Pairs trading

Gesina Gorter

December 12, 2006

Contents

Chapter 1

Introduction

1.1 IMC

IMC, International Marketmakers Combination, was founded in 1989. IMC is a diversified financial company. The company started as a market maker on the Amsterdam Options Exchange. Apart from its core business activity trading, it is also active in asset management, brokerage, product development and derivatives consultancy. IMC Trading is IMC's largest operational unit and has been the core of the company for the past 17 years. IMC Trading trades solely for its own account and benefit. IMC is active in the major markets in Europe and the US and has offices in Amsterdam, Zug (Switzerland), Sydney and Chicago. By trading a large number of different securities in different markets, the company is able to keep its trading risk to a minimum.

The dealingroom in Amsterdam is divided in two main sections: Marketmaking and Cash. Marketmaking's main focus is on option trading, a market maker for a certain option will quote both bid and offer prices on the option and make profits from the bid-ask spread. The Cash or Equity desk is dedicated to the worldwide arbitrage of diverse financial instruments. Arbitrage is a trading strategy that takes advantages of two or more securities being mispriced relative to each other. Pairs trading is one of the many trading strategies with Cash.

1.2 Pairs trading

History Pairs trading or statistical arbitrage was first developed and put into practice by Nunzio Tartaglia, while working for Morgan Stanley in the 1980s. Tartaglia formed a group of mathematicians, physicists and computer scientists to develop automated trading systems to detect and make use of mispricings in financial markets. One of the computer scientists on Tartaglia's team was the famous David Shaw. Pairs trading was one of the most profitable strategies that was developed by this team. With members of the team gradually spreading to other firms, so did the knowledge of pairs trading. Vidyamurthy [15] presents a very insightful introduction to pairs trading.

Motivation The general 'rule of thumb' in trading is to sell overvalued securities and buy undervalued ones. It is only possible to determine that a security is overvalued or undervalued if the true value of the security is known. The true value can be very difficult to determine. Pairs trading is about relative pricing, so that the true value of the security is not important. Relative pricing is based on the idea that securities with similar characteristics should be priced more or less the same. When prices of two similar securities are different, one security is overpriced with respect to its 'true value' or the other one underpriced or both.

Pure arbitrage is making risk-less use of mispricing, which is why one could call this a deterministic moneymaking machine. The most pure form of arbitrage is profitably buying and selling the exact same security on different exchanges. For example, one could buy a share in Royal Dutch on the Amsterdam exchange at € 25.75 and sell the same share on the Frankfurt exchange at € 26.00. Because shares in Royal Dutch are inter-exchangeable, such a trade would result in a flat position and thus risk-less money.

Although pairs trading is called an arbitrage strategy, it is not risk-free at all. The key to success in pairs trading lies in the identification of pairs and an efficient trading algorithm. Pairs trading is an arbitrage strategy that makes advantage of a mispricing between two securities. It involves putting on positions when there is a certain magnitude of mispricing, buying the lower-priced security and selling the higher-priced. Hence, the portfolio consists of a long position in one security and a short position in the other. The

expectation is that the mispricing will correct itself, and when this happens the positions are reversed. The higher the magnitude of mispricing when positions are put on, the higher the profit potential.

Example To determine if two securities form a pair is not trivial but there are some securities that are obvious pairs. For example one fundamentally obvious pair is Royal Dutch and Totalfina, both being European oil-producing companies. One can easily argue that the value of both companies is greatly determined by the oil price and hence that movements of the two securities should be closely related to each other. In this example, let's assume that historically, the value of one share Totalfina is at 8 times a share Royal Dutch. Assume at time t_0 it is possible to trade Royal Dutch at € 26.00 and Totalfina at € 215.00. Because 8 times € 26 is € 208, we feel that Totalfina is overpriced, or Royal Dutch is underpriced or both. So we will sell one share in Totalfina and buy 8 shares in Royal Dutch, with the expectation that Totalfina becomes cheaper or Royal Dutch becomes more expensive or both. Assume at t_1 the prices are € 26.00 and € 208, we will have made a profit of € 215 minus € 208 is € 7. We would have made the same profit if at t_1 the prices are € 26.875 (215 divided by 8) and € 215.00 respectively. In conclusion, this strategy does not say anything about the true value of the stocks but only about relative prices. In this example a predetermined ratio of 8 was used, based on historical data. How to use historical data to determine this ratio will be discussed in paragraph 2.4.

1.3 Graduation project

The goal of this project is to apply statistical techniques to find relationships between stocks in all markets that IMC is active in, based solely on the history of the prices of the stocks. The closing prices of these stocks, dating back two years, is the only data that will be used in this analysis. The goal is to find pairs of stocks whose movements are close to each other.

IMC is already trading a lot of pairs which were found by fundamental analysis and by applying their trading strategy to historical data (backtesting). No statistical analysis was made. From trading experience, IMC is able to make a distinction between good and bad pairs based on profits. IMC has provided a selection of ten pairs that are different in quality.

The main focus of this project will be modeling the relationships between stocks, such that we can identify a good pair based on statistical analysis instead of fundamental analysis or backtesting. The resulting relationships will be put in order of the strength of co-movement and profitability.

Although one could study pairs trading between all sorts of financial instruments, such as options, bonds and warrants, this project focuses on trading pairs that consist of two stocks.

1.4 Outline

In the next chapter a trading strategy for pairs will be derived, it illustrates how money is made and what properties a good pair has. In chapter 3 some basics of time series analysis is briefly stated, which we will need for the concept of cointegration. Chapter 4 discusses cointegration and two methods for testing, the Engle-Granger and the Johansen method. Also in this chapter a start is made with an alternative method. The Engle-Granger method makes use of an unit root test named Dickey-Fuller, the properties of this unit root test will be derived in chapter 5. The properties of the Engle-Granger method are found by simulation in chapter 6. IMC has provided 10 pairs for investigation. The results of the trading strategy and cointegration tests are stated in chapter 7, the pairs are also put in order of profitability and cointegration. After the conclusions in chapter 8, some suggestions for alternative trading strategies are made in chapter 9. In this chapter we will also give some recommendations for further research.

Chapter 2

Trading strategy

IMC first started to identify pairs of stock based on fundamental analysis, which means they have investigated similarities between companies in products, policies, dependencies of market circumstances, etcetera.

When a pair is identified, the question remains how to make money. In this chapter, a trading strategy is explained that is quite similar to the strategy used by IMC. It is not exactly the same strategy because IMC does not want to give away a ready-to-go-and-make-money trading strategy but also because essential parts of their strategy, like the selection of parameters, are based on 'gut-feeling' and is in the hands of the trader. That makes it at least very difficult to write down a general model of their trading strategy.

2.1 Introductory example

Assume we have two stocks X and Y that form a pair based on fundamental analysis. Also available are the closing prices of these stocks dating back 2 years, which form times series $\{x_t\}_{t=0}^{T}$ and $\{y_t\}_{t=0}^{T}$ as shown in figure 2.1. In one year there are approximately 260 trading days, so two years of closing prices form a dataset of approximately 520 observations for each stock.

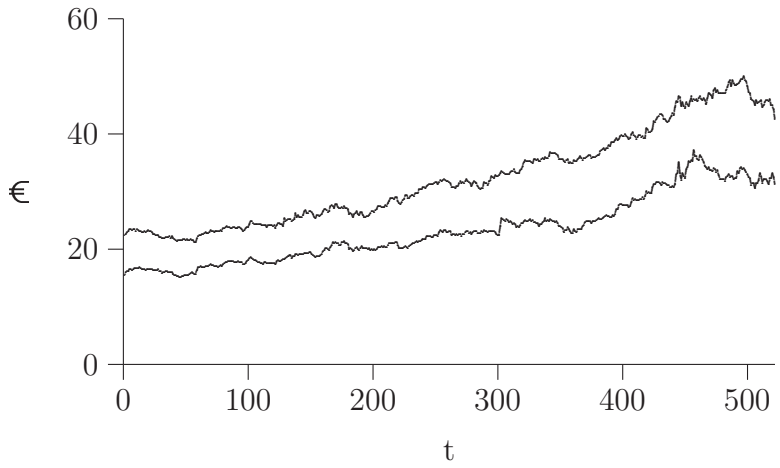

Figure 2.1: Times series x_t and y_t.

The first half of observations are used to determine certain parameters of the trading strategy. The second half are used to backtest the trading strategy based on these parameters, i.e., to test whether the strategy makes money on this pair.

The average ratio of Y and X of the first 260 observations,

$$\bar{r} = \frac{1}{260} \sum_{t=0}^{259} \frac{y_t}{x_t} \ ,$$

in this example is 1.36, which means that 1 stock of Y is approximately 1.36 stock of X during this time period. Although the average ratio is probably not the best estimator, we will use it in the trading strategy to calculate a quantity called *spread* for each value of t:

$$s_t = y_t - \bar{r} x_t.$$

If the price processes of X and Y were perfectly correlated, that is if X and Y changes in the same direction and in the same proportion (for every $t > 0$, $y_t = \alpha x_t$ for some $\alpha > 0$, so the correlation coefficient is $+1$), the spread is zero for all t and we could not make any money because X nor Y are ever over- or underpriced. However, perfect correlation is hard to find in real life. Indeed, in this example the stocks are not perfectly correlated, as we can see in figure 2.2.

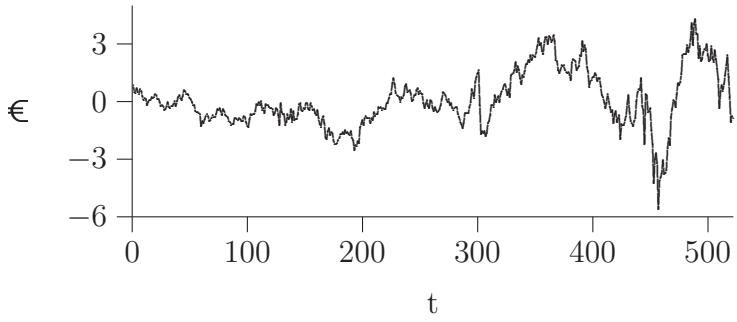

Figure 2.2: Spread s_t.

As mentioned before, we like to buy cheap and sell expensive. If the spread is below zero, stock Y is cheap relative to stock X. The other way around, if the spread is above zero stock Y is expensive relative to stock X (another way to put it is that X is cheap in comparison with Y). So basically the trading strategy is to buy stock Y and sell stock X at the ratio 1:1.36 if the spread is a certain amount below zero, which we call threshold Γ. When the spread comes back to zero, the position is flattened, which means we sell Y and buy X in the same ratio so there is no position left. In that case, we have made a profit of Γ. An important requirement is that we can sell shares we do not own, also called short selling. In summary, we put on a portfolio, containing one long position and one short, if the spread is Γ or more away from zero. We flatten the portfolio when the spread comes back to zero. Just like the average ratio, Γ is determined by the first half of observations. In this example we determined a Γ of 0.40. The way Γ has been calculated will be discussed in paragraph 2.4.

After determination of the parameters, the trading strategy is applied to the second half of observations in the dataset. This results in 13 times making a profit of Γ. In other words, the spread moves 13 times away from 0 with at least Γ and back to 0. Note that this involves 26 trading instances, since putting on and flattening a position requires two. Figure 2.3 and table 2.1 shows all 26 trading instances. The profit, made here, is at least 13Γ: We use closing prices instead of intra-day data, so we do not trade at exactly $-\Gamma$, 0 and Γ as we can see in table 2.1.

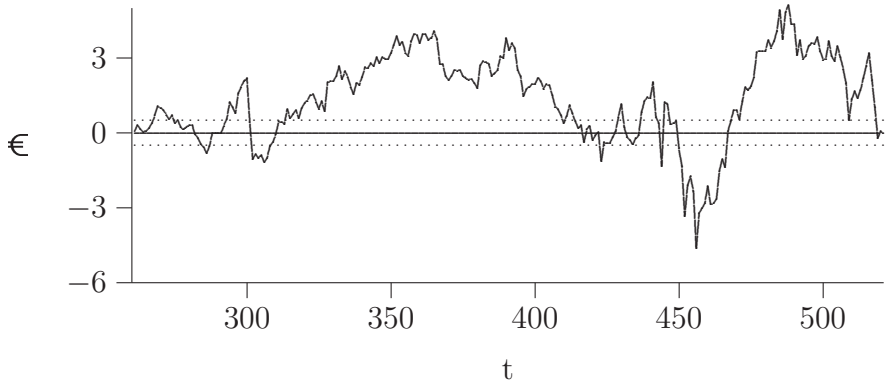

Figure 2.3: Spread s_t and Γ.

Table 2.1: Trading instances strategy I.

trade	t	s_t	position (Y,X)	price Y	price X	profit
1	268	0.69	(-1,+1.36)	31.49	22.63	-
2	282	-0.07	flat	31.37	23.11	0.76
3	284	-0.47	(+1,-1.36)	30.54	22.79	-
4	289	0.01	flat	31.43	23.10	0.48
5	293	0.55	(-1,+1.36)	32.05	23.15	-
6	300	-0.16	flat	32.81	24.23	0.71
7	302	-1.05	(+1,-1.36)	33.57	25.44	-
8	310	0.17	flat	33.56	24.54	1.22
9	311	0.45	(-1,+1.36)	33.58	24.34	-
10	420	-0.30	flat	40.33	29.85	0.75
11	423	-1.15	(+1,-1.36)	40.79	30.82	-
12	428	0.08	flat	43.15	31.65	1.23
13	429	0.65	(-1,+1.36)	43.43	31.44	-
14	432	-0.19	flat	42.60	31.45	0.84
15	434	-0.47	(+1,-1.36)	42.16	31.33	-
16	435	0.04	flat	42.61	31.28	0.51
17	437	0.82	(-1,+1.36)	42.79	30.84	-
18	440	-0.25	flat	44.01	32.52	1.07
19	444	-1.33	(+1,-1.36)	46.53	35.17	-
20	445	0.12	flat	46.17	33.84	1.45
21	446	1.24	(-1,+1.36)	46.32	33.13	-
22	449	-0.17	flat	45.89	33.85	1.41
23	450	-0.63	(+1,-1.36)	45.46	33.87	-
24	467	0.05	flat	46.19	33.91	0.68
25	468	0.48	(-1,+1.36)	47.16	34.31	-
26	519	-0.21	flat	44.95	33.19	0.69
				total profit		11.80

Rather than closing the position at 0, one could also choose to reverse the position when the spread reaches Γ in the other direction. Assume we have sold 1 Y and bought 1.36 X, because the spread was larger than Γ, we could now wait until the spread reaches $-\Gamma$ and buy 2 times Y and sell 2 times 1.36 X. As a result, we are now left with a portfolio of long 1 Y and short 1.36 X. This results in one initial trade and 12 trades reversing the position. Note that the profit of reversing the position is 2Γ, so the total profit is at least 12 times 2Γ. These trades are shown in table 2.2.

Table 2.2: Trading instances strategy II.

trade	t	s_t	position (Y,X)	price Y	price X	profit
1	268	0.69	(-1,+1.36)	31.49	22.63	-
2	284	-0.47	(+1,-1.36)	30.54	22.79	1.16
3	293	0.55	(-1,+1.36)	32.05	23.15	1.02
4	302	-1.05	(+1,-1.36)	33.57	25.44	1.60
5	311	0.45	(-1,+1.36)	33.58	24.34	1.50
6	423	-1.15	(+1,-1.36)	40.79	30.82	1.60
7	429	0.65	(-1,+1.36)	43.43	31.44	1.80
8	434	-0.47	(+1,-1.36)	42.16	31.33	1.12
9	437	0.82	(-1,+1.36)	42.79	30.84	1.29
10	444	-1.33	(+1,-1.36)	46.53	35.17	2.15
11	446	1.24	(-1,+1.36)	46.32	33.13	2.57
12	450	-0.63	(+1,-1.36)	45.46	33.87	1.87
13	468	0.48	(-1,+1.36)	47.16	34.31	1.11
				total profit		18.79

This change of strategy reduces the number of trading instances on average by a factor of 2. In doing so, we reduce trading costs. More important, if the spread moves around 0 back and forth, strategy II will be more profitable. For example, with the first trade the spread has moved above Γ, so we sell 1 Y and buy 1.36 X. When trading according to strategy I, we will flatten our position at 0 and have zero position while moving from 0 to $-\Gamma$ and not profit from this movement. When trading according to strategy II, we will still be short Y and long X while the spread moves to $-\Gamma$ (eg. X becomes more expensive relative to Y). This is shown is figures 2.4 and 2.5.

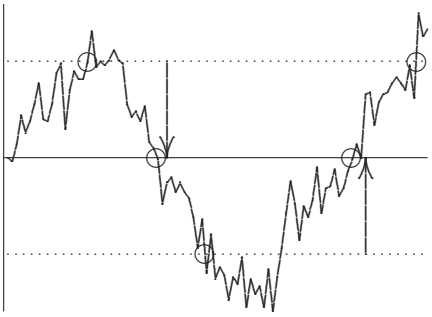

Figure 2.4: Trading strategy I.

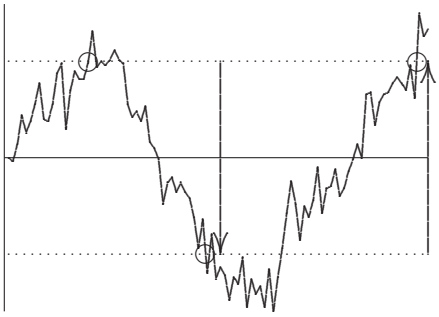

Figure 2.5: Trading strategy II.

Unfortunately, it involves a certain opportunity of loss as well. If a pair has a tendency to move between 0 and $+\Gamma$ or between 0 and $-\Gamma$, we might not be reversing our position at all, whereas strategy I will take on and flatten a position time and again and make money. This is shown in figures 2.6 and 2.7.

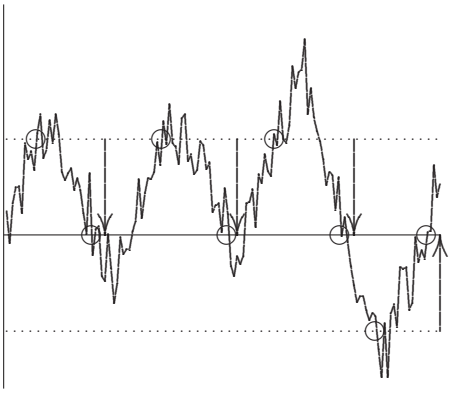

Figure 2.6: Trading strategy I.

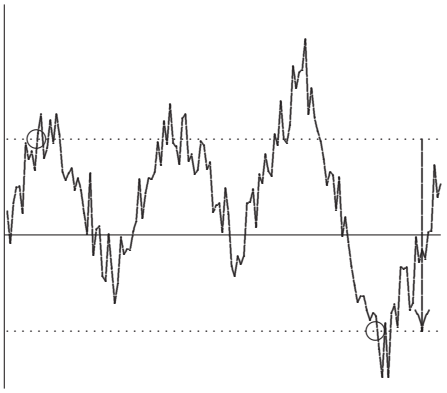

Figure 2.7: Trading strategy II.

In this report we will use a modified version of strategy II.

2.2 Data

The price data which IMC uses is provided by Bloomberg. Bloomberg is a leading global provider of data, news and analytic tools. Bloomberg provides real-time and archived financial and market data, pricing, trading, news and communications tools in a single, integrated package to corporations, news organizations, financial and legal professionals and individuals around the world.

Historical closing prices of stocks are easily extracted from Bloomberg to Excel. One issue has to be considered, namely dividend. Companies normally pay out dividend to its shareholders every year or twice a year, some companies pay out dividend four times a year. The amount of dividend is subtracted from the stock price at the day the dividend is paid out, called going ex-dividend. This usually results in a twist in the price process like in picture 2.8.

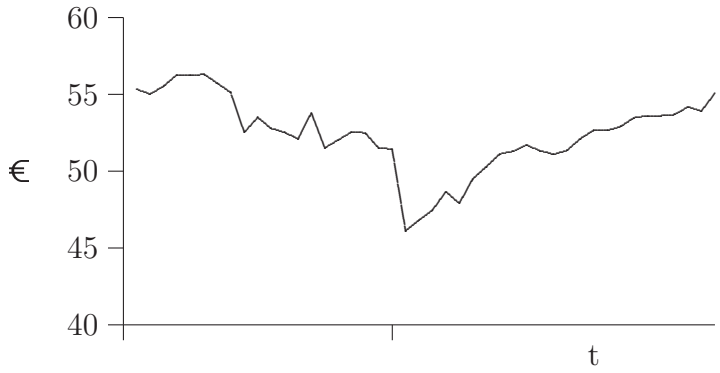

Figure 2.8: Dividend.

It is unlikely that different companies go ex-dividend at the same day. So the closing prices of stocks have to be corrected for dividend, to make a good comparison with other stocks. In this report we will assume that the dividend is re-invested in the stock. So it is not just adding the dividend up with the closing price, it is a growing amount proportionally to the growth of the stock price.

Example Consider the following the ex-dividend dates and amounts of a certain stock.

date	amount
04/28/2003	1.20
04/30/2004	1.40
04/29/2005	1.70

Suppose we want to use data of this stock starting from 03/01/2004. So we extract from Bloomberg the closing prices from this date forward, actually we start at 03/02/2004 because the first of March was a Sunday. From 03/02/2004 until 04/29/2004 we use exactly these prices, the first ex-dividend is not used. On 04/30/20004 the stock is ex-dividend for the amount of 1.40. We calculate what percentage this is of the stock price and from this date forward we keep multiplying the closing prices from Bloomberg with this percentage until the next ex-dividend date. Then we calculate the percentage of the dividend amount and adding it up to the percentage before, this is shown in table 2.3.

2.3 Properties of pairs trading

Pairs trading is almost cash neutral, we do not have to invest a lot of money. We use the earnings of short selling one stock to purchase the other stock. This usually does not exactly sum up to zero, to be precise it sums up to $\pm\Gamma$, a small positive or negative amount compared to the stock prices. An other aspect that makes pairs trading not entirely cash neutral is short selling. Short selling is selling something we do not have. The exchange on which we trade will want to be sure that we will not go bankrupt. We need to put money, called margin, aside to secure the exchange there are no risks involved with short selling. Normally, this margin is a percentage of the value of the short sale, typically between 5 and 50, depending on the credibility of the short seller. IMC's costs for short selling are relatively low, so pairs trading is almost cash neutral.

Table 2.3: Calculation of closing prices corrected for dividend.

date	Bloomberg	dividend	factor	our prices
03/02/2004	44.00	-	1	44.00
03/03/2004	43.37	-	1	43.37
⋮	⋮	⋮	⋮	⋮
04/29/2004	43.85	-	1	43.85
04/30/2004	43.04	1.40	1+1.40/43.04=1.03	1.03*43.04=44.33
05/01/2004	42.90	-	1.03	1.03*42.90=44.19
⋮	⋮	⋮	⋮	⋮
4/28/2005	51.44	-	1.03	1.03*51.44=52.98
4/29/2005	50.11	1.70	1.03+1.70/50.11=1.07	1.07*50.11=53.62
4/30/2005	50.64	-	1.07	1.07*50.64=54.18
⋮	⋮	⋮	⋮	⋮

Pairs trading is also market neutral: if the overall market goes up 10% it has no consequences for the strategy and profits of pairs trading. The 10% loss in the short stock is compensated by a 10% gain in the long stock, and the other way around if the overall market goes down. We do not have a preference for up or down movements, we only look at relative pricing.

How to make money with pairs trading was explained in the example in paragraph 2.1. The amount of money made by trading a pair is a measure for the quality of a pair. Obviously, more money is better! We make profits if the spread oscillates around zero often hitting Γ and $-\Gamma$. An important issue for the traders is that the spread should not be away from zero for a long time. Traders are humans and they tend to get a bit nervous if they have a big position for a long time. There is a chance that the spread will never return to zero and in that case it costs money to flatten the position.

Example Consider figure 2.9 of the spread of pair X, Y. We put on a position the first time the spread hits $-\Gamma$, because there Y is cheap relative to X in our opinion. We reverse our position at $+\Gamma$ and again at $-\Gamma$, making a profit of at least 4Γ. Then we like the spread to go to $+\Gamma$, but the spread is going further and further away from zero not knowing if it will ever come

back. At this time, our portfolio is worth less than when we put it on: the value of the long position in Y becomes less because Y is getting cheaper (relative to X) and the value of the short position in X is getting less because X is more expensive now (relative to Y). So, if we want to flatten our portfolio we have to sell Y for less than we bought it and/or buy X for more money than we sold it.

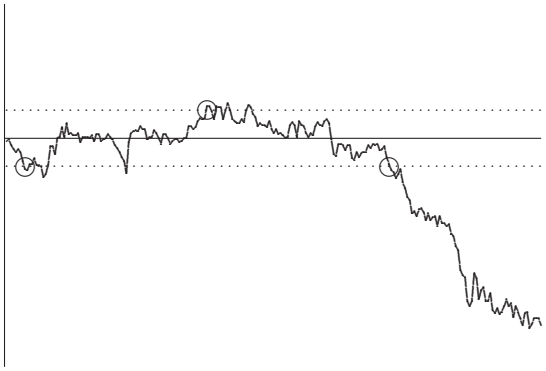

Figure 2.9: Spread s_t walks away.

In conclusion, a good pair has a spread that is rapidly mean-reverting and the price processes of the stocks in the pair are tied together, they can not get far away from each other.

2.4 Trading strategy

In this section we describe how the parameters in the introductory example (section 2.1) are determined. Then a few adjustments are made to strategy II, to get the final trading strategy that resembles the strategy from IMC. Finally, we give the assumptions made for applying this strategy.

Parameters Assume we have two datasets of closing prices of two different stocks X and Y for a certain period, roughly two years, which are corrected for dividend:

$$x_t \text{ and } y_t \ , \ \text{for } t = 0, ..., T.$$

The first half, $t = 0, ..., \lfloor T/2 \rfloor$, is considered as history and is used to determine the parameters ratio $\bar{r}$ and threshold Γ.

The second half, $t = \lfloor T/2 \rfloor + 1, ..., T$, is considered as the future and is used to determine the profit or loss that would be made trading the pair $\{X, Y\}$ with these parameters.

The ratio $\bar{r}$ is the average ratio of Y and X of the first half of observations:

$$\bar{r} = \frac{1}{\lfloor T/2 \rfloor + 1} \sum_{t=0}^{\lfloor T/2 \rfloor} \frac{x_t}{y_t}.$$

The threshold Γ is determined quite easily, we just try a few on the 'history' and take the one that gives the best profit based on the 'history'. We calculate the maximum of the absolute spread of the first half of observations, denoted as m:

$$m = \max_t \left(|y_t - \bar{r} x_t|, t = 0, ..., \lfloor T/2 \rfloor \right).$$

The values of Γ that we are going to try are percentages of m. Table 2.4 shows the percentages and the outcome for the introductory example of paragraph 2.1, where $m = 2.01$. Because of rounding to two digits it looks like there are several values of Γ which give the same largest profit, but $\Gamma = 0.40$ gives the largest profit.

The profit is calculated by multiplying number of trades minus one with two times Γ, except when no trades were made then the profit is just zero. It is the minimal profit if you always trade one spread, in this example one Y and 1.36 X. The first trading instance is to put on a position for the first time, denoted by t_1, then we do not make a profit yet:

$$t_1 = \min \left(t, \text{ such that } |s_t| \geq \Gamma \right).$$

The succeeding trading moments are:

If $s_{t_n} \geq \Gamma$:

$$t_{n+1} = \min \left(t, \text{ such that } t > t_n, s_t \leq -\Gamma \right).$$

If $s_{t_n} \leq -\Gamma$:

$$t_{n+1} = \min \left(t, \text{ such that } t > t_n, s_t \geq \Gamma \right).$$

Table 2.4: Profits with different Γ.

percentage	Γ	trades	profit
5	0.10	15	2.81
10	0.20	9	3.21
15	0.30	9	4.82
20	0.40	7	4.82
25	0.50	3	2.01
30	0.60	3	2.41
35	0.70	3	2.81
40	0.80	3	3.21
45	0.90	3	3.61
50	1.00	3	4.02
55	1.10	3	4.42
60	1.20	3	4.82
65	1.30	2	2.61
70	1.40	2	2.81
75	1.51	2	3.01
80	1.61	2	3.21
85	1.71	1	0
90	1.81	0	0

To determine Γ we simply take the one that has the largest profit based on the history, but in practice we do not take Γ larger than $0.5m$. This profit is a gross profit, no transaction costs are accounted. We neglected the transaction costs because it turned out they hardly had any influence on the value of Γ. This is because IMC does not trade one spread, which in this example was 1 Y and 1.36 X, but they trade a large number of Y and X, for example 1,000 Y and 1,360 X. The costs that IMC makes consists of two parts, a fixed amount a plus amount b times the number of traded stocks. The costs of trading 1,000 Y and 1,360 X would be $2a + 2,360b$. We always trade the same amount, no matter the value of Γ, so the costs per trade for all Γ are exactly the same. So the more trades the more costs, but the costs are really small compared to the profit. When the profits for the different thresholds are not too close to each other, the Γ when considering

the net profits is the same Γ when neglecting the costs. Unfortunately of all the pairs considered in this report, the pair from table 2.4 is the only one where accounting transaction costs would have made a difference. There are three thresholds, 0.30, 0.40 and 1.20, which result in almost the same profits. Therefor accounting the transaction costs would resulted in the threshold with the lowest number of trades, $\Gamma = 1.20$. In the remainder of this report, we will neglect transaction costs.

Modified trading strategy There are pairs of stock that work quite well for a certain time but then the spread walks away from zero and starts to oscillate around a level different from zero. We can see an example in figure 2.10. If we do not do anything, we are probably going to have a position for a long time which is not desirable as explained in paragraph 2.3. The figure shows us that the relation between the stocks in the pair has changed, the ratio $\bar{r}$, determined by the past, is not good anymore. It would be a waste to lose money on these kind of pairs by closing the position or to exclude them from trading. A better way is to replace the average ratio $\bar{r}$ with some kind of moving average ratio.

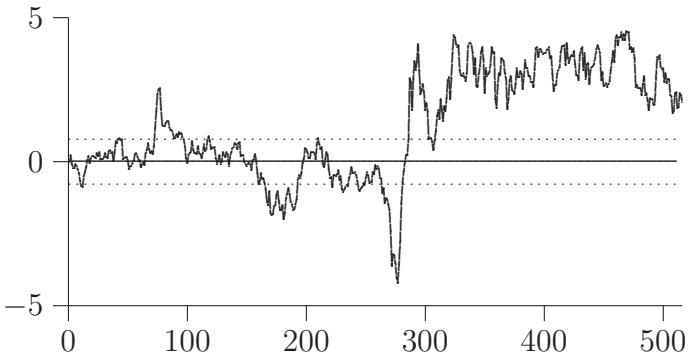

Figure 2.10: Spread oscillates around a new level.

Assume we have a dataset of closing prices, the first half is used in the exact same way as described before. So we have the average ratio $\bar{r}$ and threshold Γ. The backtest on the second half of the data set is slightly different because we use a moving average ratio $\tilde{r}_t$, instead of $\bar{r}$, to calculate the spread.

The moving average ratio we use, is:

$$\tilde{r}_t = (1-\kappa)\,\tilde{r}_{t-1} + \kappa\, r_t, \qquad t = \lfloor T/2 \rfloor + 1, ..., T,$$

with $\tilde{r}_{\lfloor T/2 \rfloor} = \bar{r}$ and where r_t is the actual ratio:

$$r_t = \frac{y_t}{x_t}, \qquad t = 0, ..., T.$$

The parameter κ is a percentage between 0 and 10% and is determined very simple with the first half of the data set. We count how many trades were made in the first half and use table 2.5 to find κ.

Table 2.5: Determining κ.

# trades	κ	# trades	κ
>15	0	4	6
10-15	1	3	7
8,9	2	2	8
7	3	1	9
6	4	0	10
5	5		

If there were a lot of trades in the first half of observations we do not expect to need a moving average ratio, the table motivates this. The use of a moving average ratio and this way of determining its value, has some disadvantages which will be discussed later on.

So the first half of the data set determines three parameters: Average ratio $\bar{r}$, threshold Γ and adjustment parameter κ. In the second half of the data set, the new spread is calculated as:

$$\tilde{s}_{\kappa,\,t} = y_t - \tilde{r}_t x_t.$$

Trading the pair goes in the same way as described before, the difference is the position in X is not equal to $\bar{r}$ anymore but it is equal to $\tilde{r}_t$. The following example will make this more clear.

We take the pair from figure 2.10, available are 520 closing prices of the two stocks. The first half of observations gives us three parameters:

$$\begin{aligned} \bar{r} &= 1.86, \\ \Gamma &= 0.77, \\ \kappa &= 5\%. \end{aligned}$$

First we look at what the strategy without the modification does on the second half of observations, table 2.6 shows the trading instances. Two trades are made with a total profit of €1.88. The strategy with the modification works better, 7 trades with a total profit of €5.21. Table 2.7 shows all trading instances. The table also shows that the position in stock X is not longer constant in absolute sense. For example, with trade number 1 we put on a position of +1 Y and -1.85 X because $\tilde{r}_t$ at this time is 1.85. With the second trade we flatten this position and put on a position the other way around, but now $\tilde{r}_t$ is 1.81 so in total we sell 2 shares of stock Y and buy 1.85+1.81=3.66 shares of stock X. The profit of these two trades is calculated with the position that is flattened, i.e., (51.81-48.70)+1.85*(26.80-28.06)=0.77.

Table 2.6: Trading instances strategy II.

trade	t	s_t	position (Y,X)	price y	price X	profit
1	263	-1.10	(+1,-1.86)	48.70	26.80	-
2	285	0.78	(-1,+1.86)	52.33	27.74	1.88
				total profit		1.88

Table 2.7 also shows that not all profits per trade are larger than Γ, one trade gave a relatively large loss. This happens because the ratio when the position was put on, differs a lot from the ratio when this position is reversed. The ratios differ a lot because the actual ratio r_t is moving a lot. We can see all the ratios in figure 2.11. The solid line is the actual ratio r_t, the dashed line is the moving average ratio $\tilde{r}_t$ and the straight dotted line is the average ratio $\bar{r}$.

Table 2.7: Trading instances modified strategy.

trade	t	$\widetilde{s}_{\kappa,\,t}$	position (Y,X)	price Y	price X	$\widetilde{r}_t$	profit
1	263	-0.99	(+1,-1.85)	48.70	26.80	1.85	-
2	281	1.07	(-1,+1.81)	51.81	28.06	1.81	0.77
3	358	-0.82	(+1,-1.97)	51.52	26.56	1.97	-2.43
4	392	0.93	(-1,+1.96)	56.38	28.23	1.96	1.57
5	407	-0.94	(+1,-1.98)	55.45	28.52	1.98	1.52
6	459	0.97	(-1,+1.98)	55.27	27.47	1.98	1.92
7	476	-1.31	(+1,-1.99)	57.20	29.38	1.99	1.86
					total profit		5.21

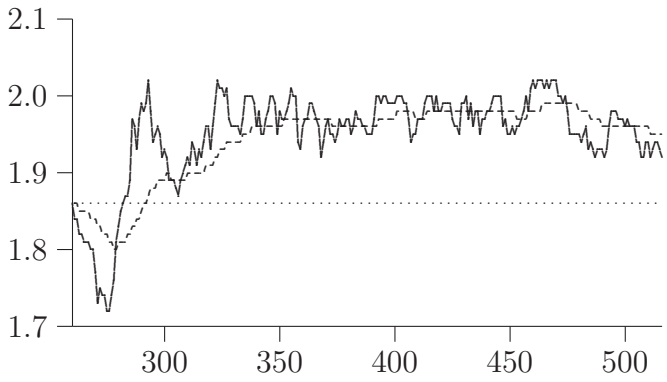

Figure 2.11: Ratios r_t, $\widetilde{r}_t$ and $\bar{r}$.

Figures 2.12 and 2.13 show the spread calculated with the average ratio $\bar{r}$ and calculated with the moving average ratio $\tilde{r}_t$ with $\kappa = 5\%$ respectively.

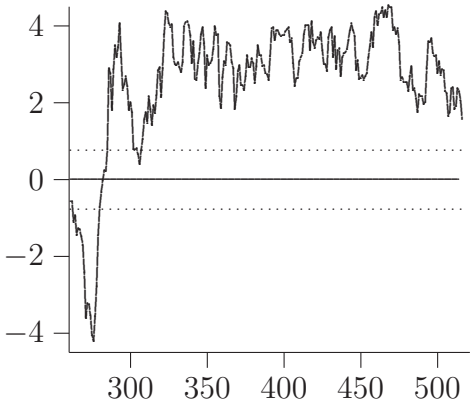

Figure 2.12: Spread s_t.

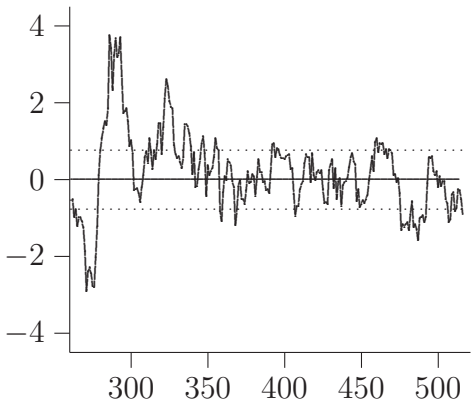

Figure 2.13: Spread $\tilde{s}_{\kappa,\,t}$, $\kappa = 5\%$.

From figure 2.10 it is clear that the average ratio $\bar{r}$ does not fit anymore, around $t = 300$ the stocks in the pair get another relation. Replacing the fixed average ratio $\bar{r}$ by an moving average ratio $\tilde{r}_t$ resolves this. As we saw in the example we can lose money if the moving average ratio, used to calculate the spread, differs a lot between trades. If there is some fundamental change, such a trade will happen once or twice and the loss that is made will be compensated by good trades from that moment on. The advantage of the modified trading strategy is when the relation between stocks in a pair changes in some fundamental way as in the example above i.e., the spread is oscillating around a new level, we are still able to trade the pair with a profit instead of making a loss by closing the position and exclude the pair from trading.

When there is no such fundamental change but we use the modified strategy, with $\kappa > 0$, it is possible we throw away money with each trade. This happens if the moving average ratio differs a lot between each two succeeding trades. We consider an example, suppose we have 520 observations.

The first half is used to determine the three parameters $\bar{r}$, Γ and κ:

$$\begin{aligned} \bar{r} &= 1.00, \\ \Gamma &= 0.62, \\ \kappa &= 7. \end{aligned}$$

Figures 2.14 and 2.15 show the spread for the second half of observations calculated with the average ratio $\bar{r}$, which is the same as $\kappa = 0$, and $\kappa = 7$ respectively.

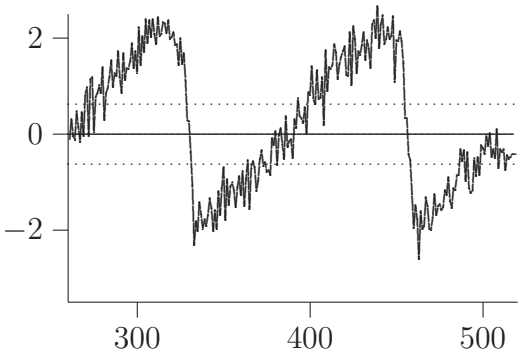

Figure 2.14: Spread $\tilde{s}_{\kappa,\,t}$, $\kappa = 0$.

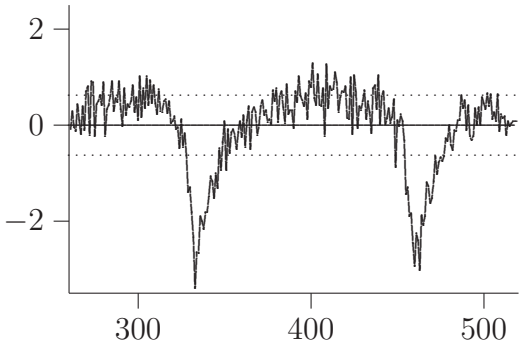

Figure 2.15: Spread $\tilde{s}_{\kappa,\,t}$, $\kappa = 7$.

Trading the spread with $\kappa = 0$ results in four trades with a total profit of €5.69. However trading the spread with $\kappa = 7$ results in five trades with a total *loss* of €4.03, table 2.8 shows the corresponding trading instances.

In this example there is a loss with every trade if we use $\kappa = 7$, but we make a substantial profit when we use $\kappa = 0$. This is a bit of an extreme example but what is often seen is that when there is no fundamental change between the stocks in the pair, the profit is less when using the modified strategy ($\kappa > 0$) then the original strategy ($\kappa = 0$). This is a big disadvantage of the modified strategy, it is at least very difficult to determine if the relation between the stocks is fundamentally changing. In spite of this disadvantage we use the modified strategy because we do not want to exclude pairs like in figure 2.10, we are willing to give up some profit on pairs who do not change much.

Table 2.8: Trading instances modified strategy.

trade	t	$\widetilde{s}_{\kappa,\ t}$	position (Y,X)	price Y	price X	$\widetilde{r}_t$	profit
1	270	0.71	(-1,+1.01)	10.72	9.94	1.01	-
2	328	-0.63	(+1,-1.18)	10.80	9.72	1.18	-0.30
3	378	0.72	(-1,+0.92)	9.93	10.50	0.92	-1.25
4	449	-0.87	(+1,-1.18)	11.59	10.55	1.18	-1.21
5	487	0.63	(-1,+0.90)	9.54	9.88	0.90	-1.27
					total profit		-4.03

Assumptions We apply the trading strategy to historical closing data to see if trading a pair of two stocks would have been profitable. This assumes that we could have traded on the closing price and that there was no bid-ask spread. It also assumes we could have traded every amount we wanted, including fractions. If it is decided to start trading a specific pair, it is going to be traded intra-day, so it would probably be better to apply the trading strategy to intra-day data but that kind of data is difficult to get and is difficult to handle. With real life trading the number of stocks have to be integers. The assumption that we are allowed to trade fractions is not that bad because when trading a pair it is about large quantities so we can round the number of stocks to an integer without completely messing up the ratios.

2.5 Conclusion

In this chapter we have derived a trading strategy that resembles the strategy IMC uses. It is not necessary anymore to do a fundamental analysis to find out if a pair of two stocks is profitable to trade as a pair. We can apply the trading strategy on historical data and see if we would have made a profit if we actually traded the pair. In this way IMC identified a lot of pairs. We would like to see if we can identify pairs in a more statistical setting, again using historical data of two stocks, not to estimate profits, but to see if the two time series exhibit behavior that could make them a good pair. We will examine the concept of cointegration, but first we need some time series basics.

Chapter 3

Time series basics

This chapter discusses briefly some basics of time series which we will need for later purposes. More information can be found in [2] and [3].

White noise A basic stochastic time series $\{z_t\}$ is *independent white noise*, if z_t is an independent and identically distributed (i.i.d.) variable with mean 0 and variance σ^2 for all t, notation $z_t \sim \text{i.i.d}(0, \sigma^2)$. A special case is *Gaussian white noise*, where each u_t is independent and has a normal distribution $N(0, \sigma^2)$.

Stationarity A time series $\{z_t\}$ is *covariance-stationary* or *weakly stationary* if neither the expectation nor the autocovariances depend on time t:

$$\begin{aligned} E(z_t) &= \mu, \\ E(z_t - \mu)(z_{t-j} - \mu) &= \gamma_j, \end{aligned}$$

for all t and j. Notice that if a process is covariance-stationary, the variance of z_t is constant and the covariance between z_t and z_{t-j} depends only on lag j. For example, a white noise process is covariance-stationary. Covariance-stationary is shortened by stationary in the remaining of this report.

A stationary process exhibits mean reverting behavior, the process tends to remain near or tends to return over time to the mean value.

MA(q) A *q-th order moving average process*, denoted MA(q), is characterized by:

$$z_t = \mu + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \theta_q u_{t-q}, \tag{3.1}$$

where $\{u_t\}$ is white noise ($\sim$ i.i.d$(0, \sigma^2)$), μ and $(\theta_1, \theta_2, \ldots, \theta_q)$ are constants. The expectation, variance and autocovariances of z_t are given by:

$$\begin{aligned} E(z_t) &= \mu, \\ \gamma_0 &= (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)\sigma^2, \\ \gamma_j &= \begin{cases} (\theta_j + \theta_{j+1}\theta_1 + \theta_{j+1}\theta_2 + \cdots + \theta_q\theta_{q-j})\sigma^2 & \text{if } j = 1, \ldots, q, \\ 0 & \text{if } j > q. \end{cases} \end{aligned}$$

So an MA(q) process is stationary.

AR(1) A *first-order autoregressive process*, denoted AR(1), satisfies the following difference equation:

$$z_t = c + \phi z_{t-1} + u_t, \tag{3.2}$$

where $\{u_t\}$ is independent white noise ($\sim$ i.i.d$(0, \sigma^2)$). If $|\phi| \geq 1$, the consequences of the u's for z accumulate rather than die out over time. Perhaps it is not surprising that when $|\phi| \geq 1$, there does not exist a causal stationary process for z_t with finite variance that satisfies (3.2). If $|\phi| > 1$ the process z_t can be written in terms of innovation in the future instead of innovations in the past, that is what is meant by 'there does not exist a causal stationary process'. If $\phi = 1$ and $c = 0$ the process is called a *random walk*. When $|\phi| < 1$, the AR(1) model defines a stationary process and has an MA(∞) representation:

$$z_t = c/(1 - \phi) + u_t + \phi u_{t-1} + \phi^2 u_{t-2} + \phi^3 u_{t-3} + \cdots.$$

The expectation, variance and autocovariances of z_t are given by:

$$\begin{aligned} \mu &= c/(1 - \phi), \\ \gamma_0 &= \sigma^2/(1 - \phi^2), \\ \gamma_j &= (\sigma^2\phi^j/(1 - \phi^2)), \qquad \text{for } j = 1, 2, \ldots \end{aligned}$$

AR(p) A *p-th order autoregressive process*, denoted AR(p), satisfies:

$$z_t = c + \phi_1 z_{t-1} + \phi_2 z_{t-2} + \cdots + \phi_p z_{t-p} + u_t. \tag{3.3}$$

Suppose that the roots of

$$1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p = 0, \tag{3.4}$$

all lie outside the unit circle in the complex plain. This is the generalization of the stationarity condition $|\phi| < 1$ for the AR(1) model. Then the expectation, variance and autocovariances of z_t are given by:

$$\begin{aligned} \mu &= c/(1 - \phi_1 - \phi_2 - \cdots - \phi_p), \\ \gamma_0 &= \phi_1\gamma_1 + \phi_2\gamma_2 + \cdots + \phi_p\gamma_p + \sigma^2, \\ \gamma_j &= \phi_1\gamma_{j-1} + \phi_2\gamma_{j-2} + \cdots + \phi_p\gamma_{j-p}\ , \qquad \text{for } j = 1, 2, \ldots \end{aligned}$$

If equation (3.4) has a root that is on the unit circle, we call that a unit root and the process that generates z_t a unit root process.

Information Criteria In chapter 4 we want to fit an AR(p) model on a given dataset, with p unknown. An *information criterion* is designed to maximize the model fit while minimizing the number of parameters, in our case minimizing p. The criterion assigns a value to each model depending on the model fit and the number of parameters in the model. The better the model fit is, the smaller the value will be. The more parameters are used, the larger the value will be. The model with the smallest value is most suitable for the data according to that criterion. There are several information criteria, they differ in the penalty they give to each extra parameter and therefore have different properties.

The *Akaike information criterion* (AIC) formula is:

$$\mathrm{AIC}(k) = -2\log L + 2k, \tag{3.5}$$

where k is the number of parameters, and L is the likelihood function. The likelihood function assumes that the innovations u_t are $N(0, \sigma^2)$.

The log likelihood for an AR(k) model is given by:

$$\begin{aligned}\log L = & -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) + \frac{1}{2}\log|\mathbf{V}_k^{-1}| \\ & -\frac{1}{2\sigma^2}(\mathbf{z}_k - \boldsymbol{\mu}_k)'\mathbf{V}_k^{-1}(\mathbf{z}_k - \boldsymbol{\mu}_k) \\ & -\sum_{t=k+1}^{T}\frac{(z_t - c - \phi_1 z_{t-1} - \cdots - \phi_k z_{t-k})^2}{2\sigma^2}\end{aligned}$$

where $\sigma^2\mathbf{V}_k$ denotes the covariance matrix of $(z_1, z_2, \ldots, z_k)$:

$$\sigma^2\mathbf{V}_k = \begin{bmatrix} E(z_1-\mu)^2 & E(z_1-\mu)(z_2-\mu) & \cdots & E(z_1-\mu)(z_k-\mu) \\ E(z_2-\mu)(z_1-\mu) & E(z_2-\mu)^2 & \cdots & E(z_2-\mu)(z_k-\mu) \\ \vdots & \vdots & \cdots & \vdots \\ E(z_k-\mu)(z_1-\mu) & E(z_p-\mu)(z_2-\mu) & \cdots & E(z_k-\mu)^2 \end{bmatrix}$$

and $\boldsymbol{\mu}_k$ denotes a $(k \times 1)$ vector with each element given by

$$\mu = c/(1 - \phi_1 - \phi_2 - \cdots - \phi_k),$$

and $\mathbf{z}_k$ denotes the first k observations in the sample, $(z_1, z_2, \ldots, z_k)$ and T denotes the sample size.

The first term in (3.5) measures the model fit, the second term gives a penalty to each parameter. The Akaike information criterion is calculated for each model AR(k), with $k = 1, 2, \ldots, K$. The k with the smallest value AIC(k), is the estimate for the model order.

Two other information criteria are the Schwarz-Baysian and the Hannan-Quint information criteria.

The *Schwarz-Baysian information criterion* (BIC) formula is:

$$\text{BIC}(k) = -2\log L + k\log(T),$$

where T denotes the number of observations in the data set.

The *Hannan-Quint information criterion* (HIC) formula is:

$$\text{HIC}(k) = -2\log L + 2k\log(\log(T)).$$

First difference operator The first difference operator Δ is defined by:

$$\Delta z_t = z_t - z_{t-1}.$$

I(d) A time series is *integrated of order d*, written as $y_t \sim I(d)$, if the series is non-stationary but it becomes stationary after differencing a minimum of d times. An already weakly stationary process is denoted as $I(0)$. If a time series generated by an AR(p) process is integrated of order d, than its autoregressive polynomial (equation (3.4)) has d roots on the unit circle.

Unit root test Statistical tests of the null hypothesis that a time series is non-stationary against the alternative that it is stationary are called *unit root tests*. In this paper we consider the *Dickey-Fuller test* (DF) and the *Augmented Dickey-Fuller test* (ADF).

Dickey-Fuller test The *Dickey-Fuller test* tests whether a time series is stationary or not when the series is assumed to follow an AR(1) model. It is named after the statisticians D.A. Dickey and W.A. Fuller, who developed the test in [4].

The assumption of the DF test is that the time series z_t follows an AR(1) model:

$$z_t = c + \rho z_{t-1} + u_t, \tag{3.6}$$

with $\rho \geq 0$. If $\rho = 1$, the series z_t is non-stationary. If $\rho < 1$, the series z_t is stationary. The null hypothesis is that z_t is non-stationary, more specific z_t is integrated of order 1, against the alternative z_t is stationary:

$$H_0 : z_t \sim I(1) \text{ against } H_1 : z_t \sim I(0),$$

which can be restated in terms of the parameters:

$$H_0 : \rho = 1 \text{ against } H_1 : \rho < 1,$$

under the assumption that z_t follows an AR(1) model.

The test statistic of the DF test S is the t ratio:

$$S = \frac{\hat{\rho} - 1}{\hat{\sigma}_{\hat{\rho}}} \ ,$$

where $\hat{\rho}$ denotes the OLS estimate of ρ and $\hat{\sigma}_{\hat{\rho}}$ denotes the standard error for the estimated coefficient.

The t ratio is commonly used to test whether the coefficient ρ is equal to ρ_0 when the time series is stationary, i.e. $\rho < 1$. Then the test statistic

$$\frac{\hat{\rho} - \rho_0}{\hat{\sigma}_{\hat{\rho}}} \ ,$$

has a t-distribution. But we do not assume that the time series is stationary, because the null hypothesis is that $\rho = 1$. So, the test statistic S does not need to have a t-distribution. We need to distinguish several cases to derive the distribution of the DF test statistic.

Case 1:
The true process of z_t is a random walk, i.e. $z_t = z_{t-1} + u_t$, and we estimate the model $z_t = \rho z_{t-1} + u_t$. Notice that we only estimate ρ and not a constant c.

Case 2:
The true process of z_t is again a random walk and we estimate the model $z_t = c + \rho z_{t-1} + u_t$. Notice that now we do estimate a constant but it is not present in the true process.

Case 3:
The true process of z_t is a random walk, but now with drift, i.e. $z_t = c + z_{t-1} + u_t$, where the true value of c is not zero. We estimate the model $z_t = c + \rho z_{t-1} + u_t$.

Although the differences between the three cases seem small, the effect on the asymptotic distributions of the test statistic are large, as we will see in chapter 5.

Augmented Dickey-Fuller test The *Augmented Dickey-Fuller test* tests whether a time series is stationary or not when the time series follows an AR(p) model. One of the assumptions of the Augmented Dickey-Fuller test is that the time series z_t follows an AR(p) model:

$$z_t = c + \phi_1 z_{t-1} + \cdots + \phi_p z_{t-p} + u_t. \tag{3.7}$$

Like the regular Dickey-Fuller test, we test:

$$H_0 : z_t \sim I(1) \text{ against } H_1 : z_t \sim I(0).$$

The null hypothesis is that the autoregressive polynomial

$$1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p = 0,$$

has exactly one unit root and all other roots are outside the unit circle. Then the unit root cannot be a complex number, because the autoregressive polynomial is a polynomial with real coefficients and if $x = a + bi$ is a unit root than so is its complex conjugate $\bar{x} = a - bi$. This contradicts the null hypothesis that there is exactly one unit root. Two possibilities remain, the unit root is -1 or 1. The first possibility gives an alternating series, which is not realistic for modeling the spread (this becomes more clear in the chapter of cointegration). Thus the single unit root should be equal to 1, which gives us

$$1 - \phi_1 - \phi_2 - \cdots - \phi_p = 0. \tag{3.8}$$

The AR(p) model (3.7) can be written as:

$$z_t = c + \rho z_{t-1} + \beta_1 \Delta z_{t-1} + \cdots + \beta_{p-1} \Delta z_{t-p+1} + u_t, \tag{3.9}$$

with

$$\begin{aligned} \rho &= \phi_1 + \cdots + \phi_p, \\ \beta_i &= -(\phi_{i+1} + \cdots + \phi_p), \quad \text{for } i = 1, \ldots, p-1. \end{aligned}$$

The advantage of writing (3.7) in the equivalent form (3.9) is that under the null hypothesis only one of the regressors, namely z_{t-1}, is $I(1)$, whereas all of the other regressors $(\Delta z_{t-1}, \Delta z_{t-2}, \ldots, \Delta z_{t-p+1})$ are stationary. Notice

that (3.8) implies that coefficient ρ is equal to 1. This leads to the same hypotheses as with the regular Dickey-Fuller test:

$$H_0 : \rho = 1 \text{ against } H_1 : \rho < 1,$$

and the same test statistic:

$$S = \frac{\hat{\rho} - 1}{\hat{\sigma}_{\hat{\rho}}} .$$

To derive the distribution of the ADF test statistic we need to distinguish the same three cases as above, but now in the appropriate AR(p) form. As we will see in chapter 5, the distributions are the same as DF distributions without any corrections for the fact that lagged values of Δy are included in the regression.

One last note: If the null hypothesis that z_t is non-stationary cannot be rejected, it does not necessarily mean that z_t is generated by a $I(1)$ process. It may be non-stationary because it is generated by a $I(2)$ process or by an integrated process of an even higher order. The next step could be to repeat the procedure but this time using Δy_t instead of y_t. That is, to test $H_0 : \Delta y_t \sim I(1)$ against $H_1 : \Delta y_t \sim I(0)$ which is equivalent to $H_0 : y_t \sim I(2)$ against $H_1 : y_t \sim I(1)$, and so on.

Chapter 4

Cointegration

Empirical research in financial economics is largely based on time series. Ever since Trygve Haavelmos work it has been standard to view economic and financial time series as realizations of stochastic processes. This approach allows the model builder to use statistical inference in constructing and testing equations that characterize relationships between economic and financial variables. The Nobel Prize of 2003 for economics has rewarded two contributions, the ARCH model and cointegration from Robert Engle and Clive Granger.

This chapter discusses the concept of cointegration and two methods for testing for cointegration, the Engle-Granger and the Johansen method. Other methods are described in, for example, [13] and [14]. In the last section of this chapter a start is made with an alternative method. In this report this alternative method is used for generating cointegrated data but not for testing for cointegration, although this is possible.

4.1 Introducing cointegration

An $(n \times 1)$ vector time series $\mathbf{y}_t$ is said to be *cointegrated* if each of the series taken individually is $I(1)$, integrated of order one, while some linear combination of the series $\mathbf{a}'\mathbf{y_t}$ is stationary for some nonzero $(n \times 1)$ vector $\mathbf{a}$, named the cointegrating vector.

Cointegration means that although many developments can cause permanent changes in the individual elements of $\mathbf{y}_t$, there is some long-run equilibrium relation tying the individual components together, represented by the linear combination $\mathbf{a}'\mathbf{y_t}$.

A simple example of a cointegrated vector process with $n = 2$, which was taken from [1], is:

$$\begin{aligned} x_t &= w_t + \epsilon_{x,t} \ , \\ y_t &= w_t + \epsilon_{y,t} \ , \\ w_t &= w_{t-1} + \epsilon_t \ , \end{aligned}$$

where error processes $\epsilon_{x,t}$, $\epsilon_{y,t}$ and ϵ_t are independent white noise processes. The series w_t is a random walk, so x_t and y_t are $I(1)$ processes, though the linear combination $y_t - x_t$ is stationary. This means $\mathbf{y}_t = (x_t, y_t)$ is cointegrated with $\mathbf{a} = (-1, 1)$.

Figure 4.1 shows a realization of this example of a cointegrated process, where the error processes are standard Gaussian white noise. Note that x_t and y_t can wander arbitrarily far from the starting value, but x_t and y_t themselves are 'tied together' in the long run. The figure also shows the corresponding spread $y_t - x_t$ of the realization.

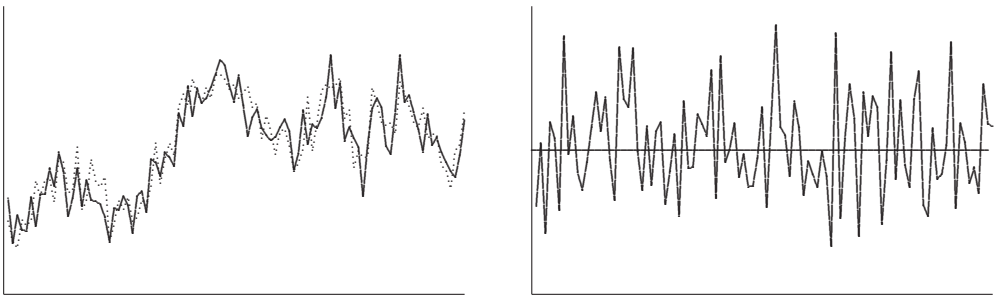

Figure 4.1: Realization of cointegrated process and spread of realization.

Correlation Correlation is used in analysis of co-movements in assets but also in analysis of co-movements in returns. Correlation measures the strength and direction of linear relationships between variables. If x_t denotes a price process of a stock, the returns h_t are defined by

$$h_t = \frac{x_t - x_{t-1}}{x_{t-1}} \ ,$$

with $\log(1+\epsilon) \approx \epsilon$ as $\epsilon \to 0$, we can approximate this by:

$$\frac{x_t - x_{t-1}}{x_{t-1}} = \frac{x_t}{x_{t-1}} - 1 \approx \log\left(\frac{x_t}{x_{t-1}}\right).$$

Correlation can refer to co-movement in the stock returns and in the stock prices themselves, cointegration refers to co-movements in the stock prices themselves or the logarithm of the stock prices. Cointegration and correlation are related, but they are different concepts. High correlation does not imply cointegration, and neither does cointegration imply high correlation. In fact, cointegrated series can have correlations that are quite low at times. For example, a large and diversified portfolio of stocks which are also in an equity index, where the weights in the portfolio are determined by their weights in the index, should be cointegrated with the index itself. Although the portfolio should move in line with the index in the long term, there will be periods when stocks in the index that are not in the portfolio have exceptional price movements. Following this, the empirical correlations between the portfolio and the index may be rather low for a time.

The simple example at the beginning of this section shows the same, that is, cointegration does not imply high correlation. For illustration purposes it is convenient to look at the differences, Δx_t and Δy_t, instead of the returns or x_t and y_t themselves because in this example they do not have constant variances. The variance of Δx_t is

$$\begin{aligned} \mathrm{Var}(\Delta x_t) &= \mathrm{Var}(x_t - x_{t-1}) \\ &= \mathrm{Var}(\epsilon_t + \epsilon_{x,\,t} + \epsilon_{x,\,t-1}) \\ &= \sigma^2 + 2\sigma_x^2, \end{aligned}$$

where σ^2, σ_x^2, and σ_y^2 denote the variances of ϵ_t, $\epsilon_{x,\,t}$ and $\epsilon_{y,\,t}$ respectively.

In the same way, $\text{Var}(\Delta y_t) = \sigma^2 + 2\sigma_y^2$. The covariance of Δx_t and Δy_t is given by

$$\begin{aligned} \text{Cov}(\Delta x_t, \Delta y_t) &= E(\Delta x_t \Delta y_t) - E(\Delta x_t)E(\Delta y_t) \\ &= E(\epsilon_t^2) - 0 \\ &= \sigma^2. \end{aligned}$$

The correlation between the difference processes is

$$\begin{aligned} \text{Corr}(\Delta x_t, \Delta y_t) &= \frac{\text{Cov}(\Delta x_t, \Delta y_t)}{\sqrt{\text{Var}(\Delta x_t)\text{Var}(\Delta y_t)}} \\ &= \frac{\sigma^2}{\sqrt{(\sigma^2 + 2\sigma_x^2)(\sigma^2 + 2\sigma_y^2)}}. \end{aligned}$$

The correlation between Δx_t and Δy_t is going to be less than 1, and when the variances of ϵ_{xt} and/or ϵ_{yt} are much larger than the variance of ϵ_t the correlation will be low while x_t and y_t are cointegrated.

The converse also holds true: there may be high correlation between the stock prices and/or the returns without the stock prices being cointegrated. Figure 4.2 shows two stock price processes which are highly correlated, namely 0.9957. The correlation between the returns is even equal to 1. But the price processes are clearly not cointegrated, they are not tied together, instead they are diverging more and more as time goes on. So, correlation does not tells us enough about the long-term relationship between two stocks: they may or may not be moving together over long periods of time, i.e. they may or may not be cointegrated.

Looking from a trading point of view, the 'pair' in figure 4.2 is not a good one. Figures 4.3 and 4.4 show the spread calculated with the average ratio $\overline{r}$ and calculated with a 10% moving average ratio $\widetilde{r}_t$ respectively. In figure 4.3 it is clear that this 'pair' is not a good one, because the spread is not oscillating around zero. Figure 4.4 looks better, but actually we are loosing money with nearly every trade because the ratios when positions were put on differ a lot from the ratios when the positions were reversed. The ratios differ a lot because the actual ratio r_t is moving a lot, which is due to the divergence between the stock prices. So, correlation is not a good way to identify pairs.

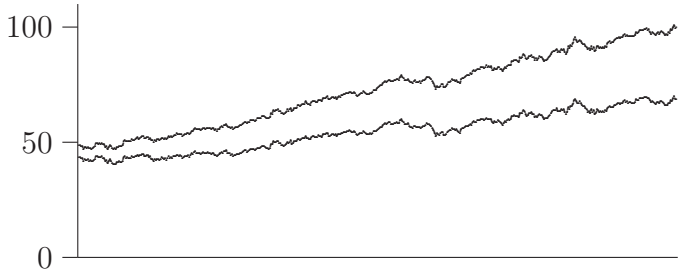

Figure 4.2: Highly correlated stock prices.

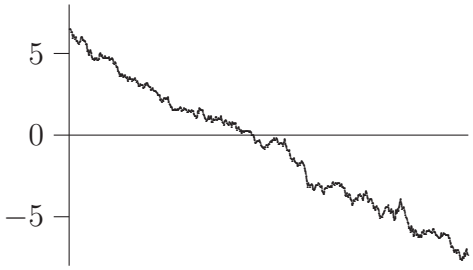

Figure 4.3: Spread s_t, $\overline{r} = 0.76$.

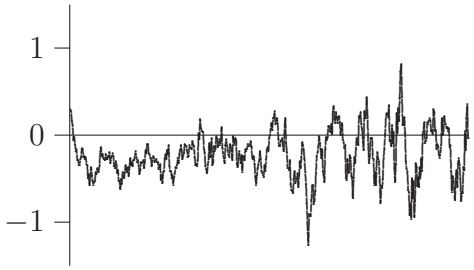

Figure 4.4: Spread $\widetilde{s}_{\alpha,t}$, $\alpha = 10\%$.

A better way to identify pairs is with cointegration, because we would like the stock prices in a pair to be tied together. If two stocks in a pair are cointegrated, a certain linear combination of the two is stationary. This implies that the spread, defined with the cointegrating vector $\mathbf{a}$ instead of the average ratio $\bar{r}$ or moving average ratio $\tilde{r}_t$, is mean-reverting. In paragraph 2.3 was explained that this property is an important one.

4.2 Stock price model

In the preceding section cointegration was introduced, the question remains how to test for cointegration. The test should be preceded by examining if each component of $\mathbf{y}_t$ is $I(1)$ because that is a requirement in the definition of cointegration. Several books and articles are written about modeling stock prices. In this section we will derive a commonly used model which can be

found in, among others, [7]. This model is famous for the use in option valuation. We will use this model to show that the logarithm of stock prices are integrated of order one and to show that it is more or less justified to assume stock prices themselves are integrated of order one.

In figure 4.5 the daily closing prices of Royal Dutch Shell are plotted. The figure shows the jagged behavior that is common to stock prices.

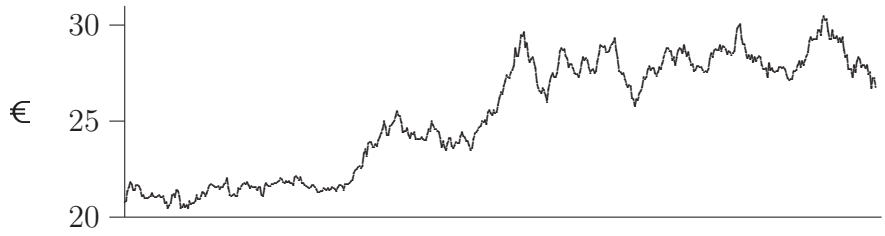

Figure 4.5: Daily Royal Dutch Shell stock prices.

We first examine the returns of the Royal Dutch Shell stock. Figure 4.6 shows the estimated density of the daily returns with the $N(0,1)$ density superimposed, figure 4.7 the empirical distribution function and figure 4.8 the normal QQ-plot. The daily returns were normalized to

$$\hat{h}_t = \frac{h_t - \hat{\mu}}{\hat{\sigma}^2} \ ,$$

where $\hat{\mu}$ and $\hat{\sigma}^2$ are the sample mean and sample variance. These figures suggest that the marginal distribution of daily returns of the Royal Dutch Shell stock is Gaussian. The QQ-plot indicates that the match is least accurate at the extremes of the range, the returns have fatter tails than the normal distribution. Figure 4.9 shows the sample autocorrelation function of the daily returns. The bounds $\pm 1.96 T^{-1/2}$ are displayed by the dashes lines, here $T = 520$. The figure strongly suggests that the returns are uncorrelated. Although uncorrelated does not implies independence, we suggest that for modeling x_t we take the returns as normally distributed i.i.d. samples, because the autocorrelation function for a sample from an i.i.d. noise sequence looks similar as figure 4.9.

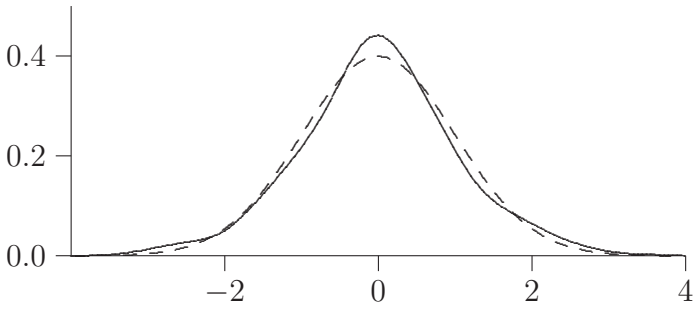

Figure 4.6: Estimated density.

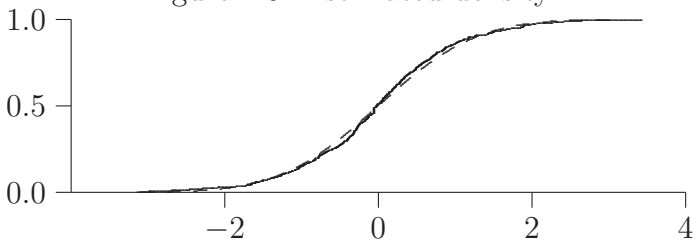

Figure 4.7: Empirical distribution function.

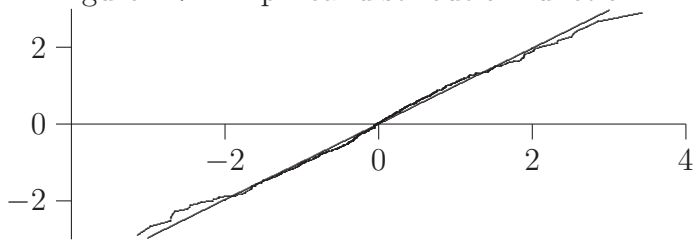

Figure 4.8: Normal QQ-plot.

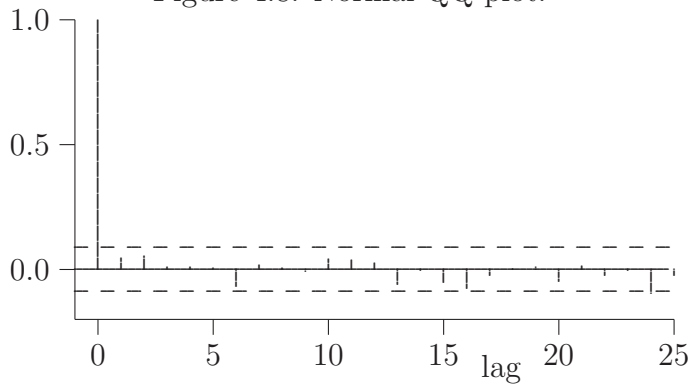

Figure 4.9: Autocorrelation function.

Given the stock price $x(0) = x_0$ at time $t = 0$, we like to come up with a process that describes the stock price $x(t)$ for all times $0 \leq t \leq T$. As a starting point for the model we note that the value of a risk-free investment D, like putting money on a savings account, changes over a small time interval δ as

$$D(t + \delta) = D(t) + \mu\delta D(t),$$

where μ is the interest rate.

There is something that is called the *efficient market hypothesis* that says that the current stock price reflects all the information known to investors, and so any change in the price is due to new information. We may build this into our model by adding a random fluctuation to the interest rate equation. Let $t = i\delta$, the discrete-time model becomes:

$$x(t_i) = x(t_{i-1}) + \mu\,\delta\,x(t_{i-1}) + \sigma\sqrt{\delta}\,u_i\,x(t_{i-1})\,, \tag{4.1}$$

where the parameter $\mu > 0$ represents an annual upward drift of the stock prices. The parameter $\sigma \geq 0$ is a constant that determines the strength of the random fluctuations and is called the *volatility*. The random fluctuations $u_1, u_2, \ldots$ are i.i.d $N(0, 1)$. Notice that the returns $[x(t_i) - x(t_{i-1})]/x(t_{i-1})$ indeed form a normal i.i.d sequence.

We consider the time interval $[0, t]$ with $t = L\delta$. Assume we know $x(0) = x_0$, the discrete model (4.1) gives us expressions for $x(\delta), x(2\delta), \ldots, x(t)$. To derive a continuous model for the stock price, we let $\delta \to 0$ to get a limiting expression for $x(t)$.

The discrete model says that over each time δ the stock price gets multiplied by a factor $1 + \mu + \sigma\sqrt{\delta}u_i$, hence

$$x(t) = x_0 \prod_{i=1}^{L} (1 + \mu\delta + \sigma\sqrt{\delta}u_i).$$

Dividing by x_0 and taking logarithms gives

$$\log\left(\frac{x(t)}{x_0}\right) = \sum_{i=1}^{L} \log(1 + \mu\delta + \sigma\sqrt{\delta}u_i).$$

We are interested in the limit $\delta \to 0$, we exploit the approximation $\log(1+\epsilon) \approx \epsilon - \epsilon^2/2 + \cdots$ for small ϵ.

$$\log\left(\frac{x(t)}{x_0}\right) \approx \sum_{i=0}^{L-1} \mu\delta + \sigma\sqrt{\delta}u_i - \frac{1}{2}\sigma^2\delta u_i^2.$$

This is justifiable because $E(u_i^2)$ is finite. We have ignored terms that involve the power of $\delta^{3/2}$ or higher.

The expectation and the variance are:

$$E(\mu\delta + \sigma\sqrt{\delta} - \frac{1}{2}\sigma^2\delta u_i^2) = \mu\delta - \frac{1}{2}\sigma^2\delta \ ,$$
$$\text{var}(\mu\delta + \sigma\sqrt{\delta}u_i - \frac{1}{2}\sigma^2\delta u_i^2) = \sigma^2\delta + \text{higher powers of } \delta \ .$$

The Central Limit Theorem, which can be found in section 5.1, suggest that $\log(x(t)/x_0)$ behaves like a normal random variable:

$$\log\left(\frac{x(t)}{x_0}\right) \sim N\left((\mu - \frac{1}{2}\sigma^2)t, \sigma^2 t\right).$$

The limiting continuous-time expression for the stock price at fixed time t becomes:

$$x(t) = x_0 e^{(\mu - \frac{1}{2}\sigma^2)t + \sigma\sqrt{t}W}, \quad \text{where } W \sim N(0,1).$$

For non-overlapping time intervals, the normal random variables that describe the changes will be independent. We can describe the evolution of the stock over any sequence of time points $0 = t_0 < t_1 < t_2 < \cdots < t_m$ by

$$x(t_i) = x(t_{i-1}) e^{(\mu - \frac{1}{2}\sigma^2)(t_i - t_{i-1}) + \sigma\sqrt{t_i - t_{i-1}}W_i}. \tag{4.2}$$

This model guarantees that the stock prices is always positive, if $x_0 > 0$. Model (4.2) is used a lot and is often referred to as geometric Brownian motion.

We like to model the daily closing prices, we assume that the time intervals $t_i - t_{i-1}$ are equally spaced. That is, we set the time between Friday

evening and Monday evening equal to the time between Thursday evening and Friday evening. We can write (4.2) as

$$x_t = x_{t-1} e^{(\mu - \frac{1}{2}\sigma^2)\delta + \sigma\sqrt{\delta}\, u_t}, \tag{4.3}$$

with δ equal to 1/260, because there are approximately 260 trading days in a year. This is basically the same as the discrete model (4.1).

From this model follows that $\log x_t$ is integrated of order one, because

$$\log x_t = \log x_{t-1} + (\mu - \frac{1}{2}\sigma^2)\delta + \sigma\sqrt{\delta}\ u_t,$$

hence

$$\begin{aligned} \log x_t - \log x_{t-1} &= (\mu - \frac{1}{2}\sigma^2)\delta + \sigma\sqrt{\delta}\ u_t \\ &= \text{constant} + \text{Gaussian white noise.} \end{aligned}$$

This difference process of $\log x_t$ is $I(0)$ and because the process $\log x_t$ itself is not, it follows that $\log x_t$ is $I(1)$. This is one of the reasons why cointegration tests are also applied to the logarithms of stock prices. Unfortunately, translating cointegration between the logarithm of two stocks into a trading strategy is less intuitively clear then translating cointegration between two stock prices themselves. When there is cointegration between the stock prices, trading the pair is very obvious. Let $\mathbf{y}_t = (x_t, y_t)$ be two stock prices processes which are cointegrated with cointegrating vector $\mathbf{a}$. We 'normalize' this vector to $(-\alpha, 1)$, so $y_t - \alpha x_t$ is a stationary process with mean zero, which means that y_t is approximately αx_t. It could be that there is a constant in the cointegrating relation, than $y_t - \alpha x_t$ does not have mean zero. This will be discussed in the next section, for now we assume that the mean is zero. We treat $y_t - \alpha x_t$ as our spread process described in chapter 2, so we trade pair (x, y) in the constant ratio $\alpha : 1$. This is exactly the same as the trading strategy, if we do not use the average ratio to calculate the spread but the least squares estimator.

If the logarithms of the stock prices x_t and y_t are cointegrated with cointegrating vector $\mathbf{b}$, we normalize this to $(-\beta, 1)$, then $\log y_t - \beta \log x_t$ is a stationary process. So $\log y_t$ is approximately $\beta \log x_t$, we cannot trade logarithms of stocks so we like to know the relation between x_t and y_t.

Let ε_t denote the residual process:

$$\log y_t - \beta \log x_t = \varepsilon_t.$$

The relation between x_t and y_t becomes:

$$y_t = x_t^{\beta} e^{\varepsilon_t}.$$

It is not clear how we can trade this relation, not with the strategy from chapter 2. This is the reason why we want to test for cointegration on the stock prices and not on their logarithms, in order to do that we need x_t and y_t to be integrated of order one. In chapter 9 we will make an attempt to come up with a trading strategy if we have cointegration between the logarithms of the stock prices.

Model (4.3) does not imply that x_t is $I(1)$, this is more easily seen in (4.1). The difference is

$$x_t - x_{t-1} = \mu\delta x_{t-1} + \sigma\sqrt{\delta} u_t x_{t-1},$$

this has not got a constant expectation, so according the derived stock price model the difference process is not $I(0)$. Fortunately, we look at the stock prices $\{x\}_{t=0}^{T}$ for fixed T, μ is a small number between 0.01 and 0.1 and typical values of σ are between 0.05 and 0.5, so it is not likely that x_{t-1} becomes very large or very small. That is why the differences divided by the mean value of x_{t-1} look a lot like the returns:

$$\frac{x_t - x_{t-1}}{x_{t-1}} \approx \frac{x_t - x_{t-1}}{\bar{x}_{t-1}} \ .$$

The returns are $I(0)$, this indicates that the difference process Δx_t are also more or less $I(0)$ and stock price process x_t more or less $I(1)$. We consider a realization of model (4.3), where we take $\mu = 0.03$, $\sigma = 0.18$ and $x_0 = 20$ shown in figure 4.10. The differences of this realizations is shown in figure 4.11, which looks like pretty stationary. This indicates that realizations of model (4.3) behave like they are $I(1)$, while strictly under the model they are not.

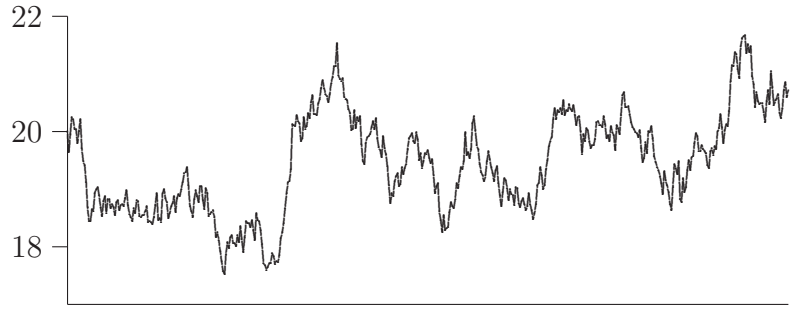

Figure 4.10: Realization of model (4.3), $\mu = 0.03$, $\sigma = 0.18$.

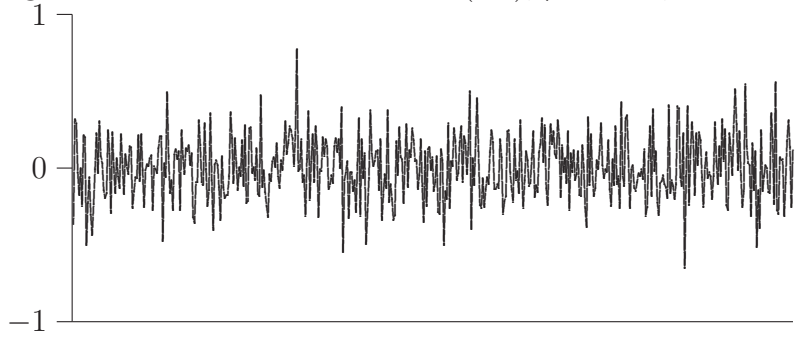

Figure 4.11: Differences of realization of model (4.3).

An other way to show that it is justifiable to assume the stock prices are integrated of order one, is to examine the differences instead of the returns. At the beginning of this section we examined the returns of Royal Dutch Shell, let us do the same for the differences $\Delta x_t = x_t - x_{t-1}$. Figure 4.12 shows the estimated density of the daily returns with the $N(0,1)$ density superimposed, figure 4.13 the empirical distribution function and figure 4.14 the normal QQ-plot. The daily differences were normalized to

$$\widehat{\Delta x}_t = \frac{\Delta x_t - \hat{\mu}}{\hat{\sigma}^2},$$

where $\hat{\mu}$ and $\hat{\sigma}^2$ are the sample mean and sample variance of the differences. Figure 4.15 shows the sample autocorrelation function of the differences.

These figures look pretty much the same as the figures for the returns, this suggests that it is justifiable to see the differences of a stock price process as normal distributed i.i.d samples. This implies that the differences are $I(0)$ and the stock prices $I(1)$.

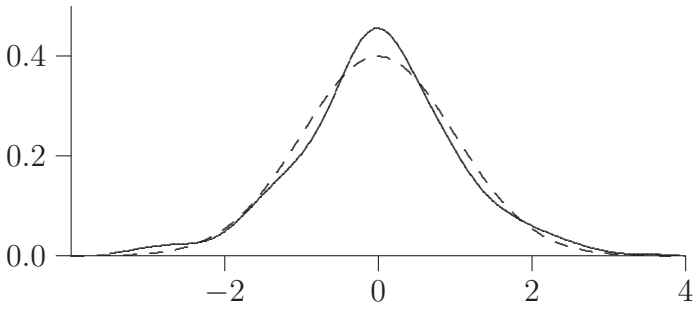

Figure 4.12: Estimated density.

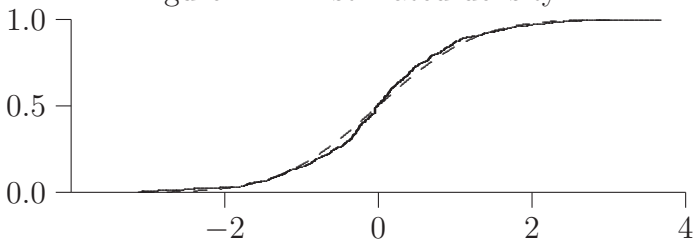

Figure 4.13: Empirical distribution function.

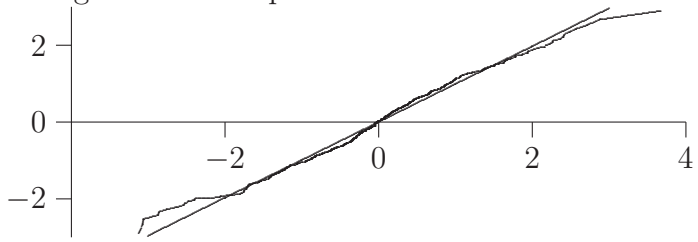

Figure 4.14: Normal QQ-plot.

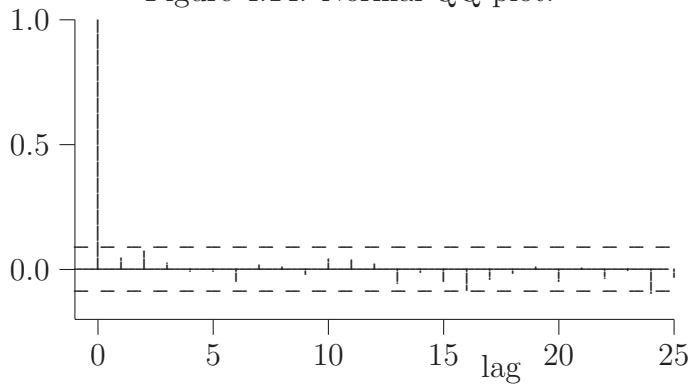

Figure 4.15: Autocorrelation function.

So far we have discussed why it is likely that stock price processes are integrated of order one, but we can also do unit root tests on the data we want to test for cointegration. The unit root test we use in this report is the (Augmented) Dickey-Fuller test, introduced in chapter 3. The first test is:

$$H_0 : \ x_t \sim I(1) \text{ against } H_1 : \ x_t \sim I(0).$$

The outcome should be to not reject H_0. The second test is:

$$H_0 : \ x_t \sim I(2) \text{ against } H_1 : \ x_t \sim I(1),$$

which is equivalent to:

$$H_0 : \ \Delta x_t \sim I(1) \text{ against } H_1 : \ \Delta x_t \sim I(0).$$

The outcome of this second test should be to reject H_0, which makes is likely that the price processes are $I(1)$. Which case of the DF-test should be used is discussed in the next section, the critical values of these tests are derived in chapter 5 and the results of these tests for data used in this report are stated in chapter 7.

4.3 Engle-Granger method

The question remains how to test for cointegration. There are several methods for testing for cointegration. R.F. Engle and C.W.J. Granger were the first to develop the key concepts of cointegration, which can be found in [5]. They received the nobel prize in economics in 2003 for their work on cointegration and ARCH models. The approach of testing for cointegration will be to test the null hypothesis that there is no cointegration among the elements of an $(n \times 1)$ vector $\mathbf{y}_t$. Rejection of the null hypothesis is then taken as evidence of cointegration.

The Engle-Granger test is a two-step process, which should be preceded by examining if each component of $\mathbf{y}_t$ is $I(1)$ which was discussed in the previous section. Let us assume that this condition is fulfilled. A vector process $\mathbf{y}_t$ is cointegrated if there exists a linear combination of its components $\mathbf{a}'\mathbf{y}_t$ that is stationary. The first step in the Engle-Granger test is to estimate $\mathbf{a}$, this is done with an OLS (Ordinary Least Squares) regression. The second step is

to test whether the residuals of the regression are stationary using a Dickey-Fuller test. Because if the residuals are stationary, the linear combination $\mathbf{a}'\mathbf{y}_t$ is stationary, which means $\mathbf{y}_t$ is cointegrated with cointegrating vector $\mathbf{a}$.

Looking from a pairs trading point of view we have two stock prices processes, $\mathbf{y}_t = (x_t, y_t)$. We like x_t and y_t to be cointegrated such that the spread $\varepsilon_t = y_t - \alpha x_t$ oscillates around zero, again we have 'normalized' the cointegrating vector $\mathbf{a}$ to $(-\alpha, 1)$. A stationary process has constant expectation but it is not necessarily equal to zero. In order to get a stationary process with mean zero, we can include a constant in the cointegration relation such that the spread becomes:

$$\varepsilon_t = y_t - \alpha x_t - \alpha_0.$$

For example, consider the pair (x_t, y_t) which is generated with the relation:

$$y_t = 2x_t + 20 + \varepsilon_t,$$

so $\alpha = 2$ and $\alpha_0 = 20$. Figure 4.16 shows x_t and y_t.

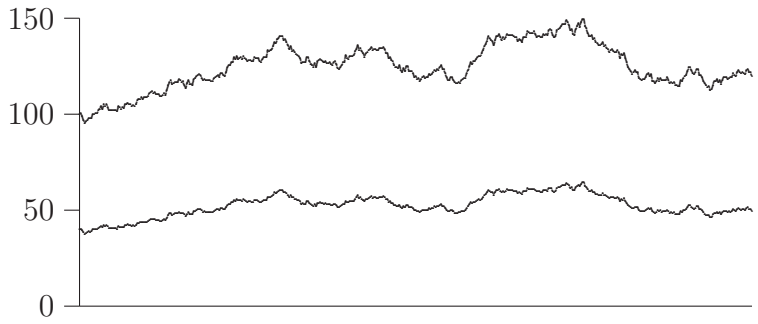

Figure 4.16: Paths x_t and y_t for $\alpha_0 = 20$.

Normally we do not know the exact value of α and α_0, so we have to estimate them. According to the Engle-Granger method we do this with OLS. We have two possibilities, regression with and without intercept. In the first situation, regression with intercept, we get

$$\hat{\alpha}_{\alpha_0} = 2.01 \quad \hat{\alpha}_0 = 19.67.$$

In the second, regression without intercept:

$$\hat{\alpha}_{\neg\alpha_0} = 2.41.$$

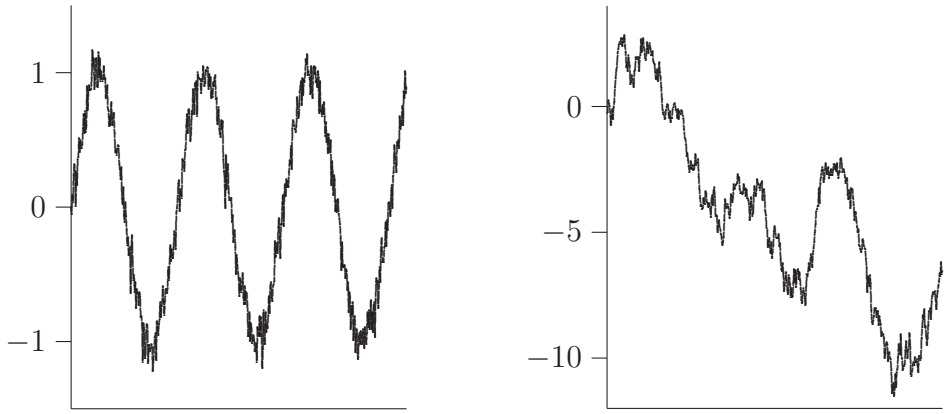

Figure 4.17: Spread with and without α_0, $\alpha_0 = 20$.

Figure 4.17 shows the corresponding spread processes. The left is $y_t - \hat{\alpha}_{\alpha_0} x_t - \alpha_0$, the right is $y_t - \hat{\alpha}_{\neg\alpha_0} x_t$. The left figure of 4.17 looks a lot better but there is a disadvantage. In section 2.3 about the properties of pairs trading was described that pairs trading is more or less cash neutral. The trading strategy is cash neutral up to Γ if we neglect costs for short selling, in other words each trade costs or provides us with Γ. The cash neutral property is a property we like to keep. If we trade the spread from the left figure of 4.17 it is not cash neutral anymore. Assume that the predetermined threshold Γ is equal to 1. The first time the spread is above 1, the value of x is € 43.73 and the value of y is € 108.59, which gives that the spread at this time equal to 108.59-2.01*43.73-19.67=1.02. Then we sell y which provides us with € 108.59 and buy 2.01 x which costs us 2.01*43.73=87.90. So we are left with a positive difference in money of € 20.69. The first time the spread is below -1, the value of x is € 52.66 and y is € 124.51, which gives that the spread at this time equal to 124.51-2.01*52.66-19.67=-1.01. Then we buy y which costs us € 124.51 and sell 2.01 x providing us 2.01*52.66=105.85. So this trade costs us € 18.66. This way of trading is not cash neutral, each trade costs or provides us approximately α_0.

A possibility to resolve this is to neglect α_0, so we trade the spread from the right figure of 4.17. In this example it is probably not worthwhile, because the spread has a clear downward trend. Let us consider the two different spreads for $\alpha_0 = 1$:

$$\varepsilon_t = y_t - 2x_t - 1,$$

shown in figure 4.18.

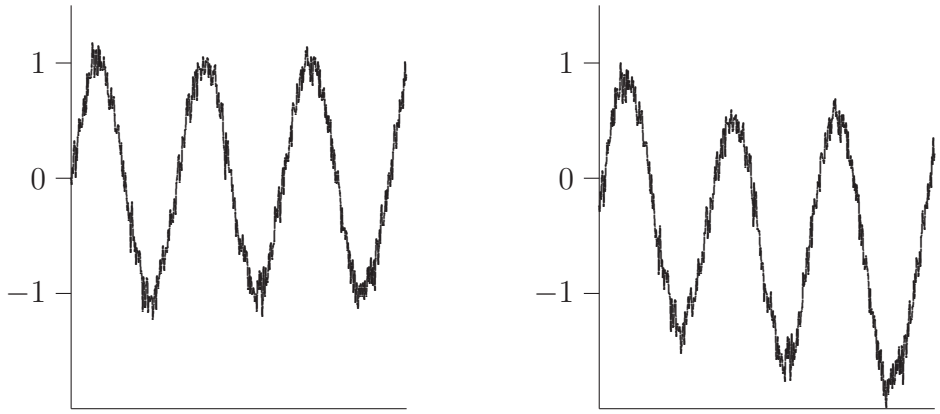

Figure 4.18: Spread with and without α_0, $\alpha_0 = 1$.

Now the spread where α_0 is neglected looks almost as good as the spread with α_0. In conclusion, in order to keep the cash neutral property and the trading strategy from chapter 2, α_0 should be close to zero such that neglecting it still gives a stationary spread process. So when testing real stock price processes x_t and y_t for cointegration, we only estimate α and test the residual process $y_t - \hat{\alpha}x_t$ for stationarity. A suggestion for an alternative trading strategy is stated in chapter 9, which is able to trade a pair when α_0 cannot be neglected.

It is not standard to do OLS regression on non-stationary data. OLS regression applied to non-stationary data is quite likely to produce spurious results. There is only one circumstance when the OLS estimation gives a consistent estimate of the cointegrating vector and that is when there is a cointegrating relation. Note that if $\varepsilon_t = y_t - \alpha x_t$ is stationary, then

$$\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t - \alpha x_t)^2 \xrightarrow{P} E(\varepsilon_t^2). \tag{4.4}$$

By contrast, if $(-1, \alpha)$ is not a cointegrating vector between x and y, then $y_t - \alpha x_t$ is $I(1)$ and from proposition 2 in section 5.6,

$$\frac{1}{T^2} \sum_{t=1}^{T} (y_t - \alpha x_t)^2 \xrightarrow{D} \lambda^2 \int_0^1 W(r)^2 \, dr,$$

where $W(r)$ is standard Brownian motion which will be defined in chapter 5 and λ is a parameter determined by the autocovariances of $\Delta\varepsilon_t$. Hence, if $(-1, \alpha)$ is not a cointegrating vector, the statistic in (4.4) would diverge to $+\infty$. This suggests that we can obtain a consistent estimate of a cointegrating vector by choosing α so as to minimize (4.4). It turns out that the OLS estimator for α, also when α_0 is included in the regression, converges at rate T. This is analyzed by Philips and Durlauf in [12].

Now that we have a method for estimating the cointegrating vector, the second step in the Engle-Granger method is examining the residuals with a Dickey-Fuller test. In chapter 3 was described that there are several cases, so the question remains which case do we use when testing for cointegration. In most literature about cointegration, it is not stated which case is used and why, but from the critical values used can be seen that case 2 is used most often. One discussion found in Hamilton [6], is the following:

> *Which case is the 'correct' case to use to test the null hypothesis of a unit root? The answer depends on why we are interested in testing for a unit root. If the analyst has a specific null hypothesis about the process that generated the data, then obviously this would guide the choice of test. In the absence of such guidance, one general principal would be to fit a specification that is a plausible description of the data under both the null and the alternative. This principle would suggest using case 4 for a series with a obvious trend and the case 2 for series without a significant trend. For example, the nominal interest rate series used in the examples in this section. There is no economic theory to suggest that nominal interest rates should exhibit a deterministic time trend, and so a natural null hypothesis is that the true process is a random walk without trend. In terms of framing a plausible alternative, it is difficult to maintain that these data could have been generated by* $i_t = \rho i_{t-1} + u_t$ *with* $|\rho|$ *significantly less than 1. If these data were*

to be described by a stationary process, surely the process would have a positive mean. This argues for including a constant term in the estimated regression, even though under the null hypothesis the true process does not contain a constant term. Thus, case 2 is a sensible approach for these data.

We do not have a specific null hypothesis, so according to this quote we should use case 2 because there is no trend in spread processes. In the next chapter we investigate the power of the three different tests, case 1 through case 3, maybe we can find another reason to use Dickey-Fuller case 2.

So far we have looked at cointegration between two stocks because of *pairs* trading. However, pairs trading with three of more stocks in a 'pair' is very interesting. Cointegration is defined for a $(n \times 1)$ vector $\mathbf{y}_t$ and the trading strategy is easily extended for three or more stocks. For example, consider a pair of three stocks $\mathbf{y}_t = (x_t, y_t, z_t)$ who are cointegrated with cointegrating vector $(-\alpha_1, -\alpha_2, 1)$. Then we calculate the spread as

$$s_t = z_t - \alpha_1 x_t - \alpha_2 y_t,$$

which we trade the same way as before. When the spread reaches Γ we sell 1 z and buy α_1 times x and α_2 times y. When the spread goes below $-\Gamma$ we reverse our position and lock in a profit of at least 2Γ. The threshold Γ is determined in the same way as in chapter 2, we just try a few on historical data and take the best one.

If the number of stocks in a pair is greater than two, $n > 2$, the Engle-Granger method has a disadvantage. We estimate the cointegrating vector with OLS regression, if $\mathbf{y}_t = (y_{1t}, y_{2t}, \ldots, y_{nt})$ we regress y_{1t} on $(y_{2t}, y_{3t}, \ldots, y_{nt})$. So the first element of the cointegrating vector is set to be unity. This normalization is not harmless if the first variable y_{1t} does not appear in the cointegrating relation at all, in other words, its coefficients is equal to zero but is set to one.

A second disadvantage, which exist also when $n = 2$, is that the method is not symmetric. Suppose $n = 2$, we regress y_{1t} on y_{2t}:

$$y_{1t} = \alpha y_{2t} + u_t.$$

We might equally well have normalized the coefficient of y_{2t}, so the regression would be

$$y_{2t} = \beta y_{1t} + v_t.$$

Then the OLS estimate $\hat{\beta}$ is not simply the inverse of $\hat{\alpha}$, meaning that these two regression will give different estimates of the cointegrating vector. Thus, choosing which variable to call y_1 and which to call y_2 might end up making a difference for the evidence one finds for cointegration.

For these reasons we discuss the Johansen method in the next section. First a summary is given for testing on cointegration with the Engle-Granger method:

- Given is $(n \times 1)$ vector $\mathbf{y}_t = (y_{1t}, y_{2t}, \ldots, y_{nt})$.
- Examine or assume that each individual variable y_{it} is $I(1)$.
- Then $\mathbf{y}_t$ is cointegrated if $\mathbf{a}'\mathbf{y}_t$ is $I(0)$ for some nonzero vector $\mathbf{a}$.
- Regress y_{1t} on $(y_{2t}, \ldots, y_{nt})$, a constant maybe included but with our trading strategy we do not want to include this constant. This regression gives the estimation $\hat{\mathbf{a}}$.
- Then the residuals of this regression, which is our spread process, are given by
$$e = y_{1t} - \hat{a}_2 y_{2t} - \cdots - \hat{a}_n y_{nt},$$
which resembles the real error process:
$$\varepsilon_t = y_{1t} - a_2 y_{2t} - \cdots - a_n y_{nt}.$$
- The Dickey-Fuller test assumes that ε_t follows an AR(p) model, with an unit root and with or without a constant term. If we use case 2, which is suggested by the above quote, we assume that the true model does not have a constant but we include a constant in the estimated model. So the Dickey-Fuller case 2 assumes the true model of ε_t is
$$\varepsilon_t = \rho\varepsilon_{t-1} + \beta_1 \Delta\varepsilon_{t-1} + \cdots + \beta_{p-1}\Delta\varepsilon_{t-p+1} + \eta_t,$$
with $\rho = 1$.
- To estimate p we fit an AR(k) model with OLS on e_t for $k = 1, \ldots, K$. The value of k with the smallest information criteria AIC(k), BIC(k) and HIC(k) is the estimate for the model order $\hat{p}$. If the information criteria give different values, we take the rounded mean.

- We use the AR($\hat{p}$) fit

$$e_t = \hat{c} + \hat{\rho} e_{t-1} + \hat{\beta}_1 \Delta e_{t-1} + \cdots + \hat{\beta}_{\hat{p}-1} \Delta e_{t-\hat{p}+1} + n_t \,,$$

  to calculate the Dickey-Fuller test statistic

$$\frac{\hat{\rho} - 1}{\hat{\sigma}_{\hat{\rho}}} \,,$$

  where $\hat{\rho}$ is the OLS estimate of ρ and $\hat{\sigma}_{\hat{\rho}}$ is the standard error for the estimated coefficient.

- Compare the outcome with the critical values of the Dickey-Fuller test.

The critical values of the Dickey-Fuller test will be derived and simulated in the next chapter. Engle-Granger is a two-step method, first we do an OLS regression and then a Dickey-Fuller test. In chapter 6 we will examine if the first step influences the critical values, in other words, are the critical values for Engle-Granger really the same as for Dickey-Fuller.

4.4 Johansen method

The Johansen method also known as 'full-information maximum likelihood' was developed by Søren Johansen in [8] and [9]. This method allows us to test for the number of cointegrating relations. An $(n \times 1)$ vector $\mathbf{y}_t$ has h cointegrating relations if there exists h linearly independent vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_h$ such that $\mathbf{a}_i' \mathbf{y}_t$ is stationary. If such vectors exist, their values are not unique, since any linear combination of $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_h$ is also a cointegrating vector. With the Engle-Granger method this was resolved by setting the first element in the cointegrating vector equal to one. As mentioned before this has some disadvantages. In this section the Johansen method is summarized, no proof or argumentation is given. These can be found in [8] and [9].

Let $\mathbf{y}_t$ be an $(n \times 1)$ vector. The Johansen method assumes that $\mathbf{y}_t$ follows a VAR(p) model

$$\mathbf{y}_t = \mathbf{c} + \mathbf{\Phi}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{\Phi}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t, \tag{4.5}$$

where $\mathbf{c}$ is an $(n \times 1)$ vector and $\mathbf{\Phi}_i$ is an $(n \times n)$ matrix.

Model (4.5) can be written as

$$\mathbf{y}_t = \mathbf{c} + \boldsymbol{\rho}\mathbf{y}_{t-1} + \boldsymbol{\beta}_1\Delta\mathbf{y}_{t-1} + \cdots + \boldsymbol{\beta}_{p-1}\Delta\mathbf{y}_{t-p+1} + \boldsymbol{\varepsilon}_t, \tag{4.6}$$

where

$$\begin{aligned} \boldsymbol{\rho} &= \boldsymbol{\Phi}_1 + \boldsymbol{\Phi}_2 + \cdots + \boldsymbol{\Phi}_p\,, \\ \boldsymbol{\beta}_i &= -\left(\boldsymbol{\Phi}_{i+1} + \boldsymbol{\Phi}_{i+2} + \cdots + \boldsymbol{\Phi}_p\right), \quad \text{for } i = 1, 2, \ldots, p-1. \end{aligned}$$

Subtracting $\mathbf{y}_{t-1}$ from both sides of (4.6) results in

$$\Delta\mathbf{y}_t = \mathbf{c} + \boldsymbol{\beta}_0\mathbf{y}_{t-1} + \boldsymbol{\beta}_1\Delta\mathbf{y}_{t-1} + \cdots + \boldsymbol{\beta}_{p-1}\Delta\mathbf{y}_{t-p+1} + \boldsymbol{\varepsilon}_t, \tag{4.7}$$

with

$$\begin{aligned} E(\boldsymbol{\varepsilon}_t) &= \mathbf{0}, \\ E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_\tau) &= \begin{cases} \boldsymbol{\Omega} & \text{for } t = \tau, \\ \mathbf{0} & \text{otherwise.} \end{cases} \end{aligned}$$

Johansen showed that under the null hypothesis of h cointegrating relations, only h separate linear combinations of $\mathbf{y}_t$ appear in (4.7). This implies that $\boldsymbol{\beta}_0$ can be written in the form

$$\boldsymbol{\beta}_0 = -\mathbf{B}\mathbf{A}'\,, \tag{4.8}$$

for $\mathbf{B}$ an $(n \times h)$ matrix and $\mathbf{A}'$ an $(h \times n)$ matrix.

If we consider a sample of $T+p$ observations, denoted $(\mathbf{y}_{-p+1}, \mathbf{y}_{-p+2}, \ldots, \mathbf{y}_T)$, and if the errors $\boldsymbol{\varepsilon}_t$ are Gaussian, the log likelihood of $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T)$ conditional on $(\mathbf{y}_{-p+1}, \mathbf{y}_{-p+2}, \ldots, \mathbf{y}_0)$ is given by

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\Omega}, \mathbf{c}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{p-1}) = \\ &\quad -\frac{Tn}{2}\log(2\pi) - \frac{T}{2}\log|\boldsymbol{\Omega}| \\ &\quad -\frac{1}{2}\sum_{t=1}^{T}\left[(\Delta\mathbf{y}_t - \mathbf{c} - \boldsymbol{\beta}_0\mathbf{y}_{t-1} - \boldsymbol{\beta}_1\Delta\mathbf{y}_{t-1} - \cdots - \boldsymbol{\beta}_{p-1}\Delta\mathbf{y}_{t-p+1})'\right. \\ &\quad \left.\times\ \boldsymbol{\Omega}^{-1}(\Delta\mathbf{y}_t - \mathbf{c} - \boldsymbol{\beta}_0\mathbf{y}_{t-1} - \boldsymbol{\beta}_1\Delta\mathbf{y}_{t-1} - \cdots - \boldsymbol{\beta}_{p-1}\Delta\mathbf{y}_{t-p+1})\right]. \end{aligned} \tag{4.9}$$

The goal is to choose $(\boldsymbol{\Omega}, \mathbf{c}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{p-1})$ so as to maximize (4.9) subject to the constraint that $\boldsymbol{\beta}$ can be written in the form of (4.8). The Johansen method calculates the maximum likelihood estimates of $(\boldsymbol{\Omega}, \mathbf{c}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{p-1})$.

The first step of the Johansen method is to estimate a VAR$(p-1)$ for $\Delta\mathbf{y}_t$. That is, regress Δy_{it} on a constant and all elements of the vectors $\Delta\mathbf{y}_{t-1}, \ldots, \Delta\mathbf{y}_{t-p+1}$ with OLS. Collect the $i = 1, 2, \ldots, n$ regressions in vector form

$$\Delta\mathbf{y}_t = \hat{\boldsymbol{\pi}}_0 + \hat{\boldsymbol{\Pi}}_1\Delta\mathbf{y}_{t-1} + \cdots + \hat{\boldsymbol{\Pi}}_{p-1}\Delta\mathbf{y}_{t-p+1} + \hat{\mathbf{u}}_t. \tag{4.10}$$

We also estimate a second regression, we regress $\mathbf{y}_{t-1}$ on a constant and $\Delta\mathbf{y}_{t-1}, \ldots, \Delta\mathbf{y}_{t-p+1}$

$$\mathbf{y}_{t-1} = \hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\chi}}_1\Delta\mathbf{y}_{t-1} + \cdots + \hat{\boldsymbol{\chi}}_{p-1}\Delta\mathbf{y}_{t-p+1} + \hat{\mathbf{v}}_t. \tag{4.11}$$

The second step is to calculate the sample covariance matrices of the OLS residuals $\hat{\mathbf{u}}_t$ and $\hat{\mathbf{v}}_t$:

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}_{\mathbf{vv}} &= \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{v}}_t\hat{\mathbf{v}}_t', \\
\hat{\boldsymbol{\Sigma}}_{\mathbf{uu}} &= \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{u}}_t\hat{\mathbf{u}}_t', \\
\hat{\boldsymbol{\Sigma}}_{\mathbf{uv}} &= \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{u}}_t\hat{\mathbf{v}}_t', \\
\hat{\boldsymbol{\Sigma}}_{\mathbf{vu}} &= \hat{\boldsymbol{\Sigma}}_{\mathbf{uv}}{}'.
\end{aligned}$$

From these, find the eigenvalues of the matrix

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{vv}}^{-1}\hat{\boldsymbol{\Sigma}}_{\mathbf{vu}}\hat{\boldsymbol{\Sigma}}_{\mathbf{uu}}^{-1}\hat{\boldsymbol{\Sigma}}_{\mathbf{uv}}, \tag{4.12}$$

with the eigenvalues ordered $\hat{\lambda}_1 > \hat{\lambda}_2 > \cdots > \hat{\lambda}_n$. The maximum value attained by the log likelihood function subject to the constraint that there are h cointegrating relations is given by

$$\mathcal{L}_0^* = -\frac{Tn}{2}\log(2\pi) - \frac{Tn}{2} - \frac{T}{2}\log|\hat{\boldsymbol{\Sigma}}_{\mathbf{uu}}| - \frac{T}{2}\sum_{i=1}^{h}\log(1-\hat{\lambda}_i). \tag{4.13}$$

The third step is to calculate the maximum likelihood estimates of the parameters. Let $\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_h$ denote the $(n \times 1)$ eigenvectors of (4.12) associated with the h largest eigenvalues. These provide a basis for the space of cointegrating relations. That is, the maximum likelihood estimate is that any cointegrating vector can be written in the form

$$\mathbf{a} = b_1\hat{\mathbf{a}}_1 + b_2\hat{\mathbf{a}}_2 + \cdots + b_h\hat{\mathbf{a}}_h \,,$$

for some choice of scalers $b_1, \ldots, b_h$. Johansen suggests normalizing these vector $\hat{\mathbf{a}}_i$ such that $\hat{\mathbf{a}}_i'\hat{\mathbf{\Sigma}}_{\mathbf{vv}}\hat{\mathbf{a}}_i = 1$. Collect the first h normalized vectors in a an $(n \times h)$ matrix $\hat{\mathbf{A}}$:

$$\hat{\mathbf{A}} = \left[\begin{array}{cccc} \hat{\mathbf{a}}_1 & \hat{\mathbf{a}}_2 & \cdots & \hat{\mathbf{a}}_h \end{array} \right] .$$

Then the maximum likelihood estimate of $\boldsymbol{\beta}_0$ is given by

$$\hat{\boldsymbol{\beta}}_0 = \hat{\mathbf{\Sigma}}_{\mathbf{uv}}\hat{\mathbf{A}}\hat{\mathbf{A}}' \,.$$

The maximum likelihood estimate of $\mathbf{c}$ is

$$\hat{\mathbf{c}} = \hat{\pi}_0 - \hat{\boldsymbol{\beta}}_0 \,.$$

Now we are ready for hypothesis testing. Under the null hypothesis that there are exactly h cointegrating relations, the largest value that can be achieved for the log likelihood function was given by (4.13). Consider the alternative hypothesis that there are n cointegrating relations. This means that *every* linear combination of $\mathbf{y}_t$ is stationary, in which case $\mathbf{y}_{t-1}$ would appear in (4.7) without constraints and no restrictions are imposed on $\boldsymbol{\beta}_0$. The value for the log likelihood function in the absence of constraints is given by

$$\mathcal{L}_1^* = -\frac{Tn}{2}\log(2\pi) - \frac{Tn}{2} - \frac{T}{2}\log|\hat{\mathbf{\Sigma}}_{\mathbf{uu}}| - \frac{T}{2}\sum_{i=1}^{n}\log(1-\hat{\lambda}_i). \qquad (4.14)$$

A likelihood ratio test of

$$H_0 : h \text{ relations} \quad \text{against} \quad H_1 : n \text{ relations},$$

can be based on

$$2(\mathcal{L}_1^* - \mathcal{L}_0^*) = -T\sum_{i=h+1}^{n}\log(1-\hat{\lambda}_i). \qquad (4.15)$$

An other approach would be to test the null hypothesis of h cointegrating relations against $h+1$ cointegrating relations. A likelihood ratio test of

$$H_0 : h \text{ relations} \quad \text{against} \quad H_1 : h+1 \text{ relations},$$

can be based on

$$2(\mathcal{L}_1^* - \mathcal{L}_0^*) = -T \log(1 - \hat{\lambda}_{h+1}). \tag{4.16}$$

Like with the Dickey-Fuller test, we need to distinguish several cases. There are also three cases for the Johansen method, but they are different than the Dickey-Fuller cases:

Case 1: The true value of the constant c in (4.7) is zero, meaning that there is no intercept in any of the cointegrating relations and no deterministic time trend in any of the elements of $\mathbf{y}_t$. There is no constant term included in the regressions (4.10) and (4.11).

Case 2: The true value of the constant c in (4.7) is such that there are no deterministic time trends in any of the elements of $\mathbf{y}_t$. There are no restrictions on the constant term in the estimation of the regressions (4.10) and (4.11).

Case 3: The true value of the constant c in (4.7) is such that one or more elements of $\mathbf{y}_t$ exhibit deterministic time trend. There are no restrictions on the constant term in the estimation of the regressions (4.10) and (4.11).

For both tests, which can be based on (4.15) and (4.16), the critical values for the three different cases can be found in [10] and [11]. Unfortunately, the critical values are for a sample size of $T = 400$. Although the data of the ten pairs IMC provided consist of 520 observations, these critical values will be used when testing the ten pairs for cointegration. I assume that the critical values are not that different for a sample size of 520. For case 1 this is very likely because Johansen showed that the asymptotic distribution of test statistic (4.15) is the same as that of the trace of matrix

$$\mathbf{Q} = \left[\int_0^1 \mathbf{W}(r) d\mathbf{W}(r)'\right]' \left[\int_0^1 \mathbf{W}(r)\mathbf{W}(r)' dr\right]^{-1} \left[\int_0^1 \mathbf{W}(r) d\mathbf{W}(r)'\right]$$

where $\mathbf{W}(r)$ is g-dimensional standard Brownian motion, with $g = n - h$.

And fortunately, case 1 is the case we will use because we do not want an intercept in the cointegrating relations, as was explained in the previous section, and we assume there is no deterministic time trend in the price processes. The Johansen case 1 test can be compared with the Dickey-Fuller case 1 test, there is no constant and we do not estimate one. There is not really a Johansen case which can be compared to the Dickey-Fuller case 2, with the Johansen case 2 test, the constant c is not necessarily equal to zero. The critical values for case 1 and $T = 400$ for both test statistics (4.15) and (4.16) are shown in tables 4.1 and 4.2 respectively.

Table 4.1: Critical values for test statistic (4.15).

	Case 1		
g	1%	5%	10%
1	6.51	3.84	2.86
2	16.31	12.53	10.47
3	29.75	24.31	21.63
4	45.58	39.89	36.58
5	66.52	59.46	55.44

Table 4.2: Critical values for test statistic (4.16).

	Case 1		
g	1%	5%	10%
1	6.51	3.84	2.86
2	15.69	11.44	9.52
3	22.99	17.89	15.59
4	28.82	23.80	21.58
5	35.17	30.04	27.62

Note that if $g = 1$, then $n = h + 1$. In this case the two tests are identical. For this reason the first rows of the tables are the same.

With two stocks in pair, we can do several hypothesis tests:

1) $H_0 : 0$ relations against $H_1 : 2$ relations,
2) $H_0 : 0$ relations against $H_1 : 1$ relation,
3) $H_0 : 1$ relation against $H_1 : 2$ relations.

For the first test, we use the second row of table 4.1. We basically test the null of no cointegration between the two stocks against the stocks themselves being stationary. Although the alternative hypothesis does not imply 'real' cointegration, because every linear combination of $\mathbf{y}_t$ is stationary since $\mathbf{y}_t$ is already stationary, rejection of the null is taken as evidence of cointegration.

For the second test, we use the second row of table 4.2. We test the null of no cointegration between the against the alternative of a single cointegration relation.

For the third test, we use the first row of either table. We test the null of one cointegrating relation against the stock prices being stationary already. Basically we test if the relation is a 'real' cointegrating relation.

If the third null hypothesis is rejected, the test indicates there are two cointegrating relations which means the stock prices themselves are stationary. As we saw in section 4.2 we do not think that stock prices are stationary, but if they are we can trade them as a pair like any other pair. We could even trade each stock as a spread process. That means, we apply the trading strategy on the price process instead of the spread process. But this would not be cash and market neutral anymore, and is seen as far more risky. So with two stocks in a pair, we would like there to be one or two cointegrating relations, but we expect there is only one. In chapter 7 the results of the different tests for the 10 pairs are given. They are compared with the results from the Engle-Granger method.

In the previous section was stated that the Johansen method has an advantage compared to the Engle-Granger method when there are more than two stocks in a pair, $n > 2$. With Johansen we do not impose the first element of the cointegrating relation to be unity, we normalize the estimated cointegrating relation such that the first element is unity or as Johansen proposed, normalizing such that $\hat{\mathbf{a}}_i' \hat{\mathbf{\Sigma}}_{\mathbf{vv}} \hat{\mathbf{a}}_i = 1$. With three stocks in a pair, we would like there to be one, two or three cointegrating relations but we expect that there are no more than two. With our pair trading strategy it does not matter how many relations there are as long as the stock are cointegrated, because we only trade one relation. This relation will be the eigenvector corresponding to the largest eigenvalue of the matrix in (4.12) because, according to Hamilton [6], this results in the most stationary spread process.

4.5 Alternative method

In this section a start is made with an alternative method. Assume, like the Engle-Granger and Johansen method, price processes x_t and y_t are integrated of order one: $x_t, y_t \sim I(1)$. Denote with $\mathbf{z}_t$ the vector of the differences of these price processes

$$\mathbf{z}_t = \begin{pmatrix} x_t - x_{t-1} \\ y_t - y_{t-1} \end{pmatrix}.$$

Then each component of $\mathbf{z}_t$ is $I(0)$, i.e. stationary. Notice that

$$\begin{pmatrix} x_t - x_0 \\ y_t - y_0 \end{pmatrix} = \sum_{i=1}^{t} \mathbf{z}_i \,.$$

Two price processes are cointegrated if a linear combination of them is stationary, i.e. constant mean, constant variance and autocovariances that do not depend on t. In this section we like to find out if $\mathbf{z}_t$ can be represented as an VAR(p) or as a vector MA(q) process.

Engle and Granger showed that a cointegrated system can never be represented by a finite-order vector autoregression in the differenced data $\Delta\mathbf{y}_t = \mathbf{z}_t$. The outline of the deduction is that if $\mathbf{z}_t$ is causal, i.e. $\mathbf{z}_t$ can be written as a linear combination of past innovations, and (x_t, y_t) are cointegrated then $\mathbf{z}_t$ is non-invertible. This implies that if (x_t, y_t) are cointegrated $\mathbf{z}_t$ cannot be represented by a VAR(p).

If we assume $\mathbf{z}_t$ to be an vector MA(q) process, we can find restrictions on the parameters of the model to ensure a linear combination of (x_t, y_t) such that it is stationary exists. Let us examine this for $q = 2$:

$$\mathbf{z}_t = \mathbf{\Theta}_2\mathbf{w}_{t-2} + \mathbf{\Theta}_1\mathbf{w}_{t-1} + \mathbf{\Theta}_0\mathbf{w}_t \,,$$

where $\mathbf{w}_t$ is i.i.d $N_2(0, \Sigma)$ and $\mathbf{\Theta}_0 = I$. Notice that a MA(q) process is always stationary.

Then

$$\begin{aligned} \sum_{i=1}^{t} \mathbf{z}_i &= \mathbf{\Theta}_2\mathbf{w}_{-1} + (\mathbf{\Theta}_2 + \mathbf{\Theta}_1)\mathbf{w}_0 + (\mathbf{\Theta}_2 + \mathbf{\Theta}_1 + \mathbf{\Theta}_0)\sum_{i=1}^{t-2} \mathbf{w}_i \\ &+ (\mathbf{\Theta}_1 + \mathbf{\Theta}_0)\mathbf{w}_{t-1} + \mathbf{\Theta}_0\mathbf{w}_t \,. \end{aligned}$$

If $\mathbf{v}$ is a cointegrating vector, i.e. a vector such that $\mathbf{v}\sum_{i=1}^{t}\mathbf{z}_i$ is stationary, than every multiple of $\mathbf{v}$ is also a cointegrating vector. We can make some kind of normalization so we can write $\mathbf{v} = [-\alpha \quad 1]$. For $t > 2$

$$\begin{aligned}
(y_t - \alpha x_t) \;-\; (y_0 - \alpha x_0) \;&=\; [-\alpha \quad 1]\textstyle\sum_{i=1}^{t}\mathbf{z}_i \\
&= [-\alpha \quad 1]\boldsymbol{\Theta}_2\mathbf{w}_{-1} + [-\alpha \quad 1](\boldsymbol{\Theta}_2 + \boldsymbol{\Theta}_1)\mathbf{w}_0 \quad \text{(begin)} \\
&+ [-\alpha \quad 1](\boldsymbol{\Theta}_2 + \boldsymbol{\Theta}_1 + \boldsymbol{\Theta}_0)\textstyle\sum_{i=1}^{t-2}\mathbf{w}_i \qquad \text{(middle)} \\
&+ [-\alpha \quad 1](\boldsymbol{\Theta}_1 + \boldsymbol{\Theta}_0)\mathbf{w}_{t-1} + [-\alpha \quad 1]\boldsymbol{\Theta}_0\mathbf{w}_t \quad \text{(end)}
\end{aligned} \tag{4.17}$$

The mean of (4.17) is constant for every $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2$ and α. The variance, however, is not. The number of terms in (begin) and (end) are the same for every t, so only the variance of the (middle) part of (4.17) is depending on t. To resolve this, $\boldsymbol{\Theta}_2, \boldsymbol{\Theta}_1$ and α have to satisfy:

$$[-\alpha \quad 1]\,(\boldsymbol{\Theta}_2 + \boldsymbol{\Theta}_1 + \boldsymbol{\Theta}_0) = 0.$$

The matrix $(\boldsymbol{\Theta}_2 + \boldsymbol{\Theta}_1 + \boldsymbol{\Theta}_0)$ must have an eigenvalue zero with eigenvector $[-\alpha \quad 1]$. Then (4.17) is a stationary process.

The same argument goes for $q > 2$. So if the difference process $\mathbf{z}_t$ is assumed to be an MA(q), then for (x_t, y_t) to be cointegrated the parameters have to satisfy:

$$\text{matrix } (\Theta_q + \Theta_{q-1} + \cdots + \Theta_0) \text{ has eigenvalue } 0. \tag{4.18}$$

The corresponding eigenvector is the cointegrating relation. Now we have a method to generate cointegrated data that is unlikely to satisfy the assumptions of the Engle-Granger method as well as the Johansen method. Engle-Granger assumes that $y_t - \alpha x_t$ is an AR(p) process and Johansen assumes that the vector (x_t, y_t) is a VAR(p). In section 6.4 we will see if the Engle-Granger method is robust enough to identify data generated in the way here described as cointegrated.

It should be possible to construct a new method for testing for cointegration. With real data it is obvious we can determine the difference process $\mathbf{z}_t$. It is, however, pretty difficult to estimate the parameters of the MA(q) with only 500 observations, specially when q becomes large. But if we could, than we could base a hypothesis test on the estimated eigenvalue closest to zero of the estimated matrices. We do not proceed with this in this report.

Chapter 5

Dickey-Fuller tests

In the literature that describes Dickey-Fuller tests there a lot of differences in the critical values. Some do not state clearly which true model is used, so the null hypothesis is not clear. Sometimes it seems that different models are used at the same time and sometimes there is the exact same model but the critical values are just different. That is why this chapter discusses the asymptotic distributions of the (Augmented) Dickey-Fuller test statistic to find the critical values for this test. In other words, this section discusses the asymptotic distributions for OLS estimated coefficients of unit root processes. They differ from those for stationary processes. The asymptotic distributions can be described in terms of functionals of Brownian motion. In the first section some notions and facts from probability theory, used to establish these distributions, are stated. In the next three sections the asymptotic distribution of the estimated coefficients for a first-order autoregression when the true process is a random walk are derived, i.e., the asymptotic distribution of the DF test statistic for case 1 to case 3. These distributions turn out to depend on whether a constant is included in the estimated regression. In section 5.5 the power of the three different cases is investigated. In section 5.6 the properties of the estimated coefficients for a pth-order autoregression are derived, i.e., distributions of the ADF test statistics. The book of Hamilton [6] is used for the derivation of the asymptotic distributions, this book clearly distinguishes the different models.

5.1 Notions/ facts from probability theory

First we need some definitions and theorems. For the following three definitions we assume that $\{X_T\}$ is a sequence of random variables, and X is a random variable, and all of them are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Convergence almost surely:
The sequence of random variables $\{X_T\}_{T=1}^{\infty}$ converges almost surely towards random variable X if

$$\mathbb{P}(\{\omega \,\epsilon\, \Omega : \lim_{T\to\infty} X_T(\omega) = X(\omega)\}) = 1.$$

Notation: $X_T \to X$ a.s.

Convergence in probability:
The sequence of random variables $\{X_T\}_{T=1}^{\infty}$ converges in probability towards random variable X if

$$\forall \varepsilon > 0 \ \lim_{T\to\infty} \mathbb{P}(\{\omega \,\epsilon\, \Omega : |X_T(\omega) - X(\omega)| > \varepsilon\}) = 0.$$

Notation: $X_T \xrightarrow{P} X$.

Convergence in distribution:
The sequence of random variables $\{X_T\}_{T=1}^{\infty}$ converges in distribution towards random variable X if for all bounded continuous functions g it holds that

$$E\, g(X_t) \to E\, g(X).$$

Notation: $X_T \xrightarrow{D} X$.

Central limit theorem:
Let $X_1, X_2, \ldots$ be a sequence of i.i.d variables such that $E\, X_1^2 < \infty$. Define $E\, X_1 = \mu$ and $\mathrm{var}(X_1) = \sigma^2$. Then

$$\sqrt{T}(\bar{X}_T - \mu) \xrightarrow{D} N(0, \sigma^2), \qquad \text{for } \ T \to \infty,$$

where $\bar{X}_T = \frac{1}{T}\sum_{t=1}^{T} X_t$.

Law of large numbers:
Let $X_1, X_2, \ldots$ be a sequence of i.i.d. variables such that $E|X_t| < \infty$, then

$$\frac{1}{T}\sum_{t=1}^{T} X_t \to E\,X_1 \quad \text{a.s.} \quad \text{for } T \to \infty.$$

Continuous mapping theorem(random vectors):
Let $\mathbf{X}_1, \mathbf{X}_2, \ldots$ be a sequence of random $(n \times 1)$ vectors with $\mathbf{X}_T \xrightarrow{D} \mathbf{X}$ and let $\mathbf{g} : R^n \to R^m$ be a continuous function, then

$$\mathbf{g}(\mathbf{X}_T) \xrightarrow{D} \mathbf{g}(\mathbf{X}).$$

A similar results hold for sequences of random functions:
Continuous mapping theorem(random functions):
Let $\{S_T(\cdot)\}_{T=1}^{\infty}$ and $S(\cdot)$ be random functions, such that $S_T(\cdot) \xrightarrow{D} S(\cdot)$ and let g be a continuous functional, then

$$g(S_T(\cdot)) \xrightarrow{D} g(S(\cdot)).$$

Definition Brownian motion:
Standard Brownian motion $W(\cdot)$ is a continuous-time stochastic process, associating each time point $t \in [0, 1]$ with the scalar $W(t)$ such that:

- (i) $W(0) = 0$,
- (ii) *For any time points $0 \le t_1 \le t_2 \le \ldots \le t_k \le 1$, the increments $[W(t_2) - W(t_1)], [W(t_3) - W(t_2)], \ldots, [W(t_k) - W(t_{k-1})]$ are independent multivariate Gaussian with $[W(s) - W(t)] \sim N(0, s - t)$,*
- (iii) *$W(t)$ is continuous in t with probability 1.*

Although $W(t)$ is continuous in t, it cannot be differentiated using standard calculus: the direction of change at t is likely to be completely different from that at $t + \delta$, no matter how small we make δ.

Now we like to derive something that is known as the *functional central limit theorem*. Let u_t be i.i.d variables with mean zero and finite variance σ^2. Given a sample size T, we can construct a variable $X_T(r)$ from the sample mean of the first rth fraction of observations, $r \in [0,1]$, defined by

$$X_T(r) = \frac{1}{T}\sum_{t=1}^{\lfloor Tr \rfloor} u_t \,,$$

where $\lfloor Tr \rfloor$ denotes the largest integer that is less than or equal to T times r. For any given realization, $X_T(r)$ is a step function in r, with

$$X_T(r) = \begin{cases} 0 & \text{for} \quad 0 \le r < 1/T, \\ u_1/T & \text{for} \quad 1/T \le r < 2/T, \\ (u_1+u_2)/T & \text{for} \quad 2/T \le r < 3/T, \\ \quad \vdots & \\ (u_1+\cdots+u_T)/T & \text{for} \quad r = 1. \end{cases}$$

Then

$$\sqrt{T}X_T(r) = \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr \rfloor} u_t = \frac{\sqrt{\lfloor Tr \rfloor}}{\sqrt{T}} \frac{1}{\sqrt{\lfloor Tr \rfloor}} \sum_{t=1}^{\lfloor Tr \rfloor} u_t \,.$$

By the central limit theorem

$$\frac{1}{\lfloor Tr \rfloor}\sum_{t=1}^{\lfloor Tr \rfloor} u_t \xrightarrow{D} N(0,\sigma^2)$$

while $\left(\sqrt{\lfloor Tr \rfloor}/\sqrt{T}\right) \to \sqrt{r}$. Hence the asymptotic distribution of $\sqrt{T}X_T(r)$ is that of $\sqrt{r}$ times a $N(0,\sigma^2)$ random variable, or

$$\sqrt{T}\left[X_T(r)/\sigma\right] \xrightarrow{D} N(0,r).$$

Consider the behavior of a sample mean based on observations $\lfloor Tr_1 \rfloor$ through $\lfloor Tr_2 \rfloor$ for $r_2 > r_1$, than we can conclude that this too is asymptotically normal

$$\sqrt{T}\left[X_T(r_2) - X_T(r_1)\right]/\sigma \xrightarrow{D} N(0, r_2 - r_1).$$

More generally, the sequence of stochastic functions $\{\sqrt{T}X_T(\cdot)/\sigma\}_{T=1}^{\infty}$ has an asymptotic probability law that is described by standard Brownian motion $W(\cdot)$:

$$\sqrt{T}X_T(\cdot)/\sigma \xrightarrow{D} W(\cdot)\,. \tag{5.1}$$

There is a difference between the expressions $X_T(\cdot)$ and $X_T(r)$, the first denotes a random function while the last denotes the value that function assumes at time r, it is a random variable. Result (5.1) is known as the *functional central limit theorem.* The derivation here assumed that u_t was i.i.d.

Proposition 1:
Suppose that z_t follows a random walk without drift

$$z_t = z_{t-1} + u_t\,,$$

where $z_0 = 0$ and u_t is i.i.d. with mean zero and finite variance σ^2. Then

$$\begin{array}{lll} (i) & T^{-1/2}\sum_{t=1}^{T} u_t & \xrightarrow{D} \sigma W(1)\,, \\ (ii) & T^{-3/2}\sum_{t=1}^{T} z_{t-1} & \xrightarrow{D} \sigma \int_0^1 W(r)dr\,, \\ (iii) & T^{-2}\sum_{t=1}^{T} z_{t-1}^2 & \xrightarrow{D} \sigma^2 \int_0^1 W(r)^2 dr\,, \\ (iv) & T^{-1}\sum_{t=1}^{T} z_{t-1}u_t & \xrightarrow{D} \sigma^2 \left(W(1)^2 - 1\right)/2\,. \end{array}$$

Proof of proposition 1:
(i) follows from the central limit theorem. $W(1)$ denotes a random variable with a $N(0,1)$ distribution, so $\sigma W(1)$ denotes a random variable with a $N(0,\sigma^2)$ distribution.

(ii): Note that $X_T(r)$ can be written as

$$X_T(r) = \begin{cases} 0 & \text{for} \quad 0 \le r < 1/T\,, \\ z_1/T & \text{for} \quad 1/T \le r < 2/T\,, \\ z_2/T & \text{for} \quad 2/T \le r < 3/T\,, \\ \quad\vdots & \\ z_T/T & \text{for} \quad r = 1\,. \end{cases}$$

The area under this step function is the sum of T rectangles, each with width $1/T$:

$$\int_0^1 X_T(r)\,dr = z_1/T^2 + \cdots + z_{T-1}/T^2.$$

Multiplying both sides with $\sqrt{T}$:

$$\int_0^1 \sqrt{T} X_T(r)\,dr = T^{-3/2}\sum_{t=1}^{T} z_{t-1}.$$

Statement (ii) follows by the functional central limit theorem and the continuous mapping theorem.

(iii): Define $S_T(r)$ as

$$S_T(r) = T\left[X_T(r)\right]^2 .$$

This can be written as

$$S_T(r) = \begin{cases} 0 & \text{for} \quad 0 \leq r < 1/T\,, \\ z_1^2/T & \text{for} \quad 1/T \leq r < 2/T\,, \\ z_2^2/T & \text{for} \quad 2/T \leq r < 3/T\,, \\ \quad \vdots & \\ z_T^2/T & \text{for} \quad r = 1\,. \end{cases}$$

It follows that

$$\int_0^1 S_T(r)\, dr = z_1^2/T^2 + \cdots + z_{T-1}^2/T^2.$$

By the continuous mapping theorem:

$$S_T(r) = \left[\sqrt{T}X_T(r)\right]^2 \xrightarrow{D} \sigma^2 \left[(W(\cdot)\right]^2 .$$

Again applying this theorem:

$$\int_0^1 S_T(r)\, dr \xrightarrow{D} \sigma^2 \int_0^1 W(r)^2\, dr\,,$$

which gives statement (iii).

(iv): Note that for a random walk

$$z_t^2 = (z_{t-1} + u_t)^2 = z_{t-1}^2 + 2z_{t-1}u_t + u_t^2\,,$$

summing over $t = 1, 2, \ldots, T$ results in

$$\sum_{t=1}^{T} z_{t-1}u_t = 1/2(z_T^2 - z_0^2) - 1/2\sum_{t=1}^{T} u_t^2\,.$$

Recall that $z_0 = 0$ and dividing by T gives

$$\begin{aligned} T^{-1}\sum_{t=1}^{T} z_{t-1}u_t &= \frac{z_T^2}{2T} - \frac{1}{2T}\sum_{t=1}^{T} u_t^2 \\ &= \frac{S_T(1)}{2} - \frac{1}{2T}\sum_{t=1}^{T} u_t^2\,. \end{aligned}$$

But $S_T(1) \to^D \sigma^2 W(1)^2$ and by the law of large numbers $T^{-1}\sum_{t=1}^T u_t^2 \xrightarrow{P} \sigma^2$ which proofs (iv).

Now we are ready to construct some asymptotic properties of OLS estimators of AR(1) processes when there is an unit root.

5.2 Dickey-Fuller case 1 test

Consider a AR(1) process

$$z_t = \rho z_{t-1} + u_t\,, \qquad \text{for } t = 1, \ldots, T\,, \tag{5.2}$$

with $\rho \geq 0$ and where $u_t \sim$ i.i.d with mean zero and finite variance σ^2. The OLS estimate of ρ is given by

$$\hat{\rho} = \frac{\sum_{t=1}^T z_{t-1} z_t}{\sum_{t=1}^T z_{t-1}^2}\,.$$

The t statistic S, used for testing the null hypothesis that ρ is equal to some particular value ρ^0, is given by

$$S = \frac{\hat{\rho} - \rho^0}{\hat{\sigma}_{\hat{\rho}}}$$

where $\hat{\sigma}_{\hat{\rho}}$ is the standard error of the OLS estimate of ρ:

$$\hat{\sigma}_{\hat{\rho}} = \left(r_T^2 / \sum_{t=1}^T z_{t-1}^2 \right)^{1/2}$$

with

$$r_T^2 = \frac{1}{T-1} \sum_{t=1}^T (z_t - \hat{\rho} z_{t-1})^2\,.$$

When (5.2) is stationary, i.e. $\rho < 1$, S has an limiting Gaussian distribution:

$$S \xrightarrow{D} N(0,1).$$

But Dickey-Fuller tests the null hypothesis that $\rho = 1$, so we like to know the limiting distribution of S when $\rho = 1$. Then we can write S as:

$$S = \frac{\hat{\rho} - 1}{\hat{\sigma}_{\hat{\rho}}} = \frac{\hat{\rho} - 1}{\left(r_T^2 / \sum_{t=1}^T z_{t-1}^2 \right)^{1/2}}\,. \tag{5.3}$$

The numerator of (5.3) can be written as:

$$\hat{\rho} - 1 = \frac{\sum_{t=1}^{T} z_{t-1} u_t}{\sum_{t=1}^{T} z_{t-1}^2} \,. \tag{5.4}$$

Substituting this in (5.3):

$$\begin{aligned} S &= \frac{\sum_{t=1}^{T} z_{t-1} u_t}{\left(\sum_{t=1}^{T} z_{t-1}^2\right)^{1/2} (r_T^2)^{1/2}} \\ &= \frac{T^{-1} \sum_{t=1}^{T} z_{t-1} u_t}{\left(T^{-2} \sum_{t=1}^{T} z_{t-1}^2\right)^{1/2} (r_T^2)^{1/2}} \,. \end{aligned}$$

Apart from the initial term z_0, which does not affect the asymptotic distributions (unfortunately it could affect the finite sample size distributions, we will see this later on), the variable z_t is the same as in proposition 1. So it follows from proposition 1 (iii) and (iv) together with $r_T^2 \xrightarrow{P} \sigma^2$, that as $T \to \infty$:

$$S \xrightarrow{D} \frac{\sigma^2 \left(W(1)^2 - 1\right)/2}{\left(\sigma^2 \int_0^1 W(r)^2 \, dr\right)^{1/2} (\sigma^2)^{1/2}} = \frac{\frac{1}{2}\left(W(1)^2 - 1\right)}{\left(\int_0^1 W(r)^2 \, dr\right)^{1/2}} \,. \tag{5.5}$$

In conclusion, when the true model is a random walk without a constant term ($\rho = 1, c = 0$) and we only estimate ρ and not a constant, basically a regression without intercept, the t statistic S has limiting distribution (5.5). This test statistic is referred to as the Dickey-Fuller case 1 test statistic. Note that $W(1)$ has a $N(0,1)$ distribution, meaning that $W(1)^2$ has a $\chi^2(1)$ distribution.

We can approximate this asymptotic distribution and the corresponding critical values by simulating a lot of paths W on the interval $[0,1]$:

- Divide the interval [0,1] in n equal pieces.
- Take $u_1, u_2, \ldots, u_n$ i.i.d from a $N(0, 1/n)$ distribution.
- Set W(0)=0.

- Build path W by: $W(\frac{i}{n}) = W(\frac{i-1}{n}) + u_i \quad \text{for } i = 1, 2, \ldots, n.$

For each path the fraction in the right-hand side of (5.5) can be calculated, approximating the integrals with Riemann sums. Then the density of S can be estimated with applying a Gaussian kernel estimator on all these values. Figure 5.1 shows the estimated density for 5,000 paths and $n = 500$.

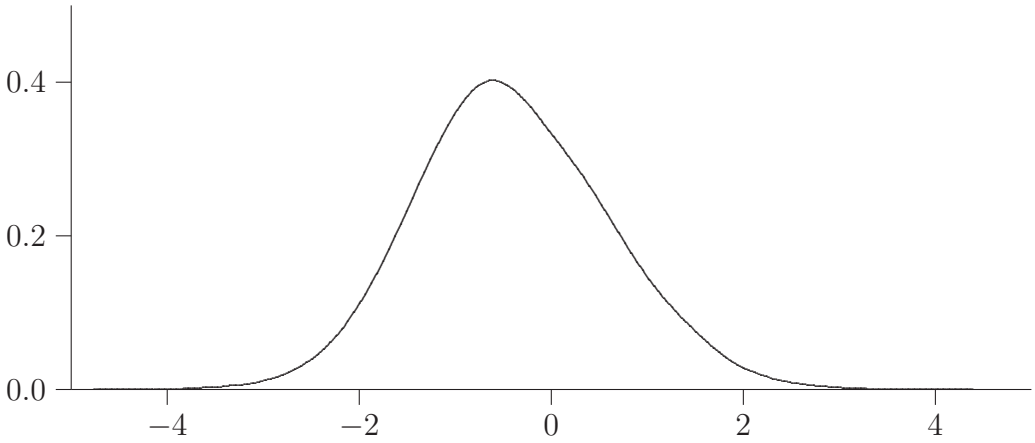

Figure 5.1: Asymptotic density of DF case 1 test statistic.

We can approximate the 1%, 5% and 10% critical values by calculating the corresponding quantiles of all the calculated fractions. Table 5.1 shows the critical values according to this simulation and the values according to Hamilton [6].

Table 5.1: Critical values for DF case 1.

	1%	5%	10%
Hamilton	-2.58	-1.95	-1.62
simulation	-2.56	-1.95	-1.60

These critical values belong to the asymptotic distribution (5.5), which describes the distribution of the DF case 1 test statistic if the sample size T goes to infinity.

We approximate the critical values for finite sample sizes T, by simulating in a different way:

- Take $u_1, \ldots, u_T$ from a $N(0, \sigma^2)$ distribution.
- Set $z_0 = 0$.
- Build path z_t: $z_t = z_{t-1} + u_t, \quad t = 1, \ldots, T$.
- Calculate $\hat{\rho}$.
- Calculate $\hat{\sigma}_{\hat{\rho}}$.
- Calculate test statistic: $(\hat{\rho} - 1)/\hat{\sigma}_{\hat{\rho}}$.
- Repeat the preceding steps 5,000 times.

We can approximate the 1%, 5% and 10% critical values by calculating the corresponding quantiles of the simulated test statistics. For finite T, the critical values are exact only under the assumption of Gaussian innovations. As T becomes large, these values also describe the asymptotic distribution for non-Gaussian innovations. Table 5.2 shows the critical values according to this simulation for different values of T and σ^2. Table 5.3 shows the critical values according Hamilton [6]. The critical values should be independent of σ. Table 5.2 shows roughly the same values for different σ^2, but as σ^2 becomes large there is more dispersion. Figure 5.2 shows the estimated density of the simulated test statistics for different values of σ^2 and $T = 500$, the graph of figure 5.1 is also displayed.

Table 5.2: Simulated critical values for DF case 1.

	$\sigma^2 = 1$			$\sigma^2 = 5$			$\sigma^2 = 10$		
T	1%	5%	10%	1%	5%	10%	1%	5%	10%
100	-2.61	-1.98	-1.63	-2.62	-1.98	-1.63	-2.55	-1.94	-1.61
250	-2.61	-1.96	-1.62	-2.56	-1.94	-1.59	-2.54	-1.93	-1.59
500	-2.56	-1.95	-1.61	-2.59	-1.95	-1.62	-2.59	-2.00	-1.63

Table 5.3: Hamilton's critical values DF case 1.

T	1%	5%	10%
100	-2.60	-1.95	-1.61
250	-2.58	-1.95	-1.62
500	-2.58	-1.95	-1.62

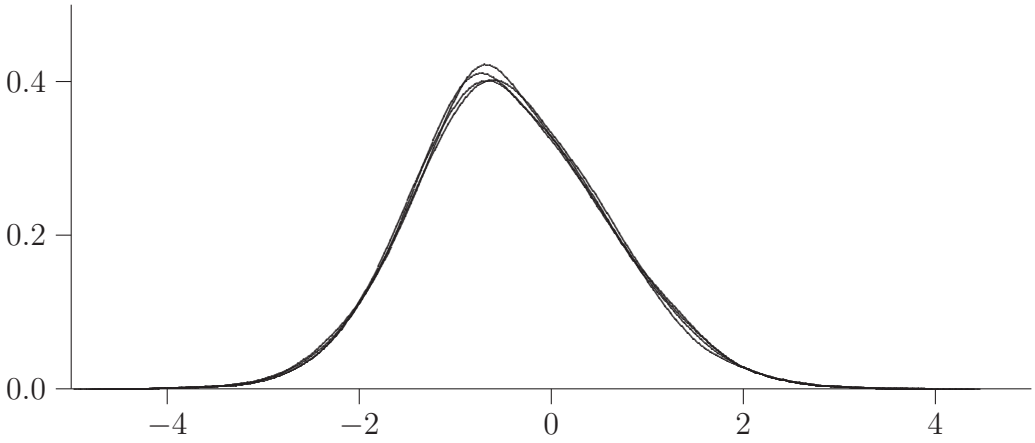

Figure 5.2: Estimated density of DF case 1 for different σ^2 and $T = 500$.

The initial term z_0 does not affect the asymptotic distribution. Unfortunately it does affect the distribution when the sample size is finite. With Dickey-Fuller case 1, we basically fit a line that goes through the origin. If the initial term is large the slope of this line, $\hat{\rho}$, is closer to one then when the initial term is small. The standard error for $\hat{\rho}$, $\hat{\sigma}_{\hat{\rho}}$ is a lot smaller for a large initial value then for a small initial value. That is why the test statistic for a large initial value is likely to be larger than the test statistic for a small initial value. The estimated densities for initial values $z_0 = 0, 1, 10, 50, 100, 500$ are shown in figure 5.3. The solid lines correspond to $z_0 = 0, 1, 10$, the dashed lines correspond to $z_0 = 50, 100, 500$. We see a shift to the right as the initial value increases. The density found with simulating Brownian motion is not displayed, it lies among the three solid lines.

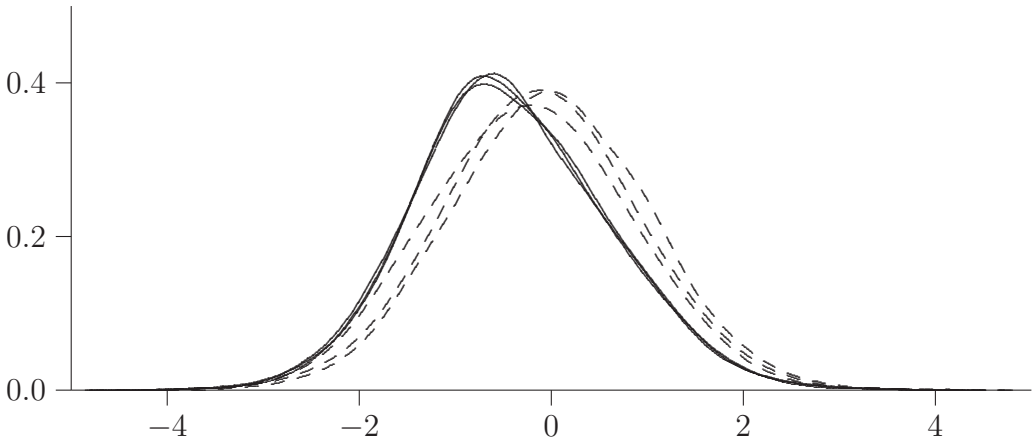

Figure 5.3: Estimated density of DF case 1 for different z_0 and $T = 500$.

5.3 Dickey-Fuller case 2 test

In this section we consider the AR(1) process with a constant:

$$z_t = c + \rho z_{t-1} + u_t\,, \qquad \text{for } t = 1, \ldots, T,$$

where $u_t \sim$ i.i.d with mean zero and finite variance σ^2. We are interested in the properties of test statistic $S = \frac{\hat{\rho}-1}{\hat{\sigma}_{\hat{\rho}}}$ under the null hypothesis that $c = 0$ and $\rho = 1$. The OLS estimates are given by

$$\begin{bmatrix} \hat{c} \\ \hat{\rho} \end{bmatrix} = \begin{bmatrix} T & \sum z_{t-1} \\ \sum z_{t-1} & \sum z_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum z_t \\ \sum z_{t-1} z_t \end{bmatrix}$$

here Σ denotes summation over $t = 1, \ldots, T$.

The deviation of the estimates from the true values is

$$\begin{bmatrix} \hat{c} \\ \hat{\rho} - 1 \end{bmatrix} = \begin{bmatrix} T & \sum z_{t-1} \\ \sum z_{t-1} & \sum z_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum u_t \\ \sum z_{t-1} u_t \end{bmatrix}. \tag{5.6}$$

The estimates $\hat{c}$ and $\hat{\rho}$ have different rates of convergence, a scaling matrix $\mathbf{Y}$ is helpful in describing their limiting distributions. Note if

$$\mathbf{v} = \mathbf{A}^{-1}\mathbf{w}\,,$$

then

$$\begin{aligned} \mathbf{Yv} &= \mathbf{YA}^{-1}\mathbf{w} \\ &= \mathbf{YA}^{-1}\mathbf{YY}^{-1}\mathbf{w} \\ &= (\mathbf{Y}^{-1}\mathbf{AY}^{-1})^{-1}\ \mathbf{Y}^{-1}\mathbf{w}. \end{aligned} \tag{5.7}$$

Here we use the scaling matrix

$$\mathbf{Y} = \begin{bmatrix} T^{1/2} & 0 \\ 0 & T \end{bmatrix}.$$

With (5.7), equation (5.6) results in

$$\begin{bmatrix} T^{1/2}\hat{c} \\ T(\hat{\rho}-1) \end{bmatrix} = \begin{bmatrix} 1 & T^{-3/2}\sum z_{t-1} \\ T^{-3/2}\sum z_{t-1} & T^{-2}\sum z_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} T^{-1/2}\sum u_t \\ T^{-1}\sum z_{t-1}u_t \end{bmatrix} \tag{5.8}$$

From the proposition in paragraph 5.1 follows that the first term of the right side of (5.8) converges to

$$\begin{aligned} \begin{bmatrix} 1 & T^{-3/2}\sum z_{t-1} \\ T^{-3/2}\sum z_{t-1} & T^{-2}\sum z_{t-1}^2 \end{bmatrix} &\xrightarrow{D} \begin{bmatrix} 1 & \sigma\int W(r)dr \\ \sigma\int W(r)dr & \sigma^2\int W(r)^2dr \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix} \begin{bmatrix} 1 & \int W(r)dr \\ \int W(r)dr & \int W(r)^2dr \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix} \end{aligned} \tag{5.9}$$

where the integral sign denotes integration over r from 0 to 1.

The second term of the right side of (5.8) converges to

$$\begin{aligned} \begin{bmatrix} T^{-1/2}\sum u_t \\ T^{-1}\sum z_{t-1}u_t \end{bmatrix} &\xrightarrow{D} \begin{bmatrix} \sigma W(1) \\ \sigma^2(W(1)^2-1)/2 \end{bmatrix} \\ &= \sigma \begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix} \begin{bmatrix} W(1) \\ (W(1)^2-1)/2 \end{bmatrix} \end{aligned} \tag{5.10}$$

Substituting (5.9) and (5.10) into (5.8) establishes

$$\begin{aligned} &\begin{bmatrix} T^{1/2}\hat{c} \\ T(\hat{\rho}-1) \end{bmatrix} \xrightarrow{D} \\ &\begin{bmatrix} \sigma & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \int W(r)dr \\ \int W(r)dr & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ (W(1)^2-1)/2 \end{bmatrix} \end{aligned} \tag{5.11}$$

The second element in the vector in (5.11) states that

$$T(\hat{\rho}-1) \xrightarrow{D} \frac{\frac{1}{2}(W(1)^2-1) - W(1)\int W(r)dr}{\int W(r)^2dr - \left[\int W(r)dr\right]^2}\,. \tag{5.12}$$

We like to know the properties of the t statistic S:

$$S = \frac{\hat{\rho}-1}{\hat{\sigma}_{\hat{\rho}}}\,,$$

where

$$\begin{aligned} \hat{\sigma}^2_{\hat{\rho}} &= r_T^2 \left[\begin{array}{cc} 0 & 1 \end{array}\right] \left[\begin{array}{cc} T & \sum z_{t-1} \\ \sum z_{t-1} & \sum z^2_{t-1} \end{array}\right]^{-1} \left[\begin{array}{c} 0 \\ 1 \end{array}\right], \qquad (5.13)\\ r_T^2 &= \frac{1}{T-2}\sum_{t=1}^{T}(z_t - \hat{c} - \hat{\rho}z_{t-1})^2\,. \end{aligned}$$

If we multiply both sides of (5.13) by T^2, the result can be written as

$$T^2\hat{\sigma}^2_{\hat{\rho}} = r_T^2 \left[\begin{array}{cc} 0 & 1 \end{array}\right] \mathbf{Y} \left[\begin{array}{cc} T & \sum z_{t-1} \\ \sum z_{t-1} & \sum z^2_{t-1} \end{array}\right]^{-1} \mathbf{Y} \left[\begin{array}{c} 0 \\ 1 \end{array}\right]. \tag{5.14}$$

From (5.8) follows

$$\begin{aligned} &\mathbf{Y} \left[\begin{array}{cc} T & \sum z_{t-1} \\ \sum z_{t-1} & \sum z^2_{t-1} \end{array}\right]^{-1} \mathbf{Y} \xrightarrow{D} \\ &\left[\begin{array}{cc} 1 & 0 \\ 0 & \sigma \end{array}\right]^{-1} \left[\begin{array}{cc} 1 & \int W(r)dr \\ \int W(r)dr & \int W(r)^2dr \end{array}\right]^{-1} \left[\begin{array}{cc} 1 & 0 \\ 0 & \sigma \end{array}\right]^{-1} \end{aligned}$$

From equation (5.14) and $r_T^2 \xrightarrow{P} \sigma^2$ follows

$$\begin{aligned} T^2\hat{\sigma}^2_{\hat{\rho}} &\xrightarrow{D} \left[\begin{array}{cc} 0 & 1 \end{array}\right] \left[\begin{array}{cc} 1 & \int W(r)dr \\ \int W(r)dr & \int W(r)^2dr \end{array}\right]^{-1} \left[\begin{array}{c} 0 \\ 1 \end{array}\right] \\ &= \frac{1}{\int W(r)^2dr - \left[\int W(r)dr\right]^2}\,. \end{aligned}$$

Finally, the asymptotic distribution of test statistic S is

$$\begin{aligned} S &= \frac{T(\hat{\rho}-1)}{[T^2\hat{\sigma}_{\hat{\rho}}^2]^{1/2}} \\ &\xrightarrow{D} \frac{(W(1)^2-1)/2 - W(1)\int W(r)dr}{\left(\int W(r)^2dr - \left[\int W(r)dr\right]^2\right)^{1/2}}. \end{aligned} \tag{5.15}$$

In conclusion, when the true model is a random walk without a constant term ($\rho = 1, c = 0$) but we do estimate a constant c and of course ρ, the t test statistic S has the asymptotic distribution described by (5.15). This test statistic is referred to as the Dickey-Fuller case 2 test statistic.

We can find this asymptotic distribution and the corresponding critical values by simulating a lot of paths W in the same way as in the preceding paragraph. The results are shown in figure 5.4 and table 5.4. Figure 5.4 shows that the distribution of the DF case 2 statistic is shifted more to the left than the DF case 1 statistic.

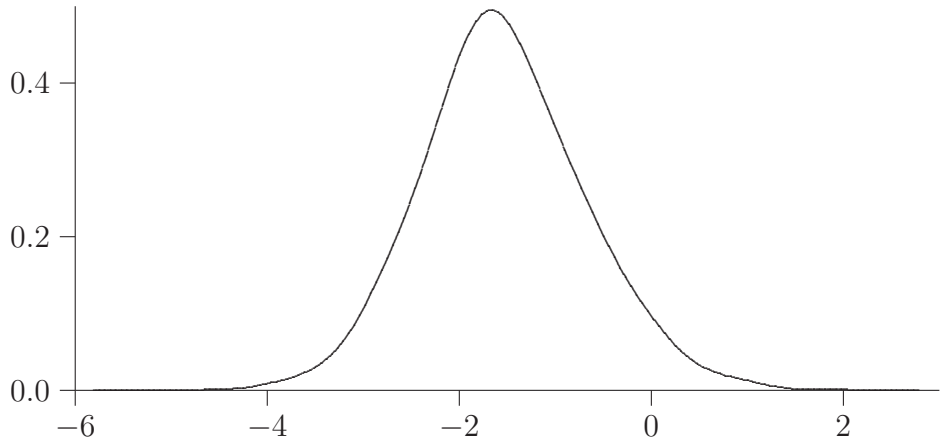

Figure 5.4: Asymptotic density of DF case 2 test statistic.

Table 5.4: Critical values for DF case 1.

	1%	5%	10%
Hamilton	-3.43	-2.86	-2.57
simulation	-3.43	-2.85	-2.59

These critical values belong to the asymptotic distribution (5.15), which describes the distribution of the DF case 2 test statistic if the sample size T goes to infinity. We find the critical values for finite sample sizes T, by simulating paths z_t as in the preceding paragraph. The only difference is the way we calculate $\hat{\rho}$. Again we simulate for different values of σ^2. The results are shown in table 5.5, table 5.6 shows the critical values for DF case 2 according to Hamilton [6]. Figure 5.5 shows the estimated density of the simulated test statistics for different values of σ^2 and $T = 500$.

Table 5.5: Simulated critical values for DF case 2.

	$\sigma^2 = 1$			$\sigma^2 = 5$			$\sigma^2 = 10$		
T	1%	5%	10%	1%	5%	10%	1%	5%	10%
100	-3.54	-2.85	-2.54	-3.50	-2.88	-2.56	-3.54	-2.89	-2.58
250	-3.40	-2.84	-2.53	-3.46	-2.87	-2.56	-3.46	-2.86	-2.56
500	-3.42	-2.86	-2.58	-3.40	-2.86	-2.55	-3.49	-2.87	-2.58

Table 5.6: Hamilton's critical values DF case 2.

T	1%	5%	10%
100	-3.51	-2.89	-2.58
250	-3.46	-2.88	-2.57
500	-3.44	-2.87	-2.57

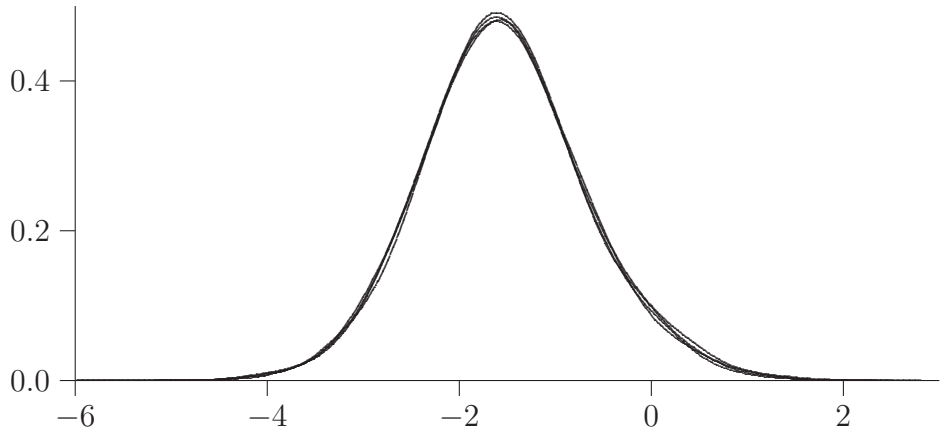

Figure 5.5: Estimated density of DF case 2 for different σ^2 and $T = 500$.

Like in the preceding section, the initial term z_0 does not affect the asymptotic distribution. And fortunately, with case 2, it does not affect the distribution when the sample size is finite as well. With case 2 we estimate a constant even though it is not present in the true model, we basically fit a line which does not have to go through the origin. Then the slope of line, $\hat{\rho}$, is not closer to one if the initial value is large, such as with case 1. That is why the test statistic for a large initial value is likely to be the same as the test statistic for a small initial value. The estimated densities for initial values $z_0 = 0, 1, 10, 50, 100, 500$ are shown in figure 5.6.

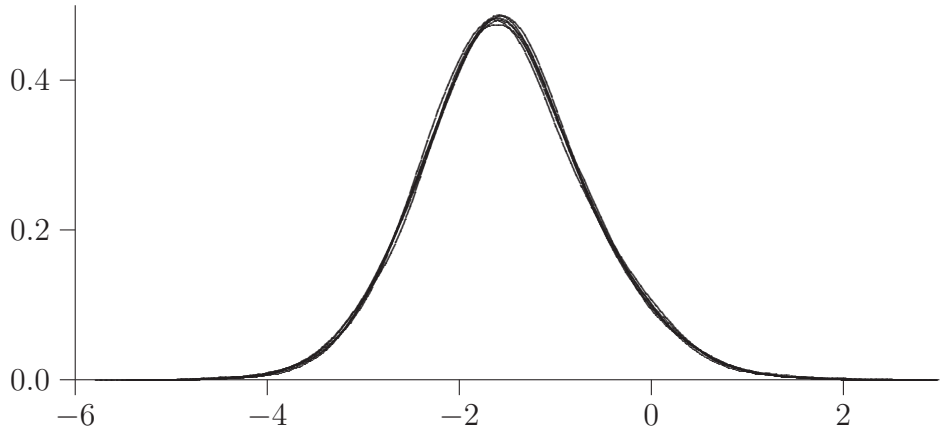

Figure 5.6: Estimated density of DF case 2 for different z_0 and $T = 500$.

5.4 Dickey-Fuller case 3 test

In this section we consider again the AR(1) process with a constant

$$z_t = c + \rho z_{t-1} + u_t\,, \qquad \text{for } t = 1, \ldots, T, \tag{5.16}$$

where $u_t \sim$i.i.d with mean zero and finite variance σ^2. We are interested in the properties of test statistic $S = \frac{\hat{\rho}-1}{\hat{\sigma}_{\hat{\rho}}}$ under the null hypothesis that $\rho = 1$ and $c \neq 0$.

The deviation of the estimates from the true values is

$$\begin{bmatrix} \hat{c} - c \\ \hat{\rho} - 1 \end{bmatrix} = \begin{bmatrix} T & \sum z_{t-1} \\ \sum z_{t-1} & \sum z_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum u_t \\ \sum z_{t-1} u_t \end{bmatrix}, \tag{5.17}$$

here Σ denotes the summation over $t = 1, \ldots, T$.

We examine the four different sum terms in the right side of (5.17) separately. First notice that (5.16) can be written as:

$$z_t = z_0 + c\,t + (u_1 + u_2 + \ldots + u_t) = z_0 + c\,t + v_t\,,$$

where

$$v_t = u_1 + \ldots + u_t\,, \quad \text{for } t = 1, \ldots, T, \quad \text{with } v_0 = 0\,.$$

Consider the behavior of the sum

$$\sum_{t=1}^{T} z_{t-1} = \sum_{t=1}^{T} \left[z_0 + c(t-1) + v_{t-1} \right]. \tag{5.18}$$

The first term in (5.18) is $T z_0$, so divided by T is a fixed value. The second term is equal to

$$\sum_{t=1}^{T} c(t-1) = (T-1)Tc/2\,.$$

In order to converge, this term has to be divided by T^2:

$$\frac{1}{T^2} \sum_{t=1}^{T} c(t-1) \to c/2\,.$$

The third term in (5.18) converges when divided by $T^{3/2}$, according to proposition 1 *ii*):

$$T^{-3/2}\sum_{t=1}^{T} v_{t-1} \xrightarrow{D} \sigma \int_0^1 W(r)dr\,.$$

The order in probability of the three terms in (5.18) is:

$$\sum_{t=1}^{T} z_{t-1} = \underbrace{\sum_{t=1}^{T} z_0}_{O(T)} + \underbrace{\sum_{t=1}^{T} c(t-1)}_{O(T^2)} + \underbrace{\sum_{t=1}^{T} v_{t-1}}_{O_p(T^{3/2})}\,.$$

The time trend $c(t-1)$ asymptotically dominates the other two components:

$$\frac{1}{T^2}\sum_{t=1}^{T} z_{t-1} \xrightarrow{P} c/2\,.$$

In the same way, we have

$$\begin{aligned}
\sum_{t=1}^{T} z_{t-1}^2 &= \sum_{t=1}^{T} \left[z_0 + c(t-1) + v_{t-1}\right]^2 \\
&= \underbrace{\sum_{t=1}^{T} z_0^2}_{O(T)} + \underbrace{\sum_{t=1}^{T} c^2(t-1)^2}_{O(T^3)} + \underbrace{\sum_{t=1}^{T} v_{t-1}^2}_{O_p(T^2)} \\
&+ \underbrace{\sum_{t=1}^{T} 2z_0 c(t-1)}_{O(T^2)} + \underbrace{\sum_{t=1}^{T} 2z_0 v_{t-1}}_{O_p(T^{3/2})} + \underbrace{\sum_{t=1}^{T} 2c(t-1)v_{t-1}}_{O_p(T^{5/2})}
\end{aligned}$$

where the order of the last term follows from

$$T^{-5/2}\sum_{t=1}^{T} t v_{t-1} = T^{-3/2}\sum_{t=1}^{T} (t/T) v_{t-1} \xrightarrow{D} \sigma \int_0^1 rW(r)dr\,.$$

The time trend $c^2(t-1)^2$ is the only term that does not vanish asymptotically if we divide by T^3:

$$\frac{1}{T^3}\sum_{t=1}^{T} z_{t-1}^2 \xrightarrow{P} c^2/3$$

From the central limit theorem follows that $\sum_{t=1}^{T} u_t$ is of order $O_p(T^{1/2})$.

And finally

$$\begin{aligned}\sum_{t=1}^{T} z_{t-1}u_t &= \sum_{t=1}^{T} \left[z_0 + c(t-1) + v_{t-1}\right] u_t \\ &= \underbrace{z_0 \sum_{t=1}^{T} u_t}_{O_p(T^{1/2})} + \underbrace{\sum_{t=1}^{T} c(t-1)u_t}_{O_p(T^{3/2})} + \underbrace{\sum_{t=1}^{T} v_{t-1}u_t}_{O_p(T)}\end{aligned}$$

from which

$$T^{-3/2} \sum_{t=1}^{T} z_{t-1}u_t \xrightarrow{P} T^{-3/2} \sum_{t=1}^{T} c(t-1)u_t \,.$$

This results in the deviation of the OLS estimates from their true values satisfy

$$\begin{bmatrix} \hat{c} - c \\ \hat{\rho} - 1 \end{bmatrix} = \begin{bmatrix} O_p(T) & O_p(T^2) \\ O_p(T^2) & O_p(T^3) \end{bmatrix}^{-1} \begin{bmatrix} O_p(T^{1/2}) \\ O_p(T^{3/2}) \end{bmatrix}$$

In this case the scaling matrix is

$$\mathbf{Y} = \begin{bmatrix} T^{1/2} & 0 \\ 0 & T^{3/2} \end{bmatrix}$$

Using (5.7) we get

$$\begin{bmatrix} T^{1/2}(\hat{c} - c) \\ T^{3/2}(\hat{\rho} - 1) \end{bmatrix} = \begin{bmatrix} 1 & T^{-2}\sum z_{t-1} \\ T^{-2}\sum z_{t-1} & T^{-3}\sum z_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} T^{-1/2}\sum u_t \\ T^{-3/2}\sum z_{t-1}u_t \end{bmatrix} \quad (5.19)$$

where the first term of the righthand side converges to

$$\begin{bmatrix} 1 & T^{-2}\sum z_{t-1} \\ T^{-2}\sum z_{t-1} & T^{-3}\sum z_{t-1}^2 \end{bmatrix} \xrightarrow{P} \begin{bmatrix} 1 & c/2 \\ c/2 & c^2/3 \end{bmatrix} = \mathbf{A} \,. \quad (5.20)$$

The second term of (5.19) satisfies

$$\begin{bmatrix} T^{-1/2}\sum u_t \\ T^{-3/2}\sum z_{t-1}u_t \end{bmatrix} = \begin{bmatrix} T^{-1/2}\sum u_t \\ T^{-3/2}\sum c(t-1)u_t \end{bmatrix} + o_p(1)$$

Therefore

$$\begin{aligned}\begin{bmatrix} T^{-1/2}\sum u_t \\ T^{-3/2}\sum z_{t-1}u_t \end{bmatrix} &\xrightarrow{D} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & c/2 \\ c/2 & c^2/3 \end{bmatrix}\right) \\ &= N(\mathbf{0}, \sigma^2\mathbf{A})\,. \end{aligned} \tag{5.21}$$

It follows from (5.19)-(5.21) that

$$\begin{bmatrix} T^{1/2}(\hat{c}-c) \\ T^{3/2}(\hat{\rho}-1) \end{bmatrix} \xrightarrow{D} N(\mathbf{0}, \mathbf{A}^{-1}\sigma^2\mathbf{A}\mathbf{A}^{-1}) = N(\mathbf{0}, \sigma^2\mathbf{A}^{-1})$$

We like to know the properties of the t statistic S:

$$S = \frac{\hat{\rho}-1}{\hat{\sigma}_{\hat{\rho}}}\,, \tag{5.22}$$

where

$$\hat{\sigma}^2_{\hat{\rho}} = r_T^2 \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} T & \sum z_{t-1} \\ \sum z_{t-1} & \sum z_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{5.23}$$

with

$$r_T^2 = \frac{1}{T-2}\sum_{t=1}^{T}(z_t - \hat{c} - \hat{\rho}z_{t-1})^2\,.$$

Test statistic (5.22) can be written as:

$$S = \frac{T^{3/2}(\hat{\rho}-1)}{T^{3/2}\hat{\sigma}_{\hat{\rho}}}\,.$$

The denominator is:

$$\begin{aligned} T^{3/2}\hat{\sigma}_{\hat{\rho}} &= \left(r_T^2 \begin{bmatrix} 0 & T^{3/2} \end{bmatrix} \begin{bmatrix} T & \sum z_{t-1} \\ \sum z_{t-1} & \sum z_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ T^{3/2} \end{bmatrix}\right)^{1/2} \\ &= \left(r_T^2 \begin{bmatrix} 0 & 1 \end{bmatrix} \mathbf{Y} \begin{bmatrix} T & \sum z_{t-1} \\ \sum z_{t-1} & \sum z_{t-1}^2 \end{bmatrix}^{-1} \mathbf{Y} \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right)^{1/2}. \end{aligned}$$

We have already shown that

$$\mathbf{Y}\begin{bmatrix} T & \sum z_{t-1} \\ \sum z_{t-1} & \sum z_{t-1}^2 \end{bmatrix}^{-1}\mathbf{Y} = \left(\mathbf{Y}^{-1}\begin{bmatrix} T & \sum z_{t-1} \\ \sum z_{t-1} & \sum z_{t-1}^2 \end{bmatrix}\mathbf{Y}^{-1}\right)^{-1}$$
$$= \begin{bmatrix} 1 & T^{-2}\sum z_{t-1} \\ T^{-2}\sum z_{t-1} & T^{-3}\sum z_{t-1}^2 \end{bmatrix}$$

converges in probability towards $\mathbf{A}$.

Because $r_T^2 \xrightarrow{P} \sigma^2$, the denominator converges towards

$$T^{3/2}\hat{\sigma}_{\hat{\rho}} \xrightarrow{P} \sigma c/\sqrt{3}\,.$$

Thus, the test statistic S is asymptotically Gaussian. The regressor y_{t-1} is asymptotically dominated by the time trend $c(t-1)$. In large samples, it is as if the explanatory variable y_{t-1} were replaced by the time trend $c(t-1)$. That is why the asymptotic properties of $\hat{c}$ and $\hat{\rho}$ are the same as those for the deterministic time trend regression. Therefore, for finite T test statistic S has a t distribution.

In conclusion, when the true model is a random walk with a constant term ($\rho = 1, c \neq 0$) and we estimate both ρ and c then the t test statistic S has an asymptotic distribution equal to the standard Gaussian distribution:

$$S \xrightarrow{D} N(0,1)$$

This test statistic is referred to as the Dickey-Fuller case 3 test statistic. The critical values for $T \to \infty$ are given in table 5.7.

Table 5.7: Critical values for DF case 3.

	1%	5%	10%
$N(0,1)$	-2.33	-1.64	-1.28

For finite T the Dickey-Fuller case 3 test statistic is t distributed, but the degrees of freedom are large so it is almost standard normal. We can also find the critical values for finite T, by simulating paths z_t as in the preceding paragraphs. Again we simulate for different values of σ^2. The results are shown in table 5.8. Figure 5.7 shows the estimated density of the simulated test statistics for different values of σ^2 while $T = 500$ and $c = 2.5$, the standard normal density is also displayed.

Table 5.8: Simulated critical values for DF case 3.

	$\sigma^2 = 1$			$\sigma^2 = 5$			$\sigma^2 = 10$		
T	1%	5%	10%	1%	5%	10%	1%	5%	10%
100	-2.36	-1.76	-1.37	-2.51	-1.82	-1.46	-2.48	-1.84	-1.49
250	-2.37	-1.68	-1.33	-2.41	-1.74	-1.35	-2.40	-1.79	-1.40
500	-2.38	-1.75	-1.33	-2.41	-1.69	-1.35	-2.45	-1.76	-1.38

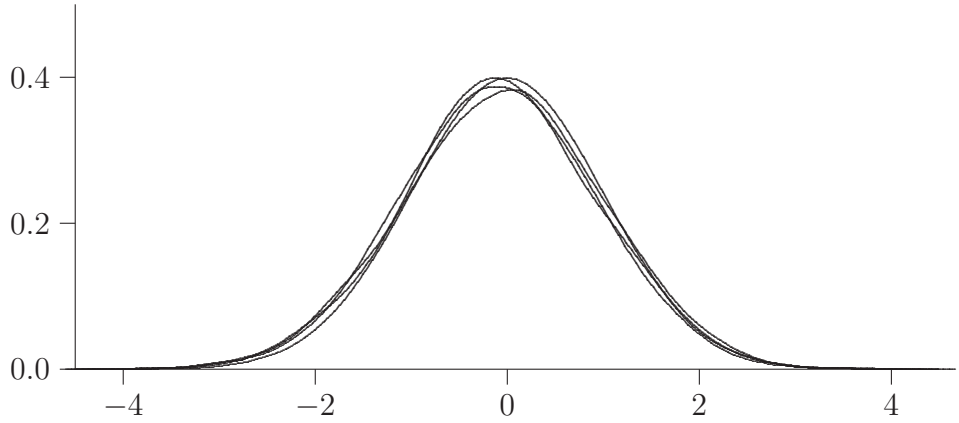

Figure 5.7: Estimated density of DF case 3 for different σ^2 and $T = 500$.

To see if the value of the constant c has an impact on finite sample distribution of the Dickey-Fuller case 3 statistic, we simulate paths z_t for $c = 0.1, 0.5, 1, 2.5, 10$. The results are shown in figure 5.8, the standard normal density is also displayed. The graph most left corresponds with $c = 0.1$.

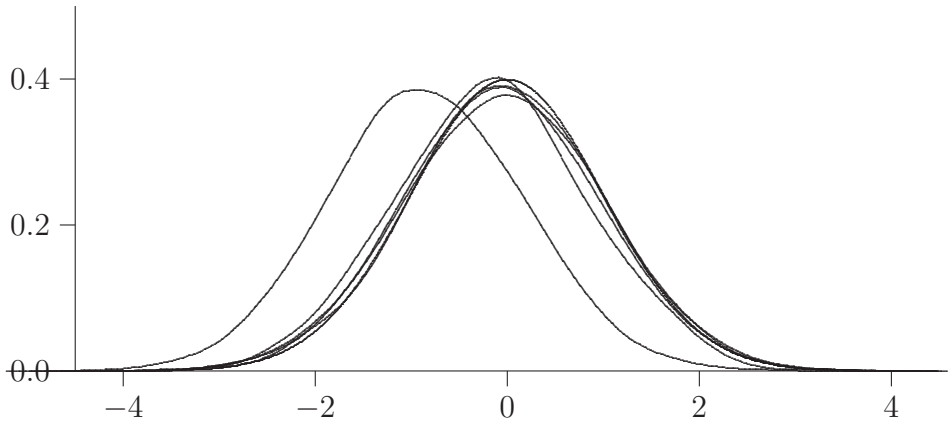

Figure 5.8: Estimated density of DF case 3 for different c and $T = 500$.

It looks like a small value of c causes a shift to the left. In figure 5.9 the estimated distribution for $c = 0.01, 0.05, 0.1$ is plotted with solid lines. The dashed line is the distribution of case 1 and the dotted line is the distribution of case 2. For $c = 0.01$ the estimated density is almost the same as the density found for case 2. This makes sense because the steps taken in the case 2 en case 3 tests are exactly the same, the only difference is the true model for case 2 has no constant and for case 3 it has. So for a decreasing constant the case 3 test statistic converges to the case 2 statistic. The other way around is also valid: for an increasing constant, in absolute value, the case 2 test statistic converges to the case 3 statistic because the tests are the same but now there is a constant in the true model.

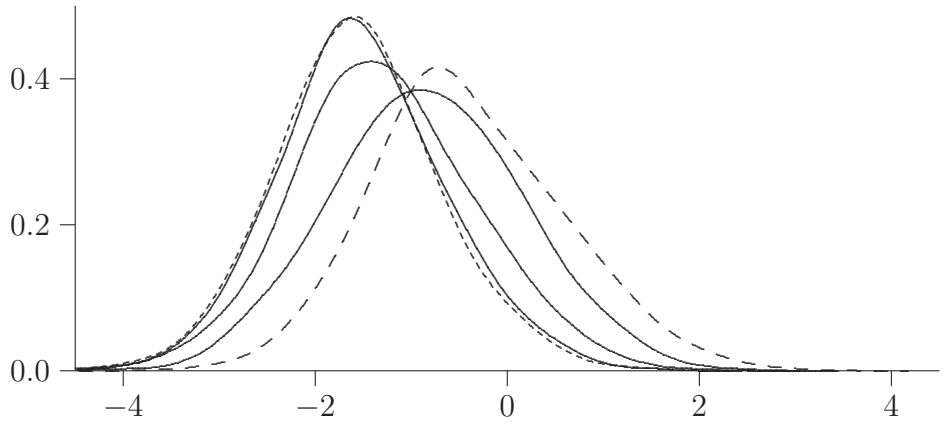

Figure 5.9: Estimated density of DF case 3 for small c and $T = 500$.

To be consistent, we also simulate for several different initial values z_0. Figure 5.10 shows the results for $z_0 = 0, 1, 10, 50, 100, 500$ while $T = 500$ and $c = 2.5$. The figure suggests that the initial value z_0 does not affect the density of the test statistic for finite sample sizes.

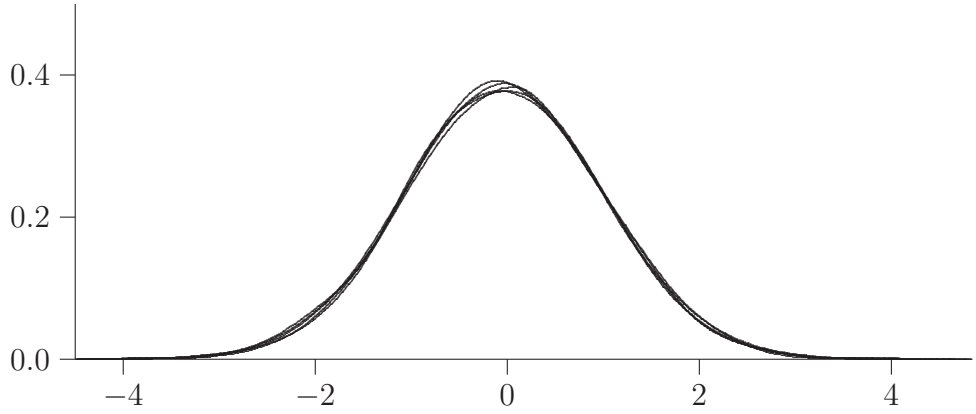

Figure 5.10: Estimated density of DF case 3 for different z_0

There also exist a case 4 for the Dickey-Fuller test, this includes a deterministic time trend in the true model. We are not interested in spread processes with deterministic trends so we do not discuss this case.

5.5 Power of the Dickey-Fuller tests

In this section we investigate how 'powerful' the different cases of the Dickey-Fuller test are. We generate paths z_t which do not have a unit root, $\rho < 1$ in

$$z_t = c + \rho z_{t-1} + u_t$$

and see if the different tests see it as stationary. In other words whether the outcome of the test is to reject the null hypothesis that there is a unit root, $\rho = 1$. For different values of c and z_0 we generate 1,000 paths z_t, $t = 0, \ldots, T$ where $T = 500$ and count the number of rejections. The results are presented in tables.

First we summarize the previous sections.

Case 1: The true model of case 1 is $z_t = z_{t-1} + u_t$ where $u_t \sim$ i.i.d with mean zero and finite variance σ^2. We estimate the model $z_t = \rho z_{t-1} + u_t$. The critical values for test statistic $S = \frac{\hat{\rho}-1}{\hat{\sigma}_{\hat{\rho}}}$ when $T = 500$ are

1%	5%	10%
-2.58	-1.95	-1.62

Case 2: The true model of case 2 is $z_t = z_{t-1} + u_t$ where $u_t \sim$ i.i.d with mean zero and finite variance σ^2. We estimate the model $z_t = c + \rho z_{t-1} + u_t$. The critical values for test statistic $S = \frac{\hat{\rho}-1}{\hat{\sigma}_{\hat{\rho}}}$ when $T = 500$ are

1%	5%	10%
-3.44	-2.87	-2.57

Case 3: The true model of case 3 is $z_t = c + z_{t-1} + u_t$ where $u_t \sim$ i.i.d with mean zero and finite variance σ^2 and $c \neq 0$. We estimate the model $z_t = c + \rho z_{t-1} + u_t$. The critical values for test statistic $S = \frac{\hat{\rho}-1}{\hat{\sigma}_{\hat{\rho}}}$ when $T = 500$ are

1%	5%	10%
-2.33	-1.64	-1.28

We have seen that the initial value z_0 does affect the finite sample distribution of Dickey-Fuller case 1 but does not affect case 2 and case 3. The value of c does affect the distribution of case 3: as c becomes smaller the distribution converges to the distribution of case 2. IMC has provided 10 pairs, the range of $\hat{c}$ of these 10 pairs is $(-0.01, 0.1)$. The absolute initial value z_0 of the 10 pairs is less than 1.5 for 9 of the 10 pairs. With one pair z_0 is 106. So we are interested in the power of the three tests for small values of c and z_0, but we will also look at large values of z_0.

We start with generating paths with $c = 0$ and $z_0 = 0$. In all following tables $T = 500$, $\sigma = 1$ and the number of generated paths is 1,000. Table 5.9 shows the number of rejections for the different tests and different values of ρ. For $\rho = 1$ we have simulated paths under the null hypothesis of case 1 and case 2, the number of rejections are in line with what we expected. The case 3 test does not perform very well, with $\rho = 1$ it rejects the null hypothesis that $\rho = 1$ 632 out of 1,000 times on the 10% level. For ρ just under 1, the case 1 test performs better than the case 2 test.

Table 5.9: Number of rejections, $c = 0$, $z_0 = 0$.

	Case 1			Case 2			Case 3		
ρ	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.9	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.95	986	1000	1000	746	966	997	1000	1000	1000
0.975	479	910	982	135	427	663	835	991	1000
0.99	73	310	528	22	116	205	338	771	918
0.995	41	144	274	21	77	148	231	617	783
1	10	48	96	12	54	105	171	456	632
1.01	0	0	0	0	0	0	1	4	12

Table 5.10 shows the number of rejections for generated paths with $c = 0$ and $z_0 = 100$. For $\rho = 1$ we simulated under the null hypothesis of case 1 and 2, the number of rejections for case 1 is small. This was expected because of figure 5.3. Again, the case 3 test does not perform very well when $\rho = 1$. The case 1 and 2 tests do perform well, with ρ slightly less than 1 they reject the null of an unit root.

Table 5.10: Number of rejections, $c = 0$, $z_0 = 100$.

	Case 1			Case 2			Case 3		
ρ	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.99	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.995	1000	1000	1000	406	683	805	874	974	995
1	8	26	47	10	52	106	157	461	631
1.01	0	0	0	0	0	0	0	0	0

Table 5.11 shows the number of rejections for $c = 0.1$ and $z_0 = 0$. For $\rho = 1$ we have simulated under the null hypothesis of case 3 but the case 3 test still rejects to many times. Because the null is fulfilled we expect the number of rejections to be around 10, 50 and 100 for the 1%, 5% and 10% levels respectively. In figure 5.8 we already saw that the case 3 test is dependent of the value of c, when $c = 0.1$ the distribution of the case 3 test statistic is shifted to the left compared to its asymptotic distribution. We see that with this setting the case 2 test performs more or less the same as with $c = 0$ and $z_0 = 0$ except when $\rho = 1$, in which case it rejects less. The null is not satisfied for the case 2 test, so this is not a bad outcome. The less rejections for $\rho = 1$ the better. The case 1 test performs less compared to case 2 test as well as the setting $c = 0$ and $z_0 = 0$.

Table 5.11: Number of rejections, $c = 0.1$, $z_0 = 0$.

	Case 1			Case 2			Case 3		
ρ	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.9	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.95	758	968	997	848	995	1000	999	1000	1000
0.975	111	434	675	157	476	729	831	996	1000
0.99	6	37	87	39	142	262	376	765	917
0.995	0	7	20	18	90	158	257	599	768
1	0	2	3	5	18	37	73	217	329
1.01	0	0	1	0	0	0	2	8	9

Table 5.12 shows the number of rejections for $c = 0.1$ and $z_0 = 100$. It is remarkable how good the case 1 test performs, it rejects almost every time when ρ is slightly below 1 and does not reject when $\rho \geq 1$ even though the null hypothesis is not satisfied. In section 5.2 was explained that this test basically fit a line through the origin and because the scatterplot starts around (100,100) it estimate ρ very accurately which makes the standard error relatively small. With this setting, we know there is an intercept of 0.1 but this is so small compared to the starting point of 100 that the test does not overestimate ρ too much. So when we generate path for $\rho < 1$, $\hat{\rho} - 1$ is negative and divided by the small standard error the test statistic is a large negative value, so the null is rejected. When generating paths with $\rho \geq 1$, $\hat{\rho}$ is always slightly above 1, so the test statistic is a large positive value, so the null is not rejected.

Table 5.12: Number of rejections, $c = 0.1$, $z_0 = 100$.

	Case 1			Case 2			Case 3		
ρ	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.975	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.99	1000	1000	1000	996	999	1000	1000	1000	1000
0.995	997	1000	1000	189	451	586	689	913	961
1	0	0	0	8	21	45	73	234	345
1.01	0	0	0	0	0	0	0	0	0

For illustration purposes, table 5.13 shows the the number of rejections for $c = 1$ and $z_0 = 0$. The value of c is now much larger than the values of $\hat{c}$ for the 10 pairs. We see that case 1 test lost all its power, the case 2 test performs well and the case 3 test is finally performing as it should when $\rho = 1$ and is very powerful.

Table 5.13: Number of rejections, $c = 1$, $z_0 = 0$.

	Case 1			Case 2			Case 3		
ρ	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.9	0	0	0	1000	1000	1000	1000	1000	1000
0.95	0	0	0	1000	1000	1000	1000	1000	1000
0.975	0	0	0	998	1000	1000	1000	1000	1000
0.99	0	0	0	1000	1000	1000	1000	1000	1000
0.995	0	0	0	997	1000	1000	1000	1000	1000
1	0	0	0	1	7	12	13	59	119
1.01	0	0	0	0	0	0	0	0	0

This section clearly indicates that the Dickey-Fuller case 3 test is not the one we should use when testing pairs for cointegration. Unfortunately it does not clearly distinguish case 1 and case 2. Case 1 performs better for $c = 0$, $z_0 = 0$ and $c = 0$, $z_0 = 100$ and $c = 0.1$, $z_0 = 100$, but case 2 performs better for $c = 0.1$, $z_0 = 0$ which is most seen in the 10 pairs. In the remainder of this report we will focus on the case 2 test because of Hamilton's view given in section 4.3 and because this section does not clearly indicate to do otherwise. Another possible reason to use case 2 instead of case 1 could be that that the first step of the Engle-Granger method, which is a linear regression to estimate α, influences the power of the two tests. This will be considered in chapter 6.

5.6 Augmented Dickey-Fuller test

So far we discussed the properties of the estimated coefficients for a first-order autoregression when there is a unit root. In this section we discuss the distribution of the estimated coefficients for a p-th order autoregression. Recall that the *Augmented Dickey-Fuller test* tests

$$H_0 : z_t \sim I(1) \text{ against } H_1 : z_t \sim I(0)\,, \tag{5.24}$$

when z_t is assumed to follow an AR(p) model

$$z_t = c + \phi_1 z_{t-1} + \cdots + \phi_p z_{t-p} + u_t\,, \tag{5.25}$$

where $u_t \sim \text{i.i.d}(0, \sigma^2)$.

This model can be written as

$$z_t = c + \rho z_{t-1} + \beta_1 \Delta z_{t-1} + \cdots + \beta_{p-1} \Delta z_{t-p+1} + u_t \,, \tag{5.26}$$

with

$$\begin{aligned} \rho &= \phi_1 + \phi_2 + \cdots + \phi_p \,, \\ \beta_i &= -(\phi_{i+1} + \cdots + \phi_p) \,, \quad \text{for } i = 1, \ldots, p-1 \,. \end{aligned}$$

The null hypothesis is that the autoregressive polynomial

$$1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p = 0 \,,$$

has exactly one unit root and all other roots lie outside the unit circle. The single unit root gives us:

$$1 - \phi_1 - \phi_2 - \cdots - \phi_p = 0$$

i.e., $\rho = 1$. This implies

$$1 - \phi_1 x - \cdots - \phi_p x^p = (1 - \beta_1 x - \cdots - \beta_{p-1} x^{p-1})(1 - x) \,. \tag{5.27}$$

Of the p values of x that make the left side of (5.27) zero, one is $x = 1$ and all other roots are assumed to be outside the unit circle. The same must be true for the right side as well, meaning all roots of

$$1 - \beta_1 x - \cdots - \beta_{p-1} x^{p-1} = 0 \,.$$

lie outside the unit circle. So, (5.24) is equivalent to

$$H_0 : \rho = 1 \text{ against } H_1 : \rho < 1 \,.$$

We are interested in the properties of test statistic $S = \frac{\hat{\rho}-1}{\hat{\sigma}_{\hat{\rho}}}$ in the three cases:

Case 1: The true process of z_t is (5.26) with $c = 0$ and $\rho = 1$, the model estimated is (5.26) except for c.

Case 2: The true process of z_t is (5.26) with $c = 0$ and $\rho = 1$, the model estimated is (5.26).

Case 3: The true process of z_t is (5.26) with $c \neq 0$ and $\rho = 1$, the model estimated is (5.26).

We can derive the asymptotic properties in a similar manner as in the preceding sections. To keep this section from being to tedious, we only derive the properties for case 2. We state the outcomes for case 1 and case 3 at the end of this section, the derivations can be found in Hamilton [6].

Before deriving the properties for Augmented Dickey-Fuller case 2, we first state a proposition.

Proposition 2:
Let $v_t = \sum_{j=0}^{\infty} \theta_j u_{t-j}$, *where* $\sum_{j=0}^{\infty} j \cdot |\, \theta_j| < \infty$ *and* $\{u_t\}$ *is an i.i.d sequence with mean zero, variance* σ^2, *and finite fourth moment. Define*

$$\gamma_j = E(v_t v_{t-j}) = \sigma^2 \sum_{s=0}^{\infty} \theta_s \theta_{s+j}\,, \qquad \textit{for } j = 0, 1, \ldots\,, \tag{5.28}$$

$$\lambda = \sigma \sum_{j=0}^{\infty} \theta_j\,, \tag{5.29}$$

$$z_t = v_1 + v_2 + \cdots + v_t\,, \qquad \textit{for } t = 1, 2, \ldots, T\,, \tag{5.30}$$

with $z_0 = 0$. *Then*

$$(i) \quad T^{-1} \textstyle\sum_{t=1}^{T} v_t v_{t-j} \xrightarrow{P} \gamma_j \qquad \text{for } j = 0, 1, \ldots$$

$$(ii) \quad T^{-1} \textstyle\sum_{t=1}^{T} z_{t-1} v_{t-j} \xrightarrow{D} \begin{cases} \frac{1}{2}\left(\lambda^2 W(1)^2 - \gamma_0\right) & \text{for } j = 0 \\ \frac{1}{2}\left(\lambda^2 W(1)^2 - \gamma_0\right) & \\ \quad +\gamma_0 + \cdots + \gamma_{j-1} & \text{for } j = 1, 2, \ldots \end{cases}$$

$$(iii) \quad T^{-3/2} \textstyle\sum_{t=1}^{T} z_{t-1} \xrightarrow{D} \lambda \int_0^1 W(r)\, dr\,.$$

$$(iv) \quad T^{-2} \textstyle\sum_{t=1}^{T} z_{t-1}^2 \xrightarrow{D} \lambda^2 \int_0^1 W(r)^2\, dr\,.$$

$$(v) \quad T^{-1/2} \textstyle\sum_{t=1}^{T} v_t \xrightarrow{D} \lambda\, W(1)\,.$$

$$(vi) \quad T^{-1} \textstyle\sum_{t=1}^{T} z_{t-1} u_t \xrightarrow{D} \frac{1}{2}\sigma\lambda(W(1)^2 - 1)\,.$$

The proof of this proposition can also found in [6].

Asymptotic distribution ADF case 2
We assume that the sample is of size $T + p$, $(z_{-p+1}, z_{-p+2}, \ldots, z_T)$ and the model is

$$\begin{aligned} z_t &= c + \rho z_{t-1} + \beta_1 \Delta z_{t-1} + \cdots + \beta_{p-1} \Delta z_{t-p+1} + u_t \\ &= \mathbf{x}_t' \boldsymbol{\beta} + u_t \,, \end{aligned}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{p-1}, c, \rho)$ and $\mathbf{x}_t = (\Delta z_{t-1}, \Delta z_{t-2}, \ldots, \Delta z_{t-p+1}, 1, z_{t-1})$.

Under the null hypothesis of exactly one unit root and the assumption that z_t follows above AR(p) model with $c = 0$ and $\rho = 1$, we show that z_t behaves like the variable z_t in proposition 2. Because z_t is integrated of order one and

$$v_t = \Delta z_t \,,$$

v_t is stationary and follows an AR($p-1$) model:

$$\begin{aligned} \Delta z_t &= \beta_1 \Delta z_{t-1} + \cdots + \beta_{p-1} \Delta z_{t-p+1} + u_t \,, \\ \Leftrightarrow v_t &= \beta_1 v_{t-1} + \cdots + \beta_{p-1} v_{t-p+1} + u_t \,. \end{aligned}$$

The autoregressive polynomial of v_t is

$$\Phi(x) = 1 - \beta_1 x - \cdots - \beta_{p-1} x^{p-1} \,,$$

and all roots of $\Phi(x) = 0$ are outside the unit circle because v_t is stationary and we assume it is causal, like all other autoregressive models in this report. Then v_t has a MA(∞) representation

$$v_t = \sum_{j=0}^{\infty} \theta_j u_{t-j}$$

which polynomial is

$$\Theta(x) = 1 + \theta_1 x + \theta_2 x^2 + \cdots$$

and because $\Phi(x)$ and $\Theta(x)$ are polynomials, we have

$$\Theta(x) = \frac{1}{\Phi(x)} \,.$$

All $p-1$ roots, which is a finite number of roots, of $\Phi(x)$ are outside the unit circle, so there exists an $\varepsilon > 0$ such that the modus of all roots are larger than $1+\varepsilon$, so $\Phi(x) \neq 0$ for $|x| < 1+\varepsilon$. Within the radius of convergence $1+\varepsilon$, the analytic function $\Theta(x)$ is differentiable:

$$\Theta'(x) = \sum_{j=1}^{\infty} j\theta_j x^{j-1}\,.$$

And because it is absolutely convergent within its radius of convergence, particularly in point 1, we have

$$\sum_{j=1}^{\infty} j \cdot |\theta_j| < \infty\,.$$

This shows we can use proposition 2 without making any further assumptions.

The deviation of the OLS estimate $\hat{\boldsymbol{\beta}}$ from the true value $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'\right]^{-1} \left[\sum_{t=1}^{T} \mathbf{x}_t u_t\right]\,. \tag{5.31}$$

With $v_t = z_t - z_{t-1}$, the terms in (5.31) are

$$\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' = \begin{bmatrix} \sum v_{t-1}^2 & \cdots & \sum v_{t-1}v_{t-p+1} & \sum v_{t-1} & \sum v_{t-1}z_{t-1} \\ \sum v_{t-2}v_{t-1} & \cdots & \sum v_{t-2}v_{t-p+1} & \sum v_{t-2} & \sum v_{t-2}z_{t-1} \\ \vdots & \cdots & \vdots & \vdots & \vdots \\ \sum v_{t-p+1}v_{t-1} & \cdots & \sum v_{t-p+1}^2 & \sum v_{t-p+1} & \sum v_{t-p+1}z_{t-1} \\ \sum v_{t-1} & \cdots & \sum v_{t-p+1} & T & \sum z_{t-1} \\ \sum z_{t-1}v_{t-1} & \cdots & \sum z_{t-1}v_{t-p+1} & \sum z_{t-1} & \sum z_{t-1}^2 \end{bmatrix},$$

$$\sum_{t=1}^{T} \mathbf{x}_t u_t = \begin{bmatrix} \sum v_{t-1}u_t \\ \vdots \\ \sum v_{t-p+1}u_t \\ \sum u_t \\ \sum z_{t-1}u_t \end{bmatrix}.$$

Like in the derivation of DF case 2 we need a scaling matrix, in this section we use the following $(p+1)\times(p+1)$ scaling matrix:

$$\mathbf{Y} = \begin{bmatrix} \sqrt{T} & 0 & \cdots & 0 & 0 \\ 0 & \sqrt{T} & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \sqrt{T} & 0 \\ 0 & 0 & \cdots & 0 & T \end{bmatrix}$$

With multiplying (5.31) by the scaling matrix Y and using (5.7) we get

$$\mathbf{Y}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left\{ \mathbf{Y}^{-1} \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \right] \mathbf{Y}^{-1} \right\}^{-1} \left\{ \mathbf{Y}^{-1} \left[\sum_{t=1}^{T} \mathbf{x}_t u_t \right] \right\} \tag{5.32}$$

Consider the matrix $\mathbf{Y}^{-1} \sum \mathbf{x}_t \mathbf{x}_t' \mathbf{Y}^{-1}$. Elements in the upper left $(p \times p)$ block of $\sum \mathbf{x}_t \mathbf{x}_t'$ are divided by T, the first p elements of the $(p+1)$th row or $(p+1)$th column are divided by $T^{3/2}$ and the element at the lower right corner is divided by T^2. Moreover,

$$\begin{array}{rcll} T^{-1} \sum v_{t-i} v_{i-j} & \xrightarrow{P} & \gamma_{|i-j|} & \text{from proposition } 2(i)\,, \\ T^{-1} \sum v_{t-j} & \xrightarrow{P} & E(v_{t-j}) = 0 & \text{from the law of large numbers}\,, \\ T^{-3/2} \sum z_{t-1} v_{t-j} & \xrightarrow{P} & 0 & \text{from proposition } 2(ii)\,, \\ T^{-3/2} \sum z_{t-1} & \xrightarrow{D} & \lambda \int W(r)\, dr & \text{from proposition } 2(iii)\,, \\ T^{-2} \sum z_{t-1}^2 & \xrightarrow{D} & \lambda^2 \int W(r)^2\, dr & \text{from proposition } 2(iv)\,, \end{array}$$

where

$$\begin{aligned} \gamma_j &= E(\Delta z_t\, \Delta z_{t-j})\,, \\ \lambda &= \sigma/(1 - \beta_1 - \cdots - \beta_{p-1})\,, \\ \sigma^2 &= E(u_t^2)\,, \end{aligned}$$

and the integral sign denotes integration over r from 0 to 1. Thus,

$$\begin{aligned} \mathbf{Y}^{-1} \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \right] \mathbf{Y}^{-1} &\xrightarrow{D} \begin{bmatrix} \gamma_0 & \cdots & \gamma_{p-2} & 0 & 0 \\ \vdots & \cdots & \vdots & \vdots & \vdots \\ \gamma_{p-2} & \cdots & \gamma_0 & 0 & 0 \\ 0 & \cdots & 0 & 1 & \lambda \int W(r)\, dr \\ 0 & \cdots & 0 & \lambda \int W(r)\, dr & \lambda^2 \int W(r)^2\, dr \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{V} & 0 \\ 0 & \mathbf{Q} \end{bmatrix}, \end{aligned}$$

with

$$\begin{aligned}\mathbf{V} &= \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-2} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-3} \\ \vdots & \vdots & \cdots & \vdots \\ \gamma_{p-2} & \gamma_{p-3} & \cdots & \gamma_0 \end{bmatrix}, \\ \mathbf{Q} &= \begin{bmatrix} 1 & \lambda \int W(r)\, dr \\ \lambda \int W(r)\, dr & \lambda^2 \int W(r)^2\, dr \end{bmatrix}. \end{aligned} \tag{5.33}$$

Next, consider the second term in the right side of(5.32)

$$\mathbf{Y}^{-1} \begin{bmatrix} \sum_{t=1}^{T} \mathbf{x}_t u_t \end{bmatrix} = \begin{bmatrix} T^{-1/2} \sum v_{t-1} u_t \\ \vdots \\ T^{-1/2} \sum v_{t-p+1} u_t \\ T^{-1/2} \sum u_t \\ T^{-1} \sum z_{t-1} u_t \end{bmatrix}. \tag{5.34}$$

The first $p-1$ elements of this vector satisfy the central limit theorem. This is because these elements are $\sqrt{T}$ times the sample mean of a martingale difference sequence whose covariance matrix is $\sigma^2\mathbf{V}$, but this is not discussed further. The result is

$$\begin{bmatrix} T^{-1/2} \sum v_{t-1} u_t \\ \vdots \\ T^{-1/2} \sum v_{t-p+1} u_t \end{bmatrix} \xrightarrow{D} \mathbf{h}_1 \sim N(\mathbf{0}, \sigma^2 \mathbf{V}).$$

The distribution of the last two elements in (5.34) can be obtained from statements (v) en (vi) of proposition 2:

$$\begin{bmatrix} T^{-1/2} \sum u_t \\ T^{-1} \sum z_{t-1} u_t \end{bmatrix} \xrightarrow{D} \mathbf{h}_2 \sim \begin{bmatrix} \sigma W(1) \\ \frac{1}{2}\sigma\lambda(W(1)^2 - 1) \end{bmatrix}. \tag{5.35}$$

This gives that the deviation of the OLS estimate from its true value is

$$\mathbf{Y}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} \begin{bmatrix} \mathbf{V} & 0 \\ 0 & \mathbf{Q} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{V}^{-1}\mathbf{h}_1 \\ \mathbf{Q}^{-1}\mathbf{h}_2 \end{bmatrix}. \tag{5.36}$$

The last two elements of $\boldsymbol{\beta}$ are c and ρ, which are the constant term and the coefficient on the $I(1)$ regressor, z_{t-1}. From (5.33),(5.35) and (5.36), their limiting distribution is given by

$$\begin{aligned} &\begin{bmatrix} T^{1/2} & 0 \\ 0 & T \end{bmatrix} \begin{bmatrix} \hat{c} \\ \hat{\rho}-1 \end{bmatrix} \xrightarrow{D} \\ &\begin{bmatrix} \sigma & 0 \\ 0 & \sigma/\lambda \end{bmatrix} \begin{bmatrix} 1 & \int W(r)\,dr \\ \int W(r)\,dr & \int W(r)^2\,dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \frac{1}{2}(W(1)^2-1) \end{bmatrix} \end{aligned} \tag{5.37}$$

The t test statistic S of the null hypothesis that $\rho = 1$ is

$$S = \frac{\hat{\rho}-1}{\hat{\sigma}_{\hat{\rho}}} = \frac{\hat{\rho}-1}{\{r_T^2 \mathbf{e}(\sum \mathbf{x}_t\mathbf{x}_t')^{-1}\mathbf{e}\}^{1/2}},$$

where $\mathbf{e}$ denotes a $p+1$ vector with unity in the last postition and zeros elsewhere. Multiplying the numerator and the denominator by T results in

$$S = \frac{T(\hat{\rho}-1)}{\{r_T^2 \mathbf{e}\mathbf{Y}(\sum \mathbf{x}_t\mathbf{x}_t')^{-1}\mathbf{Y}\mathbf{e}\}^{1/2}}.$$

But

$$\begin{aligned} \mathbf{e}\mathbf{Y}\left(\sum \mathbf{x}_t\mathbf{x}_t'\right)^{-1}\mathbf{Y}\mathbf{e} &= \mathbf{e}\left\{\mathbf{Y}\left(\sum \mathbf{x}_t\mathbf{x}_t'\right)\mathbf{Y}\right\}^{-1}\mathbf{e} \\ &\xrightarrow{D} \mathbf{e}'\begin{bmatrix} \mathbf{V}^{-1} & 0 \\ 0 & \mathbf{Q}^{-1} \end{bmatrix}\mathbf{e} \\ &= \frac{1}{\lambda^2\left\{\int W(r)^2\,dr - \left(\int W(r)\,dr\right)^2\right\}}. \end{aligned} \tag{5.38}$$

By (5.37) we have

$$T(\hat{\rho}-1) \xrightarrow{D} (\sigma/\lambda)\frac{\frac{1}{2}(W(1)^2-1) - W(1)\int W(r)\,dr}{\int W(r)^2\,dr - \left(\int W(r)\,dr\right)^2}. \tag{5.39}$$

Using (5.38) and (5.39) together with $r_T^2 \xrightarrow{P} \sigma^2$, we finally get

$$S \xrightarrow{D} \frac{\frac{1}{2}(W(1)^2-1) - W(1)\int W(r)\,dr}{\left(\int W(r)^2\,dr - \left[\int W(r)\,dr\right]^2\right)^{1/2}}, \tag{5.40}$$

which is exactly the same as the asymptotic distribution of the Dickey-Fuller case 2 test statistic. So the critical values are the same as in table 5.4 in section 5.3 without making any corrections for the fact that lagged values of Δz_t are included in the regression. This is also true for the other cases, Augmented Dickey-Fuller case 1 test statistic has the same asymptotic distribution as Dickey-Fuller case 1 and ADF case 3 the same as DF case 3.

Like in the preceding sections we can simulate the density of the test statistic for finite sample sizes, we show the results for the case 2 test when $p = 2$. We simulate for different values of σ when $T = 500$ and $\beta_1 = -0.1$, and naturally $\rho = 1$, $c = 0$. We took this value for β_1 because this value is seen a few times in the 10 pairs IMC provided. The estimated densities of 5,000 simulated test statistics for $\sigma^2 = 1, 5, 10$ are shown in figure 5.11. Also the asymptotic density we found for the case 2 test, figure 5.4, is plotted with a dashed line. The different graphs coincide nicely. With this setting, the 'original' AR model with lagged terms instead of differenced terms is:

$$z_t = 0.9z_{t-1} + 0.1z_{t-2} + u_t \,.$$

The autoregressive polynomial

$$1 - 0.9x - 0.1x^2 = 0 \,,$$

has roots 1 and -10, so the assumption of exactly one unit root is fulfilled.

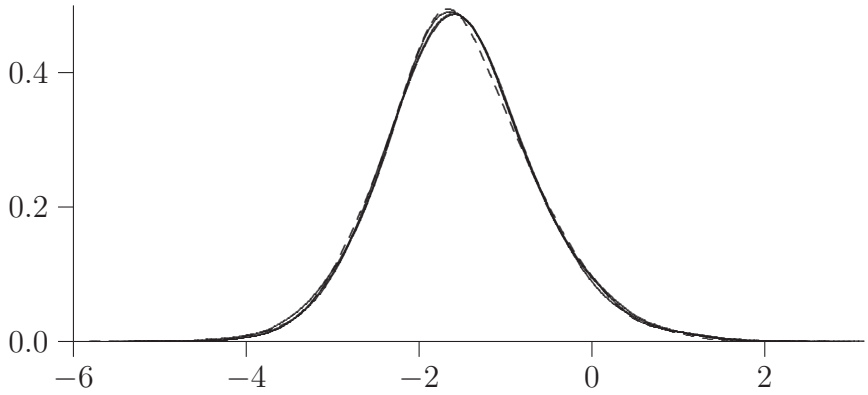

Figure 5.11: Estimated density of ADF case 2 for different σ^2, $T = 500$ and $\beta_1 = -0.1$.

To see what influence β_1 has we also vary its value while keeping σ^2 fixed at 1. The results for $\beta_1 = -0.9, -0.5, 0, 0.5, 0.9, 1$ are shown in figure 5.12. The awkward graph corresponds with $\beta_1 = 1$, the 'original' AR model is:

$$z_t = 2z_{t-1} - z_{t-2} + u_t$$

The autoregressive polynomial

$$1 - 2x + x^2 = 0$$

has twice root 1, so there are two unit roots. That is probably why the graph for $\beta_1 = 1$ does not look like the other ones. For the other values of β_1 the assumption of exactly one unit root is fulfilled. The values of β_1 for the 10 pairs when an AR(2) model is fit on the spread process are in a range of (-0.25,0.1).

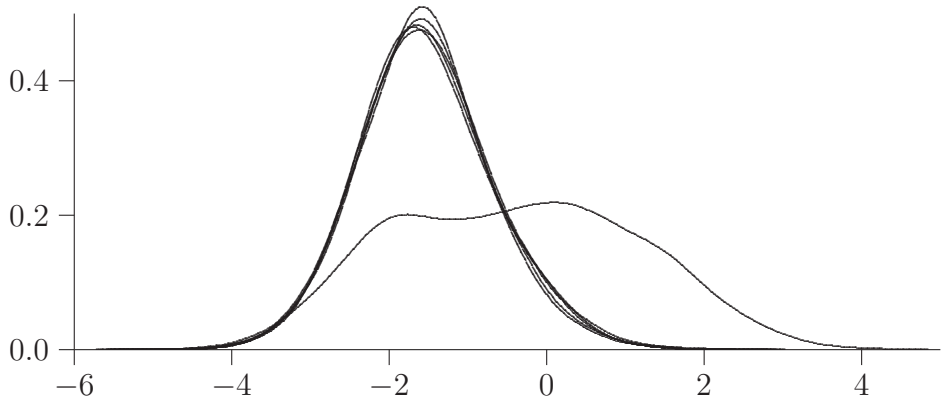

Figure 5.12: Estimated density of ADF case 2 for different β_1, $T = 500$ and $\sigma^2 = 1$.

We also show the results for two higher order models. First $p = 3$, figure 5.13 shows the estimated densities for three different settings of β_1 and β_2. We take β_2 equal to -0.1 and β_1 is -0.2, -0.1 and 0.1 successively, these are also values seen with the 10 pairs. For these values the autoregressive polynomial has exactly one unit root and the other roots are outside the unit circle, so the null hypothesis is satisfied. Also the graph of figure 5.4 is displayed, again they coincide nicely.

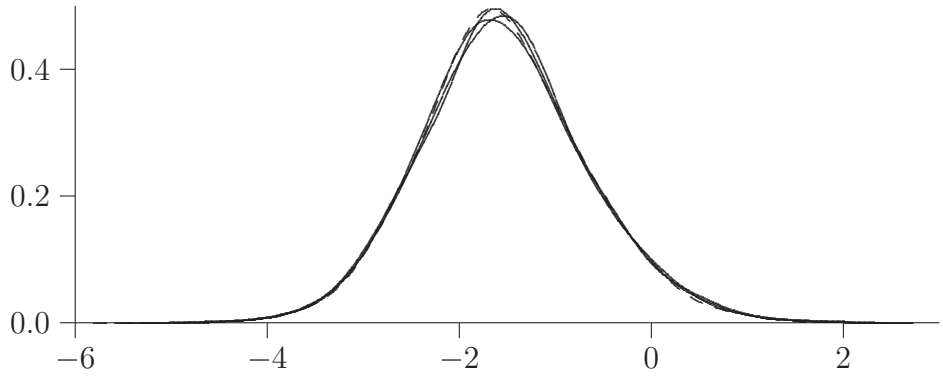

Figure 5.13: Estimated density of ADF case 2 for different β_1 and β_2, $T = 500$ and $\sigma^2 = 1$.

Lastly, we consider $p = 5$. Figure 5.14 shows the estimated densities for three parameter settings:

$$\begin{array}{llllll} \text{Setting 1}: & \beta_1 = -0.3 & \beta_2 = -0.2 & \beta_3 = -0.1 & \beta_4 = -0.05 \\ \text{Setting 2}: & \beta_1 = -0.1 & \beta_2 = -0.1 & \beta_3 = -0.1 & \beta_4 = -0.1 \\ \text{Setting 3}: & \beta_1 = 0.1 & \beta_2 = -0.1 & \beta_3 = 0.1 & \beta_4 = 0.05 \end{array}$$

These three settings also represents most of the 10 pairs and that the null hypothesis is satisfied was checked with Maple. We see that the densities show more dispersion and do not coincide with the asymptotic density as nicely as for the lower order models above.

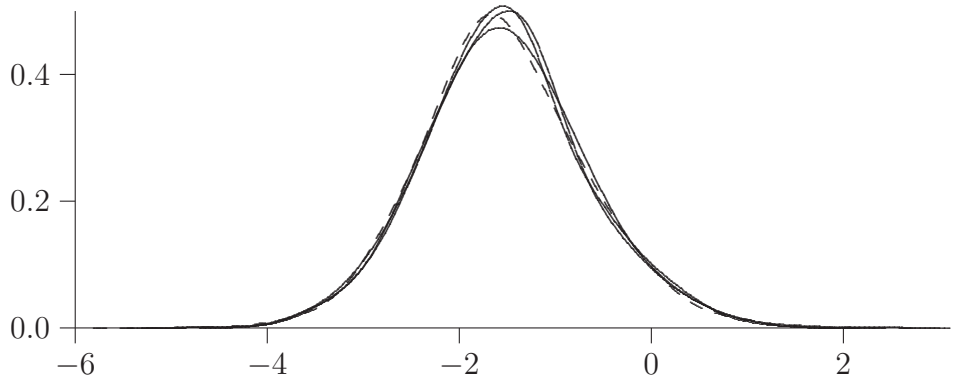

Figure 5.14: Estimated density of ADF case 2 for different β_1, $T = 500$ and $\sigma^2 = 1$.

In the next section we look again at all these models to see if the power of the Augmented Dickey-Fuller case 2 test is influenced by the value of p.

5.7 Power of the Augmented Dickey-Fuller case 2 test

In this section we briefly look at how much influence the value of p, the order of the autoregressive model, has on the power of the Augmented Dickey-Fuller case 2 test. As in section 5.5 we generate 1,000 paths for different values of ρ and see how many times the test rejects the null hypothesis of a unit root.

We start with $p = 2$, so the paths are generated according to the model

$$z_t = \rho z_{t-1} + \beta_1 \Delta z_{t-1} + u_t \,,$$

for we take u_t i.i.d standard Gaussian random variables, sample size $T = 500$ and $z_0 = 0$. Table 5.14 shows the number of rejections for several values of ρ and β_1. We use values of β_1 which are seen with the 10 pairs IMC provided. When $\rho = 1$ the null hypothesis is satisfied, the other roots of the autoregressive polynomial lie outside the unit circle, they are -4, -10 and 10 respectively for $\beta_1 = -0.25, -0.1, 0.1$. We see that under the null hypothesis the test behaves as expected. The power is quite similar to the Dickey-Fuller case 2 test in table 5.9. The power is better for the positive value of β_1.

Table 5.14: Number of rejections, $p = 2$.

	$\beta_1 = -0.25$			$\beta_1 = -0.1$			$\beta_1 = 0.1$		
ρ	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.9	986	1000	1000	998	1000	1000	1000	1000	1000
0.95	447	845	949	582	919	976	792	980	997
0.975	68	297	491	104	395	589	179	502	708
0.99	18	75	162	13	89	199	28	117	232
0.995	15	70	138	13	66	137	10	74	144
1	9	49	97	7	45	88	9	42	89
1.01	0	1	4	0	1	2	0	1	2

Table 5.15 shows the number of rejections for the AR(3) model:

$$z_t = \rho z_{t-1} + \beta_1 \Delta z_{t-1} + \beta_2 \Delta z_{t-2} + u_t \, .$$

We use the same values of β_1 and β_2 as in the previous section: $\beta_2 = -0.1$ and $\beta_1 = -0.25 - 0.1, 0.1$. For these values of β_1 and β_2 and when $\rho = 1$ the null hypothesis of exactly one unit root is satisfied. The table does not indicate that the power when $p = 3$ is much less than the power of the test when $p = 2$.

Lastly, table 5.16 shows the number of rejections for the AR(5) model:

$$z_t = \rho z_{t-1} + \beta_1 \Delta z_{t-1} + \beta_2 \Delta z_{t-2} + \beta_3 \Delta z_{t-3} + \beta_4 \Delta z_{t-4} + u_t \, ,$$

where we used the three parameter settings from the previous section. This table indicates that the power of the test with $p = 5$ is less than the power of the test for smaller values of p. Specially the first setting of parameters shows that the power of the test is less for $p = 5$.

Table 5.15: Number of rejections, $p = 3$ and $\beta_2 = -0.1$.

	$\beta_1 = -0.2$			$\beta_1 = -0.1$			$\beta_1 = 0.1$		
ρ	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.9	978	998	1000	986	1000	1000	1000	1000	1000
0.95	406	765	900	487	851	944	684	941	984
0.975	63	283	482	86	287	473	129	399	615
0.99	17	83	163	22	89	177	27	113	219
0.995	13	74	134	13	71	136	19	67	128
1	12	51	114	13	62	104	14	58	104
1.01	0	0	2	0	2	3	1	5	6

Table 5.16: Number of rejections, $p = 5$.

	setting 1			setting 2			setting 3		
ρ	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.9	782	975	999	924	995	1000	1000	1000	1000
0.95	198	527	717	307	708	874	731	904	992
0.975	37	176	313	69	253	418	186	531	743
0.99	15	83	152	15	97	179	21	114	228
0.995	14	52	115	15	52	124	14	83	163
1	12	48	99	9	50	101	12	47	97
1.01	0	1	10	0	4	9	0	1	1

Chapter 6

Engle-Granger method

In the previous chapter we derived and simulated the properties of the Dickey-Fuller test. In this chapter we would like to find the properties of the Engle-Granger method, which we use for testing stock price data for cointegration. As explained in chapter 4 the Engle-Granger method consists of two steps, a linear regression followed by the Dickey-Fuller test on the residuals of this regression. The main question is, are the critical values of the Engle-Granger method the same as those for the Dickey-Fuller test. In the first section the critical values of Engle-Granger method are found by simulating price processes x_t and y_t with random walks, then all the assumptions of the method are satisfied. The second section also simulates the critical values but now the model from section 4.2 is used for simulating the processes x_t and y_t. Then not all assumptions are completely satisfied, because x_t and y_t are not strictly integrated of order one. The third section finds the critical values with bootstrapping from real data. In the last section we simulate cointegrated price processes x_t and y_t with the alternative method from section 4.5 and find out whether the Engle-Granger method recognizes them as cointegrated. The main focus of this chapter is on the case 2 test, but we will also compare the power of this test with the case 1 test.

6.1 Engle-Granger simulation with random walks

The Engle-Granger assumes we have two prices processes $\{x_t, y_t\}_{t=0}^{T}$ where each individually is integrated of order one, $I(1)$. Then x_t and y_t are cointegrated if there exists a linear combination that is stationary:

$$x_t, y_t \text{ are cointegrated} \iff \exists \alpha, \alpha_0 \text{ such that } y_t - \alpha x_t - \alpha_0 = \varepsilon_t \sim I(0)\,.$$

As described in chapter 4, we prefer to set $\alpha_0 = 0$ because of the cash-neutral aspect. So the Engle-Granger method boils down to

- Estimate α with OLS: $\hat{\alpha} = \sum_{t=0}^{T} x_t y_t / \sum_{t=0}^{T} x_t^2$.
- Calculate spread $e_t = y_t - \hat{\alpha} x_t$.
- Test e_t for stationarity with ADF case 2 test.

In order to simulate pairs of data that are certain to be cointegrated and certain to be not, we like to simulate x_t and generate y_t belonging to x_t such that the spread process is an AR(p) process. This is because the Dickey-Fuller test assumes that the input series, in our case the spread process, is an AR(p) process. The process x_t has to be integrated of order one, then the most simple model for x_t is a random walk:

$$x_t = x_{t-1} + u_t\,, \tag{6.1}$$

with u_t i.i.d $N(0, \sigma_x^2)$ variables and x_0 an initial value. Then the difference $x_t - x_{t-1}$ is white noise, so $x_t \sim I(1)$. Now we have x_t, we like to generate y_t such that $y_t - \alpha x_t$ is AR(p) for some α and some p.

In this section we look at a few different settings, but only for $p = 1$ and find out whether the distribution and power of the Engle-Granger test statistic differs from the earlier derived distribution and power of the Dickey-Fuller case 2 test statistic. First, this is done under the null hypothesis of the DF case 2 test, this means that there is no constant in the spread process but the constant is estimated. Second, when there is a small constant present in the spread process. Last, for $p = 1$, we generate y_t with α_0 but do not regress on a constant to find out whether this is still cointegrated according to the Engle-Granger method.

AR(1) under the null hypothesis of DF case 2
We want the spread process to be an AR(1) process:

$$y_t - \alpha x_t = \varepsilon_t = \beta_0 + \beta\varepsilon_{t-1} + \eta_t \,,$$

for $\{\eta_t\}$ we take i.i.d $N(0, \sigma_\eta^2)$ variables. Then we can generate y_t like:

$$\begin{aligned} y_t &= \alpha x_t + \beta_0 + \beta\left(y_{t-1} - \alpha x_{t-1}\right) + \eta_t \,, \qquad \text{for } t = 1, \ldots, \,, & (6.2)\\ y_0 &= \alpha x_0 \,. & (6.3) \end{aligned}$$

Under the null hypothesis of the Dickey-Fuller case 2 there is a unit root and no constant in the spread process, $\beta = 1$ and $\beta_0 = 0$. The processes x_t and y_t are cointegrated if we take $\beta < 1$. Figure 6.1 shows a sample path for x and y when $\beta = 1$ and figure 6.2 for $\beta = 0.5$, with both graphs $\alpha = 0.8$, $\beta_0 = 0$, $x_0 = 25$, $\sigma_x^2 = \sigma_\eta^2 = 1$ and $T = 500$.

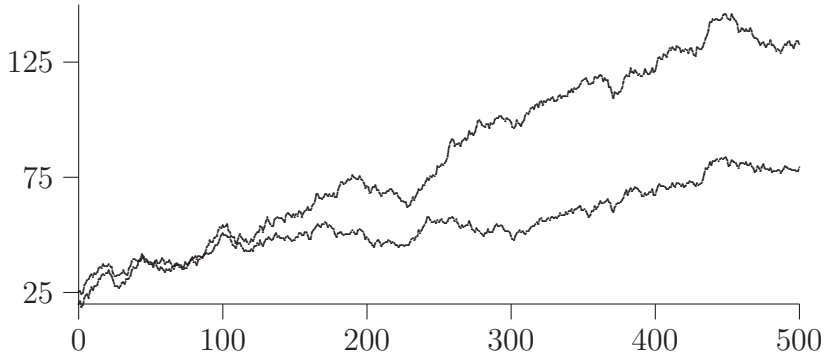

Figure 6.1: Pair x, y not cointegrated.

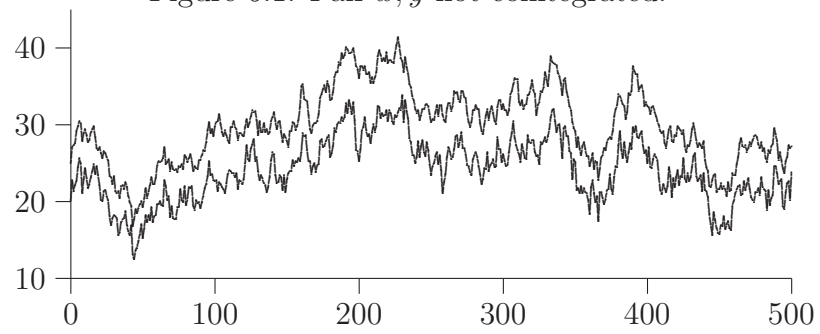

Figure 6.2: Pair x, y cointegrated.

To see whether the critical values of Engle-Granger are more or less the same as for Dickey-Fuller case 2, i.e. to see if estimating α has an affect on the critical values, we simulate a lot of paths x_t and y_t under the null hypothesis and calculate the test statistic S. The procedure is:

- Simulate x_t.
- Generate y_t according to (6.2).
- Calculate $\hat{\alpha}$.
- Calculate the spread $e_t = y_t - \hat{\alpha}x_t$.
- Calculate the spread $e_t = y_t - \hat{\alpha}x_t$
- DF-test on the spread:
  Estimate β with OLS.
  Calculate the OLS standard error for β: $\hat{\sigma}_{\hat{\beta}}$.
  Calculate test statistic $S = (\hat{\beta} - 1)/\hat{\sigma}_{\hat{\beta}}$.
- Repeat all this 5,000 times.

Then we estimate the density of the simulated test statistics, again with a Gaussian kernel estimator. This we can compare to the density we found for the Dickey-Fuller case 2 test in chapter 5.

Figure 6.3 shows the estimated densities for different values of α for $T = 500$, $x_0 = 25$, $\sigma_x^2 = \sigma_\eta^2 = 1$ and $\beta_0 = 0$. The figure also shows the Dickey-Fuller case 2 from figure 5.4. Figure 6.4 and 6.5 show estimated densities for different values of σ_x^2 and σ_η^2 respectively, for the same parameters as above and $\alpha = 1$. When the null hypothesis is completely satisfied, that is $\beta = 1$ and $\beta_0 = 0$, these densities look a lot like the density for the Dickey-Fuller case 2 density. So it looks like the preceding step to the DF test, namely estimating α, does not really affect the critical values. To see if the power of the test is not affected by the preceding step, table 6.1 shows the number of rejections for different values of β and α. It is clear from the table that the power of the Engle-Granger method is not dependent of the value of α. This table should be compared with the columns of case 2 of table 5.9, because there is no constant and the initial value of the spread process is 0. We see

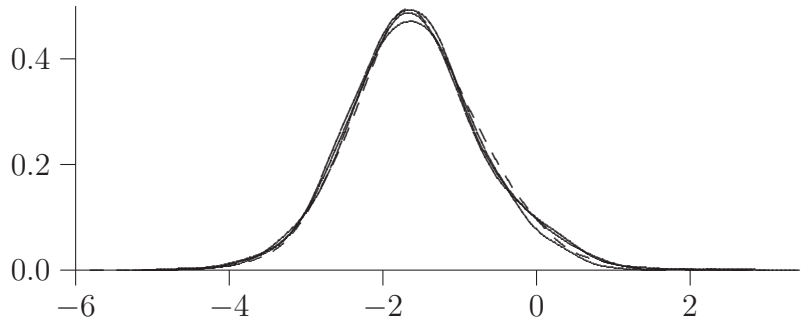

Figure 6.3: $AR(1)$ Estimated density for EG test statistic, $\alpha = 0.1, 0.5, 1$.

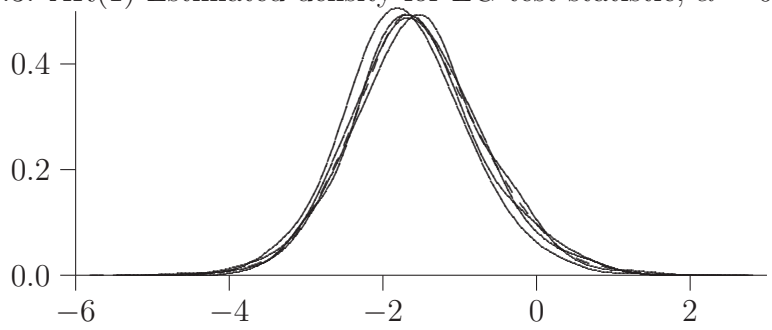

Figure 6.4: $AR(1)$ Estimated density of EG test statistic, $\sigma_x^2 = 0.1, 0.5, 1, 5$.

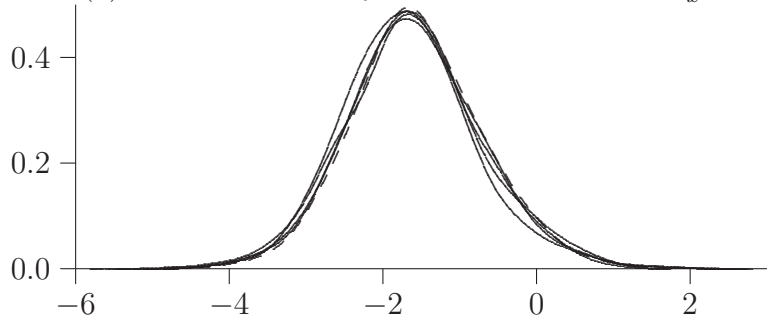

Figure 6.5: $AR(1)$ Estimated density of EG test statistic, $\sigma_\eta^2 = 0.1, 0.5, 1, 5$.

that there is practically no difference between these columns and table 6.1, which indicates that the power of the Engle-Granger method is as good as that of the Dickey-Fuller test. The estimation of α does not have a negative influence on the power of the test, which is a nice property.

Table 6.1: Number of rejections, AR(1) and $\beta_0 = 0$

	$\alpha = 1$			$\alpha = 0.5$			$\alpha = 0.1$		
β	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.9	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.95	722	975	1000	724	964	993	717	961	993
0.975	169	485	686	164	466	683	145	469	691
0.99	29	133	248	38	123	242	36	143	254
0.995	14	67	149	17	75	161	18	87	163
1	14	52	105	8	51	109	10	52	115
1.01	0	0	3	1	1	2	1	4	6

AR(1) with constant in spread

When testing for cointegration with the Engle-Granger method, we use the Dickey-Fuller case 2 test which assumes there is no constant in the spread process but does estimate one. It is interesting to see what happens to the properties of the Engle-Granger method if we do include a constant, β_0, in the spread process. From tables 5.11 and 5.12 we already saw that the power of the Dickey-Fuller case 2 test was not really affected by a small constant, when there was no unit root. But the number of rejections when there was a unit root were a bit small. To be more precise, for increasing values of the constant the Dickey-Fuller case 2 test statistic converges to the case 3 statistic, standard Gaussian, as explained in section 5.4. But with the Engle-Granger method we do a preceding step, we estimate α, maybe this first step has a restraining influence on the shift. Figure 6.6 shows the estimated density of the Engle-Granger test statistic when there is a small constant, $\beta_0 = 0.1$, in the spread process. The dashed line is the density when $\beta_0 = 0$. With both graphs the paths were generated with $T = 500$, $\alpha = 1$, $\sigma_x = 1$, $\sigma_\eta = 1$ and $\beta = 1$. There is a shift to the right, as we could expect from the properties of the Dickey-Fuller test. However, the dotted line is the density when $\beta_0 = 1000$, so there is a restraining influence on the shift. The Engle-Granger test statistic does not converge to the DF case 3 statistic for large constants. Although this is nice, for small values of β_0 we still have the same shift as for the DF case 2 statistic, so the first step in the Engle-Granger method does not have a big influence. This can also be observed in table 6.2, it shows the number of rejections for different values of β when $\beta_0 = 0.1$. The power of the Engle-Granger test is rather close to the

power of the Dickey-Fuller case 2 test with a small constant, as seen in table 5.11. So it looks like the Engle-Granger test statistic has the same properties as the Dickey-Fuller case 2 test statistic, for small constants.

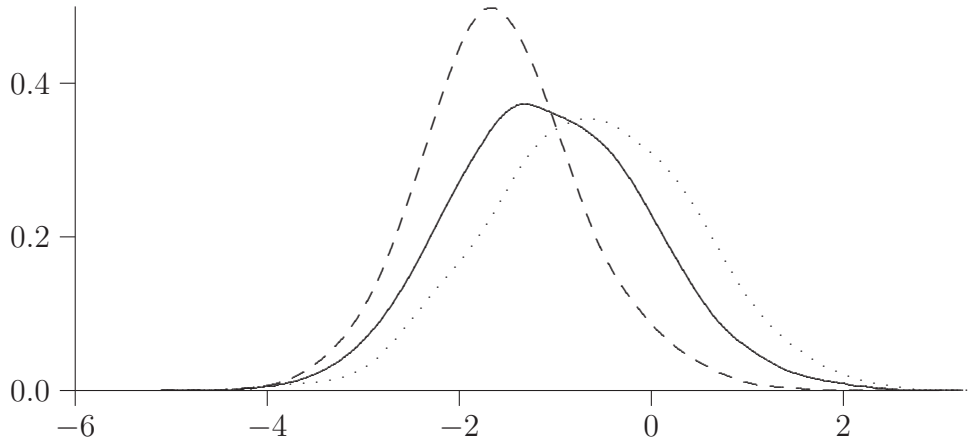

Figure 6.6: Estimated density for EG test statistic, $\beta_0 = 0.1$.

Table 6.2: Number of rejections of the null hypothesis, $\beta_0 = 0.1$.

β	1%	5%	10%
0.9	1000	1000	1000
0.95	733	969	993
0.975	144	444	650
0.99	33	120	240
0.995	25	86	170
1	6	31	62
1.01	0	0	0

Neglecting present constant in the cointegrating relation

As explained before, we like the cointegrating relation not to have a constant, α_0. However, if there is a constant we neglect it; we only fit y_t on x_t and not on a constant. In chapter 4 we saw that if there is relatively large positive α_0, neglecting it results in an overestimation of α which in turn results in a down trend in the spread process. The other way around, neglecting a negative value of α_0 results in an up trend in the spread process. Two stock price

processes with a trend in their spread process, do not form a good pair for our trading strategy. But for a small value of α_0, there is not a big trend and the price processes form a good pair as seen in figure 4.18. It is interesting to see how the Engle-Granger method performs if there is a small α_0 but it is neglected. We can generate cointegrated data where the cointegrating relation has a constant. We generate y_t with:

$$y_t = \alpha x_t + \alpha_0 + \beta_0 + \beta\,(y_{t-1} - \alpha x_{t-1}) + \eta_t\,, \qquad \text{for } t = 1, \ldots, T.$$

From this equation we can see that with this generating scheme including α_0 is a bit lame, including α_0 is the same as including a larger value of β_0. We have already seen what happens for larger values of β_0 in figure 6.6. But at last, we now have found a reason to use Dickey-Fuller case 2 instead of case 1! The power of case 1 is practically zero when there is a constant, see table 5.13. Table 6.3 shows this is also true when we perform the preceding step of estimating α. The table shows the number of rejections when we use case 1 in the Engle-Granger method and the 'normal' Engle-Granger which uses the case 2 test. When we use the DF case 1 test in the Engle-Granger method instead of the case 2 test, the power is almost zero. The paths were generated with $x_0 = 25$, $T = 500$, $\alpha = 1$, $\sigma_x = 1$, $\sigma_\eta = 1$ and $\beta_0 = 0$. For the value of α_0 used to make table 6.3 and $\beta < 1$, we do see x_t and y_t as a good pair, so we like the Engle-Granger method to see them as cointegrated.

Table 6.3: Number of rejections, $\alpha_0 = 1$.

	Case 1			Case 2		
β	1%	5%	10%	1%	5%	10%
0.9	6	51	126	578	835	918
0.95	0	0	3	196	432	577
0.975	0	0	0	139	352	492
0.99	0	0	0	77	256	415
0.995	0	0	0	56	157	264
1	0	0	0	4	14	28
1.01	0	0	0	0	0	0

When performing the Engle-Granger test on real data we do not know if there is a small constant in the cointegrating relation, so from now on we only look at the DF case 2 test. Because we use the DF case 2 test within the Engle-Granger method, this method makes the following assumptions:

- Processes x_t, y_t are integrated of order one.
- Spread process $y_t - \alpha x_t$ is an AR(p) process.
- There is no constant in the spread process.

So far, we have seen that when all assumptions of the Engle-Granger method are fulfilled, the Engle-Granger test statistic has the same distribution and power properties as the DF case 2 test statistic. In other words the first step of estimating α does not have an influence. We have seen that when there is a constant in the spread process, so not all assumptions are fulfilled, the distribution makes a limited shift to the right. With limited we mean that the Engle-Granger statistic does not converge to the DF case 3 statistic, like the DF case 2 statistic does when there is an increasing constant. Last, we have seen when there is a constant in the cointegrating relation it is better to use the DF case 2 test within the Engle-Granger method instead of the DF case 1 test. In the next section we examine what happens when the price processes x_t and y_t are not strictly integrated of order 1.

6.2 Engle-Granger simulation with stock price model

In section 4.2 we derived a stock price model which is commonly used for the valuation of options. This model is more realistic then the random walk from the preceding section. Although with this model the assumptions from the Engle-Granger method are not completely satisfied, it is interesting to find out if the method performs the same.

The approach for simulating price processes x_t and y_t is the same as in the preceding section, only the paths for x_t are simulated with the stock price model instead of random walks:

$$x_t = x_{t-1} + \mu\,\delta_t\,x_{t-1} + \sigma\sqrt{\delta_t}\,u_t\,x_{t-1}\,, \tag{6.4}$$

where u_t are i.i.d $N(0,1)$. Then x_t is not exactly integrated of order 1, there is an upward drift μ so the expectation of the differences is not constant. We look at small values of μ and for a for a finite sample size $T = 500$, so x_t is almost integrated of order 1. By simulating a lot of paths for x_t and corresponding y_t we are going to see if this effects the Engle-Granger method.

We again simulate y_t such that the spread process is AR(p) and to fulfill the remaining assumption of the method, we do not include a constant β_0 in the spread process. For $p = 1$ the results of the simulations are the same as in figure 6.3 through 6.5 and table 6.1, that is why they are not displayed. It looks like the Engle-Granger method is not sensitive for x_t not being exactly integrated of order 1. In this section we consider the situation when the spread process is an AR(2) process.

First we need to simulate x_t. Typical values of the drift parameter μ are between 0.01 and 0.1, and volatility σ between 0.05 and 0.5 when we measure time in years. We like to simulate daily stock prices, so we take $\delta_t = 1/260$ because there are roughly 260 trading days a year. We want to generate y_t such that the spread process $y_t - \alpha x_t$ is AR(2), for some α:

$$y_t - \alpha x_t = \varepsilon_t = \beta\varepsilon_{t-1} + \beta_1\Delta\varepsilon_{t-1} + \eta_t\,, \tag{6.5}$$

for n_t we take i.i.d. $N(0,\sigma_\eta^2)$ variables. Then we can generate y_t like:

$$\begin{aligned} y_t &= \alpha x_t + \beta\,(y_{t-1} - \alpha x_{t-1}) + \beta_1\Delta\varepsilon_{t-1} + \eta_t\,, \quad t = 2,\ldots,T, \\ y_0 &= \alpha x_0, \quad y_1 = \alpha x_1\,, \end{aligned} \tag{6.6}$$

where we use within each step:

$$\Delta\varepsilon_{t-1} = (y_{t-1} - \alpha x_{t-1}) - (y_{t-2} - \alpha x_{t-2})\,.$$

We take $\sigma_\eta^2 = 0.1$, because if we take the variance of η equal to 1 then y_t is much more jagged than x_t and we are trying to model the price processes more realistically. Then x_t and y_t are cointegrated if $\beta < 1$. Again we estimate the density of the Engle-Granger test statistic by simulating for different values of α in the same way as the previous paragraph. The results are shown in figure 6.7 for $T = 500$, $x_0 = 25$, $\mu = 0.05$, $\sigma = 0.20$, $\beta_1 = -0.1$ and of course $\beta = 1$. The density of the Engle-Granger test statistic is again comparable with the density of the Dickey-Fuller case 2 statistic.

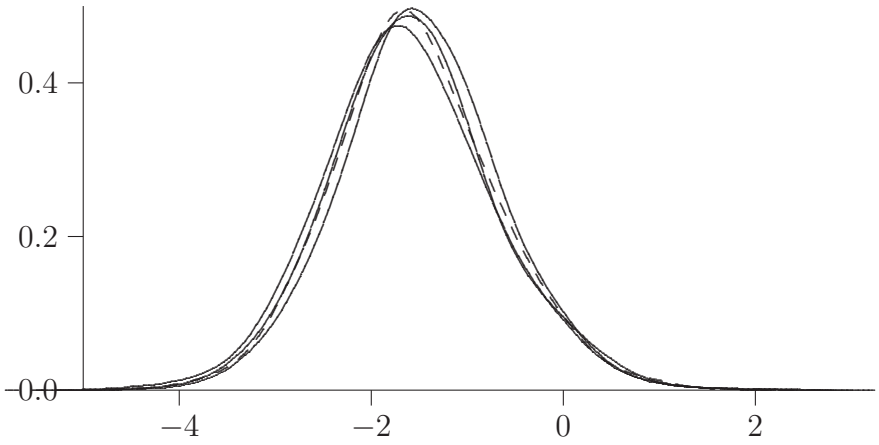

Figure 6.7: AR(2) Estimated density for EG test statistic , $\alpha = 0.1, 0.5, 1$.

Table 6.4 shows the number of rejections for three different values of β_1. Compared to table 5.14, which shows the corresponding power of the Dickey-Fuller case 2 test, the power of the Engle-Granger method has not declined. It seems that the Engle-Granger performs the same for data that is not exactly integrated of order one, as for data that is.

Table 6.4: Number of rejections, AR(2) and $\beta_0 = 0$.

	$\beta_1 = -0.25$			$\beta_1 = -0.1$			$\beta_1 = 0.1$		
β	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.9	987	1000	1000	999	1000	1000	1000	1000	1000
0.95	477	834	945	611	925	986	825	986	998
0.975	98	337	505	139	404	617	197	569	756
0.99	15	98	176	24	126	234	30	150	262
0.995	12	66	143	12	65	144	21	97	163
1	8	61	111	11	53	101	13	50	104
1.01	1	2	5	0	2	3	1	4	4

6.3 Engle-Granger with bootstrapping from real data

So far we have simulated paths x_t and y_t from scratch to find the critical values of the Engle-Granger method. In this section we build paths x_t and y_t by bootstrapping from real data. The data are the ten pairs of stocks that IMC provided. First we describe the bootstrap procedure and then we look at some results of the ten pairs.

Bootstrap procedure
Assume we have a pair that consists of two stock price processes x_t and y_t, for $t = 0, \ldots, T$, which are integrated of order one. Let us assume further that there exists an α such that $y_t - \alpha x_t$ follows an AR(p) process:

$$y_t - \alpha x_t = \varepsilon_t = \beta_0 + \beta\varepsilon_{t-1} + \beta_1\Delta\varepsilon_{t-1} + \cdots + \beta_{p-1}\Delta\varepsilon_{t-p+1} + \eta_t\,. \qquad (6.7)$$

The null hypothesis of no cointegration against the alternative that there is cointegration between x_t and y_t can be formulated as

$$H_0 : \beta = 1 \text{ against } H_1 : \beta < 0\,.$$

The first step in the bootstrap procedure is to estimate α with OLS, which results in $\hat{\alpha}$. Then we can calculate the spread process:

$$e_t = y_t - \hat{\alpha}x_t, \quad t = 0, \ldots, T,$$

this resembles the true spread process ε_t which is assumed to follow an AR(p) process.

In the preceding sections we knew the value of p but now do not, since we are working with real data. The second step is to estimate p with the information criteria described in chapter 3, which results in $\hat{p}$.

The third step is to estimate the coefficients of the AR($\hat{p}$) model with linear regression, which results in $\hat{\beta}, \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{\hat{p}-1}$. Then we can calculate the residuals:

$$n_t = e_t - \hat{\beta}_0 - \hat{\beta}e_{t-1} - \hat{\beta}_1\Delta e_{t-1} - \cdots - \hat{\beta}_{\hat{p}-1}\Delta e_{t-\hat{p}+1}, \quad t = \hat{p}, \ldots, T,$$

this resembles the true residuals η_t which are assumed to be white noise.

The fourth step is to calculate the test statistic for the real data

$$S = \frac{\hat{\beta} - 1}{\hat{\sigma}_{\hat{\beta}}},$$

where $\hat{\sigma}_{\hat{\beta}}$ is the standard error of $\hat{\beta}$.

Now we are ready to build a new path y_t^* that belongs to the original x_t. This is done in the following way:

$$y_t^* = \hat{\alpha} x_t + \varepsilon_t^*, \quad t = \hat{p}, \ldots, T,$$

where ε_t^* is built under the null hypothesis, that is $\beta = 1$ and $\beta_0 = 0$:

$$\varepsilon_t^* = \varepsilon_{t-1}^* + \hat{\beta}_1 \Delta\varepsilon_{t-1}^* + \cdots + \hat{\beta}_{\hat{p}-1} \Delta\varepsilon_{t-\hat{p}+1}^* + \eta_t^*,$$

with η_t^* is taking uniform out of n_t. We initialize the new path by:

$$\varepsilon_i^* = y_i - \hat{\alpha} x_i, \quad i = 0, \ldots, \hat{p} - 1.$$

We treat the new pair $\{x_t, y_t^*\}$ the same way as with the original pair $\{x_t, y_t\}$. That is, we calculate $\hat{\alpha}^*$ and spread process $e_t^* = y_t^* - \hat{\alpha}^* x_t$ which should follow an AR($\hat{p}$) process. Then we estimate the coefficients of this AR($\hat{p}$) process and calculate the test statistic S^*:

$$S^* = \frac{\hat{\beta}^* - 1}{\hat{\sigma}_{\hat{\beta}^*}}.$$

By building a lot of new paths y_t^* and calculating the corresponding test statistic S^*, we can calculate the density of these bootstrapped test statistics. Then we can see if the test statistic of the real pair is exceptional. The estimated density should also give an indication for the critical values of the Engle-Granger method.

Results

The ten provided pairs are named *pair I*, *pair II*,..., *pair X*. We start with a pair for which all three information criteria indicate that the spread process is AR(1), *pair II*. By spread process we mean the residuals from the first

regression, e_t. The two stocks used are the same stock but listed on different exchanges. The spread process is shown in figure 6.8. This is not necessarily the spread we trade, in chapter 2 we discussed the adjustment parameter κ which can result in a different spread. With *pair II* we will find $\kappa = 0$, so the spreads for this pair look the same. In pair trading, this is as good as it gets: we have a large number of trades, we never have a position for a long time and the risk of the two stocks walking away from each other is minimal because they are in fact the same.

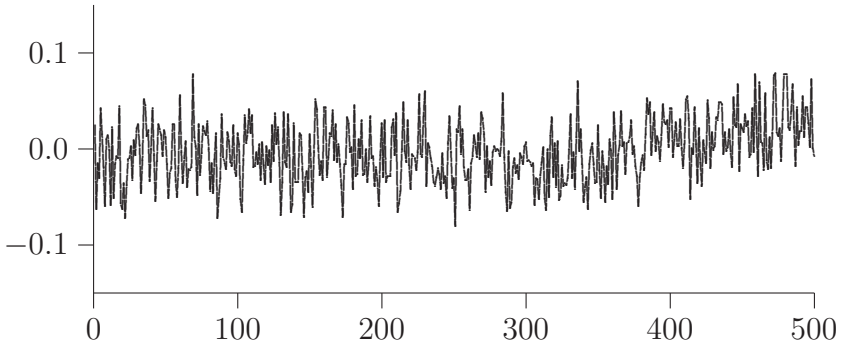

Figure 6.8: Spread process *pair II*.

The spread series look stationary and according to the Engle-Granger method the two stocks are cointegrated. The test statistic is -17.5, compared to the 1% critical value which is -3.44, we see that the null hypothesis of no cointegration is rejected. Applying the bootstrap procedure on this data set, we get figure 6.9. The dashed line is the density of the Dickey-Fuller case 2 test statistic. This figure does not give an indication that the density of the Engle-Granger test statistic differs from the Dickey-Fuller case 2 statistic.

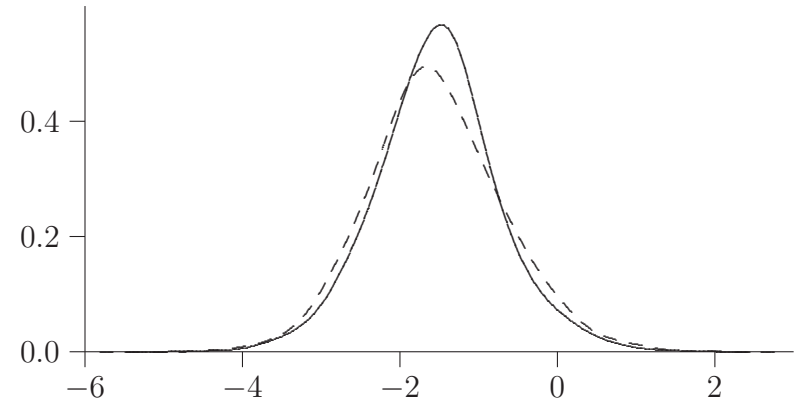

Figure 6.9: Estimated density for EG test statistic by bootstrapping from *pair II*.

Let us consider a pair for which all information criteria say that the spread process is AR(2), *pair VII*. The spread process is shown in figure 6.10. It does not look as good as figure 6.8, but this still is a good pair. According to the Engle-Granger method the stocks in this pair are cointegrated, the test statistic is -4.65. The bootstrap procedure results in figure 6.11. The estimated density coincides with the density of the Dickey-Fuller case 2 statistic.

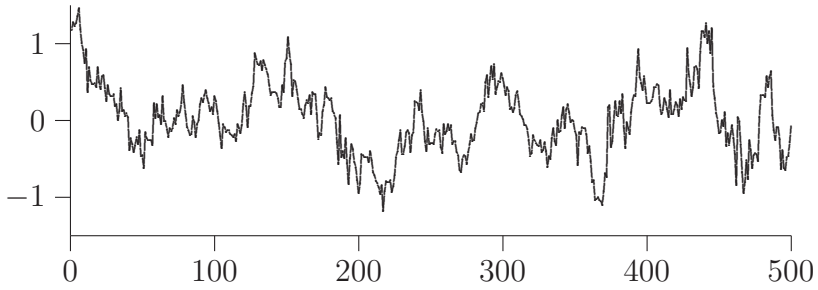

Figure 6.10: Spread process *pair VII*.

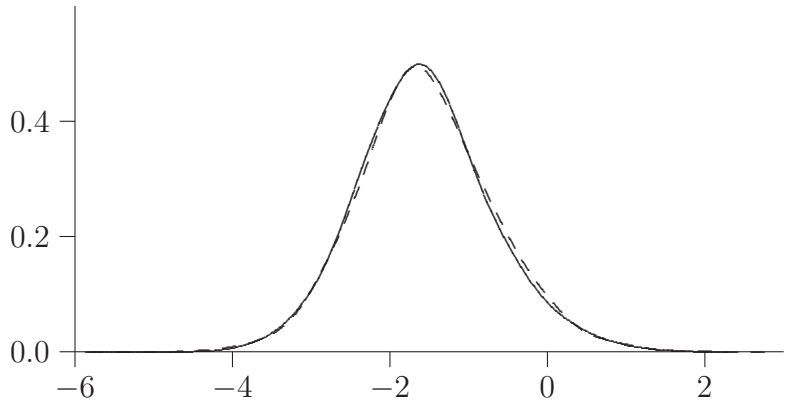

Figure 6.11: Estimated density for EG test statistic by bootstrapping from *pair VII*.

Let us consider a pair for which all information criteria indicate that the spread process is AR(3), *pair VI*. The spread process is shown in figure 6.12. This looks a lot less interesting than the previous figure: initially the spread is below zero for a long time and at the end the spread is above zero for a long time. This shows that trading the spread would have resulted in only a few trades and we would have had the same position for a long time. But this is not necessarily the spread we trade as stated, in the next chapter we will see the spread we would have really traded. According to the Engle-Granger method the stocks in this pair are not cointegrated, the test statistic is -2.23. The null hypothesis of no cointegration is not even rejected at the 10% level. The bootstrap procedure results in figure 6.13. The estimated density is a bit bumpy but still coincides with the density of the Dickey-Fuller case 2 statistic. Even when the real data is not cointegrated, according to the Engle-Granger method, the bootstrap procedure finds nearly the same density as the density of the Dickey-Fuller test statistic.

So far we have seen pairs for which all information criteria find the same small value of p. IMC also provided a pair for which the information criteria find p to be very large, *pair V*. As described in chapter 3 we fit an AR(k) model, for $k = 1, \ldots, K$, on the data and see for which k the criteria have to lowest values. For this pair, even if we set $K = 100$ the criteria have to lowest value for $p = K$. This indicates that the spread process does not follow an AR(p) model. The spread process is shown in figure 6.14. It is obvious that

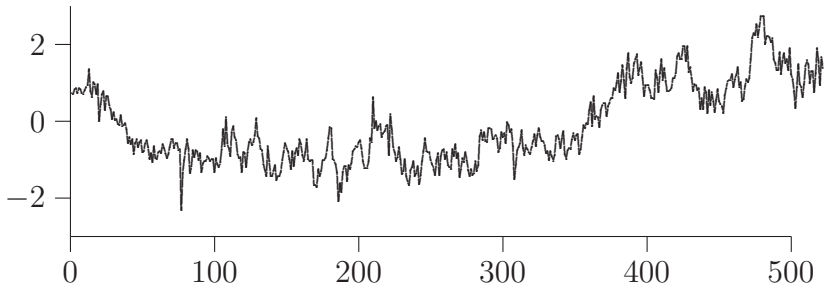

Figure 6.12: Spread process *pair VI*.

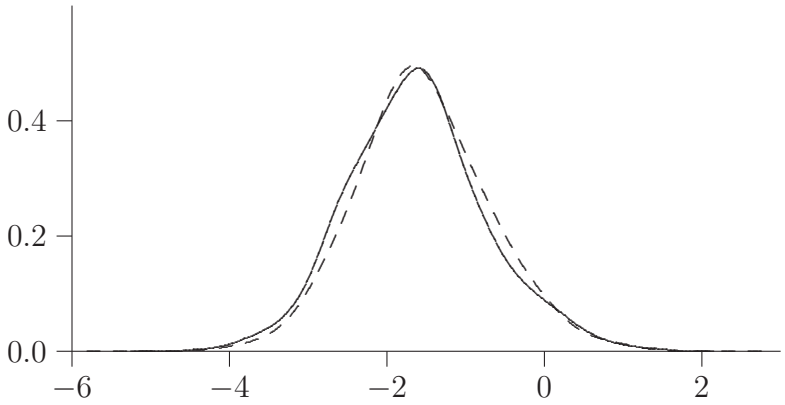

Figure 6.13: Estimated density for EG test statistic by bootstrapping from *pair VI*.

this 'pair' is not suitable for pair trading. The Engle-Granger method does not reject the null hypothesis of no cointegration because the test statistic is -1.04 when $p = 10$ and 0.63 if $p = 100$. To apply to bootstrap procedure, we set $p = 10$. The result is shown in figure 6.15, which coincides surprisingly well with the density of the Dickey-Fuller case 2 test statistic.

We examine these and the remaining pairs further in chapter 7. In this chapter, we have seen no reason to assume that the test statistic of the Engle-Granger has a different distribution than the Dickey-Fuller case 2 test statistic. The power of the two tests are also comparable. To find out if the Engle-Granger method is also 'robust', we apply the method on generated data which do not fulfill the null hypothesis.

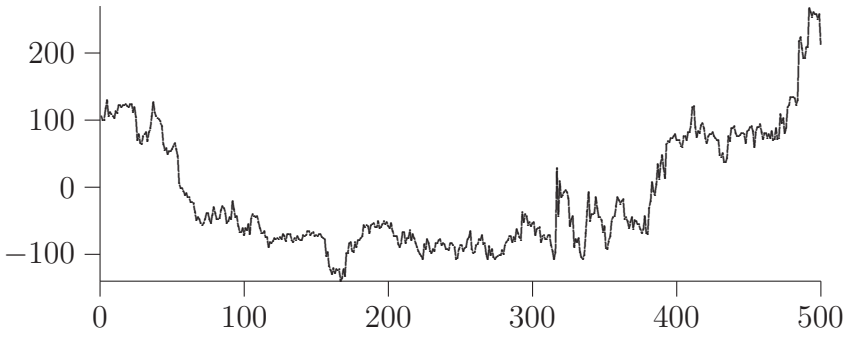

Figure 6.14: Spread process *pair V*.

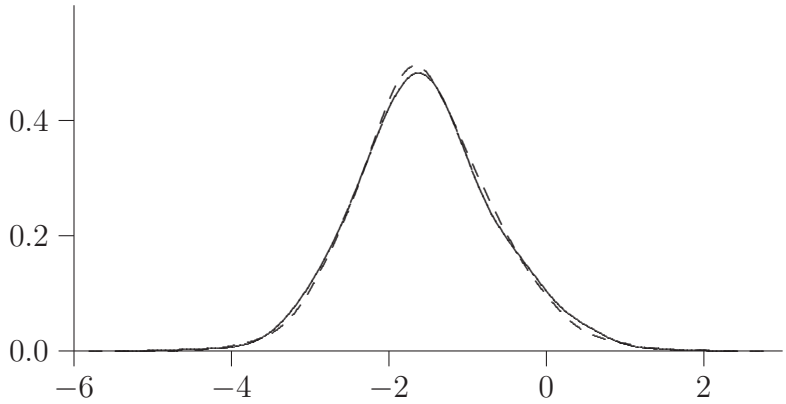

Figure 6.15: Estimated density for EG test statistic by bootstrapping from *pair V*.

6.4 Engle-Granger simulation with alternative method

In section 4.5 we found a method for generating cointegrated data which do not satisfy the assumptions of the Engle-Granger method. The generated data is integrated of order one, but the spread process is not likely to follow an AR(p) process. It is interesting to find out if the Engle-Granger method is robust enough to see this data as cointegrated.

We will generate data such that the difference process $\mathbf{z}_t$ follows an MA(2) model:

$$\begin{bmatrix} x_t - x_{t-1} \\ y_t - y_{t-1} \end{bmatrix} = \mathbf{z}_t = \Theta_2 \mathbf{w}_{t-2} + \Theta_1 \mathbf{w}_{t-1} + \Theta_0 \mathbf{w}_t \,,$$

where $\mathbf{w}_t$ is i.i.d. $N_2(0, \Sigma)$ and $\Theta_0 = I$. Then x_t and y_t are cointegrated if matrix $(\Theta_2 + \Theta_1 + \Theta_0)$ has eigenvalue zero and the corresponding eigenvector is the cointegrating relation, which we normalize to $[-\alpha \quad 1]'$. For example, the matrix

$$\begin{bmatrix} 4 & 2 \\ -2 & -1 \end{bmatrix},$$

has eigenvalue zero with eigenvector $[-1/2 \quad 1]'$. So one possibility to generate cointegrated x_t and y_t, is

$$\Theta_2 = \begin{bmatrix} 2 & 1 \\ -2 & -4 \end{bmatrix}, \quad \Theta_1 = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}, \quad \Theta_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

There are no restrictions on the covariance matrix of the innovations $\mathbf{w}_t$, Σ except that it is a covariance matrix, so it must be symmetric. We start with a diagonal matrix, so the innovations are independent. For $\Sigma = cI$, table 6.5 shows the number of rejections of the Engle-Granger test for 1,000 different paths x_t and y_t. Although the spread process in this section is not of autoregressive form, the Engle-Granger method fits an AR(p) on the spread process. The value of p is again estimated with the information criteria from chapter 3, the maximum value of p, K, was set equal to 10. The table also shows the average of the estimated values of p. Unfortunately, the Engle-Granger method does not perform very well. It sees on average 12% of the generated paths as cointegrated on a 10% level. The average of estimated p values is high, which means that it is difficult to fit a good AR(p) model on the spread process of the data, which in turn is not strange because the spread process does not follow an AR(p) model.

Table 6.5: Number of rejections, $\Sigma = cI$.

c	1%	5%	10%	$\bar{p}$
2	47	89	147	9.9
1	25	77	125	9.6
0.5	24	61	116	9.1
0.1	11	49	102	7.6

Consider the situation when the innovations are correlated, we take Σ of the form:

$$A = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Table 6.6 shows the number of rejections of the Engle-Granger test for different values of ρ. Even for $\rho = 1$ the Engle-Granger method does not perform well.

Table 6.6: Number of rejections, $\Sigma = A$

ρ	1%	5%	10%	$\bar{p}$
1	34	86	140	9.8
0.5	31	67	134	9.7

To see what happens, figure 6.16 shows the spread process of a realization x_t and y_t. This does not look stationary, there seems to be a trend in the spread process. This could mean that with this setting there a is a constant in the cointegrating relation, α_0. Figure 6.17 shows the spread process if we regress the same realization of y_t on the same x_t and a constant.

It is clear that there is a constant, α_0, in the cointegrating relation. Neglecting the constant, results in a spread process which is not stationary. That is why the Engle-Granger method does not reject the null hypothesis of no cointegration. Table 6.7 shows the number of rejection when we do not neglect α_0 and take $\Sigma = cI$. The maximum value of p is set to 5 in the remainder of this section, to reduce computation time. It is clear that the Engle-Granger method performs very well, almost every path is seen as cointegrated (which is true).

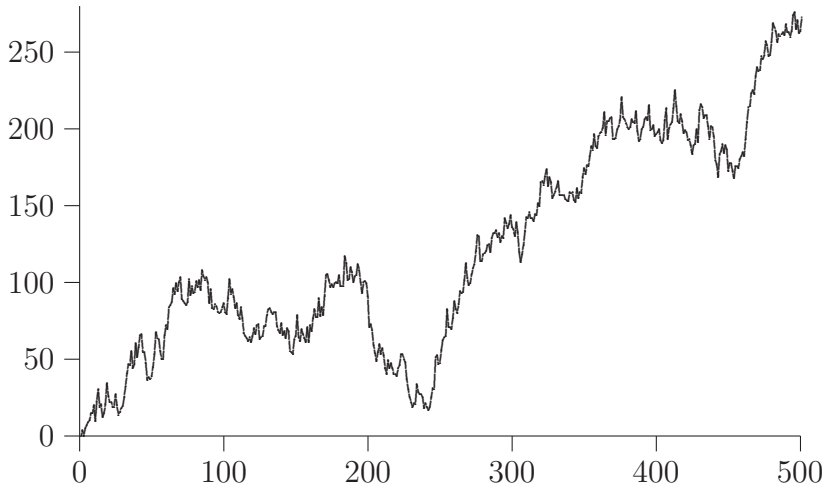

Figure 6.16: Realization spread setting 1, $\rho = 0.5$.

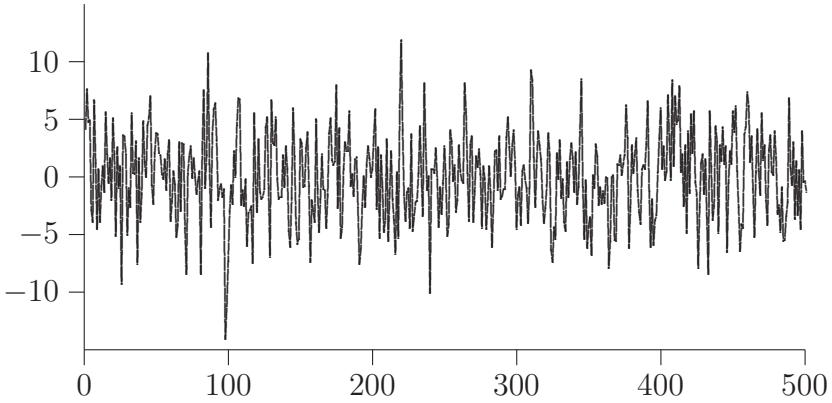

Figure 6.17: Realization spread setting 1, with α_0.

Table 6.7: Number of rejections, with α_0 and $\Sigma = cI$.

c	1%	5%	10%	$\bar{p}$
2	1000	1000	1000	4.7
1	1000	1000	1000	4.5
0.5	999	1000	1000	4.3
0.1	1000	1000	1000	3.5

So far, we have generated cointegrated data but not in the way we want it to be cointegrated, which is data with a small or no constant α_0. We look at a different setting of parameters:

$$\Theta_2 = \begin{bmatrix} 1 & -1 \\ -1 & 0 \end{bmatrix}, \quad \Theta_1 = \begin{bmatrix} -1 & 2 \\ 2 & 0 \end{bmatrix}, \quad \Theta_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The matrix $(\Theta_2 + \Theta_1 + \Theta_0)$ has eigenvalue zero with eigenvector $[-1 \quad 1]'$. Figure 6.18 shows a realization of the spread process, when y_t is only regressed on x_t and not on a constant. In other words, we neglect a possible α_0.

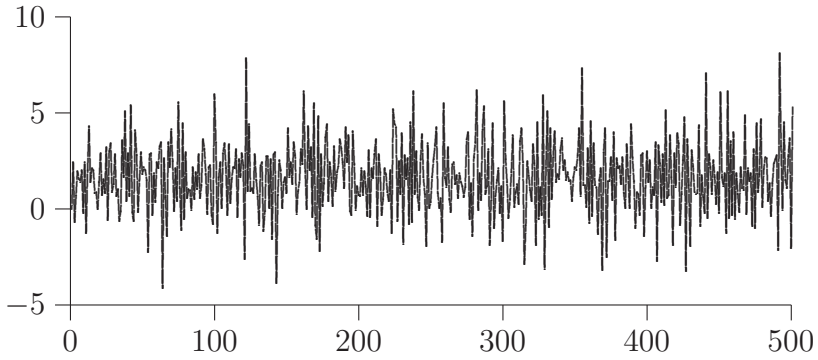

Figure 6.18: Realization spread setting 2, $\rho = 0.5$.

Neglecting α_0 with this setting is no problem, the spread process seems to be stationary. We wish Engle-Granger method to see the corresponding x_t and y_t as cointegrated. Table 6.8 and 6.9 show the number of rejections for $\Sigma = cI$ and $\Sigma = A$ respectively. The power of the Engle-Granger test is good, almost every time the null is rejected.

The Engle-Granger method performs very well, even when the spread process does not follow an AR(p) model, the test behaves exactly how we want it. If there is a large constant α_0 in the cointegrating relation, it does not reject the null hypothesis of no cointegration. Although the data is cointegrated, it is not cointegrated the way we want it, that is with a small or no α_0. If there is a small or no constant in the cointegration relation, the test has rejected the null hypothesis almost every time.

Table 6.8: Number of rejections, setting $2\ \Sigma = cI$.

c	1%	5%	10%	$\bar{p}$
2	946	980	992	4.83
1	974	988	997	4.75
0.5	981	992	995	4.60
0.1	995	998	999	3.95

Table 6.9: Number of rejections, setting $2\ \Sigma = A$.

ρ	1%	5%	10%	$\bar{p}$
1	991	996	999	4.7
0.5	970	991	996	4.6

Chapter 7

Results

In this chapter the results for the ten pairs IMC provided are discussed. To be clear, IMC provided 2 years of historical closing prices for each stock of the ten pairs. According to IMC, among these ten are some very good pairs which means they make high profits. Some are losing money and some are mediocre. In the first section we apply the trading strategy to the historical data to see which pairs would have been profitable and put the pairs in order of profitability. We like to see if the stocks in a profitable pair are cointegrated, and if the stocks in a pair that loses money are not cointegrated. In other words if profitable and cointegration coincides. We apply two different cointegration test, the Engle-Granger and the Johansen method, but first we will examine in the second section if the assumption of the price processes being integrated of order 1 is fulfilled. In the third and fourth section the results for respectively the Engle-Granger and the Johansen method are stated, the pairs are put in order of the levels of rejection of the cointegration tests.

7.1 Results trading strategy

The 10 pairs are named, *pair I*, *pair II*,..., *pair X*. In chapter 2 the trading strategy was explained. For each pair, we need the first half of observations to determine the parameters of the strategy and we apply the strategy to the second half. In order to compare the results we trade the same amount of money with each pair. With each trade we buy one stock for the amount of € 10,000 and sell the other for roughly the same amount. The sell trade is not exactly € 10,000 because of the positive or negative 'investment' of

threshold Γ, as explained in section 2.3. The results/profits are shown in table 7.1. The traded spread processes of the 10 pairs are shown in figure 7.1, these are the spreads with the adjustment ratio if present. The upper left corner is the spread for *pair I*, upper right corner for *pair II*, and so on. To be clear, the spread of the second half of observations is displayed and this is the spread which is traded. The dashed lines are the corresponding thresholds Γ.

Table 7.1: Results trading strategy.

	parameters		result	
pair	Γ	κ	# trades	profit
I	2.33	5	3	1129
II	0.02	0	25	5536
III	0.16	2	4	506
IV	0.77	5	7	1344
V	19.68	8	0	0
VI	0.48	1	4	495
VII	0.13	1	11	2293
VIII	030	2	10	2091
IX	0.12	2	4	141
X	0.30	2	12	2304

Even the highest profit may look a bit small, but recall that we do not have to invest a lot of money. On the other hand to loose the same amount as the highest profit, the two stocks have to walk 50% away from each other in the wrong direction, which has little chance of occurring. Profits above €1,000 are considered good enough to trade, profits below €1,000 are considered not to be worthwhile. But profit is not the only criteria, the number of trades is also important. Obviously, the more trades the higher the profit. But this is not the only reason, in chapter 2 was explained that traders do not want a position for a long time because that involves risk and the number of trades is an indication for this. According to IMC, *pair IV* is still a good one. We get exactly the same selection of good and bad pairs as IMC if we set the minimal amount of trades equal to 7. IMC already decided which of the 10 pairs is a good one and which is not based on trading experiences, before providing the data. A pair is considered good enough to trade if the profit is

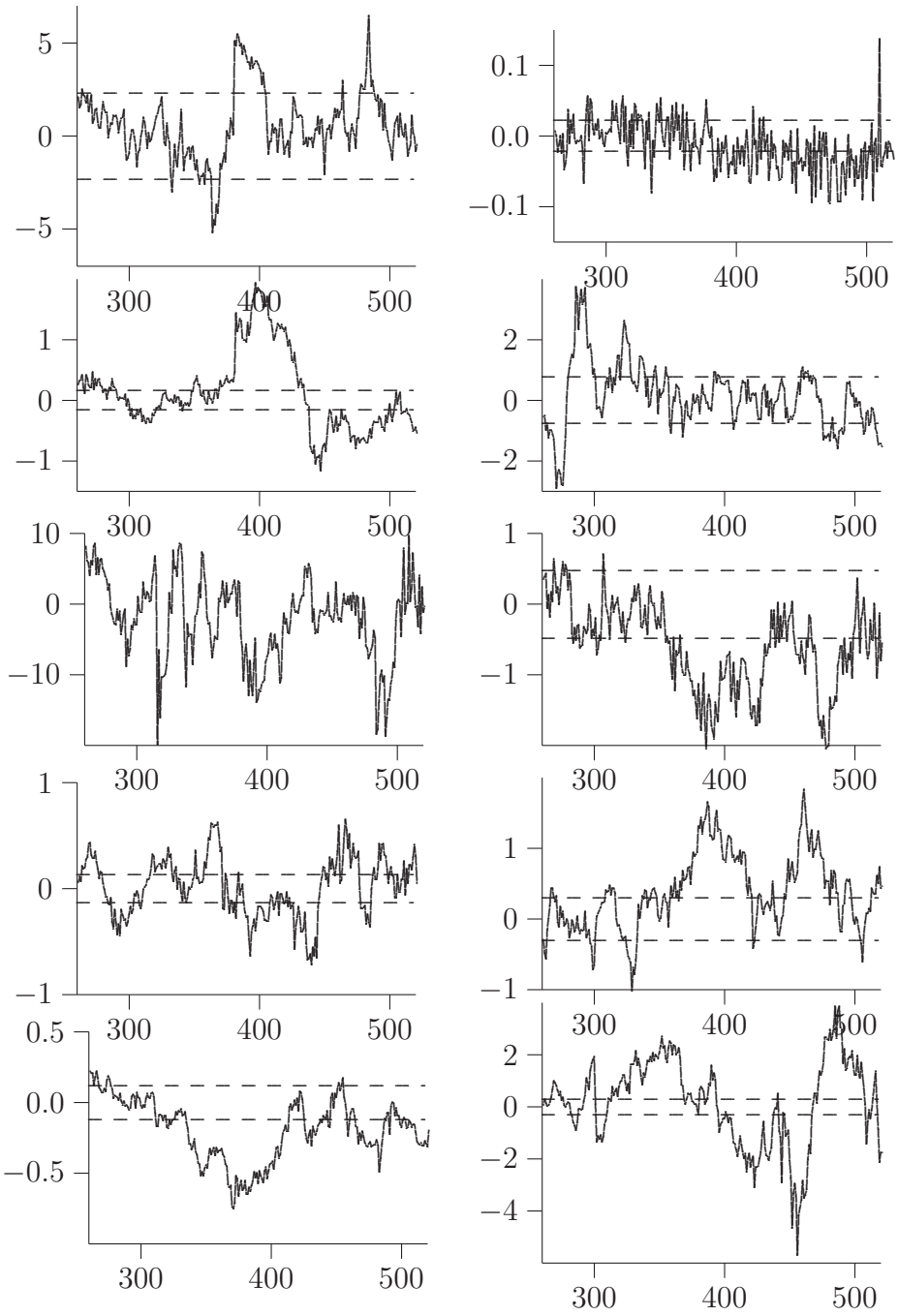

Figure 7.1: Traded spreads.

above €1,000 and the number of trades is no less than 7, otherwise the pair is considered not to be worthwhile. The ordering of the 10 pairs based purely on the results from the trading strategy, where the first five are considered good or good enough, the remaining five are not, is:

1	*pair II*
2	*pair X*
3	*pair VII*
4	*pair VIII*
5	*pair IV*
6	*pair I*
7	*pair III*
8	*pair VI*
9	*pair IX*
10	*pair V*

We briefly discuss the spreads from figure 7.1. The spread for *pair I* is rarely hitting its threshold Γ although the adjustment parameter κ is large. The spread for *pair II* looks good, but it could be better if we had used $\kappa = 1$. After $t = 425$ the spread is a relatively long time below $+\Gamma$, with $\kappa = 1$ we would have made a profit of €6721 in 36 trades. The spreads for *pair III*, *VI* and *IX* are rarely hitting their tresholds Γ, the adjustment parameter κ is small but increasing it does not have a positive effect. For *pair III* increasing κ to 5, results in a loss of €2,108. For *pair VI* the profit gets smaller, while the number of trades increases. The spread for *pair IV* shows the reason why we use an adjustment parameter, without it this pair would have traded twice with a total profit of €385. For *pair V* the threshold Γ is not displayed, because it is 19.68. Lowering the threshold results in a loss when we keep $\kappa = 8$, only when we also reduce κ to 1 or zero we get a small profit. The spreads from *pair VII*, *VIII*, and *IX* look good, they hit their thresholds Γ regularly and produce a nice profit. Changing the parameters slightly does not effect the number of trades, it affects the profits slightly.

7.2 Results testing price process I(1)

Both cointegration tests require that the stock price processes x_t and y_t are integrated of order one. In section 4.2 was derived that it is reasonable to assume stock price processes fulfill this requirement, but in this section we will do a unit root test on the stocks of the 10 pairs to see if the requirement is fulfilled. The unit root test we use is again the (Augmented) Dickey-Fuller case 2 test, we perform the test twice. The first test is:

$$H_0 : \ x_t \sim I(1) \text{ against } H_1 : \ x_t \sim I(0) \, .$$

The outcome should be not to reject H_0. The second test is:

$$H_0 : \ x_t \sim I(2) \text{ against } H_1 : \ x_t \sim I(1) \, ,$$

which is equivalent to:

$$H_0 : \ \Delta x_t \sim I(1) \text{ against } H_1 : \ \Delta x_t \sim I(0) \, .$$

The outcome of this second test should be to reject H_0, which makes it likely that the price processes are $I(1)$.

The Dickey-Fuller test fits an AR(p) model to the stock price process, we estimate p with the information criteria from chapter 3 and set the maximum value of p, which is K, equal to 10. Table 7.2 shows the outcomes of both tests, where we used the following critical values

1%	5%	10%
-3.44	-2.87	-2.57

since we have roughly 520 observations, $T = 520$. The stocks in a pair are denoted with x and y, the test statistic of the first test is stated along with whether the null hypothesis is rejected. The outcome of 'not rejected' is denoted with symbol $\neg$ otherwise the level is stated. The average value of the estimated p also stated and the results of the second test are stated in the same way.

The table shows that it is likely that all stocks from the 10 pairs are integrated of order one.

Table 7.2: Results I(1).

	Test 1			Test 2		
stock	statistic	outcome	$\bar{p}$	statistic	outcome	$\bar{p}$
I-x	-1.7	¬	4	-11	1%	4
I-y	-2.1	¬	4	-11	1%	4
II-x	-1.4	¬	1	-12	1%	3
II-y	-1.3	¬	2	-25	1%	1
III-x	-2.1	¬	1	-22	1%	1
III-y	-2.2	¬	1	-26	1%	1
IV-x	-1.4	¬	1	-24	1%	1
IV-y	-1.5	¬	1	-22	1%	1
V-x	-1.4	¬	8	-8	1%	8
V-y	-0.3	¬	10	-8	1%	10
VI-x	-0.6	¬	4	-12	1%	4
VI-y	-0.6	¬	4	-12	1%	4
VII-x	-1.1	¬	1	-25	1%	1
VII-y	-0.9	¬	1	-25	1%	1
$VIII$-x	-1.4	¬	1	-23	1%	1
$VIII$-y	-1.5	¬	1	-23	1%	1
IX-x	-1.4	¬	1	-25	1%	1
IX-y	-0.9	¬	1	-25	1%	1
X-x	-0.5	¬	2	-23	1%	1
X-y	-0.7	¬	1	-16	1%	2

7.3 Results Engle-Granger cointegration test

In this section we perform the Engle-Granger test on the 10 pairs. We have found no reason to assume the Engle-Granger test statistic has a different distribution than the Dickey-Fuller case 2 test statistic, so we will use the same critical values:

1%	5%	10%
-3.44	-2.87	-2.57

We perform the cointegration test on the whole data set, so we have 520 observations per stock. Recall that the profits were determined for the second half of observations. As stated in section 4.3, the Engle-Granger method is not symmetric. The results can be different for regressing x_t on y_t and the other way around. That is why we perform the Engle-Granger test twice. The results are stated in table 7.3.

Table 7.3: Results Engle-Granger test.

pair	statistic 1	outcome	$\hat{\alpha}_1$	$\bar{p}$	statistic 2	outcome	$\hat{\alpha}_2$	$\bar{p}$
I	-1.57	¬	1.53	6	-1.56	¬	0.65	6
II	-17.54	1%	0.997	1	-17.56	1%	1.003	1
III	-2.21	¬	0.72	3	-2.21	¬	1.37	3
IV	-2.67	10%	0.52	1	-2.61	10%	1.91	2
V	-1.04	¬	5.03	10	-1.08	¬	0.20	7
VI	-2.23	¬	1.08	3	-2.24	¬	0.93	3
VII	-4.65	1%	1.60	2	-4.64	1%	0.63	2
VIII	-2.92	5%	0.53	1	-2.91	5%	1.87	1
IX	-0.63	¬	2.36	2	-0.90	¬	0.42	1
X	-3.48	1%	0.72	1	-3.03	5%	1.39	4

We see that there is only one pair where the outcome of the two tests are different, *pair X*. The estimated cointegrating relation are for all pairs practically the same:

$$\hat{\alpha}_1 \approx 1/\hat{\alpha}_2 \,.$$

So the disadvantage of the Engle-Granger method of not being symmetric does not seem to be very harmful when testing pairs for cointegration. The pairs are put in order of the test statistic, the idea is that the lower test statistic the lower the level of rejection, which is more evidence for being cointegrated. For example, the Engle-Granger method rejects the null hypothesis for *pair II* even at 0.1% level, while *pair VIII* is only rejected at 5%. So there is more evidence that *pair II* is cointegrated than *pair VIII*, which is why we prefer *pair II*.

The ordering of the 10 pairs based on the Engle-Granger method is

1	*pair II*
2	*pair VII*
3	*pair X*
4	*pair VIII*
5	*pair IV*
6	*pair VI*
7	*pair III*
8	*pair I*
9	*pair V*
10	*pair IX*

where there is evidence for cointegration for the first five pairs, and no evidence for the remaining five. This ordering is not exactly the same as the ordering found with the trading strategy, but they coincide on what is good and what is not. The five pairs which are considered to be worthwhile trading are cointegrated and the five pairs that are not worthwhile trading are not cointegrated according to the Engle-Granger method. The first in both orderings are the same, this is the pair that consists of two equal stocks but listed on different exchanges. In the first half of the ordering, the good ones, only places 2 and 3 are switched, the others are at the same places. The second half of the two orderings differ a lot.

7.4 Results Johansen cointegration test

In this section we perform the Johansen test on the 10 pairs. As discussed in section 4.4, we perform three tests:

Test 1	$H_0 : 0$ relations	against	$H_1 : 2$ relations.
Test 2	$H_0 : 0$ relations	against	$H_1 : 1$ relation.
Test 3	$H_0 : 1$ relation	against	$H_1 : 2$ relations.

The critical values for each test are in table 7.4, these are for a sample size of $T = 400$. Although the data of the 10 pairs IMC provided consist of 520 observations, these critical values will be used when testing the 10 pairs for cointegration.

Table 7.4: Critical values for Johansen test.

Test	1%	5%	10%
1	16.31	12.53	10.47
2	15.69	11.44	9.52
3	6.51	3.84	2.86

One issue that was not addressed in section 4.4 was how to find p. The Johansen method assumes that the vector process $\mathbf{y}_t = (x_t, y_t)$ follows a VAR(p) model. In S-PLUS, the program used for all simulations and calculations in this report, exists a built-in function called 'ar' which fits a VAR model using Yule-Walker equations. The function determines the order of the VAR with the Akaike information criterion. This function is used for estimating p. We set the maximum value of p equal to 10 and the minimum value equal to 2 because the first step of the Johansen method is to fit a VAR($p-1$) on the differences $\Delta\mathbf{y}_t$. The results of the Johansen test are in table 7.5.

Table 7.5: Results Johansen test.

	Test 1		Test 2		Test 3		Parameters	
pair	stat. 1	outcome	stat. 2	outcome	stat. 3	outcome	$\hat{p}$	$\hat{\alpha}$
I	6.80	¬	6.12	¬	0.68	¬	2	1.49
II	154.6	1%	153.8	1%	0.82	¬	2	0.997
III	7.37	¬	6.56	¬	0.81	¬	2	0.71
IV	10.89	10%	10.43	10%	0.46	¬	2	0.52
V	2.86	¬	2.79	¬	0.07	¬	4	5.44
VI	6.47	¬	5.91	¬	0.56	¬	3	1.08
VII	24.25	1%	23.10	1%	1.15	¬	2	1.59
VIII	13.84	5%	13.16	5%	0.02	¬	2	0.52
IX	2.78	¬	2.68	¬	0.11	¬	2	2.55
X	14.98	5%	10.11	5%	1.98	¬	2	0.72

The Johansen method is symmetric, there is no difference if we set $\mathbf{y}_t = (x_t, y_t)$ or $\mathbf{y}_t = (y_t, x_t)$. The test statistics and the estimated cointegration relations are exactly the same. We consider the stocks of the pair being cointegrated if the null hypothesis of the first and the second test are rejected and the null hypothesis of the third test is not rejected.

The Johansen method finds the same pairs cointegrated as the Engle-Granger method, *pair II*, *IV*, *VII*, *VIII* and *X*. The levels for rejection the null hypothesis of no cointegration are the same. Only for *pair X* the results differ a bit, but this is because the Engle-Granger method had two different outcomes, the first test had rejected at 1% and the second test at 5%. The Johansen method has rejected *pair X* at 5%. There are no real differences for the cointegrated pairs, the estimated cointegrating relations are also practically the same. The biggest difference is for *pair VIII*, where the Engle-Granger method estimates α equal to 5.338 and the Johansen method 5.231. The two methods differ more for pairs that are not cointegrated, the differences between the estimates of α are larger. But according to these methods the pairs are not cointegrated so there does not exist an α such that $y_t - \alpha x_t$ is stationary.

The ordering of the 10 pairs based on the Johansen method is

1	*pair II*
2	*pair VII*
3	*pair VIII*
4	*pair X*
5	*pair IV*
6	*pair III*
7	*pair I*
8	*pair VI*
9	*pair V*
10	*pair IX*

where there is evidence for cointegration for the first five pairs, and no evidence for the remaining five. This ordering differs slightly from the Engle-Granger ordering. But most important is that the two methods coincide on which pairs are cointegrated and which are not. And this in turn coincides with the results from the trading strategy.

Chapter 8

Conclusion

The goal of this project was to apply statistical techniques to find relationships between stocks. The closing prices of these stocks, dating back two years, are the only data that have been used in this analysis.

From trading experience, IMC is able to make a distinction between good and bad pairs based on profits. In chapter 2 we derived a trading strategy that resembles the strategy used by IMC. From this strategy, we derived the important characteristics of a good pair. We saw that we like the price processes to be tied together such that their spread oscillates around zero and does not walk away.

In this report we tried to identify pairs with cointegration. If two stocks in a pair are cointegrated, a certain linear combination of the two is stationary. This implies that this linear combination, which can be seen as the spread, is mean-reverting. This is in line with the characteristics of a good pair.

In chapter 4 we introduced two methods for testing for cointegration, the Engle-Granger and the Johansen method. We have looked at the Engle-Granger method in detail. This method makes use of a unit root test, the Dickey-Fuller test. Because there is a lot of ambiguity in the literature of which Dickey-Fuller test and which critical values should be used, we discussed the different cases in chapter 5. The asymptotic distributions of the test statistics were derived and the critical values for finite sample sizes were found with simulation.

In chapter 6 we examined the properties of the Engle-Granger method, which consists of a linear regression followed by the Dickey-Fuller test on the residuals of this regression. The main question was, which Dickey-Fuller case to use and whether the critical values of the Engle-Granger method are the same as those for this Dickey-Fuller test. We saw that case 2 was the most appropriate one for the way we want to test for cointegration, that is without a constant in the cointegrating relation. There was no indication, based on simulations, that the critical values from the Engle-Granger test differ from those of the Dickey-Fuller case 2 test. Also the power of the two tests were found similar when the assumptions of the method were fulfilled. The Engle-Granger test appeared to perform well, even when some assumptions were not fulfilled. The Engle-Granger test assumes that the residuals follow an autoregressive model. When we generated cointegrated data with residuals that are not likely to be autoregressive, the method still rejects the null hypothesis of no cointegration often.

IMC has provided a selection of ten pairs that are different in quality. In chapter 7 we applied the trading strategy from chapter 2 to the historical closing prices. Based on profitability and the number of trades, we find a distinction between good and bad pairs which coincides with the distinction made by IMC. In this chapter we also tested the ten pairs for cointegration, using both the Engle-Granger as well as the Johansen method. The two methods coincide on which pairs are cointegrated and which are not. Also the estimated cointegrating relations are almost the same. All the good pairs according to the trading strategy are seen as cointegrated, according to both tests. Furthermore all bad pairs are seen as not cointegrated according to both tests.

Based on the results of this project, we may conclude that cointegration is an appropriate concept to identify pairs suitable for IMC's trading strategy.

Chapter 9

Alternatives & recommendations

In this chapter we briefly discuss some alternative trading strategies in the first section and give some recommendations for further research in the second section.

9.1 Alternative trading strategies

In this report we focused on pair trading with two stocks in a pair. Two stocks being cointegrated is easily translated in the trading strategy from chapter 2, we take the spread process as the linear combination of the two stocks corresponding to the cointegrating vector:

$$y_t - \alpha x_t.$$

If we would take $\bar{r}$ from chapter 2 as the least squares estimate instead of the average ratio, the spread process of chapter 2 would be exactly the same as the spread process found with the Engle-Granger method. That is if we use the strategy without adjustment parameter κ, i.e., $\kappa = 0$.

In section 4.3 was stated that we neglect a possible constant in the cointegrating relation, α_0. In this section we will look at a trading strategy that does not neglect the constant. We also look at what can happen if we have cointegration between the logarithms of the stock prices.

Trading strategy with constant
Consider two stock price processes, x_t and y_t, which have the relation

$$y_t - \alpha x_t - \alpha_0 = \varepsilon_t \,, \tag{9.1}$$

where ε_t is some stationary process. In other words, the two stocks are cointegrated with a constant in their relation. We could trade the pair y, x with ratio $1 : \alpha$ and give up the cash neutral property, but another possibility is to determine the trading instances with (9.1) and trade a quantity of x such that the whole trade is cash neutral. More clearly, with (9.1) we can determine whether x_t is over- or underpriced compared to y_t at time t but we do not trade this relation, we trade one stock of y and y_t/x_t stocks of x if there was a mispricing larger than Γ at time t.

Let us consider an example, let x and y be a pair with relation (9.1) where $\alpha = 2$ and $\alpha_0 = 20$ such that spread ε_t looks figure 9.1. The corresponding processes for x_t and y_t are shown in figure 9.2.

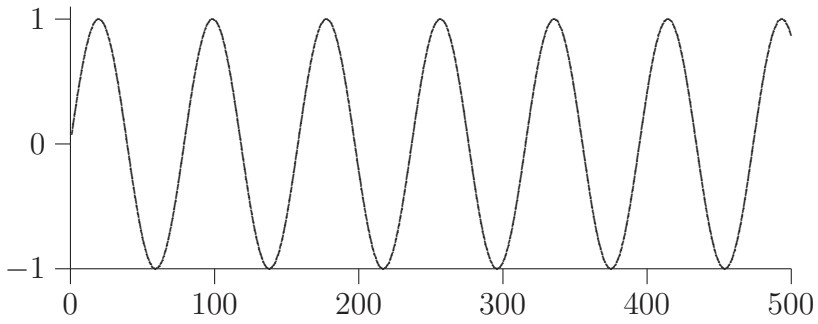

Figure 9.1: Spread ε_t.

For illustration purposes we took an artificial example. We have 500 observations, where we use the first half to determine the parameters of the strategy and the second half to see if the strategy works. We fit the first half of observations of y on the first half of observations of x and a constant, which results in:

$$\hat{\alpha} = 1.98 \quad \text{and} \quad \hat{\alpha_0} = 20.29$$

The threshold Γ is determined in the same way as in chapter 2, but now the spread process is the residuals from this fit. For this example it turned out to be that $\Gamma = 0.91$. We apply the new strategy to the second half of

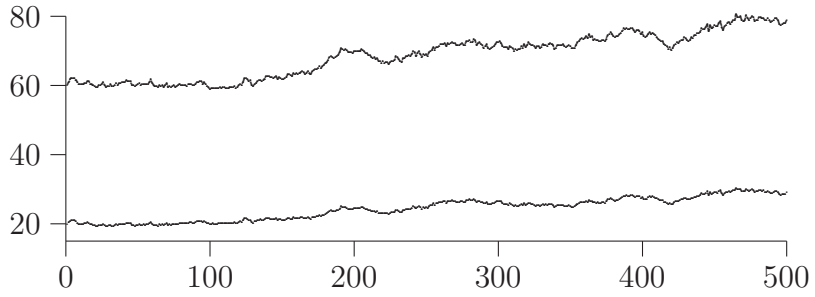

Figure 9.2: Price processes x_t and y_t.

observations. The trades are shown in table 9.1. With the first trade we put on a position for the first time, so we made no profit yet. The second trade consists of two parts, we flatten the position from the first trade which results in a profit and we put on a new position. We always trade one stock of y and we trade y_t/x_t number of stocks x so it is exactly cash neutral. The actual traded spread is not shown because it is basically the same as the right half of figure 9.1. Figure 9.3 shows the spread if we do not include a constant, i.e., if we neglected α_0.

Table 9.1: Trading instances.

trade	t	s_t	position (y,x)	price y_t	price x_t	profit
1	251	0.91	(-1,+2.82)	72.08	25.59	-
2	291	-0.93	(+1,-2.80)	66.87	23.90	0.44
3	331	0.93	(-1,+2.85)	70.20	24.63	1.27
4	370	-0.91	(+1,-2.73)	71.03	25.98	2.99
5	409	0.95	(-1,+2.74)	77.65	28.37	0.07
6	452	-0.94	(+1,-2.66)	77.05	29.02	2.37
7	487	0.93	(-1,+2.71)	79.86	29.49	1.55
				total profit		8.69

Although the profit for each trade is not at least 2Γ, as the profit for the trading strategy from chapter 2 with constant ratio was, it is still quite profitable to trade this pair this way. Specially because the trading strategy from chapter 2 would not make any money, even if we would have used a large adjustment parameter κ.

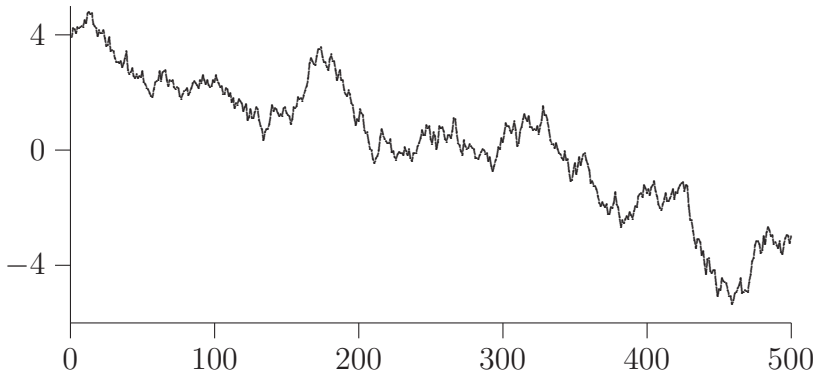

Figure 9.3: Spread when neglecting α_0.

Although this strategy can be applied for every α_0, we still do not want α_0 to be large because of the market neutral property of pair trading. If the overall market is up 50%, so x increases with 50% then we expect that y also increases with 50%. With a large α_0 compared to the stock prices, this does not hold. Actually it does not hold for any $\alpha_0 \neq 0$, but there is only a small effect when α_0 is small. The value of α_0 used in the example is actually too large, it is equal to the first observation of x. Which values of α_0 that can be used with this strategy, should be examined further.

Trading strategy for the logarithms
Assume we have two stock price processes and their logarithms are cointegrated:

$$\log y_t - \beta \log x_t = \varepsilon_t \,,$$

where ε_t is some stationary process. Then the relation between x_t and y_t becomes

$$y_t = x_t^{\beta} e^{\varepsilon_t}. \tag{9.2}$$

If $\beta = 1$, we can apply a trading strategy on the ratio process y_t/x_t instead of applying it on a spread process. An example is shown in figure 9.4, where we simulated x_t according to the model in section 4.2 and generated y_t such that ε_t follows a stationary $AR(1)$ model. A trading strategy could be to sell one stock of y and buy one stock of x when the ratio is above $1 + \Gamma$, and the other way around if the ratio is below $1 - \Gamma$. Or we could trade really cash neutral, so we trade one stock of y and y_t/x_t number of stocks of x.

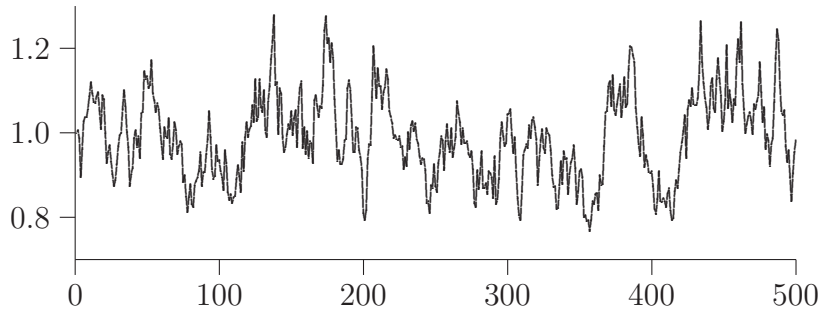

Figure 9.4: Ratio process y_t/x_t.

When $\beta \neq 1$, it is not that simple anymore. It is valid to say that when $\beta \neq 1$, $\beta > 1$. Because if it is not, we take y_t to be x_t and vice versa. Then we get the same problem as with a large α_0, the relation is not market neutral. If x increases with 50%, y increases more than 50% according to the relation (9.2). And if this happens when we have a long position in x and a short position in y, the profit in x cannot compensate the loss in y, so we lose money. Maybe this can be prevented for values of β close to 1, by adjusting the ratio in which we trade x and y but this should be examined further.

In this report we have only discussed trading strategies that trade one line, we put on a position when the spread reaches $\pm\Gamma$ and wait till the spread reaches Γ in the other direction. But it is very interesting to trade more lines. For example, if the spread reaches $+\Gamma_1$, we put on a short position in y and a long position in x. If the spread increases further and reaches Γ_2 we enlarge our short position in y and our long position in x. A trading strategy could be to trade the same amounts at each threshold, which are equally spaced, $\Gamma_2 = 2\Gamma_1$. Figure 9.5 illustrates this idea. To make this more clear, table 9.2 shows the trading instances for this strategy with two lines when we trade x and y in the ratio 1:1.

Figure 9.5: A 2 line strategy.

This strategy can be easily extended to more lines, it can even be easily extended to three or more stocks in a pair. How to choose the number of lines, the thresholds and the corresponding amount of stocks is very interesting to examine further.

9.2 Recommendations for further research

It would be nice to develop the alternative method from section 4.5 further, such that we have a new method for testing for cointegration. In order to do so, we need an accurate algorithm for the estimation of the parameters of the MA(q) model.

In this report we used closing prices. It is interesting to apply the trading strategies and cointegration tests to intra-day data because we trade during the day. This is specially interesting if we have a trading strategy with a large number of lines with the thresholds close to each other.

We could cut the cointegration test into several pieces. Suppose we have datasets containing four years of closing prices, than we could perform three tests on two years of data with an overlap of one year. More clearly, the first test is on the first and second year, the second test is on the second and

Table 9.2: Trading instances.

trade	t	s_t	position (y,x)
1	26	2.12	(-1,+1)
2	56	4.22	(-2,+2)
3	97	1.94	(-1,+1)
4	152	-0.01	flat
5	158	-2.11	(+1,-1)
6	199	-4.20	(+2,-2)
7	206	-1.89	(+1,-1)
8	221	0.13	flat
9	284	2.18	(-1,+1)
10	289	4.06	(-2,+2)
11	297	1.92	(-1,+1)
12	306	-0.13	flat

third year and the third test is on the third and fourth year. Then we can see if the stocks are cointegrated on each time interval and if the cointegrating relation changes. This could be very helpful to determine a good adjustment parameter κ.

There exists several representations for cointegrated processes, one is the VAR representation we saw briefly with the Johansen method in section 4.4. It would be interesting to see if it is possible to use one of the representations to build a monitoring system; a set of confidence intervals to see if the spread behaves according to the model and attach certain actions when the intervals are exceeded. For example, if the first confidence interval is exceeded we stop with enlarging our positions, if the second interval is exceeded we revert a part of our positions with a loss and if the third interval is exceeded we close out our entire positions and stop seeing the stocks as a pair.

Bibliography

[1] C. ALEXANDER. *Market Models*. John Wiley & Sons, 2001.

[2] P.J. BROCKWELL and R.A. DAVIS. *Introduction to time series and forecasting*. Springer-Verlag, 2002.

[3] P.J. BROCKWELL and R.A. DAVIS. *Time Series: Theory and Methods*. 1987, Springer-Verlag.

[4] D.A. DICKEY and W.A. FULLER. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74:427–431, 1979.

[5] R.F. ENGLE and C.W.J. GRANGER. Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2):251–276, 1987.

[6] J.D. HAMILTON. *Time Series Analysis*. Princeton University Press, 1994.

[7] D.J. HIGHAM. *An introduction to financial option valuation*. Cambridge University Press, 2004.

[8] S. JOHANSEN. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12:231–254, 1988.

[9] S. JOHANSEN. Estimation and hypothesis testing of cointegration vectors in guassian vector autoregressive models. *Econometrica*, 59:1551–1580, 1991.

[10] S. JOHANSEN and K. JUSELIUS. Maximum likelihood estimation and inference of cointegration - with application to the demand for money. *Oxford Bulletin of Economics and Statistics*, 52:208, 1990.

[11] M. OSTERWALD-LENUM. A note with quantiles of the asymptotic distribution of the maximum likelihood cointegration rank test statistics. *Oxford Bulletin of Economics and Statistics*, 54:462, 1992.

[12] P.C.B. PHILIPS and S.N. DURLAUF. Multiple time series regression with integrated processes. *Review of Economic Studies*, 53:473–495, 1986.

[13] P.C.B. PHILIPS and S. OULIARIS. Asymptotic properties of residual based tests for cointegration. *Econometrica*, 58(1):165–193, 1990.

[14] J.H. STOCK and M.W. WATSON. Testing for common trends. *Journal of the American Statistical Association*, 83(404):1097–1107, 1988.

[15] G. VIDYAMURTHY. *Pairs Trading*. John Wiley & Sons, 2004.