87	12.27	13.3
Oct-05	13.2	14.47	12.4	12.99
Nov-05	13.35	13.88	12.88	13.41
Dec-05	13.49	14.44	13	13.71
Jan-06	13.68	17.16	13.58	15.33
Feb-06	15.5	16.97	15.4	16.12
Mar-06	16.2	20.95	16.02	20.08
Apr-06	20.56	20.8	18.02	19.49
May-06	19.8	21.8	15.8	18.03
Jun-06	18.2	18.4	13.4	15.62
Jul-06	16	17.2	13.82	16.16
Aug-06	15.9	18.6	15.7	17.84
Sep-06	18	18.88	16.8	18.46
Oct-06	18.78	24.2	18.5	22.78
Nov-06	22.8	28.47	21.4	25.32
Dec-06	25.4	29.66	24.4	26.97
Jan-07	27.4	34.16	27	30.16

5) Inventory Studies: In a big company, such as Amazon and Flipkart, maintaining inventory is a crucial part of their business. Their business cant function if there inventories are not properly maintained, with proper tools, stocks etc. The basic factor is sales, on which the stock in the inventory depend. But inventories are also maintained by keeping the future sales in consideration. Hence maintainece of an inventory becomes a problem in which many variables have to considered, considering the time as well. For example - A detailed model depicting when the sales would excel and when the sales would be diminished, which would be used to maintain adequate stock inside the inventory, and to prevent the losses due to damage, in case of perishable items.

6) Earthquake Prediction: Now-a-days a lot of research is going in the scientific community to develop a model which could predict the happening of earthquakes in real time. Till now, earthquakes remain unpredictable, but by observing previous data, recognising patterns of tectonic plates, and other changes observed during earthquakes, certain models are being made, which could predict medium to large scale

earthquakes. A major assumption in this is that the earthquakes time series is stochastic. Various clustering techniques are also used in the process.

Origin Time	Latitude	Longitude	Depth	Magnitude	Location
2021-07-31 09:43:23 IST	29.06	77.42	5	2.5	53km NNE of New Delhi, India
2021-07-30 23:04:57 IST	19.93	72.92	5	2.4	91km W of Nashik, Maharashtra, India
2021-07-30 21:31:10 IST	31.5	74.37	33	3.4	49km WSW of Amritsar, Punjab, India
2021-07-30 13:56:31 IST	28.34	76.23	5	3.1	50km SW of Jhajjar, Haryana
2021-07-30 07:19:38 IST	27.09	89.97	10	2.1	53km SE of Thimphu, Bhutan
2021-07-30 04:39:14 IST	38.52	73.27	115	5.2	286km NE of Fayzabad, Afghanistan
2021-07-30 03:33:16 IST	27.9	94.2	10	3	48km W of Basar, Arunachal Pradesh, India
2021-07-29 18:47:30 IST	26.6	92.51	28	3.1	28km WSW of Tezpur, Assam, India
2021-07-29 14:09:29 IST	22.88	95.95	10	5.5	107km N of Burma, Myanmar
2021-07-27 16:11:30 IST	37.96	72.39	160	4.3	188km ENE of Fayzabad, Afghanistan
2021-07-27 05:54:26 IST	32.29	76.65	10	2.6	31km ENE of Dharamshala, Himachal Pradesh, India
2021-07-26 16:38:34 IST	38.1	75.45	10	4.4	399km N of Kargil, Laddakh, India
2021-07-26 14:19:17 IST	36.78	73.49	19	4.2	263km E of Fayzabad, Afghanistan
2021-07-26 10:33:31 IST	24.6	94.33	50	4.9	39km SSW of Ukhrul, Manipur, India
2021-07-26 05:00:53 IST	16	78.22	10	4	156km S of Hyderabad, Telangana, India
2021-07-25 20:39:22 IST	27.29	88.5	10	4	11km WSW of Gangtok, Sikkim
2021-07-25 04:41:22 IST	37.37	71.85	100	4.2	118km ENE of Fayzabad, Afghanistan
2021-07-24 09:55:40 IST	28.97	77.2	18	2.4	39km N of New Delhi, India
2021-07-24 01:42:50 IST	36.57	70.78	250	4.5	62km SSE of Fayzabad, Afghanistan
2021-07-24 01:28:41 IST	30.74	78.68	10	3.4	23km E of Uttarkashi, Uttarakhand, India
2021-07-23 21:31:58 IST	27.56	91.13	10	3.2	73km W of Tawang, Arunachal Pradesh, India
2021-07-23 18:42:01 IST	27.35	86.86	10	3.5	134km W of Yuksom, Sikkim, India
2021-07-23 11:46:58 IST	25.94	92.56	5	2.9	73km E of Nongpoh, Meghalaya, India

7) Wind Power Forecasting: Since the world is shifting to green energy and its usages, wind becomes an important source of energy. Hence it becomes important to correctly predict the amount of power which can be generated by wind power. The data shows some parameters, which include time stamp, system power generated, speed of the wind and its direction and pressure and temperature in the air. These factors are used to predict the wind power which could be generated. The model commonly used is ARIMA (Auto Regressive Integrated Moving Average) model.

Time stamp	System power generated | (kW)	Wind speed | (m/s)	Wind direction | (deg)	Pressure | (atm)	Air temperature | ('C)
Jan 1, 12:00 am	1766.64	9.926	128	1.00048	18.263
Jan 1, 01:00 am	1433.83	9.273	135	0.99979	18.363
Jan 1, 02:00 am	1167.23	8.66	142	0.999592	18.663
Jan 1, 03:00 am	1524.59	9.461	148	0.998309	18.763
Jan 1, 04:00 am	1384.28	9.184	150	0.998507	18.963
Jan 1, 05:00 am	1293.93	8.996	149	0.998507	19.063
Jan 1, 06:00 am	1301.63	9.016	151	0.998211	19.113
Jan 1, 07:00 am	1308.13	9.036	154	0.997815	19.163
Jan 1, 08:00 am	792.081	7.612	154	1.00028	19.363
Jan 1, 09:00 am	399.537	6.129	162	1.00295	19.963
Jan 1, 10:00 am	362.988	5.961	152	1.00048	20.763
Jan 1, 11:00 am	951.359	8.117	141	1.00068	21.063
Jan 1, 12:00 pm	1549.75	9.54	141	0.996926	21.063
Jan 1, 01:00 pm	1835.22	10.094	136	1.00028	20.763
Jan 1, 02:00 pm	1208.37	8.789	137	0.998801	20.663
Jan 1, 03:00 pm	686.154	7.286	126	0.999986	20.663
Jan 1, 04:00 pm	624.105	7.088	122	0.999591	20.563
Jan 1, 05:00 pm	678.936	7.266	121	0.997715	20.363
Jan 1, 06:00 pm	1187.51	8.739	116	0.996334	20.363
Jan 1, 07:00 pm	1881.72	10.212	126	0.994952	20.363
Jan 1, 08:00 pm	1685.91	9.797	132	0.998406	20.363
Jan 1, 09:00 pm	1188.7	8.729	149	1.00048	20.263
Jan 1, 10:00 pm	939.228	8.087	155	0.99821	20.263
Jan 1, 11:00 pm	772.632	7.563	160	0.998703	20.263
Jan 2, 12:00 am	1006.16	8.265	157	0.996432	20.163
Jan 2, 01:00 am	1150.85	8.631	157	0.998999	19.863

8) Clustering of states according to covid cases in the states. We can see the data below:

Date	Daily Conf	Total Confirmed	Daily Recovered	Total Recovered	Daily Deceased	Total Deceased
30-Jan	1	1	0	0	0	0
31-Jan	0	1	0	0	0	0
01-Feb	0	1	0	0	0	0
02-Feb	1	2	0	0	0	0
03-Feb	1	3	0	0	0	0
04-Feb	0	3	0	0	0	0
05-Feb	0	3	0	0	0	0
06-Feb	0	3	0	0	0	0
07-Feb	0	3	0	0	0	0
08-Feb	0	3	0	0	0	0
09-Feb	0	3	0	0	0	0
10-Feb	0	3	0	0	0	0
11-Feb	0	3	0	0	0	0
12-Feb	0	3	0	0	0	0

We may want to see which states have the similar trend in terms of total number of confirmed cases, total deaths and recovered. This can give us an idea of effect of climate or even government policies among different states.

9) Real time conversion of sign language to text: Sign language consists of series of images in time, where at each timestamp the number of features is equal to the number of pixels in that image. This will require classification of each image to a word depending on the previous images and current images. Hence, it is also multivariate time series data. The training data will be a set of videos. Where each video is a time series data, where each timestamp has feature values of all of its pixels. GnoSys app developed by Google uses neural networks and computer vision to recognise the video of sign language speaker and then smart algorithms translate it into speech.

10) Cryptocurrencies forecasting: More than 40 billion USD worth cryptocurrencies flow in the open source market. Cryptocurrencies are known by their fast fluctuations, which they go through. In such circumstances, it becomes important to predict the values and price of crypto currencies in the future. In such a process, an important model named ARCH (Auto Regressive Conditional Heteroskedasticity) model is used. The data below includes the history of bitcoin, along with the opening, closing prices and volume of transactions.

UNIX time stamp	Opening Price	Highest Price	Lowest Price	Closing Price	Volume of transactions
1.60943E+12	28782.01	28821.85	28763.94	28811.85	95.835795
1.60943E+12	28812.64	28822.59	28714.29	28726.62	58.516227
1.60943E+12	28728.28	28744.76	28684.69	28693.37	75.038373
1.60943E+12	28693.37	28715.15	28682.09	28690.29	37.128193
1.60943E+12	28690.29	28734.7	28680	28715.11	38.411112
1.60943E+12	28715.11	28741.26	28713.06	28735.88	25.011182
1.60943E+12	28735.89	28747.43	28720.87	28732.01	28.182584
1.60943E+12	28729.43	28729.43	28700	28700.01	24.41429
1.60943E+12	28700	28720	28682.08	28720	18.954939
1.60943E+12	28719.99	28729.2	28714.62	28719.85	18.412759
1.60943E+12	28719.6	28729.19	28701.67	28708.71	41.247486
1.60943E+12	28710.15	28719.78	28696.6	28706.24	17.996544
1.60943E+12	28706.37	28715.85	28678.19	28678.19	38.172147
1.60943E+12	28678.09	28683.25	28615.82	28653.79	152.116471
1.60943E+12	28652.42	28655.9	28623.41	28648.53	47.906837
1.60943E+12	28648.52	28653.01	28600	28603.55	48.903846
1.60943E+12	28603.55	28603.55	28555	28595.65	112.687004
1.60943E+12	28595.65	28622.49	28586.77	28606.44	48.338657
1.60943E+12	28606.44	28633.82	28605	28630.53	34.660922
1.60943E+12	28630.52	28670.73	28621.4	28669.24	38.168901
1.60943E+12	28670.46	28707.97	28667.85	28702.03	48.591758
1.60943E+12	28702.04	28703.75	28670	28671.3	40.400552
1.60943E+12	28671.3	28682.99	28640.41	28673.95	44.274367

Problem Definition

To predict the value of air pollution parameters over time based on basic weather information like temperature and humidity and measurements from 5 different sensors. The air pollution parameters to be predicted are: target_carbon_monoxide, target_benzene, target_nitrogen_oxides.

Dataset Used

The training dataset used is a multivariate time series with features: deg_C, relative_humidity, absolute_humidity, 5 sensor values, target_carbon_monoxide, target_benzene, target_nitrogen_oxides. The data duration is from 10/3/2010 to 1/1/2011. The data is collected hourly. The training dataset is shown below:

date_time	deg_C	relative_humidity	absolute_humidity	sensor_1	sensor_2	sensor_3	sensor_4	sensor_5	target_carbon_monc	target_benzene	target_nitrogen_oxid
2010-03-10 18:00:00	13.1	46	0.7578	1387.2	1087.8	1056	1742.8	1293.4	2.5	12	167.7
2010-03-10 19:00:00	13.2	45.3	0.7255	1279.1	888.2	1197.5	1449.9	1010.9	2.1	9.9	98.9
2010-03-10 20:00:00	12.6	56.2	0.7502	1331.9	929.6	1060.2	1586.1	1117	2.2	9.2	127.1
2010-03-10 21:00:00	11	62.4	0.7867	1321	929	1102.9	1536.5	1263.2	2.2	9.7	177.2
2010-03-10 22:00:00	11.9	59	0.7888	1272	852.7	1180.9	1415.5	1132.2	1.5	6.4	121.8
2010-03-10 23:00:00	11.2	56.8	0.7848	1220.9	697.5	1417.2	1462.6	949	1.2	4.4	88.1
2010-03-11 0:00:00	10.7	55.7	0.7603	1244.2	669.3	1491.2	1413	769.6	1.2	3.7	59.5
2010-03-11 1:00:00	10.3	57	0.7702	1181.4	631.7	1511.1	1359.7	715.4	1	3.4	63.9
2010-03-11 2:00:00	10.1	62.7	0.7648	1159.6	602.9	1610.6	1212.2	657.2	0.9	2.2	46.4

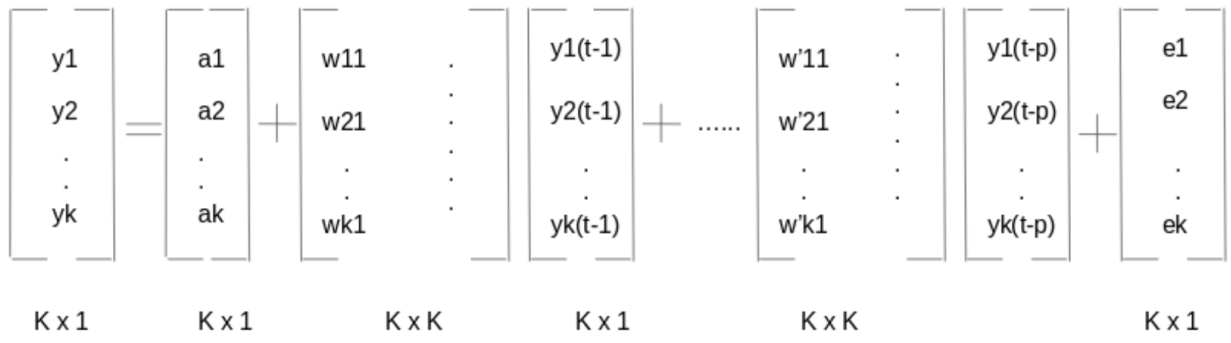
The test data is also a time series data with time from 1/1/2011 to 4/4/2011. The test data has all the columns as in the training set except the columns to be predicted, i.e. target_carbon_monoxide, target_benzene and target_nitrogen_oxides. The test data is shown below:

date_time	deg_C	relative_humidity	absolute_humidity	sensor_1	sensor_2	sensor_3	sensor_4	sensor_5
2011-01-01 0:00:00	8	41.3	0.4375	1108.8	745.7	797.1	880	1273.1
2011-01-01 1:00:00	5.1	51.7	0.4564	1249.5	864.9	687.9	972.8	1714
2011-01-01 2:00:00	5.8	51.5	0.4689	1102.6	878	693.7	941.9	1300.8
2011-01-01 3:00:00	5	52.3	0.4693	1139.7	916.2	725.6	1011	1283
2011-01-01 4:00:00	4.5	57.5	0.465	1022.4	838.5	871.5	967	1142.3
2011-01-01 5:00:00	4.5	53.7	0.4759	1004	745.5	914.2	989.1	973.8
2011-01-01 6:00:00	3.3	54.8	0.4636	940.9	738.2	816	896.8	1049.4
2011-01-01 7:00:00	3.2	60.7	0.4667	954.5	713.9	834.7	935.6	956.3
2011-01-01 8:00:00	2.5	65.7	0.4721	969.9	679.1	943.8	959.3	892

Implementation of Regression algorithm - Vector Autoregression (VAR)

Vector Autoregression (VAR)

Vector Autoregression (VAR) is one of the most commonly used methods for MVTS forecasting. It is a mathematical modelling, which focuses on making a linear function for each and every variable used in forecasting, by providing specific weights to every parameter. In a VAR model, each variable is a linear function of the past values of itself and the past values of all the other variables. Vector Autoregression (VAR) model is an extension of univariate autoregression model to multivariate time series data. A VAR model is a multi-equation system where all the variables are treated as endogenous (dependent). There is one equation for each variable as a dependent variable.
A VAR(p) model for k dependent variables looks like:



$$Iy(t) = a + w_1 * y(t-1) + w_2 * y(t-2) + ... + w_p * y(t-p) + \varepsilon$$

Where,
yi(t) - dependent (endogenous) variables
ai - constant terms
ei - error terms
yi(t-j) - jth lag term of yi
wij - weights
for 1<i<k and 1<j<p

Steps to applying VAR:

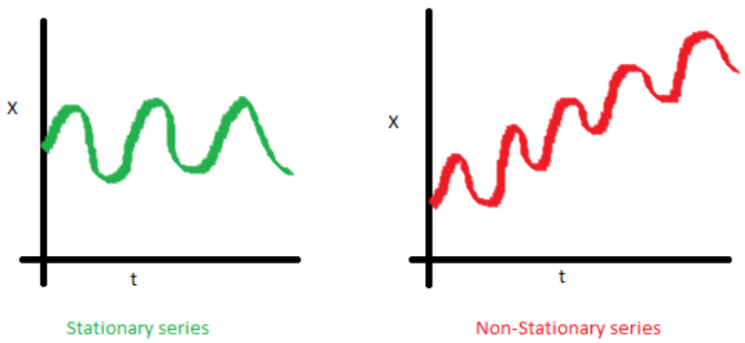
1. Check if time series is stationary or not, if not convert it to stationary time series.
2. Apply the VAR model to the time series data by selecting appropriate p (lag order).
3. Undo the transformations (used to stationarize the TS), if any.
4. Test the VAR model on a test set.

Below, we shall look at what is a stationary time series and how to stationarize the time series.

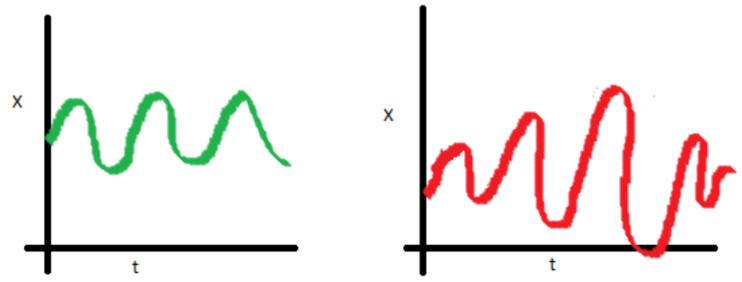
Stationary Series and Tests of Stationarity for univariate and multivariate TS

A time series model works best on stationary time series.
A time series is said to be stationary if:

a) The mean is constant with time.

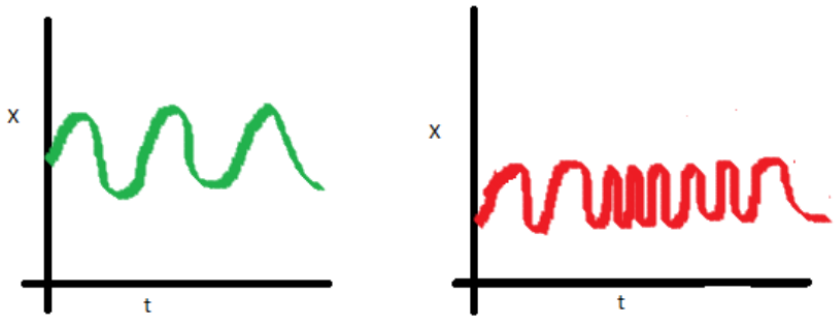


b) The variance is constant with time.



Stationary series Non-Stationary series

c) The covariance of the i th term and the (i + m) th term should not be a function of time.



Stationary series Non-Stationary series

We have 2 tests for checking stationarity of a time series:

a) Visual test: Simply plotting the time series data can give us a visual clue that the properties of time series are changing or not. However, sometime it may not always give accurate results.

b) Statistical tests: For a univariate time series:

$$yt = a*yt\text{-}1 + \varepsilon\ t$$

where yt is the value at the time instant t and ε t is the error term. In order to calculate yt we need the value of yt-1, which is :

$$y(t\text{-}1) = a*y(t\text{-}2) + \varepsilon(\ t\text{-}1)$$

Doing that for all observations, the value of yt will come out to be:

$$y(t) = (a\text{^}n)*y(t\text{-}n) + \Sigma\varepsilon(t\text{-}i)*ai$$

If the value of a is 1 (unit) in the above equation, then the predictions will be equal to the yt-n and sum of all errors from t-n to t, which means that the variance will increase with time. This is known as unit root in a time series.

Two tests which utilizes the unit root concept are:

1. ADF (Augmented Dickey Fuller) Test
2. KPSS (Kwiatkowski-Phillips-Schmidt-Shin) Test

Similarly, for multivariate time series:

$$Iy(t) = a + w_1 * y(t-1) + w_2 * y(t-2) + ... + w_p * y(t-p) + \varepsilon$$

Can be represented in terms of lag operators:

$$Iy(t) = a + w_1 * L^1 y(t) + w_2 L^2 * y(t) + ... + w_p * L^p y(t) + \varepsilon$$

Which can be written as:

$$(I - w_1 * L^1 - w_2 L^2 * -... - w_p * L^p) y(t) = a + e$$

The coefficient of y(t) is called lag polynomial represented as Φ(L):

$$\Phi(L) y(t) = a + \epsilon$$

$$y(t) = \Phi(L)^{-1}(a + e)$$

For a series to be stationary, the eigenvalues of |Φ(L)^-1| should be less than 1 in modulus. Test for multivariate time series is called Johnsen's Test.

Stationarize a Time Series

If the time series is not stationary then we can apply:

a) Differencing: In this method, we compute the difference of consecutive terms in the series. Differencing is typically performed to get rid of the varying mean. Mathematically, differencing can be written as:

yt' = yt – y(t-1)

b) Seasonal differencing: In seasonal differencing, instead of calculating the difference between consecutive values, we calculate the difference between an observation and a previous observation from the same season. For example, an observation taken on a Monday will be subtracted from an observation taken on the previous Monday. Mathematically it can be written as:

yt' = yt – y(t-n)

Where n can be 7 for the above example.

c) Transformation:Transformations are used to stabilize the non-constant variance of a series. For example, if we can clearly see that there is a significant positive trend. So we can apply transformations which penalize higher values more than smaller values. These can be taking a log, square root, cube root, etc