Predictability of Industry Daily Returns

Author: Siow Meng Low

Introduction

The data consists of daily stock returns in forty-nine different industries, from 2-Jan-2015 to 31-Jan-2017. The companies are grouped into these forty-nine industries based on their Standard Industrial Classification (SIC) codes. The detailed sector information can be found at Kenneth French's website.

The objective is to use the past daily returns to predict future returns in these industries.

Industry Stock Returns Relationships

Contemporaneous Relationships

Given that equities belong to a single asset class, the stock returns of companies from different industries should be positively correlated in general. In other words, at the same time period *t*, stock returns from different industries tend to rise or fall together. This can be seen from the pairwise correlations in the given data where almost all correlation coefficients are positive.

Furthermore, the stock returns from similar industries are more strongly correlated. For example, ***Banks*** (consists of financial services firms) and ***Fin*** (brokerages and investment management businesses) sectors are very highly correlated (correlation coefficient of 0.9391, in the given data). This is because their businesses are so closely linked that any industry news that moves the stock prices of companies in one sector will affect the other sector as well.

Another interesting fact is that the ***Gold*** industry has very low correlation coefficient with most of the other industries. This is because the gold commodity is traditionally viewed by investors as a safe haven asset and the companies involved in gold business are viewed as proxy to invest in gold commodity. As a result, it has low correlation with other industries.

Predictive Relationships

Across Industries

Lo & Mackinlay (1990) has shown that stock returns are often positively cross-autocorrelated. Since an industry is made up of many stocks operating in the same sector, we would expect the cross-autocorrelations between industries to be positive in general. In other words, higher (or lower) returns from other sectors in the previous periods will tend to predict higher (or lower) return for a particular sector. It was illustrated that this positive cross-autocorrelation can be attributed to the lead-lag effects where large-capitalisation stocks tend to lead small-capitalisation stocks (Lo & Mackinlay, 1990).

Chinco et al (2017) has also shown that the regressors selected by LASSO tend to predict future news announcement. Therefore, the stock returns of the companies, that are going to have news announcement, reflect a certain predictability of future stock return of other companies and this can be picked up by simple LASSO regression.

Within Industries

Moskowitz & Grinblatt (1999) demonstrated in their paper that there exist short-term momentums within industries, hence we expect the returns within industries are positively autocorrelated. In other words, higher (or lower) previous returns for a particular industry, tend to predict higher (or lower) future returns for the same industry.

It is also illustrated that market indexes are positively autocorrelated in general (Lo & Mackinlay, 1990). Similar to market indexes, industries are made up of large number of stocks. Thus, we would expect the returns within an industry to be positively autocorrelated as well.

Predictive Models

Model Training and Testing

Rolling Window Size

The size of the rolling window is chosen to be 80 trading days, close to 4 calendar months. This size provides sufficiently large amount of recent data for training and cross-validation purposes.

Cross-Validation Method

The daily returns data is a collection of time series data samples. Consequently, the traditional K-Fold cross validation method will not be appropriate since it might use future data for training and past data for testing. The training-test split must be performed such that the training set belongs to the earlier time period while testing set belongs to the later time period.

The scikit-learn package in Python has a function **TimeSeriesSplit()** to perform this (please refer to the Jupyter notebook). The total number of split is chosen to be ten, a good balance between model out-of-sample accuracy and training time.

Hyper-parameters

The list of hyper-parameters (of each predictive model) tuned via cross-validation, are as below:

Predictive Model	Hyper-Parameters
LASSO	λ, L1 shrinkage parameter
Elastic Net	λ_1 & λ_2 , L1 and L2 shrinkage parameters
Random Forests	Minimum number of samples required to be at leaf node
Gaussian Process Regression	Variance of the noise in the observations
Support Vector Regression	C: Penalty parameter of the error term ε: Width of the epsilon-tube where no penalty is applied

Table 1 Predictive Models Used and Their Hyper-Parameters

Benchmark Performance

To quantify how well the models perform, benchmarks will be used to judge how the models perform relative to these benchmark models. The two models used as benchmark are OLS (1-Day Lag) model and historical mean model.

OLS (1-Day Lag) Benchmark

OLS benchmark assumes linear relationships between an industry's daily return and the previous returns from all the industries. It is expressed mathematically as in the below equation:

$$R_{i,t} = \beta_0 + \beta_1 R_{1,t-1} + \beta_2 R_{2,t-1} + \cdots + \beta_{49} R_{49,t-1} + \epsilon_{i,t}$$

The observations within the rolling window are used for OLS regressions. The estimated linear model will then be used to predict the next out-of-sample daily return of a particular industry.

Historical Mean Benchmark

The historical mean benchmark predicts the next out-of-sample daily return using the average daily returns across the entire window (i.e. 80 trading days). In mathematical notation, it is:

$$\overline{R_t} = \frac{1}{n}\sum_{i=1}^{n} R_{t-i}$$

n is the number of observations inside the rolling window

In order to beat the market returns, our predictive models must perform better than the two benchmarks.

Predictive Models Used

LASSO (n-day Lag) Model

LASSO (Least Absolute Shrinkage and Selection Operator) regression extends OLS by adding a L1 penalty term to the objective function. LASSO model minimises the objective function and shrinks some of the weights to zeroes. This regularisation is to reduce overfitting and improves out-of-sample predictive performance. The optimal value of λ is selected through the use of cross-validation.

LASSO is useful in selecting a subset of informative lagged returns for prediction. For the n-day lag model, all the lag daily returns (from 1 to *n* days lag) are included as independent variables.

Elastic Net (n-day Lag) Model

Elastic Net is also a linear regression model trained with L1 and L2 regularizers. It allows a combination of coefficient shrinkage using L1 and L2 penalties. The L1 and L2 shrinkage parameters are selected using cross-validation. In certain scenarios, Elastic Net has an advantage over LASSO where the highly correlated regressors will all be retained.

Random Forest (n-day Lag) Model

Random forest is a nonlinear machine learning method where a number of regression trees are built using bootstrap samples and each node split only considers a random subset of features. It is useful in identifying the rule-based relationships between lagged returns that predicts future returns.

The minimum number of samples required to be at the leaf node is adjusted to avoid growing overly large tree (and hence overfitting). The optimal value of this hyper-parameter is selected using cross-validation (from possible values of {2, 6, 10}).

Gaussian Process Regression (n-day Lag)

Gaussian process regression is a nonparametric nonlinear regression method. It represents the data as a sample from multivariate Gaussian distribution and estimates the probabilistic predictive distribution of the outcome variable given the training data. This might be useful in this context since financial data is often noisy and the probabilistic approach factors in uncertainties of the estimates.

Gaussian process regression is modelled using covariance function. Without prior domain knowledge of the true function, the covariance function is chosen to be the Radial Basis Function (RBF) kernel so that the probabilistic distribution of the predicted returns will be smooth. Gaussian process regression model also allows us to specify the variance of the white noise in the observations. This value is selected using cross validation.

Support Vector Regression (n-day Lag)

This is an extension method of Support Vector Machine to perform regression. In this assignment, RBF kernel is used for nonlinear mapping to detect if there is any nonlinear predictive relationship between lagged returns and future return. Support Vector Regression (SVR) attempts to regress a boundary where the data points falling inside the epsilon-tube, will not affect regression. The value of epsilon is tuned via cross-validation to reduce overfitting.

Performance Analyses

Economic Significance (Daily Sharpe Ratio of S&P 500 Returns)

To quantify the economic significance of the predictive models, we would need the daily Sharpe ratio of stocks. The daily closing price of S&P 500 index (3-Jan-1950 to 18-Apr-2017) is downloaded from Yahoo Finance and used to calculate the expected daily return and its standard deviation.

Since US-based investors often use the interest rate of three-month US Treasury Bill as risk free rate (Investopedia, 2017), we obtain the annual interest rate (0.82%) from US Treasury website. The Sharpe ratio of S&P 500 daily return is then calculated to be 0.0318 (please refer to Jupyter notebook for detailed calculation).

As introduced in class, the ratio of the two expected returns (with and without predictability) is very close to R_{OOS}^2/S^2 . This ratio will be used to quantify the economic performance of the models.

Performance Evaluation (LASSO)

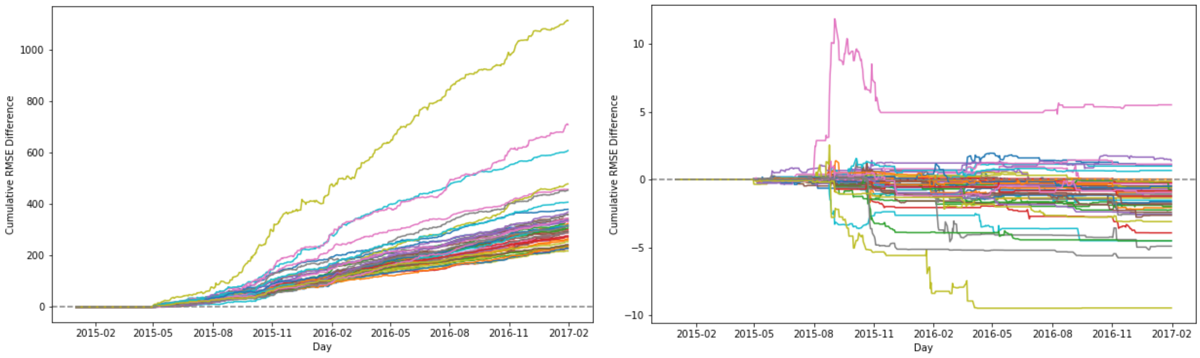

Figure 1 Cumulative RMSE Difference (LASSO-1 Day Lag), Left: OLS Benchmark, Right: Historical Mean Benchmark

The above two graphs are the cumulative RMSE difference between the two benchmarks and LASSO (1-day lag) model. We can see that LASSO outperforms OLS model by a large margin but underperforms Historical Mean model. This is because OLS is not a particularly good model for forecasting. All the regression techniques used are able to beat OLS forecast. Therefore, the rest of the section will focus on using historical mean model as benchmark.

For the graph on the right, in many of the time periods, the cumulative RMSE difference is simply a horizontal line. This means LASSO simply reverts back to the mean value for forecasting. For many of the industries, the curves fall below the zero-line, indicating LASSO underperforms historical mean.

Best Model & Best Lag

In looking for a method that performs better than historical mean, a number of regression methods (with different lags) have been run, and the average $\boldsymbol{R^2_{OOS}}$ (across all industries) are tabulated:

	1 Day Lag	2 Days Lag	3 Days Lag	4 Days Lag	5 Days Lag
LASSO	-0.0071	-0.0051	-0.0055	-0.0066	-0.0060
Elastic Net	-0.0090	-0.0061	-0.0063	-0.0088	-0.0084
Random Forest Regression	-0.1071	-0.1051	-0.0918	-0.1005	-0.0986
Gaussian Process Regression	**0.0137**	**0.0135**	**0.0138**	**0.0138**	**0.0136**
Support Vector Regression	-0.0210	-0.0189	-0.0128	-0.0110	-0.0114

Table 2 Average R^2_{OOS} for the Regression Models (with different lags), Benchmark: Historical Mean

As we can see, only Gaussian Process (GP) regression performs better than historical mean. The best lag is the one with 4 days of lagged returns, with average $\boldsymbol{R^2_{OOS}} = 0.0138$. This is economically very significant since the ratio $R^2_{OOS}\big/ S^2 = \frac{0.0138}{0.0318^2} = 1364\%$ better performance than without prediction!

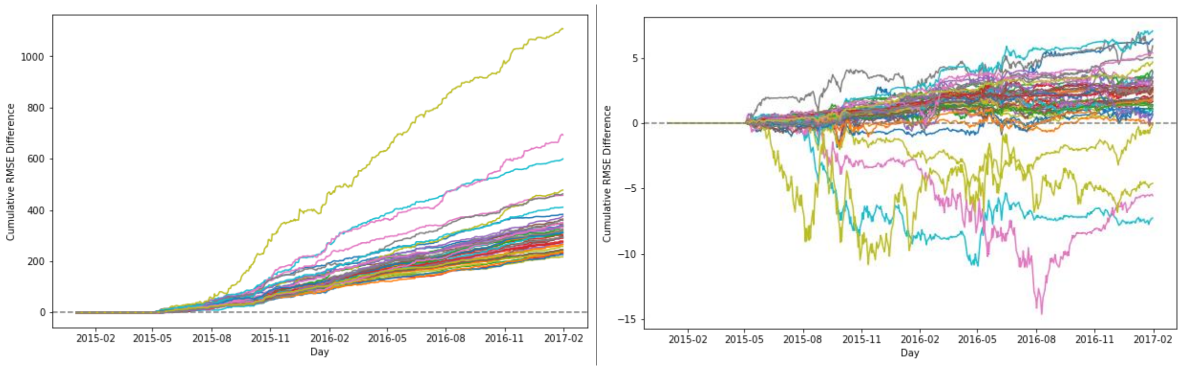

Figure 2 Cumulative RMSE Difference (Gaussian Process Regression - 4 Days Lag), Left: OLS Benchmark, Right: Historical Mean Benchmark

From the right graph, we can also see that GP regression has higher forecast accuracy than historical mean in most of the industries. The only four industries where it underperforms are sectors heavily involved in commodities: **Gold**, **Coal**, **Steel** and **FabPr** (Jupyter notebook has multiple subplots for the cumulative RMSE difference in each industry). As discussed before, investors treat these sectors as a proxy to invest in commodities. For example, gold is treated as safe haven asset which has little correlation with equities market. Therefore, it is expected that we will not get good forecasting accuracy by using the lagged returns in equity markets alone. In the Extra Section of Jupyter notebook, it is shown that the additional daily return data of Gold and Silver commodity prices do not help to improve the predictive power of the GP Regression model.

From Table 2, we can also observe that 4-Day Lag GP Regression achieves slightly higher R^2_{OOS} than 1-Day Lag GP Regression. This difference is also economically significant since it means $\frac{0.0001}{0.0318^2} = 10\%$ better performance than without prediction. Hence, we conclude than including more lagged returns helps to improve the forecasting performance slightly. However, it does not help to add too many lagged returns since the 5-Day Lag GP Regression model actually performs worse.

Prior Feature Selection using PCA

Prior feature selection may help to improve the performance of supervised learning in some occasions. In the Jupyter notebook (Extra Section), Principal Component Analysis (PCA) is used as feature selection before GP Regression. It is shown that PCA does not improve model performance in this case.

Different Lag Structures in Each Industry

The bar chart below shows the number of industries which achieves highest R^2_{OOS} while using LASSO model with different lags.

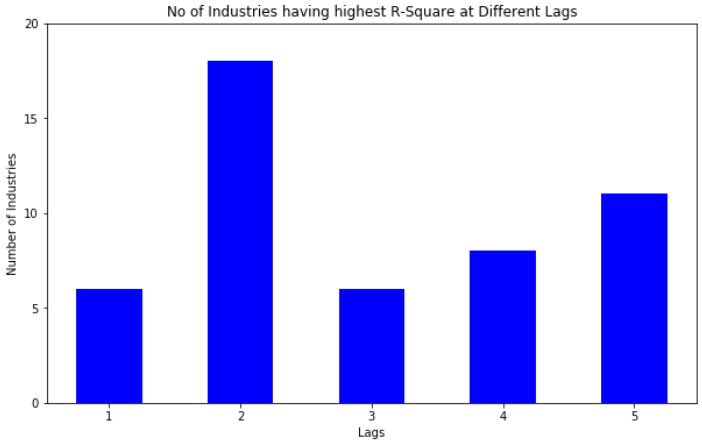

Figure 3 Number of Industries with Highest R^2_{OOS} using LASSO Model with Different Lags

It can be seen that the industries have different lag structures, and most of the industries perform best using 2 days of lagged returns. Six of industries perform best using only the previous day of industry return for forecasting. One of the six is **Oil** industry. Oil industry is mainly made up of huge oil majors and they tend to lead other small-capitalisation stocks from other industries (Lo & Mackinlay, 1990). Consequently, using more lagged returns to forecast their returns will not be useful, thus one lagged day forecasting is the best for this group of industries.

On the other hand, there are around ten industries which are best forecast using all 5 days of lagged returns, and one of them is **Retail** industry. Retail industry consists of many smaller retailers and the stock returns of these small companies can be well predicted by using the lagged returns of big companies (Lo & Mackinlay, 1990). Due to this lagging effect, the stock returns for this industry are best predicted using the lagged returns of big companies from many other industries.

Nonlinear Methods

The industry return data provided is in nonlinear representation (i.e. percentage change). Two nonlinear machine learning and regression techniques (Random Forest Regression, Gaussian Process Regression, and Support Vector Regression) have also been attempted to investigate if there is any nonlinear prediction patterns between the industry returns.

From Table 2, it can be seen that Random Forest and SVR perform badly while Gaussian Process Regression is the only method that outperforms Historical Mean benchmark. Therefore, there is some nonlinear predictability in the industry stock returns and this is aptly captured by the probabilistic model used in Gaussian Process Regression. One explanation of the poor performance of random forest is that it might not have rectified the overfitting tendencies of regression trees.

Frequently Used Industries as Forecasters

Table 3 displays the top three most frequently used industries as forecasters by the five different LASSO models. The most striking feature is that **Drugs** and **Hlth** industries always show up as useful forecasters.

	First	Second	Third
LASSO (1-Day Lag)	Drugs (Pharmaceutical Products)	Toys (Recreation)	Clths (Apparel)
LASSO (2-Days Lag)	Hlth (Healthcare Services)	Drugs (Pharmaceutical Products)	Toys (Recreation)
LASSO (3-Days Lag)	Hlth (Healthcare Services)	Drugs (Pharmaceutical Products)	FabPr (Fabricated Products)
LASSO (4-Days Lag)	Hlth (Healthcare Services)	Drugs (Pharmaceutical Products)	Clths (Apparel)
LASSO (5-Days Lag)	Hlth (Healthcare Services)	Drugs (Pharmaceutical Products)	FabPr (Fabricated Products)

Table 3 Top Three Most Frequently Used Forecasters in LASSO Models (with different sets of lags)

This agrees with the findings in Chinco et al (2017). In their paper, they demonstrated that LASSO tends to select the stocks with future news announcements. There is a lot of financial news related to pharmaceutical and healthcare companies (for example, certain drugs get approved, healthcare policy changes, certain biotechnology breakthrough, etc). It is no surprise that LASSO frequently uses the lagged returns of these two industries for prediction (since LASSO anticipates news announcements). Furthermore, these two sectors tend to have huge companies. Due to the lead-lag effects, these industries are naturally used as forecasters for other industries made up of many small companies.

Conclusion

In this assignment, a few predictive models have been built to attempt predicting future industry stock returns by using the lagged industry returns. The only model that consistently beat historical mean benchmark is Gaussian Process Regression (a nonparametric nonlinear regression method).

The underlying economic forces that result in this predictability are likely the investors' behaviours. As pointed out in Chinco et al (2017), LASSO is likely to select industries which are going to have news announcement. In other words, these companies would have certain pattern in their stock returns. This might be attributable to the investors who have access to certain information that is not yet publicly disclosed. Another factor discussed is the leading effects of large-capitalisation companies. This could be due to the fact that large companies receive greater investor interests and their price change will be more immediate than small companies.

References

Chinco, A., Clark-Joseph, A., and M. Ye, (2017), Sparse Signals in the Cross-Section of Returns, *Working paper*. Available from http://www.alexchinco.com/sparse-signals-in-cross-section.pdf

Lo, A., and C. Mackinlay (1990), When are Contrarian Profits Due to Stock Market Overreaction? *Review of Financial Studies*. Available from http://www.jstor.org/stable/2962020

Moskowitz, T., and M. Grinblatt, (1999), Do Industries Explain Momentum? *Journal of Finance.* Available from http://onlinelibrary.wiley.com/doi/10.1111/0022-1082.00146/full

French, K., (2017), *Detail for 49 Industry Portfolios*. Available from
http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_49_ind_port.html
[Accessed 18 April 2017]