MAT2040: Project #2

Lingxiao XU
119020061

The Chinese University of HongKong, shenzhen — December 29, 2020

PART 1

We need to build a model to predict the product price in this project. This project took the product price on a daily basis, which means each price was collected on consecutive days. Firstly, a data visualization of the training data was performed.

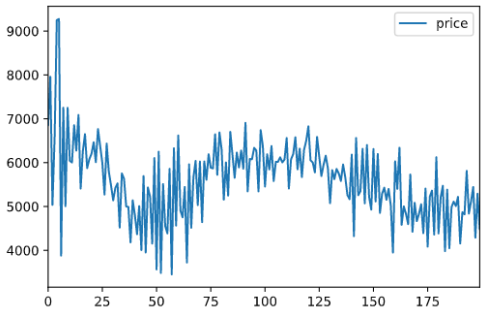

Figure 1: The data visualization of the part 1 training data

Model 1: Autoregressive Model

The Autoregressive Model Specifies that the dependent variable depends on its own previous values, which means we could use its past values to predict the future values.
The notation AR(p) indicates an Autoregressive model of order p, which means the model uses p lags of time slots to predict the dependent variable

$$X_t = c + \sum_{i=1}^{p} \varphi_t X_{t-i} + \varepsilon_t \tag{1}$$

The estimation of the coefficients in this model could be done by finding the least squares estimation of the following system

$$\begin{bmatrix} X_{p+1} \\ X_{p+2} \\ X_{p+3} \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 & X_2 & X_3 & \cdots & X_p \\ 1 & X_2 & X_3 & X_4 & \cdots & X_{p+1} \\ 1 & X_3 & X_4 & X_5 & \cdots & X_{p+2} \\ 1 & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n-p-1} & X_{n-p} & X_{n-p+1} & \cdots & X_{n-1} \end{bmatrix} \begin{bmatrix} c \\ \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \vdots \\ \varphi_p \end{bmatrix} \tag{2}$$

Info: Other methods like The Yule Walker Equations could also be used to estimate the coefficients of autoregression model.

0.0.1 Lags Selection

The key question in this model is to determine how many lags should be taken into the model, which is the value of p in the model. This project used two standards to select time lags, which are auto correlation function(ACF) and partial auto correlation function(PACF)

Auto Correlation Function (ACF)
ACF describes the autocorrelation between an observation and another observation at a prior time step that includes direct and indirect dependence information. This means we would expect the ACF for the AR(k) time series to be strong to a lag of k and the inertia of that relationship would carry on to subsequent lag values, trailing off at some point as the effect was weakened.

$$R_{xx}(t_1, t_2) = E[X_{t1}\overline{X_{t2}}] \tag{3}$$

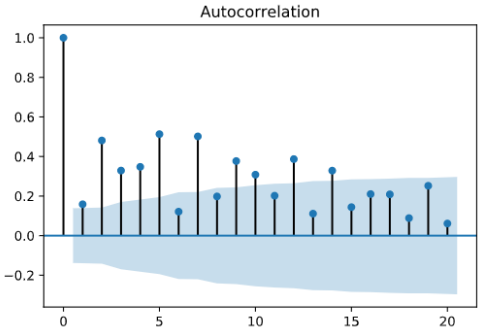

Figure 2: ACF of part 1 training data

The figure shows that the AR(10) time series has statistically significant relationship.

Partial Auto Correlation Function (PACF)
The partial autocorrelation at lag k is the correlation that results after removing the effect of any correlations due to the terms at shorter lags.
Given a time series Z_t, the partial autocorrelation of lag k, denoted as α_k, is the autocorrelation between Z_t and Z_{t+k} with the linear dependence of Z_t on Z_{t+k-1} removed; equivalently, it is the autocorrelation between Z_t and Z_{t+k} that is not accounted for by lag 1 through $t + k - 1$, inclusive.

$$\alpha(1) = corr(Z_{t+1}, Z_t), for k = 1 \tag{4}$$

$$\alpha(k) = corr(Z_{t+k} - P_{t,k}(Z_{t+k}), Z_t - P_{t,k}(Z_t)), for k \geq 2 \tag{5}$$

where $P_{t,k}(x)$ is the orthogonal projection of x onto the linear subspace spanned by $Z_{t+1}, \cdot, Z_{t+k+1}$.

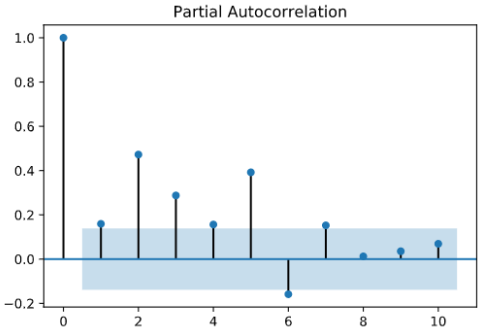

Figure 3: PACF of part 1 training data

By considering ACF and BACF, this report selected lag1, lag2, lag3, lag4, lag5, and lag6, which means p=6 in equation (1).

0.0.2 Implementation and Prediction

This project used AutoReg function from statsmodels package to implement and predict the model. The result of the regression is presented as follows:

AutoReg Model Results

Dep. Variable:	price	No. Observations:	200
Model:	AutoReg(6)	Log Likelihood	-1498.410
Method:	Conditional MLE	S.D. of innovations	547.206
Date:	Tue, 29 Dec 2020	AIC	12.692
Time:	19:54:04	BIC	12.827
Sample:	6	HQIC	12.747
	200		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	1530.6964	384.158	3.985	0.000	777.760	2283.633
price.L1	-0.2667	0.065	-4.079	0.000	-0.395	-0.139
price.L2	0.3117	0.056	5.534	0.000	0.201	0.422
price.L3	0.1812	0.058	3.100	0.002	0.067	0.296
price.L4	0.1340	0.057	2.367	0.018	0.023	0.245
price.L5	0.3812	0.055	6.924	0.000	0.273	0.489
price.L6	-0.0270	0.060	-0.447	0.655	-0.145	0.091

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-0.9914	-0.5466j	1.1321	-0.4198
AR.2	-0.9914	+0.5466j	1.1321	0.4198
AR.3	1.0792	-0.0000j	1.0792	-0.0000
AR.4	0.2629	-1.3334j	1.3591	-0.2190
AR.5	0.2629	+1.3334j	1.3591	0.2190
AR.6	14.4765	-0.0000j	14.4765	-0.0000

Figure 4: The result of the autoregression model

Based on this model, we could predict the product price in 100 days. The result is presented as follows:

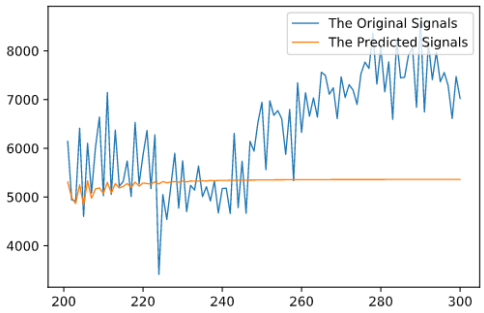

Figure 5: The predicted price and the true price

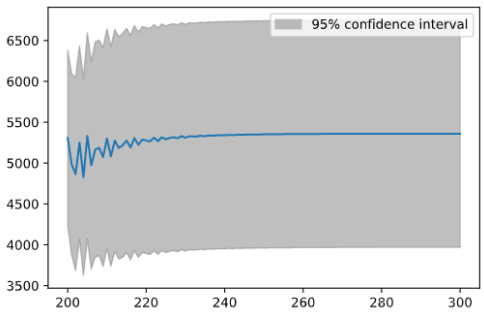

Figure 6: The Prediction Confidence Interval

The RMSE of this model is 1481.368.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i^2} \tag{6}$$

0.0.3 Validation of the Model

The least squares estimation of the autoregression model relies on several assumptions

1) **Homoscedasticity**: The variance of the residual is the same.
2) **Independence**: The error terms are not correlated.
3) **Normality**: For any fixed value of regressors, the dependent variable is normality distributed.

The following figure can well verify whether the model satisfies the above assumptions.

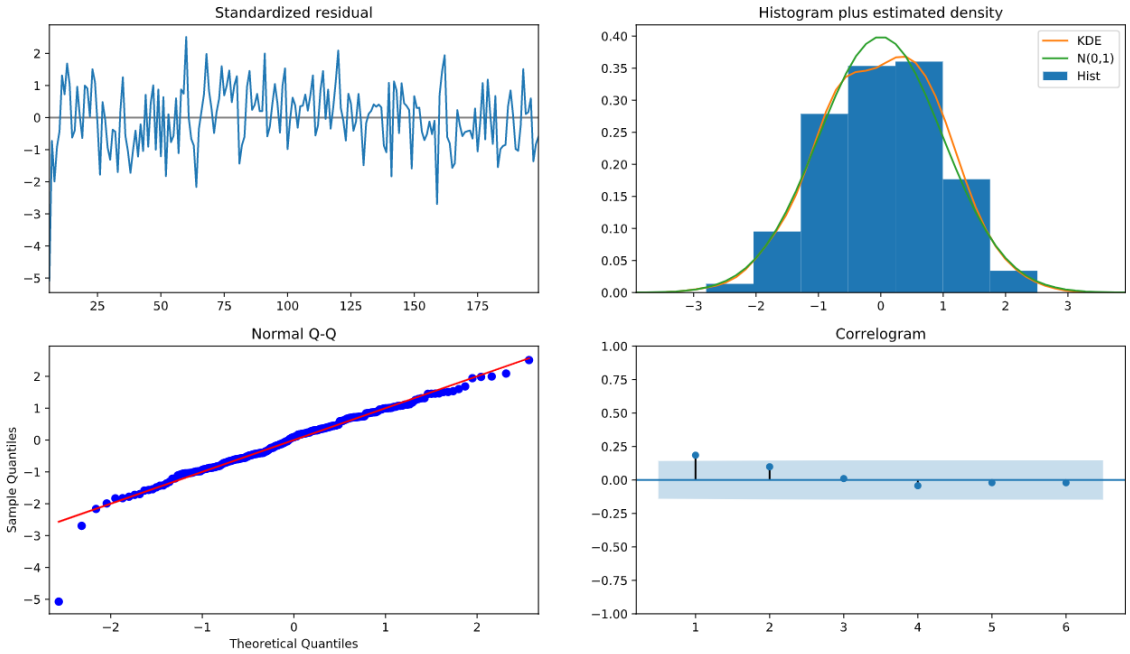

Figure 7: The Validation of the Assumptions

The first graph indicates that the residuals are not correlated with regressors The Second and third graph indicate that the dependent variable and regressors have finite fourth moments, and it meets the normality(residuals are normally distributed) as well as homoscedasticity assumptions in regression. The fourth graph indicates that there are no new auto correlated regressors should be taken into consideration(The chosen regressors are comprehensive enough)

Model 2: Fourier Series

The Fourier transform transforms a function of time and signal into a function of frequency and power. Real data often contains noise and Fourier transform makes it easy to see through the noise. Since the price of the product is a discrete time series, this report used discrete transform. The mathematics is presented as follows:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{i2\pi}{N}kn} = \sum_{n=0}^{N-1} x_n \left[\cos\left(\frac{2\pi}{N}kn\right) - i\sin\left(\frac{2\pi}{N}kn\right)\right] \tag{7}$$

In order to perform a prediction, this project firstly run a simple linear regression of the price on time. If the price has a periodic characteristic, its residuals should tend to be periodic. This project performs a Fourier transform on the residuals and take the Fourier terms into regression to calculate the Fourier coefficients. The R-squared of the regression model is 9.01e - 0.2.
Next, by performing a Fourier transformation, the dominant frequencies could be found. Here are the values these peaks correspond to.

The first simple linear regression only considers the time as regressor. After performing Fourier transform,

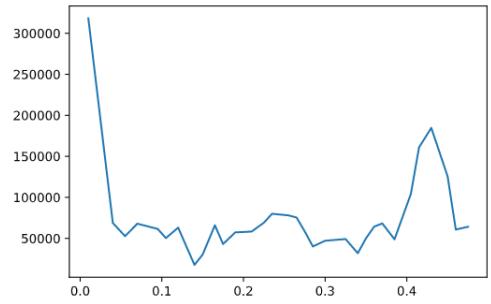

Figure 8: The Dominant Frequencies

label	fft	freq (1/365)	amplitude	phase
F_1	258057.887136-186228.431604j	0.010	318237.178612	-0.625111
F_2	181340.309169-35050.374509j	0.430	184696.606582	-0.190931
F_3	113871.798647-113617.949628j	0.415	160859.643804	-0.784282
F_4	98288.763921-78223.642969j	0.450	125616.955195	-0.672209
F_5	-71371.003425+75900.234638j	0.405	104185.727180	2.325450
F_6	70800.137613-37174.078240j	0.235	79966.065171	-0.483491
F_7	58677.008352-51446.724328j	0.255	78036.893539	-0.719836
F_8	48511.720134-57721.657474j	0.265	75400.110954	-0.871877
F_9	67798.757957-13407.395280j	0.225	69111.719908	-0.195234
F_10	58622.405537-35963.934789j	0.040	68774.930291	-0.550276
F_11	-53327.458382-42512.925659j	0.370	68199.462356	-2.468558
F_12	-25405.110005-62962.028632j	0.070	67894.305091	-1.954315
F_13	27935.287910+59693.536414j	0.165	65906.741690	1.133093
F_14	-54017.934365-34722.880599j	0.360	64215.384996	-2.570294
F_15	7856.663702-63637.874009j	0.475	64121.027541	-1.447959
F_16	-26072.744948-57452.496542j	0.120	63091.817125	-1.996817
F_17	-31320.161158-52897.749889j	0.095	61474.583677	-2.105378
F_18	-56786.104441-21528.404766j	0.460	60730.007981	-2.779220
F_19	29706.537074-50442.490299j	0.275	58539.928019	-1.038565
F_20	40805.896946+41610.504197j	0.210	58279.973276	0.795161
F_21	-378.432979+57302.824003j	0.190	57304.073592	1.577400
F_22	22952.246813-47405.655820j	0.055	52669.743093	-1.119895
F_23	3625.517070-50209.736801j	0.105	50340.461298	-1.498714
F_24	-44896.174056+21997.491695j	0.350	49995.560660	2.686006
F_25	-19300.708897-45222.804324j	0.325	49169.293211	-1.974183
F_26	-9342.018612+47958.375526j	0.385	48859.790163	1.763181
F_27	15610.305254-44454.363190j	0.300	47115.518004	-1.233094
F_28	31663.540060+29187.669733j	0.175	43063.904068	0.744733
F_29	38639.638243-10758.196374j	0.285	40109.355925	-0.271547
F_30	-29112.548336-13107.013198j	0.340	31927.014667	-2.718557
F_31	27625.041140-12370.240302j	0.150	30268.229930	-0.421015
F_32	-16756.539137+5561.372602j	0.140	17655.324099	2.821139

Figure 9: The fourier terms

fourier terms could be added into regression. The new term should be the function of the form:

$$f(t) = A\cos(\omega t + \phi) \tag{8}$$

A is the amplitude, ω is the angular frequency, and ϕ is the phase shift.
The fourier terms in figure 9 is the waves corresponding each peak. For example, the relationship between time and the first wave is presented as follows:

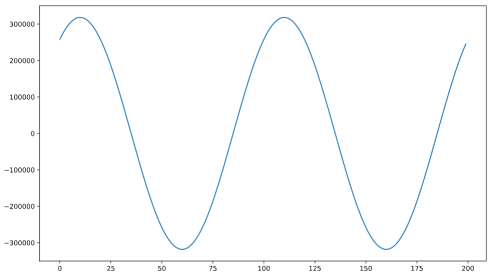

Figure 10: The relationship between time and F_1

Then, adding each waves from the Fourier transform into one column. Finally, include these terms into linear regression, the predicted value on the training data and the true value looks like:

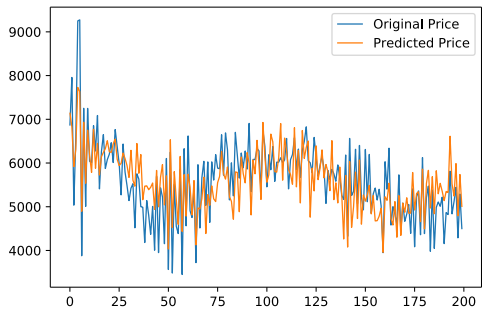

Figure 11: The predicted price on the training data

And we predict the price on the test data, which looks like:

The RMSE of this model is 1844.897.

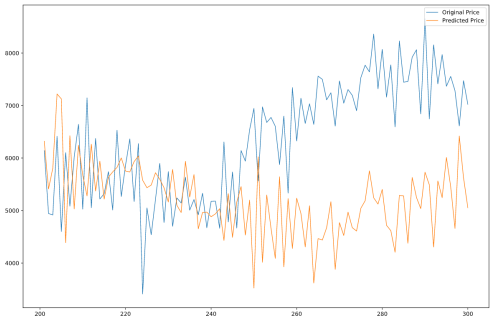

Figure 12: The predicted price on the test data

0.1 Model 3: Taylor Expansion

The Taylor Expansion requires the model looks like:

$$X_t = \sum_{i=0}^{N} a_i t \tag{9}$$

which could be done by polynomial regression. The key question in this model is to determine the value of N, which is the degree of the polynomial. This project used leave-one-out cross validation (LOOCV) to determine the degree of polynomial. Calculate the average MSE of each degree:

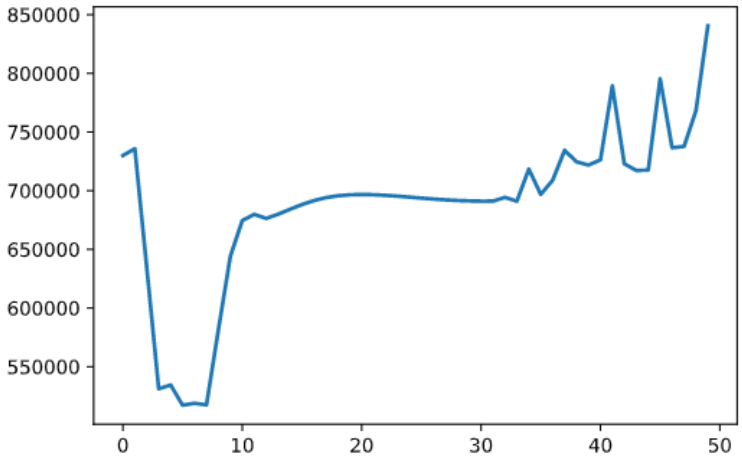

Figure 13: The average MSE of each degree

The average MSE of the 6 degree is the smallest. Thus, N = 6. A polynomial regression is performed and the result of prediction is presented as follows:

The RMSE of the Taylor series is 94479.060.

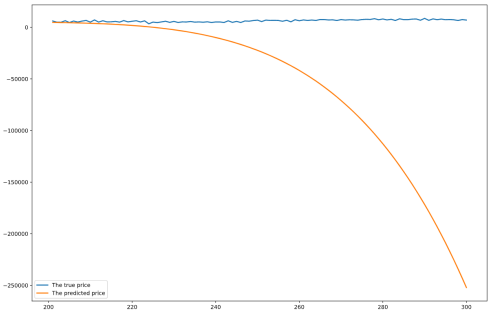

Figure 14: The result of prediction of Taylor model

0.2 Conclusion

The autoregression model gives the least RMSE. Thus, this report used Vector Auto Regression Model (VAR) in part 2 of this project.

1 Part 2

The data set of part 2 looks like: Vector autoregression (VAR) is a statistical model used to capture the

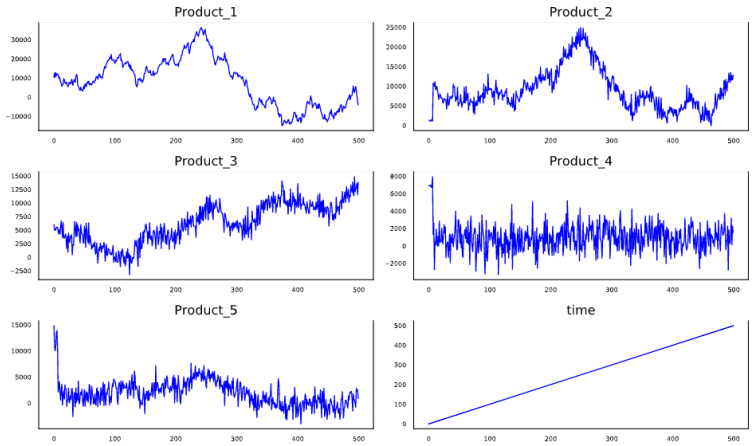

Figure 15: The data visualization of part 2

relationship between multiple quantities as they change over time. A VAR model describes the evolution of a set of k variables, called endogenous variables, over time. Each period of time is numbered, t = 1, ..., T. The k variables are modeled as a linear function of only their past values. This means we need to solve the least squares estimation of the system:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \\ \vdots \\ y_{k,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_k \end{bmatrix} + \begin{bmatrix} a^1_{1,1} & a^1_{1,2} & \cdots & a^1_{1,k} \\ a^1_{2,1} & a^1_{2,2} & \cdots & a^1_{2,k} \\ \vdots & \vdots & \vdots & \vdots \\ a^1_{k,1} & a^1_{k,2} & \cdots & a^1_{k,k} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \\ \vdots \\ y_{k,t-1} \end{bmatrix} + \cdots + \begin{bmatrix} a^p_{1,1} & a^p_{1,2} & \cdots & a^p_{1,k} \\ a^p_{2,1} & a^p_{2,2} & \cdots & a^p_{2,k} \\ \vdots & \vdots & \vdots & \vdots \\ a^p_{k,1} & a^p_{k,2} & \cdots & a^p_{k,k} \end{bmatrix} \begin{bmatrix} y_{1,t-p} \\ y_{2,t-p} \\ y_{3,t-p} \\ \vdots \\ y_{k,t-p} \end{bmatrix} \tag{10}$$

The Vector Auto Regression has a strict assumption on the stationary of the time series.Thus, one test had to be performed, which is Augumented Dickey-Fuller Test. It turned out that price of product 1, product 2, and product 3 are not stationary. We calculate the increment difference of each series and use them to perform an Augumented Dickey-Fuller Test. It turned out that five series are all stationary. Next, we need to test the correlation between each time series. The test is Granger's Causality Test. The result is presented as follows (The value in the form is the p-value of the test. The null hypothesis is the correlation

	Product_1_x	Product_2_x	Product_3_x	Product_4_x	Product_5_x
Product_1_y	1.0	0.0	0.0	0.0	0.0
Product_2_y	0.0	1.0	0.0	0.0	0.0
Product_3_y	0.0	0.0	1.0	0.0	0.0
Product_4_y	0.0	0.0	0.0	1.0	0.0
Product_5_y	0.0	0.0	0.0	0.0	1.0

Figure 16: Granger's Causality Test

between time series is 0): The key question in this model is the same as the autoregression model, which is to determine the value of p, the number of time lags we use. This report build the model from lag 1 to lag 80 and found the best lag by AIC. The best lag is 78. Finally, we use lag 78 to build and predict the model. The result is presented as follows: The RMSE for each product are:

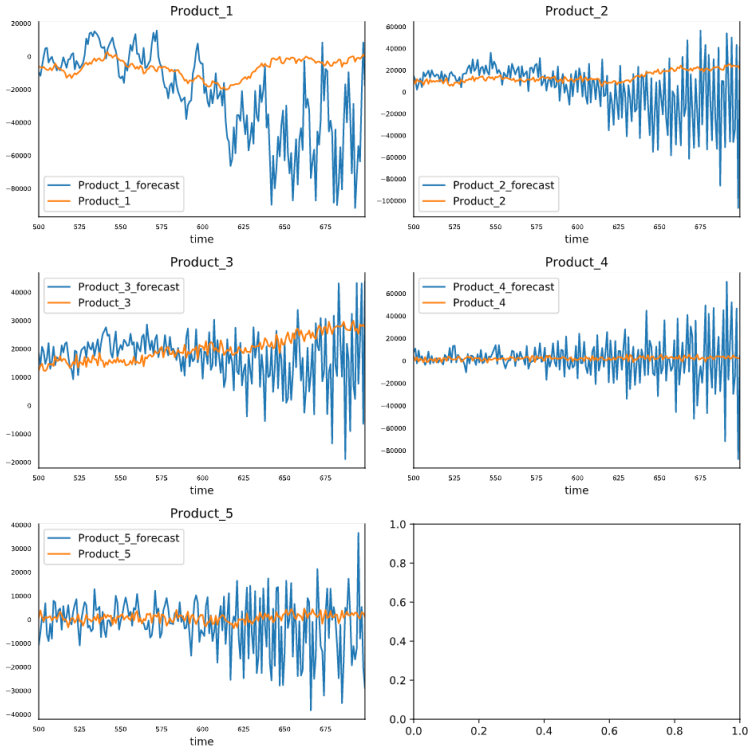

Figure 17: The prediction result of the VAR model

1) product 1: 31976.980883723056
2) product 2: 29334.36012630151
3) product 3: 10965.563702961275
4) product 4: 19445.44243437039
5) product 5: 11770.881628593599

The detailed implementation is shown in the code (written in Python on jupyter notebook)