High Frequency Statistical Arbitrage with Kalman Filter and Markov Chain Monte Carlo

by

Han Xu

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2017

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Statistical arbitrage, or sometimes called pairs trading, is an investment strategy which exploits the historical price relationships between two or several assets and profits from relative mispricing. It has a long history in hedge fund industry and variates of this kind of strategies are still profitable nowadays. The idea is simple and the source of the profit has support from fundamentals in economics and pricing theories. However, there are still many difficulties in implementing and testing such strategies in real life, which include how to select pairs, how to estimate hedge ratio, when to enter, when to exit and etc. Due to its proprietary nature, there is very few literature on this subject. This thesis is an attempt to demystify statistical arbitrage in high-frequency settings, using freely available data of Chinese commodity futures. This thesis introduces and discusses the existing research done on this subject. Also, with the help of advanced statistical inference approaches for treating time series, this thesis proposed a new model which generalizes the entire process of creating a profitable statistical arbitrage trading strategy for a given market. Several different approaches are implemented and their simulated performances in the Chinese commodity future market are compared horizontally. Unlike much other existing literature, transaction costs and market frictions have been considered thoroughly in order to make the research result more meaningful. Empirical results show that our new model delivers very competitive performance in online hedge ratio estimation.

Acknowledgements

First and foremost, I would like to thank my advisor Professor Liang-Liang Xie for his invaluable guidance, generous support, and continuous encouragement during my masters program.

Second, I would like to thank the readers of this thesis Zhou Wang and Pin-Han Ho, for taking their precious time to read my thesis and providing constructive advices.

Last but not least, I am deeply indebted to my friends and families, for their love and support.

Dedication

This is dedicated to the one I love.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ADF Augmented Dickey Fullers 8, 16, 31, 42

APT Arbitrage Pricing Theory 4, 7, 16, 17, 30, 31

ARMA Autoregressive-Moving-Average 16

CAPM Capital Asset Pricing Model 6, 7

DESP Double Exponential Smoothing-based Prediction 18, 32

FIA Futures Industry Association 4

IR Information Ratio 35, 36, 38, 40–43, 45

MCMC Markov chain Monte Carlo 2, 5, 20–23, 28, 32, 33, 36, 44, 48, 49

OLS Ordinary Least squares 9, 18, 28, 31, 32, 36, 40

OTC Over the Counter 11

VWAP Volume Weighted Average Price 34

Chapter 1

Introduction

1.1 Overview

In the financial world there are two ways to make profit, one is to long an asset and sell it later at a higher price, the other is to short an asset in advance and buy it back at a lower price. While losses occur when things happen the other way around. Speculators are those investors who believe in their ability to predict the market movement in certain scenarios and hence make money by betting on the right side. These predictions may be generated from various sources, for e.g. it can be fundamental analysis of the underlying companys financial report, the business cycle, influences of government policies; it can be technical analysis of historical changes in price and trading volume of the asset; it can also be sophisticated mathematical models which exploit high- frequency data such as transaction histories and order book imbalances. Pairs trading is a quantitative method of speculation which has a long history that can be dated back to 1980s in Wall Street [12]. The basic idea is to find a pair of assets whose prices move together. Usually, the pair of assets has some sort of underlying economic link. Profit can be made from the market misprice by getting long position on the undervalued asset and short position on the overvalued one, and then unwind the positions upon the convergence of their spread, a measure of their relative misprice. The economic link between the two assets could be weak and unstable overtime, which essentially separates statistical arbitrage from pure arbitrage. In pure arbitrage, the profit comes from mispricing of a pair of identical assets in different markets or different forms. The nature of arbitrage makes it an almost risk-free but also a very competitive strategy which relies much on execution speed of the trading system. Statistical arbitrage, on the other hand, is sometimes known as pairs trading. The profitability of this strategy highly depends on the statistical modelling and prediction of the spread time series.

Due to proprietary nature of hedge fund industry, there are very limited published work as well as many myths on this subject. Thus, the goal of this thesis is to summarize a handful of influential work on the matter of statistical arbitrage and use mathematical formulation to get rid of the myths that may confuse newbies on this subject. The thesis also proposes an innovative way of modelling the relative misprice or spread with the help of Kalman filter and MCMC model. It then compares the performance of different modelling and trading strategies on a unified dataset - Chinese commodity future. In order to define the underlying problems and the scope of this research better, the following section will explain some of the essential concepts on this subject.

1.2 Problems, Motivations and Scope

As mentioned before, the goal of investment is to buy the undervalued asset, sell the overvalued asset and make profits from that. It will be the best if we as investors know the true value of the asset. Actually, there are lots of analysts working so hard day and night trying to estimate individual company's true value by estimating its future income or make a guess on how much a commodity should worth by valuing its supply and demand. However, this kind of processes usually include tons of assumptions and are usually subjective. It is almost always the case that defining and computing the true value of an asset is extremely hard. The idea behind statistical arbitrage is to use relative pricing to alleviate this dilemma. Consider an example of two commodities X and Y, whose prices are determined by the same economical factors but are traded in different markets. Say that the prices of the two commodities always move together and the returns are exactly the same. Define prices of commodities X and Y at time t to be p_t^X and p_t^Y, respectively, then the returns in a period of time i are $(p_{t+i}^X - p_t^X)/p_t^X$ and $(p_{t+i}^Y - p_t^Y)/p_t^Y$, we have:

$$\forall t \; and \; \forall i, \; (p_{t+i}^X - p_t^X)/p_t^X = (p_{t+i}^Y - p_t^Y)/p_t^Y$$

and since:

$$log(\theta + 1) \approx \theta \; as \; \theta \to 0$$

therefore when return is small, we have, in log price:

$$\forall t \; and \; \forall i, \; logp_{t+i}^X - logp_t^X = logp_{t+i}^Y - logp_t^Y$$

$$\Leftrightarrow logp_{t+i}^X - logp_{t+i}^Y = logp_t^X - logp_t^Y = c$$

where c is a constant

In other words, the spread or the price difference between the two assets from time to time is always the same. However, in real market situations, due to market inefficiencies and mispricings, the relationship can be as following:

$$logp_t^X - logp_t^Y = \xi_t \tag{1.1}$$

where ξ_t is a stationary time series that exhibits mean reversion property. The equation 1.1 suggests that the spread between X and Y are oscillating around an equilibrium. Thus profits can be made by exploiting the deviations from the equilibrium, by shorting the overvalued asset and longing the undervalued asset at the same time, then unwind positions when equilibrium is met again. In research done by Gatev et al [5] and Nath [10], a $1:1$ ratio of returns for the pair was a presumption in pairs selection stage. In fact, for pairs trading to work, it is unnecessary for X and Y to have the $1:1$ ratio of returns. If X and Y have $b:c$ ratio of return in equilibrium, traders can simply short/long c dollar of X and concurrently long/short b dollar of Y to get the same hedging effect. This $b:c$ ratio is known as hedge ratio. The whole idea behind this is to construct a stationary time series from two non-stationary ones. Traders create such a pair and treat it as trading one asset. It was Engle and Granger [4] who first discovered that the linear combination of two non-stationary time series can be stationary. They proposed a statistical approach named cointegration for testing such property, which helped them won the Nobel Prize economics in 2003. A short version of their two-step approach is: first perform linear regression on the two time series and find the hedge ratio as well as estimated residual time series, and second use Augmented Dicker-Fuller test for testing the stationarity of the estimated residual. Vidyamurthy [12] adopts this approach in his pairs trading strategy and formulated an explanation of why it might work in terms of fundamental insights rather than pure statistical results, using APT[11]. However, there are several problems to consider in Vidyamurthys work, which will be addressed in detail in later chapters.

The objective of this research is to address the issues with existing trading models and propose a new one. The marvellous thing about developing a statistical arbitrage strategy quantitatively is that the underlying principle is the same across various markets and assets. There is no need to search for predictive indicators extensively as we may need in order to price movement of one particular financial instrument. Instead, we search for pairs that their prices tend to move together and then apply a highly disciplined approach for designing trading strategy. The potential profit solely comes from relative mispricing of the pair due to market inefficiency. In other words, a successful pairs trading approach that works in one market can be easily adjusted and applied to another market, even if the underlying financial instruments are very different. There are three reasons for choosing Chinese future markets in this thesis study. Firstly, according to the 2016 annual volume survey provided by FIA, future trading in China plays an important role in the global market. The trading volume of future contracts in China is on an upward trend. "Trading on the Shanghai Futures Exchange rose 60.0% to 1.68 billion contracts, making it the sixth largest exchange in the world, while trading on the Dalian Commodity Exchange rose 37.7% to 1.54 billion contracts, eighth in the world." [1] Secondly, unlike European and US future markets which have a few centuries of history, China had its first modern future exchange in 1991. Although the number of institutional investors is increasing rapidly over the recent years, retail investors still play the major role in Chinese future market. According to a survey done by Shanghai Future exchange in 2014, institutional investors only contributes 25.87% of trading volume in Chinese future markets [6]. Retail investors usually do not have the resource to react to the market as quickly as institutional investors. Therefore intuitively there shall be more chances to exploit market inefficiencies. Although Chinese stock markets have similar properties and much larger trading volume, due to government regulations that enforce T+1 trading policy, it is impossible to open a new position and close it within one trading day. Regulations also prohibit short selling stocks, which makes the realization of pairs trading even more difficult. Last but not least, one major reason for choosing to perform analysis on Chinese future markets in this study is the data accessibility. Companies such as Ricequant and DataYes have made their high-quality minute bar data freely available online for researchers. All data used for analysis in this thesis are provided by Ricequant.

1.3 Outline

In the first chapter, we will discuss some preliminary materials for statistical arbitrage, including some widely accepted pricing models, some essential concepts in time series as well as an introduction to Kalman filtering. We then introduce some basic knowledge in trading future contracts, plus where market frictions lie from the perspective of a modern order matching system. Reviews on related work are in the last section of the first chapter. In the second chapter, we mainly introduce our modified model for online estimation of hedge ratio using Kalman filter. We will also introduce MCMC and describe how it can be used for parameter estimation in our model. Based on the math of our model we then propose a new profitability measure for selecting pairs. The third chapter consists of details on data preparation and trading algorithm implementation for our simulations. All the parameter settings and assumptions will be discussed thoroughly. Empirical results and analysis will be covered in chapter 4. We will compare the effectiveness of popular indicators for pairs selection. We will then compare the performance of three trading models, using widely accepted measures. Chapter 5 concludes this thesis.

Chapter 2

Preliminaries of Statistical Arbitrage and Futures Trading

2.1 Mathematical Models and Tools for Statistical Arbitrage

In this section, we will introduce some basic yet important models and tools for statistical arbitrage, which will be used throughout this thesis.

2.1.1 General Asset Pricing Models

The CAPM is a widely used model for describing the relationship between systematic risk and the expected return on assets. It states that the expected return of an asset is a linear function of a theoretical risk-free return (government bond for example) and expected market return. The math formulation goes as following:

$$E(R_i) = R_f + \beta_i(E(R_m) - R_f) \tag{2.1}$$

where R_i is the return of asset i, R_f is the theoretical risk-free return, R_m is the market return and β_i describes how sensitive R_i is with the change of R_m. The CAPM model basically asserts that all the return of a particular asset comes from the risk-free rate and the market. If the asset has low risk, the β_i term is small and the expected return is closer to R_f; and if the asset has high risk, it can have potentially higher return compensating the possibility of taking a higher loss. Despite that the simple model fails in numerous empirical tests [9] and have assumptions such as the existence of a theoretical risk-free asset and using market return as the sole risk factor, it remains a popular model due to its simplicity. The CAPM model won William F. Sharpe, Harry Markowitz and Merton Miller the Nobel Memorial Prize in Economics in 1990. It is also where the wall street idioms 'beta' and 'alpha' originates from. [1]

It is natural to extend the CAPM model to account for more factors for pricing assets. Asset returns are said to follow a factor structure if:

$$R_i = a_i + b_{i1}F_1 + b_{i2}F_2 + \cdots + b_{in}F_n + \xi_i \tag{2.2}$$

where a_i is a constant term specific to asset i, F_j is a systematic risk factor, bi_j is the measure of the sensitivity of asset i to factor j, and ξ_i a random noise with mean equal to zero. The factor model enables the possibility of modelling the return of a risky asset with multiple macroeconomic factors.

Stephen Ross et al. proposed APT [11], which states that if asset returns follow factor structure, the expected return of an asset can be described with a combination of various risk factors using a linear model. Formally, it can be expressed as:

$$E(R_i) = R_f + b_{i1}\lambda_1 + b_{i2}\lambda_2 + \cdots + b_{in}\lambda_n \tag{2.3}$$

where R_f is the theoretical risk-free rate as in CAPM model 2.1, λ_j is the risk premium of a particular risk factor j. Vidyamurthy [12] explained the source of profit for pairs trading using the APT model and some other assumptions, which we will discuss in the later section.

[1] Alpha believers think they can have some predictive power on the movement of R_i sometime, and gain an extra profit term α_i which is not related to the market.

2.1.2 Stationarity, Integration and Cointegration

A time series is stationary if the stochastic process for generating the series does not have changing parameters over time. A simple example would be white a noise signal, which draws independently from a normal distribution at each step. Stationarity of a time series can be statistically tested with many approaches, and ADF test is one among widely used ones. Financial forecasters are keen on stationary time series because they are easy to forecast, due to their mean-reverting property. We would like to see the price of an asset to be a stationary series so that we can always buy at low and sell at high as mean reversion happens all the time. However the reality is after so many empirical research, many believe that prices move as a random walk, which is a typical example for mean-drifting, non-stationary time series. Random walk sequence is unpredictable, and here is where integration and cointegration come into place for financial forecasting.

Wold's decomposition theorem [13] states that any stationary time series Y_t can be decomposed into a deterministic component and a stochastic component with moving average representation:

$$Y_t = \sum_{j=0}^{\infty} b_j \varepsilon_{t-j} + \eta_t \tag{2.4}$$

where ε_t is the innovation series, b_j is the moving average weights for the innovation series, and η_t is a deterministic time series. There are several other properties for a stationary series in this representation, and one of them is that the autocorrelation of the series has to decay quickly:

$$\sum_{j=0}^{\infty} b_j^2 = 0 \tag{2.5}$$

An integration of order 0 or $I(0)$ means that the above condition is satisfied with the moving average representation of some time series, which is not necessarily stationary. Since all stationary series has this property but not vice versa, $I(0)$ is a necessary but not sufficient condition for stationarity. A time series X_t is of integration order d or $I(d)$ if and only if $(1 - L)^d X_t$ is $I(0)$, where L is the lag operator such that:

$$L(X_t) = X_{t-1} \tag{2.6}$$

In finance the log price of an asset P_t is often modelled as a random walk process, which is $I(1)$, and the increment of log price at each time step $P_t - P_{t-1}$ is $I(0)$.

Cointegration was initially introduced to solve the problem of spurious correlation which often happens in linear regression. Clive Granger and Robert Engle formalized and coined the term in their influential paper [4]. Cointegration depicts the relationship between two or more time series. A set of time series $(X_{1t}, X_{2t}, X_{3t}, \cdots X_{nt})$ is cointegrated if they are all $I(1)$, plus that there exist a linear combination of them which is $I(0)$. This tool is particularly useful in pairs trading, since if P_t^A, the price of asset A is cointegrated with P_t^B, the price of asset B, it is possible to construct a portfolio $Y_t = P_t^A - P_t^B$ such that Y_t is $I(0)$ and likely to be stationary. In fact when testing for cointegration, the widely used Engle-Granger two-step method first estimate with OLS and perform a test for stationarity on the residual.

2.1.3 Introduction to Kalman Filtering

Kalman filtering, named after one of its primary developer Rudolf E. Klmn, is a Bayesian inference algorithm which produces an estimate of one or more unknown variables, which can be unobservable, based on series of inaccurate measurements over time. The core philosophy behind it is that information is always useful, and several inaccurate measurements can be used to infer an estimate which is better than any of the single measurement alone. For example, to accurately locate a moving cell phone, utilizing solely the GPS signal is usually not enough. This is because the sensor readings are noisy and jump around all the time. However, by also taking into account the law of physics, the distance can be computed by integrating speed over time. Cell phones have accelerometers and gyroscopes which can produce measurements on both speed and direction of movement. Kalman filter combines these measurements and thereby aids to locate a cellphone accurately within few meters of error. Kalman filtering mainly finds its application in navigation and control of vehicles, aircrafts and spacecrafts. In fact, one of the first application of Kalman filter is in the Apollo 11 program which lands spacecraft on the moon. Moreover, it also finds applications in signal processing and econometrics. Mathematically it can be represented by the following equations.

Prediction stage:

$$\hat{x}_k^- = A\hat{x}_{k-1} + Bu_k \tag{2.7}$$

$$P_k^- = AP_{k-1}A^T + Q \tag{2.8}$$

Update stage:

$$K_k = P_k^- C^T (CP_k^- C^T + R)^{-1} \tag{2.9}$$

$$\hat{x}_k = \hat{x}_k^- + K_k(y_k - C\hat{x}_k^-) \tag{2.10}$$

$$P_k = (I - K_k C)P_k^- \tag{2.11}$$

Note that the algorithm follows a two-step process, namely prediction stage and update stage. In equation (2.7), $\hat{x}_{k-1}$ denotes the previous best estimate of state variables, $\hat{x}_k^-$ denotes the prior estimate of state variables before taking into account the new observation, A is the state transition matrix, and the term Bu_k represents an offset term for the transition. Equation (2.8) makes a prediction of the covariance matrix of the state variables, where P_k^- and P_{k-1} represent the prior estimation of the current covariance matrix and previous covariance matrix respectively, and Q is the error term brought by the transition. In the update stage, the algorithm first computes the so-called Kalman gain K_k (2.9), where C is the observation matrix and R is the error term or uncertainty brought by observation. In equation (2.10) the current best estimate of state variables is computed by blending the prior estimate obtained in prediction stage and the residual between observation and expected observation. Finally, equation (2.11) updates the covariance matrix of the state variable, which depicts the confidence of current estimation.

In the plain words, the Kalman filter first produces an estimate of the current state along with uncertainties, via a transition model in the prediction stage. Consider the previous example of locating the cell phone, transition model is the law of physics and velocity readings. In the update stage, an observation is first made, which in the locating cell phone example is the location reading based on GPS signal. Then the new estimated value is updated using a weighted average of results from prediction and observation, with more weight given to the one with higher certainty. In even simpler language, at each time step, the algorithm makes a prediction, take an observation and updates its belief based on prediction and observation values. The algorithm is recursive and runs online. Meaning that it makes estimates solely based on current measurement and state estimation one-step ago, requiring no more past information.

2.2 Basics in Trading Futures

2.2.1 Future Contracts

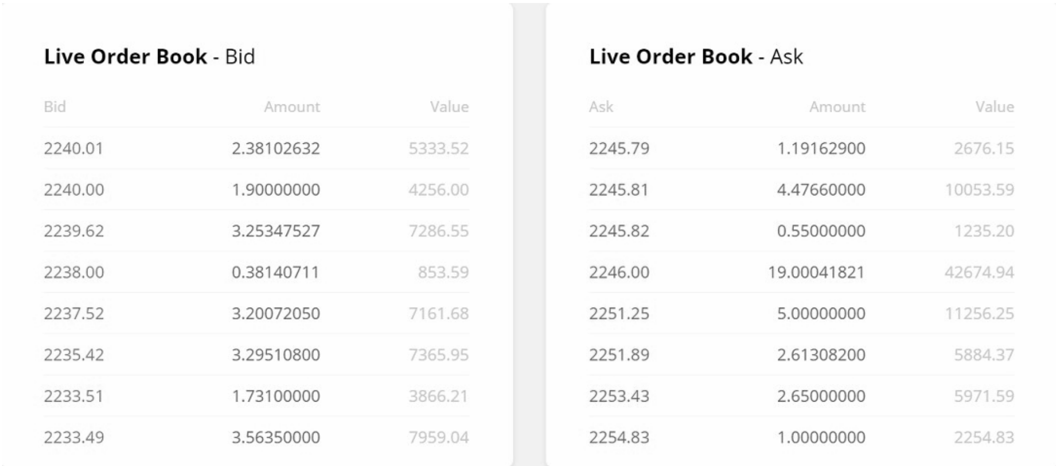
Futures, similar to forward contracts, are financial contracts that obligate a buyer to buy or a seller to sell a certain amount of underlying asset at specified date and price. The difference between futures and forward contracts is that futures are traded in and regulated by future exchanges, while forward contracts are usually agreements between two counterparties and traded in OTC markets. Exchanges also function to standardize tradings in aspects like the quantity and quality for delivery of one single contract, as well as matching deliveries. The underlying asset of a future contract can be delivered physically or settled in cash within the range of legitimate delivery dates, per specified by the contract. The creation of this financial instrument is to facilitate hedgers who want to get rid of the unknown risk of future price change for an asset. For example, in April a bean farmer knows he will produce ten tons of bean by August. However, he does not want to take the risk of price dropping of beans in this period. What he can do is to sell the beans in April with future contract using a locked in price and complete the trade in August. His counterparty, maybe a speculator, bets on the price rise in beans and takes the risk. When a future contract was created, the price may deviate quite a lot from the spot price of the underlying asset due to various reasons. However as time is getting closer to delivery, the future price will converge to the spot price. Otherwise, there is an obvious arbitrage opportunity. It is the possibility of delivery that makes future price eventually converges to spot price. But since the delivery process of a physical asset can be troublesome, only a small portion of future positions eventually ended up with delivery. Since future price will eventually converge to spot price, either hedger or speculator can do their job by closing their positions prior to the delivery date. One important factor that attracts many traders in the futures market is that future contracts are traded with marginal accounts, the future trading is thereby inherently highly leveraged.

2.2.2 Modern Order Matching Systems and Market Frictions

Frictions of the market can hamper a lot of the profit made for a trading strategy, especially in relatively high-frequency settings. It is necessary to include friction estimation in simulation backtest for a trading strategy to make the result meaningful. The two major components of market frictions are transaction fees and slippage due to market illiquidity. Transaction fees required by exchange and brokers are usually constants per trade or a percentage of the total transaction and are therefore easy to model. To understand slippage better it is essential to grasp the idea of modern electronic order matching system and have some knowledge of the microstructure. Unlike open outcry which was popular before 1980s, modern electronic order matching system does not involve shouting and use of hand signals between professionals in the exchange in order to fill a buy or sell order. In this modern system, all legitimate unfilled orders are stored in an electronic list, known as order book. Nowadays most exchanges in the world implement this kind of electronic system so that orders can be passed to exchanges server via a network and matched internally. The order book is organized by price level, showing the number of buy and sell orders at each price level. 2.1 is an illustration of order book.

Bid orders are orders with the intention to buy. At the top is the best bid offer with the highest price. Ask orders are arranged exactly in the opposite way, in which the top ask order is the one intended for sale at the lowest price. The gap between top ask price and top bid price is also sometimes called spread.[2] When there is an order crosses the spread and take the existing offer in the order book, a transaction or trade occurs. Therefore it is

[2]This is different from the spread we talked about in the context of pairs trading.



Live Order Book - Bid

Bid	Amount	Value
2240.01	2.38102632	5333.52
2240.00	1.90000000	4256.00
2239.62	3.25347527	7286.55
2238.00	0.38140711	853.59
2237.52	3.20072050	7161.68
2235.42	3.29510800	7365.95
2233.51	1.73100000	3866.21
2233.49	3.56350000	7959.04

Live Order Book - Ask

Ask	Amount	Value
2245.79	1.19162900	2676.15
2245.81	4.47660000	10053.59
2245.82	0.55000000	1235.20
2246.00	19.00041821	42674.94
2251.25	5.00000000	11256.25
2251.89	2.61308200	5884.37
2253.43	2.65000000	5971.59
2254.83	1.00000000	2254.83

Figure 2.1: An example of modern order matching system (OKcoin.com)

not hard to understand that the gap between ask and bid prices is where market friction come into place. A trader would have to give up at least the spread if she takes an order on one side and unwinds it later on the other side, and such sacrifice is the so-called slippage. Similarly, if the trader crosses the spread and makes a big take order, it will have a big impact on the market and thereby results in greater slippage.

2.3 Related Works

2.3.1 The Distance Measure Approach

Gatev et al. drafted and circulated a distance measure approach for pairs trading in 1999, mimicking hedge funds approach for pairs selection and trade execution in US equity market. In this draft, they performed walk forward tests on their strategy using daily data through the end of 1998. The researchers then finished and concluded their work in 2003, using the old model and the true out-of-sample data in the 1999-2002 period [5]. The average excess return of the fully invested portfolio of top 20 selected pairs was claimed to be 10.4% per annum, with an annual standard deviation of 3.8%. In this type of approach, distance is a measure of co-movement in a pair. Gatev et al. computed distance as the sum of squared difference of the two normalized, cum-dividend prices (cumulative total return with dividend reinvested). Their approach works like this: First, they form pairs over a 12 months formation period by selecting the top 120 pairs with lowest historical distance measures. Then the following 6 months would be used as the trading period, where the trading trigger is set to be two standard deviations of distance estimated in the previous formation period. Opened positions are unwinded if the two prices cross each other in this period. Any unclosed positions by the end of the trading period are considered into gains and losses. The researchers claimed to had all parameters like formation and trading periods, as well as trade triggering point, is arbitrarily chosen, in order to avoid data snooping. Nath [10] proposed another distance measure and trade triggering mechanism and applied the analysis to U.S. Treasury securities market. Instead of normalizing the prices by constructing a cumulative total return index, Nath normalized the prices by subtracting the sample mean and then dividing by the sample standard deviation in the in the 40 days training or formation period. An empirical distribution of price differences was recorded for each pair in the trading universe. Then in the following 40 days trading period, trading signals are triggered by certain percentile in the empirical distribution rather than in units of standard deviations. Positions are unwinded when any of the three situations happen:

- the spread narrows and crosses the median of the empirical distribution
- the spread is widen further to hit a stop loss trigger
- the last day of the trading period is reached

Several combinations of the trade and stop loss triggering values were tested and the results were compared to two benchmarks, the S&P500 and the Lehman Brothers Treasury Index. The flagship strategy which triggers at 15 percentile and stop-loss at 5 percentile was claimed to be able to outperform the benchmarks a lot and has little correlation with them. As a matter of fact, the distance measure modeling approaches are good for several things. Firstly these type of methods are economic model-free and purely exploits the statistical properties of relative pricing in a pair of assets. The methods can thereby alleviate problems and criticisms that may bring by model misspecification. Being non-parametric also help reduce the risk of overfitting the data. Secondly, these type of methods fit into the intuition of pairs trading, which is simply finding a pair of assets whose prices move together in long-term, short the overvalued asset and long the undervalued one. There are few advanced mathematical tools involved and hence easy to backtest and implement such strategies. However just as every coin has two sides, the distance methods are without exception. First of all, there is no way to estimate a reasonable holding period or a stop-loss trigger point for an open position due to the lack of a mathematical model. What traders can do is to grid search numerous combinations of the hyperparameters and seek to optimize for example risk-return ratio, which is a process prone to over-fitting historical data. Another problem is that selecting pairs by simple distance measure can only find pairs which have the property of return parity, that is, the returns of the pair move in a 1 : 1 ratio. Mathematically:

$$log(p_t^Y) = \beta log(p_t^X) + \xi_t \tag{2.12}$$

where $\beta = 1$ and ξ_t is a stationary error term with mean equals to zero. For a perfect hedge, the pairs trading strategy should always put an equal amount of money on each side of a pair. However, these type of pairs usually have very strong economic links or can even be the same underlying asset in different exchanges or different forms. In the real world, it is extremely hard to find a pair with return parity while not being arbitraged away. Last but not least, the distance measure may opt-out useful pairs even when $\beta = 1$. This situation happens when the noise level is high, or in other words, t has a large variance. The pair can be even more profitable since ξ_t is still stationary. However, the distance is quite high in the proposed distance measures and the pair may not even be selected.

2.3.2 The Cointegration Modelling Approach

Traders tend to avoid black box models because they are unexplainable and do not have many fundamental insights. Vidyamurthy [12] made efforts to model pairs trading in a more rigorous manner and tried to explain how it shall work based on a widely used asset pricing theory, namely APT. The author proposed a workflow of designing pairs trading strategy based on the idea of cointegration test. As discussed the previous section, two $I(1)$ time series are cointegrated if there exists a linear combination of them which results in a $I(0)$ time series. Vidyamurthy applied Engle-Granger two-step cointegration test on the logarithm price of two assets. The log price of asset A is regressed against the log price of asset B:

$$log(p_t^A) - \beta log(p_t^B) = \mu + \epsilon_t \tag{2.13}$$

where β is the cointegration coefficient or hedge ratio, $\mu + \epsilon_t$ represents the residual term which is then tested for stationarity using ADF test. If the residual term passed the stationarity test it means we can construct a portfolio of asset A and B that has a mean-reverting price based on historical data, and hence profitable by selling at high and buying at low if history continues. The residual term can be further modeled with ARMA process in order to find the theoretical optimal entry and exit points which leads to maximum profit. The author, however, prefers an empirical approach for modeling the residual term, using the rate of zero-crossings as a measure of mean reversion. The idea was to reduce the complexity of the model and to avoid model misspecification. Vidyamurthy also tried to relate the cointegration method to a mainstream asset pricing model, namely APT model and attempted to explain why pairs trading work and where the source of profit lies. First of all, why on earth should one asset always have return proportional to another asset? Vidyamurthy explained that asset A and B may have the same risk factors, and if there is a constant risk exposure proportionality, their log prices are cointegrated. In order to explain the profit of pairs trading, he modified the random walk model of log price for a bit and added a stationary term:

$$log(p_t) = n_t + \varepsilon_t \tag{2.14}$$

where n_t is the random walk component and ε_t is the stationary component. Then the return of an asset within a time period can also be separated into a common trend component and a specific component:

$$\begin{aligned} log(p_t) - log(p_{t-1}) &= (n_t - n_{t-1}) + (\varepsilon_t - \varepsilon_{t-1}) && (2.15) \\ r_t &= r_t^c + r_t^s && (2.16) \end{aligned}$$

where r_t^c is the common trend component due to the random walk part and r_t^s is the specific component due to the stationary part. Vidyamurthy argued that if asset A and B have risk factor exposure vectors βb and b respectively, then:

$$r_A = \beta(b_1F_1 + b_2F_2 + \cdots + b_nF_n) + r_A^s \quad (2.17)$$

$$r_B = (b_1F_1 + b_2F_2 + \cdots + b_nF_n) + r_B^s \quad (2.18)$$

By constructing a portfolio it is, therefore, possible to cancel out common trend component and leave out only the specific component:

$$r_{port} = r_{port}^c + r_{port}^s \quad (2.19)$$

where

$$r_{port}^c = r_A^c - \beta r_B^c = 0 \quad (2.20)$$

$$r_{port}^s = r_A^s - \beta r_B^s \quad (2.21)$$

The spread of the portfolio can then be computed by integrating r_{port}^s, and the spread is a stationary time series. The process of cointegration test is essentially estimate β and then perform stationarity test on the spread or residual term. There are several problems, however, associated with Vidyamurthys cointegration approach for pairs trading strategy design. Firstly, although the author managed to associate cointegration method with APT model, extra assumptions were added to explain why it works. For example in the original form of factor model that relates directly to APT model:

$$r_i = a_i + b_{i1}F_1 + b_{i2}F_2 + \cdots + b_{in}F_n + \epsilon_i \quad (2.22)$$

there is no assumption that $a_i + \epsilon_i$ can be integrated into a stationary time series. In fact, if a_i is a constant term there is no way that the integration can be stationary. A viable explanation would be that the constant term is due to risk-free rate and therefore negligible in relatively high-frequency settings. Another problem is that Vidyamurthy failed to hedge the portfolio correctly. In his work he claimed to use β as a position ratio for asset B and A, meaning that each time when a trading signal triggers, he would sell β units of B and buy 1 unit of A, or buy β units of B and sell 1 unit of A. This would work as a perfect hedge when prices of B and A are regressed against each other. However since all the works before were done using log prices, and $log(p_t) - log(p_{t-1}) \approx r_t$, a better way to hedge would be to fix the investment for B and A to be in ratio of $\beta : 1$, which is still not perfect since the approximation is accurate only when r_t is close to zero. Last but not least, the cointegration model requires estimation of hedge ratio, which in this case is done by OLS regression. LaViola [8] used DESP to do the same job. No matter which approaches above was used, the parameter has to be estimated in a rolling manner if the system has a varying hedge ratio. In which case the size of the rolling window would be another parameter to be tuned. Vidyamurthy mentioned the use of Kalman filter in his work, however, the detail on how to apply the trick for parameter estimation was opt out. Christian L [3] and Kangwhee [7] applied Kalman filter for adaptive estimation of hedge ratio in a very similar manner. The state-space equations are:

$$p_t^A = \beta_t p_t^B + \varepsilon_t \quad (2.23)$$

$$\beta_t = \beta_{t-1} + \eta_t \quad (2.24)$$

where p_t^A and p_t^B are the price series of the pair, β_t is the time-varying hedge ratio, ε_t and η_t are independent uncorrelated error terms. Equation 2.23 represents the observation equation and equation 2.24 is the state transition equation. This model suggests that the hidden state, which is the hedge ratio, follows a random walk. In order to make an online estimation of the hidden state with Kalman filter, the observation error t and transition error t need to be estimated and specified. Both papers mentioned above claimed to accomplish this by maximizing the likelihood function but omitting the details.

Chapter 3

A New Model for Statistical Arbitrage

3.1 A Modified Kalman Filter State Model

Recent literatures on adaptive hedge ratio prediction with Kalman filter [3][7] agreed on a model described by equation 2.23 and 2.24. It models the change of hedge ratio as a random walk and the spread as Gaussian white noise. However, this model neglect the intercept part of linear regression, which fails to explain the price movement between pairs in many cases. For example, consider a classic pair of statistical arbitrage - soybeans and their refined products, namely soybean meal and soybean oil. Since soybeans are crushed to produce soybean meal and soybean oil, it will be reasonable to model the price of soybean as a linear combination of the prices of soybean meal and soybean oil:

$$p_t^B = \beta_1 p_t^M + \beta_2 p_t^O + \xi_t \tag{3.1}$$

where p_t^B, p_t^M and p_t^O are log prices of soybean, soybean meal and soybean oil respectively

The economic link here is quite self-explanatory. When the price of soybean soars, the production cost for soybean meal and soybean oil increases, and hence their prices increase. If there is a high demand for the oil in one particular season, both the price and the supply of it would increase, meaning that more beans are crushed, which leads to a fall in soybean meal price because the demand for soybean has not changed. Therefore if we consider the soybean and soybean meal pair for pairs trading, the residual cannot be a stationary time series since it is governed by the price of soybean oil as well. There are many pairs like this in the world of commodity futures. Taking this into consideration we propose a new state model which introduces an intercept term:

$$p_t^Y = \beta_t p_t^X + \alpha_t + \xi_t \tag{3.2}$$

$$\beta_t = \beta_{t-1} + \eta_{1t} \tag{3.3}$$

$$\alpha_t = \alpha_{t-1} + \eta_{2t} \tag{3.4}$$

where α_t is the intercept, models as a random walk process just like β, but with independent variance.

Introducing a random walk intercept term certainly helped to take into consideration the effect of price movements of some other related assets. It also makes the parameter estimation harder, however, by introducing a new variable. In the next section, we will discuss the theoretical background of MCMC, and its application in our model.

3.2 MCMC for Parameter Estimation

In the context of Bayesian inference, often we need to deal with complex probabilistic models. In most cases, our quantity of interest is the posterior distribution of model parameters, $\mathbb{P}(\theta|x)$, where θ is our parameters, and x stands for observed data. To compute this, we resolve to the Bayesian formula.

$$\mathbb{P}(\theta|x) = \frac{\mathbb{P}(x|\theta)\mathbb{P}(\theta)}{\mathbb{P}(x)} \tag{3.5}$$

In the equation 3.5 above, the likelihood, $\mathbb{P}(\theta|x)$, and the prior, $\mathbb{P}(\theta)$, are easily accessible. The denominator $\mathbb{P}(x)$ however, can only be computed by integrating over all possible parameter values.

$$\mathbb{P}(x) = \int_{\Theta} \mathbb{P}(x, \theta)\, d\theta \tag{3.6}$$

This is the main difficulty of the Bayesian formula. For complex models, it is impossible to compute the posterior in a closed-form manner. Thus, we turn to MCMC method. This method constructs a Markov chain whose equilibrium distribution equals to our desired distribution. With this Markov chain, we can sample from our desired posterior.

Recall that a discrete-time Markov chain is a sequence of random variables $X_0, X_1, \cdots$ such that:

$$\mathbb{P}(X_t \in A|x_0, \cdots, x_{t-1}) = \mathbb{P}(X_t \in A|x_{t-1}) \equiv P(x_{t-1}, A) \tag{3.7}$$

where the function P is named the transition kernel. A probability distribution Π with density function π over the state space is called the equilibrium distribution for p if:

$$\Pi(A) = \int_{\mathbb{R}} P(x, A)\pi(x)\, dx \tag{3.8}$$

for all measurable sets A. To put the above equation in words, if the Markov chain with transition kernel P starts with state distribution Π, the unconditional state distribution remains invariant over time.

To focus on the principle of MCMC method, we omit some technical details and assume that our transition kernel P with equilibrium distribution Π is aperiodic, irreducible, and positive Harris-recurrent, we can draw the conclusion that Π is the unique equilibrium distribution for P, and the Markov chain is Harris-ergodic. Furthermore, we have the following theorem, for every x and all measurable sets A:

$$\lim_{t\to\infty} |P(X_t \in A | X_0 = x) - \Pi(A)| = 0 \tag{3.9}$$

This means that we can start the Markov chain at any state, let the chain run for sufficient time, and the final state distribution will be approximately Π. MCMC with its dedicated sampling algorithm, is able to construct such a Markov chain with the equilibrium distribution converge to the posterior distribution we desire.

3.3 A Probabilistic Model for Pairs Trading

Recall that we modelled both the hedge ratio β and intercept α as random walk process rather than as constants in equation 3.2. In finance, the hedge ratio and the intercept have intuitive interpretations. The hedge ratio β, depicts the intrinsic linear relationship between the values of X and Y. The intercept, α, stands for other factors other than X driving the price of asset Y. These factors may vary over time and intuitively should be treated as time series. As a result, we have our probabilistic model of pair trading, in a hidden Markov chain setup:

State transition:

$$\begin{pmatrix} \beta_t \\ \alpha_t \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_{t-1} \\ \alpha_{t-1} \end{pmatrix} + W_t \tag{3.10}$$

where $W_t \sim N(0, Q)$ is a two-dimensional white noise process, $Q = \begin{pmatrix} q_h & 0 \\ 0 & q_b \end{pmatrix}$

Observation:

$$Y_t = \begin{pmatrix} X_t & 1 \end{pmatrix} \begin{pmatrix} \beta_t \\ \alpha_t \end{pmatrix} + \xi_t \tag{3.11}$$

where $\xi_t \sim N(0, r)$ is a white noise process

In our model, the hedge ratio and intercept, β_t and α_t are modeled by two independent random walk processes. The state transition error, q_h and q_b, as well as the observation error, r, are to be estimated by MCMC. With these parameters estimated, we can perform online estimation of the state variables, hedge ratio β_t and intercept α_t with Kalman filter.

3.4 A New Profitability Criterion

Recall that there are two pairs trading profitability criterions that are widely used by researchers and practitioners, namely the distance measure and the two-step cointegration test model. In both models, however, a constant hedge ratio and intercept are assumed. Our model does not depend on the constant parameters assumption, and hence we can derive a better profitability criterion, taking parameter variation into consideration. We can take a brief analysis of our trading process. Suppose we are trading a pair of assets, X and Y, with the mean-reverting portfolio being:

$$\beta_t X - Y + \alpha_t \tag{3.12}$$

The time series of hedge ratio and intercept, β_t and α_t, follows our model in the previous section. Furthermore, let the volatility (standard deviation of the return series) of the price series of X, P_t^X be σ.

Our entering signal is when the spread is large, meaning that the price of the pair portfolio is far from the long time equilibrium. In this case, we implement a very simple signal: we enter our trade cycle if the deviation of spread from equilibrium reaches k times the observation error r. The observation error r, estimated by MCMC, is just the volatility of the mean reversion portfolio. Our exiting signal triggers when the spread closes, meaning that the portfolio has returned to the long-time equilibrium. See figure 3.1 for a graphical illustration. During one trade cycle, let the time we enter be 0, and the time we exit be t.

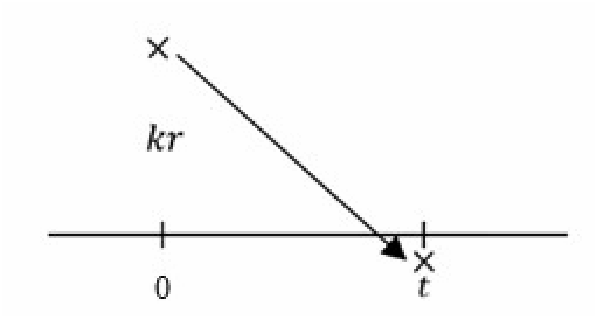


Figure 3.1: An illustration of a complete trade cycle for a pair

At time s when we enter our trade cycle, we estimate that the mean reversion portfolio is:

$$\beta_0 X - Y + \alpha_0 \tag{3.13}$$

However, when we exit our trade cycle, at time t, the mean reversion part will change to:

$$\beta_t X - Y + \alpha_t \tag{3.14}$$

The difference between the two estimations is important. Due to the variability of β_t and α_t, the portfolio we traded at time 0 is only partially mean-reverting at time t, and the part that is no longer mean-reverting need to be treated as a random walk. We can see that the changes in hedge ratio and intercept have a detrimental effect on our profitability: if they are changing too fast, a large part of our portfolio will not be mean reverting, eliminating our source of profit.

Taking this variability into account, our profit of a single trade cycle can be written as:

$$Profit = kr + (\beta_0 - \beta_t)P_t^X + (a_0 - a_t) \tag{3.15}$$

$$\mathbb{E}[Profit] = kr \tag{3.16}$$

The exit time t stands for the first time a white noise series hits below zero (when our exiting signal is triggered). Under this assumption t follows a $Geometric(1/2)$ distribution. In addition, we can compute the variance of our profit by conditioning on t:

$$Var(Profit|t) = q_h^2 t\sigma^2 t + q_b^2 t = q_h^2 \sigma^2 t^2 + q_b^2 t \tag{3.17}$$

$$Var(Profit) = \mathbb{E}[Var(Profit|t)] = 36q_h^2\sigma^2 + 2q_b^2 \tag{3.18}$$

As a result, our profit follows a normal distribution with mean kr and variance $36q_h^{22} + 2q_b^2$. We can compute the probability of a positive profit, $\mathbb{P}(Profit > 0)$ by:

$$\mathbb{P}(Profit > 0) = 1 - \Phi(-\frac{kr}{\sqrt{36q_h^2\sigma^2 + 2q_b^2}}) \tag{3.19}$$

where Φ being the standard normal cumulative distribution function.

To guarantee profitability, we will require that $\mathbb{P}(Profit > 0)$ be sufficiently large, thus giving us a way to identify profitable pairs.

Chapter 4

Data Preparation and Implementation of Trading Strategies

4.1 Data Preparation

As mentioned in the introduction section, the data used in this study is provided by Ricequant, a Shenzhen based company that aims to create crowd-sourced hedge fund and provides high-quality intra-day data for Chinese stocks and futures. Unlike most previous studies on pairs trading which utilized daily data [12][4][5], the data used for analysis in this work is in 1-minute frequency. We also use this data to generate data sampled at 5-min, 30-min frequencies. The time span of most future varieties covers entire three years period between January 01, 2014 and January 01, 2017, despite the ones that were newly created during the period. We select 23 most liquid future varieties for our analysis, and for each variety, the most liquid contract is picked for that trading period. It is worth mentioning that each future contract has a maturity date after which the contract will never be traded anymore in the market. Since we are trying to run analysis and backtest trading strategies for each future variety continuously for a quite long period, it is necessary that we handle the price jumps caused by contango and backwardation. A snapshot of the continuous price of iron ore future illustrates this problem, price jump due to switching contract highlighted by red pen, see figure 4.1. In fact, many financial data providers and websites failed to point this out when they make graphs for continuous contracts, which may mislead the investors to think that there actually exist sudden price jumps.

Figure 4.1: An example of price jump due to backwardation(finance.sina.com)

It will be easier to understand this via an example, say there is a price gap between the contract which matures in May and the contract which matures in September, due to storage cost, opportunity cost and etc. For one particular day say on March 23rd, the most liquid contract turns from the May contract to the September contract. If we simply stitch the prices together we get an unwanted price jump that does not reflect market intentions on the underlying asset. This kind of jumps can be big and happens regularly, which may cripple our data analysis and backtest results because our trading algorithm will take the price jump as a diverge from historical value and may hence trigger trading signals. The way we solve this problem is quite conventional. We monitor the daily trading volume of contracts that share the same underlying asset day by day. When a later contract surpasses the previous contract with highest trading volume by 10%, the price gap between the two contracts is computed, and then prices of all previous trades are shifted by the price gap. In this way when we stitch the prices together we can smooth out the price jump effect brought by switching contract. In this study, we employ the method of rolling forward cross-validation for separating in-sample and out-of-sample data points. For every trading day, we use the data from the previous five trading day to obtain the necessary statistics to generate trading signals. In this way, we make sure that there is no look-ahead bias for the backtest simulation.

4.2 Implementation of Statistical Arbitrage Trading Strategies

In this section, we discuss in detail the implementation of three different statistical arbitrage trading strategies. First, we depict the very naive trading strategy based on distance measure, which assumes 1 : 1 constant hedge ratio. Then we re-investigate the cointegration model and discuss the necessity of estimating hedge ratio on the flow rather than fixing it as a constant. The model of estimating hedge ratio with OLS is then described in detail. At last, we propose a model which unleash the power of Kalman Filter for on-line hedge ratio estimation, and at the same time utilizes MCMC for hidden variable estimation.

4.2.1 Naive Strategy Based on Distance Measure

A naive pairs trading strategy based on distance measure has been proved to be effective in various markets [5][10]. The distance computed from a pair of an asset is not only used for generating the trading signal but can also be used to evaluate whether or not the pair is suitable for pairs trading. These type of approaches are purely empirical and do not involve any mathematical models nor pricing theories. The idea is very simple: when the price of asset X moves to a relatively high position comparing to the price of asset Y, considering the historical distribution of the spread, we can then short X and long Y and expect the spread to revert. We perform the reverse operation when the price of Y moves to a relatively high position against the price of X. The spread between X and Y S_t^{XY} is computed as:

$$S_t^{XY} = log(P_t^X) - log(P_t^Y) \tag{4.1}$$

For a practical trading strategy, the entry signal triggering point can be set as several standard deviations/sigma away from the mean of the spread, evaluated from a past window. This approach to some extent assumes the normal distribution for the spread.

Another method for triggering entry signal is to look for certain quantiles in the empirical distribution of a past window of data. We tried both methods and did not find too much difference in performance between the two. Thereby for simplicity of the analysis, we stick with the first method. We set triggering point for entry signals to be two sigmas away from the mean, evaluated from the past 300 minutes of last trade prices. Then we set exit triggering point to be 0.5 sigmas away from the mean. Meaning that we would unwind our positions once the spread reverts below the 0.5 sigma threshold. Figure 4.2 is an illustration of how trading signals are triggered on one side. Note again that the sell signal is for selling the pair (selling X and buying Y at the same time), and the exit signal serves to close positions for both assets.

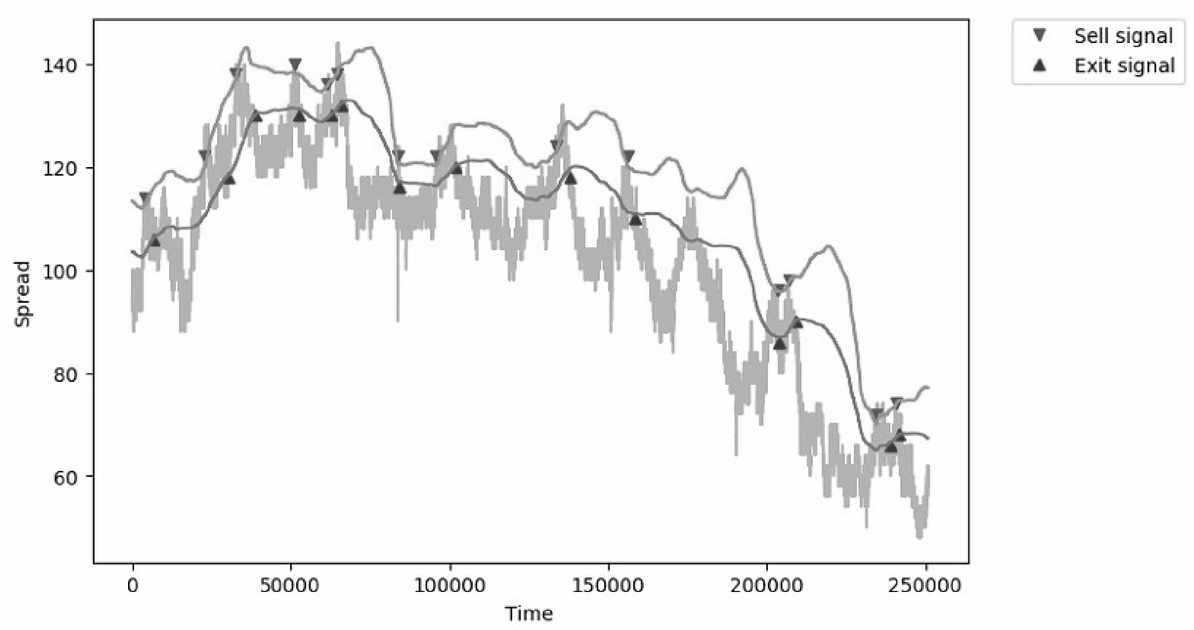


Figure 4.2: An illustration of how trading signals are triggered on one side for the naive strategy based on distance measure

4.2.2 Cointegration Model and OLS for adaptive hedge ratio estimation

Cointegration model has been widely adopted in recent efforts for designing statistical arbitrage strategies. It not only provides a more sensible way for searching suitable pairs than distance measure and Pearson correlation but also attempts to explain the source of profit for this particular trading strategy. With the help of a widely accepted pricing theory APT, Vidyamurthy [12] made contributions to make the model more rigorous. This is briefly explained in the preliminaries section with Vidyamurthy's own nomenclature as in his work. However, we feel it would be necessary to unify symbols in this thesis and fill in some missing details for Vidyamurthy's explanation. Recall that the profit of a statistical arbitrage trading strategy comes from the relative mispricing between the pair:

$$log(p_t^Y) = \beta log(p_t^X) + \xi_t \tag{4.2}$$

where the log price of the asset Y equals to the log price of an asset X plus a stationary time series ξ_t.

A cointegration test estimates the hedge ratio β for a given series of $log(p_t^Y)$ and $log(p_t^X)$, then test the residual ξ_t for stationarity. Therefore if we model the two price series like this and it passed the cointegration test, it is very likely that we can profit from the pair if history continues. But in what circumstances would equation 4.2 hold? The answer is if asset X and Y have exactly proportional risk factor exposure vectors.

Define return at time t as:

$$r_t = log(p_t) - log(p_{t-1}) \tag{4.3}$$

Current log price is the cumulative sum of all returns plus the initial log price:

$$log(p_t) = \sum_{i=1}^{t} r_t + log(p_0) \tag{4.4}$$

Then we have:

$$log(p_t^Y) - \beta log(p_t^X) = \sum_{i=1}^{t}(r_t^Y - \beta r_t^X) + log(p_0^Y) - \beta log(p_0^X) \tag{4.5}$$

Now if asset X and Y have linearly proportional risk factor exposure vectors, according to APT, $\forall t$ we have:

$$r_t^X = \sum_{j=1}^{n}(b_j F_{jt}) + r_t^{Xs} \tag{4.6}$$

$$r_t^Y = \beta \sum_{j=1}^{n}(b_j F_{jt}) + r_t^{Ys} \tag{4.7}$$

where r_t^{Xs} and r_t^{Ys} are specific components which are not related to risk factor exposure vectors. The equation 4.5 thus becomes:

$$log(p_t^Y) - \beta log(p_t^X) = \sum_{i=1}^{t}(r_t^{Ys} - \beta r_t^{Xs}) + log(p_0^Y) - \beta log(p_0^X) \tag{4.8}$$

And by considering equation 4.2 it must satisfy:

$$\sum_{i=1}^{t}(r_t^{Ys} - \beta r_t^{Xs}) + log(p_0^Y) - \beta log(p_0^X) = \xi_t \tag{4.9}$$

This means that the cumulative sum of the specific components time series must be stationary, and hence the time series cannot be $I(0)$.

The Engle-Granger two-step cointegration test fits perfectly into this model and has become a touchstone for potentially profitable trading pairs. With the historical log prices, it estimates β with OLS, and then performs ADF test on the residual for stationarity.

Asides from selecting pairs, the idea of estimating hedge ratio with OLS can also be used for generating trade signals. However, in real life scenarios, the hedge ratio of a pair of assets is unlikely to stay constant. Therefore the hedge ratio estimation should be done in a rolling manner. Christian Dunis et al. [3] compared OLS and DESP for adaptive hedge ratio estimation in high frequency settings and concluded that they perform similarly. In this thesis, we implement OLS method since it is more intuitive. With hedge ratio β, the spread is calculated as:

$$S_t^{XY} = \beta log(P_t^X) - log(P_t^Y) \tag{4.10}$$

To make the results comparable to the naive strategy described in the last section and enforce similar trading frequencies, we keep the usage of 300 minutes time window for normalizing the spread, 2 sigmas for triggering entry signal and 0.5 sigmas for exit signal. Comparing to the naive trading strategy, however, performing a rolling estimation of hedge ratio with OLS introduces another parameter, which is the length of the OLS rolling window. Christian Dunis et al. applied genetic optimization on 6 randomly chosen pairs during in sample period with the objective of maximizing the information ratio obtained on trading these pairs. For the purpose of this thesis, however, since we are dealing with all possible pairs of liquid Chinese commodity futures, we choose to omit the parameter optimization by choosing an arbitrary 3000 minutes window for OLS. This window may work well on some pairs and performs poorly on others. The whole idea is to provide a sense of how this trading strategy performs in general, and avoid over-fitting to some specific pairs.

4.2.3 Adaptive Hedge Ratio Estimation with Kalman Filter and MCMC

Recall that in chapter 3 we showed a modified state model for adaptively estimating hedge ratio with Kalman filter. We also showed that MCMC can be used for parameter estimation in this process. The implementation of our trading strategies is quite simple once we have our model. However, there are some technical details worth mentioning.

In the previous chapter, we mentioned that our entering signal triggers when the spread grows larger than kr, with r being the observation error in our model.In reality, however, the observation error r shows strong character of heteroscedasticity. It is not proper to assume r to be constant over time, so we perform a moving-window estimation of the observation error r. And just as what we did for the other two trading strategies, we make it a 300 minutes window. The detail of our trading strategy goes as following:

1. With the previous one year data of the log prices p_t^X and p_t^Y, estimate the transition error(q_h, q_b) and observation error(r) with MCMC in our model's setting.
2. Update hedge ratio and intercept with new data in our Kalman filter, acquiring β_t and α_t from β_{t-1} and $alpha_{t-1}$.
3. Compute current spread as:
$$S_t^{XY} = \beta_t p_t^X - p_t^Y + \alpha_t \tag{4.11}$$
4. Update the moving-window estimation of observation error, r.
5. If we current hold no position:
    (a) If $S_t^{XY} > kr$, short β dollar of X and long 1 dollar of Y (long the portfolio)
    (b) If $S_t^{XY} < -kr$, long β dollar of X and long 1 dollar of Y (short the portfolio)
    (c) Else do nothing
6. If we currently hold long position on the portfolio:
    (a) If $S_t^{XY} > 0$, clear our position
    (b) Else do nothing
7. If we currently hold short position on the portfolio:
    (a) If $S_t^{XY} < 0$, clear our position
    (b) Else do nothing

Chapter 5

Some Empirical Results and performance analysis

5.1 Return Calculation and Transaction Cost Estimation

Each data point for all frequencies we use in the simulation contains price information of 'Open', 'Close', 'High', 'Low' and 'VWAP' within the period. VWAP as its name suggests, is a weighted average of prices of transactions done in the specific period. As we discussed in the preliminary section of this thesis, market frictions due to illiquidity can hamper the performance of a trading strategy to a large extent. In order to make our simulations as real as possible, when the close price at time t: $P_t^{X\_Close}$ triggers a trading signal, we use the VWAP price at the next point: $P_{t+1}^{X\_VWAP}$ to fill in the trade. Therefore, the return we earned from holding a long position of future contract X from time t to time t' is computed as:

$$r_{t'}^{X} = log(P_{t'+1}^{X\_VWAP}) - log(P_{t+1}^{X\_VWAP}) \tag{5.1}$$

In the contrast, the return we earned from holding short position of future contract Y in the same period is:

$$r_{t'}^{Y} = -log(P_{t'+1}^{Y\_VWAP}) + log(P_{t+1}^{Y\_VWAP}) \tag{5.2}$$

Total return within the period is thereby:

$$r_{t'}^{total} = r_{t'}^{X} + r_{t'}^{Y} \tag{5.3}$$

For brokerage fee we considered a conservative value of 0.1% one way for trading each item. The brokerage fees actually different from contract to contract, and are usually much lower than 0.1%. We fixed the value just to make the analysis easier.

5.2 Preliminary Analysis

In this section, we will first introduce the performance measure we use in this thesis. Secondly, we discuss the performance distribution of naive trading strategy across different frequencies. Then we will list a few factors that can be used as features for selecting pairs, and compare their predictive power.

5.2.1 Information Ratio, a Performance Measure

IR is a widely used performance measure for trading strategies. It essentially measures the risk-adjusted return, as similar to another popular measure, Sharpe Ratio. The computation is done by dividing the portfolio's returns over some benchmark, to volatility of those returns:

$$IR = \frac{r_p - r_b}{s_{p-b}} \tag{5.4}$$

where r_p and r_b are annualized returns of the portfolio and the chosen benchmark, and s_{p-b} is the standard deviation of differences between the two returns.

From equation 5.4 we can see that IR rewards for high return, but also punishes for high deviation. Therefore IR is a performance measure which not only looks high returns in trading strategies, but the returns also need to have relatively high consistency. Sharpe Ratio can be computed in a similar manner, just by substituting the benchmark returns with risk-free interest rates.

Since we are considering future contracts across multiple industries and sectors, it is cumbersome and unnecessary to include various benchmarks in this study. In fact, the benchmark return terms are just offsets and do not affect our comparisons across strategies. Hence we substitute them with zero. Moreover, since we are considering daily returns in our simulation, we need to multiply a factor to compute annualized IR.

$$IR = \sqrt{252}\,\frac{r_p}{s_p} \tag{5.5}$$

where $\sqrt{252}$ is the factor we need, there are approximately 252 trading days per calendar year.

5.2.2 Does Sampling Frequencies Affect Performance?

As mentioned in the trading strategy implementation section, spreads are computed using a 300 minutes moving window for all strategies for the ease of comparison. Since we have data with 1-min, 5-min and 15-min sampling frequencies, a natural question would be whether or not sampling frequencies affect performance of trading strategies. Since parameter estimation MCMC is extremely computational expensive, we will only consider the naive distance measure strategy and the rolling window OLS strategy in this analysis. All pairs available in 2014 are used.

We can see from table 5.1 and table 5.1 that both strategies perform better overall in higher sampling frequencies, by delivering higher average IR as sampling frequency

Table 5.1: Summary of information ratio for naive distance measure trading strategy in year 2014. One way transaction cost has been considered as 0.3%

	15-Minute	5-Minute	1-Minute
Mean	-0.75	-0.49	0.40
Median	-0.72	-0.44	0.16
Max	2.18	2.81	7.08
Min	-3.76	-3.36	-3.07
Standard deviation	1.09	1.12	1.73
Skewness	-0.17	0.09	1.06
Kurtosis	0.20	0.43	1.78

Table 5.2: Summary of information ratio for OLS hedge ratio estimation trading strategy in year 2014. One way transaction cost has been considered as 0.3%

	15-Minute	5-Minute	1-Minute
Mean	-0.51	-0.17	0.83
Median	-0.45	-0.18	0.65
Max	1.93	3.46	6.27
Min	-2.95	-2.90	-1.99
Standard deviation	0.98	1.16	1.74
Skewness	-0.19	0.01	0.84
Kurtosis	-0.16	-0.17	0.49

increases. It is also not hard to spot that as sampling frequency increases, the skewness of the distribution of IR shifts from negative to positive for both strategies. This means that the majority mass of IR is negative (strategy losses money) for 15-Minute sampling frequency, while the majority mass of IR is positive (strategy makes profit) for 1-Minute sampling frequency. This result is actually not hard to anticipate. Since we are using the same time window for parameter calculation in all frequency settings, higher sampling frequency simply provides more trading opportunities and thereby making the results more robust. Note that we have already included a conservative cost for each transaction. It is not hard to imagine that if there is no or lower transaction cost, strategies with high-frequency settings would be able to deliver much better results. [1] Kurtosis also increases as sampling frequency increases, suggesting that the distribution of IR has fatter tails for higher frequency sampling rate, and thus values are more extreme. Figure 5.1 and 5.2 illustrate the distributions of IR for the two trading strategies in 2014.

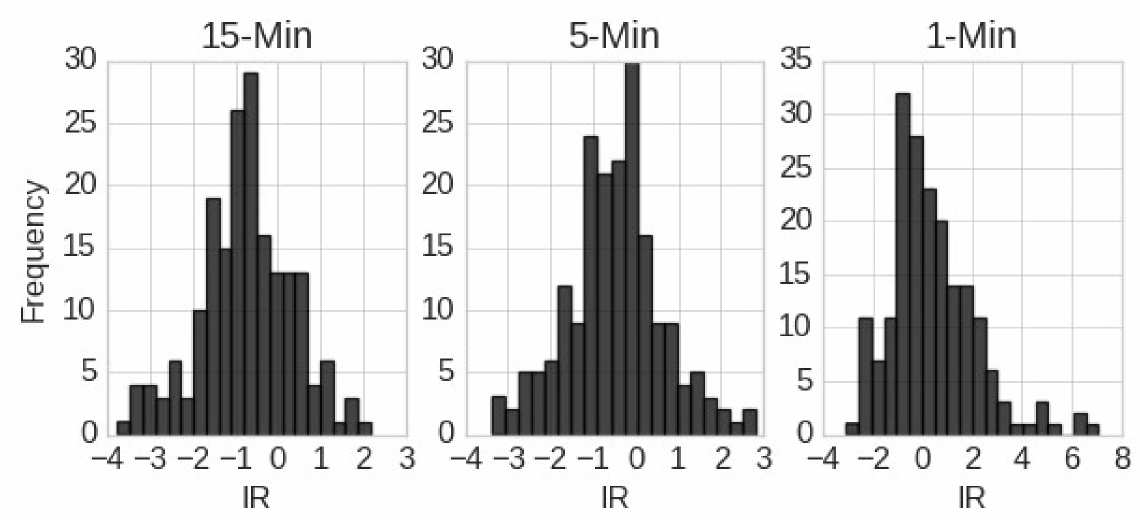


Figure 5.1: Distribution of information ratios for the naive distance measure trading strategy in year 2014

Again as we can see from the histograms, more weight shifts to higher IR regions

[1] In fact, many exchanges do provide discount on transaction fees for high-frequency traders.

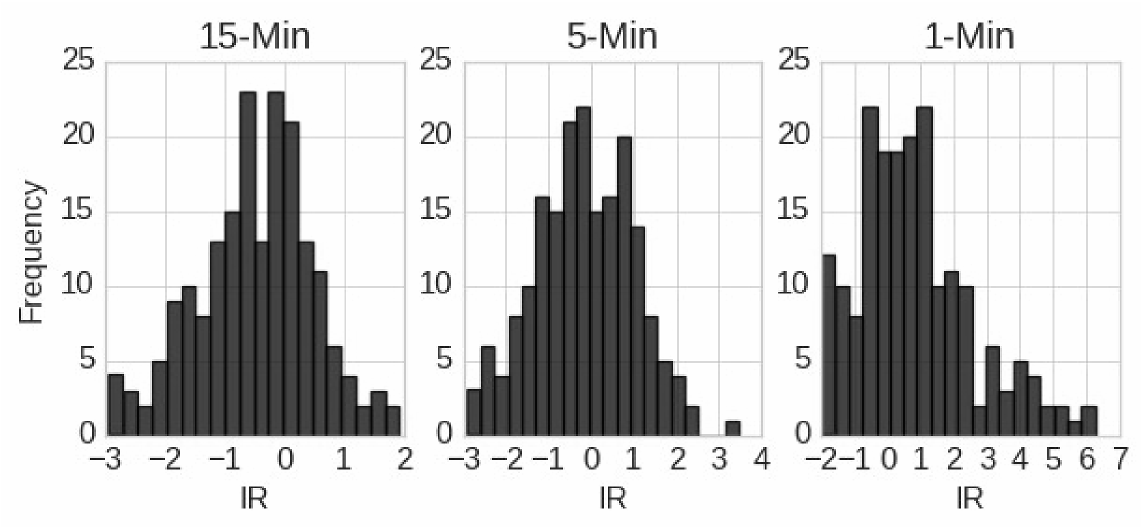


Figure 5.2: Distribution of information ratios for the OLS hedge ratio estimation trading strategy in year 2014

as sampling frequency increase for both strategies. We can conclude that the 1-Minute sampling frequency delivers the best result among the three. Thereby we will only discuss results generated from 1-Minute data for the upcoming analysis.

5.2.3 Correlation Between Distance Measure and IR

Performing exhaustive backtest simulations on trading strategies for all pairs in a market can be computationally expensive, especially with high-frequency data and a rather complicated model. Also, rely solely on simulation results for selecting pairs is a pure black box, not convincing and may subject to survival bias. Thereby it is very useful to have some indicators for pairs selection before diving into trading simulations. We implemented distance measure the same way as described in Gatev's paper [5]. The distance between X and Y in a period T is computed as the sum of squared distances of log prices of X and Y within that period:

$$D_{XY} = \sum_{t \in T} (P_t^X - P_t^Y)^2 \tag{5.6}$$

where P_t^X and P_t^Y are log prices of X and Y normalized by dividing their respective first values in that period.

We then performed an in-sample and out-of-sample test to find out if there is a correlation between distance measure and IR, obtained from both naive strategy and OLS hedge ratio estimation strategy. For the in-sample test, we compute the Pearson correlation and the p-value for testing non-correlation between distance measure and IR of the trading strategy in the same year. While for the out-of-sample test we computed the values based on the IR obtained in one year, and the distance measure in the previous year. See 5.3 and 5.4 for results.

Table 5.3: Pearson correlation and p-value computed from distance measure and IR obtained from naive strategy, in-sample

	Pearson correlation	p-value
2014	-0.184	0.011
2015	-0.174	0.016
2016	0.051	0.484

Table 5.4: Pearson correlation and p-value computed from distance measure and IR obtained from naive strategy out-of-sample

	Pearson correlation	p-value
2014	-0.062	0.395
2015	-0.106	0.145
2016	-0.054	0.463

Many researchers and traders use distance measure as the sole indicator for selecting pairs, the result, however, is quite unsatisfactory for our simulation results on naive trading strategy, which is also applied by many. There is a quite strong negative correlation between distance measure and IR in the year 2014 and 2015 for in-sample tests. The p-values are also quite significant, suggesting that it is unlikely for them to be uncorrelated. However, when it comes to the out-of-sample test, neither Pearson correlation nor p-value is significant.

We are also interested in finding out if the distance measure delivers more information for the OLS hedge ratio estimation trading strategy. See table 5.5 and 5.6

Table 5.5: Pearson correlation and p-value computed from distance measure and IR obtained from OLS hedge ratio estimation strategy, in-sample

	Pearson correlation	p-value
2014	-0.089	0.222
2015	-0.064	0.381
2016	0.055	0.448

Table 5.6: Pearson correlation and p-value computed from distance measure and IR obtained from OLS hedge ratio estimation strategy out-of-sample

	Pearson correlation	p-value
2014	-0.050	0.495
2015	0.258	0.001
2016	-0.082	0.263

It appears that the negative correlation between distance measure and IR is not strong, for both in-sample and out-of-sample test. What is more unexpected is the out-of-sample test result in 2015 end up with a very strong positive correlation, suggesting that larger distance between pairs in the past deliver better performance in the future.

5.2.4 Correlation Between p-value for Cointegration Test and IR

Recall that Engle-Granger two-step cointegration test examines the residual term for linear regression and test for stationarity using ADF test. It is a quite common test used by many researchers and practitioners in the pairs selection stage. [12][3][2]. For ADF test, a low p-value would reject the null hypothesis that the time series is non-stationary. Although the proper way to deal with p-value is to treat it as binary by selecting a cut-off value, in order to make a simple comparison with the results obtained from distance measure we computed the Pearson correlation and corresponding p-value, just like what was done in the previous section. See table 5.7 and 5.8 for results obtained from naive trading algorithm.

Table 5.7: Pearson correlation and p-value computed from p-value of cointegraton test and IR obtained from naive strategy, in-sample

	Pearson correlation	p-value
2014	-0.001	0.985
2015	-0.096	0.187
2016	-0.185	0.011

Table 5.8: Pearson correlation and p-value computed from p-value of cointegraton test and IR obtained from naive strategy, out-of-sample

	Pearson correlation	p-value
2014	-0.008	0.913
2015	0.187	0.010
2016	0.010	0.893

The results appears to be insignificant across the three year period. Table 5.9 and 5.10 used IR obtained from the OLS hedge ratio estimation trading strategy.

In-sample results show that there is a negative correlation between p-values obtained from cointegration test and IR, although not very significant. Curiously the result does

Table 5.9: Pearson correlation and p-value computed from p-value of cointegraton test and IR obtained from OLS hedge ratio estimation strategy, in-sample

	Pearson correlation	p-value
2014	-0.090	0.218
2015	-0.171	0.019
2016	-0.046	0.531

Table 5.10: Pearson correlation and p-value computed from p-value of cointegraton test and IR obtained from OLS hedge ratio estimation strategy, out-of-sample

	Pearson correlation	p-value
2014	0.123	0.091
2015	0.062	0.398
2016	0.002	0.978

not hold for out-of-sample results. We can conclude that cointegration test is not a good scheme for selecting pairs for both strategies, at least in our settings.

5.2.5 Correlation Between Previous Year IR and IR

For a particular trading strategy, it is natural to use the performance from last trading period to predict the performance of the upcoming trading period. It is not a very rigorous indicator since the performance obtained in the past can be simply an overfit and may not generalize in future. However, since we are using one set of parameters for each trading strategies and have a long testing period, the IR of a pair in the last year should be somewhat indicative for the IR of the pair in the upcoming year, if history continues. Table 5.11 and 5.12 are correlations and p-values computed using IR obtained from naive strategy and OLS hedge ratio estimation strategy respectively.

From the fact that all results deliver positive Pearson correlation and small p-value,

Table 5.11: Pearson correlation and p-value computed from previous year IR and IR obtained from naive strategy

	Pearson correlation	p-value
2014	0.193	7.50e-03
2015	0.375	9.86e-08
2016	0.320	6.73e-06

Table 5.12: Pearson correlation and p-value computed from previous year IR and IR obtained from naive strategy

	Pearson correlation	p-value
2014	0.010	9.63e-01
2015	0.132	6.84e-02
2016	0.397	1.42e-08

we can see that the performance on a pair last year is indeed a good indicator of the performance this year.

5.3 Comparison of Trading Strategies

In this section, we will compare the performance of our trading strategy which utilizes Kalman filter hedge ratio estimation with MCMC (Kalman filter hedge in short) to the naive strategy and the OLS hedge ratio estimation strategy (OLS hedge in short). Unfortunately, due to the expensive computation cost of MCMC, we were not able to perform exhaustive simulation on all 190 pairs as what we did with the previous two trading strategies. Instead, we selected 5 pairs which fundamentally make sense and also performs relatively well in the year 2014 for both naive and OLS hedge strategy, plus 5 pairs which are randomly selected in the remaining pool. Then we run simulations in the year 2015 and 2016. See table 5.13 for the detail of our selection.

Table 5.13: Selected pairs for performance evaluation of trading strategies

	item_X	item_Y	pair_contract_id
0	soybean	soybean meal	(a0, m0)
1	soybean	soybean oil	(a0, y0)
2	cotton	white sugar	(cf0, sr0)
3	soybean meal	rapeseed meal	(m0, rm0)
4	rapeseed oil	soybean oil	(oi0, y0)
5	copper	rapeseed meal	(cu0, rm0)
6	aluminum	soybean	(al0, a0)
7	corn	rapeseed oil	(c0, y0)
8	gold	polyvinyl chloride (PVC)	(au0, v0)
9	rebar	linear low-density polyethylene (LLDPE)	(rb0, l0)

From this particular selection of pairs, not only will we be able to evaluate the performance of the three trading strategies on pairs that are suitable for statistical arbitrage, but also will we be able to find out how strategies behave in worst case scenarios. Table 5.14, 5.15 and 5.16 are IR results from backtest simulations done in the 2015 & 2016 two years period.

We can see that the OLS hedge trading strategy performs a little worse than the naive strategy for the chosen 10 pairs in this particular period. This conflicts with our observation in the previous section about IR distributions, that OLS hedge trading strategy delivers IR distribution with higher mean and smaller variance. One possible explanation is the small sample size bias.

We can also see that our Kalman filter hedge trading strategy outperforms the other two strategies quite a bit by delivering significantly higher IR for the top five pairs, and no worse performance for the bottom five pairs.

Table 5.14: IR of 10 pairs for naive trading strategy, year 2015 and 2016

	pair	IR
4	(y0, oi0)	3.58
3	(m0, rm0)	2.43
9	(rb0, l0)	0.74
0	(a0, m0)	0.36
8	(au0, v0)	0.31
6	(al0, a0)	-0.15
1	(a0, y0)	-0.24
7	(c0, y0)	-0.75
5	(cu0, rm0)	-0.87
2	(cf0, sr0)	-1.16

Table 5.15: IR of 10 pairs for OLS hedge trading strategy, year 2015 and 2016

	pair	IR
4	(y0, oi0)	3.49
3	(m0, rm0)	1.18
8	(au0, v0)	0.48
2	(cf0, sr0)	0.04
0	(a0, m0)	-0.56
6	(al0, a0)	-0.62
5	(cu0, rm0)	-0.69
9	(rb0, l0)	-0.75
1	(a0, y0)	-1.17
7	(c0, y0)	-1.23

Table 5.16: IR of 10 pairs for Kalman fitler hedge trading strategy, year 2015 and 2016

	pair	IR
4	(y0, oi0)	6.71
3	(m0, rm0)	5.59
0	(a0, m0)	2.59
5	(cu0, rm0)	1.38
1	(a0, y0)	1.27
6	(al0, a0)	0.11
8	(au0, v0)	-0.32
9	(rb0, l0)	-0.49
2	(cf0, sr0)	-1.11
7	(c0, y0)	-1.26

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this thesis, we introduced the basic concepts of statistical arbitrage and future contracts trading. We introduced stationarity, cointegration, and how basic pricing theories can be used to explain the source of profit for performing statistical arbitrage. We also did a quick review of the handful research works on this subject, pointed out some problems.

As the core of this thesis, we proposed a new model for statistical arbitrage which involves Kalman filter for on-line hedge ratio estimation and MCMC for parameter estimation. We also mathematically derived a new profitability measure for pairs selection, which has benefits over existing measure like distance measure and results from cointegration test since it does not assume constant hedge ratio over the entire period.

We also implemented our trading model, along with two widely used ones and obtained some empirical results. We can conclude that higher sampling frequency does provide better performance for statistical arbitrage while other settings hold the same. This is because higher sampling frequency provides us more opportunities for finding market misprices, where our source of profit comes from. One more observation is that although distance measure and p-value from cointegration test are widely used by researchers and practitioners, these values serve as a poor indicator for pairs selection comparing to historical IR, despite the fact that they are more meaningful mathematically.

At last, we can conclude that our Kalman hedge trading strategy outperforms the other two trading strategies quite a bit, at least for our data set. Much higher IR means the performance of the trading strategy is more robust, the return curve is smoother and the return to risk ratio is significantly higher.

6.2 Future Work

Due to the complexity of MCMC and limited computing resource, we were not able to iterate through all possible pairs in the three year period as what we have done for the other two trading models. We will run more simulations on more data set in the future so that we can compare the performances of the strategies more rigorously. We can also build portfolios based on different measures for pairs selection, and compare the performances of these portfolios to some standard benchmarks such as future indexes, S&P 500 and etc.

Instead of using one measure at a time for pairs selection, we can also try blending the indicators and apply machine learning techniques to find the optimal portfolio for the next trading period. More sophisticated on-line parameter estimation and risk management approaches should also be added in the future to make our trading model more robust and scalable to other markets.

References

[1] Will Acworth. Fia 2016 annual volume survey. http://marketvoicemag.org/?q=content/2016-annual-volume-survey. Accessed: 2017-05-02.

[2] Binh Do, Robert Faff, and Kais Hamza. A new approach to modeling and estimation for pairs trading. In *Proceedings of 2006 Financial Management Association European Conference*, pages 87–99, 2006.

[3] Christian L Dunis, Gianluigi Giorgioni, Jason Laws, and Jozef Rudy. Statistical arbitrage and high-frequency data with an application to eurostoxx 50 equities. *Liverpool Business School, Working paper*, 2010.

[4] Robert F Engle and Clive WJ Granger. Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, pages 251–276, 1987.

[5] Evan Gatev, William N Goetzmann, and K Geert Rouwenhorst. Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies*, 19(3):797–827, 2006.

[6] Si-Yuan Xu Jain-Ming Yang. Institutional investors cultivation and development of china futures markets. http://www.shfe.com.cn/upload/20160816/1471335461440.pdf. Accessed: 2017-05-02.

[7] Kangwhee Kim. Performance analysis of pairs trading strategy utilizing high frequency data with an application to kospi 100 equities. 2011.

[8] Joseph J LaViola. Double exponential smoothing: an alternative to kalman filter-based predictive tracking. In *Proceedings of the workshop on Virtual environments 2003*, pages 199–206. ACM, 2003.

[9] Harry M Markowitz. The early history of portfolio theory: 1600–1960. *Financial Analysts Journal*, 55(4):5–16, 1999.

[10] Purnendu Nath. High frequency pairs trading with us treasury securities: Risks and rewards for hedge funds. 2003.

[11] Stephen A Ross. The arbitrage theory of capital asset pricing. *Journal of economic theory*, 13(3):341–360, 1976.

[12] Ganapathy Vidyamurthy. *Pairs Trading: quantitative methods and analysis*, volume 217. John Wiley & Sons, 2004.

[13] Herman Wold. *A study in the analysis of stationary time series, Second revised edition.* PhD thesis, Almqvist & Wiksell, 1954.

Glossary

backwardation A situation where futures price of a commodity is lower than the expected spot price of the commodity. 26

contango A situation where futures price of a commodity is higher than the expected spot price of the commodity. 26

long A long (or long position) is the buying of an asset with the expectation that price of the asset will rise overtime 1

short A short (or short position), is selling borrowed asset with the expectation that the price of the asset will drop overtime. 1