

Optimizing Pairs Trading of US Equities in a High Frequency Setting

Vayu Kishore

Adviser: Professor Phillipp Illeditsch

Wharton Research Scholars

April 20, 2012

Abstract

In this paper, we examine how to the performance of high-frequency pairs trading strategies are impacted by the allocation within the pair, opening and closing thresholds, restriction to daily trading, and transaction costs. We generate portfolios by applying high-frequency pairs trading strategies to the pair consisting of Exxon Mobil Corporation (XOM) and Chevron Corporation (CVX) during the year 2005. We find the following results. First, we find that a dynamic model for estimating the spread of a pair is more suitable for high-frequency trading when compared to a static model. Second, we find that allocating investment within the pair based on the ratio of their CAPM β , compared to a 1:1 dollar allocation, and allocation based on the cointegration coefficient, yields the most attractive portfolios. Third, we find that setting the opening threshold to 1.5σ and the closing threshold to 1.0σ , respectively generate portfolios with the highest Sharpe ratios when compared to portfolios constructed using the same strategy, but different threshold values. Finally we find that restricting trading to once-a-day and imposing transaction costs significantly worsens the performance of the strategy.

Introduction

High Frequency Trading

High Frequency Trading (HFT), a recent innovation in financial markets, uses high-speed connections and powerful computing to execute algorithmic trading strategies on intraday market data. HFT has come to play a significant role in the market. Zhang (2010), for example, finds that 78% of trading volume in US capital markets are due to high frequency trading.

Statistical Arbitrage and Pairs Trading

HFT may be used in conjunction with a pairs trading strategy. In pairs trading strategies, a trader takes opposing long and short positions in two assets when the difference their prices hits a certain opening threshold. These positions are then closed when a certain closing threshold is reached. The difference in the prices that the trader uses to judge when to open and close a position is referred to as the spread between the two assets. The two assets identified are expected to move together due to their status as close substitutes for each other. Examples of pairs include oil companies, large financial institutions, and credit card companies. Pairs trading strategies seek to exploit temporary mispricings of

assets within the market and thus, they rely on mean-reversion and construct market-neutral portfolios whose net market exposure is negligible. Such strategies have been shown to reap sizable profits in the past. Gatev (2006), for example, examined the application of pairs trading to daily data and finds that such strategies may generate excess profits of up to 11% for self financing portfolios, with Sharpe Ratios of 0.35-0.59.

Pairs Trading application to HFT

Recently, with the growing popularity of HFT, various studies which examine the applicability of pairs trading strategies to high frequency environments have been performed. Bowen et al. (2010) examine the sensitivity of high frequency strategies to market attributes, noting that the primary returns to their strategy arise in the first hour and last hour of trading days, when trading volume is expected to be highest. Additionally, they note that transaction costs of 15 basis points significantly deteriorate returns, as do single-interval delays in opening positions. Nath (2003), examines high-frequency pairs trading with US Treasury Securities and finds that high-frequency pairs trading strategies almost always outperform benchmarks like the Salomon Brothers Treasury Index and S&P 500.

In the studies discussed, while optimization of portfolios consisting of pairs has been examined in a high-frequency setting, optimization of asset allocation within the pair has not. For example, an equal dollar amount in each pair is invested in assets in Munir and Krishnan (2010) and Chiu et al. (2010). Allocation within the pair is an important factor of such strategies, since an important aspect of such strategies is that they strive to be market-neutral. Assessing how to allocate investment within a pair is one of the questions this paper seeks to investigate.

Moreover, in the HFT pairs trading studies examined, opening and closing thresholds are arbitrarily defined. Most often, pairs are opened when the spread exceeds 2 standard deviations of its historical mean and closed when it reverts to 0.5 standard deviations of its historical mean. These thresholds are used in both Chiu et al. (2010) and Dunis et al. (2010). As these thresholds play a key role in determining the profitability of a strategy, we also investigate the effect of changing them in this paper.

Materials and Methods

Overview

In this study, we first select an individual stock pair and test if it is compatible with a pairs trading strategy from 2004 to 2005. We define a pair is compatible with a pairs trading strategy if their CAPM β 's are within 0.15 of each other and if the logs of their price processes are cointegrated. We then study the performance of a high-frequency pairs trading strategy during 2005 according to different parameters, such as allocation ratios, thresholds, and transaction costs. In the analysis performed, the SPDR S&P 500 ETF (SPY) is used as a proxy for the market and the iShares Barclays 1-3 Year Treasury Bond Fund (SHY) is used as a proxy for the risk-free rate. The statistical programming language, R is used for the implementation of the strategy and the R packages RTAQ and PerformanceAnalytics are used for handling raw tick-by-tick data and performance analysis of portfolios, respectively.

Pair Selection

We use the “classic” pair of Exxon Mobil Corporation (XOM) and Chevron Corporation (CVX) as our candidate pair. This pair has been studied in the literature and thus provides a benchmark for pairs trading strategies. Additionally, using these large corporations ensures availability of intraday data and liquidity of their stock at an intraday level. Since these stocks are highly liquid, we also assume that we can short sell the stocks freely without additional costs. We then evaluate the two stocks for pairs trading compatibility using daily data from 2004-2005 obtained from CRSP. We test their compatibility by testing if their β 's based on the Capital Asset Pricing Model (CAPM) differ by at most 0.15 and testing if they are cointegrated. In this discussion, and in the rest of the paper, we refer to XOM as stock 1 and CVX as stock 2.

The CAPM predicts the return of an asset, i as $E(r_i) = (E(r_m) - r_f)\beta_i + r_f$, where r_f is the return on the riskfree asset, and r_m is the return on the market. From this equation, β_i can be interpreted as the correlation of a security's return to the market's return. Thus, if the ratio β_i/β_j for two separate securities, i and j is close to one, then we expect them to be affected by market movements in the same fashion, a condition that favors pairs-trading compatibility.

We then test the prices of the two assets for cointegration, an feature of attractive pairs. Cointegration, as defined by Engle and Granger is a statistical property that tests if two processes tend

move in step with each other.

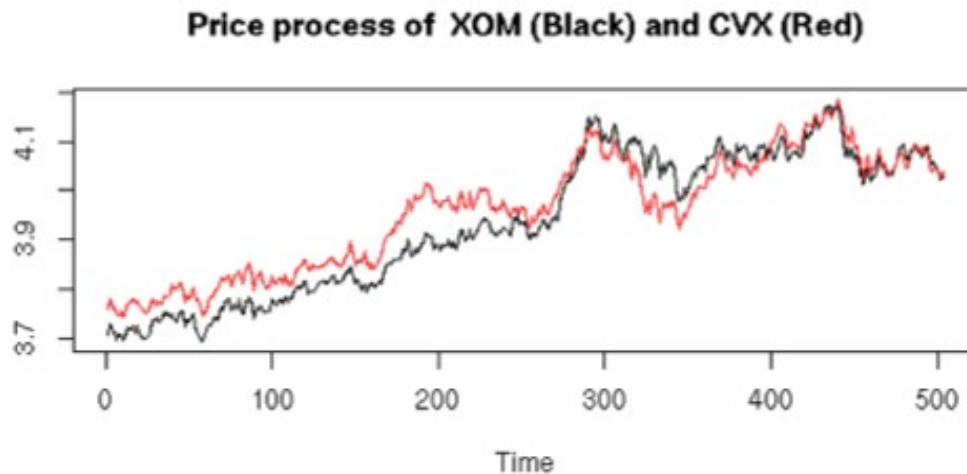


Figure 1: A time series plot of the logged prices of the “classic” pair XOM (Black) and CVX (Red), from January 2004 to December 2005. The processes of the “classic pair” appear move in step with each other, suggesting that they may be cointegrated. This is tested for by the Engle Granger two-step test for cointegration.

We test for cointegration using the Engle-Granger two step cointegration test. To do this, we performed the following steps. These steps are adapted from the procedure outlined in Vidyamurthy (2004):

- Fit a best fit line using Least Squares Linear Regression to the equation $\log(P_1) = \log(P_2) * \beta_{coint}$, where P_1 and P_2 refer to the respective prices of each stock in the pair and the fitted parameter, β_{coint} , will be referred to as the cointegration ratio. We constrain the intercept to 0 since if pair is cointegrated, then we expect 0 returns on one asset to predict 0 returns on the other.
- Construct the spread between the two assets after stripping out the effects of cointegration where the spread at time t , $S_t = \log(P_1) - \beta_{coint} * \log(P_2)$
- Test spread of pair for stationarity using an Augmented-Dickey Fuller (ADF) Test, which tests the null hypothesis that a process has a unit root (is not stationary). If the pair is cointegrated, then the spread should be stationary.

High-Frequency Data

We obtain tick-by-tick quote data on the selected pair for 2005 from the NYSE TAQ Database in monthly files. These monthly files are then split into daily files using a Python script and cleaned using the function `quotesCleanup` from the R package `RTAQ`, which performs procedures as outlined in

Brownlees and Gallo (2006). This step was used to filter quotes where either the bid or the ask is zero, merge simultaneously observed data points, remove observations with bid-ask spreads which are more than 50 times the daily median spread, and remove quotes that with bids or asks that are 10 standard deviations away from the median from a window of 50 observation, an approach taken by Chiu (2010). Additionally, since the raw TAQ data contains observations from multiple exchanges, in order to simulate a high frequency trader which operates on a single exchange, the cleanup step was used to select only quotes from the NYSE.

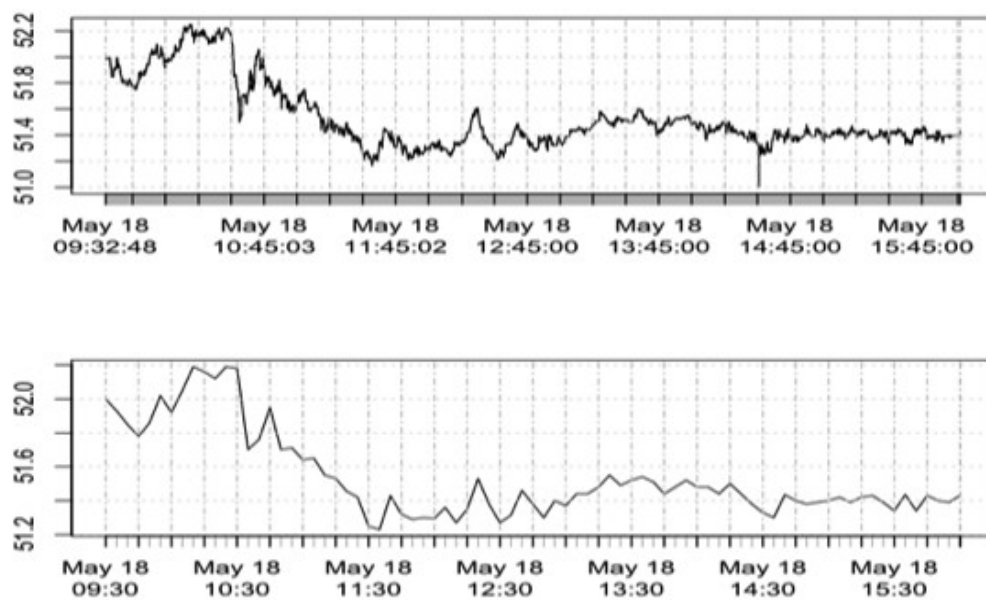


Figure 2: Visual comparison of cleaned tick-by-tick quotes (top) to cleaned and tick-by-tick quotes aggregated to a 5 minute level (bottom) for a representative day (May 18th) for XOM. Note the smoothing away of features such as the drop at 14:30 seen in the tick-by-tick data, but missing in the aggregated data.

Since the data points for each asset are observed tick-by-tick, . To rectify this, we use the `aggregateQuotes` function of RTAQ to create a synchronous time series of observations, where quotes are aggregated to the 5 minute level. This simulates an algorithm which polls the market every 5 minutes.

Construction of Pairs Trading Strategy

The primary parameters that are input into the strategy are the allocation ratio (which govern

how a position is opened) and the open, close, and stop-loss thresholds (which govern when to open or close a position). Portfolios created by this strategy begin with \$1 of wealth and are self financing. We assume that any unused cash in the portfolio is not invested and does not grow. Additionally, if the net wealth of the portfolio is ever negative, then trading is terminated. The criteria we use to form the structure of the pairs trading strategy is:

1. For each time point in the time series, calculate the risk-adjusted spread between the two assets of the pair. Calculation of the spread is specified in Table 1 and depends on the allocation ratio chosen.
2. Call the amount the spread deviates from a measure of the historical spread the “signal”. If the signal is greater than or equal to the opening threshold, open a position if not already in one. The relative amounts placed in each position are governed by the allocation ratio chosen. If the spread is above its historical mean, then we expect that stock 1 is overpriced and stock 2 is underpriced. Thus, we short-sell stock 1 and buy stock 2. On the other hand, if the spread is under its historical mean, we buy stock 2 and short-sell stock 1. We standardize the position we take in stock 1 so that we are long or short exactly 1 dollar at a given time in the stock. We adjust the position in stock 2 based on the allocation strategy.

The specific characteristics of each allocation ratio are:

Table 1: Description of how values for the high-frequency pairs trading strategy are computed for the 3 allocation ratios considered

Allocation Ratio	Calculation of spread	Magnitude, in dollars, of position in stock 1 entered when position is opened	Magnitude, in dollars, of position in stock 2 entered when position is opened	Position entered in units of stock 1 when position is opened	Position entered in units of stock 2 when position is opened
1:1 Dollar	$\log(P_1) - \log(P_2)$	1	1	$\frac{1}{P_1}$	$\frac{1}{P_2}$
CAPM β	$\log(P_1) - \frac{\beta_1}{\beta_2} \log(P_2)$	1	$(\frac{\beta_1}{\beta_2})(\frac{P_2}{P_1})$	$\frac{1}{P_1}$	$(\frac{\beta_1}{\beta_2})(\frac{1}{P_1})$
Cointegration β	$\log(P_1) - \beta_{coint} \log(P_2)$	1	$\beta_{coint}(\frac{P_2}{P_1})$	$\frac{1}{P_1}$	$\beta_{coint}(\frac{1}{P_1})$

3. If the signal is less than the closing threshold, close any existing position in the pair.

4. If a the signal is greater than the stop-loss threshold, we close the position
5. If a position is open on the last time point in the data series, we close the position.
6. We assume that the strategy begins with \$1 and terminate execution of the strategy if the net value of the portfolio created by the strategy is ever negative (no leverage is allowed).

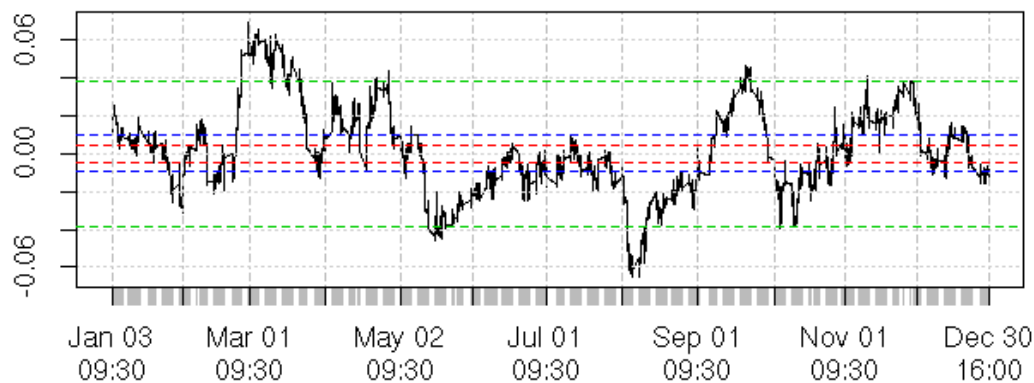


Figure 3: Illustration of how thresholds are used to determine entry and exit from positions in a pairs trading strategy. The process plotted is the $\log(\text{Price XOM}) - \log(\text{Price CVX})$ for the year 2005. In this example, the open threshold is set to 1σ (blue), the close threshold is set to 0.5σ (red), and the stop-loss threshold to 4σ (green).

Calculation of Signal

We originally considered using a static model to calculate the signal, where the mean and standard deviation of the spread are calculated from the historical daily data and remain fixed over the course of the strategy's execution. Based on preliminary results, we decided against this approach, the reasons for which are outlined in the Discussion section. Instead, we use a dynamic model to calculate the signal by recalculating the spread process on a daily basis by using the 50 day rolling means and standard deviations of the spread.

Strategies analyzed

In analyzing strategies with different parameters, we use the following parameters as our base case for comparison:

Table 2: Base case pairs-trading strategy parameters

Allocation Strategy	Open Threshold	Close Threshold	Stop Loss Threshold	Restrict Trading to Once a day?	Transaction cost
CAPM β	1 σ	0.5 σ	4 σ	No	0 bp

To conduct our analyses, we vary the parameters of the base case based on the following categories:

1. Allocation strategies: We test the three different approaches to allocation discussed earlier
2. Open and Close Thresholds: We vary the open and close thresholds of the base case on two dimensions:
 - i. We vary the “width” of the gap between the open and close thresholds. To do this, we fix the close threshold at 0.5σ and we test the open threshold at values 1σ , 1.5σ , 2σ , 3σ
 - ii. We vary the “level” of the gap by keeping the gap between the open and close thresholds constant at 0.5σ and testing the strategy at the values for (open threshold, close threshold) = $(1\sigma, 0.5\sigma)$, $(1.5\sigma, 1\sigma)$, $(2\sigma, 1.5\sigma)$, $(2.5\sigma, 2\sigma)$
3. Restricting trading to daily: To test the value of performing this strategy at a high-frequency, rather than a daily basis, we also test a strategy that can only trade once a day when the market opens.
4. Test the effect of placing a 15 basis point transaction cost on trades as done in Bowen et al. (2010)

Results and Discussion

We initially attempted to model the spread using a static model for the spread's parameters. However, this resulted in a strategy which opened and closed positions less than 10 times in a year and where some positions were held open for longer than a month. We observe that a strategy with these characteristics is not amenable to high frequency trading, and so we use the dynamic model outlined in Materials and Methods. As shown in Figure 5, this method provides a strategy which enters and exits positions many times over the course of a year and is thus more suitable for high frequency trading.

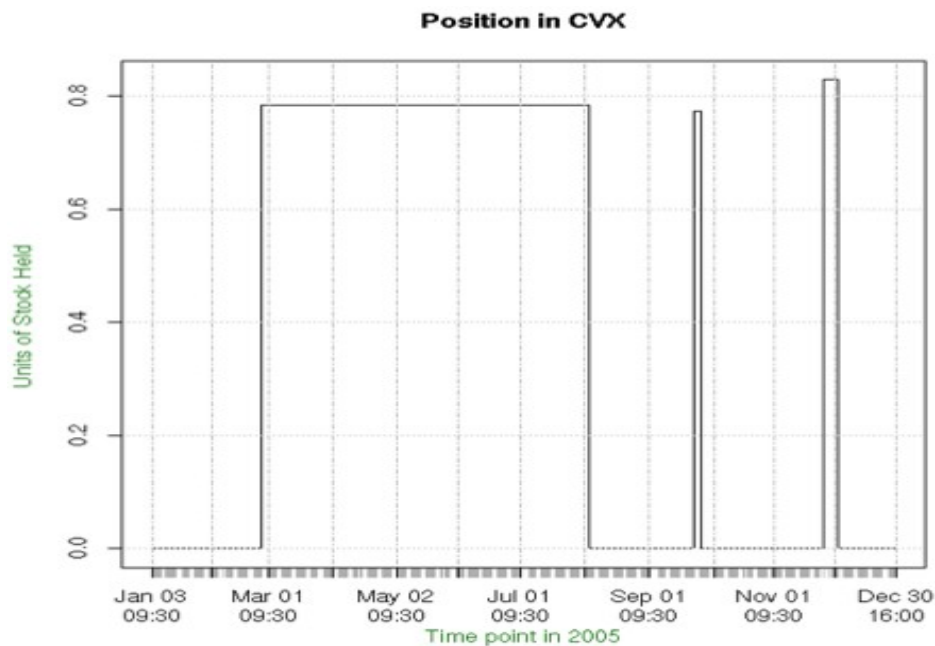


Figure 4: Representative positions held in CVX (stock 2) when a high-frequency pairs-trading strategy using a static model to estimate the signal is executed. This figure was generated using a strategy that uses the cointegration allocation strategy with open, close, and stop-loss thresholds at 1σ , 0.5σ , and 4σ

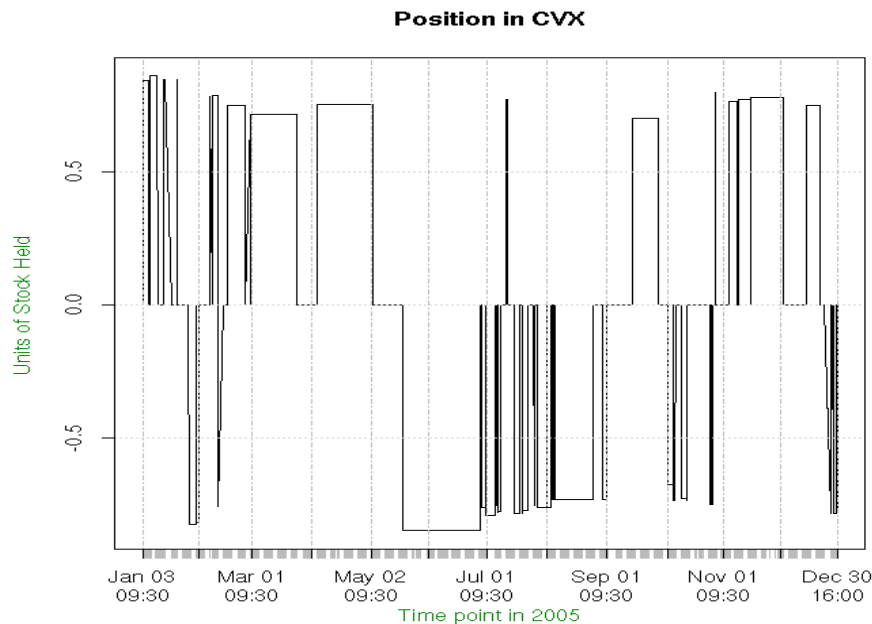


Figure 5: Representative positions held in CVX (stock 2) when a high-frequency pairs-trading strategy using a dynamic model to estimate the signal is executed. This figure was generated using a strategy that uses the cointegration allocation strategy with open, close, and stop-loss thresholds at 1σ , 0.5σ , and 4σ

Table 3: Pairs Trading Compatibility Testing Results

CAPM Test	Estimated value	Std. Err	p-value	Cointegration Test	Estimated value	Std. Err	p-value
β_{XOM}	1.0675	0.0660952	$< 2e^{-16}$	Cointegration Ratio	0.9946	0.0004968	$< 2e^{-16}$
β_{CVX}	0.9414	0.0703585	$< 2e^{-16}$	ADF Test p -Value	0.0520		
$\frac{\beta_{CVX}}{\beta_{XOM}}$	0.8819						

The pairs trading compatibility tests indicate that the β of the two stocks within the pair are close. Additionally, ADF test rejects the null hypothesis that there is a unit root for the spread between XOM and CVX indicating that the logged price processes of the pair are cointegrated. Both of these factors suggest that the pair of XOM and CVX is compatible for pairs trading.

Strategy Analysis

In the following tables, only the parameters which change within the category are shown in the table, if a parameter is not shown, then it is identical to the base case. For each portfolio, its annualized return, annualized Sharpe ratio, and 5% Value at Risk (VaR) assuming a Gaussian distribution is provided.

Table 4: Results for Allocation Costs

Allocation Strategy	Annualized Return	Sharpe Ratio (annualized)	VaR (5%) as % of initial wealth
1:1 Dollar	9.4970%	0.7106	1.0995%
CAPM β	12.3709%	1.0544	1.0202%
Cointegration β	7.7422%	0.5428	1.1062%

Allocating investment within the pair based on the ratio of the CAPM β 's has the highest

annualized return, Sharpe ratio, and lowest VaR. Surprisingly, the simple 1:1 dollar ratio provides superior performance compared to allocation based on the Cointegration β . This may be because cointegration describes a long-term phenomenon observed on the scale of months, rather than days. Thus, although the assets are cointegrated in the long run, these results may indicate that we cannot exploit this phenomenon in a high-frequency setting.

Table 5: Results for Threshold (varying width)

Open Threshold	Close Threshold	Annualized Return	Sharpe Ratio (annualized)	VaR (5%) as % of initial wealth
1 σ	0.5 σ	12.3709%	1.0544	1.0202%
1.5 σ	0.5 σ	19.5119%	1.8220	0.8371%
2 σ	0.5 σ	10.7377%	1.0166	0.7574%
3 σ	0.5 σ	-3.5647%	-0.8822	0.6847%

Varying the “width” of the thresholds indicates that there is an optimal width of the threshold gap between between 0.5σ and 1.5σ . This maxima may be the result of a risk/return tradeoff. When the opening threshold is made higher, then one may expect to obtain higher profits from convergence. However, since the gap is wider, there is a higher likelihood of price movements which result in a loss when convergence occurs and the position is closed.

Table 6: Results for Threshold (varying level)

Open Threshold	Close Threshold	Annualized Return	Sharpe Ratio (annualized)	VaR (5%) as % of initial wealth
1 σ	0.5 σ	12.3709%	1.0544	1.0202%
1.5 σ	1.0 σ	17.6155%	1.8745	0.7672%
2 σ	1.5 σ	5.4960%	0.5791	0.6151%
2.5 σ	2 σ	-7.0594%	-1.8094	0.5638%

These results indicate that, if the width of the thresholds is held fixed, then the optimal level of the open and close thresholds occurs between 1.5σ and 1.0σ respectively.

Table 7: Results for once a day constraint on trading

Restrict Trading to Once a day?	Annualized Return	Sharpe Ratio (annualized)	VaR (5%) as % of initial wealth
No	12.3709%	1.0544	1.0202%
Yes	2.5228%	0.0455	1.0177%

These results indicate that trading at a high frequency under this dynamic model provides significant improvements compared to restricting trading to once a day. Interestingly, the VaR of both portfolios is the same, indicating that restricting trading to once a day worsens returns, but does not significantly mitigate tail risk.

Lastly, we examine the effect of imposing transaction costs on our strategy. Imposing a transaction cost of 5bp decreases our returns 11.1%. Imposing a transaction cost of 15bp decreases our annualized return to about -12.9% and increase our VaR to 13% of the initial wealth. The 25.2% drop in the annualized return is greater than that found in the results of Bowen et al. (2010) who find that the returns of their strategy fell 14.3% when 15bp of transaction costs were imposed.

Table 8: Effect of transaction costs on performance of pairs trading strategy

Transaction Cost	Annualized Return	Sharpe Ratio (annualized)	VaR (5%) as % of initial wealth
0 bp	12.3709%	1.0544	1.0202%
5 bp	1.2516%	0.4053949	1.5839%
15 bp	-12.9113%	-1.3239	13.4553%

Conclusions

Our results suggest that statically estimating the parameters of the historical spread does not yield a strategy that is amenable to high-frequency trading. The failure of the statically estimated spread indicates that it may be infeasible to exploit long-term market phenomena such as cointegration of assets on an intraday basis. In other words, the characteristics of intraday price processes may be significantly different than those of longer term processes. A dynamically estimated spread which is recalculated on a daily basis can be used to address these issues of using a static model.

The results indicate that using the ratio of CAPM β provides more attractive returns for a high-

frequency pairs trading strategy compared to the other two allocation methods studied. Using an allocation strategy advised by the theory of cointegration yields the poorest returns.

Furthermore, modifying the thresholds significantly affects the portfolios generated by the strategy. The optimal open and close thresholds for the pair consisting of XOM and CVX during 2005 seem to be at 1.5σ and 1.0σ respectively.

Restricting trading to once daily significantly decreases the performance of the pairs-trading strategy, indicating that trading at high-frequency generates more attractive portfolios when compared to trading on a daily basis.

Lastly, imposing transaction costs of 5bp and 15bp decreased our returns by 11.1% and 25.2% respectively. These results indicate that the profitability of this high-frequency pairs trading strategy is highly sensitive to transaction costs.

Future Work

This paper provides an initial overview of how allocation ratios within a pairs trade and the thresholds that define such a trade can impact the profitability of a HFT strategy by examining a single asset pair during 2005. An extension of this would be to assess if the conclusions of this paper hold across a larger sample of the market, with equities from sectors of the market outside of oil.

Another possibility may be to test the findings of this paper in the context of portfolios of pairs, rather than a single pair. This would combine the findings of Chiu (2010) with those of this paper and investigate whether optimizing both within and across pairs of a portfolio provides significantly superior returns when compared to the independent optimizations. Moreover, evaluating a pool of pairs simultaneously, rather than a single pair may generate more “events” where signals cross thresholds, allowing for a greater

A third possible extension is to run the analyses in the paper on tick-by-tick data that has been aggregated to a more granular level than every 5 minutes. 5 minutes was used in this paper for computational efficiency, but aggregating to this level may have smoothed some features of the data

that could be exploited by a high-frequency strategy.

Finally, another dimension to explore may be investigating different methods to generate the dynamic estimate of the spread. We use a window of length 50 to advise the estimates. It would be instructive to observe the tradeoff between using a longer window length to incorporate more observations and a shorter window length which keeps the data used to estimate the spread as recent as possible.

Literature Cited

- Bowen, David et al. "High Frequency Equity Pairs Trading: Transaction Costs, Speed of Execution, and Patterns in Returns." (2010). Web. 9 Jan. 2012.
- Brownlees, C.T. and Gallo, G.M. (2006). Financial econometric analysis at ultra-high frequency: Data handling concerns. *Computational Statistics & Data Analysis*, 51, pages 2232-2245.
- Chiu, Jonathan et. al. "High-Frequency Trading" (2011). Web. 9 Jan. 2012.
- Dunis, Christian et al. "Optimizing Intraday Trading Models with Genetic Algorithms." *Neural Network World*, 9.3. (1999): 193-233. Web. 9 Jan. 2012.
- Engle, Robert F., Granger, Clive W. J. "Co-integration and error correction: Representation, estimation and testing", *Econometrica*, 55.2 (1987): 251-276.
- Gatev, Evan et al. "Pairs Trading: Performance of Relative Value Arbitrage Rule." (2006). Web. 8 Jan. 2012.
- Munir, Danish and Divya Krishnan. "Financial Modeling: A system to test High-Frequency Trading Strategies." (2009). Web. 9 Jan. 2012.
- Nath, Purnendu, "High Frequency Pairs Trading with U.S. Treasury Securities: Risks and Rewards for Hedge Funds." (2003). Web. 9 Jan. 2012.
- Vidyamurthy, Ganpathy. "Pairs Trading." Hoboken: Wiley Finance. 2004. Print.
- Zhang, Frank. "High-Frequency Trading, Stock Volatility, and Price Discovery." (2010). Web. 8 Jan. 2012.

Appendix 1: Sample wealth processes for strategies. The y axis measures the % change in the net wealth compared to the initial starting wealth.

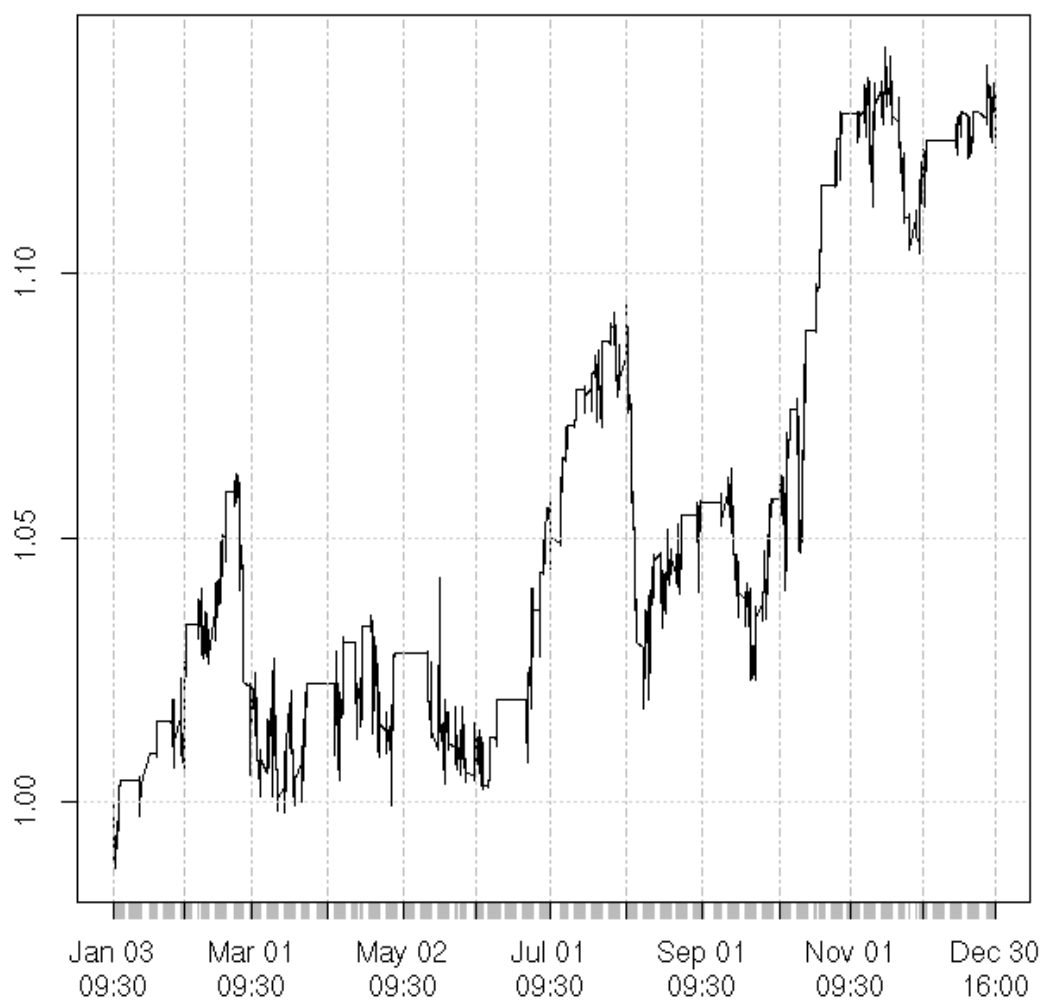


Figure 6: Wealth Process for Base Case pairs trading strategy with following parameters used

Allocation Strategy	Open Threshold	Close Threshold	Stop Loss Threshold	Restrict Trading to Once a day?
CAPM β	1σ	0.5σ	4σ	No

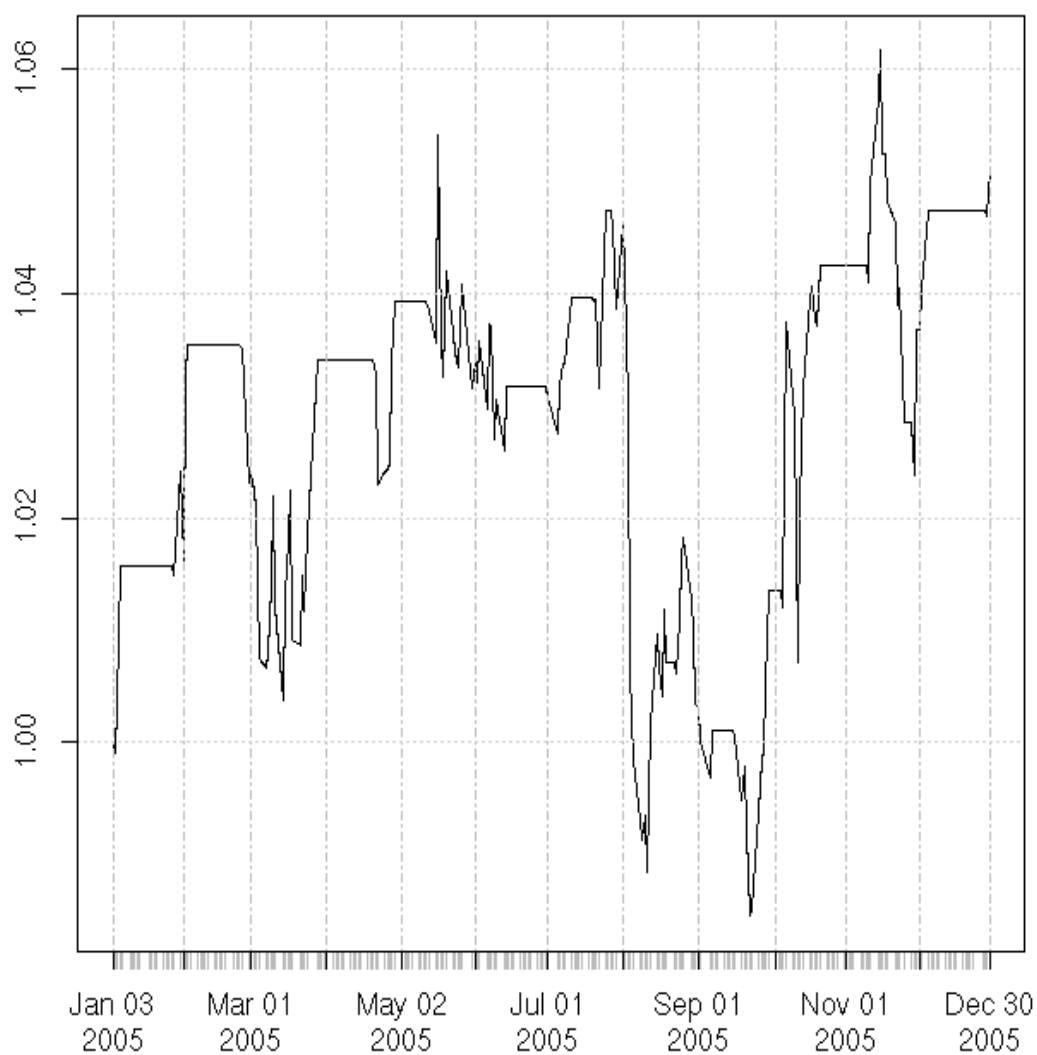


Figure 8: Wealth Process for pairs trading strategy restricted to once-a-day trading with following parameters used

Allocation Strategy	Open Threshold	Close Threshold	Stop Loss Threshold	Restrict Trading to Once a day?
CAPM β	1σ	0.5σ	4σ	Yes

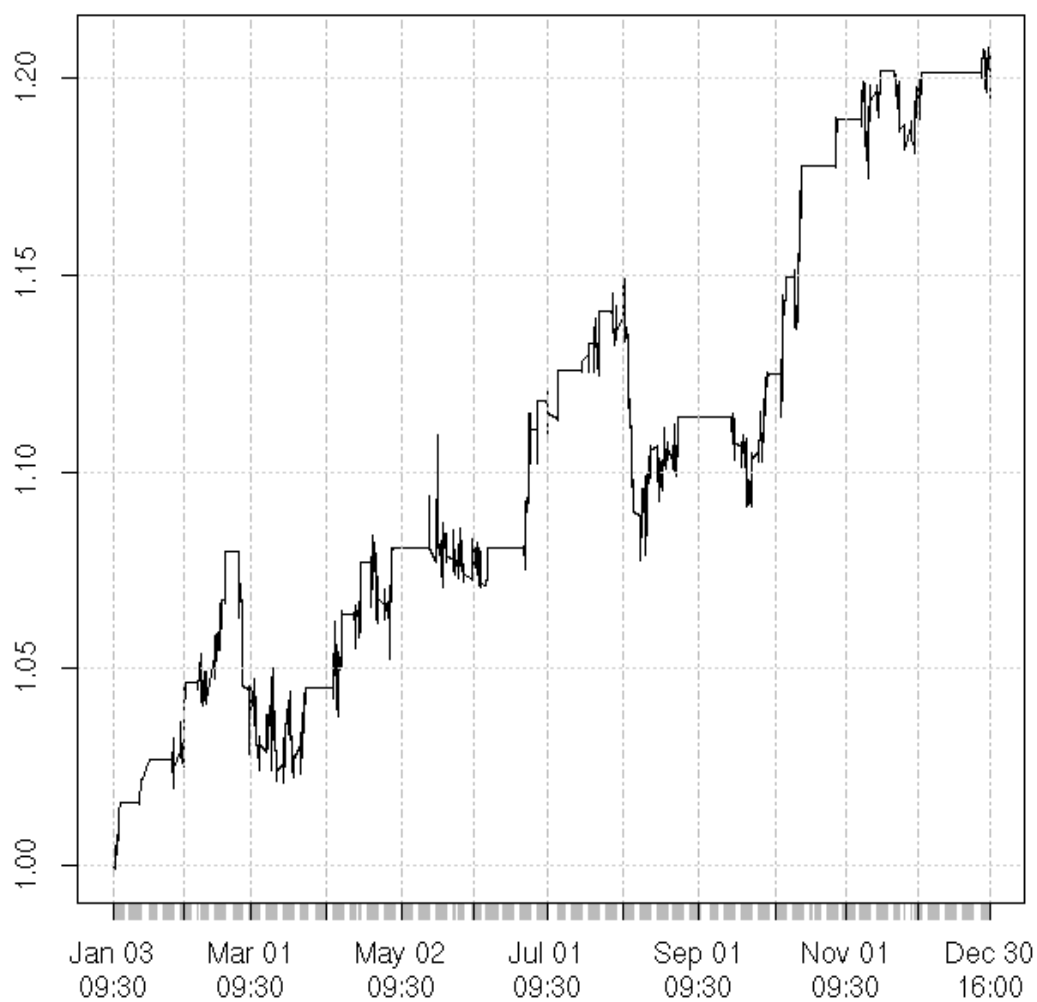


Figure 9: Wealth Process for pairs trading strategy with the following parameters

Allocation Strategy	Open Threshold	Close Threshold	Stop Loss Threshold	Restrict Trading to Once a day?
CAPM β	1.5σ	0.5σ	4σ	No

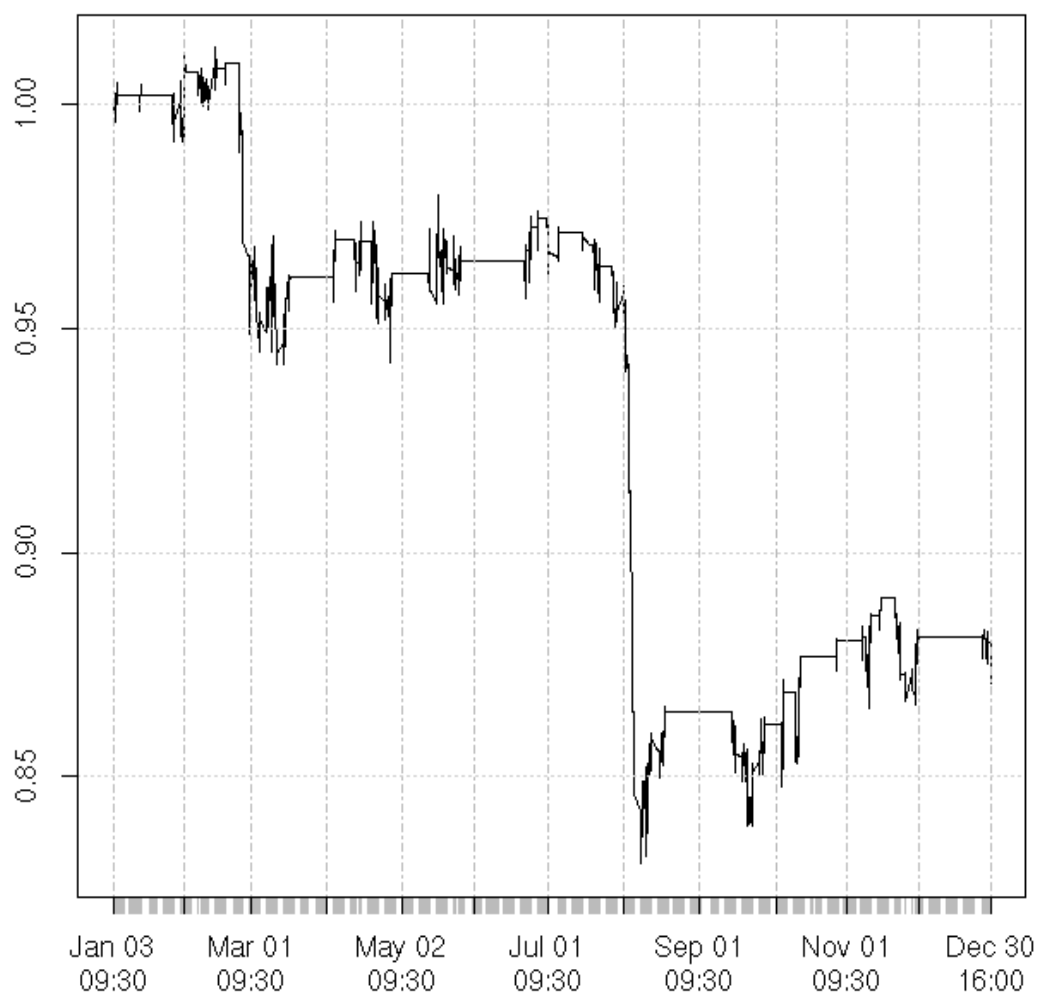


Figure 10: Wealth process with 15bp transaction cost

Allocation Strategy	Open Threshold	Close Threshold	Stop Loss Threshold	Transaction Cost
CAPM β	1σ	0.5σ	4σ	15 bp

Appendix II: Code Implementation for pair selection and compatibility testing

```

setwd("~/r/")
require(xts)
require(fUnitRoots)
x11();
tickers=c("XOM", "CVX")
startYear="2004"
endYear="2005"
dateRange=paste(startYear,endYear,sep="/");
ratioChoice=2; #think of choices as c("1:1","capm","coint")
dollarRatio=FALSE; #does ratio refer to dollar value or units? (should only really be true for 1:1 ratio)
logSpread=TRUE; #are we looking at the prices or the log of the prices to determine when to trade?
openMult=1.0;
closeMult=0.5;
#study (2,1.5), (3,2.5), (2,0.5), (3,0.5)
stopLossMult=4.0;

s1daily<-as.xts(read.zoo(file=paste(tickers[1],"Daily.csv",sep=""),head=TRUE,sep=",",format="%Y%m%d"))

#must divide raw cvx data before and including sept 10 2004 by 2 due to 1:2 stock split for cvx- (holding period
  returns are fine)
#this adjustment was done using excel

s2daily<-as.xts(read.zoo(file=paste(tickers[2],"Daily.csv",sep=""),head=TRUE,sep=",",format="%Y%m%d"))

spydaily<-as.xts(read.zoo(file="SPYDaily.csv",head=TRUE,sep=",",format="%Y%m%d"))
shydaily<-as.xts(read.zoo(file="SHYDaily.csv",head=TRUE,sep=",",format="%Y%m%d"))

##CAPM##
require(PerformanceAnalytics)
s1Beta=CAPM.beta(s1daily[dateRange]$RET,spydaily[dateRange]$RET, shydaily[dateRange]$RET)
s2Beta=CAPM.beta(s2daily[dateRange]$RET,spydaily[dateRange]$RET, shydaily[dateRange]$RET)
cat(paste(tickers[1],"beta: "),s1Beta,"\n")
cat(paste(tickers[2],"beta: "),s2Beta,"\n")
capmRatio<-s2Beta/s1Beta
cat(paste("Ratio of betas (",paste(tickers[2],tickers[1],sep="/")," )",capmRatio,"\n"));

#some guidance from http://quanttrader.info/public/testForCoint.html
#take logs of prices as per Vidyamurthy 2004
s2dailyprice<-(s2daily$PRC)
s1dailyprice<-(s1daily$PRC)
combined=merge(s1dailyprice,s2dailyprice)
names(combined)<-c("s1","s2")
#log prices for insample
insample<-log(combined[dateRange])
#no y intercept, so +0 is required
lsr1<-lm(s1~s2+0,data=insample)
cointRatio<-coef(lsr1)[1]
spread<-insample[, "s1"]-cointRatio*insample[, "s2"]
statTest <- adfTest(spread,type="nc")

cat("p-value of ADF test for pair is ",statTest@test$p.value,"\n")

cat("The value for cointegration coefficient used is ",cointRatio,"\n")

#create trading design
ratioVec=c(1,capmRatio,cointRatio); #amt of 2 to transact per unit of 1 use this to construct signal (defined as
  deviation from historical mean of signal)
hedgeRatio=ratioVec[ratioChoice];
if(logSpread){
  insample<-log(combined[dateRange])
  spread<-insample[, "s1"]-hedgeRatio*insample[, "s2"]
  signalMean=apply(spread,2,mean)
  signalDev=apply(spread,2,sd)
  getSignal<-function(s1.price,s2.price){

```

```

        signal=log(s1.price)-hedgeRatio*log(s2.price)-signalMean;
        return(signal);
    }
} else {
    insample<-combined[dateRange]
    spread<-insample[, "s1"]-hedgeRatio*insample[, "s2"]
    signalMean=apply(spread,2,mean)
    signalDev=apply(spread,2,sd)
    getSignal<-function(s1.price,s2.price){
        signal=(s1.price)-hedgeRatio*(s2.price)-signalMean;
        return(signal);
    }
}

par(mfrow=c(2,1));
ts.plot(insample,col=c(1:2))
title(paste("Price process of ",tickers[1],"(Black) and",tickers[2],"(Red)"))

#openSignal
openSignal=signalDev*openMult;
#closeSignal
closeSignal=signalDev*closeMult;
#stopLossSignal
stopLossSignal=signalDev*stopLossMult;
plot(spread-signalMean);

abline(h=openSignal,col=4,lty=2)
abline(h=-openSignal,col=4,lty=2)

abline(h=closeSignal,col=2,lty=2)
abline(h=-closeSignal,col=2,lty=2)

abline(h=stopLossSignal,col=3,lty=2)
abline(h=-stopLossSignal,col=3,lty=2)

cat("Stdev (sigma) for signal selected is ",signalDev,"\n");
cat("\tOpen at ",openMult,"sigma \n");
cat("\tClose at ",closeMult,"sigma \n");
cat("\tStop loss (close) at ",stopLossMult,"sigma \n");

#Extract p-value of beta calculation

beta.model=function (Ra, Rb, Rf = 0)
{
    Ra = checkData(Ra)
    Rb = checkData(Rb)
    if (!is.null(dim(Rf)))
        Rf = checkData(Rf)
    Ra.ncols = NCOL(Ra)
    Rb.ncols = NCOL(Rb)
    xRa = Return.excess(Ra, Rf)
    xRb = Return.excess(Rb, Rf)
    pairs = expand.grid(1:Ra.ncols, 1:Rb.ncols)
    merged = as.data.frame(na.omit(merge(xRa, xRb)))
    model.lm = lm(merged[, 1] ~ merged[, 2], merged);
    return(model.lm);
}

cat("Beta of stock 1 summary:")
summary(beta.model(s1daily[dateRange]$RET,spydaily[dateRange]$RET, shydaily[dateRange]$RET))

cat("Beta of stock 2 summary:")
summary(beta.model(s2daily[dateRange]$RET,spydaily[dateRange]$RET, shydaily[dateRange]$RET))

```

Appendix III: Code implementation of trading strategy

```

openMult=1.0;
closeMult=0.5;
stopLossMult=4.0;
ratioChoice=1; #think of choices as c("1:1","capm","coint")
dollarRatio=FALSE; #does ratio refer to dollar value or units? (should only really be true for 1:1 ratio)
require(RTAQ)
print(paste("Start time:",Sys.time()));
startTime=Sys.time();
from=timeDate("2005-01-01");
to=timeDate("2005-12-31");
frq=5; #in minutes
frqInt="mins";
#tickers=c("XOM","CVX")
recomputeDays=FALSE;
closeAtDayEnd=FALSE;
rolling=TRUE;
modifyCoInt=TRUE; #make sure to run dailyscript if this is set to false
onceADay=TRUE;
tc=0.005 #15bp is studied in http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1611623
maxMargin=-20; #should be at most 0, negative if having up to negative maxMargin balance on your books is allowed
entryMargin=0.0; #what margin can you take on while opening position
initialWealth=1;

processAttribs=c("Signal", "Bid1", "Of1", "Prc1", "Bid2", "Of2", "Prc2", "SignalDev");
pAttribs=c(paste("Position ", tickers[1]), paste("Position", tickers[2]), "Wealth", "NetWealth");

datasource=datadestination=~"/r";
#do you want to reconvert TAQ data into Rdata? true if yes, false otherwise
reconvert=TRUE;
clean=TRUE;
recomputeSignal=TRUE;

closePosition<-function(stratVec,dataVec)#pass stratVec at time just previous to current and dataVec at time current
returns stratVec at time current stratVec contains [pos1,pos2,wealth], dataVec
contains[bid1,of1,prc1,bid2,of2,prc2,sigal]
{
  toReturn<-c(0,0,0);
  wealthIncr<-0;
  transSize<-0;
  if(stratVec[1]<0){ #originally short stock 1
    wealthIncr<-wealthIncr+stratVec[1]*dataVec[2]#will pay ask for stock 1
    transSize<-transSize+abs(stratVec[1]*dataVec[2])
    wealthIncr<-wealthIncr+stratVec[2]*dataVec[4]#will get bid for stock 2
    transSize<-transSize+abs(stratVec[2]*dataVec[4])
  }else if(stratVec[1]>0){ #originally long stock 1
    wealthIncr<-wealthIncr+stratVec[1]*dataVec[1]#will get bid for stock 1
    transSize<-transSize+abs(stratVec[1]*dataVec[1])
    wealthIncr<-wealthIncr+stratVec[2]*dataVec[5]#will pay ask for stock 2
    transSize<-transSize+abs(stratVec[2]*dataVec[5])
  }else{ #not in a position
    #do nothing, defaults to previous position
  }
  toReturn[1]=0;
  toReturn[2]=0;
  toReturn[3]=stratVec[3]+wealthIncr-transSize*(tc);
  return(toReturn);
}

openPosition<-function(stratVec,dataVec)#pass stratVec at time just previous to current and dataVec at time current
returns stratVec at time current stratVec contains [pos1,pos2,wealth], dataVec
contains[bid1,of1,prc1,bid2,of2,prc2,sigal]
{
  toReturn=stratVec;
  transSize=0;
  if(stratVec[1]==0 && stratVec[2]==0){ #we're not currently in a position

```

```

if(!dollarRatio){ #hedgeRatio refers to units

    if(dataVec[7]>0){#s1 is overpriced- short s1, long s2
        toReturn[1]=-1;
        toReturn[2]=hedgeRatio;
        payoff=dataVec[1]-hedgeRatio*dataVec[5];#get 1 unit of s1 at bid,pay hedgeRatio units of s2 at ask
        transSize=abs(dataVec[1])+abs(hedgeRatio*dataVec[5]);
        if(payoff<entryMargin){ #scale ratio so payoff of opening position is more than or equal to
            newHR=(dataVec[1]-entryMargin)/dataVec[5];
            toReturn[2]=newHR;
            payoff=dataVec[1]-newHR*dataVec[5];#get 1 unit of s1 at bid,pay hedgeRatio units of s2 at
            transSize=abs(dataVec[1])+abs(newHR*dataVec[5]);#get 1 unit of s1 at bid,pay hedgeRatio
        }
        #scale down so you're short 1 dollar
        toReturn[1]=toReturn[1]/dataVec[1];
        toReturn[2]=toReturn[2]/dataVec[1];
        payoff=payoff/dataVec[1];
        transSize=transSize/dataVec[1];

        toReturn[3]=stratVec[3]+payoff-transSize*(tc);#get 1 unit of s1 at bid,pay hedgeRatio units of s2

    }else{#s2 is overpriced
        toReturn[1]=1
        toReturn[2]=-1*hedgeRatio
        payoff=-dataVec[2]+hedgeRatio*dataVec[4];#pay 1 unit of s1 at ask, get hedgeRatio units of s2 at
        transSize=abs(dataVec[2])+abs(hedgeRatio*dataVec[4]);
        if(payoff<entryMargin){ #scale ratio so payoff of opening position is more than or equal to
            newHR=(entryMargin+dataVec[2])/dataVec[4];
            toReturn[2]=newHR*-1;
            payoff=-dataVec[2]+newHR*dataVec[4];#get 1 unit of s1 at bid,pay hedgeRatio units of s2 at
            transSize=abs(dataVec[2])+abs(newHR*dataVec[4]);
        }
        #scale down so you're long 1 dollar
        toReturn[1]=toReturn[1]/dataVec[2];
        toReturn[2]=toReturn[2]/dataVec[2];
        payoff=payoff/dataVec[2];
        transSize=transSize/dataVec[2];

        toReturn[3]=stratVec[3]+payoff-transSize*(tc);
    }
}

}else{
    if(dataVec[7]>0){#s1 is overpriced- short s1, long s2
        toReturn[1]=-1/dataVec[1];#short one dollar of s1 at bid
        toReturn[2]=hedgeRatio/dataVec[5]; #long hedgeRatio dollars of s2 at ask
        payoff=1-hedgeRatio;#get 1 dollar from shorting s1,pay hedgeRatio dollars to go long s2
        transSize=abs(1)+abs(hedgeRatio);
        if(payoff<entryMargin){
            newHR=1-entryMargin;
            toReturn[2]=newHR/dataVec[5];
            payoff=1-newHR;
            transSize=abs(1)+abs(newHR);
        }
        toReturn[3]=stratVec[3]+payoff-transSize*(tc);
    }else{#s2 is overpriced
        toReturn[1]=1/dataVec[2];#long one dollar of s1 at ask
        toReturn[2]=-hedgeRatio/dataVec[4]; #short hedgeRatio dollars of s2 at bid
        payoff=-1+hedgeRatio;#lose 1 dollar from shorting s2,get hedgeRatio dollars from going long s2
        transSize=abs(-1)+abs(hedgeRatio);
        if(payoff<entryMargin){
            newHR=entryMargin+1;

```

```

        toReturn[2]=newHR/dataVec[4];
        payoff=-1+newHR;
        transSize=abs(-1)+abs(newHR);
    }
    toReturn[3]=stratVec[3]+payoff-transSize*tc;
}

}

}else{#we are in a position already
    #do nothing
}
return(toReturn);
}

if(reconvert){
    print("Converting .csv files...");
    tempTime=Sys.time();
    convert(from,to,datasource,datadestination,trades=FALSE,quotes=TRUE,ticker=tickers,dir=FALSE,extention="csv",header=FALSE,
    quotecolnames=c("SYMBOL","DATE","TIME","BID","OFR","BIDSIZ","OFRSIZ","MODE","EX"),format="%Y%m%d %H:%M:%S");
    print("Conversion complete");
    print(Sys.time()-tempTime);
}

if(recomputeDays){
    inSampleDays=timeSequence(from, to,by="days");
    inSampleDays=inSampleDays[isBizday(inSampleDays, holidays = holidayNYSE(1950:2030))];

    #add a day to to since we want to go up to and including last the date described by to, not just to midnight on to
    inSampleMins=(timeSequence(from, to+24*60*60-1,by=paste(frq,frqInt)));
    inSampleMins=inSampleMins[strftime(inSampleDays)];
    dayVec=strftime(inSampleDays)
}

if(clean){
    #clean quotes on disk
    print("Starting to clean data");
    tempTime=Sys.time();
    fromToArray=array(dim=c(length(dayVec),2))
    fromToArray[,1]=dayVec
    fromToArray[,2]=dayVec

    cleanDay<-function(x,datasource,datadesitnation,tickers,exchange){
        quotesCleanup(from=x,to=x,datasource,datadestination,ticker=tickers,exchanges=rep(exchange,length(tickers)));
    }
    sapply(dayVec,cleanDay,datasource,datadestination,ticker=tickers,"N");
    print("Data cleaning completed");
    print(Sys.time()-tempTime);
}

if(recomputeSignal){
    inxts=xts(x=array(NA,c(length(inSampleMins),length(processAttribs))),order.by=inSampleMins);
    colnames(inxts)=processAttribs;
    inxts=exchangeHoursOnly(inxts);
    #head(inxts);
    #tail(inxts);

    print("Entering loop...");
    print(paste("Time:",Sys.time()));
    tempTime=Sys.time();
    active=TRUE; #is strategy currently active?

    timePoint=NA;
    prevPoint=NA;

    for(day in dayVec){
        if(active){
            #print(paste("Processing day",day));
            if(rolling){

                #Calculate strategy parameters on a rolling basis

```



```

#go back 50 days
openWindow=timeDate(day)-51*24*60*60;
closeWindow=timeDate(day)-1*24*60*60;
dateRange=paste(openWindow,closeWindow,sep="/");
##CAPM##
require(PerformanceAnalytics)
s1Beta=CAPM.beta(s1daily[dateRange]$RET, spydaily[dateRange]$RET, shydaily[dateRange]$RET)
s2Beta=CAPM.beta(s2daily[dateRange]$RET, spydaily[dateRange]$RET, shydaily[dateRange]$RET)
capmRatio<-s2Beta/s1Beta

#some guidance from http://quanttrader.info/public/testForCoInt.html
#take logs of prices as per Vidyamurthy 2004
s2dailyprice<-(s2daily$PRC)
s1dailyprice<-(s1daily$PRC)
combined=merge(s1dailyprice,s2dailyprice)
names(combined)<-c("s1","s2")
#log prices for insample
insample<-log(combined[dateRange])
#no y intercept, so +0 is required
lsrl<-lm(s1~s2+0,data=insample)
if(modifyCoInt){
  coIntRatio<-coef(lsrl)[1];
}
spread<-insample[, "s1"]-coIntRatio*insample[, "s2"];
#print(coIntRatio);
#create trading design
ratioVec=c(1,capmRatio,coIntRatio); #amt of 2 to transact per unit of 1 use this to construct signal (defined as
deviation from historical mean of signal)
hedgeRatio=ratioVec[ratioChoice];
if(logSpread){
  insample<-log(combined[dateRange])
  spread<-insample[, "s1"]-hedgeRatio*insample[, "s2"]
  signalMean=apply(spread,2,mean)
  signalDev=apply(spread,2,sd)
  getSignal<-function(s1.price,s2.price){
    signal=log(s1.price)-hedgeRatio*log(s2.price)-signalMean;
    return(signal);
  }
}else{
  insample<-(combined[dateRange])
  spread<-insample[, "s1"]-hedgeRatio*insample[, "s2"]
  signalMean=apply(spread,2,mean)
  signalDev=apply(spread,2,sd)
  getSignal<-function(s1.price,s2.price){
    signal=(s1.price)-hedgeRatio*(s2.price)-signalMean;
    return(signal);
  }
}

}

s1Raw.ts=TAQLoad(tickers[1],day,day,trades=FALSE,quotes=TRUE,datasource);
s2Raw.ts=TAQLoad(tickers[2],day,day,trades=FALSE,quotes=TRUE,datasource);
if(day=="2005-11-25" || day=="2006-11-24" || day=="2006-07-03" ){
  s1Raw.ts=aggregateQuotes(s1Raw.ts,on=frqInt,k=frq,marketclose="13:00");
  s2Raw.ts=aggregateQuotes(s2Raw.ts,on=frqInt,k=frq,marketclose="13:00");
}else{
  s1Raw.ts=aggregateQuotes(s1Raw.ts,on=frqInt,k=frq);
  s2Raw.ts=aggregateQuotes(s2Raw.ts,on=frqInt,k=frq);
}

dayTimes=NA;
dayTimes1=time(s1Raw.ts);
dayTimes2=time(s2Raw.ts);

dayTimes=timeDate(intersect(strftime(dayTimes1),strftime(dayTimes2)));

s1Raw.ts=s1Raw.ts[dayTimes];
s2Raw.ts=s2Raw.ts[dayTimes];

```

```

s1.bidask=merge((s1Raw.ts[,3]),(s1Raw.ts[,5]))
s1.bidask=xts(sapply(s1.bidask,as.numeric),dayTimes);
s1.prices=xts(rowMeans(s1.bidask),dayTimes);
s1.ts=merge(s1.bidask,s1.prices);

s2.bidask=merge((s2Raw.ts[,3]),(s2Raw.ts[,5]))
s2.bidask=xts(sapply(s2.bidask,as.numeric),dayTimes);
s2.prices=xts(rowMeans(s2.bidask),dayTimes);
s2.ts=merge(s2.bidask,s2.prices);

colnames(s1.ts)=c(sapply(c("BID", "OFR", "PRC"),paste,tickers[1]));
colnames(s2.ts)=c(sapply(c("BID", "OFR", "PRC"),paste,tickers[2]));

signal.ts=xts(getSignal(s1.prices[,1],s2.prices[,1]),dayTimes);
colnames(signal.ts)=c("SIGNAL")
combined.ts=merge(s1.ts,s2.ts,signal.ts);
dayTimes=strftime(time(combined.ts));

for(timePoint in dayTimes){ #timePoint is a character class object since dayTimes must be a vector for us to
loop on it

    timePoint=timeDate(timePoint); #so we make it a timeDate object before proceeding
    inxts[timePoint,"Signal"]=combined.ts[timePoint,"SIGNAL"];
    inxts[timePoint,"Bid1"]=combined.ts[timePoint,paste("BID",tickers[1],sep=".")];
    inxts[timePoint,"Ofr1"]=combined.ts[timePoint,paste("OFR",tickers[1],sep=".")];
    inxts[timePoint,"Prc1"]=combined.ts[timePoint,paste("PRC",tickers[1],sep=".")];
    inxts[timePoint,"Bid2"]=combined.ts[timePoint,paste("BID",tickers[2],sep=".")];
    inxts[timePoint,"Ofr2"]=combined.ts[timePoint,paste("OFR",tickers[2],sep=".")];
    inxts[timePoint,"Prc2"]=combined.ts[timePoint,paste("PRC",tickers[2],sep=".")];
    inxts[timePoint,"SignalDev"]=signalDev;

    prevPoint=timePoint;
}

}else{
    break;
}
}

inxts=inxts[!is.na(inxts$Signal)];
}

print("Construction of signal series complete");
x11();
par(mfrow=c(2,1));
plot(inxts[, "Signal"]);

abline(h=openSignal,col=4,lty=2)
abline(h=-openSignal,col=4,lty=2)

abline(h=closeSignal,col=2,lty=2)
abline(h=-closeSignal,col=2,lty=2)

abline(h=stopLossSignal,col=3,lty=2)
abline(h=-stopLossSignal,col=3,lty=2)

tempModel=arima(inxts[, "Signal"],c(0,1,0));
acf(tempModel$resid);

times=time(inxts);
process=xts(x=array(0,c(length(times),length(pAttribs))),order.by=times);
colnames(process)=pAttribs;

process[1,"Wealth"]=initialWealth;
process[1,"NetWealth"]=initialWealth;

```

```

prevPoint=NA;
active=TRUE;
for(day in dayVec){
  print(paste("Processing strategy on ",day));
  dayTimes=strftime(time(inxts[day])); #get all times from day
  #get signalDev from day

  for(timePoint in dayTimes){
    if(active){
      timePoint=timeDate(timePoint);

      signalDev=inxts[timePoint,"SignalDev"];
      #openSignal
      openSignal=signalDev*openMult;
      #closeSignal
      closeSignal=signalDev*closeMult;
      #stopLossSignal
      stopLossSignal=signalDev*stopLossMult;
      #
      if(is.na(prevPoint)){ #skip first time point
      }else{
        curData=inxts[timePoint];
        dVec=curData[1,c("Bid1","Of1r1","Prc1","Bid2","Of1r2","Prc2","Signal")]; #dVec
        contains[bid1,of1r1,prc1,bid2,of1r2,prc2,signal]
        if((timePoint-timeDate(last(dayTimes))==0 && closeAtDayEnd) ||
        abs(curData[timePoint,"Signal"]<=closeSignal || abs(curData[timePoint,"Signal"]>=stopLossSignal){ #close if last
        time point of day and we are closing at day end, or if signal has crossed or is below the closeSignal, or if signal
        has crossed or is above the stopLossSignal
          #close
          temp=closePosition(as.vector(process[prevPoint]),as.vector(dVec));
        }else if(abs(curData[timePoint,"Signal"]>=openSignal){
          temp=openPosition(as.vector(process[prevPoint]),as.vector(dVec))#open position
        }else{
          temp=c(process[prevPoint,1],process[prevPoint,2],process[prevPoint,3],process[prevPoint,4]);
        }
        process[timePoint,1]=temp[1];
        process[timePoint,2]=temp[2];
        process[timePoint,3]=temp[3];

        process[timePoint,4]=process[timePoint,1]*dVec[1,3]+process[timePoint,2]*dVec[1,6]+process[timePoint,3];
      }

      if(process[timePoint,3]<maxMargin){#check to see if we're still active, inactivate and break
        active=FALSE;
        break;
      }

      prevPoint=timePoint;
      if(onceADay){
        break;
      }
    }else{
      break;
    }
  }
}

curData=inxts[timePoint];
dVec=curData[1,c("Bid1","Of1r1","Prc1","Bid2","Of1r2","Prc2","Signal")]; #dVec
contains[bid1,of1r1,prc1,bid2,of1r2,prc2,signal]
temp=closePosition(as.vector(process[prevPoint]),as.vector(dVec));
process[timePoint,1]=temp[1];
process[timePoint,2]=temp[2];
process[timePoint,3]=temp[3];
process[timePoint,4]=process[timePoint,1]*dVec[1,3]+process[timePoint,2]*dVec[1,6]+process[timePoint,3];

```

```

if(onceADay){
  process=subset(process,NetWealth!=0)
}

x11()
plot(process[, "NetWealth"]);
print("Evaluation of Strategy Complete");

print(paste("End time:", Sys.time()));
print("Total time taken:");
print(Sys.time()-startTime);

#load("rollingProcess.RData")
initialWealth=1;

rollingDaily=to.daily(process[, "NetWealth"])[,4];
rollingReturns=Return.calculate(rollingDaily);
rollingReturns[is.na(rollingReturns)]<-0;

dRange=paste("2005-01-01", "2005-12-31", sep="/");

rollingReturns=xts(as.vector(rollingReturns[,1]),order.by=time(shydaily[dRange]))

print("Metrics for Strategy");
print(paste("Annual return", as.vector((last(process[, "NetWealth"])-initialWealth)/initialWealth)
))
print(SharpeRatio.annualized(rollingReturns, Rf=mean(shydaily[dRange]$RET)))
vDrag=mean(rollingReturns)-1/2*var(rollingReturns)
print(paste("Volatility Drag adjusted returns(daily):", (1+vDrag)^252-1))
print(VaR(rollingReturns));

```