STA4003 Time series
Final report
黄苏鹏 118020195

# Using hybrid ARIMA-SVM model for stock price predictions

## 1.Introduction

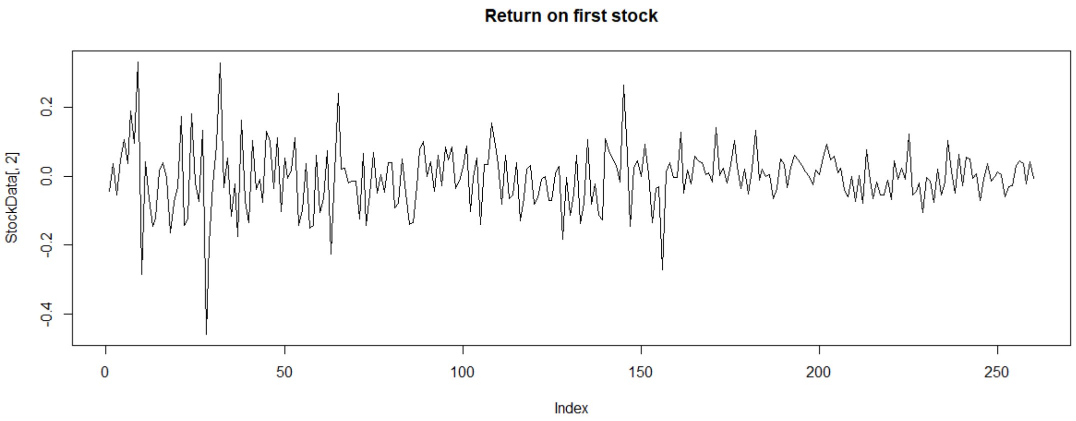
Forecasting stock prices is one of the most challenging applications of modern time series, and several models have been developed to reach a more precise prediction. The popular ARIMA model assumes a linear relationship between variables which limited its accuracy. Recently, various machine learning algorithms especially neural networks had been applied to Time Series forecast for their excellent ability to capture nonlinear components. Some researchers also try to combine these algorithms with ARIMA to capture both linear and nonlinear patterns. For example, Khashei and Bijari's model [1] integrated ARIMA and ANN, Pai and Lin[2], Chen and Wang[3] proposed a hybrid methodology to exploit the unique strength of ARIMA models and support vector machines. The hybrid model performed better in accuracy than only using the ARIMA model.

ARIMA is a linear model, while SVMs with a kernel is a non-linear model. In theory, the hybrid ARIMA and SVMs model has both a linear part and a non-linear part which may lead to a more accurate prediction. However, SVMs or other machine learning algorithms require feature inputs such as trading volume, company size and so on, considering only log return is provided in this project, applying the hybrid model is still a great challenge.

## 2.ARIMA model

### 2.1 Exploratory analysis

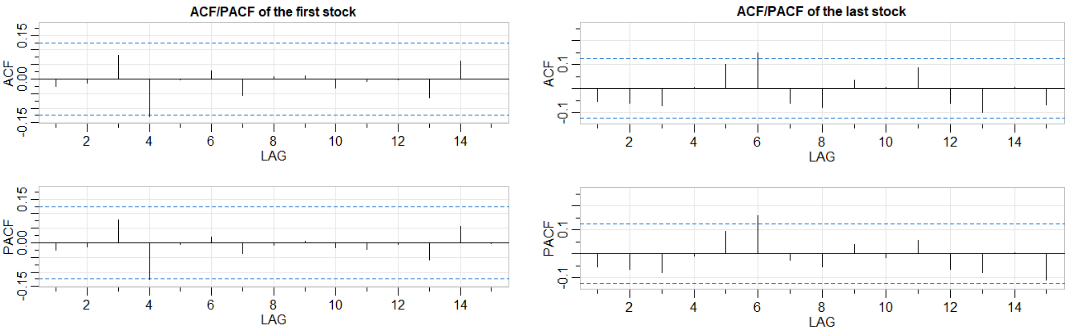
Firstly, we plot the data and check whether any transform is needed. In general, the data shows no clear trend and the variance is relatively stable. There is also no seasonal pattern. We firstly take first order difference and plot the sample ACF/PACF. The ACF and PACF decrease almost exponentially so we believe that the differencing sequence is stationary. We then conducted the augmented dickey-fuller test to check for unit roots and the p-value is all less than 1%. Therefore, we reject the null hypothesis that there is a unit root and we believe the series is stationary. This means that no differencing is needed for this data. The result is reasonable since the log return is obtained by ∇ log (Pt) which will stabilize the series and



make it stationary.

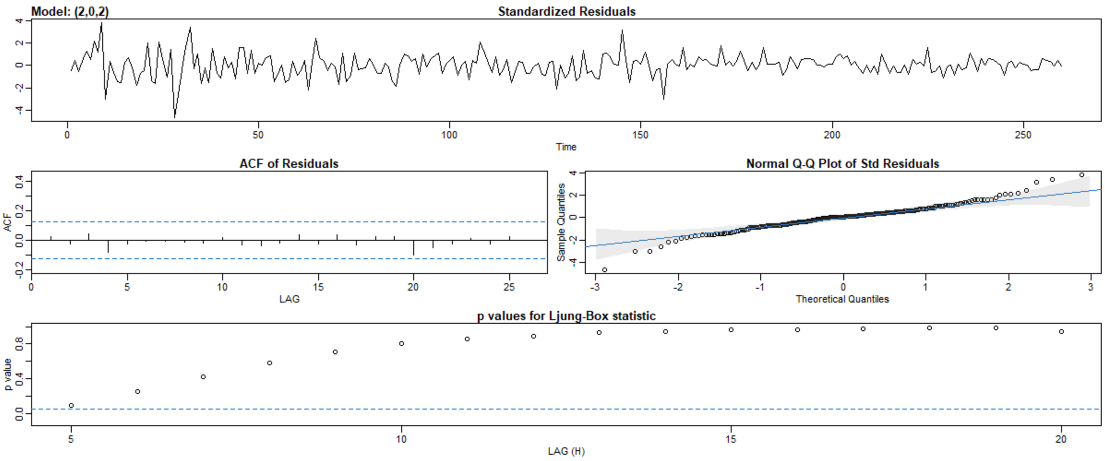
## 2.2 Model fitting

Then we try to fit an ARIMA model, after drawing the ACF and PACF, we can see that there is no significant pattern among time lags. This suggests that it may be inappropriate to determine the order using the plot so we use the AIC to determine the order of p, d, and q. As shown before, the series is stationary, so we let d equal to zero.



## 2.3 Model selection and Diagnostics

We use an algorithm to go through all possible orders and select models with minimum AIC. We exclude the ARMA(0,0) model since it gives us no time-series information. After fitting the model with minimum AIC, we need to check the model to ensure it is a good fit. If the model is correctly specified and the estimates are reasonably close to the true parameters, the residuals should behave roughly like a sequence of independent, normal random variables with zero mean and constant variance. After fitting the model, we do diagnostics and find that residuals are not normally distributed judging by the Q-Q plot so ARMA alone is not a good model. The independence assumptions seem to be hold since the residuals' ACF is close to 0 and the test statistics for Ljung-Box are not significant which means the model is appropriate. Since the return shows highly volatile periods clustered together, we also fit a GARCH (1,1) model to the ARMA.

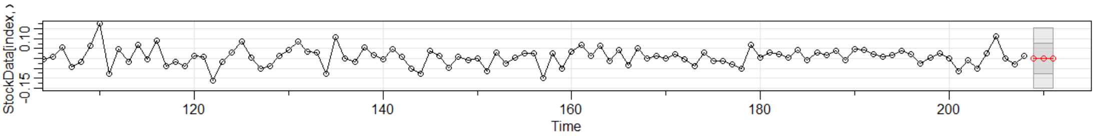


| Model | MSE1 | MSE2 | MSE3 |
|---|---|---|---|
| ARMA by AIC+GARCH(1,1) | 0.00459 | 0.00458 | 0.00458 |
| **ARMA(1,0)+GARCH(1,1)** | **0.00437** | **0.00458** | **0.00457** |
| ARMA(1,0) | 0.00463 | 0.00462 | 0.00461 |

However, the MSE of the first model is large while fitting a classic ARMA (1,0) with GARCH (1,1) can get a decent result. This may because by using the AIC we tend to choose large p or q so that we overfit the data. Since the above diagnostic process cannot be done when new data comes in, using a simple AR(1) model and a GARCH(1,1) is reasonable. We also do model checking for AR(1) on the test data and it is almost as good a fit as the model fitted by AIC. The difference in AIC is also small (around 1%). Therefore, we use the second model as the universal model to fit all the stocks across all time.

### 2.4 Discussion

The problem with the ARMA and GARCH model is that they assume a linear relationship among variables and thus can only capture very little trend movement. As shown in the figure below, its prediction is very close to the mean of the return. Therefore, to better predict the trend movement, we consider adding the Support Vector Machine to capture the nonlinear component in the trend movement.



## 3.ARIMA-SVM model

### 3.1 Introduction to the SVM model

Support vector machines construct a set of hyper-planes in a high-dimensional or infinite-dimensional space, and they can usually be used for classification, regression, or other tasks. Since linear classifiers do not work well in most scenarios, SVM enlarges the space of features by including nonlinear mappings through the kernels. After mapping the data to higher dimensions, linear regression is performed in that space.

Formally speaking, SVM is doing the optimization:

$$\underset{w,b,\xi,\xi^*}{minimize}\ \frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}(\xi_i+\xi_i^*)$$

$$s.t.\begin{cases} y_i-\langle w,\phi(x_i)\rangle-b\le\varepsilon+\xi_i \\ \langle w,\phi(x_i)\rangle-y_i+b\le\varepsilon+\xi_i^* \\ \xi_i\ge 0,\xi_i^*\ge 0, i=1,\cdots n \end{cases}$$

To minimize the error of regression with a cost constant $C$ that controls the penalty imposed on observations that lie outside the epsilon margin ($\varepsilon$) and helps to prevent overfitting (regularization). The restrictions are that all residuals having a value less than ε plus a soft margin $\xi_n$ .

The kernel function is defined as

$$K(x,y)=\sum_{i=1}^{D}\phi_i(x)\phi_i(y)=\langle\phi(x)\phi(y)\rangle$$

which leads to the nonlinear regression function:

$$f(x,\alpha,\alpha^*)=\sum_{i=1}^{D}(\alpha_i-\alpha_i^*)K(x_i,x)+b$$

### 3.2 ARIMA-SVM model

Stock movements are rarely purely linear or nonlinear so neither ARMA nor SVM can handle the data well. Hybridizing the two models can achieve a more accurate result and yield a robust method. The log return of the stock is considered to contain the linear and nonlinear parts of the movement. Intuitively, we can think that the log return is the sum of linear and nonlinear trends, which gives us the first model. Since it is difficult to define 2-step, 3-step for the hybrid model, we will only apply the hybrid model for 1-step forecast, that is using every information before the prediction date. We scale the input first by subtracting its mean and then divided by its standard deviation to enhance the performance of the SVM model.

### 3.2.1 Hybrid model 1:

$$\widetilde{N}_t = f(\varepsilon_{t-1}, \mathrm{Y}_{t-1}, \dots, \mathrm{Y}_{t-m}, x_1, \dots, x_n)$$

$$\tilde{Y}_t = \tilde{L}_t + \widetilde{N}_t$$

where $\widetilde{N}_t$ is the forecast value of the nonlinear component of the log return at time t, f is a nonlinear regression function determined by the SVM model.$x_1, \dots, x_n$ are the features we constructed from the data to improve the prediction, m is some constant. $\tilde{L}_t$ is the linear forecast obtained by the forecast from the ARIMA equation, $\mathrm{Y}_{t-1}, \dots, \mathrm{Y}_{t-m}$ are the lag returns. The residuals ε of the ARMA and GARCH will be input to the SVM models as response variables to capture the nonlinearity that failed to be captured by the ARMA-GARCH model. The lag returns should contain nonlinear components from higher time lags. At last, the final forecast results will be calculated as the sum of the linear and nonlinear forecasts.

### 3.2.2 Hybrid model 2:

$$\tilde{Y}_t = f(\varepsilon_{t-1}, \tilde{L}_t, \mathrm{Y}_{t-1}, \dots, \mathrm{Y}_{t-m}, x_1, \dots, x_n)$$

Inspired by Zhu and Wei [4], more generally, the relations of linear and nonlinear part of the trend may not be simple summation so we regard the log return as a function of linear and nonlinear components. The input of SVM then becomes the lag 1 residuals, the prediction of the return based on ARMA-GARCH model, the lag log returns up to m and constructed features. These inputs should cover all the linear and nonlinear information we can know from the dataset.

### 3.3 Fitting the Hybrid model

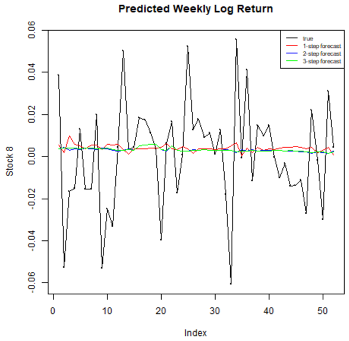
We fit the SVM based on ARMA (1,0) and GARCH (1,1). In order to utilize new coming data, we ideally want to tune the SVM model every time as the prediction windows move which will require more computational time. However, this process is time consuming so we only tune the model once using Bootstrap sampling at the first time. Additionally, we try to create several features to enhance the prediction. The following table shows some attempts we have done. However, adding these features yield to a worse result which means the features may be just some noises so we exclude them from the model.

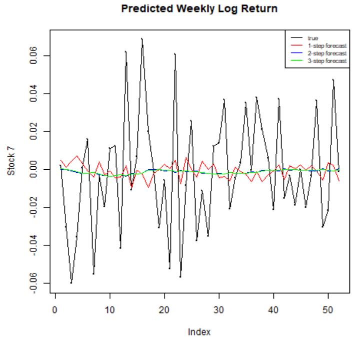
| Features | notes |
|---|---|
| Rolling Median | The rolling median of the last two week's data. |
| Rolling Standard Deviation | The rolling Standard Deviation of the last two week's data. |
| Rolling Skewness | The rolling Skewness of the last two week's data. |
| Rolling Kurtosis | The rolling Kurtosis of the last two week's data. |

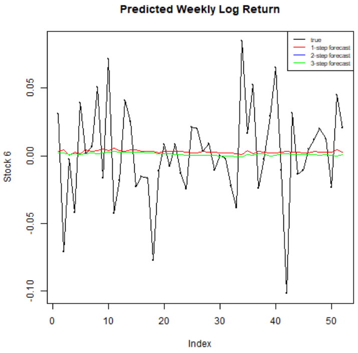
The results are shown as below:

| Model | MSE1 |
|---|---|
| Hybrid model 1 | 0.00455 |
| **Hybrid model 2** | **0.00446** |

We choose hybrid model 2 as our final model. Although the 1 stage forecast MSE is slightly higher than the single model, the hybrid model can better forecast the trend (compare red and green lines).







## 4. Portfolio Construction

We use the Markowitz's Modern Portfolio Theory (MPT) to finish the task. The maximization of the Sharpe ratio is essentially a quadratic maximization problem.

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{maximize}} & \dfrac{\mathbf{w}^T\boldsymbol{\mu} - r_f}{\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}} \\ \text{subject to} & \mathbf{1}^T\mathbf{w} = 1, \quad (\mathbf{w} \geq \mathbf{0}) \end{array}$$

Where w is the vector containing the weight of each stock and Σ is the 10x10 covariance matrix. In our case, rf is the return of sp500 and rf is obtain by the prediction of AR (1) model of SP500 return. Furthermore, we'd like to impose some restrictions on the original problem.

$$w_i \leq a \qquad \max position\, on\;\; longing,$$

$$w_i \geq b \qquad \max position\;\; on\;\; shorting\,, \quad for\;\; i = 1,2, \dots 10$$

Since our prediction is still far from capturing the movement of price, we do not want the solution weight to be concentrated on 1 or 2 stocks so we post longing and shorting limits. The covariance is calculated by multiplying an exponential decay function to the sample covariance so that weights assigned to observations are declining as they go further back in time. The smaller the λ, the faster the function will decay and when λ is equal to 1, it is the same as the sample covariance matrix. t=T is the most recent date and t=1 is the farthest date. Also, we update our covariance matrix whenever new data comes in.

$$\boldsymbol{s = \Sigma_{t=1}^{T} p_t (R_t - \bar{R})(R_t - \bar{R})^T}$$

$$\boldsymbol{p_t = \frac{\lambda^{T-t+1}}{\Sigma_{t=1}^{T} \lambda^t}}$$

By using the *solnp* function in R, we get the following result, the lambda is then set to be 0.98 as our final model. Also, as discussed above, the

| Lambda | Hybrid Sharpe ratio | ARMA-GARCH Sharpe ratio |
|---|---|---|
| 0.94 | 0.174 | 0.160 |

| | | |
|---|---|---|
| 0.96 | 0.168 | 0.173 |
| **0.98** | **0.276** | 0.209 |
| 1 | 0.220 | 0.184 |

## 5. Discussion

As stated above, the hybrid model can better describe the ups and downs of the log return, and its accuracy is slightly higher than the ARMA-GARCH model. We should always remember our goal is to create a profitable trading strategy by our forecast so that a flat predicted return curve is not what we want. Actually, if we solely predict 0 for all the stocks in the future, it will yield even lower MSE but it is impossible to construct a portfolio. By using the forecast of the hybrid model and the Modern Portfolio Theory, we build a portfolio with a satisfactory Sharpe ratio, although the MSE is slightly larger than the ARMA-GARCH model. Furthermore, this model should be robust across different data set since the ARMA(1,0)+GARCH(1,1) are very unlikely to overfit while the SVM can capture the relation among higher lag and thus solve the underfit problem. However, given only the log return which is likely to be a white noise sequence, it may be that there are just no patterns in return at all and thus all the predictions are futile.

## 6.Summary

In this project, we create a hybrid model taking advantage of the linear ARMA-GARCH model and the nonlinear SVM model. The prediction by the hybrid model can capture more trend movement and can consequently improve our portfolio performance. The model should be applicable through different data sets and can yield a good result.

References

[1]M. Khashei and M. Bijari, "A novel hybridization of artificial neural networks and ARIMA models for time series forecasting," Applied Soft Computing, vol.11, pp. 2664-2675, March 2011

[2] P.F. Pai, C.S. Lin, A hybrid ARIMA and support vector machines model in stock price forecasting, Omega 33 (2005) 497–505.

[3] K.Y. Chen, C.H. Wang, A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan, Expert Systems with Applications 32 (2007) 254–264.

[4] Zhu BZ, Wei YM. Carbon price forecasting with a novel hybrid ARIMA and least squares support vector machines methodology. Omega 2013;41:517–524..