Summary-Cubist Project

Minda Yang
minda.yang@columbia.edu

Notice: Please refer to README.txt for the codes and dnn models trained.

1. Data cleaning:

Logics:

Here the logic for data cleaning is to assume a distribution of the change in the signal and spy index. The reasoning is that the difference between two consecutive days shouldn't deviate too much from the distribution of historical difference.

Since the data is likely non-stationary, I look at the previous 20 days and calculate the mean and standard deviation of the difference between 2 consecutive days. Then if the new sample is out of the 10σ range (10 times the standard deviation), the new data point will be considered as noisy and replaced with the previous day's value.

Notice:

This method assumes the distribution of the difference to be Gaussian, which is not necessarily true, the usage of 10σ as a large margin here is an attempt to deal with the 'fat tail' effect.

Future:

Add a method to predict the real value of noisy data point, possibly LPC (linear predictive coding).

Results:

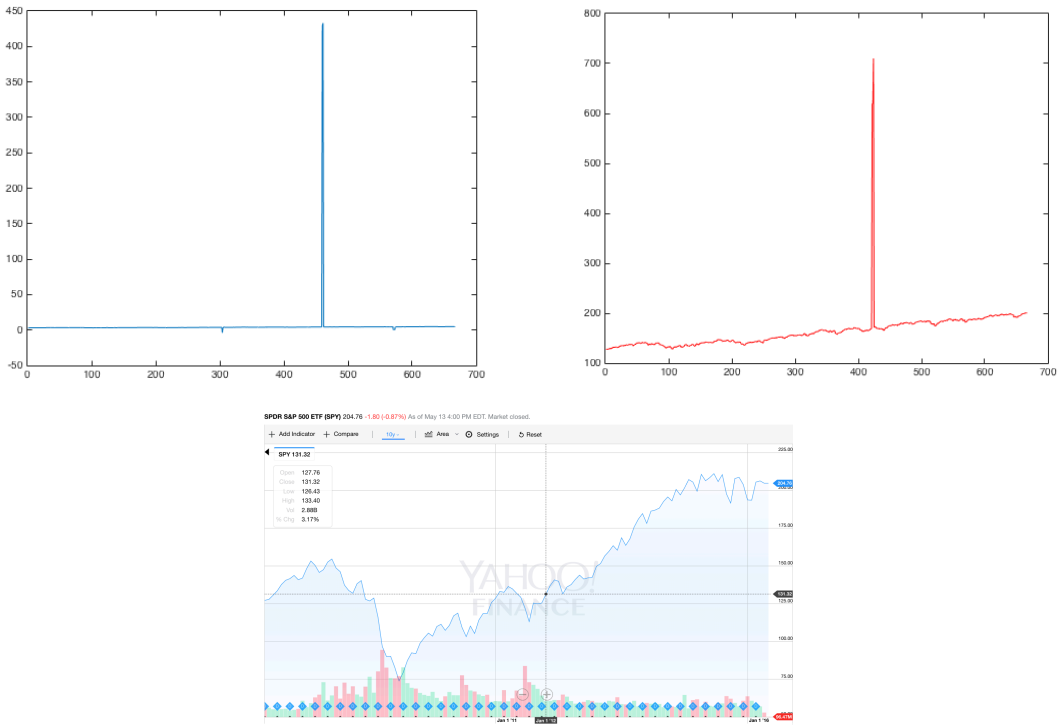


Figure 1. Before data cleaning. We can identify the sudden overshoot of the signal and

index. One good reference is to compare with the SPY from Yahoo finance.

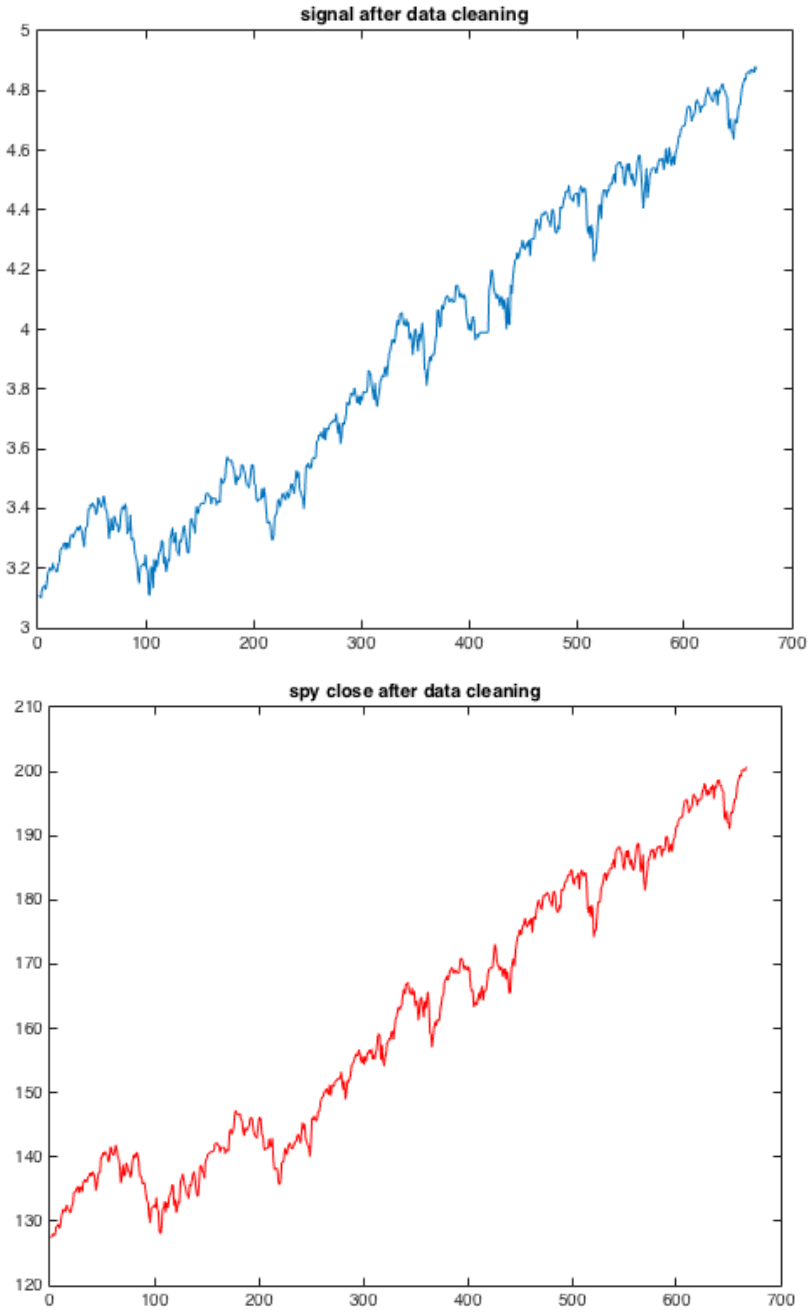


Figure 2. After data cleaning. Notice that the correlation between signal and SPY is higher than 0.99. This is also verified by the similarity of two curves.

The index identified as noisy in the signal:

304 460 461 571 572 573

The index identified as noisy in the SPY:

422 423 424

2. Linear regression model

Logics:

We receive both data at the same time, so it only makes sense to use today's signal (and signals before today) to predict future's SPY return. As a way to normalize the signal, we use the return of signals to predict return of future SPY index.

To assess whether the signal has predictive power, here I compare the predictive power of using signal vs. predictive power using spy index. This is also because the signal is highly correlated with SPY index (>0.99 correlation)

For example, to assess whether signal(1:20) can predict spy(21) better than spy(1:20). Further, whether combining signal(1:20) and spy(1:20) performs better than

Notice:

To predict return, we need to calculate the difference between today's close vs. tomorrows close.

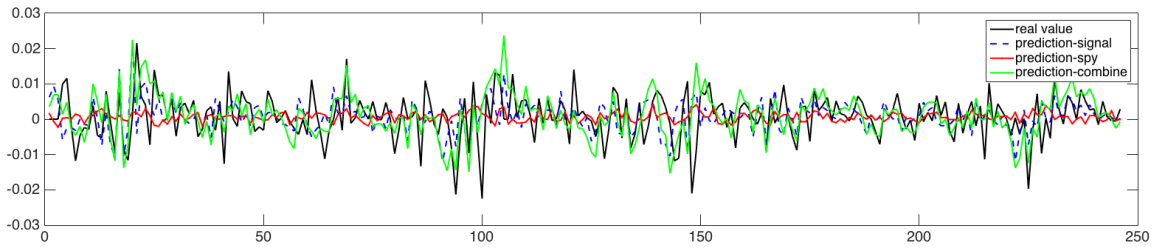
To fairly compare two approaches (using signal vs. using spy), we need to train a model in the training set and compare the prediction in the test set.

To fully take advantage of the data, we can do this in a cross-validation manner. But in this demo I'll specify the training set and testing set.

Future:

Possibly linear regression with some form of regularization (Lasso, L2).

Results:




```
Mean squared error (using signal):
   5.3739e-05

Mean squared error (using spy):
   4.2488e-05

Mean squared error (combining signal and spy):
   4.7148e-05
```

Figure 3. The real value, prediction using signal, prediction using spy and the prediction using signal and spy. The mean squared error shows that signal doesn't add value to the prediction of future SPY return.

By including the signal into the spy, we didn't better predict the SPY return based on the MSE criteria. Also, using spy alone better predicts the SPY return than using signal alone.

3. Deep learning models.

Logics:

To further assess whether the signal adds value to the prediction of future SPY return, I also implemented a deep learning model to train on the SPY index and the combination of signal and SPY index. If the combination of signal and SPY index does not outperform using the SPY index alone, then it verifies that the signal doesn't add value.

Notice:

As the way I evaluate the performance is to look at the mean squared error of the real return and predicted return, this deep learning model is designed for regression problem. To do so, I replaced the usually used softmax output layer with a hyperbolic tangent (tanh) layer so the output falls in the range of (-1, 1). To fully use this output range, I multiplied the return with a coefficient as the return is mostly a small number.

To prevent the model from overtraining I used a validation set to early stop the training of the deep learning models.

Future:

The current model is still a small scaled given the input feature size and the total amount of data. This approach could be easily applied to the minute bar data.

Results:

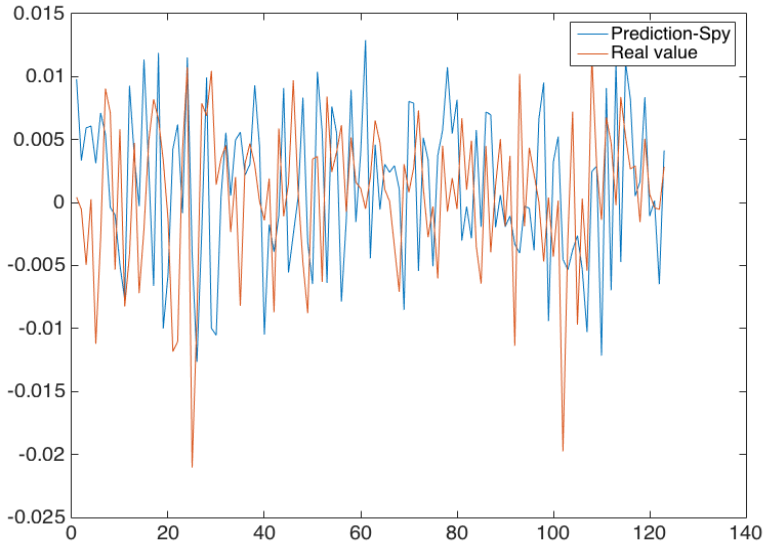


Figure 4.1 Prediction of future SPY return using past SPY return.

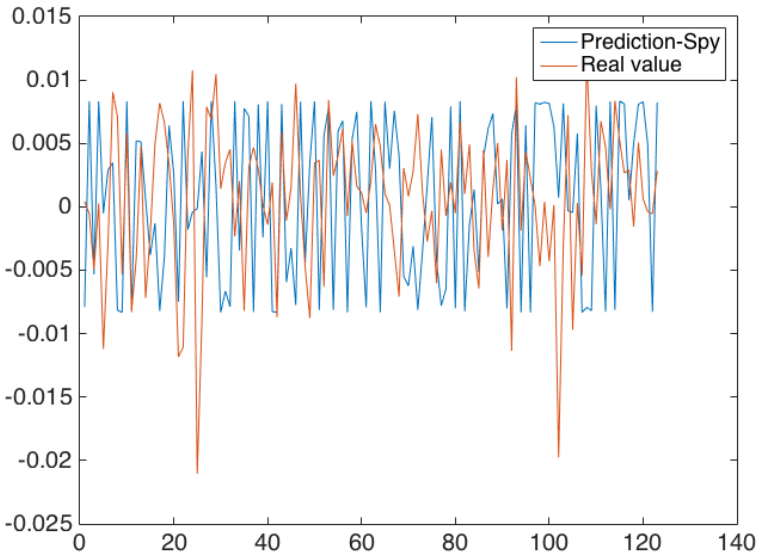


Figure 4.2 Prediction of future SPY return using past SPY return and signal return.

	Using SPY	Using SPY & signal
MSE	7.23e-05	7.25e-05

The MSE for using two approaches also reveals that by using signal we couldn't better predict the future SPY index return.

4. Conclusion

After data cleaning, the signal and the SPY index show a pretty high correlation (>0.99). This leads to the test of whether the signal adds value to the prediction of the future SPY return. I compared two models (linear regression and deep learning) on using the past SPY return and combining the past SPY return and signal return.

The results show that the signal doesn't add value to the prediction of the SPY return.