Time Series

SUPERVISOR :

MR. AGHAEI

AUTHORS :

HASAN ROKNABADI
HOSSEIN YAHYAEE

Time series

1.Time series :

Set of ordered data values observed as successive points in time.

(It visualize data across time)

2.Cross-Sectiond :

Set of data values observed at fixed point in time, or where time has no significance.

(time doesn't matter)

➔ In cross-sectioned were looking for how two features are related to each other across the time , or we neither look for how does a feature change across time. (mean or center of the feature might change across time)

3.Transactione data :

Which has both:

1.Index across time (like time series)

2.index across observation (like cross-sectional)

Example : like customer transactions, each customer is independent collection of dependent time points

so how to handle or analyze data like thus :

1. Transactional data → cross-sectional
   Example : in time series first video : 3:50

2. Transactional data → time-series
   Example : in time series first video : 4:00

Time series Decomposition

Based on statistical forecasting : Time series = Signal + Noise

That signal shows Forecasts extrapolate Signal partition of model and Noise shows

Confidence intervals account for uncertainty

Signal :

1.Trend / Cycle the way of / calculating it taking moving averages.

2.Seasonal way of / calculating it taking overall or general average.

Noise :

Erore – Reminder – Irregular

Notice : the hole time series can now be thought of like the equation below.

Additive : $Y_t = T_t + S_t + E_t$

Emplitude doesn't change.

Multiplicative : $Y_t = T_t \times S_t \times E_t$

Emplitude does change.

-Classical way of decomposition :

.taking moving averages for detecting trends.

.taking overall or general average for detecting seasonal.

-Modern techniques:

.Locally estimated scatterplot smoothing regression. (LOESS-Regression)

ETS and ARIMA

Have a disadvantage which is lacked of scalability for larger datasets with move complex temporal patterns .

Time series data have two important properties :

1. Data is measured sequentially.
2. Each time unit has at most one data measurement.

Time series data can be :

3. Regular : the measurements are equally spaced in time
4. Irregular : the measurements may not happen at fix time intervals.

In time series forecasting, usually we have two goal :

- Average method
- Moving Average method
- Naïve method

- Average method :

This type of methods assume that a future event is best described by the average Of all past events.

- Moving Average method :

This methods are based on simple average method but instead of using all past event To take the average it will predicts a new event based on predefined number of Recent events.

- Naïve method :

It assumes that next event will be equal to the most recent one.

NOTICE : these methods are not made for many fluctuation that are usually presented In time series.

Time Series Patterns

Trend :

it describes the general behaviors of a time series.

- Time series could manifests a positive long-term slope over time(which means having upward trend)
- Time series could manifests a negative long term slope over time
- Or the overall trend may also change direction like : up to down trend or down to up trend.

Seasonality :

it's any kind of change in a time series which always has fixed

Frequency (this means some events will happen in the same time

Interval again and again and again)

Like : selling rate of clothing store that sells heavy counts in every year, it will

Decrease during the summer and it will increase during winter.

Cycle :

cyclical pattern in a time series are rises and falls with non-fixed magnitudes this

Can last for long time but this change is not related to seasonal factors.

. This kind of patterns are not repetitive.

. usually they result from external factors which make them much harder to

Predict.

. not all time series are predictable (some of them presents no predictable patterns

In long term)

- To forecast this kind of data we usually use the random walk model. (typically used with financial and economic data)

Exponential smoothing Models :

Time series data relies on the assumption that the observation at a certain point

Depend on previous observations in time.

- We can apply a weighting schema that decreases exponentially the further back in time we go :

$$\hat{y}_{t+1} = \theta y_t + \theta(1-\theta) y_{t-1} + \theta(1-\theta)^2 y_{t-2} + \theta(1-\theta)^3 y_{t-3} + \ldots + \theta(1-\theta)^n y_{t-n}$$
$$and \rightarrow 0 \le \theta \le 1$$

So in case of finding best θ for the model with predefined n we must minimize one less with a loss function like :

$$SSE = \sum_{t=1}^{t} (y_t - \hat{y}_t)^2$$

Holt Exponential smoothing Model :

Level : $L_t = \theta y_t + (1-\theta) L_{t-1}$

Trend : $T_t = \gamma (L_t - L_{t-1}) + (1-\gamma) T_{t-1}$

Season : $S_t = \partial (y_t - L_t) + (1-\partial) S_{t-p}$

Winter Exponential smoothing model :

$$\hat{y}_{t+k} = L_t + kT_t + S_{t-p+k}$$

constancy of distribution

can't have independence but still want some consistency.

Notice: distribution depends only on difference in time not location in time.

we need consistency to be able to build model

Strong stationarity

as a window of a fixed size move across time that distribution of the data inside of this window be the same.

Weak stationarity

Mean, variance and autocorrelation will be the same for any window of some size.

consistency of mean and variance :

for mean constancy:

to correct for trend and seasonality → take differences

for variance consistency:

model the lack of consistency in various → ARCH/GARCH

.read about a stationary model …

Stationary model:

1. Autoregressive(AR) models :

   Tries to forecast a series based on only the past values in the series called lags.

AR(1) model that depends only on one leg in the past :

$$y_t = \omega + \phi y_{t-1} + e_t$$
$$y_t = T\arg et$$
$$\omega = Inrercept$$
$$\phi = Coedicient$$
$$y_{t-1} = lagged - t\arg et$$
$$e_t = Error$$

If we continue to go back in time well see this recursion :

$$y_t = \omega + \phi y_{t-1} + e_t$$
$$y_{t-1} = \omega + \phi y_{t-2} + e_{t-1}$$

This whole idea of recursion called long memory model

Let's get final version of AR(1) :

$$y_t = \omega(\sum_{i=0}^{n-1} \phi^i) + \phi^n y_{t-n} + \sum_{i=0}^{n-1} \phi^i e_{t-i}$$

. So the effect of shocks that happen long ago have little effect on the present IF:

$$|\phi| \leq 1$$

. We can do the same process for more lags too. AR(p)

2.Moving average (MA) models:

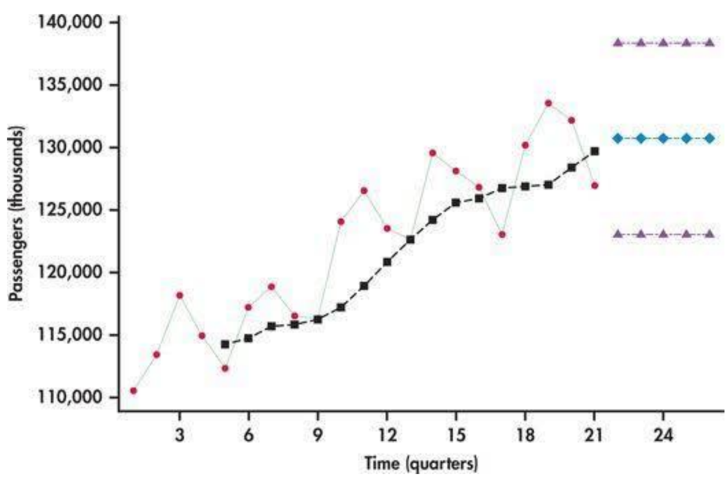
Tries to forecast a series based on only the past errors in the series called error lags.

MA(1) that depends only on one lag of error in the past :

$$y_t = \omega + \phi e_{t-1} + e_t$$

e_{t-1} is lagged error.

Hold on!!! but weren't errors unseen?



.Some unknown predicted shock that shifted you off of where you expected today to be is what actually affecting the current time point.

. in case of having a predicted value for our first observation well take the average of all serie then consider this value as a prediction.

. Starting point for this kind of models won't matter because this kind of models known as short memory model which means these errors won't last long into the future.

Let's take a look at greater detail:

$$y_t = \omega + \phi e_{t-2} + e_{t-1}$$

$$y_t = \omega + \phi e_{t-1} + e_t$$

No more sign of e_{t-2}

$$y_t = \omega + \phi e_t + e_{t+1}$$

No more sign of e_{t-1}

- Best moving average model is going back to this idea of stationarity
  - the dependency of previous observation declines overtime.

- In moving average model as we moving through time the dependency of observations on previous errors are completely gone.

Lets have closer look :

AR(1) = MA(infinite)

$$y_t = \omega + \phi y_{t-1} + e_t$$

$$y_t = (1+\phi)\omega + \phi^2 y_{t-2} + \phi e_{t-1} + e_t$$

.

.

$$y_t = (\sum_{i=0}^{n-1} \phi^i)\omega + \phi^n y_1 + \phi^{n-1} e_2 + \phi^{n-2} e_3 + \ldots + e_t$$

$$\phi^{n-1} e_2 + \phi^{n-2} e_3 + \ldots + e_t \rightarrow Large - Moving - averaget - term$$

- As conclusion we understood that for making a long term memory we will use bunch of short term memory.
- Instead of having large AR(p) model or large MA(q) model we can combine them together.

3.Autoregressive moving average models (ARMA)

$$y_t = \omega + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \phi_q y_{t-1} + \ldots + \phi_q y_{t-q} + e_t$$

How to select how many P lags and Q lags should we use?

1) Plotting patterns in correlation (ACF, PACF)
2) Automatic selection techniques

Automatic selection techniques:

Common ways are :

- MINIC : (minimum information criterion)

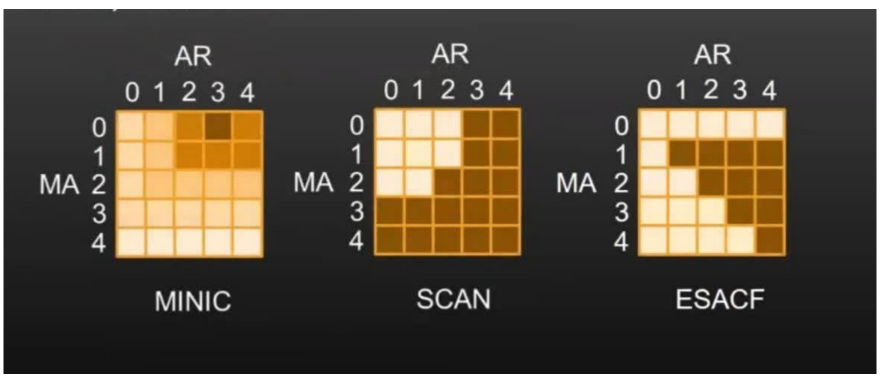
  - It's builds many combinations of models across a grid search of AR&MA terms and chooses the model with lowest basin information criterion.

- SCAN : (squirrel canonical correlation)
  1) looking at correlation matrix of our data.
  2) looking at eigen values of this correlation matrix.

  - The goal is finding the combination of AR and MA terms that produce canonical correlation of approximately Zero.

- ESACF : (extended sample auto correlation function)

  - by filtering all AR terms still we get or have only MMA terms.



4.Auto regressive integrated moving average (ARIMA)

- It's like ARMA model with extra term which is integrate and this extra term will take care of data being stationary.
  ARIMA(p(Of AR terms), d(off first differences), q(of MA terms))

5. seasonal ARIMA:

ARIMA Models are a stationary models but seasonal data is not stationary by design.

hold on! Differencing will make our data stationary but it doesn't remove seasonal correlation. Those be need seasonal ARIMA model.

ARIMA(p(AR terms),d(of first differences),q(MA terms))

(P(Seasonal 8R terms),D(seasonal differences),Q(seasonal MA terms))

S(length of season)

Example: ARIMA(1, 0, 1)(2, 1, 0) 1 2

Length 1 2 : $y_t - y_{t-12} = \omega_t$

First 1 : $\phi_1 \omega_{t-1}$

Second 1 : $\phi_1 \omega_{t-1}$

2 : $\phi_2 \omega_{t-12}$

Third 1 : $y_t - y_{t-12} = \omega_t$

Total :

$$\omega_t = \omega + \phi_1 \omega_{t-1} + \phi_1 \omega_{t-1} + \phi_2 \omega_{t-12} + \phi_3 \omega_{t-24} + \phi_1 \omega_{t-1} + e_t$$

Neural Network time series models:

Statistical focused :

- exponential smoothing(simple, holt, winters, holt, …)
- ARIMA
- Seasonal ARIMA
- Dynamic linear models
- ARCH, GARCH
- Vector AR

Mathematics/ML focused :

- autoregressive neural networks
- Eourier transforms
- TBATS
- Prophet
- LSTN
- bootstrapping and bagging
- time series clustering

.Autoregressive neural Nets :

These neural Nets take predefined number of lags as it's input and then do its networks architecture on these lags.

.Prophet model : (forecasting universe univariate or multivariate)

It will forecast a time series by breaking it down to pieces signal and noise

Time Series = Signal + Noise

. The goal is to tease out the signal and forecast that signal to the future.

- But every technique has different approach to getting that signal

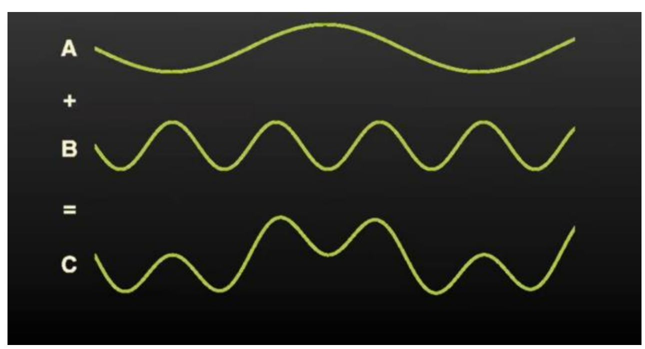
----> Signal : 1) Trend 2) Season 3) Holiday

Trend :

. Profit models use trend lines or time itself as a regressor in the model
For example using piecewise regression to detect the trend it's time Series data.

Seasonal :

. Profit models use fourier transforms



.Greater Detail :

$$X_{1,t}=\sin(\frac{2\pi t}{S})-X_{3,t}=\sin(2\times\frac{2\pi t}{S})-X_{5,t}=\sin(3\times\frac{2\pi t}{S})-\ldots$$

$$X_{2,t}=\cos(\frac{2\pi t}{S})-X_{4,t}=\cos(2\times\frac{2\pi t}{S})-X_{6,t}=\cos(3\times\frac{2\pi t}{S})-\ldots$$

$$Y_t=\beta_0+\beta_1X_{1,t}+\beta_2X_{2,t}+\beta_3X_{3,t}+\beta_4X_{4,t}+\ldots+e_t$$

S : length of seasonal variance

T : time

Fourier transforms in prophet model are two :

- Yearly season set to 10 terms by default:

$$X_y=\cos(\frac{2\pi t}{365.25})+\sin(\frac{4\pi t}{365.25})+\ldots$$

- Weakly season set to 3-terms by default:

$$X_w=\cos(\frac{2\pi t}{7})+\sin(\frac{4\pi t}{7})+\cos(\frac{6\pi t}{7})$$

Holiday:

- It's point intervention variable.

  all are points intervention variable is a binary indicator for a certain day

  - For all conclusion prophet model is basically a curve fitting.

- Multivariate regression is trying to predict multiple targets were varied at the same

  Time. which we already know it means predicting target variable based

  On its relation with other variables.

. Vector Autoregressive models:

$$VAR(1,2) \Rightarrow \begin{bmatrix} Y_{t,1} \\ Y_{t,2} \end{bmatrix} = \begin{bmatrix} a_{0,1} \\ a_{0,2} \end{bmatrix} + \begin{bmatrix} a_{11,1} & a_{12,1} \\ a_{21,1} & a_{22,1} \end{bmatrix} \begin{bmatrix} Y_{t-1,1} \\ Y_{t-1,2} \end{bmatrix} + \begin{bmatrix} e_{t,1} \\ e_{t,2} \end{bmatrix}$$

Which is AR(1) with the two target variables.

- this VAR models will predict multiple target variables at once based on its relation with other variables(each target variable has a relation with other variables)

.As we saw we can have VAR we can even have VARMA models to :

$$Y_t = \begin{bmatrix} Y_{1,t} \\ Y_{2,t} \\ \vdots \\ Y_{n,t} \end{bmatrix} = \begin{bmatrix} a_{0,1} \\ a_{0,2} \\ \vdots \\ a_{0,n} \end{bmatrix} + \begin{bmatrix} a_{11,1} & \cdots & a_{1n,1} \\ \vdots & \ddots & \vdots \\ a_{n1,1} & \cdots & a_{nn,1} \end{bmatrix} \begin{bmatrix} Y_{t-1,1} \\ Y_{t-1,2} \\ \vdots \\ Y_{t-1,n} \end{bmatrix} + \begin{bmatrix} \beta_{11,1} & \cdots & \beta_{1n,1} \\ \vdots & \ddots & \vdots \\ \beta_{n1,1} & \cdots & \beta_{nn,1} \end{bmatrix} \begin{bmatrix} e_{t-1,1} \\ e_{t-1,2} \\ \vdots \\ e_{t-1,n} \end{bmatrix} + \begin{bmatrix} e_{t,1} \\ e_{t,2} \\ \vdots \\ e_{t,n} \end{bmatrix}$$

$$Y_t = \alpha_0 + A_1 Y_{t-1} + e_t + \beta_1 e_{t-1}$$

- In general we can have : VARMA(p,q)

$$VAR(p,q) \Rightarrow Y_t = \alpha_0 + \sum_{i=1}^{p} A_i Y_{t-i} + e_t + \sum_{j=1}^{p} \beta_j e_{t-j}$$

VAR vs VARMA :

VAR :

- Move parameters to estimate to be very predictive.
- easier to estimate computationally.
- simple MLE and LS.

VARMA:

- Fewer parameters to estimate to be very predictive.
- Harder to estimate computationally.
- Iterative MLE with partial derivatives.

Bayesian Autoregressive models :

We assume some notation of a distribution around the parameters were trying to Estimate

Bayesian View

. *But the way of estimating the parameters is :*

Using Markov chain Monte Carlo :

- (MCMC)Monto Carlo : is the practice of estimating random samples from the distribution
  Example : instead of calculating the mean of distribution directly from the distribution equation Monto Carlo approach would like to use mean of samples it's easier.

- (MCMC)Markov chain : is the idea that vandon samples are generated by a special sequential process.(chain property means that each random sample depends on one before itself but not before the previous one)

.So BAR are just like normal AR model with extra assumption.(assuming some kind of distribution around our data parameters and model variants)

. When we are forecasting values from a bayesian AR model V actually getting the distribution of a forecasts for each time point.

- We can extend the best VBAR models to

Unconditional variance:

it's typical variance we know :

$$E(X - E(X))^2 = VAR(X)$$

Conditional variance :

conditional variance is the measure of our uncertainty about a variable given a set of information.

.Heteroscedasticity - variance depends on external factors.

$$VAR(X \mid I) = \partial^2{}_{cond} = E(X - E(X \mid I))^2$$

. We are interested to model volatility to manage our risk such that leads to better decision making

Autoregressive conditional heteroscedasticity:

. It's an auto regressive approach to model volatility.

$$V_t^2 \rightarrow Actual - today$$

$$\partial^2{}_{t+1} \rightarrow volatility - tomarrow$$

$$\partial^2{}_{t+1} = a_0 + a_1 V_t^2$$

$$\partial^2{}_t = \frac{1}{t}\sum_{i=1}^{t}(r_t - \overline{r})^2 = \frac{1}{t}\sum_{i=1}^{t}(r_t - 0)^2 = \sum_{i=1}^{t}(r_t)^2 = (r_t)^2$$

. We can have many lags in our ARCH model.

$$\partial^2{}_{t+m} = a_0 + a_1 r_t^2 + a_2 r_{t-1}{}^2 + \ldots + a_q r_{t-q+1}{}^2 = ARCH(q)$$

. As we had for AR models we need large number of Q to have effective accuracy so let's Extend it:

Generalized ARCH

$\partial_{t+1} \to Volatility - tomorrow$

$r_t \to Actual - today$

$\hat{\partial}_t \to Forecasted - Volatility - of - today$

$$\partial^2{}_{t+1} = a_0 + a_1 r_t^{\,2} + \beta_1 \hat{\partial}_t{}^2$$

Model evaluation:

. This curve is scale dependent and symmetric.

- Mean absolute error(MAE)
- $MAE = \frac{1}{n}\sum_{t=1}^{n} | Y_t - \hat{Y}_t |$

- mean squared error(MSE)
- $MSE = \frac{1}{n}\sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2$

- Mean absolute percentage error (MAPE)
- $MAPE = \frac{1}{n}\sum_{t=1}^{n} | \frac{Y_t - \hat{Y}_t}{Y_t} |$

- Symmetric MAPE(sMAPE)
- $sMAE = \frac{1}{n}\sum_{t=1}^{n} | \frac{Y_t - \hat{Y}_t}{|Y_t| + |\hat{Y}_t|} |$

End.