A BAYESIAN COINTEGRATION TESTING PROCEDURE WITH CONDITIONAL PRIOR ODDS

Analysis of pairwise cointegration in the US stock market

Master’s Thesis
Matti Karjalainen
Aalto University School of Business
Finance
Spring 2020
Abstract

In this thesis I introduce and evaluate the empirical performance of a Bayesian cointegration testing procedure. In contrast with previous Bayesian cointegration tests, the introduced procedure takes into account cross sectional variation in the prior probability of cointegration, and models this variation as a function of similarity metrics between company attributes. The procedure is motivated both by arbitrage pricing theory, as well as previous empirical findings on the attributes of companies with similarly behaving stock prices.

I analyze the daily closing prices of a sample of US-listed common stock during a ten-year period between 2009 and 2018. The former half of this sample is used for analyzing the cross sectional variation in the probability of pairwise cointegration for non-overlapping 12-month periods. This variation is explained by a set of similarity metrics based on the industry, book-to-market -ratios and market capitalizations of companies. I show that similarities between two companies’ industries and book-to-market -ratios positively affect the probability of pairwise cointegration. In contrast, similarity with respect to market capitalization is negatively associated with the probability of cointegration. I derive a logistic model for computing conditional prior odds of cointegration based on the similarity of company attributes, where the conditionality is with respect to an unobservable parameter representing the average cointegration probability.

Using the latter half of the sample, I test the ability of the introduced procedure in identifying persistent pairwise cointegrating relations. I classify pairs as cointegrated based on conditional posterior odds after the first six months of every 12-month period, and subsequently test whether the pair persists to be classified as cointegrated during the remaining six months. I show that the procedure is able to identify persistent pairwise cointegration even before the introduction of conditional prior odds. Further, when conditional prior odds are introduced, the empirical performance of the procedure is improved, especially when the classification threshold is high.

The research contribution of this thesis is threefold. Firstly, a conceptual contribution is made by introducing the concept of conditional prior odds, the first of its kind in pairwise cointegration testing between stocks. Secondly, the thesis offers an empirical contribution by analyzing the effects of company attributes on cointegration probability. Thirdly, the thesis offers practitioners an empirically tested procedure for the identification of persistent pairwise cointegration between stocks, that can be utilized in a pairs trading context.

Keywords Bayesian, cointegration, pairs trading, statistical arbitrage
Tekijä Matti Karjalainen

Työn nimi Bayesilainen yhteis integroitu vuuestausmenetelmä ehdollisella prioritodennäköisyydellä

Tutkinto Kauppatieteiden maisteri

Koulutusohjelma Rahoitus

Työn ohjaaja Vesa Putonnen

Hyväksymisvuosi 2020 Sivumäärä 55 Kieli englanti

Tiivistelmä

Tämä tutkielma esittelee bayesilaisen yhteis integroitu vuuestausmenetelmän, ja analysoi sen empiristä suorituskykyä osakeparikaupankäynnin näkökulmasta. Esiteltävän menetelmän poikkeaa aiemmista bayesilaisista testeistä siten, että se huomioi osakeparikohtaiset erot yhteis integroitu vuuestauksennäköisyydessä, mallintaa näitä yritysten samankaltaisuus mittareiden funktiona. Menetelmä pohjautuu arbitraašhinnoitteluteoriaan, sekä aimpin tutkimuksiin joissa on havaittu yhteis integroitu vuustauksen olevan yleisemmän samankaltaisten yritysten osakkeiden välillä.


Keywords bayesilaiset menetelmät, yhteis integroitu vuustauk, osakeparikaupankäynti, tilastollinen arbitraasi
Contents

Abstract ii
Tiivistelmä iii
Contents iv

1. Introduction 1

2. Review of literature and theoretical background 3
   2.1 Cointegration and shared exposure to stochastic trends 5
   2.1.1 Relation between company similarity and cointegration in previous empirical research 6
   2.2 Theoretical limitations of cointegration testing 7
   2.3 Persistence of identified pairwise cointegration in the stock market 8
   2.4 Bayesian inference in a cointegration context 9
   2.4.1 Bayesian inference in previous literature on pairs trading 11
   2.4.2 Prior odds in previous research 12
   2.4.3 Approximation of cross-sectional variation in prior odds 12
   2.5 Objectives and hypotheses 14
   2.5.1 Research questions and hypotheses 14

3. Methodology 17
   3.1 Bayes factor of cointegration 17
   3.1.1 Generative model for cointegration 18
   3.1.2 Prior and posterior of $\phi$ 19
3.1.3 Estimation of $\alpha$, $\beta$ and $\sigma^2$ ........................................... 21
3.1.4 Likelihood under competing hypotheses ................................. 22
3.1.5 Algorithm for Bayes factor .................................................. 22
3.2 Conditional prior odds of cointegration ........................................ 23
3.2.1 Attributes for similarity calculations ...................................... 23
3.2.2 Metrics for measuring attribute similarity ................................. 24
3.2.3 Logistic model for conditional prior odds .................................. 25
3.3 Conditional posterior odds of cointegration ................................. 27
3.4 Persistence of cointegration ..................................................... 28
3.5 Description of data ............................................................... 29
3.5.1 Retrieval and processing of data ............................................ 29
3.5.2 Partition of data into samples ............................................... 30

4. Empirical findings ................................................................. 32
4.1 Logistic model for conditional prior odds ................................. 32
4.1.1 Computation of Bayes factors in training sample ..................... 32
4.1.2 Estimation of relation between company similarity and Bayes factor 34
4.1.3 Parameterization of logistic model ........................................ 38
4.2 Persistence of cointegration .................................................... 39

5. Discussion and conclusions ...................................................... 45
5.1 Discussion on empirical findings ............................................. 45
5.1.1 Logistic model for conditional prior odds ............................... 45
5.1.2 Empirical performance of the introduced procedure .................. 46
5.1.3 Re-visiting the hypotheses .................................................. 48
5.1.4 Limitations and possible improvement areas of the study .......... 49
5.2 Concluding remarks ............................................................ 50

Bibliography ................................................................. 51

A. Appendix ................................................................. 54
A.1 System of equations for finding $\alpha$, $\beta$ and $\sigma^2$ ................... 54
A.2 Separate regressions of testing-implied cointegration on similarity metrics 55
1. **Introduction**

Identifying pairs of stocks whose prices tend to move similarly is an important step in the statistical arbitrage strategy known as pairs trading. This identification most often consists of simply analyzing the past behavior of stock prices. However, drawing inference of co-movement by simply looking at the past behavior of stock prices is likely to result in the identification of pairs where the observed co-movement is spurious, and not likely to persist. Spuriously identified co-moving relations result in excessive risk to a pairs trader, since the stock prices might start to diverge significantly.

Combining economic intuition with modern asset pricing theory, one could argue that persistently co-moving stocks should be similarly exposed to risk factors. Therefore, when analyzing co-movement between stock prices, one should also account for the similarity of their risk exposures. In this thesis, I present a Bayesian cointegration testing procedure that accounts not only for the historical co-movement between a pair of stocks, but also the similarity of their estimated risk exposures, in a novel way.

The utilization of Bayesian methods in some facets of economic research is challenging, since many quantities and related statistical models are unobservable. This is also true in the case of cointegration testing, as a cointegrating relation cannot be observed with absolute certainty. In the proposed procedure, the similarity of the risk exposures of two companies therefore needs to be accounted for with a novel approach. I parameterize a logistic model for the computation of what could best be described as conditional prior odds. The introduced approach allows taking into consideration the cross sectional variation in the probability of cointegration, driven by the degree of similarity between two companies. The prior odds are nonetheless conditional on the unobservable parameter of average cointegration probability, hence the term conditional prior odds.

Using data on US-listed common stock, I find evidence supporting a hypothesis that the
performance of a cointegration testing procedure is improved when the similarity between companies is taken into account with the introduction of conditional prior odds. In addition, by analyzing the attribute-specific effects of company similarity on cointegration probability, I find significant evidence suggesting that similarity with respect to market capitalization is, perhaps surprisingly, negatively associated with the probability of cointegration.

The contribution of this thesis to existing research is threefold. Firstly, a conceptual contribution is made by introducing the concept of conditional prior odds, that allow us to analyze cross sectional variation and its sources in an unobservable parameter. To the best of my knowledge, this thesis is the first to utilize any type of prior odds in testing for pairwise cointegration between stock prices. Secondly, the thesis offers an empirical contribution by analyzing the effects of company attributes on cointegration probability in a novel way. As far as I am aware, previous works on the applications of cointegration testing in pairs trading have accounted for company attributes only by restricting the formation of pairs. Thirdly, the thesis offers practitioners an empirically tested procedure for the identification of persistent pairwise cointegration between stocks, and could be utilized in a pairs trading context as is.

The remainder of this thesis is structured as follows. First, in Chapter 2, I review the theoretical background and previous literature relevant to the topic, specifying the research questions and hypotheses of the thesis. Then, in Chapter 3, I show how the introduced procedure is derived, how its performance is tested empirically, and describe the data used for the empirical analysis. Following this, I report the empirical findings of fitting the model for computing conditional prior odds, and testing the performance of the introduced procedure in Chapter 4. Chapter 5 discusses the findings and their implications, followed by concluding remarks.
2. **Review of literature and theoretical background**

The concept of pairs trading involves identifying a pair of securities, such as stocks, whose prices have exhibited a degree of co-movement. The observed co-moving relation is assumed to persist, and this relation is assumed to follow some statistical model where the prices tend to revert towards an equilibrium. The idea behind pairs trading is simple: the strategy aims to identify when the relation between two co-moving securities has temporarily diverged from the equilibrium of the estimated statistical model, and seeks to profit from its reversion towards the equilibrium. The profit from this reversion is captured by constructing a long-short portfolio of the two securities. As such, pairs trading is considered a statistical arbitrage strategy, and has been known to be used among investment professionals since at least the 1980’s (Gatev et al., 2006).

As Gatev et al. (2006) point out, such a simple strategy should not yield positive abnormal returns under the efficient market hypothesis. However, the authors find largely positive and statistically significant historical risk-adjusted returns for the simple strategy, with average annualized abnormal returns of 12% for the best performing long-short portfolios. Consistent with the hypothesis of Jegadeesh and Titman (1995), Gatev et al. (2006) hypothesize that the profits from pairs trading, much like other contrarian investment strategies, could arise from undisciplined over-reaction of individual investors.

While pairs trading strategies are both market neutral and self-financing, they are of course not devoid of risk. As suggested by Gatev et al. (2006), arguably the most prominent risk in pairs trading is the possibility that the pair of securities might not revert towards the equilibrium during the holding period. Even worse, the pairs might diverge further from the equilibrium, incurring potentially substantial or even catastrophic losses to the holder of the long-short portfolio.

Gatev et al. (2006) estimate co-movement between stocks non-parametrically by minimizing the squared deviations between normalized price paths. Krauss (2017) argues that this approach
is econometrically rather unstable, and therefore the risk of divergent trades is large. Moreover, Do and Faff (2010) suggest that the approach of Gatev et al. (2006) has lost its profitability since the early years of the 21st century. Some of the more recent research [see for example Caldeira and Moura (2013); Ardia et al. (2016); Cummins and Bucca (2012); Elliott et al. (2005)] models co-movement through cointegration. According to Krauss (2017), the main benefit of using cointegration is the econometrically more reliable equilibrium relation between the identified pairs. However, Rad et al. (2016) demonstrate that the number of trading opportunities has decreased from the year 2009 onwards when using cointegration methods for the selection of pairs, which would slightly hinder the usability of such strategies in the industry. Wu (2013) and Xie et al. (2016) model stock co-movement with copulas. Rad et al. (2016) however suggest that modeling co-movement with copulas tends to result in a high proportion of trades that fail to revert towards the estimated equilibrium.

While possible reasons for the lack of reversion are numerous, they can be broadly assigned into two distinct categories:

1. The dynamics of the relation change abruptly, if one of the securities experiences an idiosyncratic shock with permanent effects on the security price.

2. The observed historical co-movement relation between two securities was in fact spurious, and simply due to chance.

Unfortunately, hedging against permanent idiosyncratic shocks is difficult, if not impossible for a single pair of stocks. However, the identification of spurious co-movement relations between stock pairs could potentially be mitigated by an improved test for co-movement. Indeed, testing for co-movement between stock prices with a brute force data-mining approach, one is likely to find a number of stock pairs that seem to have exhibited a degree of co-movement historically, with no other factor driving the co-movement than pure chance. This issue is even more serious when analyzing shorter time windows.

By taking into account economic intuition, one could potentially decrease the number of pairs that are spuriously identified as co-moving. The main idea behind such an approach is simple: if one cannot explain why the prices of a pair of stocks should move together, one has more reason to believe that the identified co-movement is spurious. On the other hand, if one
believes that the two stocks are for instance similarly exposed to a set of risk factors, their prices should also behave in a similar manner and thus one has less reason to doubt the validity of the identified relation. In addition to such a simple qualitative reasoning, we can also support this view mathematically as illustrated in the following section.

2.1 Cointegration and shared exposure to stochastic trends

While there are numerous ways to measure the degree of co-movement, let us turn our focus to testing for cointegration between unit-root processes from hereon. Unit-root processes are said to be cointegrated when a linear combination of them results in stationary time series. Theoretically, this definition is fulfilled when the processes share exposure to common stochastic trends (also called factors). For a pair of stock prices, exposure to common factors could be interpreted as shared exposure to the development of a certain industry, or similar loadings on a set of risk factors. The idea of stock prices and returns being driven by a set of common factors dates back to the seminal papers on arbitrage pricing theory [see Ross (1976) for a theoretical derivation and Roll and Ross (1980) for an empirical analysis] and its extensions.

The relation between exposure to common factors and cointegration has been discussed in a variety of contexts in previous literature. Bossaerts (1988) and Kazi (2008) extend this relation to the analysis of cointegration between prices of common stock. Stock and Watson (1988) and Gonzalo and Granger (1995) apply such a relation to testing for cointegration between macroeconomic variables. Hasbrouck (1995) shows how the relation determines the prices of cross-listed securities. While Gonzalo and Granger (1995) rigorously illustrate the vector error correction model relation between common factors and cointegrated macroeconomic variables mathematically [utilizing proofs from Johansen and Juselius (1991)], the basic intuition between cointegration and factor exposure can also be illustrated through the following, excessively simplified example.

Suppose that vector $p_t \in \mathbb{R}^n_+$ contains the prices of $n$ of risk factors at each time $t$, where each factor follows a separate unit-root process. Let stock prices $x_t$ and $y_t$ follow

$$x_t = \alpha_x + \beta'_x p_t + \nu_t, \quad y_t = \alpha_y + \beta'_y p_t + \eta_t,$$

(2.1)

where $\alpha_x$ and $\alpha_y$ are scalar constants, $\beta_x, \beta_y \in \mathbb{R}^n$ are constant vectors representing the factor loadings for the two stocks respectively, and $\nu_t$ and $\eta_t$ are stationary univariate processes. Let $s_t$
be a linear combination of $x_t$ and $y_t$, more specifically

$$s_t = \gamma x_t - y_t = \gamma \alpha_x + \gamma (\beta'_x p_t) + \gamma \nu_t - \alpha_y - \beta'_y p_t - \eta_t.$$  \hspace{1cm} (2.2)

If we subtract all stationary elements, we are left with

$$\gamma (\beta'_x p_t) - \beta'_y p_t.$$ \hspace{1cm} (2.3)

Setting this to zero, we get the relation

$$(\gamma \beta'_x) p_t = \beta'_y p_t,$$ \hspace{1cm} (2.4)

and therefore when $\beta_x$ is a linear transformation of $\beta_y$, the process $s_t$ is stationary. However, the stationary but stochastic elements $\nu_t$ and $\eta_t$ could also behave in a manner that makes $s_t$ appear stationary, especially when analyzing shorter periods. While the example is extremely simplified, it nonetheless illustrates the relation between risk-exposure and cointegration, as well as the possible sources of spurious cointegration. The example is also in line with economic intuition, implying that similarity between companies (e.g. their risk exposures) is closely related to cointegration between their stocks.

When testing for cointegration using historical price paths, a test is likely to identify a number of seemingly cointegrated pairs where the economic intuition for cointegration is hard to justify. For example, a pair of stocks of companies in vastly different industries might be identified as cointegrated, without any plausible explanation as to why the stocks would share exposure to a stochastic trend. If the identified relation is indeed due to chance, failure to identify such spuriousness could result in a diverged trade, and possibly significant losses to the holder of a long-short portfolio.

### 2.1.1 Relation between company similarity and cointegration in previous empirical research

The link between company similarity and co-movement between stock prices is also consistent with empirical findings from previous research. Gatev et al. (2006) identify co-moving pairs by analyzing the squared deviations between their price paths, and form pairs trading portfolios by finding pairs where the squared deviations are minimized. In addition to evaluating the abnormal
returns of such portfolios, Gatev et al. (2006) also analyze the characteristics of firms that form the pairs with the smallest squared deviations between normalized price paths. The authors find that companies behind co-moving stock pairs have a tendency to belong to the same industry, and be of similar size in their market capitalizations. More specifically, the authors find that the most profitable pairs consist largely of utility and industrial stocks, and the market capitalizations of co-moving stocks differ by one CRSP\(^1\) decile on average. In addition, almost 70% (90%) of co-moving pairs belong to the top 30% (50%) largest stocks in the CRSP database, based on market capitalization. While the authors report findings suggesting that company similarity and stock co-movement appear to be somewhat interdependent, their approach for finding co-moving pairs draws inference from historical price time-series data alone. Such an approach does not attempt to alleviate the issue of potentially having no economic link between the stock pairs.

Figuerola-Ferretti et al. (2018) analyze the performance of pairs trading portfolios in the European stock market, using a cointegration approach in selecting co-moving pairs. In their baseline procedure, the authors restrict the formation of pairs to stocks within the same industry. The authors find that when pairs are formed by also applying univariate restrictions on their book-to-market -ratios and trading volume, the economic performance of the portfolios improves. On the contrary, no improvement is observed when restricting pairs by their market capitalizations. The authors do not perform the analysis past bivariate restrictions, due to limitations in the sample size, but their findings could nonetheless be seen to advocate for the benefits of accounting for economic intuition, by considering company similarity in the cointegration testing setup.

2.2 Theoretical limitations of cointegration testing

It is important to clarify the distinction between the presence of cointegration with absolute certainty, and the presence of cointegration implied by the result of a statistical test. Since the actual statistical model that determines whether or not a pair of stocks is cointegrated is unobservable, we cannot observe the set of pairs that are cointegrated with absolute certainty either. This raises a number of issues, such as not being able to determine the specificity or sensitivity of a cointegration test, not knowing the true proportion of stock pairs that are in fact cointegrated, or even if pairwise cointegration between stocks exists altogether.

When using cointegration testing for identifying co-movement in pairs trading, the underlying

\(^{1}\text{Center for Research in Security Prices}\)
assumption is that the identified cointegrating relations persist. Since it is not possible to observe whether a cointegration relation actually exists, we must draw inference of this persistence in some other way, mainly by evaluating the persistence of cointegration as implied by a statistical test. As demonstrated by the following example, the inability to observe actual cointegrating relations leads to issues in identifying the persistence of pairwise cointegration.

Consider the following. We identify a pair of stocks to be cointegrated in some period, based on the result of a cointegration test. We then repeat the cointegration test in a subsequent period, and observe that the test no longer suggests the pair to be cointegrated. There are many possible reasons for observing such an event. Firstly, it is possible that the test correctly identified the pair to be cointegrated in the first period and not cointegrated in the second period, implying that the pair was cointegrated initially, but the cointegrating relation did not persist until the end of the second period. The lack of persistence could arise from one stock experiencing a permanent idiosyncratic shock during the second period, for example. Secondly, it is also possible that the test results were inconsistent with the actual state of the relation in either period, and the pair persisted to be cointegrated or not cointegrated through both of the periods. Especially when a cointegration test has low power, this is a plausible explanation.

Fortunately, a pairs trader might not be concerned with being able to identify cointegration with absolute certainty. It should suffice for a trader that the pairs of stocks that were initially identified as cointegrated, generally tend to revert towards the equilibria of the estimated cointegrating relations. Such reversion is likely to result in a positive test result for cointegration in the subsequent period, even if the true state of affairs is still unknown. Therefore, while the formulation of a cointegration testing procedure is faced with significant theoretical obstacles, such a procedure might nonetheless be of value for a practitioner.

### 2.3 Persistence of identified pairwise cointegration in the stock market

The fundamental assumption in identifying co-moving stock pairs by analyzing their historical price paths, regardless of the method used for identifying co-movement, is that the co-movement identified in past data will persist in subsequent periods as well. If this assumption did not hold, the identification of past co-movement would not be informative for constructing a pairs trading portfolio. Therefore, I find it striking how little emphasis previous research on pairs trading has placed on explicitly measuring the persistence of the identified co-movement between pairs.
On one hand, the reported positive abnormal returns from back-testing pairs trading strategies could be seen to indicate that co-movement does in fact persist. On the other hand, such returns only imply that on average, the pairs did move towards the estimated equilibrium during the holding period. However, this observation does not directly imply that the pair would have been identified as co-moving during the holding period.

Clegg (2014) explicitly analyzes the persistence of co-movement between stock prices using a variety of frequentist cointegration testing procedures, and finds little evidence supporting the hypothesis of persistent pairwise cointegrating relations within US equities. This creates a puzzle with respect to the source that drives the reported profitability of pairs trading strategies that rely on cointegration for modeling co-movement. The failure to find persistent cointegrating relations could imply that cointegration is not an optimal way of modeling co-movement between stock prices. Additionally, the author tests for cointegration by evaluating consecutive 12-month periods. Finding pairs of stocks where the cointegration relation persists for a total of 24 months could be seen as unlikely, due to the high probability of permanent idiosyncratic shocks to either stock during such a long evaluation period. However, Clegg (2014) also points out that the failure to detect persistent cointegrating relations is not an indication that such relations do not exist, but suggests that the findings could also arise from the low power of the cointegration testing procedures used in the empirical analysis. This is a key motivating observation advocating for the need to formulate a cointegration testing procedure with higher power and robustness when compared to traditional frequentist cointegration testing procedures.

2.4 Bayesian inference in a cointegration context

The idea behind Bayesian hypothesis testing differs distinctly from that of its frequentist counterpart. In frequentist hypothesis testing, we evaluate the probability of encountering the observed data under a null-hypothesis. If the probability of "observing a sample that is equally or more extreme than the data we have observed" is lower than a pre-determined threshold, we reject the null-hypothesis. Note that this approach does not estimate the probability of a hypothesis $H$ being true given the observed data $P(H|D)$, but rather $P(D^+|H)$, where $D^+$ refers to a sample of data as or more extreme than the one observed.

In Bayesian inference the testing setup is reversed, and we evaluate the probability of a
hypothesis, conditional on the observed data. We have the conditional probability

\[ P(H|D) = \frac{P(D|H)P(H)}{P(D)}, \]  

(2.5)

where \( D \) and \( H \) denote the data and hypothesis respectively. The term \( P(D|H) = \mathcal{L}(H|D) \) is called the likelihood function, which evaluates the likelihood of the hypothesis, conditional on the observed data. The above can be expanded to comparing the probability of two competing hypothesis by computing their ratio, canceling out \( P(D) \).

\[ \frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)P(H_1)}{P(D|H_2)P(H_2)} = \frac{P(D|H_1)}{P(D|H_2)} \times \frac{P(H_1)}{P(H_2)}. \]  

(2.6)

This ratio is called the posterior odds, and it is simply the product of the ratio of the likelihood functions i.e. the Bayes factor\(^2\), multiplied by the ratio of the prior probabilities of the hypotheses, i.e. the prior odds. When testing for cointegration, the posterior odds are given by

\[ \frac{P(C|D)}{P(RW|D)} = \frac{P(D|C)}{P(D|RW)} \times \frac{P(C)}{P(RW)}, \]  

(2.7)

where cointegration and random walk are denoted by \( C \) and \( RW \) respectively. This is equal to the Bayes factor of cointegration, multiplied by the prior odds of cointegration\(^3\). The Bayes factor of cointegration is the ratio between the likelihood functions under a cointegration- and a random walk -hypothesis. A cointegration hypothesis implies that a linear combination of the two stock prices results in a stationary process, while the random walk hypothesis implies that any linear combination of the two prices is a random walk. When the posterior odds are greater than unity, it is more likely that the two time series are cointegrated.

Since the actual statistical model that determines cointegration is unobservable for any pair of stocks, the quantities \( P(C) \) and \( P(RW) \) cannot be observed for pairs of common stock. Therefore, the prior odds of cointegration is also an unobservable quantity. Perhaps not surprisingly, previous research on Bayesian cointegration methods in the context of pairs trading has neglected the prior odds parameter altogether. Instead, inference is drawn from the Bayes factor alone, and the threshold for identifying a pair as cointegrated through the Bayes factor is set higher

\(^2\)Note that the Bayes factor is not the ratio of the maximum likelihood estimates but rather the ratio of the likelihood functions integrated over the full parameter space of the model.

\(^3\)For the Bayes factor, as well as prior and posterior odds, it is customary to call the ratios by the hypothesis corresponding to the numerator. In this case, comparing cointegration to its complementary event, i.e. a random walk, the ratios are simply called the Bayes factor of cointegration, prior odds of cointegration, or the posterior odds of cointegration.
than unity. Theoretically, this approach implies that we hypothesize that the true proportion of cointegrated pairs in the stock market is unknown, but lower than 50%.

2.4.1 Bayesian inference in previous literature on pairs trading

While some of the most notable publications in the area of detecting co-movement in pairs trading strategies take a frequentist approach to the concept of probability, Bayesian approaches have also been explored, especially when dealing with cointegration methods. Bayesian approaches are applied for both residual-based cointegration tests (Bracegirdle and Barber, 2012; Furmston et al., 2013) as well as error-correction models (Ardia et al., 2016). Arguably the main benefit of the introduced Bayesian approaches is the ability to evaluate the entire parameter space of possible cointegrating relations, as opposed to a single system determined by coefficient point estimates.

Evaluating the full parameter space of a cointegrating relation between a pair of stock prices has the advantage of allowing us to obtain posterior distributions for each parameter of the statistical model, and therefore also allowing us to account for the uncertainty of each parameter. However, this approach also increases the computational complexity of the analysis considerably (Bracegirdle and Barber, 2012). Since the dependence structure of parameters in a cointegration system is often increasingly complex, analytic derivation of the parameter distributions is difficult, if not impossible. Therefore, instead of integrating over each parameter of the system to obtain likelihood estimates for the data, one often needs to resort to Markov-Chain Monte Carlo (MCMC) approaches. When the distribution of interest is complex, the convergence of the Markov Chain is often slow. Furmston et al. (2013) show that a fully Bayesian approach with full posterior distributions for all parameters performs better than a partially Bayesian approach with some distributions replaced by point estimates, when the procedures are tested on simulated data where the data generating process is known. The authors point out that when the residual process of the cointegrating relation is assumed to follow a higher order autoregressive process, MCMC sampling is necessary for performing posterior inference. However, Furmston et al. (2013) also show that when run on simulated data, the partially Bayesian approach clearly outperforms a typical two-step cointegration test⁴, and does not fall far behind the fully Bayesian model.

⁴The two-step cointegration test consists of running an ordinary least squares (OLS) regression on the two time-series, and subsequently using (an augmented) Dickey-Fuller test for the stationarity of the residuals from the first-stage regression (see for example Granger (1986); Harris and Sollis (2003)).
2.4.2 Prior odds in previous research

Previous research on the applications of Bayesian cointegration testing in pairs trading such as Ardia et al. (2016), Bracegirdle and Barber (2012), Furmston et al. (2013) have utilized the Bayes factor for drawing inference about a possible cointegrating relation, rather than computing the posterior odds of cointegration. This approach is equivalent to using prior odds equal to unity for each pair, and it makes two assumptions about the probability of cointegration that could be challenged.

Firstly, from a theoretical standpoint, using prior odds equal to unity implies that a priori, it is assumed equally likely that a pair of stocks is cointegrated than it is for them not to be cointegrated. However, since it is not possible to observe the statistical model that determines the presence of cointegration, the correct level of prior odds can not be determined. To address this issue, the threshold for classifying a pair of stocks as cointegrated based on the Bayes factor is set high. In fact, from a practical standpoint, increasing this threshold is equivalent to decreasing the prior odds, assuming that prior odds are constant across all pairs. Generally, such issues arising from the inability to observe statistical models could be seen as a major obstacle in using Bayesian methods in certain facets of economic research.

Secondly, using identical prior odds for all pairs of stocks implies that a priori, all stock pairs are assumed equally likely to be cointegrated. However, as per the findings of Gatev et al. (2006) and Figuerola-Ferretti et al. (2018), as well as intuition, cointegration is more likely to exist between stocks of similar companies. Therefore, in addition to accounting for the average probability of cointegration for pairs of stocks, it could also be argued that proper prior odds should account for the cross sectional variation with respect to the similarity of pairs of companies, and its effects on the probability of cointegration.

2.4.3 Approximation of cross-sectional variation in prior odds

Choosing not to utilize prior odds in cointegration testing results in omitting information not only about the unobservable average probability of cointegration, but also the cross sectional variation in the probability of cointegration. However, the inability to determine the average absolute level of cointegration probability does not prohibit us from observing the effects of an explanatory variable on the relative cointegration probability between pairs of stocks. In other words, while the actual cointegration probabilities for pairs of stock are unknown, we can nonetheless observe
whether some characteristics, such as the similarity of the companies, lead to higher or lower than average cointegration probabilities.

Consider the following simplified example. Suppose that the output of a similarity function $S$ between two companies $X$ and $Y$ is a binary variable, where $S(X, Y) = 1$ implies that the companies are similar and $S(X, Y) = 0$ implies that the companies are dissimilar. Let $T_C$ (cointegrated) and $T_{RW}$ (random walk) denote the results of a cointegration test, and $C$ (cointegrated) and $RW$ (random walk) represent the true state of the relation. The ratio

$$
\frac{P(T_C|S(X, Y) = 1)}{P(T_C|S(X, Y) = 0)}
$$

is an observable quantity, and takes a value greater than unity if similar companies are more likely to be identified as cointegrated than dissimilar companies. Simplifying the notation by denoting $S(X, Y) = 1$ by $S_1$ and $S(X, Y) = 0$ by $S_0$ and incorporating the actual nature of the relation into the expressions, the ratio can be expressed by

$$
\frac{P(T_C|C, S_1)P(C|S_1) + P(T_C|RW, S_1)P(RW|S_1)}{P(T_C|C, S_0)P(C|S_0) + P(T_C|RW, S_0)P(RW|S_0)}.
$$

(2.8)

Assume further that the sensitivity and specificity of the cointegration test are independent of similarity, that is

$$
P(T_C|C) = P(T_C|C, S_1) = P(T_C|C, S_0),
$$

$$
P(T_C|RW) = P(T_C|RW, S_1) = P(T_C|RW, S_0),
$$

(2.10)

and given that

$$
P(RW|S_1) = 1 - P(C|S_1),
$$

$$
P(RW|S_0) = 1 - P(C|S_0),
$$

(2.11)

dividing both the numerator and denominator by $P(T_C|RW)$ gives us the ratio

$$
\frac{R_{TF} \times P(C|S_1) + (1 - P(C|S_1))}{R_{TF} \times P(C|S_0) + (1 - P(C|S_0))} = \frac{P(T_C|C)}{P(T_C|RW)},
$$

(2.12)

where $R_{TF}$ is the true positive rate divided by the false positive rate. Asymptotically,

$$
\lim_{R_{TF} \to \infty} \frac{(R_{TF} - 1)P(C|S_1) + 1}{(R_{TF} - 1)P(C|S_0) + 1} = \frac{P(C|S_1)}{P(C|S_0)}.
$$

(2.13)
Therefore, when the power of the cointegration test is high,

\[
\frac{P(T_C|S_1)}{P(T_C|S_0)} \sim \frac{P(C|S_1)}{P(C|S_0)},
\]  

(2.14)

The effects of company similarity on the probability of cointegration can therefore be approximately observed, as demonstrated by the simplified example above. While a proper functional form for the prior odds cannot be estimated since the true set of cointegrated stock pairs cannot be observed, we can nonetheless approximately infer the cross sectional variance in the cointegration probabilities and the relative effects of similarity on the cointegration probability.

Should we observe any significant relation between similarity and the relative probability of cointegration, this relation should be taken into account in the prior odds. Prior odds that consider such cross sectional variation in the cointegration probability, but whose absolute value is conditional on the unobserved average cointegration probability, could be called conditional prior odds, as they contain information about the prior probability conditional on the average level. Using prior odds equal to unity could be called non-informative, since the odds do not attempt to provide information about the cross-sectional variation in the prior probability of cointegration, or the aggregate level of prior probability. To the best of my knowledge, none of the previous literature on Bayesian cointegration testing in the context of pairs trading attempt to utilize conditional prior odds.

### 2.5 Objectives and hypotheses

The objective of this thesis is to propose a robust Bayesian procedure for cointegration testing between stocks, and test its performance in a pairs trading context empirically. More specifically, the goal is to derive a procedure that accounts for the plausibility of an economic link between the companies, as well as the likelihood of the observed stock price path under a cointegration hypothesis. The economic link considers the similarity between companies, using data on their fundamental attributes.

#### 2.5.1 Research questions and hypotheses

In order to formulate the procedure and to test it’s empirical performance in a pairs trading context, one must answer the following questions.
1. Which attributes are generally similar between companies with stock prices that are estimated to be cointegrated, and how significant are the effects of the estimated similarity on testing-implied cointegration (in-sample)?

2. Is testing-implied cointegration between stock prices a persistent quality?

3. Can the estimated similarity between fundamental attributes of companies improve the identification of testing-implied persistent cointegration?

The above questions lead to the definition of the two main hypotheses that I set out to test in this thesis. The first hypothesis is related to the idea that stocks of similar companies should more likely appear cointegrated than stocks of dissimilar companies. The null and alternative hypothesis are defined as:

- **Hypothesis 1₀**: Similarity between companies is not associated with the probability of testing-implied cointegration.

- **Hypothesis 1₁**: Similarity between companies is associated with the probability of testing-implied cointegration.

The second hypothesis is related to the persistence of testing-implied cointegration relations, and its null and alternative hypothesis are respectively defined as:

- **Hypothesis 2₀**: Pairs of stocks that are classified as cointegrated by a Bayesian cointegration testing procedure, are not more likely to be classified as cointegrated in subsequent periods.

- **Hypothesis 2₁**: Pairs of stocks that are classified as cointegrated by a Bayesian cointegration testing procedure, are more likely to be classified as cointegrated in subsequent periods.

It is worth pointing out, that testing the performance of a pairs trading strategy that uses the proposed Bayesian procedure for the selection of cointegrated pairs would also be an interesting area of research. However, testing for possible abnormal returns of such a strategy would not
only consist of testing the effects of using the Bayesian procedure, but also jointly testing the effects of the exact specifications of the particular pairs trading strategy. In order to maintain a clear focus, analyzing the returns of a pairs trading strategy based on the proposed procedure is thus out of scope for this thesis.
3. Methodology

This section introduces the methodology used in formulating the Bayesian cointegration testing procedure, as well as the methodology for testing its empirical performance. To start the chapter, I show the derivation of the algorithm used for computing the Bayes factor of cointegration, based on Bracegirdle and Barber (2012). I then present the steps of the empirical analysis, which include analyzing the in-sample effects of company similarity on the Bayes factor of cointegration, parameterization of a logistic model for the computation of conditional prior odds, and testing for the persistence of cointegration in subsequent periods, when inference of cointegration is drawn from conditional posterior odds.

3.1 Bayes factor of cointegration

Following Bracegirdle and Barber (2012), I opt to use a residual-based cointegration model in the computation of the Bayes factor of cointegration. This approach simplifies the computational process by using point estimates for some of the (non-key) parameters instead of their full distributions, and could thus be described as partially Bayesian. The reason for choosing this approach is mostly due to the relatively low computational complexity of such a partially Bayesian approach, and the fact that previous research [see Furmston et al. (2013)] shows the empirical performance of a fully Bayesian approach to be only slightly superior. The following subsections are largely based on Bracegirdle and Barber (2012).
3.1.1 Generative model for cointegration

Consider two non-stationary time series of stock prices $x_t$ and $y_t$. The two time series are said to be cointegrated if there exists a linear combination that results in a stationary series $\epsilon_t$. A simple parameterization for $\epsilon_t$ is a first order stationary autoregressive [AR(1)] process with Gaussian innovations.

$$y_t - \beta x_t - \alpha = \epsilon_t$$

$$\epsilon_t = \phi \epsilon_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \sigma^2), \quad |\phi| < 1$$ (3.1)

It is worthy of pointing out that while an assumption of normality is made for the innovations in $\epsilon_t$, the processes of $x_t$ and $y_t$ are not specified, and thus do not rely on any assumptions of normality. To continue avoiding such strong assumptions, let us start with a generative model for the observations $y_t$ and latent residuals $\epsilon_t$.

$$p(y_{1:T}, \epsilon_{1:T} | x_{1:T}) = \prod_t p(y_t | x_t, \epsilon_t) p(\epsilon_t | \epsilon_{t-1}), \quad \epsilon_0 = \emptyset$$ (3.2)

The transition $p(y_t | x_t, \epsilon_t)$ is degenerate, and can be modeled with the Dirac delta distribution.

$$p(y_t | x_t, \epsilon_t) \sim \delta(y_t - \alpha - \beta x_t - \epsilon_t)$$ (3.3)

and given that our specification for the stationary AR(1) residual process assumes Gaussian innovations, the transition $p(\epsilon_t | \epsilon_{t-1})$ is

$$p(\epsilon_t | \epsilon_{t-1}) \sim \mathcal{N}(\phi \epsilon_{t-1}, \sigma^2),$$ (3.4)

where $\sigma^2$ is finite and constant. The likelihood for $y_{1:T}$ is then given by integrating over the residuals

$$p(y_{1:T} | x_{1:T}) = \int_{\epsilon_{1:T}} p(y_{1:T}, \epsilon_{1:T} | x_{1:T}).$$ (3.5)

---

1 Clegg (2014) suggests that practitioners might use either ordinary prices or log prices. In theory, the two approaches model slightly different phenomena. When using log prices, we are essentially modeling cointegration between cumulative returns of two assets, and conversely, using regular price we are modeling cointegration between their price paths. For a pairs trader, this distinction relates to the weights of the assets in the portfolio, which should depend on market values when modeling cointegration with log prices, and the number of shares when modeling cointegration between ordinary prices.

2 The process for $\epsilon_t$ could also be modeled by a more complex model, but for simplicity the subsequent derivations are based on an AR(1) process.
Factorizing the integrand according to 3.2, we get

$$p(y_{1:T}|x_{1:T}) = \prod_{t=1}^{T} \int_{\epsilon_t} p(y_t|x_t, \epsilon_t)p(\epsilon_t|\epsilon_{t-1}).$$ \hspace{1cm} (3.6)

Given that $p(y_t|x_t, \epsilon_t)$ is degenerate and $p(\epsilon_t|\epsilon_{t-1})$ is Gaussian, we have

$$\int_{\epsilon_t} p(y_t|x_t, \epsilon_t)p(\epsilon_t|\epsilon_{t-1}) = \mathcal{N}(y_t - \alpha - \beta x_t|\phi(y_{t-1} - \alpha - \beta x_{t-1}), \sigma^2)$$

$$= \mathcal{N}(\epsilon_t|\phi \epsilon_{t-1}, \sigma^2),$$ \hspace{1cm} (3.7)

and therefore the likelihood can be expressed as a product of the Gaussian residual transitions.

By factorizing away the initial residual, the likelihood $p(y_{1:T}|x_{1:T})$ can be expressed by

$$p(y_{1:T}|x_{1:T}) = p(\epsilon_1) \times \prod_{t=2}^{T} \mathcal{N}(\epsilon_t|\phi \epsilon_{t-1}, \sigma^2)$$

$$= p(\epsilon_1) \times \exp \left( \frac{1}{2\sigma^2} \sum_{t=2}^{T} (\epsilon_t - \phi \epsilon_{t-1})^2 \right) \times (\log 2\pi \sigma^2)^{-\frac{T-1}{2}}$$ \hspace{1cm} (3.8)

$$= p(\epsilon_{1:T}),$$

and the log likelihood is

$$\log p(\epsilon_{1:T}) = \log p(\epsilon_1) - \frac{1}{2\sigma^2} \sum_{t=2}^{T} (\epsilon_t - \phi \epsilon_{t-1})^2 - \frac{T-1}{2} \log 2\pi \sigma^2.$$ \hspace{1cm} (3.9)

Thus, if we assume $\phi$ to be zero, maximizing the log likelihood is achieved by minimizing the squared residuals. However, when $\phi$ is not equal to zero, a maximum likelihood approach might yield results that are different from those of the OLS regression. In other words, if the residual process is autocorrelated, using OLS might no longer be optimal.

3.1.2 Prior and posterior of $\phi$

Let us shift our focus to the distribution of the autoregressive parameter $\phi$, which is treated as a latent variable. Under the cointegration hypothesis, $\phi$ must fulfill the stationarity requirement $|\phi| < 1$, and therefore its prior distribution, denoted $p(\phi)$, can be assumed to follow $\mathcal{U}(-1,1)$.

The posterior for $\phi$ can be obtained by first formulating a joint probability model for $y_{1:T}, \epsilon_{1:T}$
and \( \phi \), conditional on \( x_{1:T} \).

\[
p(y_{1:T}, \epsilon_{1:T}, \phi | x_{1:T}) = p(y_{1:T} | \epsilon_{1:T}, x_{1:T}) p(\epsilon_{1:T} | \phi)p(\phi)
\]  
(3.10)

Integrating over \( \epsilon_{1:T} \) we get the marginal model \( p(y_{1:T}, \phi | x_{1:T}) \).

\[
p(y_{1:T}, \phi | x_{1:T}) = p(\phi) \int_{\epsilon_{1:T}} p(y_{1:T} | \epsilon_{1:T}, x_{1:T}) p(\epsilon_{1:T} | \phi)
\]
(3.11)

Finally, the posterior distribution for \( \phi \) then fulfills

\[
p(\phi | y_{1:T}, x_{1:T}) \propto p(\phi)p(\epsilon_{1:T} | \phi).
\]
(3.12)

To ensure stationarity of \( \epsilon_{1:T} \) we need its unconditional expectation and variance to be constant. As the residuals must be unbiased, it holds that \( E(\epsilon_t) = 0 \) \( \forall \ t \). The variance recurrence \( E(\epsilon_t^2) = \phi^2 E(\epsilon_{t-1}^2) + \sigma^2 \) satisfies the stationarity requirement when the variance of the initial residual fulfills \( E(\epsilon_1^2) = \frac{\sigma^2}{1-\phi^2} \). Choosing a Gaussian distribution for the initial residual, the posterior distribution of \( \phi \) iterates to

\[
p(\phi | \epsilon_{1:T}) \propto p(\phi) \sqrt{1 - \phi^2} \exp \left\{ \frac{-1}{2\sigma^2} \epsilon_1^2 (1 - \phi^2) + \sum_{t=2}^{T} (\epsilon_t - \phi \epsilon_{t-1})^2 \right\},
\]
(3.13)

which can be shown to be equivalent to

\[
p(\phi | \epsilon_{1:T}) \propto p(\phi) \sqrt{1 - \phi^2} \mathcal{N} \left( \phi \left| \frac{\hat{\epsilon}_{12}}{\hat{\epsilon}_1}, \frac{\sigma^2}{\hat{\epsilon}_1^2} \right. \right),
\]
(3.14)

where

\[
\hat{\epsilon}_1 \equiv \sum_{t=1}^{T-1} \epsilon_t^2, \quad \hat{\epsilon}_{12} \equiv \sum_{t=2}^{T} \epsilon_t \epsilon_{t-1}.
\]

Therefore, given that the posterior of \( \phi \) is proportional to the product of the uniform prior and a Gaussian likelihood (and the prefactor \( \sqrt{1 - \phi^2} \)), the posterior follows a truncated Gaussian distribution.

Finally, combining the posterior distribution of \( \phi \) and the observation likelihood from 3.8, the
likelihood of the observations under a cointegration hypothesis is

\[
p(\epsilon_{1:T}) = (2\pi \sigma^2)^{-\frac{T-1}{2}} \exp \left( -\frac{1}{2\sigma^2} \left( \sum_{t=1}^{T} \epsilon_t^2 - \frac{(\hat{\epsilon}_{12})^2}{\hat{\epsilon}_1} \right) \right) \times \frac{1}{\sqrt{\hat{\epsilon}_1}} \int_{-1}^{1} \frac{1}{2} \sqrt{1 - \phi^2} N\left( \frac{\hat{\epsilon}_{12}}{\hat{\epsilon}_1}, \frac{\sigma^2}{\hat{\epsilon}_1} \right)
\]

where the integral in the final term is with respect to \( \phi \).

### 3.1.3 Estimation of \( \alpha, \beta \) and \( \sigma^2 \)

To find \( \alpha, \beta \) and \( \sigma^2 \), we maximize the likelihood

\[
p(y_{1:T}|x_{1:T}, \alpha, \beta, \sigma^2) = \int_{\phi} p(\psi)p(\epsilon_{1:T}|\phi).
\]

Using the Expectation Maximization Algorithm [See Dempster (1977) or Moon (1996) for an introduction] by to \( \alpha \) and \( \beta \), the maximizing equation is given by

\[
E \left( \log p(\epsilon_1|\phi) \right)_{p(\phi|\epsilon_{1:T}, \alpha_{-1}, \beta_{-1}, \sigma^2_{-1})} + \sum_{t=2}^{T} E \left( \log p(\epsilon_t|\epsilon_{t-1}, \phi) \right)_{p(\phi|\epsilon_{1:T}, \alpha_{-1}, \beta_{-1}, \sigma^2_{-1})},
\]

where \( \alpha_{-1}, \beta_{-1} \) and \( \sigma^2_{-1} \) refer to the coefficient estimates from the previous iteration. The above equation can be approximated by

\[
-\frac{1}{2\sigma^2} \left( E(1 - \phi_{-1}^2)\epsilon_t^2 + \sum_{t=2}^{T} E \left( (\epsilon_t - \phi_{-1}\epsilon_{t-1})^2 \right) \right) - \frac{T}{2} \log 2\pi \sigma^2.
\]

Differentiating by \( \alpha \) and \( \beta \), and solving the resulting system of linear equations involving the first two non-central moments of \( \phi_{-1} \) (see Section A.1 in the Appendix for details), we can update our estimates of \( \alpha \) and \( \beta \).

Finally, differentiating by \( \sigma^2 \) and setting the derivative to 0 we get

\[
\hat{\sigma}^2 = \frac{1}{T} \left[ E(1 - \phi_{-1}^2)\epsilon_t^2 + \sum_{t=2}^{T} E \left( (\epsilon_t - \phi_{-1}\epsilon_{t-1})^2 \right) \right]
\]

for the residual variance.
3.1.4 Likelihood under competing hypotheses

The Bayes factor compares the likelihood of data under a cointegration hypothesis to that under the RW hypothesis. For the RW hypothesis, we require that \( \phi = 1 \). The observation likelihood under the RW hypothesis is then

\[
p(\epsilon_{1:T}) = p(\epsilon_1) \prod_{t=2}^{T} \mathcal{N}(\epsilon_t | \epsilon_{t-1}, \sigma^2),
\]

(3.20)

which is equal to

\[
p(\epsilon_{1:T}) = p(\epsilon_1) \times \exp \left( \frac{1}{2\sigma^2} \sum_{t=2}^{T} (\epsilon_t - \epsilon_{t-1})^2 \right) \times \left( \log 2\pi\sigma^2 \right)^{-\frac{T}{2}}.
\]

(3.21)

The choice of prior for \( \epsilon_1 \) is a non-trivial problem, but a wide interval uniform distribution can be used for simplicity. Consistent with Bracegirdle and Barber (2012), the estimates for \( \alpha \) and \( \beta \) are chosen to be equal under the two hypothesis. The Bayes factor of cointegration is given by

\[
\frac{p(y_{1:T} | x_{1:T}, \phi \in [1, 1])}{p(y_{1:T} | x_{1:T}, \phi = 1)} = \frac{p(\epsilon_{1:T} | \phi \in [1, 1])}{p(\epsilon_{1:T} | \phi = 1)},
\]

(3.22)

where the numerator is given by 3.15 and the denominator by 3.21.

3.1.5 Algorithm for Bayes factor

The algorithm for computing the Bayes factor is given by Algorithm 1.

**Algorithm 1: Routine for computing Bayes factor of cointegration**

\[
\alpha_0, \beta_0, \sigma_0^2 \leftarrow \text{OLS}(y_{1:T}, x_{1:T});
\]

\[
\text{while } \exists x \in \{\alpha_t, \beta_t, \sigma_t^2\} : \Delta x \neq 0 \text{ do}
\]

\[
\epsilon_{1:T} \leftarrow y_{1:T} - \alpha - \beta x_{1:T};
\]

\[
\{\text{Observation Likelihood, } E(\phi), E(\phi^2)\} \leftarrow \text{INFEERENCE}(\epsilon_{1:T}, \sigma^2);
\]

\[
\{\alpha, \beta, \sigma^2\} \leftarrow \text{EXPECTATIONMAXIMIZATION}(x_{1:T}, y_{1:T}, \phi);
\]

end while

\[
\text{return } \text{Observation Likelihood} \times (p(\epsilon_1) \prod_{t=2}^{T} \mathcal{N}(\epsilon_t | \epsilon_{t-1}, \sigma^2))^{-1};
\]

Where the subroutine INFEERENCE refers to equations 3.14 and 3.15.
3.2 Conditional prior odds of cointegration

In this section I describe how the similarity metrics between companies are calculated, and explain the parameterization of the logistic model for computing conditional prior odds.

3.2.1 Attributes for similarity calculations

Based on the findings of Gatev et al. (2006) and Figuerola-Ferretti et al. (2018), it would be reasonable to assume that stocks in the same industry would generally have a higher than average probability of cointegration. This seems intuitive, as companies operating in the same industries should share exposure to shocks related to industry-wide demand, supply, regulation as well as changes in the competitive environment. However, one should also consider factor exposures that have high variation within an industry.

According to arbitrage pricing theory, asset prices are determined by their loadings on common factors (Ross, 1976). Arguably one of the most widely accepted asset pricing factor models is the Fama-French 3-factor model (Fama and French, 1993), that considers not only the market beta of a stock, but also a stock’s exposure to the size premium or the SMB-factor, as well as the value premium or the HML-factor.\(^3\) If asset prices are determined by their exposures to these factors, then the similarity between market capitalizations and the book-to-market -ratios of two companies should also affect the probability of cointegration for a pair of stocks. The findings of Figuerola-Ferretti et al. (2018) support this hypothesis regarding the book-to-market -ratio.

It is worth pointing out, that instead of measuring the similarity of the underlying attributes, one could also measure the similarity between the actual factor loadings. A stock’s loading on a factor (such as the SMB) is not always fully in line with the value of the related attribute (market capitalization). By estimating factor loadings, one could also include information on the market betas of companies into the analysis. However, utilizing factor loadings as a proxy for stock similarity poses the need for a number of assumptions, such as choosing the length of the evaluation period. In fact, based on preliminary tests of measuring similarity between two stocks by estimating their factor loadings, proves to be a rather unstable approach: using sample windows of varying lengths (e.g. trailing three months vs. six months vs. 12 months) might yield drastically different implied loadings for individual stocks, with even potentially

---

\(^3\) SMB is an acronym for Small Minus Big, and the SMB factor is the return difference between small and large firms, measured by market capitalization. HML is an acronym for High Minus Low, and the HML factor is the return difference between high book-to-market and low book-to-market companies. See Fama and French (1993) for details on the exact specifications of the factors.
contradicting loading signs. In addition, information on the estimated factor loadings would likely be contained already in the price path used for computing the Bayes factor. Therefore, I choose to approach the similarity calculations by considering the actual underlying attributes, rather than factor loadings.

### 3.2.2 Metrics for measuring attribute similarity

The computation of attribute similarity can be made rather straightforward for the book-to-market -ratio and market capitalization, as they are continuous numeric variables. My approach is to consider the decile rank distance, i.e. the absolute difference between the ranks of attribute-specific deciles for the two companies.

For each month in the sample, I compute 9 decile breakpoints for both attributes by sorting the cross section of stocks on the attribute and finding the $n^{th}$ largest values where

$$n \in \{ \left\lfloor \frac{N}{10} \right\rfloor \times 1, \left\lfloor \frac{N}{10} \right\rfloor \times 2, \ldots, \left\lfloor \frac{N}{10} \right\rfloor \times 9 \}, \quad (3.23)$$

and $N$ is the number of stocks in the full sample for the given month. I then assign decile ranks to the stocks based on the value of the attribute and the rank of the closest decile breakpoint above the attribute value.\(^4\) Using decile ranks, as opposed to absolute or relative differences of the attribute values, has the benefit of being relatively invariant to the different distribution shapes of the attributes. Using decile ranks is also relatively insensitive to any potential extreme outliers.

I parameterize the decile-based similarity function $S_a$ between two stocks $x_1$ and $x_2$ based on a numeric attribute $a$ as

$$S_a(x_1, x_2) = 10 - |D(a_{x_1}) - D(a_{x_2})|, \quad (3.24)$$

where the expression $D(a_{x_i})$ denotes the ranking of the decile corresponding to attribute $a$ for stock $x_i$. The function simply considers the absolute difference between the decile ranks, and subtracts it from the maximum value of the similarity metric (chosen arbitrarily to be equal to ten). The minimum (maximum) attribute-specific similarity for a pair of stocks is thus one (ten).

However, when considering industry similarity, the decile distance cannot be used as industry

\(^4\)For both the breakpoints and the attribute values, the latest information that would have been available in the previous month is used.
is not a numeric variable. I choose to compute industry similarity based on the North American Industry Classification System (NAICS). The NAICS-code for each stock consists of six characters\(^5\), specifying the sector, subsector, industry group, industry and national industry of the company. The meaning of the characters is conditional on the values of the previous characters. Therefore, similar industries have common sequences of characters in corresponding positions when moving from left to right.

Since there is no widely accepted method for computing industry similarity, I propose my own method. More specifically, the similarity function \(S_i\) of two industries \(i_1\) and \(i_2\) is calculated based on their NAICS codes by

\[
S_i(i_1, i_2) = \sum_{j=2}^{6} [z(N_{1_{j,j}}, N_{2_{j,j}}) \times w_{j-1}]
\]

where \(z(x, y) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{otherwise,} \end{cases}\) (3.25)

and the expression \(N_{1_{j,j}}\) refers to the first \(j\) characters of the NAICS code for company 1. The weights \(w_j\) could be chosen arbitrarily, but for simplicity, and due to the cumulative dependence structure of the coding, I choose equal weights for all characters.

It is worth pointing out that the proposed similarity metric assumes that the effect of each additional common character in the NAICS-codes is linear throughout the code, when the weights \(w_i\) are set as constant. In other words, the increase in similarity from a pair of stocks in the same sector to a pair of stocks in the same subsector, is equal to the increase in similarity from a pair of stocks in the same subsector to a pair of stocks in the same industry group, and twice as large as the increase from a common sector to a common industry group. The validity of this assumption could be evaluated by testing for the effects of each possible value of the similarity metric separately, using binary encoding.

3.2.3 Logistic model for conditional prior odds

To derive a functional form for computing the conditional prior odds, I must first fit a model for the cointegration probability on a sample of historical data. This model and its estimated coefficients can then be used to compute the conditional prior odds of cointegration for pairs of stocks, when testing the empirical performance of the Bayesian procedure. In this thesis, I

\(^5\)The first two characters indicate the sector.
choose to model cointegration probability with a logistic model. As the aim is to explain the outcome of a binary classifier based on a cointegration test, on the similarity metrics of the pairs of stocks, a plethora of other possible models could be utilized for such a task. These include support vector machines, random forests or multilayer perceptrons. For a general introduction on the classification methods, see for example Cortes and Vapnik (1995), Breiman (2001) and Murtagh (1991).

The choice of a logistic model is due to the interpretability of the model coefficients. As the aim of this work is to propose a cointegration testing procedure that accounts for economic intuition, it is not only important to be able to predict the probability of cointegration between pairs of stocks, but also evaluate the specific effects of the similarity measures on the probability of cointegration, so that their economic intuitiveness can be analyzed. It is important to understand that other classification models could provide better predictive power over the probability of cointegration, but this predictive performance is not to be sacrificed for the interpretability of the model in this context.

The binary response variable in the logistic model is the result of a cointegration test. A natural and obvious choice for the response variable is a binary embedding of the Bayes factor. The Bayes factors of stock pairs are embedded into binary space by classifying the pair as cointegrated when the Bayes factor exceeds a predetermined threshold. The choice of the threshold is a nontrivial problem. On one hand, as illustrated by equation 2.13, the approximation accuracy of the effects of a similarity metric improve when the ratio between the true positive rate and the false positive rate of the test is high. This would advocate for a high threshold, as the false positive rate is likely to decrease faster than the true positive rate in relative terms. On the other hand, when the threshold is set high, the number of pairs that are classified as cointegrated decreases, and therefore the effects of the similarity metrics become more difficult to infer from the data. A reasonable starting point is to use a threshold equal to unity, which implies classifying all pairs that have a higher likelihood under a cointegration hypothesis than under a random walk hypothesis as cointegrated, and repeat the analysis by increasing the classification threshold to see whether the effects remain similar.

In the logistic model the, probability of being classified as cointegrated is

$$p(c_{ij} = 1) = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{n=1}^{N} \beta_n S_n(x_i, x_j)\right)}}$$  (3.26)

where $c_{ij}$ is a binary embedding of the Bayes factor, that equals 1 when $i$ and $j$ are classified as
cointegrated, 0 otherwise, \( S_a \) is the similarity function for attribute \( a \), and \( A \) is the number of attributes in the model. After fitting the model, the conditional prior odds of cointegration are then given by

\[
\frac{p(c_{ij} = 1)}{1 - p(c_{ij} = 1)} \times \frac{1 - \overline{C}}{\overline{C}} = \exp \left( \beta_0 + \sum_{a=1}^{A} \beta_a S_a(x_i, x_j) \right) \times \frac{1 - \overline{C}}{\overline{C}},
\]  

(3.27)

where

\[
\overline{C} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} p(c_{ij} = 1)}{N(N - 1) \times 0.5}
\]  

(3.28)

is the average probability of being classified as cointegrated. The parameter \( \overline{C} \) is conditional on three unobserved parameters, which are the average probability of cointegration, as well as the sensitivity and specificity of the test. If the (unknown) ratio between the true positive and false positive rate of the cointegration test is high, then the conditional prior odds is asymptotically a scaled version of the true prior odds, so that the average odds are equal to unity.

### 3.3 Conditional posterior odds of cointegration

Building on the routine for computing the Bayes factor of cointegration in Algorithm 1, we can now add our conditional prior odds to the routine. As the conditionality of the prior odds is carried over into the posterior odds, the product of the Bayes factor and the conditional prior odds is therefore called the conditional posterior odds. Algorithm 2 gives the routine for computing the conditional posterior odds of cointegration.

**Algorithm 2: Routine for computing conditional posterior odds of cointegration**

\[\begin{align*}
& \alpha_0, \beta_0, \sigma_0^2 \gets \text{OLS}(y_{1:T}, x_{1:T}); \\
\textbf{while} & \exists x \in \{\alpha_i, \beta_i, \sigma_i^2\} : \Delta x \neq 0 \text{ do} \\
& \quad \epsilon_{1:T} \leftarrow y_{1:T} - \alpha - \beta x_{1:T}; \\
& \quad \{\text{Observation Likelihood, } E(\phi), E(\phi^2)\} \gets \text{INFERENCE}(\epsilon_{1:T}, \sigma^2); \\
& \quad \{\alpha, \beta, \sigma^2\} \gets \text{EXPECTATIONMAXIMIZATION}(x_{1:T}, y_{1:T}, \phi); \\
\textbf{end while}
\end{align*}\]

\[\begin{align*}
& \text{Bayes Factor} \leftarrow \text{Observation Likelihood} \times (\prod_{t=2}^{T} N(\epsilon_t | \epsilon_{t-1}, \sigma^2))^{-1}; \\
& \text{Conditional Prior Odds} \leftarrow \exp \left( \beta_0 + \sum_{a=1}^{A} \beta_a S_a(x_i, x_j) \right) \times \frac{1 - \overline{C}}{\overline{C}}; \\
& \textbf{return} \text{Bayes Factor} \times \text{Conditional Prior Odds};
\end{align*}\]
3.4 Persistence of cointegration

To evaluate the empirical performance of the introduced procedure, I test the persistence of testing-implied cointegration in consecutive periods of equal length. The length of the periods is a nontrivial decision both in this thesis, as well as the selection of pairs trading opportunities in real life. On one hand, using a shorter evaluation period increases the difficulties in detecting spurious cointegration. On the other hand, the dynamics of the relations between stocks might change considerably when using longer windows. Such changes might make the estimated parameters, and therefore inference of cointegration obsolete. Ardia et al. (2016) use a six-month evaluation period, while Clegg (2014) and Gatev et al. (2006) opt for 12 months.

In this thesis, I utilize the following methodology. In an initial period, called the classification period, pairs of stocks are classified as cointegrated (random walks), if their classification variable exceeds (does not exceed) a predetermined classification threshold. The classification variable is either the Bayes factor or the conditional posterior odds. Then, during the subsequent period, called the testing period, I test whether the Bayes factor exceeds a predetermined testing threshold during the testing period. The testing threshold exceedance proportions are then compared between the group of pairs that were classified as cointegrated, and the group that were classified as random walks. As the absolute level of the conditional prior odds is unobservable, the optimal levels for both the classification and the testing thresholds are also unknown. Therefore, I perform the classification and testing with a variety of threshold levels.

The workflow of testing for the persistence of cointegration is thus the following. I compute the Bayes factor for each pair of stocks, using classification periods of six months, and perform the classification. I then re-compute the Bayes factor of each pair in the subsequent six-month testing period, and observe whether the testing threshold is exceeded. If the proportion of pairs that exceed the testing threshold is significantly higher for the group of pairs classified as cointegrated, then we have reason to believe that testing-implied cointegration is a persistent quality.

I then repeat this process by using the conditional posterior odds for the classification in the classification period, as opposed to the Bayes factor. The subsequent steps are kept identical, meaning that the exceedance of the testing threshold is still based on the Bayes factor, allowing for a fair comparison between the classifications. The Bayes factor is also a better representation

---

6The classification periods are non-overlapping, with the first one starting on the first trading day of January 2014, and the last one ending on the last trading day of June 2018. This implies a total of nine classification periods in the testing sample. See Section 3.5.2 for further details.

7The similarity metrics for the conditional prior odds are calculated based on the latest information that would have been available at the start of the classification period.
of the in-sample behavior of the pair, which would be of interest to a pairs trader, given that the pairs are already selected at this stage. If the proportion of pairs that exceed the testing threshold is significantly higher when using conditional posterior odds for the classification, we have reason to believe that using conditional prior odds improves the ability of the procedure in identifying persistent pairwise cointegrating relations.

The described testing setup can be thought of as evaluating the pairwise cointegration of stocks during a 12-month period. After an initial six months have passed, the procedure attempts to predict whether pairs that have thus far exhibited cointegration will persist to exhibit cointegration during the remaining six months of a 12-month period. Therefore, I argue that the correct model for computing the conditional prior odds is found by using evaluation periods of 12 months in the training sample. However, I also fit the logistic model using evaluation periods of six and 30 months to analyze whether the coefficient estimates remain similar.

3.5 Description of data

In this thesis, I train and test the introduced Bayesian cointegration procedure in an empirical setting, using historical data of US-listed common stock. This section describes the sources of data, issues arising from matching data between two databases, potential erroneous outliers, and the partitioning of the data between training and testing samples.

3.5.1 Retrieval and processing of data

Sources of data

I obtain daily data on total returns, closing prices, share turnover, shares outstanding and NAICS classification codes for each common stock listed on the New York Stock Exchange (NYSE), American Stock Exchange (AMEX) and NASDAQ found in the CRSP-database between December 2008 and December 2018. I obtain data on the book value per share from Thomson Reuters DataStream.

Matching between databases

Since CRSP and DataStream utilize slightly different keys as security identifiers, matching data between the two databases requires an intermediate step. Firstly, I obtain the entire available universe of stocks from the CRSP-database. To retrieve the corresponding data from the Thomson-
database, I transform the 8-character CUSIP identifier used in CRSP into the 9-character CUSIP identifier used in DataStream. The missing check digit can be deterministically calculated from the 8-character CUSIP using the check digit -algorithm. While the transformed CUSIP codes are successfully identified by DataStream in virtually all instances, the availability of the required variables is not complete, which results in slight reductions to the number of evaluated stock pairs (see Table 4.1 for details).

Data transformations

Since the cointegration testing procedure is performed on the price paths of the securities, it is important to also account for any possible distributions, such as dividends, that are not reflected in the closing prices of securities, but do result in cash flows to the holder. Therefore, I construct a "True Price" -time-series for each stock in the sample, based on the initial closing price of the security, and the reported daily total returns of the stock (including any potential distributions).

I compute the market capitalization time series for each company by multiplying the number of shares outstanding by the closing price. Additionally, I calculate the book-to-market -ratio for each stock by dividing the book value of equity per share obtained from DataStream, by the market price obtained from CRSP.

Potential erroneous outliers and data winsorization

The data obtained from Thomson Reuters contain some peculiar values that seem out of the ordinary. For example, clusters of book-to-market -ratios are found at around 150 million. Such values would be difficult to justify by logic, especially when closer inspection reveals that in all cases the values are preceded by values several millions smaller for the preceding days. Therefore, I suspect such values to be erroneous, and decide to discard them. Specifically, I decide to remove observations where the book-to-market -ratio is above 500 or below -100, which results in the winsorization of 2.7% of the data. Additionally, I decide to exclude so called "penny stocks", i.e. stocks with a market price of less than one dollar, due to their innate tendency to lack liquidity (Liu et al., 2015).

3.5.2 Partition of data into samples

As I first fit a logistic model for the computation of the conditional prior odds, and subsequently test for the empirical performance of a procedure utilizing this model, it is important to separate

---

8See [https://www.cusip.com/pdf/CUSIP_Intro_03.14.11.pdf](https://www.cusip.com/pdf/CUSIP_Intro_03.14.11.pdf) for details.
the data into distinct training and testing samples. I select a sample length of 5 years for both the training and testing sample, so that the training (testing) sample runs for 60 months between January 2009 (2014) and December 2013 (2018).  

*Size and selection criteria of training and testing samples*

Since the amount of possible 2-element combinations from a set increases (almost) quadratically as a function of set size, the set of evaluated stock pairs should be limited for the sake of computational complexity. I choose to limit the maximum number of evaluated stocks in the training and testing samples to 200 at any time, resulting in the maximum number of evaluated pairs being  \( \frac{200 \times (200 - 1)}{2} = 19900 \) for every evaluation period.

The sample used for fitting the logistic model should be kept separate from the testing sample to ensure that we are able to test the predictive power of the model out of sample. Should the hypothesis of persistent cointegration hold, we should also assume that pairs of stocks would carry over their cointegrating relations from the training sample to the testing sample. Therefore, I decide to partition the training and testing samples not only with respect to time, but also to use a different selection criterion for the stocks used in the two samples. This reduces the amount of overlapping stock pairs near the transition between the training and testing samples.

Since implementing a pairs trading strategy requires the investor to be able to open and close positions frequently (daily, if needed), the securities used for building the long-short portfolios should be liquid. Therefore, I choose liquidity, measured by share turnover ratio\(^{10}\) in the previous trading month as a measure for narrowing down the population of selectable securities for each sample. More specifically, the training (testing) samples consist of stocks where the share turnover ratio in the month prior to the start of the period was above that of the 401\(^{st}\) (201\(^{st}\)) highest security in the CRSP database, but below or equal to that of the 201\(^{th}\) (1\(^{st}\)) security. While the monthly share turnover ratio measure is not a perfect measure of liquidity, as it does not account for any possible liquidity clustering within the trading month, it is nonetheless an adequate measure to be used as a selection variable in building the samples. It is also worth pointing out, that this method for the selection of stocks into the samples restricts the evaluated stock pairs in a similar manner to Figuerola-Ferretti et al. (2018), with respect to the trading volume.

---

\(^9\)Note that data for December 2008 is needed for the company attributes that are based on the value at the end of the previous month.

\(^{10}\)Number of shares traded divided by the average number of shares outstanding during the month.
4. Empirical findings

In this chapter, I present the findings from fitting the logistic model for the conditional prior odds, and the empirical performance of the proposed Bayesian cointegration testing procedure in identifying persistent cointegrating relations.

4.1 Logistic model for conditional prior odds

I start the empirical analysis by fitting a logistic model for the testing-implied cointegration probability in the training sample. I first calculate the Bayes factors for each pair of stocks, followed by computing the similarity between the companies, for each evaluation period.

4.1.1 Computation of Bayes factors in training sample

Utilizing non-overlapping evaluation periods of six, 12 and 30 months in length, I estimate the Bayes factors of cointegration for pairs of stocks in the training sample, and calculate the proportion of pairs where the Bayes factor exceeds a predetermined threshold. Table 4.1 presents an overview of this process, using thresholds equal to unity, five, ten and 20.

The contents of Table 4.1 are mostly consistent with intuition. The proportion of pairs where the Bayes factor exceeds the threshold decreases as the evaluation period increases in length. This is likely explained by two separate effects. Firstly, using longer evaluation periods should decrease the number of pairs that the test spuriously identifies as cointegrated. Secondly, as the evaluation period increases in length, the probability of either stock experiencing non-transient idiosyncratic shocks also increases, thus reducing the actual number of cointegrated pairs.

The proportion of pairs that are identified as cointegrated varies slightly between the periods. What is also worthy of noting, is that when using 12-month evaluation periods, the proportion of pairs identified as cointegrated decreases almost monotonically through the years for any choice
Table 4.1. Overview of Bayes factors in the training sample

This table presents an overview of estimating the Bayes factors (BF) of cointegration for pairs of stock prices, using daily data and a set of evaluation period lengths. Panels A, B and C represent the estimation of Bayes factors using non-overlapping evaluation periods of six, 12 and 30 months respectively. In each panel, I report the number of evaluated stocks and stock pairs, along with the percentage of pairs where the Bayes factor exceeds a predetermined threshold. Thresholds equal to unity, five, ten and 20 are reported. The sample consists of US common stock listed on the NYSE, AMEX and NASDAQ, fulfilling the liquidity requirements specified in Section 3.5.2, spanning a 60-month period between January 2009 and December 2013. Penny stocks are excluded. H1 and H2 denote the first and second half of a calendar year respectively. Y denotes the full calendar year.

**Panel A: Training with Non-overlapping six-month periods**

<table>
<thead>
<tr>
<th>Period</th>
<th>Stocks</th>
<th>Pairs</th>
<th>BF &gt; 1</th>
<th>BF &gt; 5</th>
<th>BF &gt; 10</th>
<th>BF &gt; 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1 2009</td>
<td>164</td>
<td>13366</td>
<td>19.8%</td>
<td>7.8%</td>
<td>4.7%</td>
<td>2.8%</td>
</tr>
<tr>
<td>H2 2009</td>
<td>196</td>
<td>19110</td>
<td>10.2%</td>
<td>3.2%</td>
<td>2.0%</td>
<td>1.2%</td>
</tr>
<tr>
<td>H1 2010</td>
<td>176</td>
<td>15400</td>
<td>17.1%</td>
<td>6.1%</td>
<td>3.6%</td>
<td>2.3%</td>
</tr>
<tr>
<td>H2 2010</td>
<td>195</td>
<td>18915</td>
<td>11.9%</td>
<td>3.4%</td>
<td>2.0%</td>
<td>1.2%</td>
</tr>
<tr>
<td>H1 2011</td>
<td>179</td>
<td>15931</td>
<td>16.8%</td>
<td>4.3%</td>
<td>2.3%</td>
<td>1.6%</td>
</tr>
<tr>
<td>H2 2011</td>
<td>188</td>
<td>17578</td>
<td>18.7%</td>
<td>6.3%</td>
<td>3.9%</td>
<td>2.5%</td>
</tr>
<tr>
<td>H1 2012</td>
<td>189</td>
<td>17766</td>
<td>14.8%</td>
<td>5.2%</td>
<td>3.3%</td>
<td>2.3%</td>
</tr>
<tr>
<td>H2 2012</td>
<td>190</td>
<td>17955</td>
<td>16.3%</td>
<td>4.8%</td>
<td>3.1%</td>
<td>2.3%</td>
</tr>
<tr>
<td>H1 2013</td>
<td>180</td>
<td>16110</td>
<td>16.7%</td>
<td>5.7%</td>
<td>3.5%</td>
<td>2.0%</td>
</tr>
<tr>
<td>H2 2013</td>
<td>195</td>
<td>18915</td>
<td>10.4%</td>
<td>3.2%</td>
<td>1.8%</td>
<td>1.0%</td>
</tr>
</tbody>
</table>

**Panel B: Training with Non-overlapping 12-month periods**

<table>
<thead>
<tr>
<th>Period</th>
<th>Stocks</th>
<th>Pairs</th>
<th>BF &gt; 1</th>
<th>BF &gt; 5</th>
<th>BF &gt; 10</th>
<th>BF &gt; 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y 2009</td>
<td>163</td>
<td>13203</td>
<td>12.4%</td>
<td>5.1%</td>
<td>3.6%</td>
<td>2.5%</td>
</tr>
<tr>
<td>Y 2010</td>
<td>170</td>
<td>14365</td>
<td>11.4%</td>
<td>4.0%</td>
<td>2.5%</td>
<td>1.5%</td>
</tr>
<tr>
<td>Y 2011</td>
<td>165</td>
<td>13530</td>
<td>10.2%</td>
<td>3.0%</td>
<td>1.7%</td>
<td>1.0%</td>
</tr>
<tr>
<td>Y 2012</td>
<td>187</td>
<td>17391</td>
<td>9.1%</td>
<td>3.1%</td>
<td>1.8%</td>
<td>1.0%</td>
</tr>
<tr>
<td>Y 2013</td>
<td>176</td>
<td>15400</td>
<td>7.0%</td>
<td>2.7%</td>
<td>1.7%</td>
<td>0.8%</td>
</tr>
</tbody>
</table>

**Panel C: Training with Non-overlapping 30-month periods**

<table>
<thead>
<tr>
<th>Period</th>
<th>Stocks</th>
<th>Pairs</th>
<th>BF &gt; 1</th>
<th>BF &gt; 5</th>
<th>BF &gt; 10</th>
<th>BF &gt; 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1 2009 - H1 2011</td>
<td>151</td>
<td>11325</td>
<td>6.3%</td>
<td>2.2%</td>
<td>1.5%</td>
<td>1.0%</td>
</tr>
<tr>
<td>H2 2011 - H2 2013</td>
<td>150</td>
<td>11175</td>
<td>7.5%</td>
<td>3.3%</td>
<td>2.1%</td>
<td>1.3%</td>
</tr>
</tbody>
</table>
of threshold. While this could be interpreted as a sign of fewer pairs trading opportunities when modeling co-movement with cointegration, it is difficult to make any conclusions based on such a short period of time, especially when the other evaluation period lengths do not clearly support this observation. However, the observation is in line with Rad et al. (2016), who argue that the number of trading opportunities has decreased from the year 2009 onwards when modeling co-movement through cointegration.

4.1.2 Estimation of relation between company similarity and Bayes factor

The next step in formulating the logistic model is to run regressions for the binary embedding of the Bayes factors on firm-level similarity metrics. I run the regressions with evaluation periods of six, 12 and 30 months, using pooled samples containing all observations from the non-overlapping evaluation periods. As regressors, I use the similarity metrics with respect to industry, market capitalization and book-to-market -ratios at the start of the period, as described in section 3.2.2. I run the regressions using the similarity metrics both collectively, as well as separately. Comparing the coefficient estimates from the separate regressions to those of the collective regressions could help identify any potential multicollinearity between the regressors. Table 4.2 present the coefficient estimates of the similarity metrics, along with their standard errors in the collective regressions. The coefficients of the individual regressions are similar, and are reported in table 1.1 of Section A.2 in the Appendix.

When interpreting the coefficients in Table 4.2, it is important to remember the following notions. Firstly, regressions utilizing different threshold levels are not independent of each other. For example, a pair of stocks with a Bayes factor over 20 will be classified as cointegrated in each regression, and thus the similarity metrics of this pair will influence the coefficients of each regression. This implies that a collective meta-analysis of the coefficients by using a Fischer’s combined probability test or other similar approaches, is not feasible.

Secondly, observing coefficients with differing signs and magnitudes between different evaluation period lengths could arise from slow mean reversion speeds of the cointegrating relations. Pairs with a slow mean reversion speed might not appear cointegrated when evaluating shorter periods of time, and effects of some similarity metrics on the reversion speed could differ.

Thirdly, as the length of the evaluation period increases, the less informative the company attributes observed at the start of the evaluation period become. For instance, two companies that belong to the same size decile at the start of the evaluation period might end up in vastly different
Table 4.2. Logistic regressions of testing-implied cointegration on company similarity

This table presents the coefficient estimates and standard errors from three logistic regressions, where the dependent variable is the binary result of a cointegration test between a pair of stock prices. The cointegration test considers whether the Bayes factor (BF) of cointegration is above a predetermined threshold. Regressions using threshold levels of unity, five, ten, and 20 are reported. The explanatory variables of the regressions consist of the similarities between the pair of companies whose stock prices are evaluated, with respect to their industries, market capitalizations, and book-to-market ratios. Panels A, B and C report the results of the regressions with varying evaluation periods for the Bayes factor, where the evaluation periods are six, 12, and 30 months respectively. The evaluation periods are non-overlapping, constituting a 60-month period between January 2009 and December 2013. The sample consist of moderately liquid US common stock listed on the NYSE, AMEX and NASDAQ. Penny stocks are excluded. Statistical significance is denoted by *, ** and *** at the 5%, 1% and 0.1% levels respectively.

### Panel A: Six-month evaluation periods

<table>
<thead>
<tr>
<th>Attribute</th>
<th>BF &gt; 1</th>
<th>BF &gt; 5</th>
<th>BF &gt; 10</th>
<th>BF &gt; 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.709***</td>
<td>-2.654***</td>
<td>-2.979***</td>
<td>-3.319***</td>
</tr>
<tr>
<td></td>
<td>(0.033)</td>
<td>(0.054)</td>
<td>(0.068)</td>
<td>(0.084)</td>
</tr>
<tr>
<td>Industry</td>
<td>0.030***</td>
<td>0.069***</td>
<td>0.067***</td>
<td>0.078***</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.014)</td>
<td>(0.018)</td>
<td>(0.022)</td>
</tr>
<tr>
<td>Market capitalization</td>
<td>-0.024***</td>
<td>-0.059***</td>
<td>-0.080***</td>
<td>-0.088***</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.006)</td>
<td>(0.008)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>Book-to-market</td>
<td>0.015***</td>
<td>0.005</td>
<td>-0.002</td>
<td>-0.013</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.005)</td>
<td>(0.006)</td>
<td>(0.007)</td>
</tr>
</tbody>
</table>

### Panel B: 12-month evaluation periods

<table>
<thead>
<tr>
<th>Attribute</th>
<th>BF &gt; 1</th>
<th>BF &gt; 5</th>
<th>BF &gt; 10</th>
<th>BF &gt; 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.513***</td>
<td>-3.416***</td>
<td>-3.600***</td>
<td>-3.683***</td>
</tr>
<tr>
<td></td>
<td>(0.065)</td>
<td>(0.102)</td>
<td>(0.120)</td>
<td>(0.145)</td>
</tr>
<tr>
<td>Industry</td>
<td>0.050**</td>
<td>0.087***</td>
<td>0.091***</td>
<td>0.084*</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.022)</td>
<td>(0.027)</td>
<td>(0.034)</td>
</tr>
<tr>
<td>Market capitalization</td>
<td>0.005</td>
<td>-0.023*</td>
<td>-0.066***</td>
<td>-0.135***</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.011)</td>
<td>(0.013)</td>
<td>(0.016)</td>
</tr>
<tr>
<td>Book-to-market</td>
<td>0.031***</td>
<td>0.039***</td>
<td>0.046***</td>
<td>0.071***</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.008)</td>
<td>(0.010)</td>
<td>(0.013)</td>
</tr>
</tbody>
</table>

### Panel C: 30-month evaluation periods

<table>
<thead>
<tr>
<th>Attribute</th>
<th>BF &gt; 1</th>
<th>BF &gt; 5</th>
<th>BF &gt; 10</th>
<th>BF &gt; 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.759***</td>
<td>-2.950***</td>
<td>-3.251***</td>
<td>-3.256***</td>
</tr>
<tr>
<td></td>
<td>(0.149)</td>
<td>(0.204)</td>
<td>(0.238)</td>
<td>(0.273)</td>
</tr>
<tr>
<td>Industry</td>
<td>0.061</td>
<td>0.080</td>
<td>0.075</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>(0.037)</td>
<td>(0.052)</td>
<td>(0.062)</td>
<td>(0.076)</td>
</tr>
<tr>
<td>Market capitalization</td>
<td>0.039*</td>
<td>-0.059</td>
<td>-0.018</td>
<td>-0.054</td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
<td>(0.006)</td>
<td>(0.027)</td>
<td>(0.031)</td>
</tr>
<tr>
<td>Book-to-market</td>
<td>-0.012</td>
<td>-0.043**</td>
<td>-0.041**</td>
<td>-0.049**</td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.013)</td>
<td>(0.016)</td>
<td>(0.018)</td>
</tr>
</tbody>
</table>
deciles by the end of a 60-month evaluation period. Further, if we assume that companies of similar size would be more likely to exhibit cointegration, drastic changes in the relative size of either company could terminate the cointegrating relation even if the companies were of similar size and cointegrated during the start of the evaluation period.

Finally, I discuss the coefficient-implied effects of each similarity metric on the probability of cointegration in further detail in the following paragraphs.

Industry
The coefficients for industry similarity support the view that companies in similar industries do indeed tend to exhibit cointegration more often than a randomly selected pair. The coefficient for industry similarity is positive in all regressions. Using evaluation periods of six (12) months, the coefficients are statistically significant at the 0.1% (5%) -level for all thresholds. However, using evaluation periods of 30 months, the coefficients become statistically insignificant.

It is nonetheless important to note that the proposed industry-similarity metric depends on a variety of assumptions. For instance, it could well be that cointegrating relations are clustered within certain industries [Gatev et al. (2006) find evidence that most of the best performing long-short portfolios are formed by pairs of utility or industrial stocks], and that similarity for stocks outside such industries does not increase the probability of cointegration. Also, the parameterization of the similarity metric assumes a linear increase in similarity in the NAICS code, as discussed in Section 3.2.2. Overall, while this parameterization of industry similarity appears to be positively associated with cointegration probability, an improved metric might capture the industry effects even better. The formulation of such a metric could be the topic of another thesis altogether, and thus the introduced metric suffices for the purposes of this thesis.

Market capitalization
Based on the coefficients, it appears that similarity with respect to market capitalization has a negative effect on cointegration probability. The coefficient for market capitalization similarity is negative in ten out of 12 regressions. Using six-month evaluation periods, the coefficient is negative and statistically significant at the 0.1%-level in for all threshold levels. Using 12-month evaluation periods, the coefficient is negative and statistically significant at the 5%-level with a threshold of five, becoming significant at the 0.01%-level for thresholds ten and 20. Similarly to industry similarity, the coefficients become less significant when using evaluation periods of 30-months.
The negative coefficients could be interpreted as inconsistent with the findings of Gatev et al. (2006), although the authors do not model co-movement through cointegration. Figuerola-Ferretti et al. (2018) fail to document improvement in the performance of long-short portfolios, when pair formation is restricted by market capitalization. The findings in this thesis could provide an explanation to those of Figuerola-Ferretti et al. (2018), as the similarity with respect to market capitalization appears to be in fact negatively associated with the probability of cointegration.

**Book-to-market**

Based on the estimated coefficients, the effects of book-to-market similarity on cointegration probability are somewhat inconclusive. Using six-month evaluation periods, the coefficient changes from positive to negative as the threshold for the Bayes factor is increased. With 12-month evaluation periods, however, the coefficients are strictly positive and significant at the 0.01%-level for all thresholds. This implies book-to-market similarity having a positive effect on cointegration probability, and such findings are in line with Figuerola-Ferretti et al. (2018). Additionally, book-to-market similarity is the only variable with a statistically significant coefficient in any of the regressions using 30-month evaluation periods. However, the coefficients are strictly negative. Despite such conflicting findings, the effects of book-to-market similarity on cointegration probability seem to be positive and statistically significant for the 12-month evaluation that is of interest.

**General observations**

Overall, it appears that the effects of different kinds of company similarity on 30-month cointegration probability seem to be weak. This could arise from the fact that many pairs likely experience non-transient idiosyncratic shocks during such a long evaluation period, and therefore true cointegration relations are rare for such a long period.

For six and 12-month cointegration probability, the effects of industry and market capitalization similarity seem consistent through the threshold levels. When it comes to the effects of book-to-market similarity, the coefficients of the 12-month regressions are not only consistent throughout the threshold levels, but also positive and therefore in line with the findings of Figuerola-Ferretti et al. (2018). As discussed in Section 3.4, regressions with 12-month evaluation periods are theoretically consistent with the testing setup of this thesis. Therefore, I proceed to parameterize the logistic model based on the coefficients of the 12-month evaluation period regressions. In the following section I analyze the appropriate level for setting the threshold.
4.1.3 Parameterization of logistic model

Given the relation in Equation 2.13, the false positive ratio of the cointegration test should be low for accurate approximation of the true effects of a similarity metric on cointegration probability. If the false positive ratio decreases at a relatively higher pace than the true positive rate when the threshold of the test is increased, increasing the threshold would make the approximation of the true effects of similarity on cointegration probability more accurate. However, as increasing the threshold also decreases the set of pairs that are classified as cointegrated, inference on the effects of the similarity metrics becomes more difficult.

As illustrated in Table 4.2, the coefficient for industry similarity appears to stay approximately constant when the threshold is increased from five to 20. However, the coefficients for market capitalization and book-to-market similarity seem to increase or decrease monotonically as the threshold is increased up to 20. This could imply, as given by relation in equation 2.13, that the coefficient estimates are asymptotically approaching their true values, and therefore the threshold should be further increased to improve the accuracy of the approximation.

To analyze this, I compute the coefficient estimates by increasing the threshold gradually to see whether the coefficients seem to converge. Figure 4.1 illustrates the convergence of the coefficient estimates for threshold values between unity and 250.

![Figure 4.1. This figure illustrates the coefficient estimates of similarity metrics as a function of the classification threshold in the logistic regressions introduced in Section 3.2.3. MktCap and BtM denote the market capitalization and book-to-market coefficients respectively. The darker (lighter) shaded areas denote the 50% (95%) confidence interval.](image)
Based on Figure 4.1, the coefficient of market capitalization similarity seems to converge to approximately -0.42, when the Bayes factor threshold is increased up to 250. The shape of the coefficient curve is rather smooth, which could be an indication of the coefficient converging towards its true value, as suggested by Equation 2.13. Based on the shape of the curve, it does therefore appear that market capitalization similarity is negatively associated with cointegration probability, and the strongly negative coefficients in Table 4.2 are not due to chance.

On the other hand, the coefficients for industry and book-to-market similarity seem to fluctuate around 0.09 and 0.08 respectively, and do not appear to show clear signs of convergence. This type of behavior is also consistent with Equation 2.13, from which we observe that when the effects of the similarity metrics are small (and therefore the ratio in Equation 2.13 is close to unity), the coefficients should be rather insensitive to the $R_{TF}$ -parameter and therefore varying the threshold is likely to result in random fluctuation around the true value, rather than convergence. In addition, the coefficient for industry similarity becomes less significant as the classification threshold is increased. This could be an indication that industry similarity could be excluded from the model, as its effects are not significant enough. However, based on the strict positiveness of this coefficient in all the regressions, I argue that industry similarity is an informative attribute, and thus decide to include it in the model.

I proceed to parameterize the logistic model for computing the conditional prior odds based on the regression with a threshold of 250. Using this threshold ensures that the coefficient of market capitalization similarity has converged sufficiently towards its hypothesized true value. Table 4.3 presents the coefficients of the model. As the scale of the similarity metrics is between one and ten for market capitalization and book-to-market, and between 1 and 5 for the industry, the coefficients in Table 4.3 suggest that market capitalization similarity has the largest effect on cointegration probability out of all the similarity metrics.

### 4.2 Persistence of cointegration

I proceed to test the empirical performance of the introduced Bayesian procedure with and without using conditional prior odds. First, I classify pairs of stocks as cointegrated or random walks during initial non-overlapping six-month classification periods based on the Bayes factor, and then repeat the classification using conditional posterior odds, for a set of classification thresholds. The subfigures in Figure 4.2 visualize the proportion of pairs where the Bayes factor
Table 4.3. Coefficients of the logistic model for computing conditional prior odds of cointegration

This table presents the coefficients of the similarity metrics in the logistic model for computing the conditional prior odds of cointegration. The metrics are introduced in detail in Section 3.2.2. The parameterization of the logistic model is given by Equation 3.27. The scaler-parameter is conditional on three unobserved parameters which represent the average cointegration probability, as well as the sensitivity and specificity of the Bayes factor-based cointegration classification with a classification threshold of 250.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>(Similarity) Metric</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaler</td>
<td>((1 - \hat{C})/\bar{C})</td>
<td>226.2</td>
</tr>
<tr>
<td>Intercept</td>
<td>-</td>
<td>-3.053</td>
</tr>
<tr>
<td>Industry</td>
<td>(\sum_{j=2}^{6} \left[ z(N_{1j}, N_{2j}) \right] )</td>
<td>0.090</td>
</tr>
<tr>
<td>Market capitalization</td>
<td>(10 -</td>
<td>D(MC_{x1}) - D(MC_{x2})</td>
</tr>
<tr>
<td>Book-to-market</td>
<td>(10 -</td>
<td>D(BP_{x1}) - D(BP_{x2})</td>
</tr>
</tbody>
</table>

exceeds a predetermined testing thresholds of unity, five, ten and 20 in the subsequent six-month testing periods, for both classification methods. The results are based on a pooled sample of all evaluation periods belonging to the testing sample.

Figure 4.2 illustrates that regardless of the classification method, the proportion of pairs where the Bayes factor exceeds the testing threshold in the latter six-month testing period is considerably larger for pairs that were classified as cointegrated, compared to the pairs that were classified as random walks. This suggests that Bayesian testing-implied cointegration is a persistent quality, and supports the hypothesis that actual, unobservable pairwise cointegration is also a persistent quality.

Additionally, using conditional posterior odds for the classification results in an equal or larger proportion of pairs where the Bayes factor exceeds the testing threshold in the testing period, for almost all classification threshold levels. The benefits of using conditional posterior odds are clear especially when the classification threshold is above the mid twenties.

To test whether the differences between the proportions are statistically significant, I perform a set of two-proportion Z-tests. First, I test the differences between the proportion of testing threshold exceedances for pairs classified as cointegrated, versus the proportion of testing threshold exceedances for pairs classified as random walks. The tests are performed using both the Bayes factor and conditional posterior odds classification. Then, I test for the differences between the proportions of testing threshold exceedances for pairs classified as cointegrated, between the two classification methods. The Z-scores are illustrated in Figure 4.3, along with critical values.
Figure 4.2. The four subfigures illustrate the proportion of Bayes factors above a predetermined testing threshold for the remaining six months (the testing period) of the evaluation periods. Subfigures (a), (b), (c) and (d) present the proportions exceeding testing thresholds of unity, five, ten and 20 respectively. The proportion of Bayes factors is reported for pairs classified as cointegrated or random walks, using classification based on the Bayes factors (BF), as well as the conditional posterior odds (CPO) in the preceding six-month period (the classification period). The classification threshold of the classification period is on the horizontal axis. The lighter (darker) shaded areas represent 95% (50%) confidence intervals.
(a) Z-scores of proportion difference between cointegrated and random walk pairs (BF classification)

(b) Z-scores of proportion difference between cointegrated and random walk pairs (CPO classification)

(c) Z-scores of proportion difference between cointegrated pairs (CPO vs. BF classification)

Figure 4.3. The figures illustrate the Z-scores of the differences between testing threshold exceedance proportions. Subfigure (a) illustrates the differences between pairs classified as cointegrated versus those classified as random walks based on the Bayes factor (BF) -classifications. Subfigure (b) illustrates the differences between pairs classified as cointegrated versus those classified as random walks based on the conditional posterior odds (CPO) -classifications. Subfigure (c) illustrates the differences between pairs classified as cointegrated by the CPO-classifications versus pairs classified as cointegrated by the BF-classifications. Testing threshold levels of unity, five, ten and 20 are illustrated. Critical values are denoted by the gray horizontal lines.
When comparing the testing threshold exceedance proportion differences between pairs classified as cointegrated and pairs classified as random walks, the differences are significant well beyond all customary levels, regardless of the testing threshold or classification threshold used. When comparing the testing threshold exceedance proportions between pairs classified as cointegrated by the two classification approaches, the differences are not statistically significant for classification thresholds below 23. However, when using classification threshold levels above 34, the differences are significant at the 1% -level or lower, regardless of the testing threshold.

Finally, I test for differences between the testing threshold exceedance proportions of the in-sample optimal Bayes factor classification, against each conditional posterior odds classification. This analysis consists of first finding the classification threshold resulting the highest testing threshold exceedance proportion, when the classification is based on the Bayes factor, and then testing for the difference between this proportion to that of any conditional posterior odds classification. If I find a conditional posterior odds classification where the proportion is significantly higher, this implies that even the ex-post optimal classifications of the two methods are different. The test is performed for testing threshold levels of unity, five, ten and 20. The Z-scores of these tests are illustrated in figure 4.4

![Z-score graph](image)

**Figure 4.4.** This figure illustrates the Z-scores of the proportion difference tests between the ex-post optimal Bayes factor (BF) classification, and conditional posterior odds (CPO) classifications with varying classification threshold levels. The proportion of interest is the testing threshold exceedance. Testing threshold levels equal to unity, five, ten and 20 are reported. Critical values are denoted by the gray horizontal lines.

As illustrated in Figure 4.4, none of the conditional posterior odds classifications result in testing threshold exceedance proportions that are significantly greater than that of the ex-post optimal Bayes factor classification. However, this type of ex-post optimization would not be possible for a
practitioner, who would have to specify the classification threshold ex-ante. Therefore, the failure to document significant differences does not directly imply that the empirical performance of the conditional posterior odds-classification is not superior to that of the Bayes factor classifications. When fixing the classification threshold ex-ante, the differences between the two approaches are significantly higher for a variety of classification thresholds, as illustrated by Figure 4.3.
5. Discussion and conclusions

In this final chapter, I discuss the empirical findings and their implications for future researchers as well as practitioners in further detail. This is followed by concluding remarks on the findings and contributions of this thesis.

5.1 Discussion on empirical findings

The following sections discuss the implications of the empirical findings presented in the previous chapter. The findings can be divided into two categories. Firstly, I discuss the model for the conditional prior odds. Then, I discuss the performance of the introduced procedure in identifying persistent pairwise cointegration between stocks.

5.1.1 Logistic model for conditional prior odds

Arguably the most significant contribution of this thesis is the model for conditional prior odds of cointegration. As the true probability of cointegration is unobservable when testing for pairwise cointegration between stocks, I derive a model based on the cross sectional variation around an unobservable average probability. I model this cross sectional variation by testing for the effects of a set of similarity metrics between companies, where the similarity is calculated based on fundamental attributes. The reasoning for this approach, is that the similarity metrics should act as a proxy for factor exposure, and therefore similar companies should also have similar price-generating processes.

The findings on the relationship between the similarity metrics are somewhat counterintuitive. While similarities with respect to the industry and the book-to-market -ratios of the companies appear to be positively associated with cointegration probability (albeit rather weakly), the findings suggest that similarity between the market capitalizations of two companies is negatively
associated with cointegration probability. This relation appears to be the most significant driver of cointegration probability, and appears robust in a statistical sense.

My findings on this negative relation could potentially explain those of Figuerola-Ferretti et al. (2018), who fail to document improvement in the performance of pairs trading strategies, when the formation of pairs is restricted to only allowing pairs with similar market capitalizations. However, the reason for the observed negative relation between market capitalization similarity and cointegration probability is nonetheless puzzling. One possible explanation could be the method used for the formation of the training and testing samples. The evaluated stocks in the samples are selected based on their liquidity, measured by share turnover volume, in order to ensure that the evaluated stocks would be available for a liquidity demanding trading strategy in real life as well. However, this measure of liquidity is positively correlated with market capitalization. In fact, the average market capitalization similarity between evaluated stocks in the training sample is 7.85 on a scale of one to ten. For a sample with a wider dispersion of market capitalizations, the findings on the effect might turn out to be different, even if the findings in this thesis appear to be quite robust. Further investigation of the negative relation between market capitalization similarity and cointegration probability, and its possible reasons, would be an interesting topic for future research.

In addition to the findings on the effects of the similarity metrics on cointegration probability, a significant contribution of this thesis is the model for computing conditional prior odds in itself. In the introduced model, prior odds for an unobservable quantity are modeled by considering their value conditional on an unobserved average probability. The approach is novel, and therefore could surely be improved by investigating the relations between company attributes and cointegration further. Additional attributes could be considered, and the parameterization of the introduced similarity metrics could also be analyzed and potentially improved.

5.1.2 Empirical performance of the introduced procedure

The empirical performance of the introduced procedure is measured by its ability to identify persistent pairwise cointegration between stocks. As the statistical model determining whether or not a pair of stocks is cointegrated is unobservable, this implies that the persistence of cointegration is also an unobservable quality. Inference on the persistence of cointegration is therefore drawn from the results of a cointegration test, which in this case is based on the Bayes factor of cointegration and a predetermined testing threshold. However, such testing-implied
cointegration should often suffice for a practitioner attempting to profit from a pairs trading strategy, as testing-implied cointegration should be associated with the reversion of the stocks towards an estimated equilibrium.

Despite the theoretical limitations of cointegration testing, the empirical analysis suggests that the introduced procedure is capable of identifying persistent cointegrating relations between stocks, even when inference of cointegration is drawn from the Bayes factor alone, without considering the conditional prior odds. However, when conditional prior odds are introduced, and inference is thus drawn from the conditional posterior odds, the performance of the procedure improves. This improvement appears to be most significant for high classification threshold values. This is especially important, as the highest proportions of Bayes factor exceedances are obtained using classification thresholds well over 30, regardless of the testing threshold that is used. High conditional posterior odds of cointegration appears to be a more robust indicator of a persistent cointegrating relation between a pair of stocks, compared to a high Bayes factor of cointegration. This is evident from the subfigures in Figure 4.2, which illustrate that increasing the classification threshold has a distinctly different effect on the testing threshold exceedance proportions, when using conditional posterior odds for the classification rather than the Bayes factor. This observation implies that the higher the Bayes factor, the more important the prior odds -implied similarity between the companies becomes in identifying persistent pairwise cointegration.

While the introduced procedure appears to be capable of identifying persistent pairwise cointegration between stocks, it nonetheless also identifies a large proportion of pairs for which the identified cointegration does not persist. Given the unobservability of the statistical models determining the pairwise relation between stocks, we cannot infer whether this issue is due to spuriously identified cointegrating relations, or the termination of cointegration during the evaluation period. The termination of cointegration is a plausible explanation, since stocks are likely to experience non-transient idiosyncratic shocks when the evaluation period is long, causing the estimated cointegration relation to terminate. If the cointegrating relation ceases to exist (or was in fact spuriously identified to begin with), the expected return from the trade should be neutral, but catastrophic losses are nonetheless also possible. However, the power of the introduced procedure could also be improved by using a different method for calculating the Bayes factor. As shown by Furmston et al. (2013), a fully Bayesian approach to the Bayes factor is more powerful than the partially Bayesian approach used in this thesis, even if the
Also, a practitioner would likely apply a pairs trading strategy using a large variety of pairs. Therefore, the risks arising from divergence that is caused by a terminated or a spuriously identified cointegrating relation, can be diversified. Even if the majority of pairwise cointegrating relations identified by the procedure fail to persist, the relations that do persist, can nonetheless induce substantial profits to a pairs trader.

5.1.3 Re-visiting the hypotheses

In Section 2.5.1, I specified the two hypotheses tested in this thesis. The first is related to the idea that stocks of similar companies should more likely appear cointegrated than stocks of dissimilar companies, and the null and alternative hypothesis are defined as:

- **Hypothesis 1**: Similarity between companies is not associated with the probability of testing-implied cointegration.

- **Hypothesis 1**: Similarity between companies is associated with the probability of testing-implied cointegration.

Based on the findings, I reject the null hypothesis for the similarity metrics concerning the industry, book-to-market -ratio and market capitalization of companies. The findings suggest that the two former metrics are positively associated with the probability of testing-implied cointegration, while market capitalization is negatively associated with the probability of testing-implied cointegration.

The second hypothesis is related to the persistence of testing-implied cointegration relations, and its null and alternative hypothesis are respectively defined as:

- **Hypothesis 2**: Pairs of stocks that are classified as cointegrated by a Bayesian cointegration testing procedure, are not more likely to be classified as cointegrated in subsequent periods.

- **Hypothesis 2**: Pairs of stocks that are classified as cointegrated by a Bayesian cointegration testing procedure, are more likely to be classified as cointegrated in subsequent periods.
Based on the empirical findings, we can reject the null hypothesis with high levels of confidence. The introduced procedure identifies persistent cointegrating relations with and without conditional prior odds. When conditional prior odds are introduced, the performance of the procedure is further improved, especially when using high classification threshold levels. The findings are contrary to those of Clegg (2014), who fails to find significant persistence in testing-implied pairwise cointegration. The conflicting findings in this thesis might be explained by the shorter evaluation periods, as well as the differences between the Bayesian cointegration procedure used in this thesis, and the classical frequentist cointegration tests used by Clegg (2014).

5.1.4 Limitations and possible improvement areas of the study

As with any novel approach to modeling a phenomenon, the introduced procedure is likely to be far from optimal. Arguably the most significant limitation of the procedure is the small amount of similarity metrics used, as well as their rather simple parameterizations. It is very well possible, that adding additional sources of similarity, such as those corresponding to attributes that have been used in the construction of novel risk factors in more recent asset pricing literature, could be beneficial. As the effects of the similarity metrics on the book-to-market -ratio and industry similarity are rather weak, the introduction of additional sources of similarity could potentially make the two metrics obsolete. In addition, the effects of adding interaction terms between the similarity metrics would be recommended for future endeavors on improving the introduced procedure.

Secondly, the logistic model chosen for the computation of conditional prior odds might not be an optimal choice for maximizing the performance of the procedure. As the logistic model is chosen mostly due to its easy interpretability, the predictive performance might be compromised. More advanced models, that also allow for more complicated non-linear dependence structures, would be an interesting area of exploration. New types of predictive models and their empirical applications are and active area of research, and therefore could provide valuable insights into improving the procedure introduced in this thesis.

Thirdly, the tests performed concerning the persistence of the identified pairwise cointegrating relations do not measure the changes in the parameters of the relation between the periods. If the pair of stocks is classified as cointegrated in both the former and the latter six-month period of a 12-month evaluation period, but the parameters of the estimated cointegrating relation change considerably during the evaluation period, a pairs trader might not be able to profit from the
testing-implied cointegration in the latter period as the weights of assets in the portfolio are not correctly set.

Finally, the method of calculating the Bayes factors could likely be improved. The method used in this thesis is based on the partially Bayesian, but computationally swift approach introduced by Bracegirdle and Barber (2012). A fully Bayesian model with full posteriors for each parameter could improve the power of the introduced procedure. In addition, the parameterization of the residual process could be relaxed, which could potentially capture more complex types of mean reversion in the residual process. It is nonetheless important to keep in mind, that making the model more complicated comes with a cost in computational complexity.

5.2 Concluding remarks

In this thesis I introduce a Bayesian procedure for pairwise cointegration testing between stocks. The procedure utilizes a novel approach of calculating conditional prior odds of cointegration. The conditional prior odds model the cross sectional variation in the cointegration probability as a function of a set of similarity metrics between the pairs of companies, with the absolute level of the prior odds being conditional on the unobserved parameter of average cointegration probability. I parameterize a logistic model for the computation of the conditional prior odds, that accounts for the similarities with respect to the industries, market capitalizations as well as the book-to-market -ratios of the pair of companies.

I find that the similarity between the market capitalizations of two companies is negatively associated with their cointegration probability. On the contrary, the relation is positive for the industry and book-to-market -ratio similarities. The negative relation between market capitalization similarity and cointegration probability is strong and statistically robust, and is the most significant driver of cointegration probability out of the three metrics included in the introduced logistic model.

I test the empirical performance of the introduced procedure on a sample of US-listed common stock. The introduced procedure is capable of identifying persistent pairwise cointegration relations between stocks, even without accounting for the conditional prior odds. The introduction of the conditional prior odds improves the empirical performance of the procedure further, and this improvement is especially significant when using a high classification threshold for the identification of cointegration.
Bibliography


A. Appendix

A.1 System of equations for finding $\alpha$, $\beta$ and $\sigma^2$

In Equation 3.18 the term $E \left( (\epsilon_t - \phi_{-1} \epsilon_{t-1})^2 \right)$ can be completed as

$$E \left( (\epsilon_t - \phi_{-1} \epsilon_{t-1})^2 \right) = \epsilon_t^2 - 2\epsilon_t \epsilon_{t-1} \mu_1 + \epsilon_{t-1}^2 \mu_2.$$  \hspace{1cm} (1.1)

where $\mu_1$ and $\mu_2$ denote the first and second non-central moments of $\phi$. Differentiating 3.18 by $\beta$ and $\alpha$ we get

$$\beta = \frac{ae - cb}{b^2 - ad},$$  \hspace{1cm} (1.2)

$$\alpha = \frac{3d}{b} - \frac{e}{b},$$

where

$$a = 2(T - 1) \mu_1 - (T - 1) \mu_2 - (T - 1)$$

$$b = \mu_1 \sum_{t=2}^{T} (x_t + x_{t-1}) - \mu_2 \sum_{t=1}^{T-1} x_t - \sum_{t=2}^{T} x_t$$

$$c = - \mu_1 \sum_{t=2}^{T} (y_t + y_{t-1}) + \mu_2 \sum_{t=1}^{T-1} y_t + \sum_{t=2}^{T} y_t$$

$$d = 2\mu_1 \sum_{t=2}^{T} (x_t x_{t-1}) - \mu_2 \sum_{t=1}^{T-1} x_t^2 - \sum_{t=2}^{T} x_t^2$$

$$e = \mu_2 \sum_{t=1}^{T-1} (x_t y_{t-1}) - \mu_1 \sum_{t=2}^{T} (y_t x_{t-1} + x_t y_{t-1}) - \sum_{t=2}^{T} x_t y_t.$$  \hspace{1cm} (1.3)

With estimates for $\alpha$ and $\beta$, we recalculate $\epsilon_t$, differentiating 3.18 by $\sigma^2$ we get

$$\sigma^2 = \frac{\sum_{t=2}^{T} \epsilon_t^2 - 2\mu_1 \sum_{t=2}^{T} \epsilon_t \epsilon_{t-1} + \mu_2 \sum_{t=1}^{T-1} \epsilon_t^2}{T - 1}.$$  \hspace{1cm} (1.4)
## A.2 Separate regressions of testing-implied cointegration on similarity metrics

<table>
<thead>
<tr>
<th>Table 1.1. Logistic regressions of cointegration on company similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>This table presents the coefficient estimates and standard errors from three logistic regressions, where the dependent variable is the binary result of a cointegration test between a pair of stock prices. The cointegration test considers whether the Bayes factor (BF) of cointegration is above a predetermined threshold. Regressions using threshold levels of unity, five, ten and 20 are reported. The explanatory variables of the regressions consist of the similarities between the pair of companies whose stock prices are evaluated, with respect to their industries, market capitalizations, and book-to-market-ratios. Panels A, B and C report the results of the regressions with varying evaluation periods for the Bayes factor, where the evaluation periods are six, 12, and 30 months respectively. The evaluation periods are non-overlapping, constituting a 60-month period between January 2009 and December 2013. The sample consist of moderately liquid US common stock listed on the NYSE, AMEX and NASDAQ. Penny stocks are excluded. Statistical significance is denoted by *, ** and *** at the 5%, 1% and 0.1% levels respectively.</td>
</tr>
</tbody>
</table>

### Panel A: Six-month evaluation periods

<table>
<thead>
<tr>
<th>Attribute</th>
<th>BF &gt; 1</th>
<th>BF &gt; 5</th>
<th>BF &gt; 10</th>
<th>BF &gt; 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.787***</td>
<td>-3.066***</td>
<td>-3.597***</td>
<td>-4.076***</td>
</tr>
<tr>
<td>(0.007)</td>
<td>(0.012)</td>
<td>(0.015)</td>
<td>(0.019)</td>
<td></td>
</tr>
<tr>
<td>Industry</td>
<td>0.031***</td>
<td>0.069***</td>
<td>0.067***</td>
<td>0.077***</td>
</tr>
<tr>
<td>(0.009)</td>
<td>(0.014)</td>
<td>(0.018)</td>
<td>(0.022)</td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-1.617***</td>
<td>-2.607***</td>
<td>-2.974***</td>
<td>-3.373***</td>
</tr>
<tr>
<td>(0.029)</td>
<td>(0.047)</td>
<td>(0.058)</td>
<td>(0.072)</td>
<td></td>
</tr>
<tr>
<td>Market capitalization</td>
<td>-0.021***</td>
<td>-0.058***</td>
<td>-0.080***</td>
<td>-0.090***</td>
</tr>
<tr>
<td>(0.004)</td>
<td>(0.006)</td>
<td>(0.008)</td>
<td>(0.009)</td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-1.871***</td>
<td>-3.053***</td>
<td>-3.518***</td>
<td>-3.824***</td>
</tr>
<tr>
<td>(0.021)</td>
<td>(0.035)</td>
<td>(0.044)</td>
<td>(0.053)</td>
<td></td>
</tr>
<tr>
<td>Book-to-market</td>
<td>0.013***</td>
<td>0.000</td>
<td>-0.009</td>
<td>-0.021**</td>
</tr>
<tr>
<td>(0.003)</td>
<td>(0.005)</td>
<td>(0.006)</td>
<td>(0.007)</td>
<td></td>
</tr>
</tbody>
</table>

### Panel B: 12-month evaluation periods

<table>
<thead>
<tr>
<th>Attribute</th>
<th>BF &gt; 1</th>
<th>BF &gt; 5</th>
<th>BF &gt; 10</th>
<th>BF &gt; 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.243***</td>
<td>-3.300***</td>
<td>-3.724***</td>
<td>-4.185***</td>
</tr>
<tr>
<td>(0.013)</td>
<td>(0.020)</td>
<td>(0.025)</td>
<td>(0.031)</td>
<td></td>
</tr>
<tr>
<td>Industry</td>
<td>0.052***</td>
<td>0.089***</td>
<td>0.094***</td>
<td>0.086*</td>
</tr>
<tr>
<td>(0.015)</td>
<td>(0.022)</td>
<td>(0.027)</td>
<td>(0.034)</td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.317***</td>
<td>-3.162***</td>
<td>-3.262***</td>
<td>-3.236***</td>
</tr>
<tr>
<td>(0.057)</td>
<td>(0.089)</td>
<td>(0.103)</td>
<td>(0.120)</td>
<td></td>
</tr>
<tr>
<td>Market capitalization</td>
<td>0.011</td>
<td>-0.015</td>
<td>-0.057***</td>
<td>-0.121***</td>
</tr>
<tr>
<td>(0.007)</td>
<td>(0.011)</td>
<td>(0.013)</td>
<td>(0.016)</td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.469***</td>
<td>-3.561***</td>
<td>-4.005***</td>
<td>-4.595***</td>
</tr>
<tr>
<td>(0.040)</td>
<td>(0.064)</td>
<td>(0.078)</td>
<td>(0.100)</td>
<td></td>
</tr>
<tr>
<td>Book-to-market</td>
<td>0.032***</td>
<td>0.038***</td>
<td>0.041***</td>
<td>0.057***</td>
</tr>
<tr>
<td>(0.005)</td>
<td>(0.008)</td>
<td>(0.010)</td>
<td>(0.013)</td>
<td></td>
</tr>
</tbody>
</table>

### Panel C: 30-month evaluation periods

<table>
<thead>
<tr>
<th>Attribute</th>
<th>BF &gt; 1</th>
<th>BF &gt; 5</th>
<th>BF &gt; 10</th>
<th>BF &gt; 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.538***</td>
<td>-3.385***</td>
<td>-3.728***</td>
<td>-4.085***</td>
</tr>
<tr>
<td>(0.026)</td>
<td>(0.038)</td>
<td>(0.045)</td>
<td>(0.053)</td>
<td></td>
</tr>
<tr>
<td>Industry</td>
<td>0.061</td>
<td>0.077</td>
<td>0.072</td>
<td>0.048</td>
</tr>
<tr>
<td>(0.037)</td>
<td>(0.052)</td>
<td>(0.062)</td>
<td>(0.077)</td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.822***</td>
<td>-3.193***</td>
<td>-3.485***</td>
<td>-3.536***</td>
</tr>
<tr>
<td>(0.137)</td>
<td>(0.189)</td>
<td>(0.221)</td>
<td>(0.232)</td>
<td></td>
</tr>
<tr>
<td>Market capitalization</td>
<td>0.036*</td>
<td>-0.022</td>
<td>-0.028</td>
<td>-0.067*</td>
</tr>
<tr>
<td>(0.016)</td>
<td>(0.023)</td>
<td>(0.027)</td>
<td>(0.031)</td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.461***</td>
<td>-3.018***</td>
<td>-3.371***</td>
<td>-3.644***</td>
</tr>
<tr>
<td>(0.082)</td>
<td>(0.111)</td>
<td>(0.131)</td>
<td>(0.152)</td>
<td></td>
</tr>
<tr>
<td>Book-to-market</td>
<td>-0.008</td>
<td>-0.044***</td>
<td>-0.043**</td>
<td>-0.054**</td>
</tr>
<tr>
<td>(0.010)</td>
<td>(0.013)</td>
<td>(0.016)</td>
<td>(0.018)</td>
<td></td>
</tr>
</tbody>
</table>