Energy Commodity Price Analysis and Trading Strategies

by

Bo Liu

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

November, 2015

© Bo Liu 2015

All rights reserved
Abstract

This dissertation studies the three main energy commodities—oil, coal and natural gas—against a backdrop of the US shale revolution and the recent oil market crash. We focus on one fuel in each of the three main chapters, using different methodologies given the particular features of that product.

First we investigate whether the US and UK gas markets had been moving toward integration during the period 2005 to 2014. By introducing a novel measure of distance between the forward curves of the Henry Hub (HH) and National Balancing Point (NBP) indexes, we take full advantage of the information of the highly liquid gas futures markets, to complement the classical cointegration analysis which relies solely on spot prices.

Next we study the world coal markets and their co-movements with oil and gas markets. We apply the principal component analysis (PCA) and pairwise cointegration analysis on a comprehensive coal data set of ten prominent markets around the globe over the years 2004–2015, in order to find out if there is any regional divide between the Atlantic and Pacific basins. We examine the cointegration among the three
energy fuels in different periods of time in both the US and Europe, exhibiting the implications of the US shale boom.

Finally we turn to the relationship between commodity prices and mining companies’ shares prices, in the case of oil sector with its large amount of heavily traded stocks. In order to benefit from “the leverage” of commodity equities, we introduce intraday pairs trading strategies and demonstrate high and stable profitability, especially during period of sharp oil price movements. By using high frequency stock price data, we are able to design trading strategies that capture very short-term market inefficiencies and achieve remarkable returns, even accounting for transaction costs.

Primary Reader: Hélyette Geman
Secondary Reader: Lo-Bin Chang
Acknowledgments

My greatest gratitude goes to my Ph.D. advisor Hélyette Geman. As someone who has aspired to financial mathematics research, I feel quite lucky to have started my time at Hopkins around the same time as Prof. Geman, a world famous expert in the field. I’m honored to be her very first Ph.D. advisee in the US (Nassim Taleb also lives in the US, however!). Her broad knowledge, high efficiency with a hectic schedule, impressive work ethics, fabulous teaching, and super kindness to people all amaze me. I have learned an inordinate amount from her.

Prof. Lo-Bin Chang has been working with me on part of my research during the past year when he was a visiting professor at our department (he is now an assistant professor at Ohio State University). As a Ph.D. graduated only a few years ago, he gave me a lot of useful advices and valuable information on doing research and academia in general. I thank him for all the help and guidance, and flying back for my defense.

I also thank the prestigious economist Prof. Steve Hanke and the renowned statistician Prof. Daniel Naiman for being kind enough to serve on my dissertation
ACKNOWLEDGMENTS

defense committee and GBO exam committee.

My parents named me (my given name approximately means Ph.D. in Chinese) and raised me with high hopes. They supported me unconditionally throughout my school years including graduate study, both financially and spiritually. I’m indebted to them.

Finally, I’m grateful to my wife who quit her job to come to be with me in Maryland for my Ph.D. study. I sincerely thank her for bearing with me and encouraging me the numerous occasions I encountered difficulties and became frustrated.
# Contents

Abstract ii

Acknowledgments iv

List of Tables x

List of Figures xi

1 Introduction 1

2 Are world natural gas markets moving toward integration? Evidence from the HH and NBP forward curves 7

2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

2.2 World natural gas market overview . . . . . . . . . . . . . . . . . . 11

2.2.1 Natural-gas-producing countries . . . . . . . . . . . . . . . . . 11

2.2.1.1 United States . . . . . . . . . . . . . . . . . . . . . . 11

2.2.1.2 Canada . . . . . . . . . . . . . . . . . . . . . . . . . 13
## CONTENTS

4.3 Trading rules ................................................. 82

4.4 Simulation .................................................. 85
   4.4.1 Simulating $L_i$'s and $Y_i$'s ......................... 85
   4.4.2 Choosing parameters .................................. 87
   4.4.3 Simulation trading results ............................. 88

4.5 Back-testing on real data .................................. 90
   4.5.1 The data ................................................. 90
   4.5.2 Transaction costs and return calculation ............ 92
   4.5.3 Empirical results ...................................... 96
     4.5.3.1 June 2013–April 2015 ............................ 96
     4.5.3.2 The year 2008 ...................................... 98

4.6 Conclusion .................................................. 105

A Derivation of likelihood function for $Y(t)$ and the maximum likelihood estimation 107

Bibliography .................................................. 114

Vita .......................................................... 121
List of Tables

2.1 Unit Root Tests .......................................................... 28
2.2 Bivariate Johansen Cointegration Tests ................................ 29
2.3 Stationary Tests for Distances .......................................... 33
2.4 Distances ($/mmBtu) across analysis period, 1 bin=19 weeks ... 35
3.1 Biggest Coal Trade Routes in 2013 and 2014 ......................... 46
3.2 A summary of all data series for this study .......................... 51
3.3 Augmented Dickey–Fuller test statistics ............................... 54
3.4 PCA results: All ten price series ....................................... 55
3.5 PCA results: No US CAPP and China ................................. 57
3.6 Pairwise Johansen cointegration tests results ....................... 61
3.7 Pairwise correlations between the three main fossil fuel prices in the US 63
3.8 Pairwise correlations between the three main fossil fuel prices in Europe 65
3.9 Johansen cointegration tests p-values ................................. 65
4.1 Simulation Trading Results (Top half: the good pair; bottom half: the bad pair) .......................................................... 89
4.2 Oil Company Stocks Descriptions and Statistics (as of December 2014) 91
4.3 Jun’13–Apr’15 Trading Results for Different Pair Selections and Transaction Costs .................................................. 97
4.4 Jun’13-Apr’15 Trading Results by Quarter (Top 50 pairs, transaction costs = 10 bps) .................................................... 100
4.5 2008 Trading Results for Different Pair Selections and Transaction Costs 102
4.6 2008 Trading Results by Quarter (Top 50 pairs, transaction costs = 10 bps) .......................................................... 102
## List of Figures

2.1 Henry Hub (HH) spot prices, January 2005 to April 2014 .......................... 12
2.2 JKM and HH monthly spot prices, April 2012 to March 2014 .................... 17
2.3 NBP spot prices, January 2005 to April 2014 ........................................... 20
2.4 NBP (dark) and HH (light) forward curves .............................................. 22
2.5 NBP (dark) and HH (light) forward curves, continued ............................ 23
2.6 $F$ values for NBP (dark) and HH (light) .................................................. 26
2.7 Spot price values for NBP (dark) and HH (light) ....................................... 27
2.8 Hypothetical forward curves ................................................................. 33
2.9 Distances between NBP and HH from August 13, 2010 to April 10, 2014 .... 34

3.1 The three major coal price indexes, January 2007 to September 2015 ........ 47
3.2 The three major coal price indexes, January 2014 to August 2015 ............. 48
3.3 Peabody Energy and Arch Coal share prices, January 2014 to August 2015 .. 49
3.4 NYMEX coal index vs Peabody and Arch share prices, January 2014 to August 2015 ................................................................. 49
3.5 Percentage changes of NYMEX coal index vs Peabody and Arch Share Prices since January 2014 ................................................................. 50
3.6 Price trajectories for all coal data series, January 2006 to December 2014 ... 53
3.7 The first two principal component loadings of all ten price series (left panel) and the eight series excluding US and China (right panel) .............. 56
3.8 A comparison between the spot prices of US crude oil ($/barrel), natural gas ($/MMBtu) and coal ($/tonne) ................................................. 63
3.9 A comparison between the spot prices of European crude oil ($/barrel), natural gas ($/MMBtu) and coal ($/tonne) ..................................... 64

4.1 Illustration of formation period and trading period .................................. 83
4.2 Simulated 30 days’ $Y_i$’s and $L_i$’s (circles) for the good pair ............... 87
4.3 Back-testing performance for period Jun’13–Apr’15 (top 50 pairs, threshold=98%, transaction costs=10 bps) ................................................. 99
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4</td>
<td>WTI Spot Price 2007–2015</td>
<td>101</td>
</tr>
<tr>
<td>4.5</td>
<td>Back-testing performance for period 2008 (top 50 pairs, threshold=98%,</td>
<td>104</td>
</tr>
<tr>
<td></td>
<td>transaction costs=10 bps)</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

We study in this dissertation the three main energy sources, crude oil, coal and natural gas. As of 2014 they represented together over 80% (oil 31%, coal 29% and natural gas 21%) of global energy supply. They are used for everything from heating and power generation to steel-making and transportation. With increasingly tougher emissions standards, natural gas and low-carbon fuels are gradually replacing oil and coal in the energy mix. The three fossil fuels are projected to each account for about a quarter of global energy demand by 2040, according to the International Energy Agency (IEA).


The biggest story over the past decade in the energy world is the shale revolution, thanks to one of the most innovative technologies of the 21st century: horizontal
CHAPTER 1. INTRODUCTION

drilling and hydraulic fracturing, or fracking, methodologies to unlock natural gas and oil from shale rocks, previously seen as inaccessible. The US shale boom had profound economic and political implications, enhancing US energy independence and reshaping world geopolitical balance, in addition to creating thousands of jobs and boosting local economies.

The first shale oil well was drilled in South Texas’s Eagle Ford field in September 2008, at which time US crude oil production was only 4.7 million barrels a day. From 2009 to 2011, global oil prices recovered from below $40 to around $100 a barrel amid rising demand post financial crisis. At the same time, the US shale industry was booming. New drilling rigs started blanketing shale fields around the country, from Texas to Pennsylvania to North Dakota. US oil production has been steadily increasing, and reached a peak of 9.6 million barrels a day in April 2015. The Organization of the Petroleum Exporting Countries (OPEC) now only exports half the amount of oil to the US as it did at the time of the first shale well. The displaced oil has gone elsewhere in the world, creating a global supply glut.

Besides oil, US natural gas industry is the biggest beneficiary of the shale revolution. There has been a 72% increase in US shale gas production since 2007. The share of shale gas in total US gas production jumped from less than 2% in 2001 to 40% in 2014 Q1, and is projected to grow to 53% by 2040. In 2011, IEA published a special report entitled “Are we entering a golden age of gas?”, suggesting natural gas could play a more prominent role in the global energy mix in the near-term, with surging post-crisis
CHAPTER 1. INTRODUCTION

demand recovery, ample supply and expansion of liquified natural gas (LNG) trades. US gas production has been so strong that old LNG importing terminals have been transformed into exporting terminals. Several projects have applied and acquired federal approvals since 2011. Cheniere Energy’s Sabine Pass terminal in Louisiana will start exporting in late 2015, the first new export project in 46 years.

We study the global natural gas markets in Chapter 2 with a focus on the US and UK markets, home to the two most important natural gas indexes, Henry Hub (HH) and National Balancing Point (NBP). After giving an overview of the world gas markets in 2014, we delve into the question of whether the US and UK markets had been moving towards integration during the period 2005 to April 2014. We contribute to the literature by proposing and defining novel distance measures between forward curves, which are available for commodities with liquid future contracts such as oil and natural gas but not used in classical cointegration studies. Even though the distance measures are defined for natural gas markets, they can be easily extended to other commodities and they have strong implications for spread trading across locational markets.

Oil Market Crash (July 2014 – now)

Besides the booming supply, also contributing to the oil market crash was a weak global demand. In July 2014, oil prices began free-falling, triggered by the opening of two Libyan export terminals. Both the US benchmark WTI and world benchmark
CHAPTER 1. INTRODUCTION

Brent tumbled from around $110 in June 2014 to less than $50 by January 2015. They temporarily rebounded in spring 2015 to above $60 before further plunged to six-year lows in August 2015, with WTI dipping below $40, driving down the broader Bloomberg Commodity Index to the lowest point of this century.

In its November 2014 meeting, OPEC made a somewhat surprising decision not to cut production to save the market, which they did several times in the past in the role of market stabilizer. This time, the cartel member countries, led by the world’s number one oil exporter Saudi Arabia, were determined to endure short-term pain in order to protect their market share from US shale producers. Saudi’s tough stand was opposed by OPEC members in South America and Africa which have higher break-even producing costs, such as Venezuela and Nigeria. In its June 2015 meeting, OPEC again decided to keep output level unchanged at 30 million barrels a day, this time with less resistance within the group.

OPEC appears to have won the price war, at least for now. US crude production fell to about 9 millions a day in October 2015 from the 9.6 millions all-time high reached in April\(^1\) and crude oil imports climbed for three straight months from May to July. 90% of the overall decline comes from Eagle Ford in South Texas, the second most prolific US shale field behind the Permian Basin field in West Texas. In fact, Permian is the only major US shale region where production is still increasing, thanks

\(^1\)Interestingly, the number of US oil drilling rigs in October 2015 have dropped drastically by 64% from the peak of 1609 a year ago, disproportionately to the production falloff. This is because producers have become more efficient and elected to only keep rigs with high outputs in a low price environment.
to its unique combination of plentiful oil, highest quality shale rocks and advanced shipping and storage infrastructure. The global oil supply glut has not been alleviated by US production cutbacks, because other major producers such as Saudi Arabia and Russia are still pumping at high levels. With China’s slowing industrial activities and Iran’s expected ramping-up of production when sanctions against it are lifted after the nuclear deal, oil prices are forecast to average between $50 and $60 in 2016.

Also collapsing are the prices of natural gas and coal. The similar factors are driving the slump: rising supply and weakening global economy, partly due to the slowing expansion speed of China, the world’s growth engine for the past decade, also world’s largest coal producer and consumer and third largest gas consumer and LNG importer. According to the 2015 BP Energy Report, China’s coal import declined in 2014, after five consecutive years’ growths at a breakneck 59% average annual rate. Its 2014 consumption only rose less than 0.1%, compared with the 6.8% annual clip for the past decade. Global coal prices reached eight-year low in late September 2015, driving the share price of Glencore (which acquired in 2013 with a large amount of debt the company Xstrata and its coal assets) to its all time low. The persistent coal market downturn has crushed coal mining companies, forcing miners to close or file for bankruptcy protections. During the six-month period from May to October 2015, the coal mining sector has lost nearly 60% market value.

The fascinating behavior of share prices of commodity mining companies was first investigated by Tufano \[48\] in the case of gold and further studied by Geman
CHAPTER 1. INTRODUCTION

and Vergel [32] for fertilizer-mining companies and Geman [29] for other types of agriculture-related companies. We confirm this line of analysis by examining the sensitivity of coal mining companies shares to coal prices, in Chapter 3, after presenting an outlook of global coal markets as of 2015. We also contribute to the literature by analyzing the integration among coal markets around the globe, as well as between coal and the other two fossil fuel markets, during the period 2004 to September 2015. Principal component analysis (PCA) and novel pairwise cointegration tests were performed on a comprehensive and rich dataset.

We further pursue the idea of commodity prices driving share prices in Chapter 4 in the case of oil, through intra-day pairs trading strategies on twenty-six oil mining companies with largest market capitalizations. We test the hypothesis that the strategies perform better during periods of sharp oil price movements by backtesting over the periods of 2008 and June 2013–April 2015. Our other major contribution in this chapter is to exhibit highly profitable intra-day trading strategies, taking advantage of the recent emergence of high frequency stock price data, which are not available for crude oil prices. A so called ‘doubly mean-reverting’ model is designed from conditional modeling techniques, aiming to capture transient market inefficiencies that are beyond traditional pairs trading strategies with daily frequency.

In the three energy commodity markets, we believe we contributed to the existing literature by choosing a focus in agreement with the economic actuality and the recent financial trading strategies.
Chapter 2

Are world natural gas markets moving toward integration?

Evidence from the HH and NBP forward curves

We present the outlook of the world gas markets in 2014 and investigate whether the US and UK gas markets are moving toward integration. In order to do this, we classically study the cointegration of the Henry Hub (HH) and National Balancing Point (NBP) indexes, but we also introduce the novel concept of distances between forward curves and propose three of them. From all these perspectives, and by analyzing a database covering the period January 2005 to April 2014, we conclude
that no convergence has yet occurred between these two markets.

2.1 Introduction

In recent years, the emergence of fracking and horizontal drilling has made possible the exploitation of huge previously inaccessible shale gas reserves, particularly in the United States. In other countries with shale gas, the production is limited by other factors such as environmental concerns and social acceptance (eg, France) or geology and infrastructure (eg, China).

Between 2006 and 2013, the US total gas production increased by 32%, most of which was shale gas, and the United States is shifting to being an exporting country. In fact, some old liquified natural gas (LNG) importing ports are being transformed into exporting ports. As of March 2014, six projects were authorized by the US Department of Energy to export LNG to countries that do not have free-trade agreements with the United States, such as Japan, India and European Union countries. The first one was Sabine Pass in Louisiana, approved in 2011. Since May 2013 the process has been speeding up: Freeport in Texas, Lake Charles in Louisiana and Dominion Cove Point in Maryland received approvals in 2013. The most recent are Sempra Energys Cameron project in Louisiana and Veresens Jordan Cove in Oregon, approved in February and March 2014, respectively.

These six LNG terminals have a total combined capacity of 9.27 billion cubic
feet (Bcf) per day. There are more than twenty applications under consideration totaling another 19.7 Bcf per day. European governments have pressed the Obama administration to use shale gas export as a way to weaken Russia’s geopolitical power and release its pressure on Europe. On the other hand, there is also resistance to the United States exporting LNG, especially from large chemical companies that have greatly benefited from low gas prices. Until recently, the United States was only exporting natural gas to Canada and Mexico via pipelines.

Over the last decade, there have been a number of studies investigating the price relations between natural gas markets in different regions of the world. Silverstovs et al. [46] were the first to study the degree of integration of international gas markets. They conclude that the North American markets were not integrated with European or Japanese ones during the period 1993–2004. Neumann et al. [42] focused on European spot gas markets for the period 2000–2005. They show that UK and continental European markets had converged via the Interconnector, but individual markets within continental Europe had not. Neumann [41] found an increasing convergence of transatlantic natural gas spot prices from 1999 to 2008 and argued that LNG was the key driver of this convergence. Asche et al. [3] studied European spot gas markets during the period after deregulation, particularly three major hubs: National Balancing Point (NBP) in the United Kingdom, Zeebrugge in Belgium and

---

1On July 30, 2014, a tanker of oil from Texas set sail for South Korea: the first unrestricted sale of unrefined US oil since the 1970s. Analysts say future exports appear wide open, with as much as 800,000 barrels a day coming from just one of the many US fields pumping light oil. According to federal data, US oil production has jumped by about 55% from October 2011, to about 9.1 million barrels per day in October 2014.
CHAPTER 2. NATURAL GAS

the Title Transfer Facility (TTF) in the Netherlands. Their results indicate a highly integrated market, probably through the impact of the oil price.

Several other papers examined the relationship between natural gas and oil prices. Asche et al. [4] study the UK gas market around the time the Interconnector was built. They show that gas prices were integrated with oil and electricity before the Interconnector, but got decoupled afterward. Using 1997–2007 data, Brown and Yücel [10] show that natural gas prices in the United States are driven by movements in crude oil prices, after taking into account the effects from additional factors such as seasonality and weather. Brown and Yücel [11] expand the methodology to both the US and UK markets, concluding that the comovement between gas prices in the United States and the United Kingdom is mediated through crude oil prices. Hartley et al. [34] also find a link between natural gas and crude oil prices in the United States. Unlike most of the literature, they find the link is indirect, acting through the competition between natural gas and residual fuel oil. Ramberg and Parsons [45] examine the notion that natural gas and crude oil prices in the United States had been decoupled. Their results show that the relationship between the two prices is not stable over time: the prices may be related, but the cointegrating relationship could change over time.

Except for Neumann et al. [42] and Neumann [41], who apply the Kalman filter, all of the studies mentioned above use the Johansen cointegration test to investigate price relationships.
CHAPTER 2. NATURAL GAS

We wish to revisit the level of integration of natural gas markets by looking at the two most liquid ones, namely the United States and the United Kingdom. Section 2.2 reviews the current production and flows of natural gas around the world. Section 2.3 looks at the cointegration of the Henry Hub (HH) and National Balancing Point indexes over the period 2005–14. Section 2.4 introduces alternative definitions of distances between commodity markets and utilizes them in the case of the HH and NBP indexes in order to identify a possible march to convergence of the US and UK natural gas markets. Section 2.5 concludes.

2.2 World natural gas market overview

2.2.1 Natural-gas-producing countries

The United States, Argentina, Algeria, Canada and Mexico hold an estimated 80% of the world’s documented shale gas deposits. Canada, which already gets 15% of its natural gas from shale, is ramping up its production.

2.2.1.1 United States

Fracking has started an energy revolution in the United States, and the ripple effects have been felt throughout the world. There has been a 72% increase in US
shale gas production since 2007; in 2014 Q1 shale gas made up 40% of total US gas production, up from less than 2% in 2001; this figure is projected to grow to 53% by 2040. We also note a 47% increase since 2005 in US electricity generated by natural gas, together with a shift away from coal despite the lobbying efforts of the coal industry.

US production growth has been so strong that, upon full political approval, the country can export LNG and turn the regasification terminals which were under construction into liquefaction ones, in order to benefit from the arbitrage opportunities created by the large differences between US prices and those in Europe and Asia.
Figure 2.1 shows the US Henry Hub spot price trajectory over the last decade.

2.2.1.2 Canada

Roughly a dozen LNG projects aimed at Asian buyers have been designed by Canada in the Pacific province of British Columbia.

Canada has two major advantages over its southern neighbor. First, the cold weather there saves large amounts of energy in the production of liquid gas: the average temperature of several planned Canadian LNG sites is only 7°C, compared with 22°C in Louisiana on the US Gulf Coast. This difference means that roughly 25% less electricity is required to transform the gas into liquid. Second, the West Coast’s geographic proximity to Asia means much lower shipping cost. It takes less than eight days to travel from British Columbia to Japan, compared with nineteen days from Henry Hub.

2.2.1.3 Australia

International oil companies such as Chevron and Exxon Mobil are investing over $160 billion in seven Australian LNG projects, including the $52 billion Gorgon megaproject. These terminals under construction are expected to add around 60 million metric tons per year of capacity, propelling Australia to surpass Qatar in 2018 as the world’s largest LNG exporter.

The Gorgon project experienced some work delays and budget blowup in early
CHAPTER 2. NATURAL GAS

2013, due to rising labor costs, harsh weather and a weakening Australian dollar. The first shipments from Gorgon, whose export capacity of LNG will be 15 million metric tons per year, are expected in late 2015.

2.2.1.4 Russia

A traditional natural-gas-exporting powerhouse, Russia is facing challenges in the global gas market as the US shale boom emerges and political changes are shaking the former Soviet Union. First, European gas demand is weakening, due to poor economic performance, cheap coal and renewable subsidies. Moreover, in the wake of the Ukraine crisis, many European countries are trying to reduce their gas dependence on imports from Russia, which supplies 30% of Europe’s gas (51% in the case of Austria and 40% for Hungary). Eastern European countries are especially concerned about the possibility that their neighbor may close down its gas supply during political disputes, as it did in 2009. They are turning to LNG and domestic shale gas: Lithuania is constructing a floating LNG terminal called Independence on the Baltic Sea; Poland has stepped up its effort to develop its 148 trillion cubic feet (Tcf) shale gas reserves, the largest in Europe.

As a result, Russia is turning its focus to Asia, the biggest and fastest growing LNG market in the world, in an effort to compete with supply from Australia and North America.

In 2012, only 7% of Russia’s gas was exported to Asia (all in the form of LNG).
CHAPTER 2. NATURAL GAS

But there are three projects in the planning or construction phases: Gazprom’s Vladivostok project, Novatek’s Yamal project and Rosneft’s Sakhalin project. These would increase Russia’s LNG capacity from 10 million metric tons per year to 45 million metric tons by 2020.

The Yamal LNG project will be the first Russian project to link Arctic gas to Asia. Novatek plans to start production at the planned liquefaction plant by 2017; this will have an annual capacity of 16.5 million metric tons. At a cost of $20 billion, Yamal LNG has been jointly developed by the China National Petroleum Corporation and France’s Total, which each hold a 20% stake.

2.2.1.5 Mozambique

The East African country Mozambique is emerging as a future LNG-exporting hub with substantial reserve discoveries. About 150 Tcf of gas have been found offshore, and the government thinks there is potential to double that estimate.

The US company Anadarko is planning an LNG project off the coast of Mozambique with an eventual capacity of 50 million metric tons a year. The company is on track to ship the first cargo from there in 2018. It plans to start with four liquefaction plants, also known as trains, and build as many as ten trains. Anadarko appears to be open to new pricing mechanisms.
CHAPTER 2. NATURAL GAS

2.2.1.6 Qatar

Qatar is the world’s largest LNG exporter and has the world’s largest non-associated gas field, North Field. Since 1997, Qatar has delivered nearly 2000 LNG cargoes to Japan, the world’s largest LNG importer. Qatar’s Qatargas and RasGas are the two largest LNG-producing companies in the world.

2.2.2 Natural-gas-importing countries

Asia represents an increasing share of world’s LNG demand, with Japan and Korea being the largest importing countries and China and India the fastest growing ones.

2.2.2.1 Japan

All Japan’s nuclear power plants were shut down in the wake of the Fukushima nuclear disaster in March 2011. Japan had to sharply increase its LNG imports to generate electricity to make up for the idle nuclear reactors. During the two years following the disaster, Japan’s annual LNG imports rose by 23%. In 2012, Japan paid $64 billion for a record 87.3 million metric tons of LNG, costing an average $16.70 per MMBTU, which is almost six times as much as the $2.83 average price of US gas in 2012.

At the time of writing, Japan planned to list the world’s first futures contract for LNG on the Tokyo Commodity Exchange in March 2015. The futures will be cash
settled in US dollars and be based on a price index for spot LNG cargoes delivered to Japan. The contracts will also allow exchange futures for physical (EFP) transactions that allow buyers to swap their futures positions for a physical holding.

The Japan Korea Marker (JKM) is the LNG benchmark price assessment published by Platts for physical cargoes delivered into Japan and South Korea. Figure 2.2 shows the JKM price over the past two years, compared with HH spot price.
2.2.2.2 China

As stated above, China and India are the two largest LNG growth markets. The LNG imports increased from 137 Bcf and 352 Bcf for China and India, respectively, in 2007, to 706 Bcf and 708 Bcf in 2012. China is expected to add 853 Bcf of new regasification capacity by 2016, raising its total capacity to 1930 Bcf per year. India is expected to add 731 Bcf per year of capacity to its existing total of 975 Bcf by 2017.

The Chinese government has vowed to address the air pollution problem, largely due to the huge coal combustion, in their major cities. Natural gas will play a major role in that effort. By 2018 China will account for 30% of the growth of global gas demand, which will make it increasingly dependent on imports despite its progress in domestic production. According to the International Energy Agency, in the next five years China will absorb the entire production increase from Central Asia as well as one-third of the global increase in LNG supply.

2.2.3 The case of the United Kingdom

Natural gas from fields in the North Sea and Irish Sea provides around 40% of the United Kingdom’s gas supply. However, production from these fields is now in decline and the United Kingdom is importing more and more gas from abroad.

Currently, a major way of importing gas is through the four pipelines that run
CHAPTER 2. NATURAL GAS

under the sea from mainland Europe to the British mainland:

(1) the UK–Belgium Interconnector, which has an import capacity of 901 Bcf/yr;

(2) the UK–Netherlands pipeline, with an import capacity of 502 Bcf/yr;

(3) the pipeline between Scotland and Norway, with the same 502 Bcf/yr capacity;

(4) the Langeled pipeline from Norway to Yorkshire, with a 929 Bcf/yr capacity.

Britain has a remarkable storage capacity of 141 Bcf in depleted fields and salt caverns. It also has three LNG import facilities, together capable of meeting nearly 50% of annual demand. Total gas demand in 2014 was 851 terawatt-hours (TWh) (or 2783 Bcf), out of which 424 TWh (1388 Bcf), roughly half the annual consumption, came from national production; 103 TWh (335 Bcf), or one-eighth of the annual consumption, came from LNG imports, and the remainder came from Norway and Belgium through pipelines. Because of its central location between the United States and Asia, the United Kingdom has a remarkably diverse choice in terms of satisfying its natural gas demand. Figure 2.3 shows the UK National Balancing Point spot price trajectory over the last decade.
Figure 3: National Balancing Point (NBP) spot prices
Jan 2005 to Apr 2014

Figure 2.3: NBP spot prices, January 2005 to April 2014
2.3 Data and cointegration analysis

In this section we perform a cointegration analysis of a data set of US and UK prices over a period that, despite being fairly recent, is large enough to be conclusive.

2.3.1 Data

The data we used for this study was downloaded from the Bloomberg terminal as the NBP (United Kingdom) and HH (United States) daily forward curve data. The full analysis period ranges from January 3, 2005 to April 10, 2014. Weekly data were calculated as simple arithmetic averages of the four or five trading days within the week for each maturity on the forward curves.

There were at least 61 liquid maturities for HH in the entire period and for NBP only after August 13, 2010. Before that date, only 10–18 months of maturities are available.

The spot price data is not available for the NBP. We used the first nearby as a proxy for the spot price in both cases for consistency, even though HH spot data were available on the US Energy Information Administration (EIA) website.

In figure 2.4–2.5 we plot the two forward curves using data collected twice a year in order to account for the seasonality in both gas markets; the higher ones represent the NBP throughout the period.

We note that the seasonality depicted in the NBP forward curves becomes much
Figure 2.4: NBP (dark) and HH (light) forward curves
Figure 2.5: NBP (dark) and HH (light) forward curves, continued
smoother over the course of the year 2011, reflecting a greater liquidity in the trading of NBP futures.

The New York Mercantile Exchange (NYMEX) Henry Hub forward curves have exhibited a smooth seasonality for more than a decade.

2.3.2 Introducing the average forward price \( \bar{F} \)

In order to model commodities’ forward curve dynamics, Borovkova and Geman introduced a new state variable representing the average of liquid forward prices. This quantity, denoted by \( \bar{F} \) and capturing the entirety of the forward curve, was meant to replace the previously widely used spot price \[ 33, 25 \] when managing a portfolio of futures positions.

In agreement with the geometric Brownian motion based reference models, \( \bar{F} \) was defined as the geometric average (quite close to the arithmetic average, in fact) of all liquid forward contract prices:

\[
\bar{F}(t) = \sqrt[N]{\prod_{T=1}^{N} F(t, T)},
\]

where \( N \) is a multiple of twelve in the case of a seasonal commodity.

Compared with the spot price, \( \bar{F} \) has several advantages. First, it is the “center of gravity” of the forward curve, a more robust quantity able to capture all the information available at time \( t \) in the quite liquid forward market. Second, spot prices
CHAPTER 2. NATURAL GAS

may not be available or may be opaque for some commodities. Third, $\bar{F}$ absorbs the
seasonality if we use a multiple of twelve maturities when calculating the average.
Lastly, it is the quantity (up to the discount factors) that underlies a commodity
swap spanning the period up to the maturity of the swap; and commodity swaps are
experiencing a gigantic liquidity in energy markets. In the same way, $\bar{F}$ reflects the
anticipation of the average of the spot prices at stake in the Asian options traded on
a number of commodity exchanges.

The $\bar{F}$ values for NBP and HH are plotted together for the full period in Figure
2.6 while spot prices are displayed in Figure 2.7; unsurprisingly, the spot trajectories
are more jagged. Figure 2.7 also illustrates that the NBP and HH prices were much
closer in the years 2005–9; they have gone their separate ways since 2010 because of
the US shale boom.

2.3.3 Unit root tests

Before any cointegration study, the degree of integration is needed for each univariate
time series. A series is called integrated with degree $d$, i.e., an $I(d)$ process, if it is
stationary after differencing exactly $d$ times. $I(1)$ processes are also called unit root
processes.

We apply the augmented Dickey–Fuller test to four series $S_{NBP}, S_{HH}, \bar{F}_{NBP}$ and
$\bar{F}_{HH}$, on both the level and first differences. As shown in Table 2.1 the tests fail to
reject the hypothesis that the data have unit roots but reject the hypothesis that the
Figure 2.6: $\bar{F}$ values for NBP (dark) and HH (light)
Figure 7: S_NBP and S_HH (light)

Figure 2.7: Spot price values for NBP (dark) and HH (light)
 CHAPTER 2. NATURAL GAS

Table 2.1: Unit Root Tests

(Weekly data, Jan 4, 2005 to Apr 7, 2014)

<table>
<thead>
<tr>
<th>Augmented Dickey–Fuller Tests</th>
<th>Variables</th>
<th>Levels</th>
<th>First Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$S_{NBP}$</td>
<td>-2.8273+</td>
<td>-9.4740**</td>
</tr>
<tr>
<td></td>
<td>$S_{HH}$</td>
<td>-2.1295</td>
<td>-6.8470**</td>
</tr>
<tr>
<td></td>
<td>$\bar{F}_{NBP}$</td>
<td>-2.3778</td>
<td>-9.6334**</td>
</tr>
<tr>
<td></td>
<td>$\bar{F}_{HH}$</td>
<td>-1.1695</td>
<td>-3.8158**</td>
</tr>
</tbody>
</table>

+, * and ** denote significance at better than 10%, 5% and 1% respectively.

first differences have unit roots. Therefore, all four series are unit root processes.

2.3.4 Johansen cointegration tests

Several nonstationary series are called cointegrated if some linear combination of the series is an integrated series of a lower order. The most common case is that where two difference-stationary series are cointegrated; that is, a linear combination of two $I(1)$ processes is stationary.

We are interested in the cointegration relations of the two pairs of natural gas prices: the spot pair ($S_{NBP}$ and $S_{HH}$) and the $\bar{F}$ pair ($\bar{F}_{NBP}$ and $\bar{F}_{HH}$).

We use the Johansen cointegration procedure, as detailed in [37]. The results are given in Table 2.2. Neither the spot prices nor the $\bar{F}$ pair are cointegrated at any significance level.

In order to reinforce or challenge the message provided by the cointegration analysis of the spot and average forward prices, we turn to other ways of quantifying
CHAPTER 2. NATURAL GAS

Table 2.2: Bivariate Johansen Cointegration Tests

(January 2005 to April 2014)

<table>
<thead>
<tr>
<th></th>
<th>$S_{NBP}$ and $S_{HH}$</th>
<th>$\bar{F}<em>{NBP}$ and $\bar{F}</em>{HH}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$: rank $p$</td>
<td>Trace Statistics</td>
<td>Max Eigenvalue Statistics</td>
</tr>
<tr>
<td>$p = 0$</td>
<td>5.5258</td>
<td>4.6826</td>
</tr>
<tr>
<td>$p \leq 1$</td>
<td>0.8431</td>
<td>0.8431</td>
</tr>
</tbody>
</table>

Neither the spot prices nor the $F$ pair are cointegrated at any significance level.

the “distance” of the US and UK natural gas markets.

2.4 Introducing measures of distances between two commodity markets

Forward markets play a key role in commodity trading activities. Hence, a proper analysis of the integration of two different markets should capture the message provided by the corresponding forward curves. More generally, in this section we propose different ways of measuring whether two commodity markets are close or distant.
CHAPTER 2. NATURAL GAS

2.4.1 Distances between two commodity markets

We propose three definitions of distances between two commodity markets and will apply them to those of NBP and HH indexes.

Let $F_1(t, i)$ and $F_2(t, i)$ be the forward prices of two commodities with a maturity of $i$ months at date $t$; let $S_1(t)$ and $S_2(t)$ be spot prices at date $t$.

We first define the $\bar{F}$ distance

$$D_1(t) = |\bar{F}_1(t) - \bar{F}_2(t)|.$$  

This is the absolute value of the difference between the two curves’ averages. In the period with full data (after August 2010), the NBP forward curve is always higher than the HH forward curve. Hence, the absolute value disappears and $D_1$ is approximately the arithmetic average of the distances between each pair of forward prices of the same maturity: a very natural distance to introduce.

The second distance is the simple spot distance

$$D_2(t) = |S_1(t) - S_2(t)|,$$

where again the absolute value can be deleted for the period after August 2010.

In order to account for possible differences in slopes of forward curves, we add a
“slope term” to the first distance. We first define the slope as the difference between the average of the twelve furthermost points and the twelve most nearby points on the forward curve. It is expressed in dollars, hence is additive to $\bar{F}$ or spot prices:

$$\text{dollar slope} = \sqrt{\frac{1}{12} \prod_{T=50}^{61} F(t,T) - \frac{1}{12} \prod_{T=2}^{13} F(t,T)}.$$ 

We define the “dollar slope” as the difference between the average of the long-term futures prices and the average of the short-term futures prices in order to benefit from the large number of liquid maturities available as well as remove any undesirable seasonality effect. Note that in the case of an approximately linear forward curve, this dollar slope coincides with the famous spread analyzed in the commodities literature. The spread between a distant futures contract and the spot was first introduced by Working [50] as a representation of inventory—the spread being negative in the case of a low inventory. It was later used by Fama and French [26], who analyzed twenty-one commodity markets in which the inventory numbers were not available. Geman and Ohana [31] studied a database of US oil and gas prices and inventories, and showed that there is a high correlation between spreads and inventories.

We now introduce a third distance that mixes both the average forward prices and the shapes of the forward curves, namely

$$D3 = |\bar{F}_1(t) - \bar{F}_2(t)| + \frac{1}{2} |\text{dollar slope}_1 - \text{dollar slope}_2|.$$ 

2The first nearby is avoided as it is very volatile.
Note the following (see Figure 2.8).

(a) In the case of two parallel forward curves, the distance $D_3$ is identical to $D_1$ since the second term is zero.

(b) In the case of approximately linear and parallel forward curves, both increasing for instance, the distance $D_3$ is again reduced to $D_1$.

(b’) In the case of approximately linear forward curves with one increasing and the other decreasing, the first term in $D_3$ is zero and the value of $D_3$ is the same as that obtained in (b), a satisfactory result since the forward curves are in both cases not “integrated”, but for different reasons. The division by 2 of the dollar-slope term in $D_3$ was necessary to recover these same values of $D_3$ in (b) and (b’), a property we view as desirable.

2.4.2 The case of HH and NBP: were distances declining over the period 2010–14?

As mentioned before, the sixty-one months of complete NBP forward curve data do not include the period January 2005–August 2010. In order to be able to calculate $D_3$, and to keep the approach consistent across the time horizon for $D_1$, we focus our distance analysis on the period with full NBP data, namely August 13, 2010 to April 10, 2014.
In Figure 2.9, we plot all distances using weekly data in the usual unit of $/MMBTU. Compared with the brisk spot prices moves, the distances seem more stable. The augmented Dickey–Fuller tests show that they are all stationary, as displayed in Table 2.3: $D_1$ and $D_2$ at the 5% significance level, and $D_3$ at the 10% significance level.

The fact that these distances are stationary even though all $F$ and spot series of

<table>
<thead>
<tr>
<th>Variables</th>
<th>Levels</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_1$</td>
<td>-3.2648</td>
<td>0.0222*</td>
</tr>
<tr>
<td>$D_2$</td>
<td>-3.1151</td>
<td>0.0324*</td>
</tr>
<tr>
<td>$D_3$</td>
<td>-2.9153</td>
<td>0.0520+</td>
</tr>
</tbody>
</table>

+ , * and ** denote significance at better than 10%, 5% and 1% respectively.
Figure 2.9: Distances between NBP and HH from August 13, 2010 to April 10, 2014
CHAPTER 2. NATURAL GAS

Table 2.4: Distances ($/mmBtu) across analysis period, 1 bin=19 weeks

<table>
<thead>
<tr>
<th>Bin</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_1$</td>
<td>3.73</td>
<td>5.32</td>
<td>6.38</td>
<td>5.82</td>
<td>6.53</td>
<td>6.07</td>
<td>6.23</td>
<td>5.73</td>
<td>6.09</td>
<td>6.07</td>
</tr>
<tr>
<td>$D_2$</td>
<td>3.74</td>
<td>5.03</td>
<td>5.02</td>
<td>6.05</td>
<td>6.76</td>
<td>6.25</td>
<td>7.12</td>
<td>6.13</td>
<td>7.11</td>
<td>5.66</td>
</tr>
<tr>
<td>$D_3$</td>
<td>4.06</td>
<td>5.52</td>
<td>6.55</td>
<td>6.20</td>
<td>7.13</td>
<td>6.43</td>
<td>6.74</td>
<td>6.32</td>
<td>6.74</td>
<td>6.34</td>
</tr>
</tbody>
</table>

the same period are not, is strong evidence that HH and NBP markets are neither converging nor diverging.

This can also be seen in Figure 2.9: in the last three years, $D_1$ and $D_3$ oscillate around the same levels. The big drop at the end for $D_2$ is a momentary anomaly due to the combined effects of a warm winter in the United Kingdom (hence excessive inventory and declining gas prices) and the opposite situation in the United States.

Lastly, to gain a clear perspective of the situation in numbers, we report a summary of distances. The analysis period of 190 weeks is equally divided into ten “bins” of 19 weeks. The average distances within each bin are shown in Table 2.4. The distances are in fact lowest at the beginning of the period, as shown in Bin #1. Then the two trajectories, for both the spot and the $\bar{F}$, start to decouple, as shown in Figure 2.6 on page 26 and Figure 2.7 on page 27. By the third period, all distances remained relatively constant across bins: $D_1$ around $6$, $D_3$ around $6.5$ and $D_2$ between $5$ and $7.2$. 

35
CHAPTER 2. NATURAL GAS

2.5 Conclusion

In this chapter we asked whether the march toward integration of world natural gas markets has already started, by investigating the United States and the United Kingdom, the most liquid and developed gas markets. To this end, we analyzed a database of spot and forward curves of the Henry Hub and National Balancing Point Indexes over the period January 2005 to April 2014. In order to go beyond the cointegration analysis used in the existing literature on the subject, we introduced the novel concept of distance between two commodity markets. In fact, we defined three distances, two of them meant to strongly capture the signals provided by the forward curves. Using these different perspectives, our answer, as of April 2014, is no.

Obviously, the ongoing increase in US shale gas production, the building of new fleets of LNG tankers and changes in the US export policy will change the situation over time.
Chapter 3

World Coal Price Analysis and Co-Movements with Oil and Natural Gas Markets

The goal of this chapter is fourfold: i) describe the world coal markets; ii) introduce the joint analysis of coal mining companies’ share prices and coal prices; iii) investigate over the period 2004–2015 whether coal markets are integrated and whether there is any regional divide within the changing landscape of global energy economy; iv) analyze over the same period the co-movements of coal, natural gas and crude oil in the US and in Europe. Principal component analysis and pairwise cointegration analysis are performed on a comprehensive data set of ten prominent regions around the globe on both the supply and demand sides. Empirical results show that the world
coal market is generally integrated with no clear separation between the Atlantic and Pacific basins. From the cross-fuel analysis of the three main energy sources crude oil, coal and natural gas in both the US and Europe, we conclude that they mostly decoupled in 2010, except for US natural gas and US coal, which are cointegrated probably because of the recent dramatic decline in both commodity prices.

### 3.1 Introduction

Coal is the predominant fuel for electricity generation worldwide and a main source for global energy supply. In 2012, coal accounted for 29% in world total primary energy supply, following only oil with 31% [36]. Coal-fired plants’ share in worldwide electricity generation was 40% in 2010 and is projected to fall to 36% in 2040 [23], as the world aims to reduce the reliance on coal in power generation amid environmental concerns and a boom in shale gas supply. In the US, the figure was 39% in 2013 and is similarly forecasted to drop to 34% in 2040 [24]. Yet, coal will remain a key element in the world energy picture in the foreseeable future.

There are many types of coal depending on the contents of carbon, ash, sulfur and moisture. Generally speaking, the main ranks of coal are, from high to low quality, anthracite, bituminous, sub-bituminous and lignite. The first two are called hard coal and the last two brown coal. A more common classification method is by usage: steam coal (also known as thermal coal) for electricity generation and coking coal (also
known as metallurgical coal or met coal) for steel production. Only the bituminous
can be used as coking coal, which must first be “coked” to remove volatile
components before the industrial process, hence the name. Steam coal comprises
anthracite, other bituminous, and sub-bituminous coal. Its production dominates
coking coal and lignite coal, making up 70-75% of global yearly production in the last
decade, while the other two coal types contribute 10-15% each. The dominance of
steam coal is even stronger in the US, with an over 85% share. Consequently, we
focus on steam coal for price analysis. In the world market overview section however,
we use total coal since fundamental data (production, consumption and trades) are
usually reported in aggregate.

The world coal market is traditionally divided into the Atlantic region and the
Pacific region. The latter has been playing a more important role in international
coal trade market in recent years, with the emergence of China and India on the
demand side and Indonesia on the supply side. In this chapter, we perform a
comprehensive analysis of world coal markets using price data from markets around
the globe, in order to find out if the recent shale boom in oil and gas industry had any
fundamental impact on the coal market. Specifically, we want to answer the question:
are the two regions divided or integrated?

Our data for this study covers ten countries: China, Japan, Korea, Australia and
Indonesia in the Pacific market and USA, Colombia, South Africa and the Netherlands
in the Atlantic market. There are two data series for Russia—Vostochny in the
CHAPTER 3. COAL

Pacific region and Baltic in the Atlantic region. This list of countries includes all the important players on both the demand and supply sides, with the only exception of India. Netherlands is included in our data despite its not being a top importer or exporter, because the API 2 index—a CIF index including cost of insurance and freight to the ARA (Antwerp-Rotterdam-Amsterdam) region of Northwestern Europe—is the most traded coal index. Compared to the Pacific market, the Atlantic market is more dominated by derivatives and paper trades instead of spot physical trades [43]. In 2010, the size of paper trade in the Atlantic region was around 10 times as big as the physical market [35]. Along with API 2, another equally important coal index is API 4 which is based on FOB (free on board) prices at Richards Bay in South Africa, also in the Atlantic region.

There have been several papers in the past decade investigating the integration of the international steam coal market. In two early studies, Ekawan and Duchêne [21] and Ekawan et al. [22] describe the hard coal markets in the Atlantic and Pacific regions respectively, and allude to the integration between the two markets, without empirically testing the hypothesis. In 2006, Wårell [52] examines whether the European and Japanese markets were integrated over the period 1980–2000 for both the steam coal and coking coal markets using quarterly import data. For steam coal, the result supported the hypothesis of an integrated market. In order to test market integration over time, the author also applied cointegration tests to two sub-periods, the 1980s and the 1990s and concluded that no integration could be confirmed for
CHAPTER 3. COAL

the 1990s. Li et al. [38] investigated the hypothesis of a single economic market for the international steam coal industry. They used multiple methods including cointegration tests and the Kalman filter, with monthly FOB export data in six countries (Australia, China, Colombia, Indonesia, Poland and South Africa) between 1995–2007. They concluded that, in general, the international steam coal market could be seen as integrated over that time period. Zaklan et al. [53] in 2012 analyze the integration of international steam coal trade using weekly import, export and transportation data from 2001 to 2009. They use a principal component analysis (PCA) and Johansen cointegration tests and concluded that there was a significant, yet incomplete, integration. Papież and Śmiech [43] in 2013 use instead a causality methodology to investigate the international steam coal market integration, especially dependencies between different markets. They used weekly export and import data in seven markets (Indonesia, Australia, Japan, Korea, Netherlands, South Africa and Colombia) from 2002 to 2011. They concluded that the most important factor in the Pacific market was Australian coal; while in the Atlantic market, ARA and Richard Bay prices had the biggest influence on other prices—results which confirm the role of Australia and South Africa as a major producing country in each region. In a recent study of 2015, Papież and Śmiech [44] conducted rolling cointegration analysis on a weekly data set of six markets from 2001 to 2014 and found that these markets were integrated, especially during the period when freight costs were lower.

This chapter contributes to the literature in several aspects. First, we used a more
CHAPTER 3. COAL

comprehensive and richer data set than previous studies. The ten countries we cover include emerging players in global coal trade market such as China, as well as big exporters which provide coal to both Asian and European markets, namely Russia and the US. These countries are often omitted in existing literature. By using the prompt month Future price as a proxy for the spot price, we are able to incorporate several daily price series, especially the US, for which an import or export index is unavailable. Second, the pairwise cointegration analysis we conduct has two major advantages over the conventional multivariate cointegration tests. The test results of 49 pairs provide us with a deeper understanding of the inter-relationship of markets in the Atlantic and Pacific basin. Another advantage of using pairwise tests over multivariate tests is that it allows us to utilize the maximal amount of data for each pair. Lastly, by studying the cointegration between crude oil, coal and natural gas, we shed some light on the changing landscape of global energy markets amid falling world coal demand and the booming shale industry in crude oil and natural gas.

The rest of the chapter is organized as follows. Section 3.2 gives an overview of today’s world coal markets. Section 3.3 analyzes the joint behavior of coal companies’ shares prices and coal prices, in the line of our discussion in Chapter 1 for all commodities and in Chapter 4 for crude oil. Section 3.4 describes our coal prices database and tests the stationarity of the data. In section 3.5 and 3.6 respectively, a PCA analysis and cointegration analysis are conducted separately to find out whether coal Pacific and Atlantic regions are integrated. Price relationships among the three
3.2 World coal market overview

During the period 2004–2014, and despite the signature of the Kyoto Protocol by a number of countries outside the US, world coal production increased by 42.2%, from 5743.6 Mt (Million tonnes) to 8164.9 Mt, according to [1]. The top three producers in this period were China, the US and Australia, followed by India, Indonesia, Russia and South Africa. These seven countries contributed at least 84% of global production every year. China’s production steadily increased from 38.2% of world total in 2004 to 46.9% in 2014. Meanwhile, the US share decreased from 20.6% to 12.9%, largely due to the shale boom in oil and natural gas industries, as many power plants switched from coal to natural gas for their choice of fuel.

On the demand side, China, US, India and Japan were the top four countries, in this order, in terms of consumption every year. They consumed more than 70% of the world total. In most years, the next four were Russia, South Africa, South Korea and Germany in some order.

The international coal trade volume increased by 65.7% from 760.1 Mt (or 13.2% of world production) to 1296.4 Mt (15.9% of world production) over the same period.\footnote{Based on the UN Comtrade database at: http://comtrade.un.org/data/1} Japan and Korea were two of the largest importers of coal, each consisting of at
least 10% of global trade volume every year. Meanwhile, China and India gained an increasing presence in international coal trade market, both cracking into top four in imports in 2007 and stayed there since. These four energy-hungry Asian countries imported over 60% of global trade volume in 2014, led by China with 19.5% of world total. In fact, China was one of the top \textit{exporting} countries in the early years of our analysis period (number three exporter in 2004 for instance) and was still a net coal exporter as recently as 2008. It then successively overtook Korea as the number two importer in 2009 and Japan as the number one in 2011.

The exporting market is more concentrated. Australia and Indonesia have been the top two exporters, followed by Russia, USA, South Africa and Colombia (plus China in the early years). The percentage of global coal exports from these six countries was around 90% in the past few years, rising from 74% in 2004. During the period, Indonesia gained the most market share, surpassing Australia as world’s biggest exporter in 2011 after exporting less than half (13.9%) of the then-world-leading Australia (29.6%) in 2004. In 2014, Australia regained the number one status from Indonesia: the two countries—both in the Pacific region—exported 29.9% and 27.5% of global trade volume respectively.

To have a clearer picture of the relative contributions of the Pacific and Atlantic regions to the world coal trade volume, we compiled a list of biggest trade routes\footnote{A trading route here means a unique pair of countries. The volume of each route is the total coal export from the first country to the second.} ranked by trade volume in both 2013 and 2014, shown in table 3.1. There are 38 routes
CHAPTER 3. COAL

In 2013 and 33 in 2014 with more than 0.5% of global trading volume of the year, each making up three quarters of world’s total. Not surprisingly, the biggest routes are from the top two exporters (Indonesia and Australia) to the top four importers (China, Japan, India and Korea)—all within the Pacific region. In fact, the eight routes between these countries are the top eight routes in 2013 and eight of the top nine in 2014, accounting for about 44% of the global trade volume in each year.

In the Atlantic region, the biggest routes are Russia to UK in 2013 (1.8% of world total) and Russia to Switzerland in 2014 (7.6%). South Africa to India is the biggest cross-regional routes in both years, with 1.7% and 2.5% world total respectively. Overall, among the routes listed, the Pacific market contributes 51.3% of two year’s global trade volume while 12.8% is from the Atlantic market. Cross-regional routes have a smaller presence of 6.0% of world’s total and the rest are non-seaborne routes (such as Mongolia to China), totaling 4.6% of global volume. This illustrates the dominance of the Pacific market over the Atlantic market in physical coal trading—while the relationship is the opposite in the paper market, as said before.

Prices during the years 2007–2015 are displayed in figure 3.1.
Table 3.1: Biggest Coal Trade Routes in 2013 and 2014

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>MT</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>Japan</td>
<td>124.6</td>
<td>9.7% Pacific</td>
</tr>
<tr>
<td>Indonesia</td>
<td>Japan</td>
<td>116.8</td>
<td>9.1% Pacific</td>
</tr>
<tr>
<td>Indonesia</td>
<td>China</td>
<td>89.8</td>
<td>7.6% Pacific</td>
</tr>
<tr>
<td>Australia</td>
<td>China</td>
<td>88.1</td>
<td>6.8% Pacific</td>
</tr>
<tr>
<td>Australia</td>
<td>Rep. of Korea</td>
<td>49.9</td>
<td>3.9% Pacific</td>
</tr>
<tr>
<td>Indonesia</td>
<td>Japan</td>
<td>37.8</td>
<td>2.9% Pacific</td>
</tr>
<tr>
<td>Indonesia</td>
<td>Rep. of Korea</td>
<td>36.1</td>
<td>2.8% Pacific</td>
</tr>
<tr>
<td>Australia</td>
<td>India</td>
<td>34.8</td>
<td>2.7% Pacific</td>
</tr>
<tr>
<td>Kazakhstan</td>
<td>Russian Fed.</td>
<td>25.3</td>
<td>2.0% Non*</td>
</tr>
<tr>
<td>Russian Fed.</td>
<td>China</td>
<td>25.1</td>
<td>1.9% Pacific**</td>
</tr>
<tr>
<td>Russian Fed.</td>
<td>UK</td>
<td>23.4</td>
<td>1.8% Atlantic</td>
</tr>
<tr>
<td>South Africa</td>
<td>India</td>
<td>21.3</td>
<td>1.7% Cross*</td>
</tr>
<tr>
<td>Mongolia</td>
<td>China</td>
<td>18.2</td>
<td>1.4% Non</td>
</tr>
<tr>
<td>Indonesia</td>
<td>Malaysia</td>
<td>17.1</td>
<td>1.3% Pacific</td>
</tr>
<tr>
<td>Colombia</td>
<td>Netherlands</td>
<td>15.6</td>
<td>1.2% Atlantic</td>
</tr>
<tr>
<td>Russian Fed.</td>
<td>Rep. of Korea</td>
<td>14.5</td>
<td>1.1% Pacific</td>
</tr>
<tr>
<td>Indonesia</td>
<td>Philippines</td>
<td>14.5</td>
<td>1.1% Pacific</td>
</tr>
<tr>
<td>Indonesia</td>
<td>Thailand</td>
<td>14.3</td>
<td>1.1% Pacific</td>
</tr>
<tr>
<td>South Africa</td>
<td>China</td>
<td>13.1</td>
<td>1.0% Cross</td>
</tr>
<tr>
<td>Indonesia</td>
<td>China, Hong Kong</td>
<td>12.9</td>
<td>1.0% Pacific</td>
</tr>
<tr>
<td>Russian Fed.</td>
<td>Japan</td>
<td>12.5</td>
<td>1.0% Pacific</td>
</tr>
<tr>
<td>USA</td>
<td>UK</td>
<td>12.3</td>
<td>1.0% Atlantic</td>
</tr>
<tr>
<td>USA</td>
<td>Netherlands</td>
<td>11.5</td>
<td>0.9% Atlantic</td>
</tr>
<tr>
<td>Canada</td>
<td>China</td>
<td>11.1</td>
<td>0.9% Cross</td>
</tr>
<tr>
<td>Russian Fed.</td>
<td>Ukraine</td>
<td>10.6</td>
<td>0.8% Non</td>
</tr>
<tr>
<td>Canada</td>
<td>Japan</td>
<td>10.2</td>
<td>0.8% Cross</td>
</tr>
<tr>
<td>Colombia</td>
<td>UK</td>
<td>9.0</td>
<td>0.7% Atlantic</td>
</tr>
<tr>
<td>Russian Fed.</td>
<td>Turkey</td>
<td>9.0</td>
<td>0.7% Non</td>
</tr>
<tr>
<td>Netherlands</td>
<td>Germany</td>
<td>8.3</td>
<td>0.6% Non</td>
</tr>
<tr>
<td>Colombia</td>
<td>Turkey</td>
<td>8.2</td>
<td>0.6% Atlantic</td>
</tr>
<tr>
<td>USA</td>
<td>Brazil</td>
<td>7.8</td>
<td>0.6% Atlantic</td>
</tr>
<tr>
<td>USA</td>
<td>Rep. of Korea</td>
<td>7.6</td>
<td>0.6% Cross</td>
</tr>
<tr>
<td>Colombia</td>
<td>Chile</td>
<td>7.6</td>
<td>0.6% Atlantic</td>
</tr>
<tr>
<td>Canada</td>
<td>Rep. of Korea</td>
<td>7.5</td>
<td>0.6% Cross</td>
</tr>
<tr>
<td>USA</td>
<td>China</td>
<td>7.5</td>
<td>0.6% Cross</td>
</tr>
<tr>
<td>South Africa</td>
<td>Netherlands</td>
<td>7.0</td>
<td>0.5% Atlantic</td>
</tr>
<tr>
<td>Australia</td>
<td>Netherlands</td>
<td>6.8</td>
<td>0.5% Cross</td>
</tr>
<tr>
<td>Colombia</td>
<td>USA</td>
<td>6.7</td>
<td>0.5% Atlantic</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>MT</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indonesia</td>
<td>India</td>
<td>137.5</td>
<td>10.4% Pacific</td>
</tr>
<tr>
<td>Australia</td>
<td>Japan</td>
<td>119.7</td>
<td>9.2% Pacific</td>
</tr>
<tr>
<td>Indonesia</td>
<td>China</td>
<td>98.0</td>
<td>7.6% Atlantic</td>
</tr>
<tr>
<td>Australia</td>
<td>China</td>
<td>93.4</td>
<td>7.2% Pacific</td>
</tr>
<tr>
<td>Australia</td>
<td>Rep. of Korea</td>
<td>54.9</td>
<td>4.2% Pacific</td>
</tr>
<tr>
<td>Indonesia</td>
<td>China</td>
<td>49.8</td>
<td>3.8% Pacific</td>
</tr>
<tr>
<td>Australia</td>
<td>India</td>
<td>46.9</td>
<td>3.6% Pacific</td>
</tr>
<tr>
<td>Indonesia</td>
<td>Japan</td>
<td>35.6</td>
<td>2.7% Pacific</td>
</tr>
<tr>
<td>Kazakhstan</td>
<td>Russian Fed.</td>
<td>35.6</td>
<td>2.7% Pacific</td>
</tr>
<tr>
<td>Russian Fed.</td>
<td>Cyprus</td>
<td>34.1</td>
<td>2.6% Atlantic</td>
</tr>
<tr>
<td>South Africa</td>
<td>India</td>
<td>31.9</td>
<td>2.5% Cross</td>
</tr>
<tr>
<td>Mongolia</td>
<td>China</td>
<td>19.5</td>
<td>1.5% Non</td>
</tr>
<tr>
<td>Indonesia</td>
<td>Malaysia</td>
<td>17.4</td>
<td>1.3% Atlantic</td>
</tr>
<tr>
<td>Colombia</td>
<td>Thailand</td>
<td>16.2</td>
<td>1.2% Pacific</td>
</tr>
<tr>
<td>Indonesia</td>
<td>Philippines</td>
<td>15.0</td>
<td>1.2% Pacific</td>
</tr>
<tr>
<td>Indonesia</td>
<td>Malaysia</td>
<td>14.5</td>
<td>1.1% Pacific</td>
</tr>
<tr>
<td>Indonesia</td>
<td>China, Hong Kong</td>
<td>12.5</td>
<td>1.0% Pacific</td>
</tr>
<tr>
<td>USA</td>
<td>Netherlands</td>
<td>12.0</td>
<td>0.9% Atlantic</td>
</tr>
<tr>
<td>USA</td>
<td>Brazil</td>
<td>9.1</td>
<td>0.7% Atlantic</td>
</tr>
<tr>
<td>Canada</td>
<td>Japan</td>
<td>8.9</td>
<td>0.7% Cross</td>
</tr>
<tr>
<td>Russia Fed.</td>
<td>UK</td>
<td>8.3</td>
<td>0.6% Atlantic</td>
</tr>
<tr>
<td>Canada</td>
<td>USA</td>
<td>8.3</td>
<td>0.6% Atlantic</td>
</tr>
<tr>
<td>Colombia</td>
<td>Netherlands</td>
<td>7.8</td>
<td>0.6% Cross</td>
</tr>
<tr>
<td>Russia Fed.</td>
<td>Turkey</td>
<td>7.7</td>
<td>0.6% Cross</td>
</tr>
<tr>
<td>Netherlands</td>
<td>Germany</td>
<td>7.5</td>
<td>0.6% Cross</td>
</tr>
<tr>
<td>Colombia</td>
<td>Turkey</td>
<td>7.2</td>
<td>0.6% Atlantic</td>
</tr>
<tr>
<td>USA</td>
<td>Brazil</td>
<td>7.2</td>
<td>0.6% Pacific</td>
</tr>
<tr>
<td>USA</td>
<td>Rep. of Korea</td>
<td>6.7</td>
<td>0.5% Cross</td>
</tr>
<tr>
<td>Colombia</td>
<td>Chile</td>
<td>6.6</td>
<td>0.5% Non</td>
</tr>
<tr>
<td>Canada</td>
<td>Rep. of Korea</td>
<td>6.7</td>
<td>0.5% Atlantic</td>
</tr>
<tr>
<td>USA</td>
<td>China</td>
<td>7.5</td>
<td>0.6% Cross</td>
</tr>
<tr>
<td>South Africa</td>
<td>Netherlands</td>
<td>7.0</td>
<td>0.5% Atlantic</td>
</tr>
<tr>
<td>Australia</td>
<td>Netherlands</td>
<td>6.8</td>
<td>0.5% Cross</td>
</tr>
<tr>
<td>Colombia</td>
<td>USA</td>
<td>6.7</td>
<td>0.5% Atlantic</td>
</tr>
</tbody>
</table>

Note(*): Cross = Cross-regional, Non = Non-seaborne.
Note(**): Part of Russia’s export to China is non-seaborne. (See http://www.suekag.com/deliverydirections/)

Source: UN Comtrade database (http://comtrade.un.org/data)
Figure 3.1: The three major coal price indexes, January 2007 to September 2015

3.3 Coal prices and mining companies’ share prices

In this section, we examine the relationship between coal prices and share prices of coal mining companies. In the past few years, world coal prices have been steeply declining, reaching an eight-year low in September 2015. NYMEX first nearby futures prices and the two API indexes are shown in figure 3.2.

From 2010 to 2012, around the time of coal’s previous peak in 2011, mining companies such as Arch Coal, Peabody Energy, Alpha Natural Resources and Walter Energy invested billions for mines, trying to cash in the high prices amid strong demand from China. The ensuing global supply glut was met by cooling demand,
resulting in the current downturn, which caused Walter and Alpha to file for bankruptcy protection in July and August 2015 respectively. Meanwhile, the share prices of two of the largest US coal miners Peabody and Arch Coal have plummeted, as displayed in figure 3.3.

In figure 3.4, we plotted the Nymex coal price against the two share prices for the period January 2014 to August 2015, showing a clear relationship. The correlation between Nymex coal index and Peabody and Arch reached 94.7% and 91.3% respectively. The sensitivities of commodity mining company’s shares to the commodity price were first analyzed by Tufano [48] in the case of gold, and further investigated by Geman and Vergel [32] for fertilizers.

In terms of percentage changes, the coal index lost about 30% since the start of
CHAPTER 3. COAL

Figure 3.3: Peabody Energy and Arch Coal share prices, January 2014 to August 2015

Figure 3.4: NYMEX coal index vs Peabody and Arch share prices, January 2014 to August 2015
3.4 Data

3.4.1 Description of the data set

A summary of the data we use for this study is shown in table 3.2. The data include prices from 11 markets around the globe: six in the Pacific region (Newcastle
CHAPTER 3. COAL

Table 3.2: A summary of all data series for this study

<table>
<thead>
<tr>
<th>Data source/type</th>
<th>Data series</th>
<th>Frequency</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>McCloskey coal marker</td>
<td>Japan CIF</td>
<td>monthly</td>
<td>1/31/2001</td>
<td>12/31/2014</td>
</tr>
<tr>
<td></td>
<td>Asia CIF</td>
<td>monthly</td>
<td>1/31/2001</td>
<td>12/31/2014</td>
</tr>
<tr>
<td></td>
<td>China Qinhuangdao FOB</td>
<td>bi-weekly</td>
<td>10/15/2004</td>
<td>1/9/2015</td>
</tr>
<tr>
<td></td>
<td>Colombia Puerto Bolivar FOB</td>
<td>bi-weekly</td>
<td>10/15/2004</td>
<td>1/9/2015</td>
</tr>
<tr>
<td></td>
<td>Russia West (Baltic) FOB</td>
<td>bi-weekly</td>
<td>10/15/2004</td>
<td>1/9/2015</td>
</tr>
<tr>
<td></td>
<td>Russia East (Vostochny) FOB</td>
<td>bi-weekly</td>
<td>10/15/2004</td>
<td>1/9/2015</td>
</tr>
<tr>
<td></td>
<td>NW Europe (ARA, or API2) CIF</td>
<td>weekly</td>
<td>12/16/2005</td>
<td>1/9/2015</td>
</tr>
<tr>
<td></td>
<td>Richards Bay (RB, or API4) FOB</td>
<td>weekly</td>
<td>10/15/2004</td>
<td>1/9/2015</td>
</tr>
<tr>
<td></td>
<td>Newcastle FOB</td>
<td>weekly</td>
<td>10/15/2004</td>
<td>1/9/2015</td>
</tr>
<tr>
<td></td>
<td>NYMEX first nearby</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NW Europe (ARA, or API2)</td>
<td>daily</td>
<td>7/17/2006</td>
<td>9/18/2015</td>
</tr>
<tr>
<td></td>
<td>ICE first nearby</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Richards Bay (RB, or API4)</td>
<td>daily</td>
<td>7/17/2006</td>
<td>9/18/2015</td>
</tr>
<tr>
<td></td>
<td>Newcastle</td>
<td>daily</td>
<td>1/2/2009</td>
<td>9/18/2015</td>
</tr>
<tr>
<td></td>
<td>McCloskey coal marker</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Indonesian FOB</td>
<td>daily</td>
<td>8/18/2011</td>
<td>9/18/2015</td>
</tr>
</tbody>
</table>

of Australia, Japan, Asia\(^3\), China, Russia East and Indonesia), five in the Atlantic region (Northwest Europe, Richards Bay of South Africa, Colombia, US Central Appalachian and Russia Baltic).

Different data come at different frequencies, from daily to monthly, and for different time ranges. ARA, Richards Bay and Newcastle have both weekly spot data for a longer time period and daily first nearby future data for a shorter period. The former are used for the PCA analysis and the latter for cointegration analysis. Both time series are present in table 3.2. All coal prices in this study are in, or are converted to, US dollars per tonne.

In the PCA analysis, we use all data except Indonesia which only has four years of data. Weekly and bi-weekly data series are transformed into monthly data as PCA

\(^3\)McCloskey Asian Steam Coal marker represents the CIF Asian price by equally collating prices in Korea, Taiwan and Japan.
requires all the variables to have the same length with no missing data. The period of analysis is the intersection of all time ranges, January 2006 to December 2014. There are 10 variables with 108 months, plotted in figure 3.6. From the figure, we see similar features: a huge spike in mid-2008, another peak throughout 2011 and gradual drops afterwards.

In the cointegration analysis, we do pairwise cointegration tests for all pairs of prices for two main reasons: to understand in depth the intricate dynamics between coal markets, especially by comparing intra- and inter-regional integration; to take advantage of the uneven nature of the data set by using the maximum amount of data available for each pair, which can be weekly or daily, and can be longer than the period used in the PCA. Additionally, the Indonesian price can be used for pairwise tests versus price series with daily data.

3.4.2 Unit root tests

As a first step of any integration study, stationarity needs to be tested. A series is called integrated with degree $d$, i.e., an $I(d)$ process, if it is stationary after being differenced exactly $d$ times. $I(1)$ processes, also called unit root processes, are the most common financial time series.

We apply the augmented Dickey–Fuller test [18] to all 11 of our variables, on both the level and the first difference. As shown in table 3.3, the tests fail to reject the hypothesis that the data have unit roots at 5% level for all the variables, but reject
CHAPTER 3. COAL

Figure 3.6: Price trajectories for all coal data series, January 2006 to December 2014
CHAPTER 3. COAL

Table 3.3: Augmented Dickey–Fuller test statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Level</th>
<th>First difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asia</td>
<td>-2.4112</td>
<td>-7.1171 **</td>
</tr>
<tr>
<td>Japan</td>
<td>-2.6569 +</td>
<td>-5.5062 **</td>
</tr>
<tr>
<td>China</td>
<td>-2.3100</td>
<td>-6.6623 **</td>
</tr>
<tr>
<td>Colombia</td>
<td>-2.7715 +</td>
<td>-5.7417 **</td>
</tr>
<tr>
<td>Russia Baltic</td>
<td>-2.8570 +</td>
<td>-5.9577 **</td>
</tr>
<tr>
<td>Russia East</td>
<td>-2.3285</td>
<td>-7.6185 **</td>
</tr>
<tr>
<td>ARA</td>
<td>-1.7388</td>
<td>-12.5377 **</td>
</tr>
<tr>
<td>Richards Bay</td>
<td>-1.6591</td>
<td>-23.5162 **</td>
</tr>
<tr>
<td>Newcastle</td>
<td>-0.6877</td>
<td>-37.7006 **</td>
</tr>
<tr>
<td>US CAPP</td>
<td>-2.2368</td>
<td>-44.9284 **</td>
</tr>
<tr>
<td>Indonesia</td>
<td>-1.4184</td>
<td>-31.5732 **</td>
</tr>
</tbody>
</table>

+, * and ** denote significance at better than 10%, 5% and 1% respectively

the hypothesis that the first differences have unit roots. Therefore, all series are unit root processes.

3.5 Principal component analysis

In order to see if the world coal market is integrated or separated into the Atlantic and Pacific regions, we first perform the more visually explanatory principal component analysis (PCA) of the ten price series.

PCA is a statistical transformation method to convert a number of variables into linearly independent ones, or principal components. The first component (PC1) has the largest variance possible, and each remaining PC’s has the largest variance
possible subject to the linearly independence (orthogonal) requirement.

Therefore, one way to analyze PCA results is to see how many principal components are needed to explain a certain large amount, say 95%, of total variance in the data. If most of the variance can be explained by only one variable, then the original variables are mostly integrated. Secondly, we can plot the first two principal component loadings on the factorial plane to see if there is grouping due to regional divide. PCA has been used in energy market integration analysis [46, 53].

Our results in table 3.4 with all 10 price series show that the first component itself explains 91.7% of the total variance; while 96.7% are explained by the first two components. This suggests the world coal markets are not fully integrated as two major factors are needed to represent the price series. In the factorial plane (figure 3.7 left panel), all the variables except US have similar PC1 loadings while their PC2 loadings spread out much more, with China being distant to others. A grouping of the eight variables is evident from the plot. However, there is no divide between the Atlantic region ones (denoted by +) and Pacific region ones (denoted by \times).

A likely reason for the separation of US prices from most others is the nature of the price series. All prices used for the PCA are McCloskey coal marker for FOB (export)
or CIF (import) prices, whereas the US CAPP (Central Appalachian) is the first nearby price for an exchange-traded contract. As for China, its particular situation is due to the huge domestic demand/supply and the relatively less importance of its international trades. Additionally, the government’s measures to protect domestic coal industry such as subsidies and tariffs also play a part in the seeming deviation of China’s price from other markets.

Now we remove the US and China series, and run the PCA again using the other eight variables. Table 3.5 shows that the first component alone explains over 95% of the total variance among the eight variables. This can be again confirmed in the right panel of figure 3.7. All of them have similar PC1 loadings but diverse PC2 loadings.
### Chapter 3. Coal

**Table 3.5:** PCA results: No US CAPP and China

<table>
<thead>
<tr>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
<th>PC6</th>
<th>PC7</th>
<th>PC8</th>
</tr>
</thead>
<tbody>
<tr>
<td>76.53</td>
<td>11.82</td>
<td>9.27</td>
<td>5.40</td>
<td>3.99</td>
<td>3.61</td>
<td>2.80</td>
<td>2.32</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Standard deviation</th>
<th>76.53</th>
<th>11.82</th>
<th>9.27</th>
<th>5.40</th>
<th>3.99</th>
<th>3.61</th>
<th>2.80</th>
<th>2.32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion of Variance</td>
<td>95.2%</td>
<td>2.3%</td>
<td>1.4%</td>
<td>0.5%</td>
<td>0.3%</td>
<td>0.2%</td>
<td>0.1%</td>
<td>0.1%</td>
</tr>
<tr>
<td>Cumulative Proportion</td>
<td>95.2%</td>
<td>97.4%</td>
<td>98.8%</td>
<td>99.3%</td>
<td>99.6%</td>
<td>99.8%</td>
<td>99.9%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

There is still no regional divide shown—the Atlantic and Pacific prices are mixed up.

### 3.6 Cointegration analysis

We now perform the more formal and quantitative Johansen cointegration tests \[37\]. Specifically, we test the cointegration relationships between each pair of data series using richer data than in the PCA, to see if the results are consistent with the PCA. The Johansen procedure has been used in the literature for integration analysis not only in the coal market \[38, 53, 44\] but also in other commodities such as natural gas \[46, 30\].

#### 3.6.1 Model specification

For a \(K\)-dimensional VAR(\(p\)) process

\[
y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + \epsilon_t
\]

(3.1)
It can be transformed into a vector error correction model (VECM) representation

\[
\Delta y_t = \Pi y_{t-1} + \Gamma_1 \Delta y_{t-1} + \cdots + \Gamma_{p-1} \Delta y_{t-p+1} + \epsilon_t \tag{3.2}
\]

where \( \Pi := -(I_K - A_1 - \cdots - A_p) \) denotes the long-run impact, \( \Gamma_i := -(A_{i+1} + \cdots + A_p) \) is the short-run impact matrix for lag \( i = 1, 2, \ldots, p - 1 \).

We are mainly interested in the long-run impact matrix \( \Pi \). Assuming all components in equation (3.1) are either \( I(0) \) or \( I(1) \) series, then all terms in equation (3.2) except \( \Pi y_{t-1} \) are stationary. So \( \Pi y_{t-1} \) is stationary as well.

(i) If \( \Pi \) has full rank \( K \), we can left multiply \( \Pi^{-1} \) to \( \Pi y_t \) and still have the stationarity. Hence \( y_t \) is stationary.

(ii) If \( \Pi = 0 \), rank is 0. There is no cointegration.

(iii) If the rank of \( \Pi \) is \( r \) such that \( 0 < r < K \), it is the case of cointegration. \( \Pi \) can be decomposed as \( \Pi = \alpha \beta' \) where \( \alpha \) and \( \beta \) are both \( K \times r \) matrices. Left multiplying \( \Pi y_t \) by \( (\alpha' \alpha)^{-1} \alpha' \), the resulting \( \beta' y_t \) is still stationary.

The cointegration matrix \( \beta \) describes the long-run equilibrium of the system. Each column of \( \beta \) is a cointegration vector. The loading matrix \( \alpha \) describes the short-run response to the long-run equilibrium deviations. The null hypothesis of Johansen's full-information maximum likelihood test is that the cointegrating rank is \( \leq r \) for the \( K \)-dimensional processes \( y_t \). There are two variations of the test. The trace test's
alternative hypothesis is that cointegrating rank $\leq K$; the *maximum eigenvalue test’s* alternative is that the cointegrating rank be $\leq r + 1$.

The test statistics is

$$-T \sum_{i=r+1}^{K} \log(1 - \lambda_i)$$

for the trace test and

$$-T \log(1 - \lambda_{r+1})$$

for the max eigenvalue test, where $\lambda_i$ in both versions are the estimated eigenvalues of $\Pi$ and $T$ is the sample size.

The results reported below relate to the trace tests.

### 3.6.2 Cointegration tests results

The results from multivariate Johansen cointegration tests for different $r$ indicate that, the hypothesis ‘cointegration rank $\leq 2$’ is rejected at 5% level but the hypothesis ‘rank $\leq 3$’ is not rejected, which together mean the ten series are cointegrated with rank 3. However, this information alone does not provide us with too much insight into how the global markets are integrated. Consequently, we proceed to perform pairwise cointegration tests, in order to shed some light on the complex relationships among different markets, and to utilize maximal data for each pair. Recall that we converted daily and weekly data into monthly—losing some information during the
CHAPTER 3. COAL

process—for the PCA and the multivariate Johansen test which require equal lengths of all data series. As a result, the pairwise tests are also more reliable.

The results of the pairwise Johansen cointegration tests are shown in table 3.6. In panel A, we report the pairwise p-values. For better demonstration, significance levels are presented in panel B using + and * notations. Among the 11 price series, there are 49 pairs of prices tested, out of which 29 pairs are cointegrated at a significance level of better than 10%. We categorize the pairs into three types: within the Atlantic region, within the Pacific region, and cross-regional.

There are seven cointegrated pairs in the Atlantic region (upper left quadrant of the table) and four pairs in the Pacific region (lower right corner). That is 11 cointegrated pairs out of the 21 intra-regional price pairs. In contrast, there are 18 cointegrated pairs out of the 28 cross-regional pairs we tested (lower left corner). There is no evidence of a stronger integration within either region compared to the whole world market.

This confirms our result from the PCA that the two regions are integrated, with no clear regional separation.

---

4Indonesian prices are tested with price series of daily data available, due to its short history.


Table 3.6: Pairwise Johansen cointegration tests results

A: p-values

<table>
<thead>
<tr>
<th></th>
<th>The Atlantic Region</th>
<th></th>
<th>The Pacific Region</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Colombia</td>
<td>Russia Baltic</td>
<td>ARA</td>
<td>Richards Bay</td>
</tr>
<tr>
<td></td>
<td>0.0138</td>
<td>0.1790</td>
<td>0.3402</td>
<td>0.0947</td>
</tr>
<tr>
<td>The Atlantic</td>
<td>Russia Baltic</td>
<td>0.0330</td>
<td>0.2954</td>
<td>0.0105</td>
</tr>
<tr>
<td>Region</td>
<td>ARA</td>
<td>0.0442</td>
<td>0.0010</td>
<td>0.0014</td>
</tr>
<tr>
<td></td>
<td>Richards Bay</td>
<td>0.1597</td>
<td>0.4066</td>
<td>0.0334</td>
</tr>
<tr>
<td></td>
<td>US CAPP</td>
<td>0.8532</td>
<td>0.0092</td>
<td>0.6940</td>
</tr>
<tr>
<td></td>
<td>Asia</td>
<td>0.3402</td>
<td>0.0105</td>
<td>0.0014</td>
</tr>
<tr>
<td></td>
<td>Japan</td>
<td>0.4780</td>
<td>0.1135</td>
<td>0.1597</td>
</tr>
<tr>
<td></td>
<td>China</td>
<td>0.0329</td>
<td>0.0105</td>
<td>0.0014</td>
</tr>
<tr>
<td></td>
<td>Russia East</td>
<td>0.0322</td>
<td>0.0010</td>
<td>0.0014</td>
</tr>
<tr>
<td></td>
<td>Newcastle</td>
<td>0.0273</td>
<td>0.1597</td>
<td>0.0334</td>
</tr>
<tr>
<td></td>
<td>Indonesia</td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0014</td>
</tr>
</tbody>
</table>

B: significance levels

+ , *, ** denote significance at better than 10%, 5% and 1% respectively; ‘o’ denotes that there’s no cointegration between the pair
3.7 Cointegration between oil, coal and natural gas spot prices

We now turn our attention to the relationship between the three main fossil fuel markets, namely oil, natural gas and coal, in both the US and European markets. The case of the US was examined in detail in Bachmeier and Griffin [6]. They tested pairwise cointegration between the three energy prices using data from 1990 to 2004, and found that the extent of market integration was the strongest between oil and natural gas among the three pairs. They concluded that there was not a primary energy market since the three markets were very weakly integrated.

Daily front month futures prices for crude oil and natural gas are used in this section. WTI (West Texas Intermediate) oil, Brent oil and HH (Henry Hub) natural gas contracts are traded on the NYMEX; NBP (National Balancing Point) is traded on the ICE. WTI and Brent are quoted in USD/Barrel. HH is quoted in USD/MMBtu. NBP is quoted in GBP/therm, and converted to USD/MMBtu for comparison. All the data are downloaded from Bloomberg terminal.

We first look at the US market. The spot prices for WTI crude oil, HH (Henry Hub) natural gas and CAPP coal are plotted pair by pair in figure 3.8. HH prices are multiplied by 10 to be in the same scale as the other two. From the figure, we can suspect that prices behave differently in different periods: they moved together before 2009, then seemed decoupled in the following few years, finally all crashed since the
We divide the whole period into three parts: 2006–2009, Jan 2010–Jun 2014 and Jul 2014–Sep 2015, marked by the vertical dashed lines in figure 3.8. We calculate the correlations in the sub-periods as well as the full period, and display them in table 3.7.

The first period exhibits much bigger correlations than the second period. It was the time when the “financialization” of commodity markets was much greater, with an enormous amount of money invested in commodity indexes such as the GSCI or
Figure 3.9: A comparison between the spot prices of European crude oil ($/barrel), natural gas ($/MMBtu) and coal ($/tonne)

DJ-UBS Commodity Indexes, as discussed by Tang and Xiong [47]. In the second period, the gigantic production of shale gas in the US triggered a large reduction of coal use, hence a joint rapid decline of both commodities. The third period witnessed commodity bear markets caused by the oversupply in US shale industry, a strong US dollar and weak demand in emerging markets such as China. The correlations are the largest in the third period as all three prices free-fell.

The situation of Brent oil, ARA coal and NBP natural gas in Europe is similar albeit to a lesser degree, as displayed in figure 3.9 and table 3.8. The lower correlations between NBP natural gas and other two fuels reflect the absence of shale gas linking the three markets in Europe—together with the fact that NBP is a gas index which only covers the United Kingdom.

We then perform Johansen cointegration tests, both the pairwise version and the multivariate version with all three energy prices, in the US and European markets.
CHAPTER 3. COAL

Table 3.8: Pairwise correlations between the three main fossil fuel prices in Europe

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Brent</td>
<td>ARA</td>
<td>NBP</td>
<td>Brent</td>
</tr>
<tr>
<td>Brent</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>ARA</td>
<td>0.594</td>
<td>1</td>
<td></td>
<td>0.882</td>
</tr>
<tr>
<td>NBP</td>
<td>0.660</td>
<td>0.631</td>
<td>1</td>
<td>0.657</td>
</tr>
</tbody>
</table>

Table 3.9: Johansen cointegration tests p-values

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>WTI, CAPP</td>
<td>0.0096**</td>
<td>0.0460**</td>
</tr>
<tr>
<td></td>
<td>CAPP, HH</td>
<td>0.0528</td>
<td>0.0505</td>
</tr>
<tr>
<td></td>
<td>WTI, HH</td>
<td>0.1193</td>
<td>0.0451*</td>
</tr>
<tr>
<td></td>
<td>WTI, CAPP, HH (with r = 1)</td>
<td>0.0047**</td>
<td>0.0451*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Europe</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Brent, ARA</td>
<td>0.6302</td>
<td>0.1392</td>
</tr>
<tr>
<td></td>
<td>ARA, NBP</td>
<td>0.1195</td>
<td>0.0227*</td>
</tr>
<tr>
<td></td>
<td>Brent, NBP</td>
<td>0.0028**</td>
<td>0.0111*</td>
</tr>
<tr>
<td></td>
<td>Brent, ARA, NBP (with r = 1)</td>
<td>0.0887</td>
<td></td>
</tr>
</tbody>
</table>

* and ** denote significance at better than 5% and 1% respectively.

Since the third period (Jul 2014-Sep 2015) is too short for cointegration analysis, we merge it into the second period. Test results are shown in Table 3.9.

Overall, the cointegration in the US market is slightly stronger than the European market, evidenced by the significant p-value of the full-period multivariate test of null hypothesis that the cointegration rank \( r = 0 \). By rejecting the null hypothesis at 1% level, we conclude that \( r = 1 \) for WTI, CAPP and HH in the whole period. Another observation is that the cointegration in the first period is much more significant than in the second period, which confirms the observation from the price trajectories.

There is only one significant cointegration relationship in the second half, between
the US CAPP coal and US HH natural gas. Interestingly, they aren’t cointegrated for the whole period at 5% significance level, but are cointegrated in both sub-periods. This can happen because their cointegration relations differ considerably in the two periods, which are (1, -9.0887) and (1, -16.6442) respectively, meaning on average, a one dollar increase (decrease) in natural gas prices corresponds to a $9.09 increase (decrease) in coal prices in 2006–2009, but to a $16.64 increase (decrease) in coal prices in 2010–2015. Or formally,

\[ 1 \times CAPP - 9.0887 \times HH \] is stationary in the 1st period and

\[ 1 \times CAPP - 16.6442 \times HH \] is stationary in the 2nd period.

The smaller p-value (0.0149) in the second period indicates a more significant cointegration relationship than in the first period (0.0460). This stronger integration between coal and natural gas in the US market after 2010 is due to the fact that the sharp decrease of prices of natural gas in recent years have strongly affected coal prices, as the two fuels compete in electricity generation. The property that US coal and natural gas pair is the only cointegrated relationship in the 2010–Sep 2015 period shows that the three energy markets are not integrated on either side of the Atlantic Ocean.
3.8 Conclusion

In this chapter, we investigate the cointegration in world coal market. Using a comprehensive and rich price data set of ten important countries worldwide for the past decade, we aim to find out if and how the markets are integrated, especially between the two major regions, Atlantic and Pacific basin. By applying the intuitive PCA and the pairwise Johansen cointegration tests, we conclude in favor of world market cointegration in general, consistent with the findings of [52, 38, 53, 44]. Moreover, our unique method enables us to maximally exploit the rich data set, and to compare the degrees of integration of intra- and inter-regional pairs. The results show no evidence of regional separation, thanks largely to the many trade routes connecting the two basins.

We also study the market integration between the three main fossil fuel prices, crude oil, coal and natural gas, in the US and Europe. Results of cointegration analysis agree with the findings of [6], that there is not a primary energy market. However, we find evidence of cointegration between US natural gas and US coal, especially after 2010. We believe this is the consequence of the recent US shale gas boom and the inherent substitute relationship of the two commodities in power generation industry.

For future research, shipping costs such as the Baltic Dry Index (BDI) can be used to adjust for exporting prices and importing prices. With shipping price indexes, we can also test whether the degrees of cointegration in different periods depend on
CHAPTER 3. COAL

the prevailing transporting costs. Finally, the PCA and cointegration tests can be extended to coking coal—for both the integration between geographical regions and among different fuel types—pending available data.
Chapter 4

Intraday Pairs Trading Strategies on High Frequency Data: The Case of Oil Companies

This chapter introduces novel ‘doubly mean-reverting’ processes based on conditional modeling to model spreads between pairs of stocks. Intraday trading strategies using high frequency data are proposed based on the model. This model framework and the strategies are designed to capture ‘local’ market inefficiencies that are elusive for traditional pairs trading strategies with daily data. Results from real data back-testing for two periods show remarkable returns, even accounting for transaction costs, with annualized Sharpe ratios of 3.9 and 7.2 over the periods June 2013–April 2015 and 2008 respectively. By choosing the particular sector of oil companies, we also confirm
the observation that the commodity price is the main driver of the share prices of commodity-producing companies at times of spikes in the related commodity market.

4.1 Introduction

The idea of pairs trading is quite popular across various asset classes and based on the property that, since companies within a sector are highly correlated, some pairs of price returns exhibit strong similarity. We can model the return differences of these pairs as mean-reverting processes. If they deviate too far from the mean, we short/long the pair by simultaneously buying one and short selling another. We keep the position until it reverts back to the mean level.

Let A and B be a pair of closely related stocks, $S_A(t)$ and $S_B(t)$ their prices at time t. The cumulative log return difference—or spread—for this pair is

$$Y(t) = \log \left( \frac{S_A(t)}{S_A(0)} \right) - \log \left( \frac{S_B(t)}{S_B(0)} \right)$$  (4.1)

In pairs trading literature, the spread $Y(t)$ or its variation has been modeled as a mean-reverting process, oscillating either around zero or around a linear function of time. For example, Avellaneda and Lee [5] define the spread as the difference between $\log \left( \frac{S_A(t)}{S_A(0)} \right)$ and $\beta \cdot \log \left( \frac{S_B(t)}{S_B(0)} \right)$, where $\beta$ is calculated by regressing the cumulative log return of one stock on another in a certain period. The best pairs are commonly identified as either the ones with smallest distance measures defined as the sum of
CHAPTER 4. PAIRS TRADING

squared deviations [27, 9] or using cointegration relationships [49, 39].

Both the distance method and the cointegration method have their limitations. Consider two stocks A and B with cumulative log returns both being 0 at the beginning. Suppose \( \log \left( \frac{S_A(t)}{S_A(0)} \right) \) goes to a large positive value \( \alpha \) in a short amount of time then stays around that level while \( \log \left( \frac{S_B(t)}{S_B(0)} \right) \) remains around 0. Then no matter how synchronized they move afterwards, this pair would not likely be identified by the simple distance measure as they have a large average distance of \( \alpha \). Now, further assume A and B co-move only in a subperiod during the whole analysis period, then the pair is not likely to be selected according to cointegration method. Yet there are clearly profits to be made in this scenario. The problem is that both those methods imply a static relationship between two stocks during the training period whereas the relationship may very well be changing from one day to the next.

The strategy we propose in this chapter is designed exactly to capture this kind of ‘local’ statistical arbitrage opportunities, by searching for temporary market mispricing inefficiencies. The idea is to seek the pairs of which \( Y(t) \) can be characterized by the following modeling procedure: model the long term trend of \( Y(t) \), denoted as \( L(t) \), as a stochastic process, and then model \( Y(t) \) via a mean-reverting process around this long term stochastic trend \( L(t) \) using the conditional modeling technique. If the mean reversion speed of \( Y(t) \) is fast enough, we can make profit by making intraday pairs trades.

The utilization of two mean-reverting stochastic processes on the same series is
CHAPTER 4. PAIRS TRADING

partly inspired by Fourier series expansion. $L(t)$ can be regarded as the first term of a Fourier series with the largest period and lowest frequency; imagine that $Y(t)$ is approximated by the Fourier series, with higher-frequency local oscillation being added on top of the lower-frequency waves. To visualize this, imagine a long rope lying on the ground straightened out. If we hold onto one end of the rope and shake it horizontally, then it will display a wavy pattern. Now, pick two points on the rope that are close to each other, pin their locations, then shake the segment between them. The shaken part will likely become a more pronounced local wave. Repeat this to the whole rope segment-by-segment. The resulting rope would look like the $Y(t)$ process while all the pinned positions make up $L(t)$. The rational for this novel doubly mean-reverting model is that, if we can identify pairs with relative stable $L(t)$ and volatile $Y(t)$, then intraday pairs trading should perform well a priori. $Y(t)$ in these cases would return at the end of a trading day—hopefully after wild swings—to more or less the same level as daily open.

The key to add up local oscillation is through a framework of ‘conditional modeling and conditional inference’ (see the overview in [13]). This technique has been applied to analyze the dynamics of financial time series (e.g. the waiting time invariance of return sequences in [16], aggregation theorem in [15]) as well as to other research fields (e.g. [2], [14]). Nevertheless, the technique has not been utilized for designing trading strategies in the literature.

In this study, we focus on the oil sector and look at times when the underlying
commodity price is experiencing sharp moves, making it the major factor driving the share prices. Geman and Vergel [32] showed, in the case of the fertilizer commodity, that shares of fertilizer-mining companies are very sensitive to the commodity price at times of high moves of this price. Geman [29] extended this property to other types of agriculture-related companies. The literature on the subject counts as a founding paper Tufano [48] who analyzes the price sensitivities of gold mining companies’ shares. Our results confirmed the observation that crude oil price is the main driver of oil company stock prices during market turbulence.

In the literature, most pairs trading studies use daily data and a daily trading frequency [27, 39, 5, 17, 7, 54]. Bowen et al. [9], which use 60 minute return series, is one of the very few that uses data with frequency higher than daily. With the availability of tick data, we are able to use five-minute series for $Y(t)$ in the case of some highly liquid oil company stocks. To our best knowledge, we are the first academic study on pairs trading to use such high frequency data, and the first one on intraday pairs trading strategies.

The rest of the chapter is organized as follows. Section 4.2 details the model and its calibration. Trading rules are discussed in section 4.3. We perform simulations in section 4.4 in order to validate the model and the strategies. Results from real data back-testing over two periods are presented and analyzed in section 4.5. Section 4.6 concludes the chapter and provides directions for future research.
4.2 The model

4.2.1 Model specification

In order to exploit intraday pairs trading profits, we use high frequency data with interval length of five minutes to model \( Y(t) \) defined in equation (4.1). There are 78 five-minute intervals every day during trading hours from 9:30 AM to 4:00 PM, hence 79 \( Y(t) \)'s. We denote the 79 observed values of \( Y(t) \) in day \( i \) as

\[
Y_{79(i-1)+1}, Y_{79(i-1)+2}, \ldots, Y_{79i} \quad i = 1, 2, \ldots, N
\]

where \( N \) is the number of days. The subscript in this chapter refers to discretized observations of stochastic processes. Moreover, we assume that the long term trend \( L(t) \) is identified by the daily opening and closing values of the process \( Y(t) \), namely for day \( i \),

\[
L_{2i-1} = Y_{79(i-1)+1} \quad \text{and} \quad L_{2i} = Y_{79i}. \quad (4.2)
\]

The stochastic process \( L(t) \), with two observed data points per day, is preferred to have a small variance. In this study, we model \( L(t) \) as an Ornstein–Uhlenbeck (OU) process, with mean 0

\[
dL(t) = -\theta_L L(t)dt + \sigma_L dW^L_t. \quad (4.3)
\]
CHAPTER 4. PAIRS TRADING

Next by the definition of conditional distribution, the joint distribution of $Y_1, Y_2, \ldots, Y_{79N}$ can be written as the product of the distribution of $Y_{79(i-1)+1}$’s and $Y_{79i}$’s and the conditional distribution of $Y_i$’s given $Y_{79(i-1)+1}$’s and $Y_{79i}$’s

$$f(Y_1, Y_2, \ldots, Y_{79N})$$

$$= f(Y_{79(i-1)+1}, Y_{79i}, i = 1, \ldots, N) f(Y_1, Y_2, \ldots, Y_{79N} | Y_{79(i-1)+1}, Y_{79i}, i = 1, \ldots, N)$$

$$= f(L_1, L_2, \ldots, L_{2N}) f(Y_1, Y_2, \ldots, Y_{79N} | L_1, L_2, \ldots, L_{2N})$$

(4.4)

Note that due to equation (4.2), the last equality is valid, and the joint distribution of $L_i$’s can be obtained by discretizing the process (4.3).

Now to model the conditional distribution of $Y_i$’s given $L_i$’s, we use the conditional modeling technique by introducing an auxiliary process $\tilde{Y}(t)$ that follows:

$$d\tilde{Y}(t) = \theta \left( \tilde{L}(t) - \tilde{Y}(t) \right) dt + \sigma dW_t^{\tilde{Y}}$$

(4.5)

where the mean process is $\tilde{L}(t) = \frac{L_{2i-2}+L_{2i-1}}{2}$ and $i = i(t)$ refers to the day of time $t$ ($2i - 2$ refers to the closing of day $i - 1$ and $2i - 1$ refers to the opening of day $i$). We assume the conditional distribution

$$f(Y_1, Y_2, \ldots, Y_{79N} | L_1, L_2, \ldots, L_{2N})$$

is the same as the conditional distribution of $\tilde{Y}_i$’s given $\tilde{Y}_{79(i-1)+1} = L_{2i-1}$ and $\tilde{Y}_{79i} = \ldots$
CHAPTER 4. PAIRS TRADING

$L_{2i}, i = 1, \ldots, N$, where $\tilde{Y}_i$'s are the corresponding discretization of process \([4.5]\).

In other words, conditional on given daily opening and closing values of the process, $Y(t)$ is the same as $\tilde{Y}(t)$ in distribution. Hence, to simplify the notation, we use $Y(t)$ and $Y_i$ in place of $\tilde{Y}(t)$ and $\tilde{Y}_i$ in the rest of the chapter.

By defining the mean process $\bar{L}(t)$ as the average of $L_{2i-2}$ and $L_{2i-1}$, we assume in any trading day, the mean level that the spread process reverts to is the average of the opening value of the current day and the closing value of the previous trading day.

To recap, the distribution of the $Y(t)$ process is defined by first specifying the dynamics of $L(t)$ from equation \([4.3]\), then the distributions of the in-between points are given indirectly by equation \([4.5]\), via the conditional relationship \([4.4]\), hence the name conditional modeling.

For $Y(t)$, there are 79 observations per day thus 79 time intervals, the lengths of which are not equal: 78 short five minute periods and a long overnight period. The similar problem exists for $L(t)$: the time span between a day’s open and close is different from between the day’s close and next day’s open.

The lengths in real time of the trading hours per day are 6.5 hours while the overnight periods are at least 17.5 hours (from 4 pm market close to next day’s market open 9:30 am, or longer in the case of weekends and holidays). However, the amount of information and market movements during daytime trading hours are
CHAPTER 4. PAIRS TRADING

much richer than that during overnight periods. Therefore, we estimate the lengths of both periods in effective time instead of real time in the following way: since the relative (effective) lengths of the trading day and the overnight period are unknown and unequal, we need two time steps. \( \delta_1 \) is the length of trading hours 9:30 AM and 4:00 PM; \( \delta_2 \) is the length between market close and next day’s open. Hence,

\[
\delta_1 + \delta_2 = 1 \text{ day} = \frac{1}{250}
\]  (4.6)

and the algorithm to estimate \( \delta_1 \) and \( \delta_2 \) is based on the ratio of variances of intraday and overnight changes as detailed in section 4.2.2.1.

All the model parameters \( \theta_L, \sigma_L, \theta, \sigma, \delta_1 \) and \( \delta_2 \) from equation (4.3)–(4.6) are calibrated daily using maximum likelihood estimation (MLE).

4.2.2 Model calibration

For better parameter estimation, the calibration is updated every day using a moving window, also called the pairs formation period [27], the length of which is chosen properly. If the duration is too short, the calibration is unreliable due to the lack of training data; if the duration is too long, estimated parameters do not accurately reflect the present situation because of non-stationarity of market dynamics. In our model, the frequencies of \( L(t) \) and \( Y(t) \) are different, thus requiring training periods of different lengths. We use the past 100 days’ daily open and close
prices to calibrate the process $L(t)$ and use the past 30 days’ five-minute prices to calibrate the process $Y(t)$. Both lengths were decided at the beginning of the study and has not been tuned based on data, to avoid data-snooping biases \[40\].

### 4.2.2.1 Calibration for $L(t)$

For a general OU process with constant mean

$$dS(t) = \theta (\mu - S(t)) dt + \sigma dW_t,$$

the discretization $\{S_i\}$ satisfies

$$S_{i+1} = S_i e^{-\theta \delta} + \mu (1 - e^{-\theta \delta}) + \sigma \sqrt{\frac{1 - e^{-2\theta \delta}}{2\theta}} Z_i, \quad (4.7)$$

for all $i$, where $\delta$ is the time step in discretization and $Z_i$'s are i.i.d. $N(0, 1)$.

For our $L(t)$, the discretized series is the combined daily opening and closing cumulative return differences. $\mu$ is assumed to be 0. Equation \[4.7\] leads to two equations.

The intraday changes

$$L_{2i} = L_{2i-1} e^{-\theta_L \delta_1} + \sigma_L \sqrt{\frac{1 - e^{-2\theta_L \delta_1}}{2\theta_L}} Z_{2i} \quad i = 1, \ldots, N; \quad (4.8)$$
and the overnight changes

\[ L_{2i+1} = L_{2i}e^{-\theta_L \delta_2} + \sigma_L \sqrt{\frac{1 - e^{-2\theta_L \delta_2}}{2\theta_L}} Z_{2i+1}, \quad i = 1, \ldots, N - 1 \]  \hspace{1cm} (4.9)

where \( Z_i \)'s are i.i.d. \( N(0,1) \).

To estimate \( \delta_1 \) and \( \delta_2 \), we use the following equation obtained from equations (4.8) and (4.9) with the variances approximated by empirical variances:

\[
\frac{\text{Var}(L_{2i} - L_{2i-1}e^{-\theta_L \delta_1})}{\text{Var}(L_{2i+1} - L_{2i}e^{-\theta_L \delta_2})} \approx \frac{1 - e^{-2\theta_L \delta_1}}{1 - e^{-2\theta_L \delta_2}}
\] \hspace{1cm} (4.10)

Notice that solving this equation for \( \delta_1 \) and \( \delta_2 \) requires \( \theta_L \). Therefore, we develop the following algorithm to iteratively calibrate \( \delta_1, \delta_2, \theta_L \) and \( \sigma_L \) together:

1. Initialize \( \delta_1 \) and \( \delta_2 \).

2. Using \( \delta_1 \) and \( \delta_2 \) values, calibrate \( \theta_L \) and \( \sigma_L \) using MLE.

3. Plug \( \theta_L \) into equation (4.10). Then \( \delta_1 \) and \( \delta_2 \) can be solved together with equation (4.6).

4. Repeat steps 2 and 3 until \( \delta_1, \delta_2, \theta_L \) and \( \sigma_L \) all converge.

**Step 1:**

Since both \( \delta_1 \) and \( \delta_2 \) are small, from equation (4.10),

\[
\frac{\text{Var}(L_{2i} - L_{2i-1})}{\text{Var}(L_{2i+1} - L_{2i})} \approx \frac{1 - e^{-2\theta_L \delta_1}}{1 - e^{-2\theta_L \delta_2}} \approx \frac{\delta_1}{\delta_2}
\] \hspace{1cm} (4.11)
CHAPTER 4. PAIRS TRADING

where the variance of intraday return $\text{Var}(L_{2i} - L_{2i-1})$ and the variance of overnight return $\text{Var}(L_{2i+1} - L_{2i})$ are estimated empirically. Then initial $\delta_1$ and $\delta_2$ can be obtained by solving equation (4.6) and (4.11).

Step 2:

The conditional densities of $L(t)$ are

$$f(L_{2i}|L_{2i-1}; \theta_L, \hat{\sigma}_1) = \frac{1}{\hat{\sigma}_1 \sqrt{2\pi}} \exp\left(-\frac{(L_{2i} - L_{2i-1}e^{-\theta_L \delta_1})^2}{2\hat{\sigma}_1^2}\right)$$

$$f(L_{2i+1}|L_{2i}; \theta_L, \hat{\sigma}_2) = \frac{1}{\hat{\sigma}_2 \sqrt{2\pi}} \exp\left(-\frac{(L_{2i+1} - L_{2i}e^{-\theta_L \delta_2})^2}{2\hat{\sigma}_2^2}\right)$$

where

$$\hat{\sigma}_1 = \sigma_L \sqrt{\frac{1 - e^{-2\theta_L \delta_1}}{2\theta_L}} \quad \text{and} \quad \hat{\sigma}_2 = \sigma_L \sqrt{\frac{1 - e^{-2\theta_L \delta_2}}{2\theta_L}}$$

The log-likelihood function of $(L_1, \ldots, L_{2N})$ is then

$$\mathcal{L}(\theta_L, \sigma_L) = \mathcal{L}(\theta_L, \hat{\sigma}_1, \hat{\sigma}_2) = \sum_{i=1}^{N} \ln f(L_{2i}|L_{2i-1}; \theta_L, \hat{\sigma}_1) + \sum_{i=1}^{N-1} \ln f(L_{2i+1}|L_{2i}; \theta_L, \hat{\sigma}_2)$$

$$= -\frac{N}{2} \ln(2\pi) - N \ln(\hat{\sigma}_1) - \frac{1}{2\hat{\sigma}_1^2} \sum_{i=1}^{N} (L_{2i} - L_{2i-1}e^{-\theta_L \delta_1})^2$$

$$- \frac{N-1}{2} \ln(2\pi) - (N - 1) \ln(\hat{\sigma}_2) - \frac{1}{2\hat{\sigma}_2^2} \sum_{i=1}^{N-1} (L_{2i+1} - L_{2i}e^{-\theta_L \delta_2})^2$$

The MLE for $\theta_L$ and $\sigma_L$ are solved numerically from this equation using a quasi-Newton optimization algorithm called limited memory BFGS [12].

\[1\] This initial approximation turns out to be pretty close. On average, the approximates are only off by 0.3% from the final converged values.
Step 3 and 4 are straightforward, and from our experiments, the algorithm converges fast (generally only 2 to 4 iterations are needed to reach a tolerance of $10^{-6}$).

### 4.2.2.2 Calibration for $Y(t)$

For $Y(t)$, the constant mean $\mu$ in the OU process is replaced by $\tilde{L}_t = \frac{L_{2i-2} + L_{2i-1}}{2}$.

Equation (4.7) becomes

$$Y_{79(i-1)+j} = Y_{79(i-1)+j-1}e^{-\theta \delta} + \frac{L_{2i-2} + L_{2i-1}}{2}(1 - e^{-\theta \delta}) + \sigma \sqrt{\frac{1 - e^{-2\theta \delta}}{2\theta}} Z_{79(i-1)+j}, \quad \forall i, j$$

where $i = 1, 2, \ldots, 30$ denotes days; $j = 2, \ldots, 79$ denotes 5-minute periods; $\delta = \frac{\delta}{78}$ is the effective length of a five-minute interval; $Z_{79(i-1)+j}$’s are i.i.d. $N(0, 1)$. Let

$$a = e^{-\theta \delta}$$

$$b_i = \frac{L_{2i-2} + L_{2i-1}}{2}(1 - e^{-\theta \delta}) = \frac{L_{2i-2} + L_{2i-1}}{2}(1 - a), \quad i = 1, \ldots, 30$$

$$\hat{\sigma} = \sigma \sqrt{\frac{1 - e^{-2\theta \delta}}{2\theta}} = \sigma \sqrt{\frac{1 - a^2}{2\theta}}$$

where $L_0$ is defined to be 0.

Then

$$Y_{79(i-1)+j} - aY_{79(i-1)+j-1} = b_i + \hat{\sigma} Z_{79(i-1)+j}, \quad \forall i = 1, \ldots, 30, \forall j = 2, \ldots, 79$$
CHAPTER 4. PAIRS TRADING

Since \( \{L_i|i=1,\ldots,2N\} \) is a subsequence of \( \{Y_i|i=1,\ldots,79N\} \), we have

\[
f(\vec{Y};\theta,\sigma) = f(\vec{L},\vec{Y}) = f(\vec{L}) f(\vec{Y}|\vec{L})
\]

The log-likelihood for \( Y(t) \) is \( \ln f(\vec{L}) + \ln f(\vec{Y}|\vec{L}) \). The first term \( \ln f(\vec{L}) \) does not depend on \( \theta \) and \( \sigma \). We only focus on the second term

\[
\ln f(\vec{Y}|\vec{L}) = \ln f(\vec{Y}|Y_{79(i-1)+1} = L_{2i-1}, Y_{79i} = L_{2i}, \forall i = 1,\ldots,30) \\
= \ln f(Y_1, Y_2, \ldots, Y_{79}|Y_1 = L_1, Y_{79} = L_2) + \ln f(Y_{80}, Y_{81}, \ldots, Y_{79\times2}|Y_{79} = L_2, Y_{80} = L_3, Y_{79\times2} = L_4) + \cdots + \ln f(Y_{79\times30+1}, \ldots, Y_{79\times30}|Y_{79\times29} = L_{58}, Y_{79\times29+1} = L_{59}, Y_{79\times30} = L_{60})
\]

The last equality is due to the definition of \( Y_i' \)\'s. The remaining derivation of MLE formula is rather cumbersome, thus given in appendix A.

4.3 Trading rules

As mentioned before, training periods of 100 days and 30 days are fixed for the calibration of \( L(t) \) and \( Y(t) \) respectively. For an intraday trading strategy, the trading period is one day. The three periods are illustrated in figure 4.1.
After getting the parameters and consequently the variance estimations of both $L(t)$ and $Y(t)$, we select a set of ‘best’ pairs to be trading candidates for that day. Our ideal trading candidate pair will have a large $Y(t)$ variance and a small $L(t)$ variance. A large $Y(t)$ variance is preferred because more volatile intraday movements lead to more trading opportunities. The preference of small $L(t)$ variances is to ensure that the long-term value of the spread is not volatile. The most desirable situation would be $L(t)$ remaining constant over time while $Y(t)$ fluctuating a lot during the day but always coming back to the constant level.

The procedure to select the ‘best’ pairs is: first remove all the pairs with negative $\theta_L$, then rank all remaining pairs by $L(t)$’s short term variance

$$\frac{\sigma_L}{2\theta_L} \left(1 - e^{-2\theta_L(\delta_1+\delta_2)}\right)$$

in ascending order, record the ranking $r_L$, then rank them again by $Y(t)$’s short term
CHAPTER 4. PAIRS TRADING

variance

\[ \frac{\sigma}{2\theta} \left( 1 - e^{-\frac{2\theta}{78}} \right) \]

in descending order, record the ranking \( r_Y \) and finally select the top 25 or 50 or 100 pairs with smallest \( r_L + r_Y \). In section 4.5 we test different pair selection criteria on real data by varying the number of pairs selected.

During day \( i \), for each candidate pair stock A and stock B, we make a trade immediately when \( Y(t) \), the cumulative return difference between stock A and B, goes out of a ‘confidence band’. More specifically, we

1. short the pair (simultaneously short A and long B) if \( Y(t) \) exceeds \( \frac{L_{2i-2} + L_{2i-1}}{2} + \epsilon \);

2. long the pair (simultaneously long A and short B) if \( Y(t) \) drops below \( \frac{L_{2i-2} + L_{2i-1}}{2} - \epsilon \).

The value \( \epsilon \) is the 98% percentile of the absolute daily change in \( L(t) \) values in the past 100 days. If \( \epsilon \) is too large, we miss out trading opportunities by executing only few trades; if \( \epsilon \) is too small, excessive trading leads to the profits of many trades being dwarfed by transaction costs. In simulations, we also used two other \( \epsilon \) levels (95% and 90% percentiles) for comparison. In real data back-testing, we stick with the 98% percentile for better performances².

For each pair-trade, we buy $1 worth of one stock and short $1 worth of the

²It is possible that the optimal threshold is not 98%. We did not experiment on real data to find out the exact optimal value. Zeng and Lee [54] derived optimal thresholds for maximum profitability per unit of time in a single OU process model framework.
other stock. For example, when we long the pair, we buy \( \frac{s_1}{P_A} \) shares of stock A and simultaneously short \( \frac{s_1}{P_B} \) shares of stock B, where \( P_A \) and \( P_B \) are their respective prices. Our net position at the outset of each pair-trade is zero.

The open position is closed by making the opposite trades (selling the stock bought, buying back and returning the stock shorted) when either (a) \( Y(t) \) reverts back to \( \frac{L_{2i-2} + L_{2i-1}}{2} \), or (b) the market closes for the day at 4pm, whichever happens first.

### 4.4 Simulation

In this section, we demonstrate the validity of our doubly mean-reverting model and the proposed trading strategy using simulation. The goal is to simulate \( L_i \)'s and \( Y_i \)'s for a whole year using a set of parameters \( \theta_L, \sigma_L, \delta_1, \theta, \sigma \), then apply the strategy on the simulated data. For simplicity, we assume the parameters remain constant in the simulation, although we update them daily for trading on real data. If the model and strategy are well designed, the profit should be robust for a ‘good’ set of parameters but not for a ‘bad’ one.

#### 4.4.1 Simulating \( L_i \)'s and \( Y_i \)'s

First we simulate \( L_k \), \( \forall k = 1, \ldots, 2N \) by equation (4.8) and (4.9) in section 4.2.2.1.
CHAPTER 4. PAIRS TRADING

For each day $i$, define $X_{78(i-1)+j} = Y_{79(i-1)+j+1} - a Y_{79(i-1)+j}$ for all $j = 1, \ldots, 78$, where $a = e^{-\theta \delta}$. Now given $L_k$’s, we first generate $X_{78(i-1)+j}$, $j = 1, \ldots, 78$ using the conditional distribution [derived in appendix equation (A.1)-(A.3)]:

$$f(X_{78(i-1)+1}, \ldots, X_{78i}|Y_{79(i-1)} = L_{2(i-1)}, Y_{79(i-1)+1} = L_{2i-1}, Y_{79i} = L_{2i}) = \left(\frac{1}{\sigma \sqrt{2\pi}}\right)^{77} \exp \left[ -\frac{1}{2\sigma^2} \left( \sum_{j=1}^{78} (x_{79(i-1)+j} - b_i)^2 + (L_{2i} - a^{78} L_{2i-1} - \sum_{j=1}^{78} a^{78-j} x_{78(i-1)+j} - b_i)^2 \right) \right] \sqrt{\frac{1-a^2}{1-a^{156}}} \exp \left[ -\frac{1}{2\sigma^2} \frac{1-a^2}{1-a^{156}} \left( L_{2i} - 1 + a^{78} L_{2i-1} - \frac{1-a^{78}}{2} L_{2i-2} \right)^2 \right]$$

where $b_i$ and $\sigma$ are given in section 4.2.2.2. In fact, the above conditional density is a multivariate normal density with mean

$$\mu = \begin{bmatrix} 1 + a^{154} & a^{153} & \ldots & a^{78} \\ a^{153} & 1 + a^{152} & \ldots & a^{77} \\ \vdots & \vdots & \ddots & \vdots \\ a^{78} & a^{77} & \ldots & 1 + a^2 \end{bmatrix}^{-1} \begin{bmatrix} b_i + a^{77} (L_{2i} - a^{78} L_{2i-1} - b_i) \\ b_i + a^{76} (L_{2i} - a^{78} L_{2i-1} - b_i) \\ \vdots \\ b_i + a (L_{2i} - a^{78} L_{2i-1} - b_i) \end{bmatrix}$$

and variance

$$\Sigma = \sigma^2 \begin{bmatrix} 1 + a^{154} & a^{153} & \ldots & a^{78} \\ a^{153} & 1 + a^{152} & \ldots & a^{77} \\ \vdots & \vdots & \ddots & \vdots \\ a^{78} & a^{77} & \ldots & 1 + a^2 \end{bmatrix}^{-1}$$

Hence, we can generate multivariate normal random variables $X_i$’s, from which $Y_i$’s can be computed straightforwardly.
The simulated $L_i$'s and $Y_i$'s of one sample of 30 days are shown in figure 4.2 with $L_i$'s indicated by circles.

**4.4.2 Choosing parameters**

Choosing particular parameters to make the variances of $L(t)$ and $Y(t)$ small and large respectively can easily result in astronomically high profits. However, the simulated spread process trajectory may simply not be achievable by real stock pairs. Therefore, to be more realistic, we use calibrated parameters $\theta_L, \sigma_L, \delta_1, \theta, \sigma$ from real data for simulation.

The dataset is described in detail in section 4.5. We used the ranking method described in the trading rules to select a ‘good pair’ and a ‘bad pair’ as the ones with
best and worst parameters respectively in the first trading day.

The good pair’s parameters are

$$\theta_L = 1.626237, \sigma_L = 0.229202, \delta_1 = 0.0032902, \theta = 123.22211, \sigma = 0.341101;$$

and the bad pair’s parameters are

$$\theta_L = 9.328820, \sigma_L = 0.198623, \delta_1 = 0.0014689, \theta = 482.18402, \sigma = 0.205307.$$
CHAPTER 4. PAIRS TRADING

Table 4.1: Simulation Trading Results (Top half: the good pair; bottom half: the bad pair)

<table>
<thead>
<tr>
<th>$\epsilon$ level</th>
<th>avg num of trades in 250 days</th>
<th>avg num/percentage of profitable trades</th>
<th>avg num/percentage of trades that revert back to mean</th>
<th>avg pnl</th>
<th>max pnl in 400 simulations</th>
<th>min pnl in 400 simulations</th>
<th>annual Sharpe ratio</th>
<th>annual return</th>
<th>profit per trade in bp (breakeven transac. cost)</th>
</tr>
</thead>
<tbody>
<tr>
<td>98%</td>
<td>29.4</td>
<td>22.8</td>
<td>77.5%</td>
<td>1.4</td>
<td>0.230</td>
<td>0.610</td>
<td>0.028</td>
<td>3.325</td>
<td>80.2%</td>
</tr>
<tr>
<td>95%</td>
<td>54.5</td>
<td>41.4</td>
<td>75.9%</td>
<td>4.6</td>
<td>0.404</td>
<td>0.800</td>
<td>0.052</td>
<td>4.447</td>
<td>180.9%</td>
</tr>
<tr>
<td>90%</td>
<td>91.9</td>
<td>68.1</td>
<td>74.1%</td>
<td>13.1</td>
<td>0.644</td>
<td>1.111</td>
<td>0.320</td>
<td>5.668</td>
<td>409.5%</td>
</tr>
<tr>
<td>98%</td>
<td>27.5</td>
<td>13.4</td>
<td>48.8%</td>
<td>0.1</td>
<td>-0.003</td>
<td>0.095</td>
<td>-0.063</td>
<td>-0.165</td>
<td>-0.7%</td>
</tr>
<tr>
<td>95%</td>
<td>48.2</td>
<td>23.7</td>
<td>49.3%</td>
<td>0.6</td>
<td>0.000</td>
<td>0.107</td>
<td>-0.112</td>
<td>-0.039</td>
<td>0.0%</td>
</tr>
<tr>
<td>90%</td>
<td>77.7</td>
<td>38.6</td>
<td>49.7%</td>
<td>2.6</td>
<td>0.005</td>
<td>0.134</td>
<td>-0.136</td>
<td>0.074</td>
<td>1.2%</td>
</tr>
</tbody>
</table>

Meanwhile for the bad pair, although the number of trades triggered are comparable to the good pair, the profitability is much worse. Slightly less than half the trades are winning ones compared with well over 70% for the good pair. As a result, the Sharpe ratios and annual returns are close to zero.

These simulation results validated our model by showing that a good pair identified by the model can indeed provide stable profits while a bad pair cannot.

For the good pair, as the threshold $\epsilon$ is lowered from 98% to 90% percentile, the average number of trades per simulation is more than tripled from 29 to 92. Annual Sharpe ratio and annual return increased significantly due to the larger number of trades triggered. However, both the winning percentage and the profit per trade dropped slightly, from 77.5% and 78 bps to 74.1% and 70 bps respectively. This is expected since lowering the threshold means lowering the ‘standard’ in identifying trading opportunities.

The results in this section are presented without transaction costs. But as will be
CHAPTER 4. PAIRS TRADING

discussed and analyzed in detail in section 4.5, the profits for the good pair are large enough to cover any reasonable transaction costs estimation. Note that there is only one pair in the simulation trading for simplicity, but in real data trading strategy we have at least 25 potential pairs every day.

4.5 Back-testing on real data

4.5.1 The data

The data for this study are from the NYSE Trade and Quote database on Wharton Research Data Services (WRDS) platform. Tick data for 26 oil company stocks during trading hours 9:30 AM to 4:00 PM are downloaded, then processed to be 5-minutes time series by extracting the first tick price right after each 5-minute mark (i.e., 9:30:00, 9:35:00 etc.).

Initially, 31 stocks with the largest market capital in the Oil Refining & Marketing industry group traded on NYSE and/or NASDAQ were downloaded. After processing tick prices, we removed five stocks with too many missing data\(^3\) all of them non-US companies. The information of the 26 stocks are shown in table 4.2. Notice that our trading universe comprises stocks with extremely large market caps and liquidity, compared with most studies in the pairs trading literature. All the companies have

\(^3\)A missing data refers to the situation where there is no trade in a five-minute interval. There are only 11 total missing data points among the remaining 26 stocks in the whole analysis period. We interpolate these 11 missing points.
### CHAPTER 4. PAIRS TRADING

**Table 4.2: Oil Company Stocks Descriptions and Statistics (as of December 2014)**

<table>
<thead>
<tr>
<th>NYSE Ticker</th>
<th>Market Cap in Billion</th>
<th>Annual Revenue from Google Finance (in M)</th>
<th>Avg Daily Volume from Google Finance (in M)</th>
<th>Avg Daily # of Trades from Tick Data</th>
<th>Min Daily # of Trades from Tick Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exxon Mobil Corporation</td>
<td>XOM</td>
<td>393.87</td>
<td>420,836</td>
<td>12.71</td>
<td>62977</td>
</tr>
<tr>
<td>Royal Dutch Shell plc (ADR)</td>
<td>RDSA</td>
<td>215.73</td>
<td>451,235</td>
<td>2.5</td>
<td>10478</td>
</tr>
<tr>
<td>Chevron Corporation</td>
<td>CVX</td>
<td>207.98</td>
<td>220,264</td>
<td>7.29</td>
<td>42279</td>
</tr>
<tr>
<td>Total SA (ADR)</td>
<td>TOT</td>
<td>132.18</td>
<td>227,969</td>
<td>1.4</td>
<td>5773</td>
</tr>
<tr>
<td>BP plc (ADR)</td>
<td>BP</td>
<td>119.95</td>
<td>370,136</td>
<td>6.78</td>
<td>22895</td>
</tr>
<tr>
<td>ConocoPhillips</td>
<td>COP</td>
<td>82.85</td>
<td>56,185</td>
<td>8.07</td>
<td>34616</td>
</tr>
<tr>
<td>Occidental Petroleum Corporation</td>
<td>OXY</td>
<td>61.82</td>
<td>24,561</td>
<td>6.6</td>
<td>28982</td>
</tr>
<tr>
<td>Statoil ASA(ADR)</td>
<td>STO</td>
<td>57.53</td>
<td>87,781</td>
<td>2.85</td>
<td>7061</td>
</tr>
<tr>
<td>Petróleo Brasileiro Petrobras SA (ADR)</td>
<td>PBR</td>
<td>57.49</td>
<td>141,462</td>
<td>48.82</td>
<td>65939</td>
</tr>
<tr>
<td>EOG Resources Inc</td>
<td>EOG</td>
<td>49.35</td>
<td>14,290</td>
<td>6.21</td>
<td>24409</td>
</tr>
<tr>
<td>Suncor Energy Inc. (USA)</td>
<td>SU</td>
<td>44.34</td>
<td>35,398</td>
<td>4.56</td>
<td>19876</td>
</tr>
<tr>
<td>Anadarko Petroleum Corporation</td>
<td>APC</td>
<td>39.74</td>
<td>14,581</td>
<td>6.05</td>
<td>30750</td>
</tr>
<tr>
<td>Phillips 66</td>
<td>PSX</td>
<td>38.83</td>
<td>171,596</td>
<td>4.7</td>
<td>23528</td>
</tr>
<tr>
<td>Canadian Natural Resource Ltd (USA)</td>
<td>CNQ</td>
<td>35.85</td>
<td>11,182</td>
<td>4.33</td>
<td>16446</td>
</tr>
<tr>
<td>Valero Energy Corporation</td>
<td>VLO</td>
<td>25.47</td>
<td>10,074</td>
<td>7.05</td>
<td>46740</td>
</tr>
<tr>
<td>Marathon Petroleum Corp</td>
<td>MPC</td>
<td>25.19</td>
<td>100,248</td>
<td>3.62</td>
<td>26820</td>
</tr>
<tr>
<td>Devon Energy Corp</td>
<td>DVN</td>
<td>24.16</td>
<td>10,397</td>
<td>4.34</td>
<td>22804</td>
</tr>
<tr>
<td>Apache Corporation</td>
<td>APA</td>
<td>22.31</td>
<td>16,054</td>
<td>4.53</td>
<td>22623</td>
</tr>
<tr>
<td>Hess Corp.</td>
<td>HES</td>
<td>21.87</td>
<td>22,247</td>
<td>3.75</td>
<td>18808</td>
</tr>
<tr>
<td>Pioneer Natural Resources</td>
<td>PXD</td>
<td>20.69</td>
<td>3,506</td>
<td>2.63</td>
<td>14786</td>
</tr>
<tr>
<td>Marathon Oil Corporation</td>
<td>MRO</td>
<td>18.88</td>
<td>14,959</td>
<td>7.75</td>
<td>32701</td>
</tr>
<tr>
<td>Plains All American Pipeline, L.P.</td>
<td>PAA</td>
<td>18.27</td>
<td>42,249</td>
<td>1.66</td>
<td>6058</td>
</tr>
<tr>
<td>Cenovus Energy Inc (USA)</td>
<td>CVE</td>
<td>15.60</td>
<td>16,389</td>
<td>1.85</td>
<td>7051</td>
</tr>
<tr>
<td>Continental Resources, Inc.</td>
<td>CLR</td>
<td>13.18</td>
<td>3,455</td>
<td>3.72</td>
<td>10450</td>
</tr>
<tr>
<td>EQT Corporation</td>
<td>EQT</td>
<td>12.91</td>
<td>1,862</td>
<td>1.82</td>
<td>12835</td>
</tr>
<tr>
<td>Cabot Oil &amp; Gas Corporation</td>
<td>COG</td>
<td>12.83</td>
<td>1,746</td>
<td>6.11</td>
<td>35182</td>
</tr>
</tbody>
</table>

Market caps over $12 billion. On average, there are 68.0 tick prices per minute for each stock. The availability of such a high frequency database is critical for this study.

We first did the back-testing on the period of January 2, 2013 to April 29, 2015 (579 business days). Then, in order to examine the performance of the strategy during market turmoil, we back-tested on an earlier period of July 2, 2007 to December 31, 2008 (374 business days). It is worth pointing out that although the data periods in this study may seem short compared with prior literature, the high-frequency nature of the data set makes it actually larger, in terms of numbers of data points per stock.
579 days with 79 data points per day are equivalent to \(579 \times 79/252 = 181\) years of daily data.

For the more recent 2013–15 period, we used the 26 stocks described before. For the earlier period however, five stocks (CLR, CVE, MPC, PAA, PSX) either had not started trading or did not have enough liquidity. We used the other 21 stocks for the 2007–08 period.

### 4.5.2 Transaction costs and return calculation

Pairs trading strategies aim to capture stable and modest profits from market mispricing. As a result, transaction costs can have a major impact on the profitability. Bowen et al. \[9\] found that a moderate level of 15 basis points (bps) transaction costs\(^4\) would reduce the excess returns by more than 50% on one year’s data of 100 UK stocks. Testing on the US equity market in the period 1963–2009, Do and Faff \[20\] found that profitability of the simple algorithm from the original Gatev et al. paper \[27\] was largely diminished after various transaction costs.

The magnitude of transaction costs depends on many factors such as the type and size of the investor (institutional vs. retail), liquidity of the particular security and the size and timing of the order. Gatev et al. \[27\] estimated a large transaction cost of 162 bps per pair per round-trip for the period 1962–2002. However, the figure has been vastly reduced in recent years due to technology advances. Avellaneda and Lee

---

\(^4\)It was unclear to us whether these transaction costs were per round trip or per trade.
CHAPTER 4. PAIRS TRADING


Transaction costs mainly consist of commissions, bid–ask spreads, and short selling costs [20]. In this study where we consider trading a pool of highly liquid large cap US stocks from the perspective of hedge funds, commissions and short selling costs are negligible. The bid ask spread—also known as bid ask bounce, slippage, or market impact—can be estimated both directly and indirectly. Since all stocks in our investment universe are highly liquid, we used one of the lower estimates in literature as our baseline number, 10 bps per round-trip per pair-trade as in Avellaneda and Lee [5]. An alternative way to estimate bid–ask spreads is to use delayed trading as a proxy. As argued by Gatev et al. [27], if a trade is made one period (one day in their case) after the divergence signal is identified, instead of immediately, the drop in return would be a rough estimate of half the round-trip transaction cost. As will be seen in detail in the next subsection, the proxy result is consistent with our selection of 10 bps.

Another tricky issue in comparing pairs trading studies is the return calculation, which warrants two considerations. The first is the leverage ratio. Pair trading, as a market neutral strategy, has a zero net initial investment (long $1 and short $1 for example) in theory. But it is not zero in practice. In order to short stocks, we need to put margin deposits in the brokerage account. The number of dollars of market
CHAPTER 4. PAIRS TRADING

exposure allowed for every dollar in the margin account is called the margin leverage ratio. We compute our returns as profit or loss divided by the margin, as does the literature.

Avellaneda and Lee [5] used a 4:1 leverage, which means $2 long and $2 short is permitted for every dollar deposited. Gatev et al. [27] defined excess return as the profit/loss for each $1 long–$1 short pair trade. This implied a 2:1 leverage. Large institutional investors can generally get large leverages. In this study, we assumed a 5:1 leverage. Note that although return numbers largely depend on the leverage selection, Sharpe ratios do not, hence are more suited to be compared across studies.

Annualized Sharpe ratio is calculated as

\[
\text{Annualized Sharpe ratio} = \frac{\text{Annualized return}}{\text{Annualized volatility}} = \frac{E(\text{daily return}) \times 252}{sd(\text{daily return}) \times \sqrt{252}}
\]

The second consideration in return calculation is return on committed capital versus return on actual employed capital [27]. The former uses capital related to all the selected pairs for a day; the latter only uses capital related to those pairs that are traded in the day. Gatev et al. [27] argue that ‘... to the extent that hedge funds are flexible in their sources and uses of funds, computing excess return relative to the actual capital employed may give a more realistic measure of the trading profits.’ When discussing results in the next subsection, we refer to the return on actual employed capital (but we displayed both types of returns and corresponding Sharpe
CHAPTER 4. PAIRS TRADING

ratios in tables 4.3–4.6).

To illustrate all the above points, consider an example where we selected 20 pairs each trading day based on calibration results. The margin deposit required for each $1 long–$1 short trade is

\[
\frac{\text{Gross market exposure}}{\text{leverage}} = \frac{25}{5} = \$0.4
\]

Assuming a 10 bp transaction cost per round trip pair trade, then

(a) daily return on **committed** capital

\[
\frac{\text{daily net PNL}}{\text{margin position}} = \frac{(\text{daily PNL}) - 0.001 \times (\# \text{ of trades})}{0.4 \times (\# \text{ of pairs})}
\]

(b) daily return on **actual employed** capital

\[
\frac{\text{daily net PNL}}{\text{margin position}} = \frac{(\text{daily PNL}) - 0.001 \times (\# \text{ of trades})}{0.4 \times (\# \text{ of trades})}
\]

Consider the following three scenarios:

1. If in one day the daily profit is $0.3 with 2 trades, return (a) is \(\frac{0.3-0.002}{0.4\times2} = 3.725\%\); return (b) is \(\frac{0.3-0.002}{0.4\times2} = 37.25\%\)
2. If the daily loss is -$1.2 with 5 trades, return (a) is \(- \frac{1.2 - 0.005}{0.4 \times 20} = -15.06\%\); return (b) is \(- \frac{1.2 - 0.005}{0.4 \times 5} = -60.25\%\).

3. If no trade in one day, the daily return is 0 for both returns (a) and (b).

### 4.5.3 Empirical results

#### 4.5.3.1 June 2013–April 2015

As described in section 3, for each day, we use the previous 100 days’ data to select best pairs (i.e. the formation period is 100 days, trading period is one day). Therefore the first five months (January to May 2013) in the data is left out for calibration and the trading starts from June 2013. We report the trading results for different pair selection criteria and transaction cost levels in table 4.3. As in the simulation, among three threshold \(\epsilon\) levels (98%, 95% and 90% percentiles), the 98% threshold yields the highest profit per trade. Thus, the results reported in this section are based on threshold \(\epsilon = 98\%\).

The top part of table 4.3 shows the results for four different pair selection criteria. As discussed, for each trading day we skipped the pairs with negative estimated \(\theta_L\) and then selected top 25/50/75/100 pairs as trading candidates according to their rankings.

As expected, the average number of trades per day depends on the number of pairs
CHAPTER 4. PAIRS TRADING

Table 4.3: Jun’13–Apr’15 Trading Results for Different Pair Selections and Transaction Costs

<table>
<thead>
<tr>
<th>Pair selection criteria</th>
<th>Transaction cost</th>
<th># of total trades</th>
<th>avg # of trades per day</th>
<th>% of winning trades</th>
<th>Total PNL($) after TC</th>
<th>Ann. Sharpe based on employed capital</th>
<th>Ann. return on employed capital</th>
<th>Ann. Sharpe based on committed capital</th>
<th>Ann. return on committed capital</th>
<th>PNL per trade in bp (break-even TC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>top 25</td>
<td>10 bp</td>
<td>1050</td>
<td>2.20</td>
<td>53.9%</td>
<td>1.946</td>
<td>3.353</td>
<td>148.1%</td>
<td>1.250</td>
<td>10.4%</td>
<td>17.6</td>
</tr>
<tr>
<td>top 50</td>
<td>10 bp</td>
<td>1897</td>
<td>3.97</td>
<td>54.2%</td>
<td>3.710</td>
<td>3.885</td>
<td>187.8%</td>
<td>1.738</td>
<td>10.1%</td>
<td>19.6</td>
</tr>
<tr>
<td>top 75</td>
<td>10 bp</td>
<td>2694</td>
<td>5.64</td>
<td>53.2%</td>
<td>4.371</td>
<td>4.138</td>
<td>192.7%</td>
<td>1.694</td>
<td>7.9%</td>
<td>16.2</td>
</tr>
<tr>
<td>top 100</td>
<td>10 bp</td>
<td>3507</td>
<td>7.34</td>
<td>52.9%</td>
<td>4.879</td>
<td>4.078</td>
<td>190.8%</td>
<td>1.592</td>
<td>6.6%</td>
<td>13.9</td>
</tr>
<tr>
<td>top 50</td>
<td>0 bp</td>
<td>1897</td>
<td>3.97</td>
<td>59.7%</td>
<td>5.607</td>
<td>5.386</td>
<td>346.8%</td>
<td>2.609</td>
<td>15.7%</td>
<td>29.6</td>
</tr>
<tr>
<td>top 50</td>
<td>10 bp</td>
<td>1897</td>
<td>3.97</td>
<td>54.2%</td>
<td>3.710</td>
<td>3.885</td>
<td>187.8%</td>
<td>1.738</td>
<td>10.1%</td>
<td>19.6</td>
</tr>
<tr>
<td>top 50</td>
<td>20 bp</td>
<td>1897</td>
<td>3.97</td>
<td>49.0%</td>
<td>1.813</td>
<td>2.338</td>
<td>85.2%</td>
<td>0.850</td>
<td>4.7%</td>
<td>9.6</td>
</tr>
<tr>
<td>top 50</td>
<td>wait one period</td>
<td>10 bp</td>
<td>1897</td>
<td>3.97</td>
<td>53.1%</td>
<td>2.481</td>
<td>125.3%</td>
<td>1.244</td>
<td>6.6%</td>
<td>13.1</td>
</tr>
</tbody>
</table>

selected: the more pairs we select, the higher number of trades per day. The profit per trade ranges from 14 to 20 bps after deducting the 10 bps transaction costs. We selected the ‘top 50’ as our baseline strategy since it has the best overall performance metrics.

The middle part of table 4.3 shows the impact of transaction costs for the baseline ‘top 50’ selection method. Without transaction costs, the profit per trade is 30 bps. In other words, the break-even transaction cost is 30 bps on this dataset period. Even if we relax the estimate to a more conservative 20 bps per trade, we still have a 10 bps per trade profit and a 2.338 annualized Sharpe ratio, compared with the Sharpe ratio of 1.51 from 2003 to 2007 by Avellaneda and Lee [5].

Finally, we rerun the baseline strategy but with the wait-one-period constraint mentioned before. Gatev et al. [27] argued that when a spread is identified, it is more likely that the winner stock price is an ask price and the loser stock price is
CHAPTER 4. PAIRS TRADING

a bid price. After waiting a period, five minutes in our strategies, the prices are presumably equally likely to be bid or ask prices. Therefore the drop in PNL after waiting for five minutes as opposed to making the trade at the moment of divergence signal, would be a proxy of half the bid ask bounce—the other half happening at convergence, in the same vein. Of course, part of this drop could also be attributed to the natural mean-reversion in prices. Comparing the second and the bottom lines in Table 4.3, the drop in profit per trade is 6.5 bps. If the drop was exclusively due to the bid–ask bounce, the proxy would be $6.5 \times 2 = 13$ bps, which is consistent with our direct estimation of 10 bps.

To see the stability of the strategies over time, we plotted the PNL and returns over the trading period of almost two years, and reported the results by quarter. As seen from Figure 4.3, both the PNL and returns increase quite stably over the period.

In Table 4.4, quarterly results are reported for the baseline ‘top 50’ strategy with 10 bps transaction costs. Out of the seven quarters in the period, six are winning quarters and one breaks even.

4.5.3.2 The year 2008

As a contrarian strategy, pairs trading tends to perform better during markets downturns [19]. The results during the months of recent oil market crash (Jul’14–Jan’15)\(^5\)

\(^5\)Out of the 3.97 average trades per day, only 0.21 or about 5% trades converged, i.e., reverting back to $L_{2i-2}^2 + L_{2i-2}^2$ level before day’s close. Most other spreads were on their way toward $L_{2i-2}^2 + L_{2i-2}^2$ when market closed. As stated in the trading rules, we close all pairs at market’s close. Hence, the waiting-one-period proxy apply to most pairs for only the opening half of the trade.
CHAPTER 4. PAIRS TRADING

(a) Cumulative Daily PNL

(b) Cumulative Daily Return on Actual Employed Capital

(c) Cumulative Daily Return on Committed Capital

Figure 4.3: Back-testing performance for period Jun’13–Apr’15 (top 50 pairs, threshold=98%, transaction costs=10 bps)
CHAPTER 4. PAIRS TRADING

Table 4.4: Jun’13-Apr’15 Trading Results by Quarter (Top 50 pairs, transaction costs = 10 bps)

<table>
<thead>
<tr>
<th>Quarter</th>
<th># of total trades</th>
<th>% of winning trades</th>
<th>Total PNL ($)</th>
<th>Annl Sharpe based on employed capital</th>
<th>Annl return on employed capital</th>
<th>Annl Sharpe based on committed capital</th>
<th>Annl return on committed capital</th>
<th>PNL per trade in bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013 Q3</td>
<td>209</td>
<td>53.1%</td>
<td>0.070</td>
<td>1.96</td>
<td>52.4%</td>
<td>0.39</td>
<td>1.3%</td>
<td>3</td>
</tr>
<tr>
<td>2013 Q4</td>
<td>277</td>
<td>49.5%</td>
<td>0.661</td>
<td>4.48</td>
<td>475.6%</td>
<td>2.82</td>
<td>14.3%</td>
<td>24</td>
</tr>
<tr>
<td>2014 Q1</td>
<td>212</td>
<td>55.2%</td>
<td>0.304</td>
<td>0.96</td>
<td>27.1%</td>
<td>3.06</td>
<td>6.5%</td>
<td>3</td>
</tr>
<tr>
<td>2014 Q2</td>
<td>188</td>
<td>54.8%</td>
<td>0.690</td>
<td>5.56</td>
<td>222.8%</td>
<td>2.50</td>
<td>14.6%</td>
<td>37</td>
</tr>
<tr>
<td>2014 Q3</td>
<td>230</td>
<td>53.0%</td>
<td>0.174</td>
<td>4.43</td>
<td>132.8%</td>
<td>1.34</td>
<td>3.5%</td>
<td>8</td>
</tr>
<tr>
<td>2014 Q4</td>
<td>500</td>
<td>56.8%</td>
<td>1.489</td>
<td>5.41</td>
<td>459.9%</td>
<td>2.43</td>
<td>34.3%</td>
<td>30</td>
</tr>
<tr>
<td>2015 Q1</td>
<td>204</td>
<td>50.0%</td>
<td>-0.007</td>
<td>3.91</td>
<td>172.0%</td>
<td>-0.06</td>
<td>-0.2%</td>
<td>0</td>
</tr>
</tbody>
</table>

show a promising performance: the average monthly PNL without transaction costs during this seven-month span is 43% higher than the whole period. To further test this hypothesis and verify our strategies, we repeated the analysis on the whole year 2008, during which the oil market spiked to an all-time high of $145 per barrel in July then crashed to $30 in December amid global financial crisis, as shown in the US crude oil benchmark index West Texas Intermediate (WTI) history price chart (figure 4.4).

The original 26 stocks we selected were not all available for the period July 2007–Dec 2008 (the last five months of 2007 were needed for calibration). Three (CVE, MPC, PSX) had not started trading; two (CLR, PAA) had too many missing data due to low liquidity. Therefore we used 21 stocks and \( \binom{21}{2} = 210 \) total pairs for this period.

The results for 2008 are presented in table 4.5. The average number of trades per
day for the baseline ‘top 50’ strategy remarkably increases to 6.26, from 3.97 in the 2013–15 period. The higher numbers of trading opportunities were driven by higher volatilities in stock prices.\(^6\) The annual volatilities of all stocks in the 2013–15 period range from 11.1% to 37.7% with mean 23.1%. In 2008, they range from 31.2% to 63.0% with mean 50.3%. The higher volatilities in stock prices were driven by crude oil’s volatile movements.\(^{29}\) WTI’s volatilities were 27% in Jun’13–Apr’15 and 55% in 2008.

Furthermore, the strategies’ returns are much larger in 2008. For the baseline ‘top 50’ strategy with a 10 bps transaction cost, the annualized return is 187.8% for the

---

\(^6\)We recorded the number of times each stock is selected and traded over the whole period, to see if there are any discrepancies. Not surprisingly, they have a strong relationship with stocks volatility. The correlation between a stock’s volatility and number of times it being selected and traded are 88.2% and 86.6% respectively.
recent period and 1787.6% for 2008. The 7.17 Sharpe ratio of 2008 also dominates the 3.89 in 2013–15. The breakeven transaction cost is 88 bps for 2008 compared with 30 bps for 2013–15. More impressively, the quality performance in 2008 is consistent throughout the year as shown in the quarterly breakdown in table 4.6. In fact, the returns before transaction costs are positive in every month.

Lastly in table 4.5, the drop in profit per trade when we apply the wait-one-period
constraint is $78 - 54 = 24$ bps. As in section 4.5.3.1 less than 13 bps of the drop is likely due to the bid–ask bounce, while the rest is presumably caused by the convergence of the spread, which is more prominent in the more volatile 2008. The fact that our strategy is still profitable in both periods after posing the wait-one-period constraint shows its robustness to the speed of execution [9].

Within 2008, the trading strategy performed extremely well in the second half of year (figure 4.5), coinciding with the nosedive of WTI price. This is clear from table 4.6: the order of the strategy performance of the four quarters is the exact reverse order of WTI’s quarterly performance. The monthly returns of the baseline trading strategy (not reported) and WTI index are strongly negatively correlated, with a correlation coefficient -0.78. This figure is only -0.04 in the 2013–15 period. We identified two reasons for this significant difference. The first is again volatility: higher WTI and stock volatilities can much better translate the plummeting prices into trading profits through more trading opportunities. After all, pairs trading fundamentally relies on temporary relative mispricing of two stocks. The second reason is that the oil market crash in 2008 was more dramatic and more impactful. In 2008, the WTI index plunged almost 80% in less than six months, compared with a drop of almost 60% in seven months from mid-2014 to early 2015. The stocks in our trading universe lost 49% on average during the 2008 crash and only 29% in the recent market turmoil.
Figure 4.5: Back-testing performance for period 2008 (top 50 pairs, threshold=98%, transaction costs=10 bps)
CHAPTER 4. PAIRS TRADING

4.6 Conclusion

This chapter introduces a doubly mean-reverting process to model stock price spreads. We developed intraday pairs trading strategies using high frequency data with five-minute intervals on oil company stocks. Results from both simulations and real data back-testing display significant realized profits. In particular, we are able to achieve a 3.9 annualized Sharpe ratio and a 188% annualized return after transaction costs for the period June 2013 to April 2015. We also tested the hypothesis that pairs trading strategies perform better in market turmoil by back-testing on 2008 data. The impressive Sharpe ratio and annualized return of 7.2 and 1788% respectively in that year underpin this theory as well as the fundamental relationship that oil company stocks are driven by crude oil price. We also showed that the strategy is robust to both speed of execution and reasonable transaction costs.

There are several possible directions for future research. First, the frequencies of the two processes may be changed. To utilize stock price data with high liquidity, intervals smaller than five minutes could be used as the frequency for $Y(t)$. On the other hand, we can increase the interval length of $L(t)$, to make the holding period longer, enabling overnight positions. Second, some details in the strategy implementation may be refined to achieve higher returns, such as the optimal thresholds to enter and exit a trade, and the training windows of 100 and 30 days. This has to be done in a careful manner to avoid over-fitting and data-snooping biases. Third, other types of data available on the NYSE Trade and Quote database can
be included in the model. In particular, volume data may be used to adjust for different trading intensities throughout the day. Lastly, the model can be extended (a) from pairs trading to groups trading (also known as generalized pairs trading) with the simultaneous buying and selling of more than two stocks which co-move in some pattern and (b) from stock pairs within the Oil Refining industry group to cross-industry pairs, e.g., those in the highly related Oil Services & Equipment industry group.
Appendix A

Derivation of likelihood function

for $Y(t)$ and the maximum likelihood estimation

The 30 summands in equation (4.12) are similar. Define $Y_0 = L_0 = 0$ so that all 30 terms have the form $\ln f(Y_{79(i-1)+1}, Y_{79(i-1)+2}, \ldots, Y_{79i}|Y_{79(i-1)} = L_{2(i-1)}, Y_{79(i-1)+1} = L_{2i-1}, Y_{79i} = L_{2i})$. We work on the first term for now. Since $(Y_1, Y_2, \ldots, Y_{79})$ are not
APPENDIX A. DERIVATION OF LIKELIHOOD FUNCTION FOR Y(T) AND THE MAXIMUM LIKELIHOOD ESTIMATION

jointly normal, we perform change of variables. Let

\[ X_1 = Y_2 - aY_1 \]
\[ X_2 = Y_3 - aY_2 \]
\[ \ldots \]
\[ X_{77} = Y_{78} - aY_{77} \]
\[ X_{78} = Y_{79} - aY_{78} \]

Then \( X_1, \ldots, X_{78} \) are i.i.d. \( N(b_i, \hat{\sigma}^2) \), where

\[ a = e^{-\theta \delta} \]
\[ b_i = \frac{L_{2i-2} + L_{2i-1}}{2} (1 - e^{-\theta \delta}) = \frac{L_{2i-2} + L_{2i-1}}{2} (1 - a), \quad i = 1, \ldots, 30 \]
\[ \hat{\sigma} = \sigma \sqrt{\frac{1 - e^{-2\theta \delta}}{2\theta}} = \sigma \sqrt{\frac{1 - a^2}{2\theta}} \]

Use these 78 equations to recursively cancel out \( Y_2 \) to \( Y_{78} \), and express \( Y_{79} \) using \( X \)'s

\[ L_2 = Y_{79} = X_{78} + aX_{77} + a^2X_{76} + \cdots + a^{77}X_1 + a^{78}Y_1 \]
\[ = X_{78} + aX_{77} + a^2X_{76} + \cdots + a^{77}X_1 + a^{78}L_1 \]
APPENDIX A. DERIVATION OF LIKELIHOOD FUNCTION FOR $Y(T)$ AND THE MAXIMUM LIKELIHOOD ESTIMATION

Then, the conditional likelihood

$$
\ln f(X_1, X_2, \ldots, X_{78}|Y_0 = L_0, Y_1 = L_1, Y_{79} = L_2) = \ln \left( \frac{f(x_1) \cdots f(x_{77}) f(L_2 - a^{78}L_1 - a^{77}x_1 - \cdots - ax^{77})}{f_U(L_2 - a^{78}L_1)} \right) = \ln(*) \quad (A.1)
$$

where

$$
U = X_{78} + aX_{77} + a^2X_{76} + \cdots + a^{77}X_1 \\
\sim N(b_1(1 + a + \cdots + a^{77}), \hat{\sigma}^2(1 + a^2 + \cdots + a^{154})) = N(b_1 \frac{1 - a^{78}}{1 - a}, \hat{\sigma}^2 \frac{1 - a^{156}}{1 - a^2})
$$

The $f$ without subscript is the density for $N(b_1, \hat{\sigma}^2)$.

The numerator in (*) is

$$
\left( \frac{1}{\hat{\sigma}\sqrt{2\pi}} \right)^{78} \exp \left[ -\frac{1}{2\hat{\sigma}^2} \left( (x_1 - b_1)^2 + \cdots + (x_{77} - b_1)^2 + (L_2 - a^{78}L_1 - a^{77}x_1 - \cdots - ax^{77} - b_1)^2 \right) \right] \quad (A.2)
$$

The denominator in (*) is

$$
\frac{1}{\hat{\sigma}\sqrt{2\pi}} \sqrt{\frac{1 - a^2}{1 - a^{156}}} \exp \left[ -\frac{1}{2\hat{\sigma}^2} \frac{1 - a^2}{1 - a^{156}} \left( L_2 - a^{78}L_1 - b_1 \frac{1 - a^{78}}{1 - a} \right)^2 \right] = \frac{1}{\hat{\sigma}\sqrt{2\pi}} \sqrt{\frac{1 - a^2}{1 - a^{156}}} \exp \left[ -\frac{1}{2\hat{\sigma}^2} \frac{1 - a^2}{1 - a^{156}} \left( L_2 - \frac{1 + a^{78}}{2}L_1 - \frac{1 - a^{78}}{2}L_0 \right)^2 \right] \quad (A.3)
$$
APPENDIX A. DERIVATION OF LIKELIHOOD FUNCTION FOR $Y(T)$ AND THE MAXIMUM LIKELIHOOD ESTIMATION

Plug equation (A.2) and (A.3) into (A.1). Then plug (A.1) and 29 other similar terms into equation (4.12),

$$
\mathcal{L} = \ln f(Y | \tilde{Y}) = \sum_{i=1}^{30} \ln f(Y_{79(i-1)+1}, \ldots, Y_{79i} | Y_{79(i-1)} = L_{2(i-1)}, Y_{79(i-1)+1} = L_{2i-1}, Y_{79i} = L_{2i})
$$

$$
= \sum_{i=1}^{30} \left\{ -77 \ln(\hat{\sigma}) - \frac{77}{2} \ln(2\pi) - \frac{1}{2\hat{\sigma}^2} \left[ (x_{78(i-1)+1} - b_i)^2 + \cdots \right] + (x_{78i-1} - b_i)^2 + (L_{2i} - a_{78} L_{2i-1} - a_{78} x_{79(i-1)+1} - \cdots - a x_{79i-1} - b_i)^2 \right\}
$$

(plug in $\hat{\sigma}$)

$$
= \sum_{i=1}^{30} \left\{ -77 \ln(\sigma) - \frac{77}{2} \ln\left(\frac{1-a^2}{1-a^2}\right) - \frac{77}{2} \ln(\pi) - \frac{\theta}{\sigma^2 (1-a^2)} \left[ \sum_{j=1}^{78} (x_{78(i-1)+j} - b_i)^2 \right] + \frac{1}{2} \ln\left(\frac{1-a^{156}}{1-a^2}\right) + \frac{\theta}{\sigma^2 (1-a^{156})} \left( L_{2i} - \frac{1+a_{78}}{2} L_{2i-1} - \frac{1-a_{78}}{2} L_{2i-2} \right)^2 \right\}
$$

(plug in $b_{i-1}$ and $x$)

$$
= \sum_{i=1}^{30} \left\{ -77 \ln(\sigma) - \frac{77}{2} \ln\left(\frac{1-a^2}{\theta}\right) - \frac{77}{2} \ln(\pi) - \frac{\theta}{\sigma^2 (1-a^2)} (**) \right\}
$$

$$
+ \frac{1}{2} \ln\left(\frac{1-a^{156}}{1-a^2}\right) + \frac{\theta}{\sigma^2 (1-a^{156})} \left( L_{2i} - \frac{1+a_{78}}{2} L_{2i-1} - \frac{1-a_{78}}{2} L_{2i-2} \right)^2 \right\} \quad (A.4)
$$

where
APPENDIX A. DERIVATION OF LIKELIHOOD FUNCTION FOR $Y(T)$ AND THE MAXIMUM LIKELIHOOD ESTIMATION

\[
(**) = \sum_{j=1}^{78} \left( Y_{79(i-1)+j+1} - aY_{79(i-1)+j} + (a - 1) \frac{L_{2i-2} + L_{2i-1}}{2} \right)^2 \\
= \sum_{j=1}^{78} Y_{79(i-1)+j+1}^2 + a^2 \sum_{j=1}^{78} Y_{79(i-1)+j}^2 + 78(a - 1)^2 \frac{(L_{2i-2} + L_{2i-1})^2}{4} \\
+ (a - 1)(L_{2i-2} + L_{2i-1}) \sum_{j=1}^{78} Y_{79(i-1)+j+1} - a(a - 1)(L_{2i-2} + L_{2i-1}) \sum_{j=1}^{78} Y_{79(i-1)+j} \\
- 2a \sum_{j=1}^{78} Y_{79(i-1)+j+1} Y_{79(i-1)+j} 
\]  
(A.5)

Plug (A.5) into (A.4) and expand the summations,

\[
\mathcal{L}(\theta, \sigma) = -77 \times 30 \ln(\sigma) - \frac{77 \times 30}{2} \ln \left( \frac{1 - a^2}{\theta} \right) - \frac{77 \times 30}{2} \ln(\pi) + \frac{30}{2} \ln \left( \frac{1 - a^{156}}{1 - a^2} \right) \\
- \frac{\theta}{\sigma^2(1 - a^2)}(Aa^2 + Ba + C) + \frac{\theta}{\sigma^2(1 - a^{156})}(Da^{156} + Ea^{78} + F) 
\]  
(A.6)
APPENDIX A. DERIVATION OF LIKELIHOOD FUNCTION FOR \( Y(T) \) AND THE MAXIMUM LIKELIHOOD ESTIMATION

where

\[
A = \sum_{i=1}^{30} \left[ \sum_{j=1}^{78} Y_{79(i-1)+j}^2 + 78 \left( \frac{(L_{2i-2} + L_{2i-1})^2}{4} \right) - (L_{2i-2} + L_{2i-1}) \sum_{j=1}^{78} Y_{79(i-1)+j} \right]
\]

\[
B = \sum_{i=1}^{30} \left[ \left( -156 \frac{(L_{2i-2} + L_{2i-1})^2}{4} + (L_{2i-2} + L_{2i-1}) \sum_{j=1}^{78} Y_{79(i-1)+j+1} + (L_{2i-2} + L_{2i-1}) \sum_{j=1}^{78} Y_{79(i-1)+j+1} \right) \right]
\]

\[
C = \sum_{i=1}^{30} \left[ \sum_{j=1}^{78} Y_{79(i-1)+j+1}^2 + 78 \left( \frac{(L_{2i-2} + L_{2i-1})^2}{4} \right) - (L_{2i-2} + L_{2i-1}) \sum_{j=1}^{78} Y_{79(i-1)+j+1} \right]
\]

\[
D = \sum_{i=1}^{30} \left[ \left[ \frac{1}{4} L_{2i-1}^2 + \frac{1}{4} L_{2i-2}^2 - \frac{1}{2} L_{2i-1} L_{2i-2} \right] \right]
\]

\[
E = \sum_{i=1}^{30} \left[ \left[ \frac{1}{2} L_{2i-1}^2 - \frac{1}{2} L_{2i-2}^2 - L_{2i} L_{2i-1} + L_{2i} L_{2i-2} \right] \right]
\]

\[
F = \sum_{i=1}^{30} \left[ L_{2i}^2 + \frac{1}{4} L_{2i-1}^2 + \frac{1}{4} L_{2i-2}^2 - L_{2i} L_{2i-1} - L_{2i} L_{2i-2} + \frac{1}{2} L_{2i-1} L_{2i-2} \right]
\]

In the expression of \( \mathcal{L} \) in (A.6), the two parameters are \( \theta \) and \( \sigma \). \( A, B, C, D, E \) and \( F \) are functions of \( L_i \) and \( Y_i \); \( a = e^{-\theta \delta} \) contains parameter \( \theta \); \( \delta = \frac{\delta}{78} \) is the discretization step size.

Setting first order derivatives of \( \mathcal{L} \) with respect to \( \sigma \) to zero

\[
\frac{\partial \mathcal{L}}{\partial \sigma} = -77 \times 30 \frac{1}{\sigma} - \frac{2}{\sigma^3} \left[ \frac{\theta}{1 - a^{156}} (Da^{156} + Ea^{78} + F) - \frac{\theta}{1 - a^2} (Aa^2 + Ba + C) \right] = 0
\]

(A.7)

The optimal pair \((\theta, \sigma)\) that maximizes \( \mathcal{L} \) satisfies \( \frac{\partial \mathcal{L}}{\partial \sigma} = 0 \). From (A.7), we can express
APPENDIX A. DERIVATION OF LIKELIHOOD FUNCTION FOR $y(T)$ AND
THE MAXIMUM LIKELIHOOD ESTIMATION

$\sigma$ using $\theta$

$$\sigma(\theta) = \sqrt{\frac{2}{77 \times 30} \left[ \frac{\theta}{1-a^2} (Aa^2 + Ba + C) - \frac{\theta}{1-a^{156}} (Da^{156} + Ea^{78} + F) \right]} \quad (A.8)$$

Plug (A.8) into (A.6), $L(\theta, \sigma)$ becomes $L^*(\theta)$:

$$L^*(\theta) =$$

$$- 77 \times 30 \ln(\sigma(\theta)) - \frac{77 \times 30}{2} \ln\left(\frac{1-a^2}{\theta}\right) - \frac{77 \times 30}{2} \ln(\pi) + \frac{30}{2} \ln\left(\frac{1-a^{156}}{1-a^2}\right) - \frac{77 \times 30}{2}$$

The maximal solution for $L(\theta, \sigma)$ is the maximal solution for $L^*(\theta)$, which is solved numerically.
Bibliography


BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


Vita

Bo Liu was born on June 10, 1985 in Liaoning, China. Bo received his B.S. in Mechanical Engineering from Fudan University in Shanghai, China in 2007. He then went to study at Claremont Graduate University and obtained two M.S. degrees in Financial Engineering and Mathematics, in 2009 and 2010 respectively, before starting the Applied Math and Statistics Ph.D. program at Johns Hopkins University. At Hopkins, Bo was the teaching assistant for 15 graduate level courses for the AMS department. He was awarded the Paul V. Renoff Fellowship in 2010–2011 and the Professor Joel Dean Award for Excellence in Teaching twice, in 2012 and 2015. He specializes mainly in financial mathematics but has also done work on statistical analysis of the Alzheimer’s disease and published two papers on NeuroImage and Neurobiology of Aging.

In college, Bo was a member of Fudan’s road cycling team and mountaineering team. During his time at Hopkins, Bo finished Baltimore marathon four times (full, half and team relay). Bo is also a powerlifting aficionado.