Statistical Arbitrage
Carry Trade in FX Market

Justin Alick, Jingyi Guo, Pranav Hanagal, Shenchao Qiu, Jingxin Xi

December 16, 2017

Contents

1 Executive Summary

The global FX market is consequentially the largest market in the financial realm, with belief that it is less an efficient market than equity market, leaving the chance for statistical arbitrage.

In this project, we propose a modification on the popular carry trade in currency markets that takes advantage of the empirical failure of Uncovered Interest Rate Parity (UIP). If we assume exchange rates as units of domestic currency per unit of foreign currency then the empirical failure of UIP suggests longing currencies that trade at a forward discount and shorting currencies that trade at a forward premium, in the forward market. Profits can be potentially made by taking the opposite position in the spot exchange market in the following time period.

We look beyond the simple simplistic decision making process that only looks at the forward premium, and developed a ensemble model-based statistical arbitrage strategy. Due to the limited universe in FX market in comparison with the equity market, the number of observation at each formation period is seriously small. Under this constraint we designed a scheme of building up model through a series of feature extraction, feature selection, model selection and ensemble learning, so that we reduced the dimension of the raw feature, and aggregate the prediction power of several weak models to balance carefully between accuracy and variance, which works well in such limited sample size.

We used FX data from Bloomberg dating from as early as 1970 up till 2017 as our source data, and segmented our training period as pre-crisis (1970-2008) and validation period as post-crisis(2009-2017). We also validate the horizon of our formation period as hyperparameter that is influential to the strategy performance.

As a result, we found that our strategy in profitable in general, and works better in our validation period than the pre-crisis training period. 3 month formation period will give better out-sample performance, but it also has a larger difference between its in-sample and outsample performance.

To testify if our strategy is qualified as statistical arbitrage, we performed hypothesis test following JTTW paper. Unfortunately we did not pass all the sub-tests and our strategy can't be qualified as statistical arbitrage under current assumption.

The transaction cost in the FX market is relatively considered to be a dark data point. Bid and offer data is not readily available to the general public. However, we can look to contract specifications such as those from the CME to identify the "minimum tick fluctuation." Furthermore, we can extrapolate for higher spreads associated with pairs with less liquidity. Finally, it is critical to consider the implications of the market microstructure in FX Markets. Particularly that known as "Last Look:" option market-makers enjoy reneging on trades within 50 milliseconds of filling them.

For future improvement we are looking to gather more relevant data and revise the model building with a larger set of model, given more computational power.

2 Introduction

2.1 FX Market

The global FX market is consequentially the largest market in the financial realm, it's daily volume has been quoted at estimates of \$5.1 Trillion by Nasdaq[1]. The majority of this volume is concentrated in trading among the obvious pairs such as EUR/USD and GBP/USD. This massive daily volume is the result of a combination between leveraged trading and small tick sizes that the exchanges endorse and liquidity takers enjoy.

2.2 Carry Trade

The carry trade is motivated by two well known theoretical conditions related to FX Markets: Covered Interest Parity (CIP), and Uncovered Interest Parity (UIP). CIP essentially implies that money cannot be made by longing a currency in the forward market and simultaneously borrowing the currency from abroad and investing it domestically.

Formally, the following approximation holds:

$$r_t - r_t^i = f_t^i - s_t^i$$

where r_t^i is the log foreign risk free rate at time t, r_t is the log domestic risk free rate at time t, f_t^i is the log forward exchange rate at time t, which is expressed as units of domestic currency per unit of foreign currency, and s_t^i is the log spot foreign exchange rate at time t.

UIP asserts that on average, money cannot be made by borrowing in the country with the lower risk free rate and lending in the country with the higher risk free rate. Exchange rate movements should on average erase any potential profit. Formally, the following roughly holds in theory:

$$r_t - r_t^i = s_{t+1}^i - s_t^i$$

where r_t, s_t^i are defined as above. This suggests that if one country has a higher risk free rate than the other, then the currency of this country should depreciate. However, it is well-documented that this does not necessarily hold. The Forward Premium Puzzle is an empirical finding that the currency should actually appreciate under these conditions. If we combine the two equations, which should hold in theory, we find that the change in exchange rates should equal the forward premium.

We can rewrite the previous claim as:

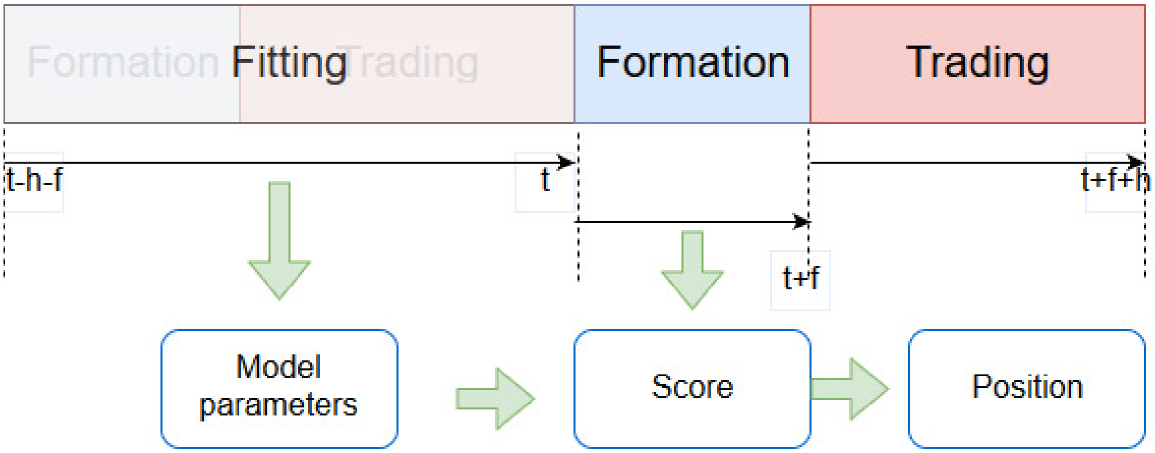
$$s_{t+1}^i - s_t^i - (f_t^i - s_t^i) = 0.$$

However, due to the Forward Premium Puzzle, this equality will not hold in practice and suggests the following popular strategy: Long currencies in the forward market with large forward discounts and short currencies in the forward market with large forward premiums. Close out the position in the next time period by taking the opposite position in the spot exchange rate market. The Forward Premium Puzzle says that this strategy should produce a profit since we should tend to make money off both the interest rate spread and the exchange rate movement. However, slippage due to "Last Look" in FX markets and unexpected changes in exchange rates do pose a risk to our profits.

3 Strategy

3.1 Trade implementation

We follow the workflow demonstrated below:



At each step, we set up non-overlapping fitting periods and out of sample formation-trading periods. The formation and trading period also do not overlap, which guarantees we don't fall into the look ahead pitfall in this strategy.

In the fitting phase, we assume we take in the formation data and try to predict the look-ahead true log-return of the carry trade. And with the fitted parameter, we go forward to score the out-sample formation data, based on which we make trading decision that take effect in the following month.

3.2 Statistical Modeling

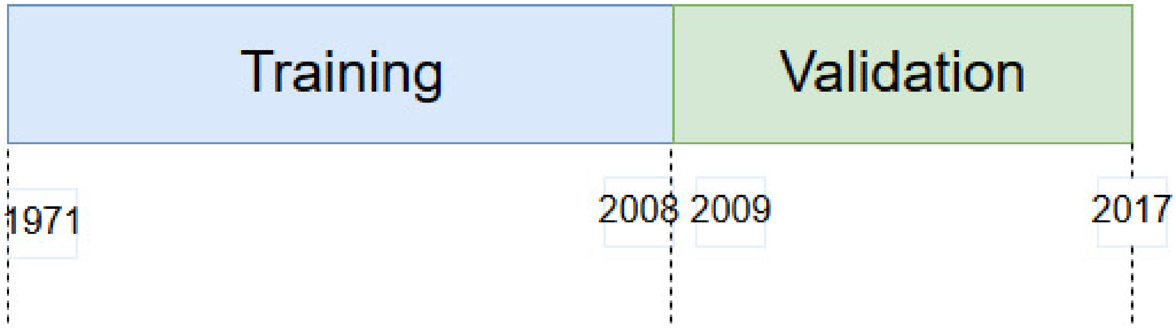
Essentially at each fitting period we wish to construc a model that accurately predict what the return will be over the next trading period via carry trade, and the raw feature input we have at each formation period is just the spot and forward exchange rate time series within this period.

$$s_{t+1} - f_t = \Psi(\{s_u : u = t, t-1, ..., t-F\}, \{f_u : u = t, t-1, ..., t-F\})$$

And we will enter long forward position for currencies with high predicted return and short forward position otherwise.

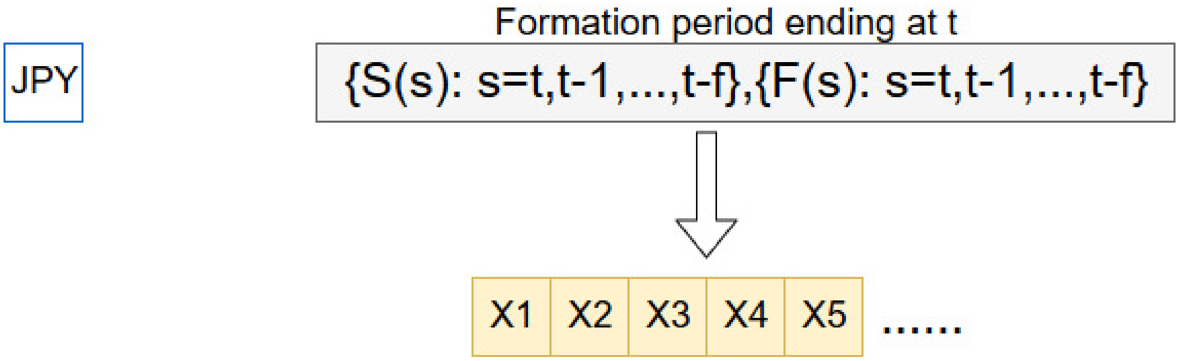
Note that the universe of currencies is very limited, which is a pain for model fitting and portfolio construction.

We construct the model with a segmentation of training and validation among our data. The rationale of taking such a segmentation is that we wish the strategy to be able to generalize well under future market conditions. This means that we need to avoid overfitting by picking the set of hyperparameters that controls model's robustness. And the goodness of each set of hyperparameters is evaluated by comparing their performance on the training set and validation set.



3.2.1 Feature extraction

As stated above, the raw features at the fitting period are times series of currency spot exchange rate, and forward exchange rate. Given the limited response number, directly using the raw features as fitting input is infeasible, therefore feature extraction is crucially important in our case.


Formation period ending at t
JPY
{S(s): s=t,t-1,...,t-f},{F(s): s=t,t-1,...,t-f}
X1
X2
X3
X4
X5
......

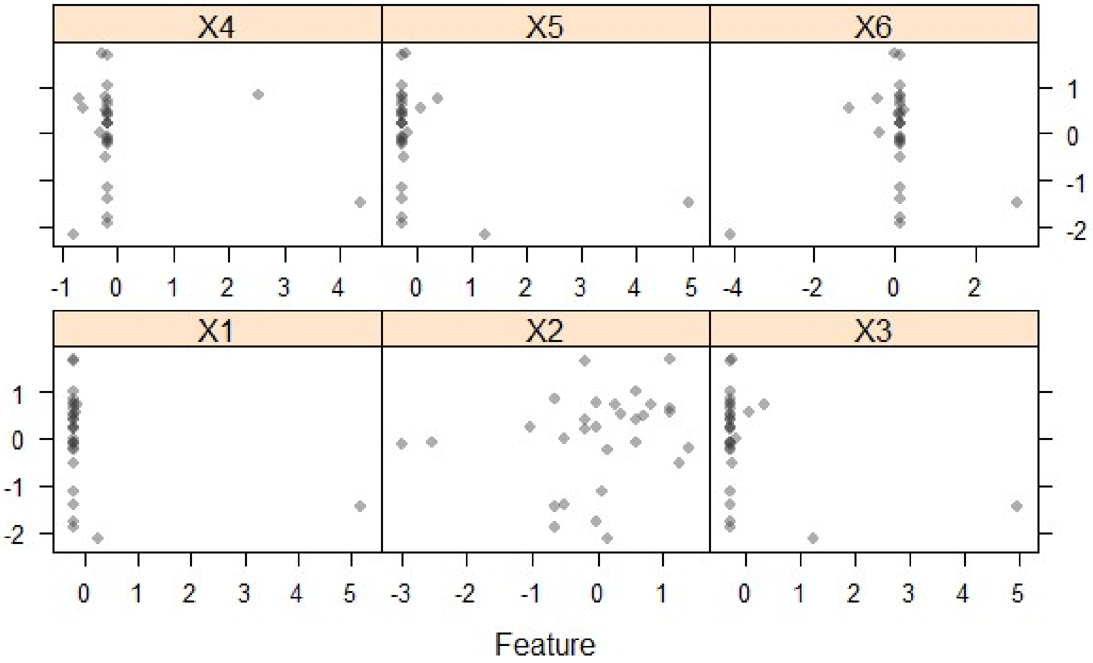
Therefore we applied some feature extraction methods to compute features out of those time series. From simplest ones such as aggregated momentum, deviation from mean, to more sophisticated ones such as entropy and information/noise ratio.

3.2.2 Feature Selection

After constructing the features, we still need to further select features that are relevant in return prediction. Intuitively, the feature extraction is completely unsupervised to represent the information included in those time series. Yet we need to select the features that can be good predictors.

Given the necessity of feature selection, the question is now which feature selection techniques to use. Many models have built in, or implicit feature selection mechanisms, such as random forests and lasso regression. However these are model-dependent, that it still requires enough data to infuse the intelligence. So here we manually selected features by getting some human judgment of the visualization result.

The plot below shows the relationship between some extracted features and the look-ahead true return. Apparently we should select the feature X2, that itself has a proper amount of variance, and there is potential interaction we can exploit in later model fitting.



Following this selection criteria, we finally selected features that are both representative of the original time series, and are strongly relevant.

3.2.3 Ensemble Learning

Even with the selected feature, our limited number of observations at each formation period remains to be a concern. On this limited playground we are not able to use models with high complexity.

Recall Sauer' Theorem:

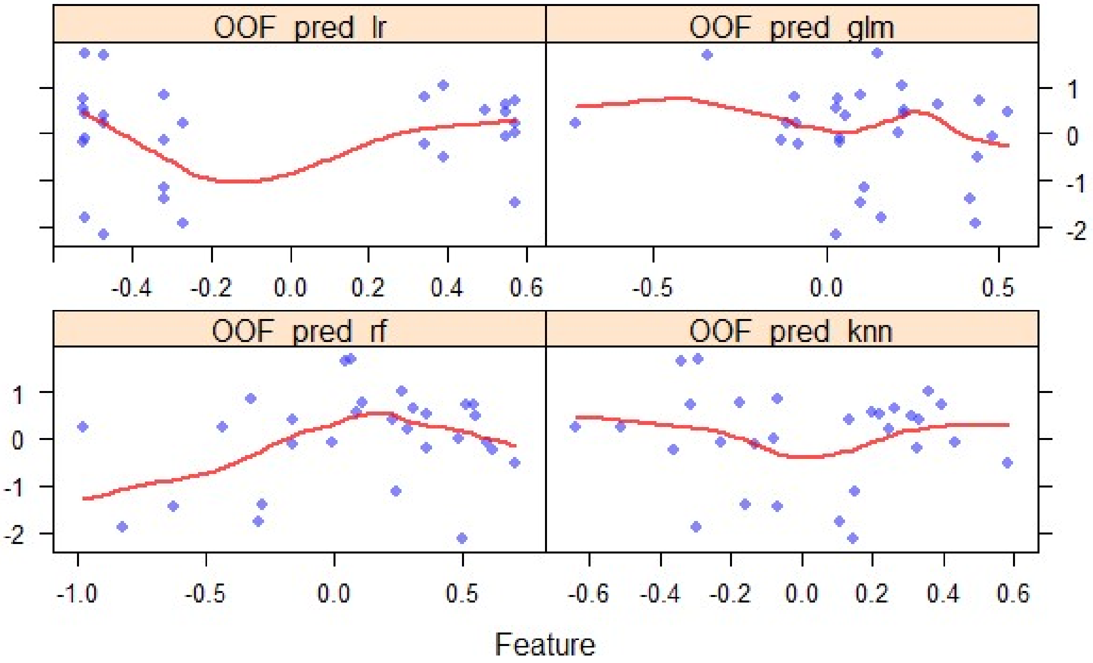
$$N = O(\frac{1}{\epsilon}[VCdim(H)log(\frac{1}{\epsilon}) + log(\frac{1}{\delta})])$$

Indeed, according to statistical learning theory, the higher the model complexity, the more data we need to sufficiently train the model to give any reasonable guess on the response.

As a result, the complex, high accuracy models has to be excluded from our scope, such as the entire artificial neural network family.

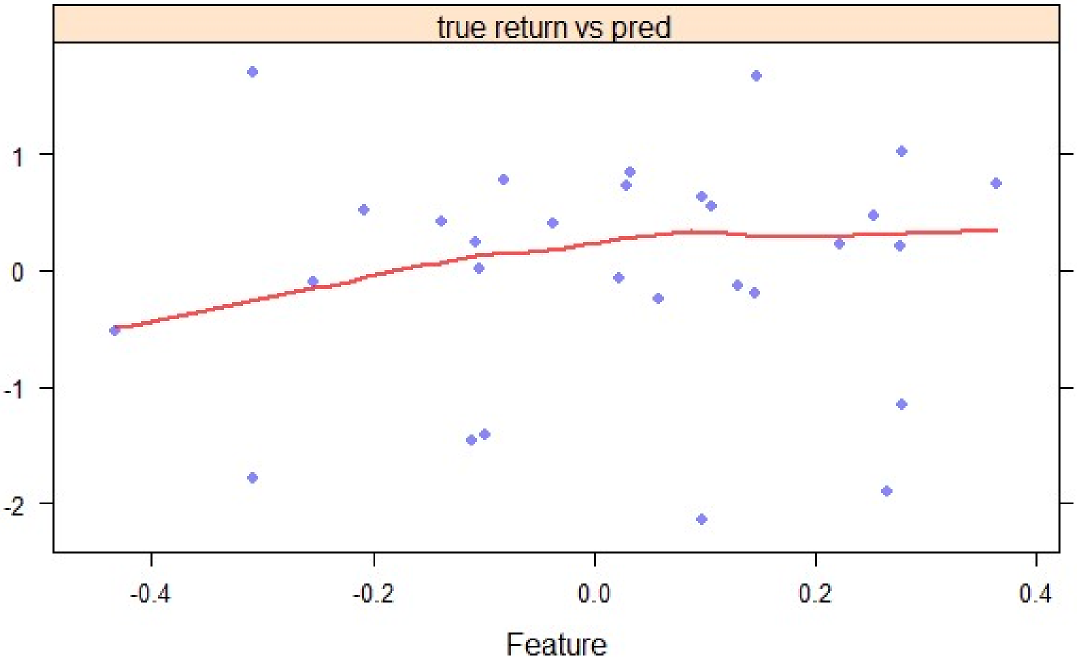
Therefore we focus our model selection under the criteria of picking robust, stable, low variance models, which on the other hand may result in sacrificing accuracy. To jump out of this dilemma, we apply ensemble learning to aggregate the prediction power of many weak predictors.

We picked 5 weak machines: lasso regression, knn, Gaussian process with radial kernel, linear regression and a shallow random forest. All these models are first trained independently to produce a first layer of prediction result:

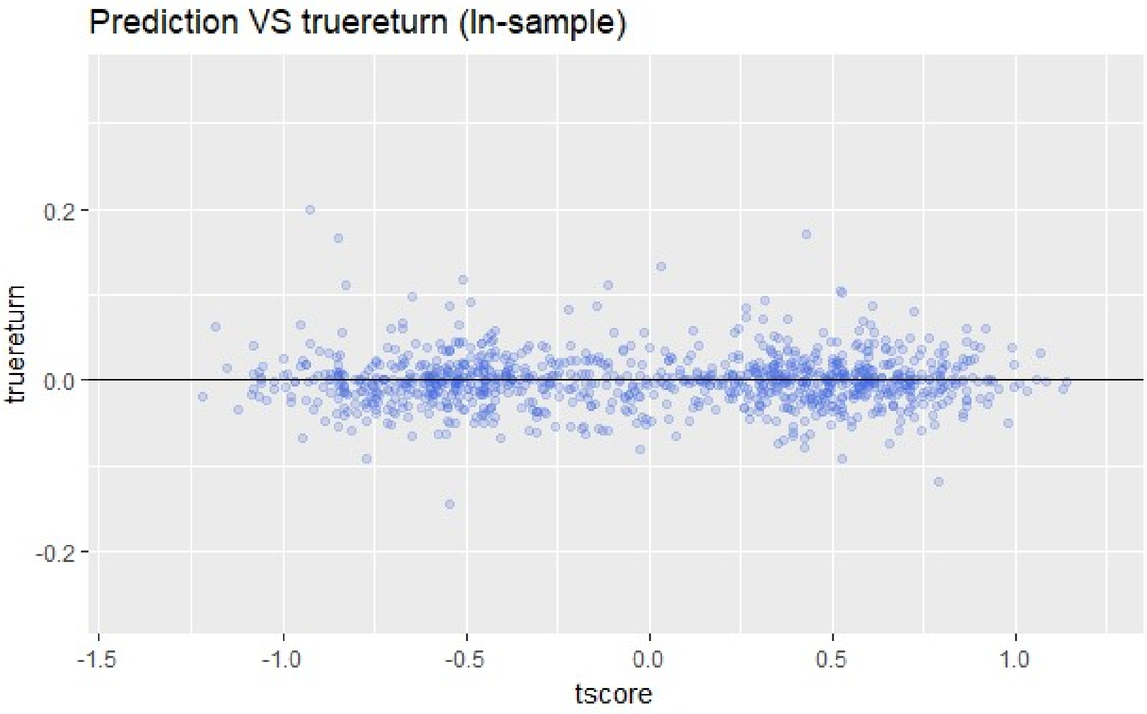


Observe from the prediction vs true return plots, for each weak model, the result trends differ, further verifying the fact that each weak model is not very accurate, and that they each capture different aspects of the interaction between the predictors and the response, which makes ensemble methods a sensible next step.

We trained a top-layer random forest that takes the output of the 5 weak models and again predicted the true return, and it produced the following graph(for just random one formation period):



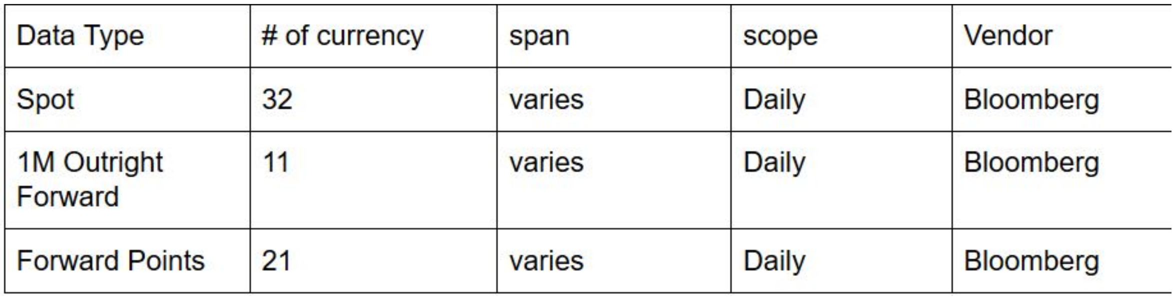
And as we move 1 period forward, there we fill fit a new set of parameters for the new formation period, and make new prediction. Along the way the in-sample prediction and the true return plot is as follows:



4 Data

4.1 Description

We retrieved forward and spot exchange rate data of 32 countries from Bloomberg terminal, dating from January 1971 to December 2017. Due to availability of forward contract we only retrieved 1 month forward exchange rate and used 1 month as our trading horizon.



Data Type	# of currency	span	scope	Vendor
Spot	32	varies	Daily	Bloomberg
1M Outright Forward	11	varies	Daily	Bloomberg
Forward Points	21	varies	Daily	Bloomberg

Note that FX spot and outright forward are cross rate denoted by foreign currency per US dollar. Forward points are the number of base points (bps) added to or subtracted from the current spot of a currency to determine the forward rate. And all forward rate are 1 month forward rate.

4.2 Preprocessing & Cleaning

- Missing values in formation period: First we use dimension reduction method to acquire an feature matrix (predictor) and response vector. If the length of response vector is small than 10, we don't trade in this period and move on next month. Otherwise, we use "imputer" to insert missing data with mean value in the feature matrix but no change in response vector. The model will return a vector with the same size as the response vector.
- Missing values in trading period: if the score give signals for opening positions but have missing data, we treat it as follows: for spot missing data, we fill with the previous value; for the forward missing data, we fill with the next available data in the current month.

4.3 Segmentation for Train and Validation

- Train Period: 1971/01/01 to 12/31/2008 (Due to long missing value series, the effective train period starts from 2000)
- Validation Period: 1/1/2009 to 12/08/2017

5 Result

5.1 In Sample Prediction

At each fitting period, we not only obtain model parameters that we use out-sample in the imminent formation-trading period, but also gain for free an insample score for the current fitting period. We can then compute the post-hoc profit, in comparison with the out-sample profit in the next part, show how bad the model overfitting is.

We tested our results with different formation horizon(F=1,3), since we only have complete forward data for 1-month contracts, the trading horizon is fixed at 1 month.

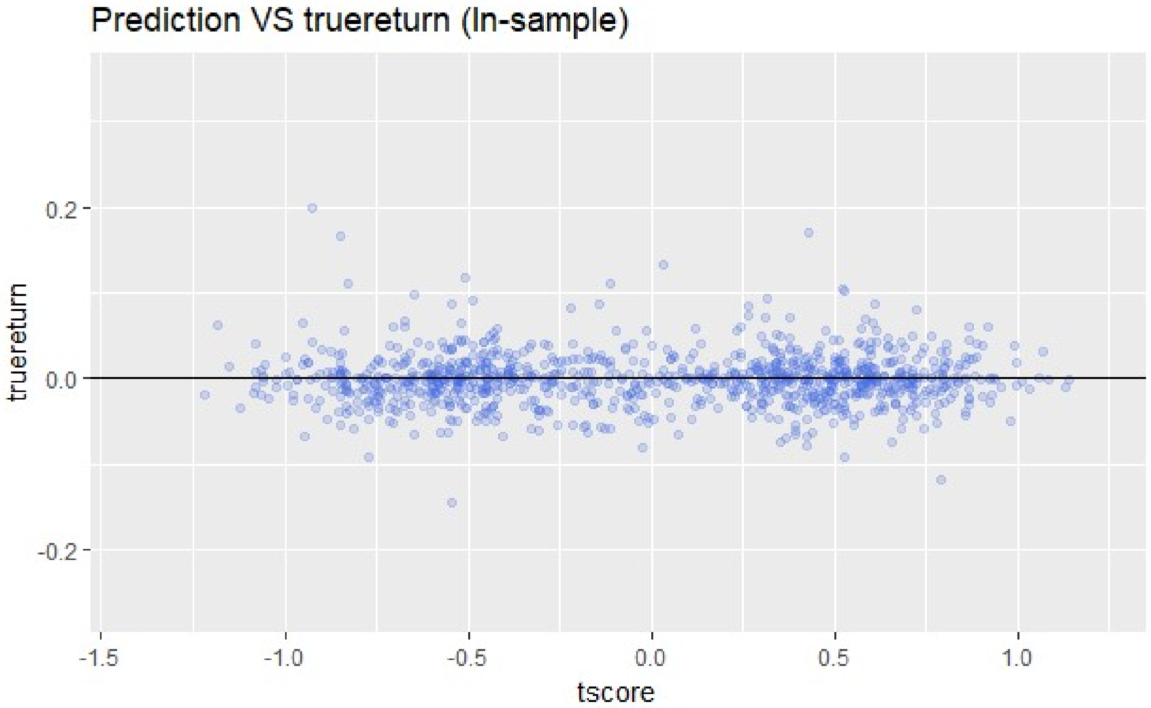


Figure 1: In Sample Prediction VS True return

Revisit this plot in part 3, the points with higher score is slightly denser where above the zero line.

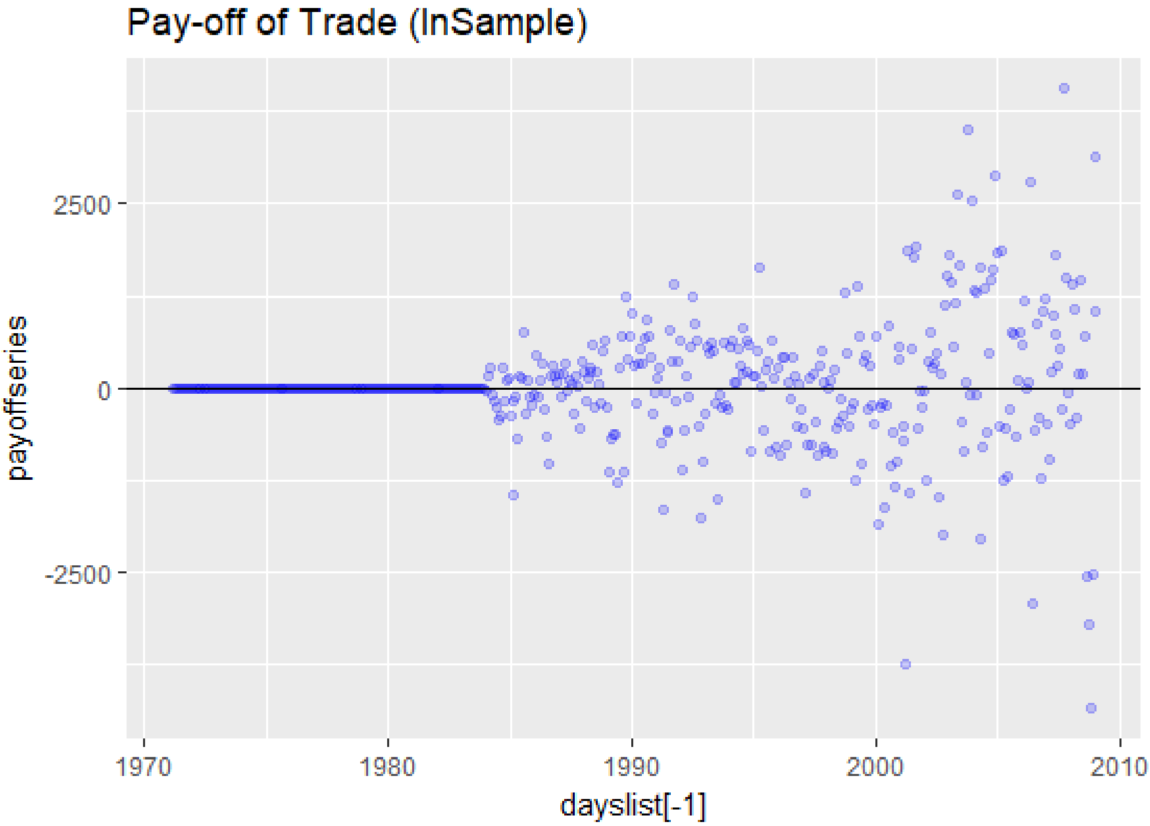


Figure 2: In Sample Trade Profit

And as for the trade, there are slightly more trades ended up with positive payoff. This difference might be visually subtle, but is we look further into the discounted wealth:

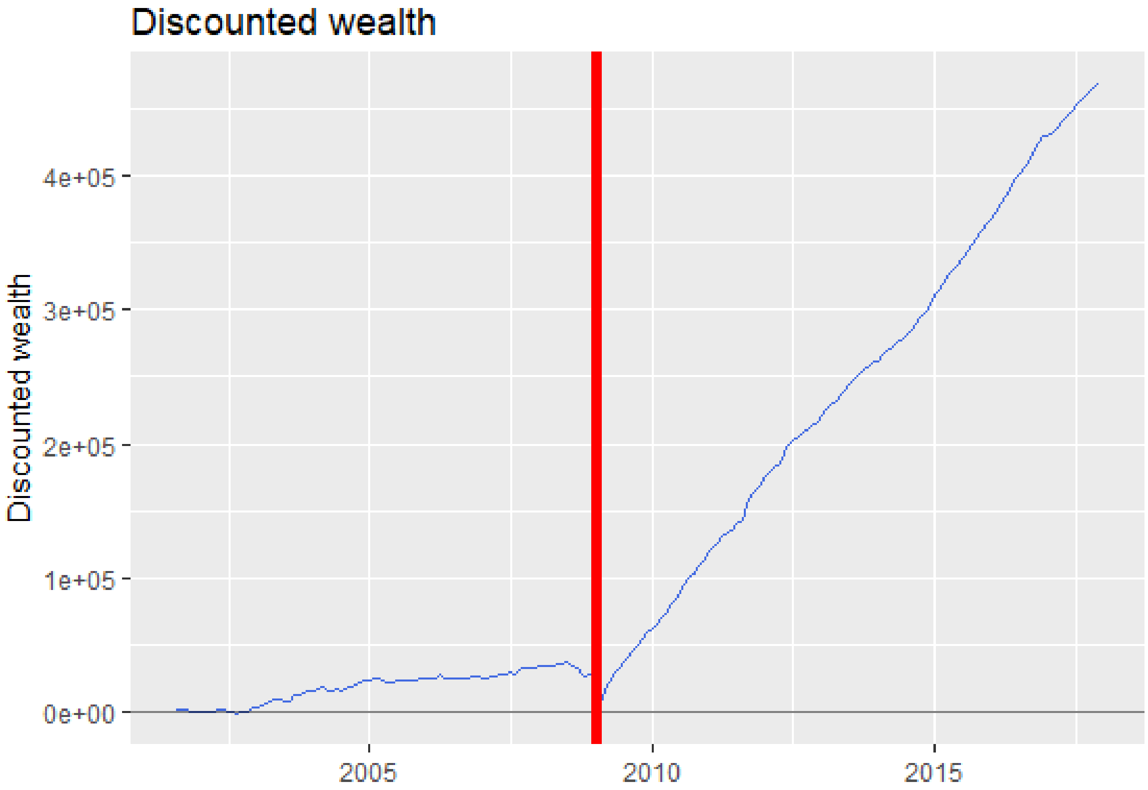


Figure 3: In Sample Discounted Wealth

Note that the red vertical line indicated the segmentation of our training period (pre-crisis and pre-EURO)and validation period (post-crisis and post-EURO).

We observe a crazily upward slope, which indicates two facts:

- Our model overfits in-sample.
- Our model works better under market conditions after 2008.

We also inspect the in-sample profitability of our model with 3 months formation period:

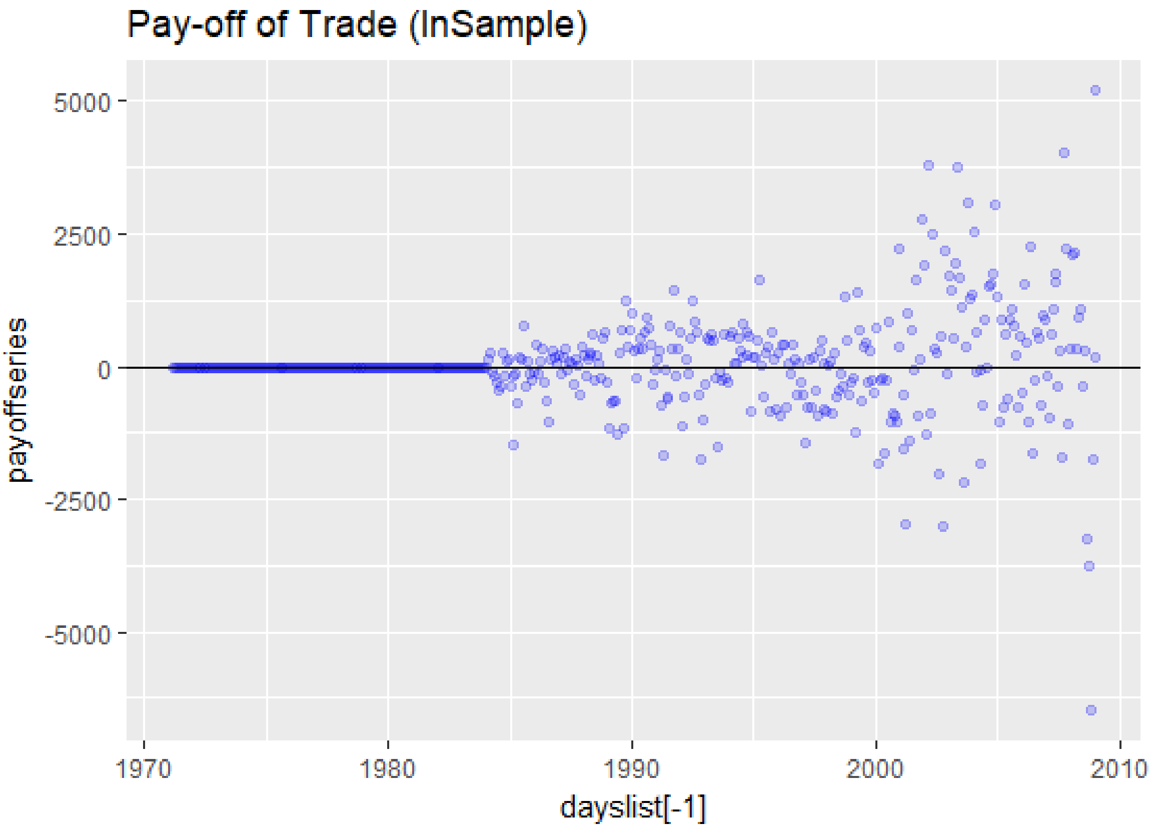


Figure 4: In Sample Trade Profit

discount.png

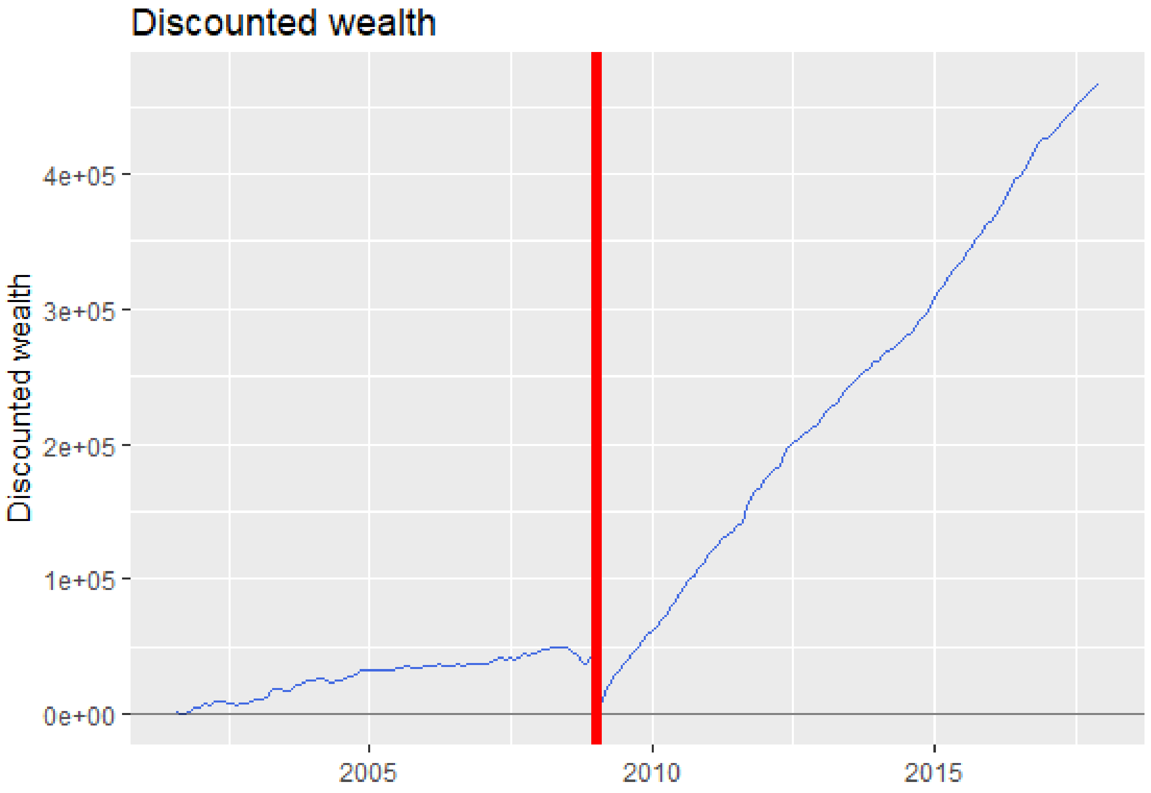


Figure 5: In Sample Discounted Wealth

It also show similar pattern of overfitting in sample data, and working better under market condition after 2008.

5.2 Out Sample Prediction

Out-Sample prediction is what really measures the profitability of the strategy. Having observed the overfitting pattern in the previous section. We are very concern about the actual performance of our model.

Again we tested the relationship between our out-sample prediction score and the true return:

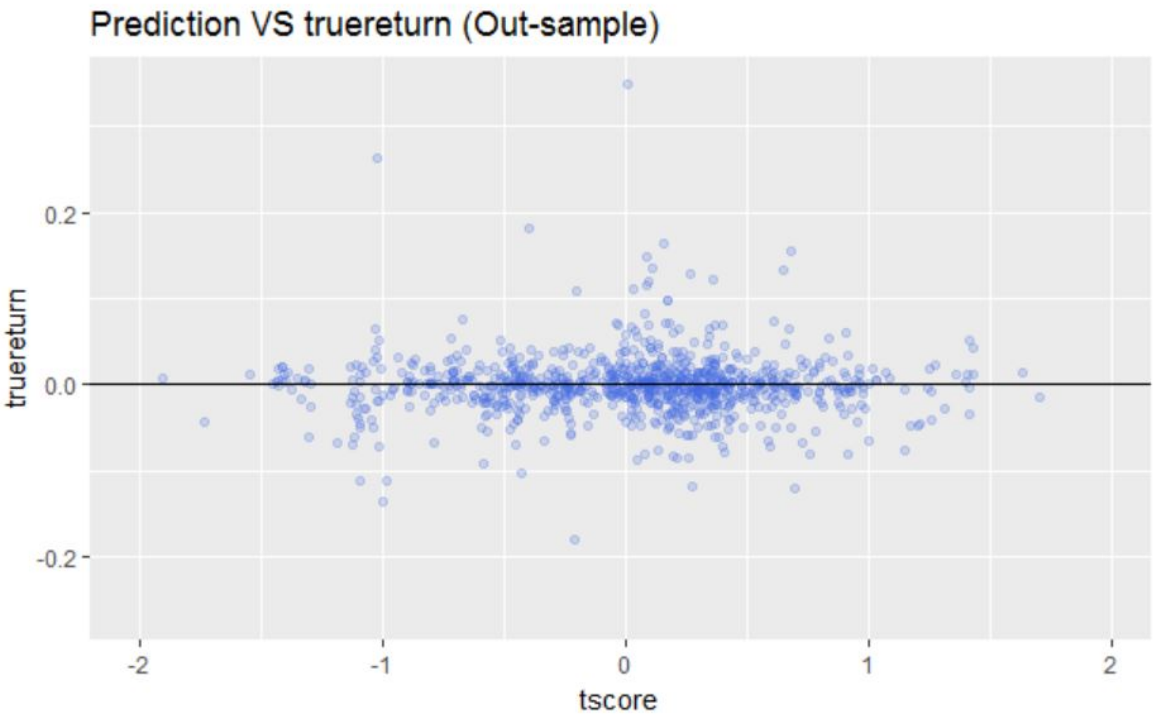


Figure 6: Out Sample Prediction VS True return

It shows similar pattern, but the portion with a higher prediction score sank lower to the zero line. And then we look at the out-sample trade payoff out.png

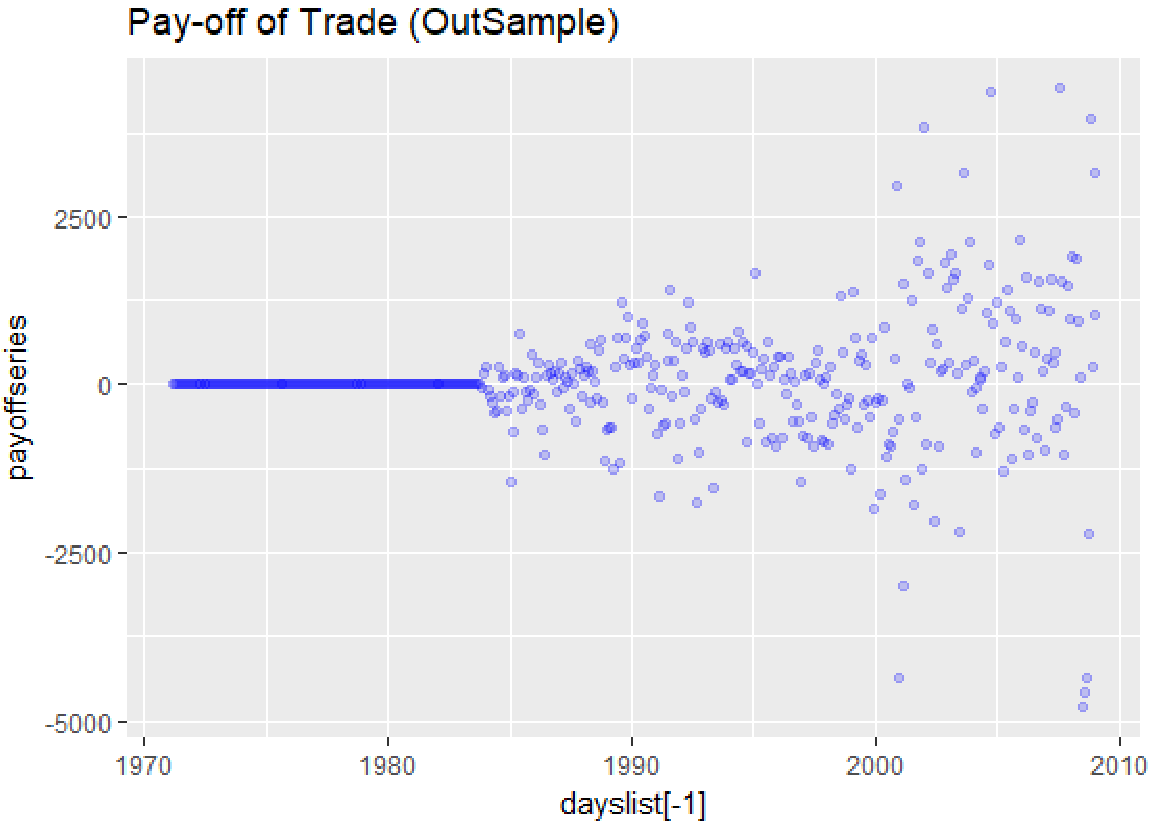


Figure 7: Out Sample Trade Payoff

And finally the discounted wealth process:

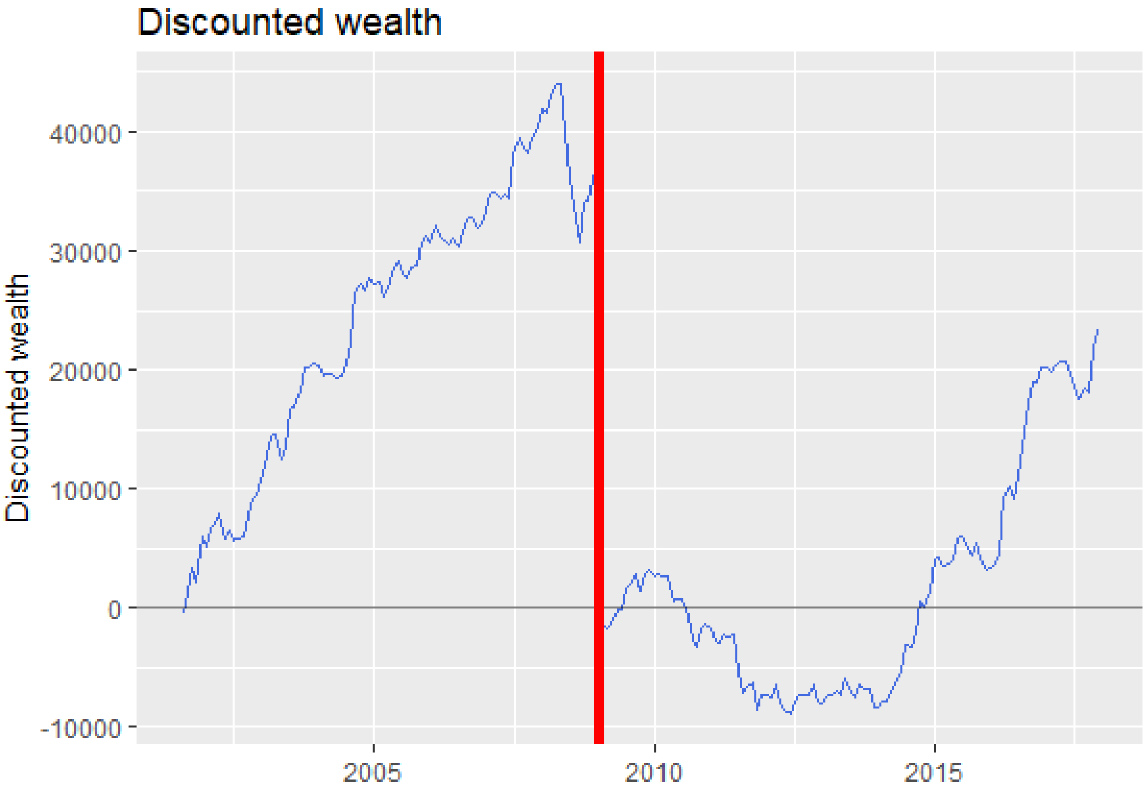


Figure 8: Out Sample Discounted Wealth

This time the wealth process is apparently more realistic than the insample case. The big difference indicates that our model overfits in-sample data, and thus the considerable model variance. However on the other hand, the discounted wealth process generally sloped upward, which make us optimistic about the model accuracy.

Similarly we testify the out sample trade payoff and discounted wealth process with formation period 3 months in comparison with the above result of 1 month.

out.png

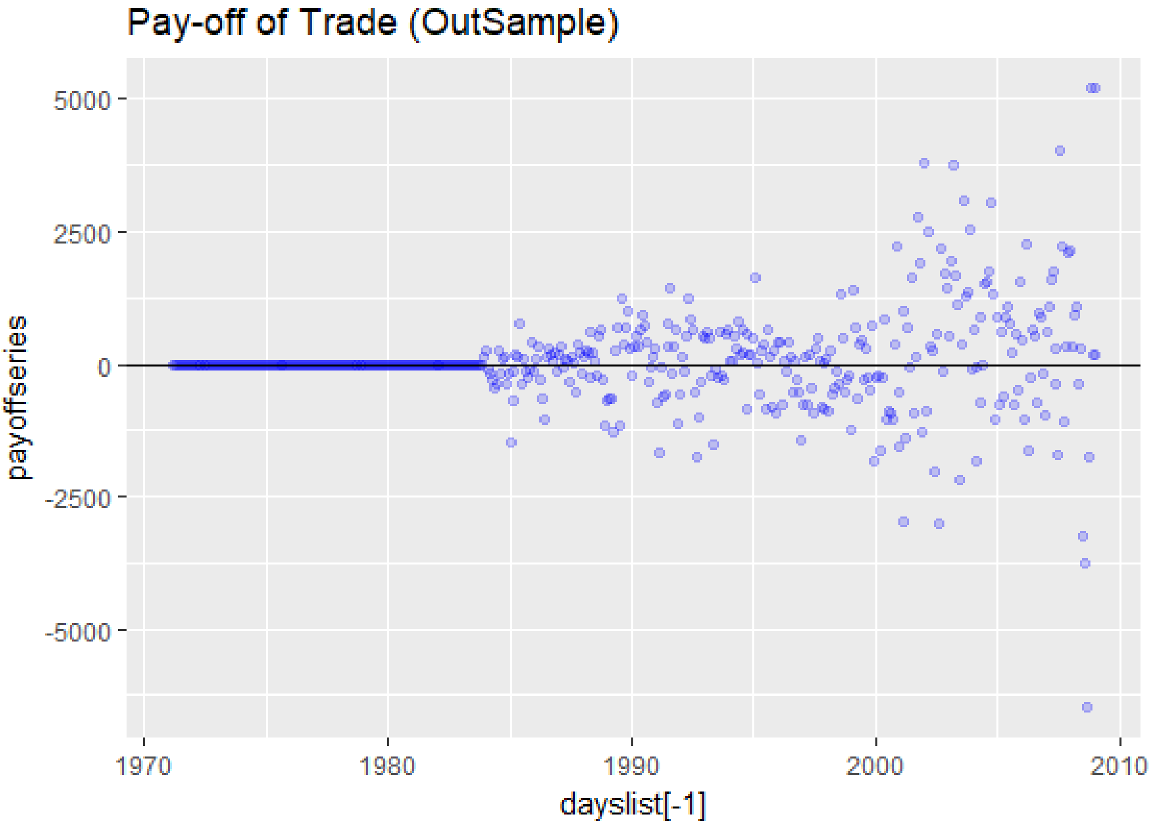


Figure 9: Out Sample Trade Payoff

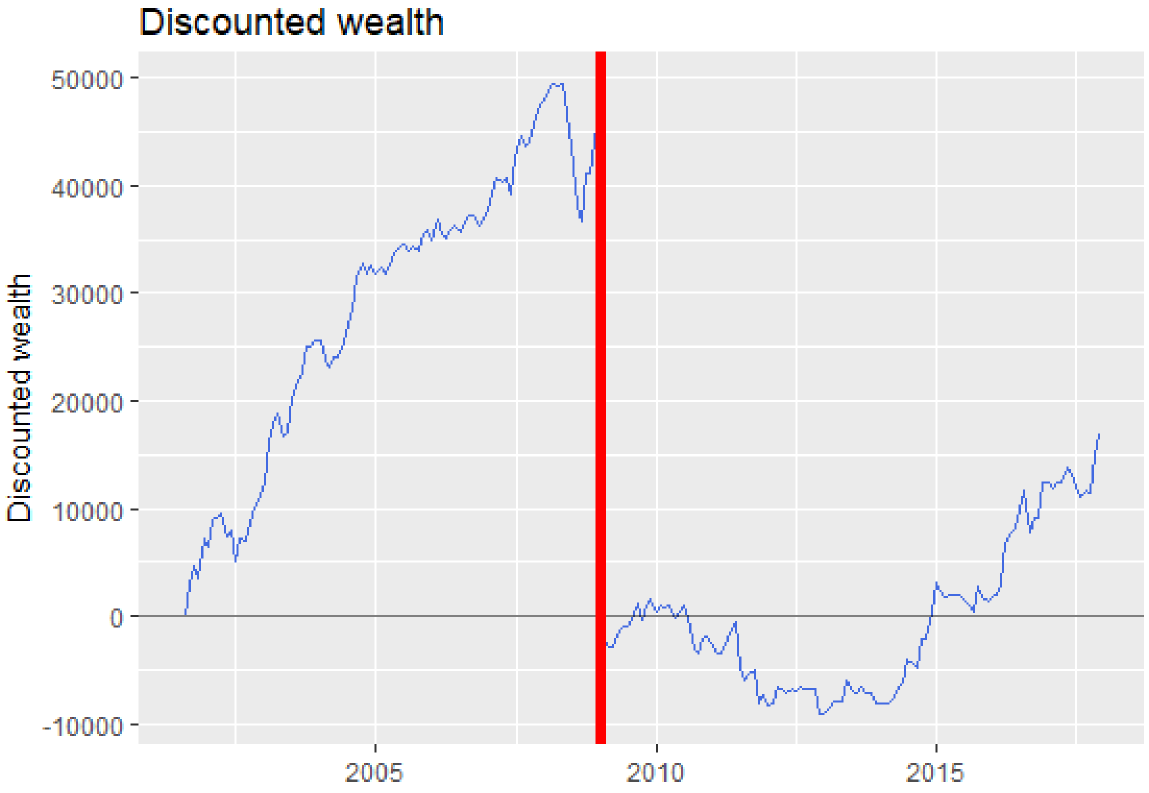


Figure 10: Out Sample Discounted Wealth

More trades ended up to be profitable, and the discounted wealth process has less big drawback compared with the one with 1-month formation horizon.

5.3 Hyperparameter Validation

From the analysis and result demonstrated above, we found formation horizon, as a hyperparameter, is influential to the strategy result. The testimony on training set indicates that 3-month formation is better than 1-month formation, but the performance on validation set indicates the case is not necessarily.

Indeed, we collected the following performance metrics on both training and validation set:

Sharpe Ratio:

F		1970-2008	2009-2017
1	insample	-0.0943245	2.45245
3	insample	2.105938	2.475493
1	outsample	0.09468	0.1860792
3	outsample	0.2456	0.1378378

Sortino Ratio:

F		1970-2008	2009-2017
1	insample	-0.1317433	2396.82
3	insample	33.179	2399.312
1	outsample	0.15178	0.3142233
3	outsample	0.55154	0.18053

Judging from in-sample over fitting, F=3 has a worse overfitting condition, i.e. the difference between in-sample performance and out-sample performance is larger. On the other hand, purely considering the out-sample performance, F=3 has an over all better performance, although not very stable.

6 Statistical Arbitrage Test

6.1 Theory

When testing a strategy, we observe specific profit/loss values from the strategy at equally spaced times $t_1, t_2, ...$ Let $p(t_i)$ be the net profit/loss that arrives at time t_i. Let $r(t_i)$ be the risk-free rate in effect in the time interval leading up to t_i after t_{i-1}. Then we have V(0)=0. And For each i¿0,

$$V(t_i) = exp[r(t_i)]V(t_{i-1}) + p(t_i)$$

and

$$v(t_i) = V(t_i)exp(\sum_{j=1}^{i} r(t_j))$$

Under the model described above, we can create a test for a statistical arbitrage as follows:

$$H_1 : \mu > 0$$

$$H_2 : \lambda < 0$$

A statistical arbitrage exists if and only if both H_1 and H_2 are true.

To perform a statistical test of arbitrage, we can use maximum likelihood to fit the model to data from any trading strategy. To calculate asymptotic standard errors for parameters that are obtained using maximum likelihood, the typical approach begins by estimating the matrix of second partial derivatives of the log-likelihood evaluated at the MLE.

To validate the hypothesis testing result, we need to conduct residual analysis to check the model fit. $\hat{z}_i = \frac{\Delta v_i - \hat{\mu}}{\hat{\sigma} i^{\hat{\lambda}}}$ should resemble i.i.d. standard normal random variables if the model fits well. We can use Q-Q plot to see how normal they look, and use ACF and PACF to see how i.i.d they look. If the residuals do not seem to satisfy the i.i.d. standard normal distribution, we'll need to refit using AR, MA noise model or changing to t distribution, which will be discussed in details in the second part.

6.2 Result

First, using the profit and the risk-free rate series from our strategy, we can calculate the two p-values for the two sub-hypotheses H_1^C and H_2^C under the assumption that residuals conform to i.i.d. standard normal distribution.

$$p_1 = 0.004029457$$

$$p_2 = 0.717825$$

Since p_2 is relatively big, we don't seem to have statistical arbitrage opportunity here.

Then, we take a look at the residuals of the standard model. From the normal Q-Q plot and the ACF and PACF plots below, we find the z_i noise might have a heavier-tailed distribution.

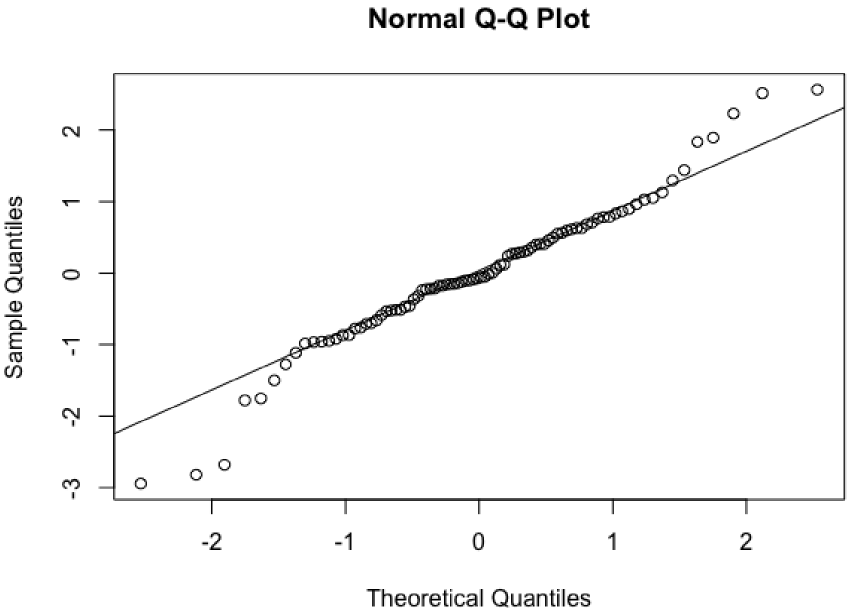


Figure 11: Normal QQ Plot

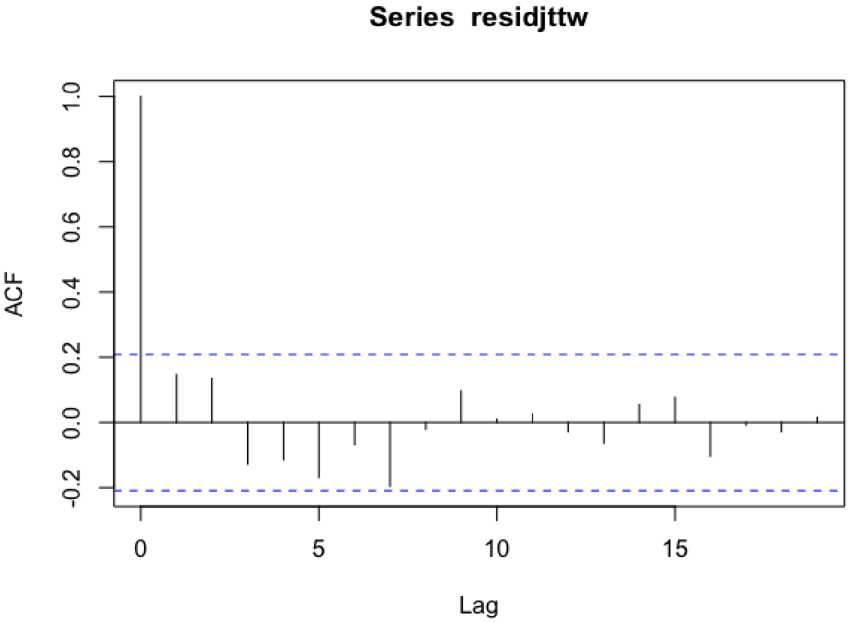


Figure 12: Normal ACF Plot

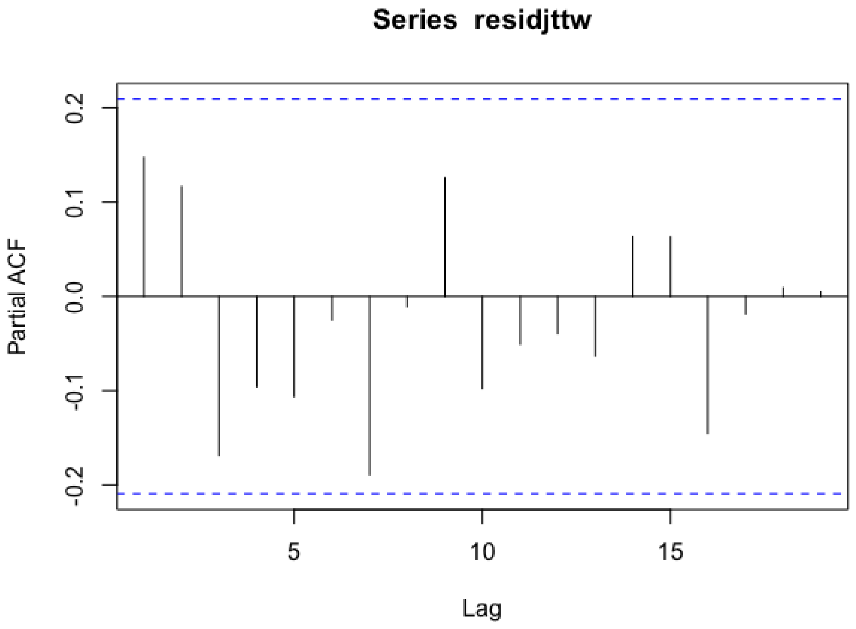


Figure 13: Normal PACF Plot

According to the plots above, the noise terms seem to have t distribution, and we do not need to use AR nor MA models for the z_i's. So we replace all of the normal distributions with t distributions, and fit one more parameter for the degrees of freedom.

Refitting the model, we have:

$$p_1 = 0.002381593$$

$$p_2 = 0.7035756$$

$$p_3 = 4.37591e^{-14}$$

Checking the residuals of the new model, we have:

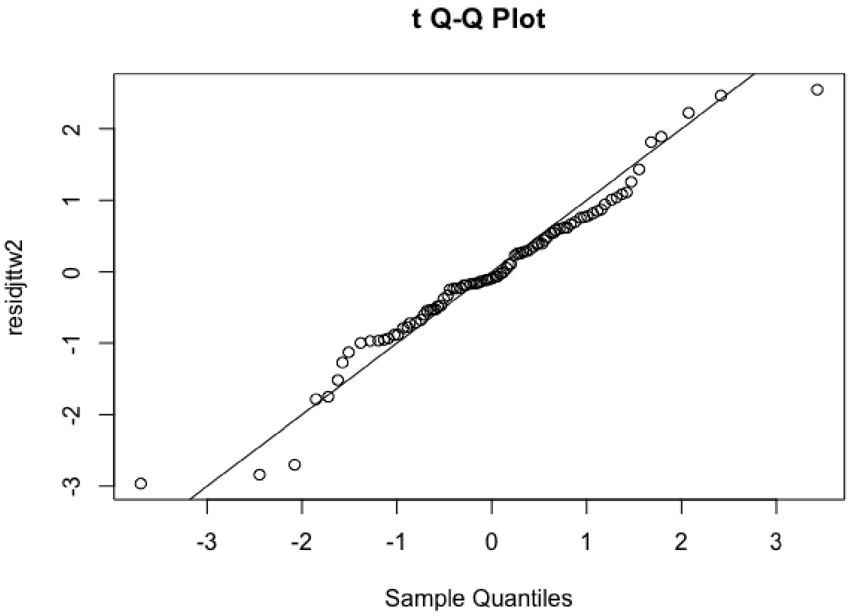


Figure 14: t QQ Plot

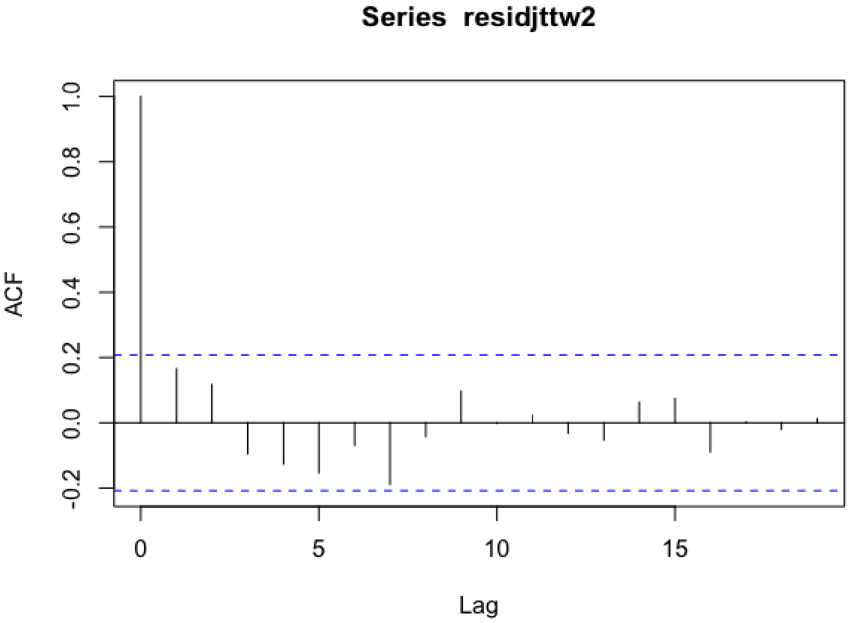


Figure 15: t ACF Plot

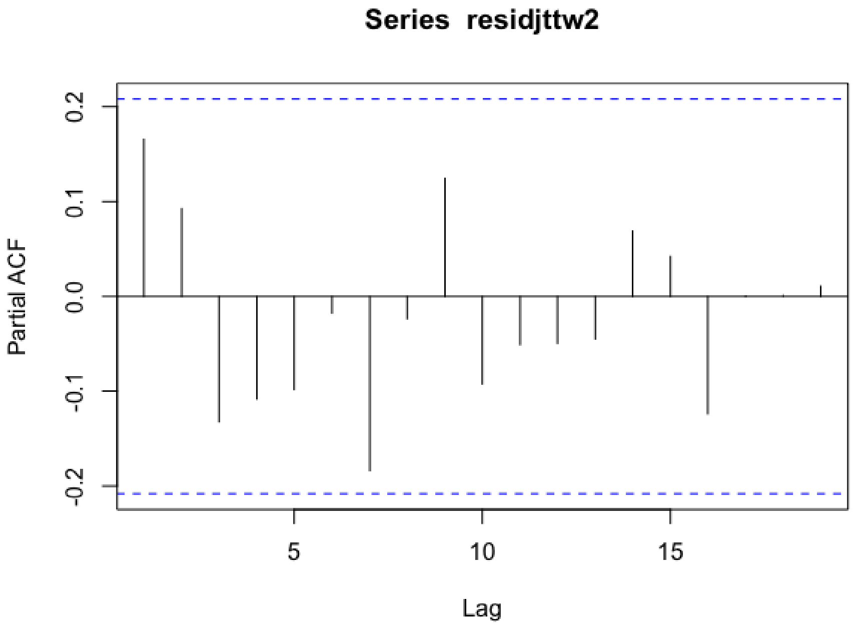


Figure 16: t PACF Plot

From the three plots above, the noise terms seem to fit with a t distribution model well and we do not need more complicated models. Since p_1 and p_3 are sufficiently small but p_2 is as large as 0.7035756, we cannot reject H_2^C. Therefore, the strategy is not statistical arbitrage.

7 Market Caveat

7.1 Transaction Cost and Spread

As one of extremely liquid markets, FX market has strikingly low tick sizes due to the stability often associated with fiat currencies. In futures exchanges such as the CME, the 6E futures contract (EUR/USD pair) has a tick size of $0.00005 USD per EUR. This is known as a half-pip, a full pip being $0.0001. In order to make trading worthwhile, market participants will often use this aforementioned leveraging to accelerate returns and hence also their losses. Leveraging in FX markets can even reach levels of 100:1[3]!

7.1.1 As a Model Factor

Due to the lack of easily and cheaply available bid/ask price data in the FX market, we decided to assess the impact on our strategy over a range of potential cost scenarios. Initially, we applied a zero cost whenever a transaction is made for our account. Next, we wanted to see the effectiveness of our strategies where the spreads were: a half-pip, a full pip, or two pips regardless whether we are "lifting the offer" or "hitting the bid" in a trade. We deduce this is a fair estimate of the spread and the impact we can have on markets. Additionally, we wanted to distinguish the scenario where certain pairs have wider spreads compared to others. We wanted to explore the scenario where liquid markets like EUR/USD have a transaction cost of one pip and more illiquid markets such as USD/RUB have two pip spreads.

7.2 Last Look

Shockingly to most, even in the financial services sector is the concept of "last look." A controversial feature in FX markets privy to the market-makers in the space. Last look is an option a market-maker (a liquidity provider) has to back out of a trade within fifty milliseconds. For example, suppose we take liquidity in a market by lifting the offer, and the market-maker on the other end gets swept –when the quantity at the best ask clears out and the new ask increases by a tick– the market-maker proceeds to exercise last look and cancels that trade in order to keep his edge and maintain profitability in a typically non-volatile market. Given the lack of limit order book data to identify the exercise rate of last look and the need to factor this possibility into our model, we look to assume this happens 50% of the time. When it does occur, we would pay an additional tick value in price impact.

8 Conclusion and Further Discussion

8.1 Conclusion

In conclusion, the carry trade strategy we developed, although not passing the statistical test to qualify as a statistical arbitrage, still has profitability in general.

This stands in support of the carry trade strategy and the forward premium puzzle, although we employed a more sophisticated decision process than just observing the forward premium. This to some extend indicates that the FX market is not as information efficient as equity market, and more statistical arbitrage to be exploited.

8.2 Further Discussion

In terms of the model goodness, although we carefully build model to lower the variance and avoid overfitting, it still shows overfitting pattern, which is in need of future improvement.

This return prediction and all of our analysis so far are based on the goal to predict the value of the true return, i.e. we are treating the problem as a regression problem. Nonetheless, the ultimate goal of our statistical arbitrage strategy as about making decisions, so if we simplified the problem to just making confident guess of long or short position, the limited sample problem could be further addressed.

9 Reference

- http://www.nasdaq.com/forex/education/foreign-exchange-market-overview.aspx

- http://www.cmegroup.com/trading/fx/g10/euro-fx_contract_specifications.html

- https://www.bofaml.com/content/dam/boamlimages/documents/PDFs/last_look_and_other_disclosures_final_copy.p

- http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.470.3846rep=rep1type=pdf