Report

Mathematical context

As discussed previously, we aimed to forecast the series using ARIMA models via a Box-Jenkins approach:

$$y_t = \mu + \alpha_1 y_{t-1} + \ldots + \alpha_p y_{t-p} + \theta_1 \epsilon_1 + \ldots + \theta_q \epsilon_{t-q}$$

Where μ is the drift, α is $AR(p)$ and θ is the $MA(q)$ constants to be estimated. The $AR(p)$ and the $MA(q)$ models must follow stationarity.

The partial autocorrelation function (PACF, ϕ) is the autocorrelation function (ACF, ρ), minus the common correlations between the lags. For example: $PACF(lag = 2) = ACF(lag = 2) - Corr(X_{t-1})$. This allows us to evaluate the autoregressive order by observing the significance of the lags. Using Yule-Walker equations, the matrix for the constnats can simultaneously be calculated by:

$$\boldsymbol{\gamma} = \boldsymbol{\Gamma}\boldsymbol{\alpha}$$

where $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_h]^T$ and $\Gamma_{ij} = \gamma_{i-j}$. Numerically, $\boldsymbol{\gamma}$ is evaluated as:

$$\widehat{\gamma}_h = \frac{1}{n} \sum_{j=1}^{n-h} (x_{j+h} - \overline{x})(x_j - \overline{x})$$

and the coeeficients can be found by inversing $\boldsymbol{\Gamma}$.

After fitting the $AR(p)$, the $\boldsymbol{\theta}$ coefficients are also evaluated using the autocovariance function (ACVF). Given that the model is well assumed and the residuals, e, are white noise, the ACVF can be equated to:

$$\gamma_h = \sigma^2 \sum_{k=|h|}^{q} (-\theta_{k-|h|})(-\theta_k)$$

And the white noise variance is given by:

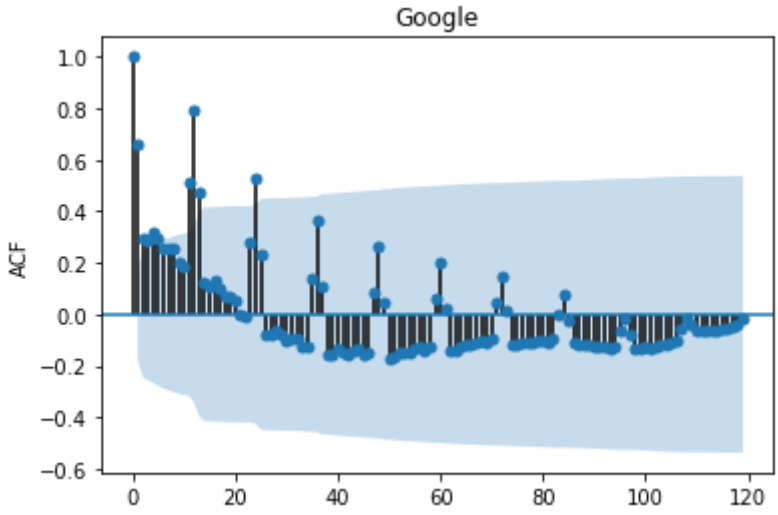
$$\sigma^2 = \frac{1}{T-p} \sum_{i=p+1}^{T} e_i^2$$

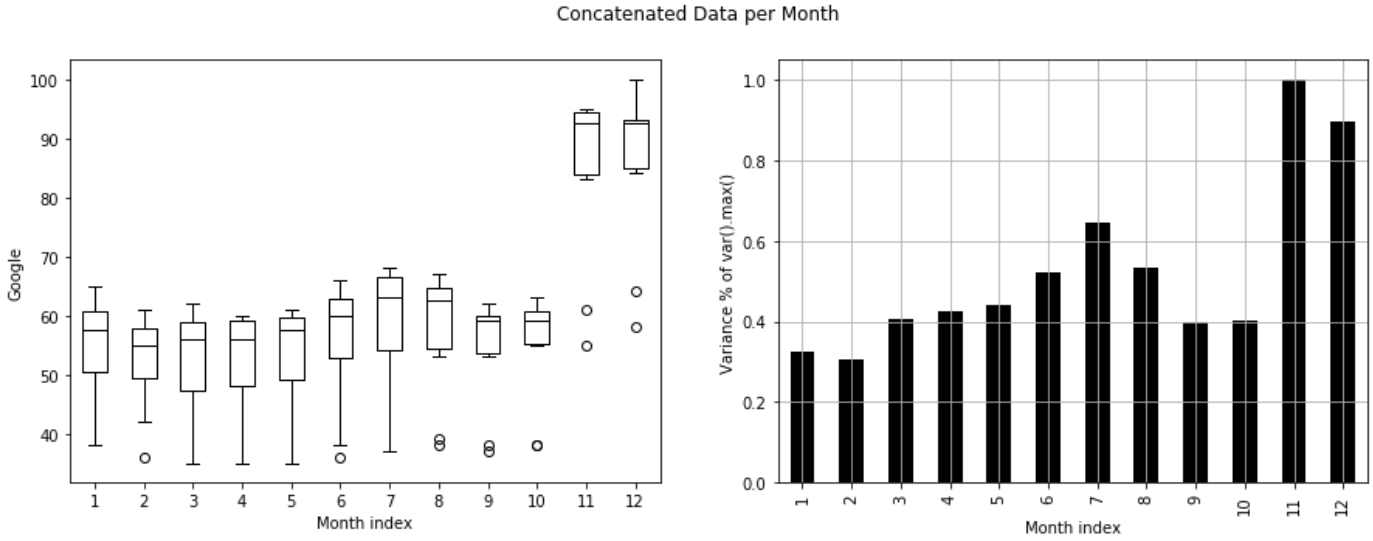
The system of p and q set of equations can be solved with zero degrees of freedom for all model parameters.

Seasonality and differencing

Google

We analysed and tested the data for stationarity and normality. The autocorrelation function (ACF) of "Google Trends" data confirmed seasonality. Each spike is at a lag multiple of 12, indicating seasonality for monthly data. The data was also parsed and monthly data was concatenated over the 9 years to show clear increase in mean and volatility of searches in the months of November and December.

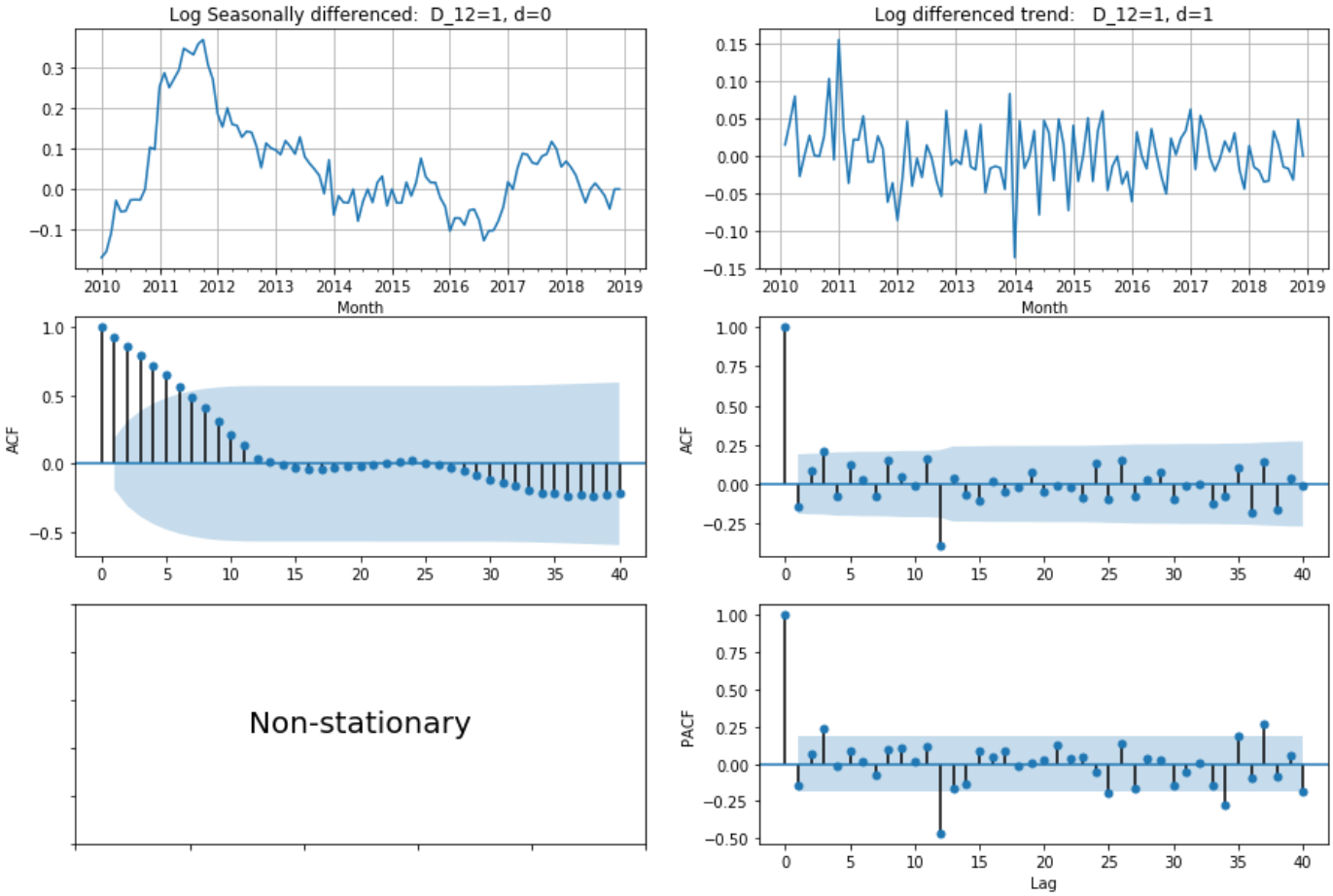




Thus, google data has been log transformed to normalise the variance, seasonally differenced (D=12) and tested for unit root using Augmented Dickey Fuller (ADF) test. The data was not yet stationary, so it was differenced once more.

Here an issue arose: ADF resulted in type 1 errors, rejecting the null hypothesis and declaring that there is no unit root present. However, it was later observed that the SARIMAX (seasonal ARIMA) model in statsmodels package was unable to carry out the forecast as it observed non-stationarity of the data. **The issue will be looked into but an assumption is that the method used for evaluating the ARIMA coefficients takes into consideration too many lags. Accordign to Enders 2014 [XXX] given a sample size of T, it is suggested to consider T/4 lags of the PACF.**

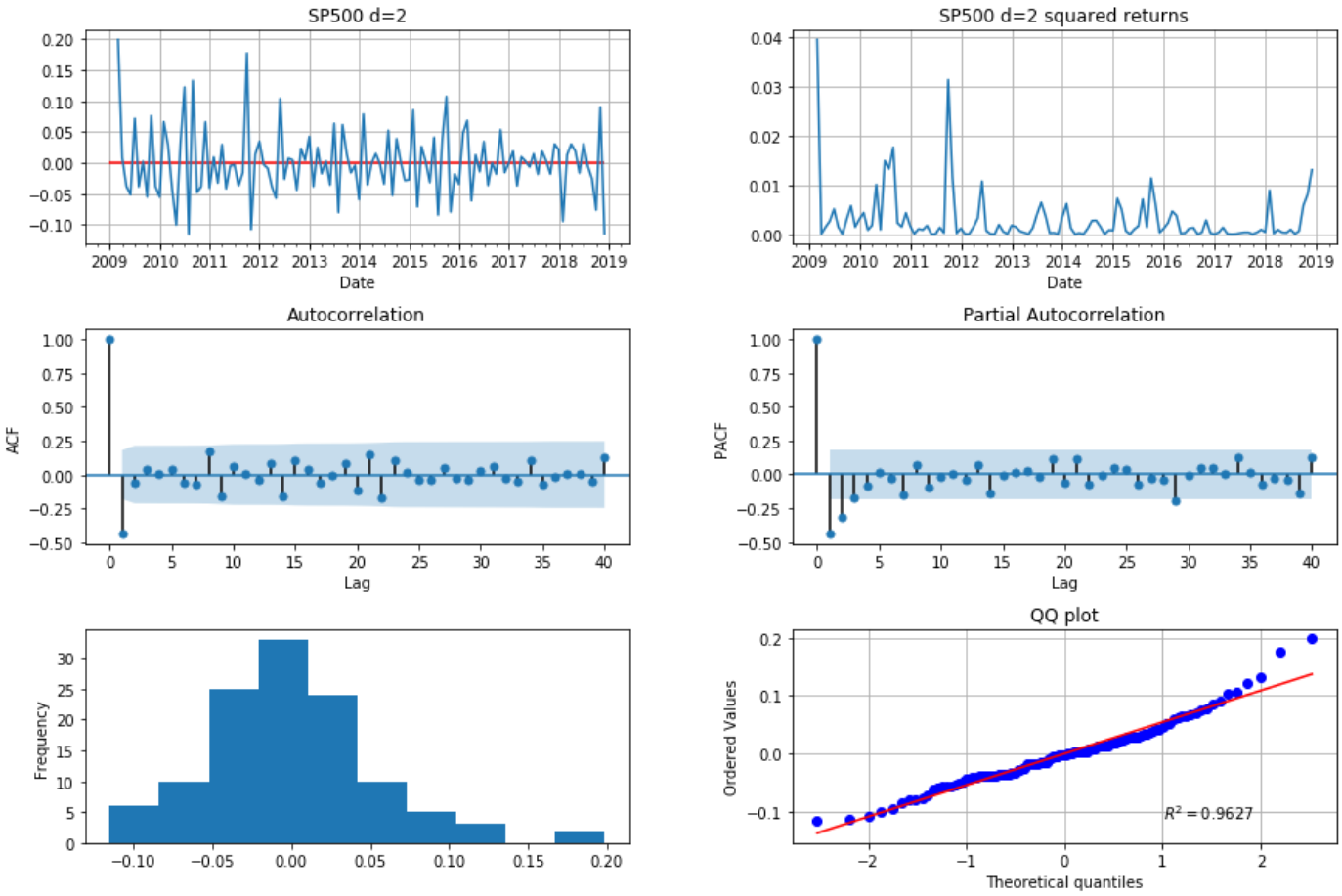
Google Trends



SP500

Once again the data was tested. The square of the returns showed volitility of some significance in some years within the given period and was therefore log transformed. The first difference of the data resulted in ADF test statistics of -11.89 (Critical values for 5% significance is -2.88) which is highly suspicious. It was once again discovered that the ADF had resulted in a type 1 error and rejected the null falsely. Hence, the second difference of the data was taken and the results are as shown.

SP500

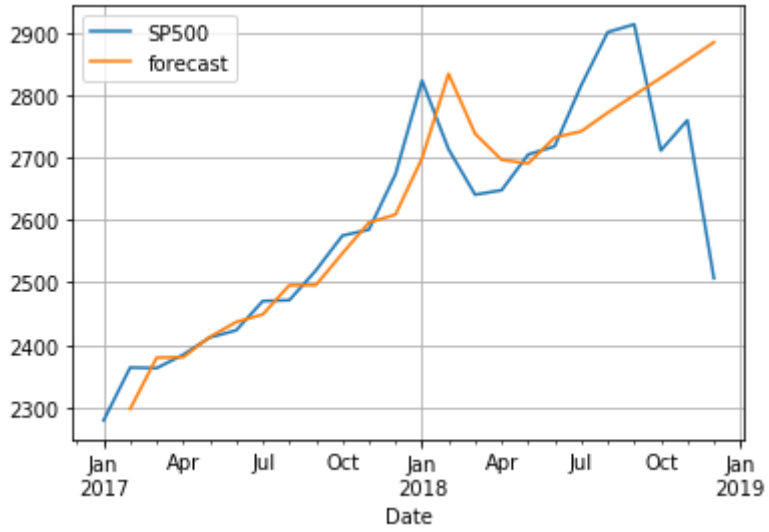


Forecasting

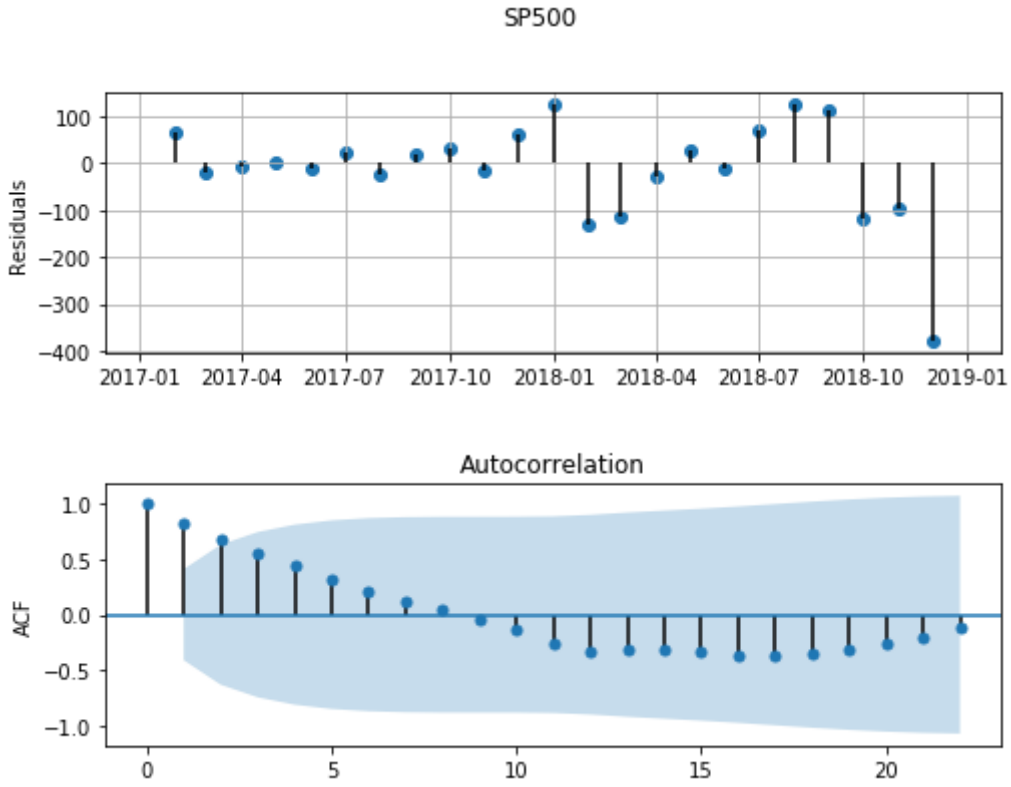
The data were first forecasted individually and then using a Vector Auto Regressive Moving Average (VARMAX).

SP500

The ACF and PACF showed a strting model of ARIMAX(3,2,1). 25 models were then iterated with $1 < p < 5$ and $1 < q < 5$ and the Akaike, Bayesian (aka Schwarz) and Hannan-Quinn information criteria (AIC, BIC/SIC, HQIC) was minimised. AIC recommended the order (2,2,1) and BIC and HQIC both showed fit of order (1,2,1) to be the best fit model.



The lagged SP500 suggests that the data is not fully utilised and the plot of the residuals shows that the model has been ineffective; since there is clear serial correlation and non-stationarity:



This can be due to omitted variables. This would therefore be a good reason for applying a VARMAX which would take into account the co-linearity of google trends data with SP500.

Another possibility is functional misspecification. This can arise from having non-stationary data where we try to fit an ARIMA model on it. Having differenced the SP500 data twice with a log transform diminishes the possibility of this error. As a fail safe, the data was tested with higher orders of differencing where the serial correlation in the residuals still presisted.

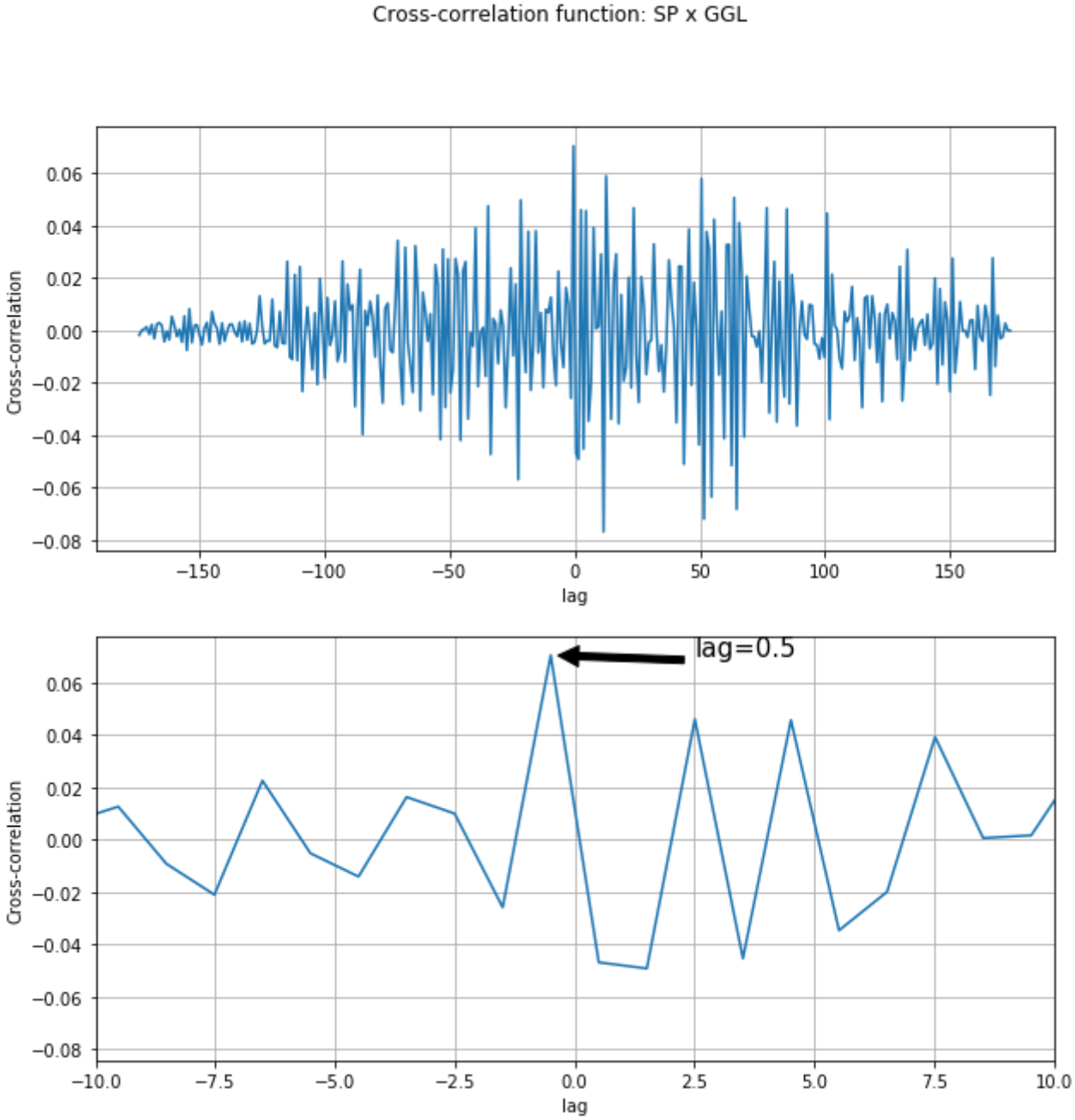
Having seen the slight deviation from normality of the original data with heavy tails in the distribution can indicate the reason to be clustering and biased standard errors. As previously mentioned, the analysis has been indicative of heteroskedacity in the data and having fit the model using the YW equations to the data, we also assume that the Gass-Markov assumptions are true which would be false. Therefore the BLUE (best linear unbiased estimator) is no longer unbiased and the MA terms need to be re-evaluated.

Finally, underlying issues such as measurement error can be neglected since no electronic data cannot suffer from this.

Cross-correlation

The cross-correlation function was applied to the second order log differenced data (after detrending the google data), by lagging the following:

$$r_{xy} = \frac{\sum((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$



VAR(p,q)

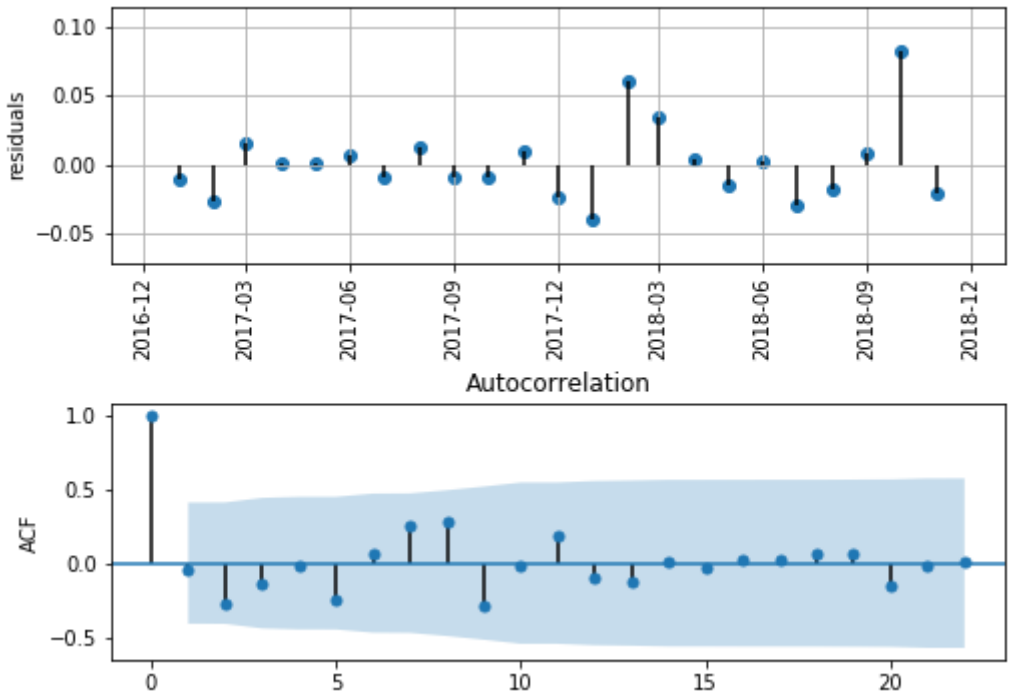
VAR follows exactly the same model as ARMA(p,q) in a vector form:

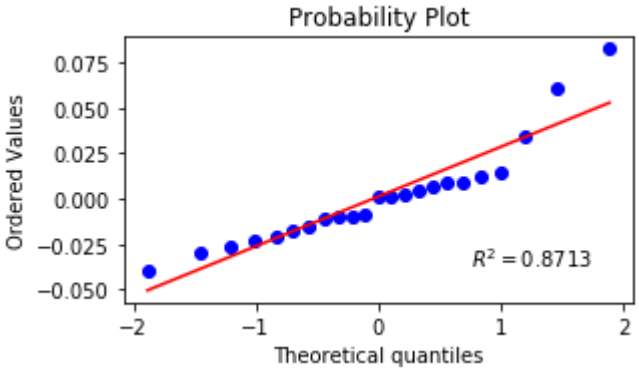
$$y_t = \alpha_1 . x_{t-1} + \ldots + \alpha_p . x_{t-p} + \theta_1 . \omega_{t-1} + \ldots + \theta_p . \omega_{t-q}$$

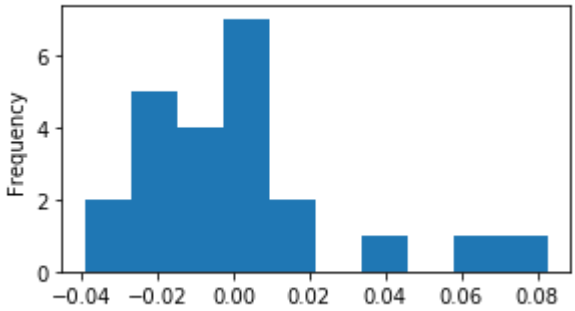
The VAR was iterated over $1 < p < 5$ and $1 < q < 5$ and for each the data was shifted and then the information criteria was calculated. Even though the cross correlation showed that google has a half a step predictability, this would only be useful if we had more data points at higher frequency. The VAR was iterated over a range of shifts in the GGL data to evaluate the model fit. All test results show an improvement when the google data is not shifted.

shift	-AIC	-BIC
-4	776	740
-3	780	742
-2	779	737
-1	778	738
→**0**	**783**	**747**
1	780	744
2	776	740
3	772	736
4	770	733

The plot residuals also show satisfactory conditions for the model however the errors are not normally distributed iid. In fact, the fat tail shown in the QQ-plot shows dangerous over-estimations by the model which need to be addressed. Although this could also be due to the small sample size where few outliers can cause significant skew in the data distribution. **volatility forecast → GARCH**







Remarks

The addition of moving averages into the models provides a high dependency on sample size. However, the current model still provides an acceptable estimation for the in sample forecasts.

It can be suggested for the model to be improved by fitting a GARCH(p,q) model to also forecast future volatilities which can be observed in the data over the second half of 2018.

Issues and questions

1. ADF gives type 1 error when SP500 is log differenced once. Why?
2. Is it alright to normalise the variance by the maximum variance in that period?
3. how much should we focus on the maths behind the analysis? because there is a lot of analysis and therefore maths involved and due to the report limit, not all of it can be covered. From what I recall, you mentioned the audience having highschool level knowledge but the maths involved requires more than that. There are a lot of references made to for example YW or BC or BJ and i cant explain all of the tests done and why i chose that over other tests.
4. What should we talk about when we carryout the KMC? what aspect of it are we analysing?
5. Need to extract original data from
6. HIGHLY suggested that GARCH be fitted to VAR as residuals show heteroskedasticity

In [ ]: