Master's Thesis: Parameter Sensitivity Analysis in Pair Trading Strategies

Erasmus University Rotterdam

Janko Huic
Supervisor: Prof. Sebastian Vogel
Word Count: 9381

11/04/2023

Abstract

Pair trading strategies have been well-documented for almost two decades. Naturally, much of the literature has gone towards increasingly complex mathematical models. Often however, certain parameters of the trading aspect are taken as a given, following in the footsteps of a now 17 year-old paper. The present analysis follows a different approach: I take the pair formation methods as given, and analyze the extent to which the performance of this strategy varies with input parameters. I find that, although the in-sample optimum of about 1.6σ for the trigger rule is close to the canonical 2σ rule, the common reversion threshold of 0σ in the literature is a far departure from the optimal value of above 1σ, depending on formation method.

Introduction

Much attention has been devoted to devising many kinds of highly mathematical methods of identifying undervalued stocks, and to the construction of sophisticated trading algorithms. However, there is a decades-old Wall Street trading strategy which seems to produce high returns, with comparatively little complexity or risk.

In the seminal paper by Gatev, Goetzmann, and Rouwenhorst (2006), henceforth referred to as GGR, the authors found significant and robust positive returns to the so-called pairs trading strategy, and the authors emphasized its simplicity and straightforward applicability. Without requiring prohibitive computational requirements, incurring excessive trading costs, or using complicated synthetic instruments, it was shown that one could earn significant excess returns on market-neutral positions with a set of simple trading rules, which are as follows. Firstly, we need to identify a pair of stocks whose prices move in unison (or at least as close as we can find). Then we observe their prices, and once they diverge from their common trajectory, we buy the stock which has fallen, and short the stock which has risen in price. We then wait for them to revert to their common price levels and pocket the arbitrage profits. The main idea behind this strategy is that, if two stocks move very similarly to one another, they represent assets which provide similar payoffs across different states of the world. If that is indeed the case, our bet is that the divergence in prices is random in nature, and in absence of further disruptions, the two prices should revert back to their long-term levels which are close together.

Despite its striking simplicity, this strategy has been widely corroborated by the literature since the 2006 paper brought it into the academic spotlight, although a downward trend in performance over time has been noted by the literature. The strategy has been expanded and advanced in many directions, most of which tend towards complexity, and the literature has naturally tended towards expanding the mathematical frameworks for modeling pairs rather than examining the intricacies of the strategy in its simplest form, as it was in GGR. These include, but are not limited to, a cointegration model with a logistic mixture auto-regressive equilibrium error of Cheng et al., (2011), and a three-regime threshold autoregressive GARCH models of Chen et al. (2014). The present paper aims to go the other way - to keep the strategy steps and rules as simple as possible, but to scrutinize the particular strategy parameters which play a key role in the realization of profits. This is an attempt to keep the pair trading strategy in the realm of "disarmingly simple," as described by GGR in their seminal paper, and to also scrutinize a side of the strategy which tends to be overlooked - return sensitivity to the trading parameters.

Specifically, in GGR they use two rules to determine the opening and closing of the appropriate positions in any given pair: when the two stocks deviate by more than two standard deviations of their price difference during the formation period, they enter the position by buying the falling stock and shorting the rising stock (in equivalent amounts). Then, they unwind the positions once the stocks' prices cross once again. Furthermore, if the prices do not cross within the trading period, then the position is closed at the last day of said trading period. The authors do not specifically elaborate on the reasons behind these particular thresholds. They are sensible and understandable, of course, but there is an effectively infinite parameter space to choose from. Moreover, although whole-number multiples of the standard deviation are very natural to humans, there is no reason to believe this particular threshold is either more or less conducive to successful execution of this strategy than any other. For this reason, the present analysis will look to examine the sensitivity of the pairs trading strategy performance on variations in the strategy parameters, namely the thresholds for closing and opening a position. Additionally, I will also analyze how the number of top pairs included in the strategy affects the strategy profits. Lastly, I will compare different methods of obtaining pairs to one another.

Literature

To identify the stock pairs which move along in unison, Gatev et al. (2006) calculate the sum of squared deviations between normalized price levels of the stocks for a period of time, and choose the pairs with

the lowest values. Then, they use a trading rule such that they open the appropriate positions in each stock of the pair once they diverge by more than two historical standard deviations, and unwind the positions when the prices next cross one another. They find that trading the top pairs can result in 11% annualized return. However, they keep those trading rules constant throughout the paper, and they do not analyze further how the performance changes depending on the trading rules. This specifically is the area where the present paper aims to contribute to the literature - to vary the trading strategy parameters and analyze how the profits change.

In addition to calculating the sum of squared deviations to identify pairs, another tempting methodology is to consider stocks which have the highest return correlation between one another. One of the benefits would be the fact that this method would identify pairs which move up or down in unison, but to differing extents. For instance, if stock B moves 2% for each 1% of stock A's movement, these two would be unlikely to be identified by the SSD approach, even if they had perfect correlation. However, due to the inherent randomness, daily return data may not be ideal for this purpose. Rather, in the literature it is more common to use lower-frequency data. For instance, Chen et al. (2019) use monthly return data to compute the pairwise return correlations. They corroborate the results of GGR, and find evidence that those returns can not be explained by common risk factors, investment based factors, liquidity risk, or leverage factors. Rather, they find that the pair trading strategy profits are related to, and can be partially explained by, short-term reversal and the one-month version of the industry momentum.

Krauss (2017) provides an overview of the literature on the pairs trading strategy, outlining five general groups of methodologies of identifying pairs: distance approach, cointegration approach, time-series approach, stochastic control approach, and other approaches. The SSD measure used in the present analysis, as well as GGR, falls into the distance approach. The author further describes it as follows: "In the formation period, distance metrics are leveraged to identify comoving securities. In the trading period, simple nonparametric threshold rules are used to trigger trading signals. The strategy is simple and transparent, allowing for large-scale empirical applications. With these studies, pairs trading is established as profitable across different markets, assets, classes, and time frames." Therefore, this approach is characterized by the simplicity and transparency of both the pair identification, and the trading stages. This is quite important, as the trading rules respond well to simple, linear variations in the inputs, and pair identification is relatively straightforward. Moreover, the arbitrary nature of threshold choices make the literature in this area particularly prone to selection bias and p-hacking. Specifically, the wide range of choices for strategy parameters makes it such that there are bound to be parameter combinations which result in high excess returns, while being outliers in the true parameter space. This is an additional contribution of the present analysis to the literature.

Much of literature on the topic of pair trading strategies focuses mainly on one aspect of the process - the specific mathematical methods of identifying and ranking stock pairs. Specifically, most of the more recent attention appears to be on copula-based and cointegration methods which employ sophisticated statistical models to identify stock pairs. In addition to finding comoving prices of stocks, the additional complexity of these models is sometimes also motivated by the goal of distinguishing between true pairs and those whose pairing is more spurious in nature.

Smith & Wu (2017) is among the few academic papers which addresses the issue of a lack of analysis with respect to parameter-return variability. Unlike most literature which uses the standard of two standard deviations of historic price difference as the trigger rule, they calculate the returns of their formed pairs over a set of seven different values of the divergence threshold, and find that the 2σ trading rule, albeit common in the literature, is not the optimal trigger value. However, they keep the reversion threshold constant with the standard rule of closing positions when the prices cross. Moreover, they note that the return profile of top pairs depends on the divergence threshold as a lower threshold means more trades, but a comparatively smaller gap between the prices from which to draw profit. Conversely, a high trigger would cause relatively few trades to go through; however, those few trades would be more profitable on average. Furthermore, they also attempt to tackle the important question of how many of the top pairs it is most profitable to include in the strategy. In addition to the trigger rule

analysyis, they also find variability of returns with respect to the trading period duration. Similarly to Huck (2013), they find high sensitivity of the returns to changes in the timing of the formation periods.

In summary, while the formation of pairs is increasing in complexity, with some papers also employing genetic algorithms (Huang et al., 2015), in much of the literature the trading rules which govern the opening and closing of positions in the subsequent period remain similar to, or exactly the same as, the earliest papers on the topic (namely GGR). Specifically, the trading rule where one opens the positions in either leg of the pairs once the difference in prices crosses two standard deviations of price difference from the previous period, and closes the positions either when the prices cross, or when they revert to a sufficiently low difference. This is the gap in literature which the present paper attempts to address. Namely, I seek to explore and document the extent to which pair based trading strategy performance depends on specific parameters, most of which tend to get overlooked by a lot of the literature. Parameters in question would be divergence and reversion thresholds, data frequency, time period variations, and number of top pairs traded.

Data & Methodology

The present paper is based on the CRSP Compustat merged database of US stocks between January 2010 and June 2022. For different purposes in the analysis, I will use either daily or weekly price and return data. All of the prices are adjusted to include cash equivalent distributions, reinvestment of dividends, and the compounding effect of dividends on reinvested dividends, as this more closely aligns with the kinds of returns a representative trader faces in the real world.

Due to the computational requirements of calculating pairwise statistics such as sum of squared deviations, Pearson correlations, and other, on a large number of stocks and over a long period of time, the considered timeline is only after 2010; other concessions had to be made for the same reason. As there are n * (n - 1) / 2 pairs of stocks for which these statistics need to be calculated, where n is the total number of stocks considered, the complexity increases exponentially with n. Ideally, there are many more dimensions of analysis which the present paper could have explored. However, due in large part to the aforementioned computational constraints, these could not be included, and the analysis will take a more concise approach than optimal.

The stocks have been selected such that they are few enough for the computational requirements to allow for extensive analysis, yet they represent the bulk of the market capitalization in the US stock market. The stocks are selected such that at the beginning of each formation period, the included stocks comprise 95% of the total market capitalization in the CRSP Compustat database. This makes for an average of 1545 stocks per formation period, with 1452 in January 2010, and 1489 in January 2021. Although comparatively fewer than the set of stocks which most of the literature on this topic takes into consideration, the fact that these stocks comprise 95% of total market capitalization of US stocks means that these stocks represent the vast majority of trading profit available to institutional investors. Because the market capitalization of individual stocks around the 95-percentile is relatively low compared to the larger stocks which are included, potential stock pairs may not be of practical value for institutional investors due to the price impact of large trades for small stocks, as well as low trading volume, and potentially high bid-ask spreads. It is worth noting that in order for a successful trade of stock pairs, both stocks need to be sufficiently liquid, the chance of which decreases as we start to include many more smaller stocks for marginal increases in total considered market capitalization. Therefore, this stock selection was made as a middle ground between including as many stocks as possible, while keeping the computational requirements manageable.

For the formation and trading of pairs, I follow the approach by GGR and much of the subsequent literature - first, there is a 12-month formation period in which stock pairs with the lowest (highest) sum of squared deviations (correlations or cross count) are identified. Then, there is the 6-month trading period, during which the identified pairs are traded according to the strategy rules. In the present analysis, there are 23 such formation-trading periods beginning with January 2010 and ending with June 2022, staggered by 6 months. This is done such that the 6-month trading periods are

followed end-to-end by one another, with sequential 12-month formation periods overlapping each other by 6 months. For instance, the first formation period is during the whole calendar year of 2010, with its corresponding trading period being in the following 6 months, that is January to June (inclusive) of 2011, while the second formation period is staggered by 6 months, such that it takes place between July 2010 and June 2011 (inclusive), with the corresponding trading period taking place between July and December 2011 (inclusive). For each such formation period, I take the price data (adjusted for reinvested dividends) on all stocks in the sample, and exclude stocks which record any missing observations. Then, I normalize all stock prices such that the starting price for that period equals 1. Finally, I compute the sum of squared deviations for each of the n*(n-1)/2 pairs in the sample, and rank the pairs by SSD (cross count) in ascending (descending) order. For the pairs formed on correlation, no normalization is necessary, and the coefficients are computed on the basic adjusted daily and weekly return data, after which I rank the pairs in descending order.

Pearson correlations are calculated on weekly, as well as daily, return data, because this measure perhaps does not benefit from as high frequency data as the sum of squared deviations. Small perturbations on a short-term scale can significantly change how this metric evaluates pairs of stocks, even if they tend to move alike over a longer timeline. With this line of reasoning, it may be possible that correlations based on lower-frequency data could produce better pairs, because they capture pairs which have a stronger long-run correlation, rather than those which exhibit similar day-to-day idiosyncrasies. Many academic papers on pair trading do use correlation based on high frequency data to identify pairs, though. The difference is that in those papers, their trading windows are proportionally brief (e.g. Kim, 2011). In the present paper, however, both the formation and the trading windows could be too long for high-frequency correlation to be particularly useful. Therefore, I use weekly return data to identify Pearson correlation-based pairs in addition to the daily return data, in order to compare the two approaches.

The last method is based on a simple measure which to the best of my knowledge has not been included in the literature on pair trading. This is simply a count of how many times two stocks cross each other during the formation period, also referred to in the following analysis as "cross count". These are the methods which will result in four different sets of identified pairs for each considered time period. These four methods of identifying pairs will henceforth be referred to as formation methods, with their respective statistics (SSD, cross count, and Pearson correlation) being referred to as formation statistics. Although these sets of pairs are obtained by different means, the subsequent analysis of how these pairs perform will be the same across all methods.

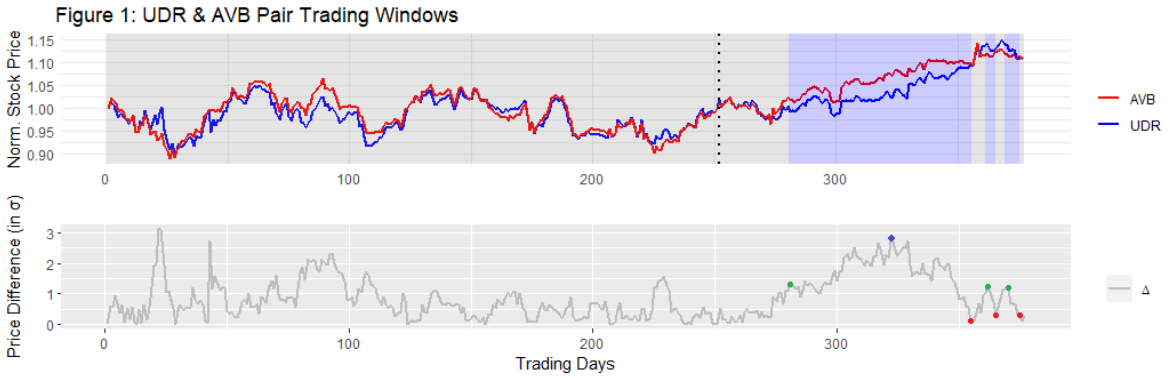
The next step is applying the trading strategy on the identified pairs in the subsequent 6 months of price data. The trigger rules are simply a set of thresholds which, if reached, result in us opening or closing a position. They are expressed as a multiple of the standard deviation of the difference in the normalized prices during the formation period. Then, the prices are calculated given these rules for opening and closing the positions. Specifically, once the divergence threshold is met, I calculate as if one unit of currency was invested in a long position of the fallen stock, and as if one unit of currency worth of the rising stock was shorted. Once the positions are closed, I add up the return of the long leg and the short leg to get the return for that position. Note that I do not make any adjustments to the strategy profit for the cost of transaction or shorting. The analysis is instead focused on the theoretical variability in performance, and accurate details of a real-world application of the pair trading strategy is out of the scope of the present paper.

During a trading period, there may be multiple occasions on which the positions open and close - in that case, the total profit for a traded pair is simply the sum of all such positions. This may result in positions which are open throughout the trading period, or positions which are traded for only a few days, or none at all. Therefore, this measure of profit would evaluate equally a pair which yields return r during only a single day of trading as a pair which yields an identical return r over the full 6 months of trading. This could be seen as a flaw of the measure because intuitively, one would much prefer for the same return to be realized over a shorter period of time. Due to the fact that the long leg and the short leg of the trades offset each other in terms of expected return and risk, this measure

of profit is interpreted as excess return (GGR). In fact, literature shows that risk-adjusted returns of this strategy tend to be very close to raw returns, as most of the excess return is market neutral (Smith and Wu, 2017); still, the positions are not risk-free because their return variance is non-zero. Furthermore, because the short leg of the position fully finances the long leg (at least in theory), the capital requirements for this trading strategy are low, and the duration of the positions is less of a factor than it would be in traditional stock trading due to the lower capital employment. Nevertheless, realizing the same return in a shorter time span is strictly preferred to realizing it over a longer time span. For this reason, in addition to the simple return which was explained earlier in this paragraph, I will also include the annualized return in the analysis and discussion of the results. This is calculated simply by dividing the total return of the strategy by the proportion of total trading days in a year during which the position was open (e.g. if the trade lasted 50 trading days, the adjustment factor would equal 50 / 252). This calculation is based on an arithmetic, rather than a geometric, annualization because the latter has a tendency to disproportionately inflate the positive returns compared to the negative returns, which results in an inaccurate skewing of the return distribution and an erroneous increase in the mean return. Still, this annualized return value should not be taken as a representative measure for the actual realizable yearly return. The returns from this strategy are highly opportunistic, and are therefore not realistically scalable to their mathematical yearly return. Rather, this is more of an informative measure meant to assist us in comparing between different formation methods.

While GGR use two standard deviations of historical price difference as the trigger to open the positions and the prices crossing as trigger to close the positions, the point of the present analysis lies in not conforming to one single strategy set. Rather, I iterate over a range of values both for the opening, and for the closing of positions. Specifically, the range is expressed in multiples of the standard deviation which is taken as the deviation/reversion threshold.

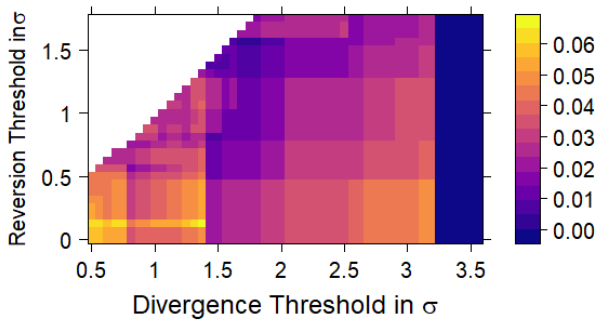
To better understand the interplay between deviation and reversion thresholds with the strategy performance, below is Figure 1 which plots one example pair between United Dominion Realty Trust and AvalonBay Communities Inc. The top figure plots the normalized price levels of the two stocks for the period of Jan 2016 to June 2017, with the vertical dashed line signifying the cutoff point between the formation period, which is the calendar year of 2016, and the trading period, which is the first six months of the following year, 2017. The bottom figure plots the absolute value of the difference in the stock prices, as measured in standard deviations of the historical difference during the formation year. Lastly, the shaded area in the top figure denotes the period during which the positions are open with the ex-post optimal combination of deviation and reversion thresholds, while the points highlighted in green (red) denote the points at which we open (close) the positions.



Looking at just the line denoting the price difference in terms of our trading strategy, we can tell that we should be looking at points of opening and closing the trades with the largest vertical drop. With this simple interpretation in mind, we would ideally want to open the position at the topmost inflection point (marked by the blue dot), and continue until the same point at which we close the first position. However, it so happens that setting this high threshold would result in us missing out on the profit which we realize during the subsequent two positions.

To visualize the relationship between these thresholds and the total profit during the trading period, I iterate over 1500 parameter sets to populate the heatmap in Figure 2. Note that the top-left corner is whited-out for all y > x, as a strategy where the reversion threshold is lower than the divergence threshold is not logically consistent with our strategy.

Figure 2: UDR & AVB Return Heatmap



The single bright yellow point at $x = 1.36\sigma$ and $y = 0.37\sigma$ is the optimal combination of trading thresholds. Even the slightest nudge to the right, to the adjacent node representing (x, y) = (1.41, 0.37), the return drops from 6.49% to just 2.15%. From there, we can see that a further increase of the divergence decreases the profit, although we allow the trade to kick in at a higher difference between the prices. Looking back at Figure 1, this would be equivalent to pushing the position opening towards the highest point in the difference graph. Increasing further, the profit increases until we surpass the point of the largest price difference, after which no positions would be opened, and we would realize zero profit. The variation in total return relative to the reversion threshold is not so clear. There is a tendency for lower thresholds to perform better, as it allows the prices to revert as much as possible before we close the position, realizing a higher profit; however, higher thresholds tend to result in lower losses, since the probability of closing the positions on an above-zero profit is lower. It is not immediately clear which of these two effects prevails over a sample of many pairs.

It is important to note, however, that the characteristics of the pair trajectory which result in this heatmap are highly particular to the specific price trajectories which were recorded during this period. As such, the insights gained from this specific pair analysis are not meant to guide our overall trading strategy. Rather, the figure is here to visualize one example of how the divergence and reversion thresholds may impact the total realized profit of the pair trading strategy during one trading period. In this particular example, the optimum would be impossible to precisely guess ex-ante, and the analysis is merely an ex-post consideration. However, the value of this example comes from showing how the specific stock movements can create intriguing relationships in the payoff structure. Much of the analysis in the present paper, therefore, will aim to investigate whether there exist patterns in these relationships which we can use to tune our trading rules in order to maximize the performance of the pair trading strategy.

Results

To summarize, the results in this section are obtained from the top 100 pairs identified in each of the 23 time periods from January 2010 until December 2021. Therefore, the periods during which the stocks are traded range between January 2011 to June 2022. There are four methods of identifying pairs. The first one is the sum of squared deviations (SSD), second is the number of times the stock prices cross during the period (cross count), and last are two Pearson correlation coefficients, one based on daily return data, and the other on weekly return data. The following results, unless specified otherwise, will always be based on the simple threshold rules as in GGR (enter at two standard deviations of divergence and exit at price crossing), and will cite the simple (non-annualized) pair return. Still, in some of the following analysis I will vary these thresholds systematically to better understand the relationship between strategy parameters and strategy return. I will analyze pair trading performance for a spectrum of values both for the divergence, and for the reversion thresholds.

Parameter Sensitivity Analysis

Following the computation of all $n*(n-1)/2$ pairwise formation statistics, for each of the 23 formation-trading periods, I take the 25 pairs with the lowest (highest) SSD (correlation or cross count) for each period, and compute the trading profit over 1500 different sets of strategy parameters for each pair during the trading period, to obtain sufficient data points to populate the heatmap (50 different values of divergence threshold between 0.5σ and 3.5σ, and 30 different values of the reversion threshold between 0σ and 1.75σ. I then average over all of the 575 considered pairs (25 pairs for each of the 23 periods) to obtain the aggregate heatmap for each of the four methods of obtaining pairs. Therefore, the heatmaps in Figures 3 to 6 represent the total aggregate profit one would have realized, were they to trade the 25 top pairs in each formation-trading period, conditional on the trigger rules (divergence and reversion thresholds) used in the trading strategy. The four separate heatmaps represent pairs obtained from the four considered pair formation methods.

This is done with the intention of being able to visualize how the profit of the pair trading strategy is related to the specific parameters which we use as part of the trading trigger rules. With the purpose of it being that if we can identify certain kinds of patterns in the relationship, we can make better decisions with regards to setting the trigger parameters which we employ as part of the trading rules.

Figure 3: Top 25 Pairs From Each Period: SSD

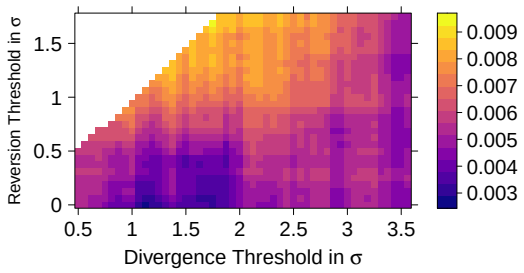


Figure 4: Top 25 Pairs From Each Period: Cross Count

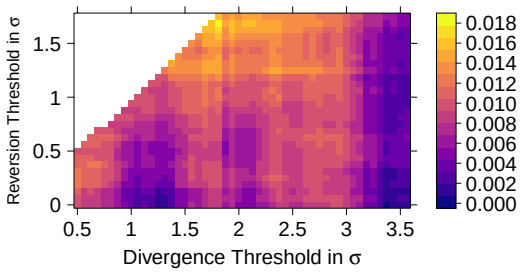


Figure 5: Top 25 Pairs From Each Period: Daily Correlation

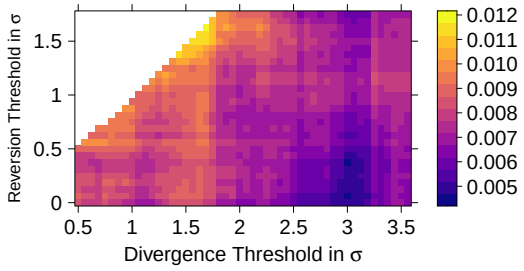
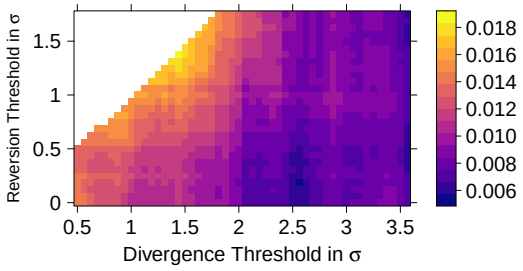


Figure 6: Top 25 Pairs From Each Period: Weekly Correlation

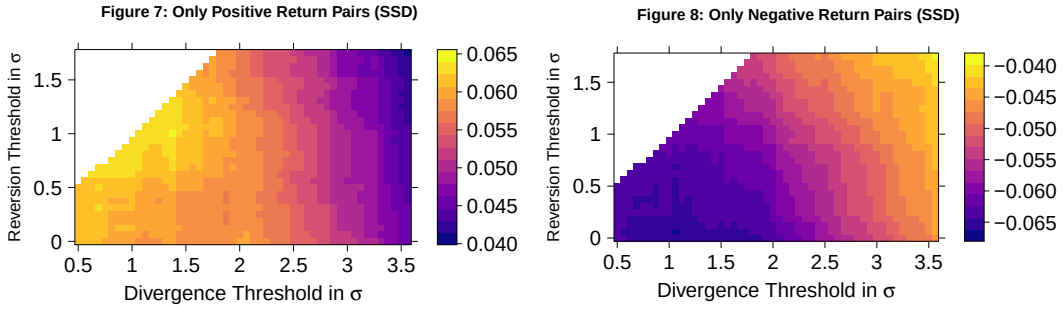


Perhaps the most noticeable pattern in the figures is that across all of the four formation methods, the highest profits tend to lie relatively far towards the upper parts of the graph. That is, it appears that higher reversion thresholds result in better performance of the pair trading strategy. The in-sample uptimal reversion threshold values are 1.68σ, 1.68σ, 1.37σ, and 1.62σ for pairs formed on SSD, cross count, weekly return correlations, and daily return correlations, respectively. For one, this is somewhat unexpected because nearly all of the literature on the topic uses the rule of closing positions once the prices cross one another, following in the footsteps of GGR. This strategy is in the middle of the bottom-most row of the heatmaps, as it is de facto represented by a reversion threshold of 0. Furthermore, most of the highest performing points tend to be close to the top-left edge, after which lies the area where the divergence thresholds are lower than the reversion thresholds. This means that the highest performing strategies are those where the divergence and reversion thresholds are very close to one another in magnitude, which tends to result in many brief trades, each of which presumably

only realizes a small amount of profit. Moreover, these kinds of trades tend to be relatively safe in that a small divergence naturally has a larger probability of reversion than a large divergence.

As for the divergence threshold, the optimal value tends to be below the generally accepted two standard deviations; however, there is somewhat more variation across formation methods than there is for the reversion threshold. Nevertheless, the local optima of the four figures all lie around the value of 1.5, which is reasonably close to the literature standard. Specifically, these are 1.7σ, 1.76σ, 1.4σ, and 1.64σ for pairs formed on SSD, cross count, weekly return correlations, and daily return correlations, respectively. Therefore, the evidence suggests that the canonical trigger value of two standard deviations is not far from the optimum. However, considering the two-dimensional parameter space, getting one of the two parameters right does not necessarily lead to the strategy as a whole being close to the optimum.

Looking back at the example in Figure 1, we know that from the profit side of things, we would ideally open the position at maximum divergence, and close as late as possible upon reversion, in order to maximize the vertical distance which we cover in terms of the bottom graph, representing the difference in prices. The downside, however, is that an exceedingly high divergence threshold results in the strategy missing out on smaller, safer profits. In case of a loss, we would again want to open at maximum divergence, but close as early as possible, to conversely minimize the exact same calculation. Furthermore, the lower our reversion threshold, the more likely that an open position in a pair will not reach the threshold, and will instead diverge until the end of the period when we close all positions. Therefore, we can think of the tradeoff in trigger rules as managing the two ends of the spectrum - maximizing profit in case of reversion, and minimizing losses in case of divergence. To visualize this relationship, I dissect Figure 3 representing the SSD pairs into two figures: one which only includes ex-post profitable pairs, and one which only includes ex-post unprofitable pairs.



We get two clear pictures in Figures 7 & 8 which quite well match our expectations with one important observation: while the unprofitable figure exhibits a smoothly monotone gradient towards larger threshold values on either axis, the figure which includes only profitable paris exhibits its local maximum at a divergence threshold of 1.35 standard deviations and reversion threshold of 1 standard deviation. This indicates that there is more to capturing the profit of a successful pair than is immediately obvious. Smith and Wu (2017) put forward an interpretation of this relationship between the trigger rule and pair profit. As they find a hump-shaped relationship between the two, they explain it by pointing out that although higher trigger values result in profit increases of pairs which diverge by a greater amount, this comes at the cost of foregoing trading the pairs which would not diverge by that much. Therefore, a high trigger value produces relatively few highly profitable trades, while a low trigger value produces a larger number of moderately profitable trades, and it is not always clear which effect is more powerful. It is worth noting that, because the present analysis lacks the adjustments for transaction cost, the optimal strategy as currently identified is likely biased towards the lower end of the divergence threshold, as it ignores the increased cost of higher frequency trading, which the literature has found to impact pair trading strategies (Do & Faff, 2012).

Portfolio Size

Perhaps the second most important factor in the construction of a robust pair trading strategy is one relating to the scalability of this approach to trading. Namely, depending on the market capitalization and liquidity of the top stock pairs, there is limited profit to be made from such a strategy. While for a retail investor, an investment of a few tens or hundreds of thousands of dollars is perfectly acceptable, institutional investors which are the most likely entities to perform these strategies tend to look for greater earning potential. This issue is further exacerbated by the fact that the potential of any individual pair is constrained by the component with the lowest liquidity or market capitalization out of the two stocks, and that both stocks of the pair should ideally be highly liquid, medium to large-cap stocks for the strategy to be at its most profitable. The most obvious solution to this constraint is to increase the number of pairs which we consider for trading. However, the farther down we go the list of top pairs, the less we should be confident that the considered stocks do in fact act as true pairs, and the more likely it is that their correlation is instead spurious. This brings up one of the issues which come up in the practical application of the pair trading strategy - determining ex-ante how many of the top pairs we should choose to open positions in.

Table 1 presents the overall mean returns across all considered years of the top pairs in each period, for a total of 2300 pairs for each formation method. The rows signify how many of each period's top pairs were considered in the calculation of the mean - top 5, top 10, top 50, and top 100 pairs. It is important to note that, as previously discussed, the annualized returns in this calculation are deceptively high because of the difference in trade duration between profitable and unprofitable pairs. Due to the nature of the strategy, some of the positive trades are very brief, as low as one trading day, whereas the only way an unprofitable pair can be that brief is if the position opens very close to the end of the period. Therefore, the distribution of trade duration means that brief profitable positions are unproportionately inflated, while the unprofitable pairs are adjusted by a lower factor on average. Therefore, the annualized returns are not meant to represent a realizable return of scaling the strategy; they are presented because they add another dimension of comparison between formation methods, as they represent an imprefect measure of return which takes into account total trade length as well as the simple trade profits.

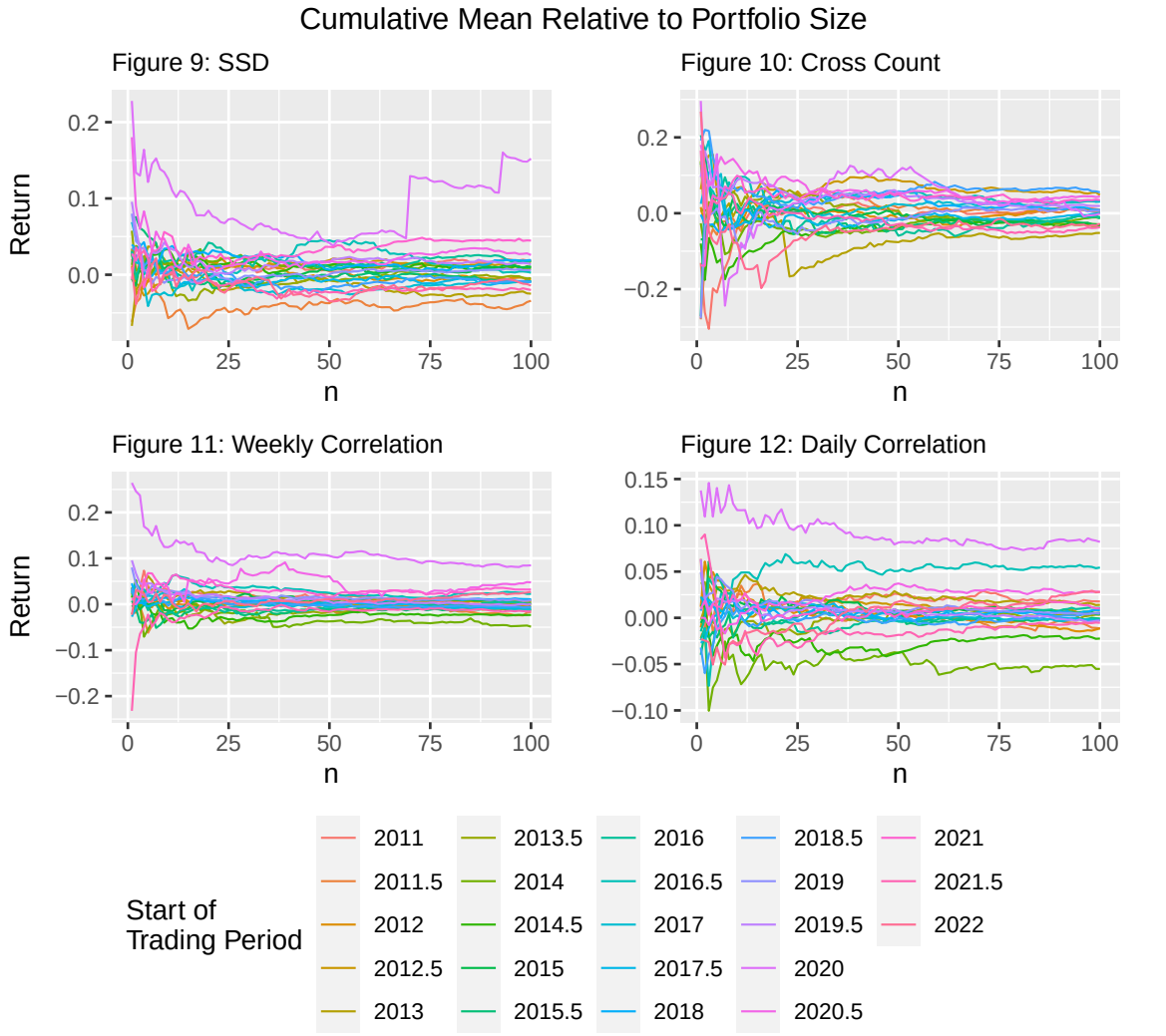
Table 1: Summary of Top Pair Mean Performance: Simple & Annualized Returns

Top n Pairs	SSD	Cross Count	Corr Weekly	Corr Daily	SSD ann.	Cross Count ann.	Corr Weekly ann.	Corr Daily ann.
Top 5	0.0172	0.0115	0.0136	0.0100	0.2811	0.2730	0.2190	0.0947
Top 10	0.0130	-0.0029	0.0110	0.0100	0.1963	0.1813	0.1621	0.0844
Top 50	0.0031	0.0087	0.0061	0.0089	0.1138	0.2648	0.1446	0.1145
Top 100	0.0090	0.0022	0.0045	0.0070	0.1298	0.2385	0.1344	0.1082

The first notable observation is that, as one would expect, the average profit per pair tends to go down as we trade more and more of the top pairs. The intuitive explanation would be that the top-ranking pairs are also of highest quality, so to speak, and are therefore the most viable options for a profitable pair trade, while the lower we go on the list of top pairs, the more idiosyncratic the relationships become, and therefore the lower the mean return. Secondly, despite the previously discussed criticism of the difference method of obtaining pairs, it appears to be the top performing pair formation method. Although, its performance does drop off relatively steeply as we begin to trade more and more pairs. Furthermore, comparing the performance between the correlation method based on daily and weekly data, it appears that weekly data does indeed offer some advantage over daily data. As mentioned previously, daily returns are highly random, and one ought not to expect similar stocks to necessarily behave similarly over a period of time as short as one trading day. Rather,

similar stocks tend to move alike one another over slightly longer periods of time. Taking weekly data also seems to be a good middle ground between taking a period of time longer than a day, while also providing enough data points during a formation period (52 weekly returns in a typical year) compared to monthly returns (12 data points in a year).

Moreover, the annualized return of weekly correlation exceeds the return of the daily pairs to an even greater extent. Although the average returns are very similar at 1.36% and 1.00% respectively, it happens to be the case that pairs formed on weekly return correlation have shorter trade windows on average, which makes for an annualized return which is over twice as large as that of the daily return pairs in some brackets. Another interaction which could be at play is the fact that pair trades with negative returns tend to have their positions open for a longer amount of time. This comes about as a direct consequence of the trading strategy - the only time we choose to close a position is when the stocks decrease in their divergence from one another (which results in a positive return by definition); conversely, negative return trades come about only when the trade period expires before the stocks revert back over a given threshold. This is evidenced by the fact that, on average, pairs which result in a negative return have around a 30-40% longer total time of positions being opened than pairs that result in a positive return. For instance, among the top 100 pairs ranked on their respective formation statistics, profitable pairs ranked on daily return correlations exhibit 28% higher position durations, while those ranked on SSD exhibit 43% higher position durations than unprofitable pairs across all 23 periods.



To get a more fine-grained understanding of the specifics of the relationship between portfolio size and overall strategy performance, I examine how the strategy's cumulative return varies with the number of pairs we decide to execute the trades on. Figures 9 to 12 plot the cumulative mean return of the trading strategy for each year on the y-axis, with the x-axis being the total number of top pairs included in the trades; the four figures represent the sets of pairs obtained by the four different formation methods. Importantly, the year signifies the start of the trading window for that period. For instance, the highest performing year was 2020, which was due to the profit gained during the first months of 2020 with pairs which were formed during the whole of 2019. As it is generally regarded that periods of high turbulence allow for high arbitrage profits (Do and Faff, 2010; Bowen et al. 2010), the pair trading strategy is no exception. It is further notable that by the end of the graph at the 100th pair, there are usually only about half of the years which end in a positive return.

The cumulative mean trajectories are about as we would expect - large returns at the beginning of the graph, and gradually converging to a mean as we add additional pairs. However, it is difficult to say whether the average marginal return to each additional pair increases or decreases with n - while there tends to be a downward trajectory to most of the lines in these figures, there are also periods which exhibit upward trajectories. For this reason, I plot the marginal return to each additional pair in Figures 13 to 16. The plotted lines most resemble a random walk, and do not appear to exhibit any clear trends or patterns. A potential explanation could be that, through the erosion of arbitrage profits over time, the performance of this strategy has become less robust, and more random. However, the speculation on the causes of such a trend and the evaluation of standalone performance is beyond the scope of the present paper, and we shall turn again towards examining the parameter variability.

Marginal Return of Trading an Additional Pair, as Function of Portfolio Size

Figure 13: SSD

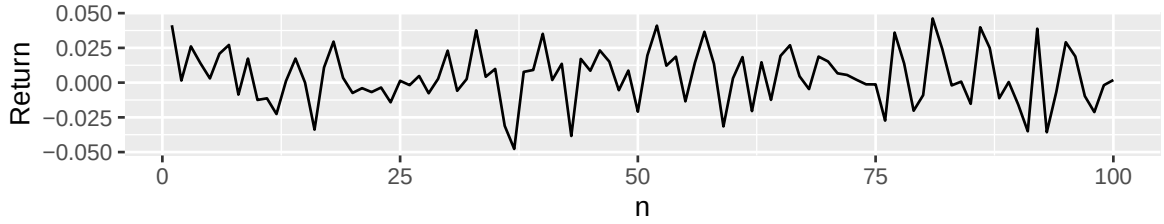


Figure 14: Cross Count

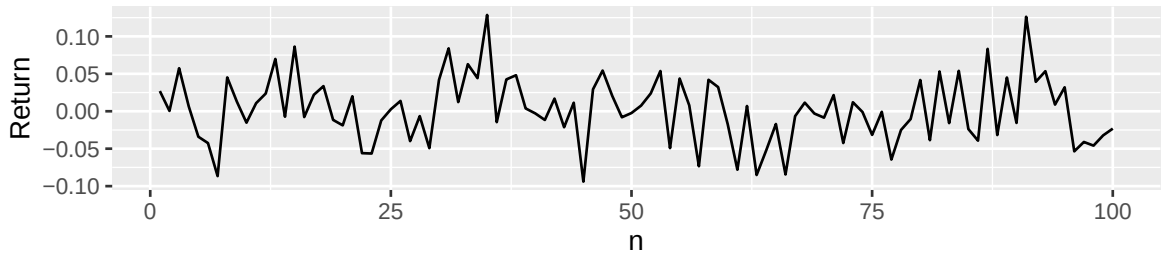
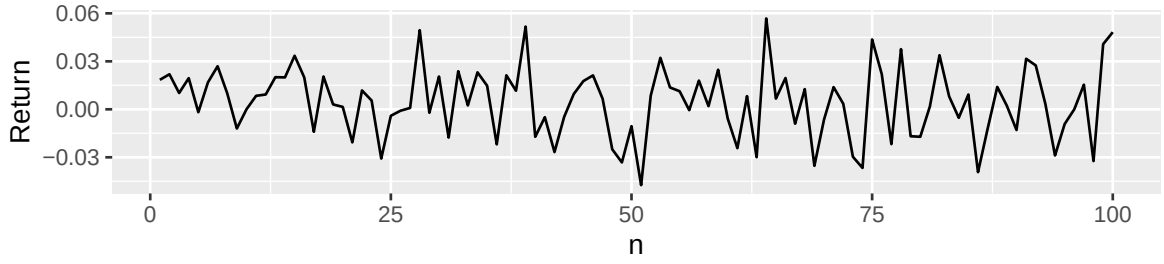
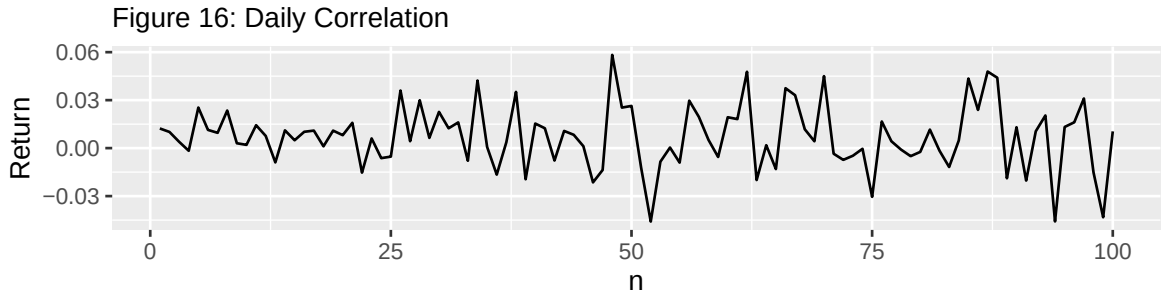
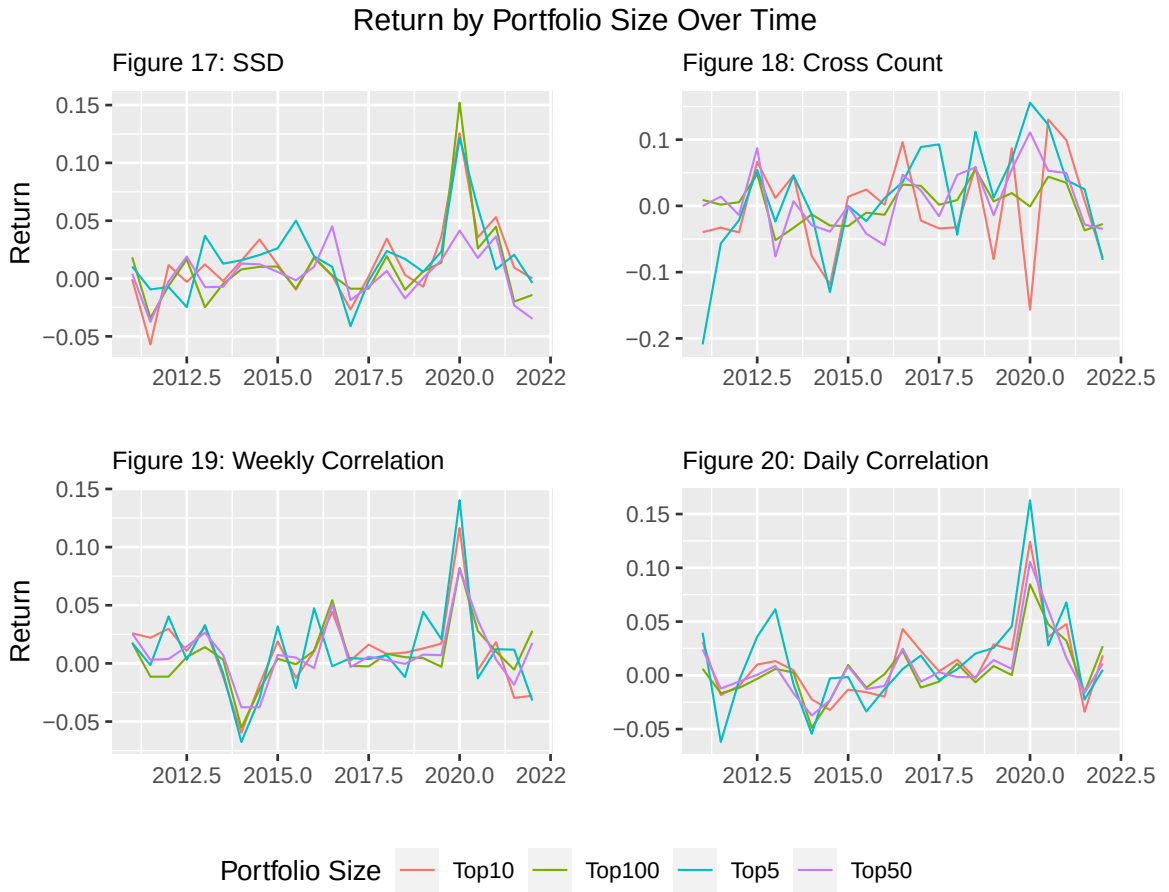


Figure 15: Weekly Correlation





What is potentially muddying the clarity which these figures could ideally provide, is the fact that we are taking an aggregate look, averaged over 23 periods, despite the fact that it is unlikely for all years to be alike in the number of tradeable true pairs. That is, since the strategy is based on capitalizing on a divergence which is random in nature, and that the specific price movements which allow the strategy to realize a positive return are in turn random themselves, we should therefore not expect that the distribution of profits across the top pairs to be the same in every year (for instance, in year t the optimum may be 50 pairs, while in year t+1 it may be 75 pairs), and these effects could average out over time.



Therefore, I calculate the values as shown in Table 1 (cumulative return based on portfolio size), but stratified based on trading period, and plot the resulting time series in Figures 17 to 20. First off, these figures clearly corroborate what has come to be widely accepted in the literature - that pair trading strategies perform best in times of market turbulence, with the convincingly best-performing period being during the pandemic-induced crisis of early 2020. Further, it is worth noting that the

year 2020 also exhibits the most well-defined performance order of portfolio sizes - that is, with only a few exceptions, the smallest portfolios tend to also be the best performing. This is in contrast with other years, where the order appears to be somewhat more random. The outlier out of the four figures appears to be that for cross count - while the other three figures exhibit similar structures, Figure 18 appears rather random in comparison. This could be an indication that cross count, as a formation method, is not as accurate as the other three in identifying what would traditionally be considered "true" pairs. Regardless, it still performs comparatively well, so it is difficult to rule it out completely.

The lack of a clear order of portfolio sizes with respect to their performance indicates that there is not a strong relationship between pair ranking and pair return. Rather, pairs even in the top 100 can potentially provide results comparable to the very top pairs. After all, even the largest considered category of 100 pairs is small compared to the total number of possible stock pairs. With an average 1545 stocks per period in the sample, there exist close to 1.2 million possible pairs of stocks, which makes the top 100th pair sit right above the 99.99th percentile.

With the observation in mind that there is time-variation in pair quality, perhaps there is a way of judging a pair by its formation statistic, rather than simply by its ranking compared to other pairs in the same period. The most straightforward of these would be to analyze whether there is perhaps a threshold, or a range of values, of the SSD or correlation coefficient, after which pairs tend to drop off in performance. This analysis will take place in the following section.

Evaluating Pair Formation Methods

As was noted by Krauss (2017), the method of identifying stock pairs used by GGR is not without its shortcomings. While simple, easy to check and implement, and intuitively clear, the sum of squared deviations has one distinct drawback. This is that, by its very nature, it "punishes" stocks which deviate significantly from each other. Mathematically, the ideal pair according to the sum of squared deviations criterion is one where the component stocks do not deviate from each other at all. Clearly, this would allow no profit to be made from cleverly trading such a pair of stocks. Moreover, the squared term in the formula disadvantages pairs of volatile securities. This is unfortunate for our specific application because the potential profit which can be made from trading one pair of stocks is directly proportional to the deviation which the stocks' prices make in relation to one another.

Some evidence to corroborate this hypothesis comes from observing the return distributions of traded pairs. Figure 21 plots the density distribution of the four employed pair formation methods overlaid over one another. Specifically, it plots the distribution of top 100 pairs identified in each formation-trading period, for a total of 2300 pairs of stocks over the considered time period. What we can clearly see is that the distribution of returns of pairs obtained by the sum of standard deviations (SSD) exhibits the lowest kurtosis among the four distributions. This provides evidence that this measure of pair likeness does indeed suffer from the side-effect of identifying low-volatility pairs which lack the capacity to be a highly profitable trade. Although this lack of potential upside is a shortcoming of the measure, the same appears to be the case on the negative side of the return distribution, meaning that it only results in lower variance of pair returns, while still potentially providing good returns on average. Therefore, it is possible that the low-variance pairs which tend to result from the difference approach are a feature, rather than a drawback of this strategy. Figure 22 plots the same relationship, but with annualized returns instead. For the most part, this figure is very similar to the previous one - with only SSD distribution being the outlier in terms of shape. However, the remaining three distributions appear to align more closely than in Figure 21.

Figure 21: Simple Return Distributions

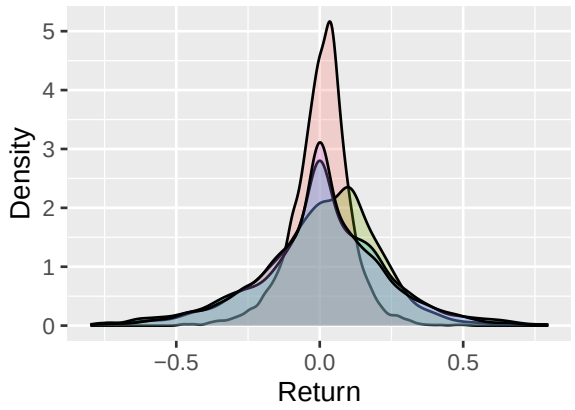
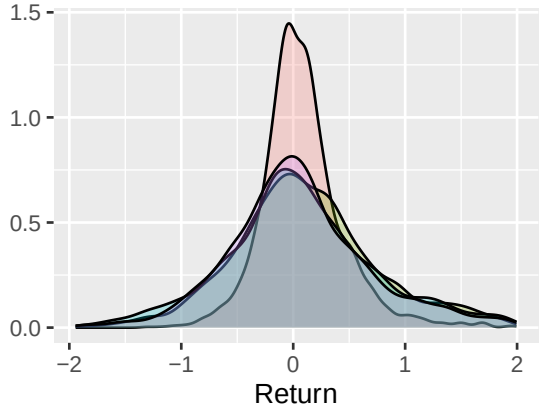


Figure 22: Annualized Return Distributions



Figures 23 to 26 plot the relationship between the pair formation measures and the return of all pairs in the top 100 of each period in the sample. Each point represents one pair, which corresponds to a combination of its formation statistic (i.e. SSD, cross count, or correlation) and the return of the pair. Furthermore, pairs where the stocks belong to the same industry are colored blue, while those where that is not the case are colored red. This is determined by whether the stocks in a pair share the same Standard Industrial Classification, better known as the SIC code.

Formation Statistics Against Return

Figure 23: SSD

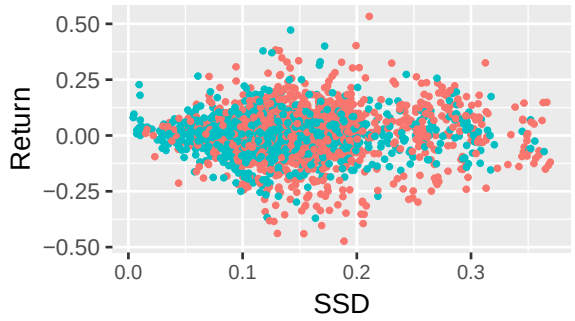


Figure 24: Cross Count

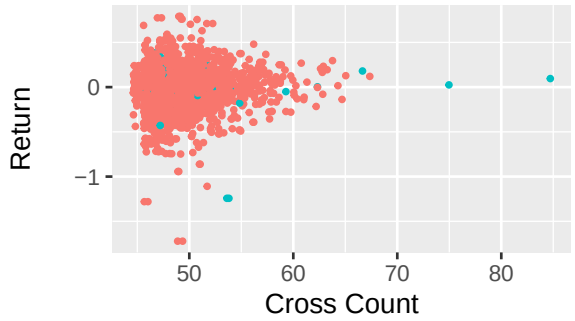


Figure 25: Weekly Correlation

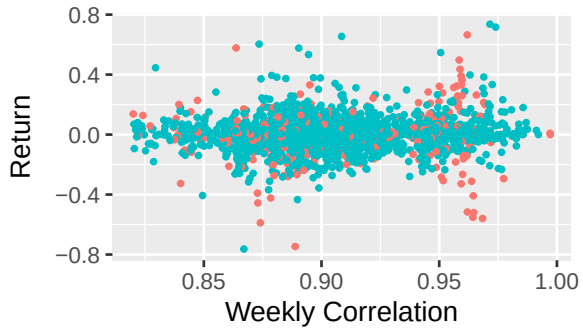
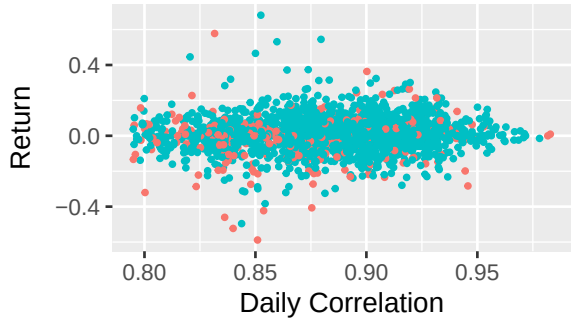


Figure 26: Daily Correlation



As mentioned previously, one of the (negative) features of the sum of squared deviations is that it tends to prefer pairs which shouldn't be very lucrative by virtue of their tendency to not diverge from one another. If this is indeed the case, we should see a positive relationship between SSD and pair return - since we would expect low-SSD pairs to provide a proportionally low return from the pair trading strategy. However, looking at Figure 23, there does not appear to be a clear relationship between the values, much less a positive one. For a more rigorous test, however, I a run simple univariate regression of pair return to its respective SSD, and the slope is not found to be statistically significant. Moreover, while GGR find that pairs which belong to the same industry result in higher returns than those which do not, the present analysis finds little evidence to support this. While the average return is indeed slightly higher (1.52% compared to 0.41 %), the difference is not found to be statistically significant (p-value = 0.21). The significance of the difference in means here and in the following discussion is computed using Welch's t-test, as these two groups do not have the same variance.

Figure 24 plots the same relationship, but for pairs formed on the cross count method. What is immediately striking is that this method appears to produce a far lower proportion of pairs which belong to the same industry. This could be an indication that the method tends to pick up more on random fluctuations than stocks which share fundamental attributes. That is, assuming that pairs which share industries tend to be similar in more fundamental ways than those which do not. Either way, the difference in returns between pairs of the same and different industries is not statistically significant, likely due to the small sample size of industry-sharing pairs (n = 66). Combined with the insights gathered from Figure 18, this method again appears to be the outlier among the four, seemingly identifying more spurious cases of stock likeness than the other ones, and missing those pairs which do possess a higher degree of similarity in their fundamentals.

Lastly, the two scatter plots in Figures 25 and 26 picture again the same relationships, but now with pairs formed on Pearson correlation, one using weekly return data, and the other using daily return data. Counter to the previous figure of the cross count pairs, the correlation method of pair formation appears to identify a far higher proportion of industry-sharing pairs. 83% of pairs across the whole sample identified by the weekly return correlation share industries with one another, while the same is true for 86.5% for those based on daily return correlations. Although I have previously hypothesized that weekly return correlations may do a better job of identifying true stock pairs, this observation does not corroborate this idea. Moreover, out of the four methods, pairs from daily return correlations are the only group which exhibits the return of same-industry pairs to be statistically significantly higher than those of pairs which do not share industries, with a p-value of 0.0059 (p-value for the same statistic of weekly return pairs is 0.39).

All in all, this section has shown little indication of patterns between the statistics which we use to measure the likeness of pairs, and the return of those same pairs, at least when restricted to the top 100 pairs per period. Perhaps this is too stringent of a restriction, and we ought to look at a larger number of top pairs to spot any meaningful patterns. Whether or not this is the case, we have seen little evidence to suggest that these measures can be used to aid us in determining desirable portfolio sizes for the pair trading strategy

Conclusion

Throughout the analysis of the present paper, I have taken a deeper look at the parameter-dependent variability of the pair trading strategy. While the trend in the literature is to mostly focus on one main aspect, which is how best to mathematically model pair-like price movements, and how best to statistically deduce which pairs of stocks behave like one another, I have kept the formation and strategy rules simple, and more deeply scrutinized the specific details. Although that aspect of the literature is certainly important, I have taken a different approach by analyzing, given simple formation methods, to what extent the profit of trading the resulting set of pairs changes by altering the trading rules which govern the mechanisms of opening and closing the positions in either leg of the trade, along with other elements of the strategy.

While the accepted trading rule in pair trading literature is entering the positions once the stocks diverge past two standard deviations of normalized-price deviation during the formation period, and exiting the positions once the prices cross one another, in the present paper I systematically iterate over a set of 1500 parameter sets to analyze the relationship between trigger rules and pair profit. What I find is that, although the standard trigger rules work reasonably well in some situations, they are not optimal in most cases, which is in line with the findings of Smith and Wu (2017). Although, one of the main shortcomings of the analysis is the lack of an adjustment for transaction cost, which could lead to a bias towards lower thresholds because the increased cost of higher frequency trading is not accounted for. While the common divergence threshold of 2σ is close to the in-sample optimum of around 1.6σ (depending on formation method), the reversion threshold of price crossing appears too stringent. In fact, the widely accepted reversion threshold of 0 is a far departure from the apparent optimum. Further, there does not appear to be a statistically driven motivation for the literature-adopted thresholds; rather, they are likely just rules of thumb which are based on a statistical tradition of sorts, where the measure of two standard deviations is often taken as a heuristic for a statistically significant deviation from the mean. Although intuitive and not incorrect per se, I conclude that the literature accepted the common rules without taking a more objective approach to the issue.

Aside from trigger rules for the trading itself, perhaps the second most important element of a successful practical implementation of pair trading is the number of top pairs one chooses to include in the trades. While this part of the strategy would in principle most likely include input from the trader, relying on human subjectivity steers the strategy more towards informed stock picking, and away from the ideal of a simple, quantifiable, and practical trading strategy which pair trading has earned a reputation for in the literature. In that regard, one ought to have a solid idea of how far down the list of top pairs to go in order to maximize volume and investable capital, while still making a profit from the trades. What I find is that, although profits tend to decrease as one includes more of the top pairs in the strategy, there is too much pair-to-pair and year-to-year variability to make conclusive statements about the optimal number of pairs one should trade per period.

One of the findings which corroborates previous literature is the fact that the sum of standard deviations measure does indeed tend to identify pairs with low volatility and, in turn, low earning potential. However, it is not as clear whether this is strictly a downside, because the negative end of the return spectrum is similarly constrained. What results is a portfolio of pairs which realizes a return of lower variance, but similar mean, compared to other methods. Moreover, I provide further evidence to solidify what is generally accepted in the literature, that pair trading is most lucrative in times of market turbulence. Specifically, considering the lack of recent publications, what the present paper contributes to the literature is application of the strategy during the pandemic-induced market crisis of early 2020.

Lastly, I introduce a so-far unrepresented formation method to the literature, cross count. Although it provides returns of comparable magnitude to the other methods, it exhibits some peculiar features which I point out in the analysis - for instance, the time-series of top pair brackets appears wholly unstructured and somewhat random when compared to all three other methods. It also does not appear to identify almost any pairs which belong to the same industry, which brings questions to how "true" the pair-like behaviour of its identified pairs really is at a more fundamental level.

The analysis has brought up some further questions which could be worthwhile considerations in future literature. Firstly, the variation in strategy performance across parameter sets is significant enough to warrant further analysis. Perhaps what could be the case is that we should not treat all stock pairs equally. Rather, we should identify key attributes of the component stocks in order to fine-tune the trading rules accordingly, along the lines of Papadakis & Wysocki (2007). Secondly, when deciding on portfolio size, it would be worth investigating how liquid the top pairs tend to be, and specifically how this would impact the scaling of this strategy in a real-world application. This is most important for institutional investors, as they are the most likely to invest amounts which would be problematic in case of small-cap or illiquid stocks. Therefore, in order to expand the strategy to a larger number of pairs, it would pay to have an idea of the scalability of this strategy. Lastly, a more

thorough analysis of the part played by the formation statistic, as well as other determinants of pair performance, should be considered. This relates in part to the aforementioned scalability issue - that is, having a solid ex-ante idea of what kinds of pairs we would or would not like to commit to trading could pay dividends to the overall performance of the pair trading strategy.

Bibliography

Bowen, D. *et al.* (2010) High-frequency equity pairs trading: Transaction costs, speed of execution, and patterns in returns. *The Journal of Trading*, 5(3): 31–38.

Chen, C. W. *et al.* (2014) Pairs trading via three-regime threshold autoregressive GARCH models. In: *Modeling dependence in econometrics: Selected papers of the seventh international conference of the thailand econometric society, faculty of economics, chiang mai university, thailand, january 8-10, 2014*. (s.l.): Springer. 127–140.

Chen, H. *et al.* (2019) Empirical investigation of an equity pairs trading strategy. *Management Science*, 65(1): 370–389.

Do, B. and Faff, R. (2010) Does simple pairs trading still work? *Financial Analysts Journal*, 66(4): 83–95.

Do, B. and Faff, R. (2012) Are pairs trading profits robust to trading costs? *Journal of Financial Research*, 35(2): 261–287.

Gatev, E. *et al.* (2006) Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies*, 19(3): 797–827.

Huang, C.-F. *et al.* (2015) An intelligent model for pairs trading using genetic algorithms. *Computational Intelligence and Neuroscience*, 2015 16–16.

Huck, N. (2013) The high sensitivity of pairs trading returns. *Applied Economics Letters*, 20(14): 1301–1304.

Kim, K. (2011) Performance analysis of pairs trading strategy utilizing high frequency data with an application to KOSPI 100 equities. *Available at SSRN 1913707*,

Krauss, C. (2017) Statistical arbitrage pairs trading strategies: Review and outlook. *Journal of Economic Surveys*, 31(2): 513–545.

Papadakis, G. and Wysocki, P. (2007) Pairs trading and accounting information. *Boston university and mit working paper*,

Smith, R. T. and Xu, X. (2017) A good pair: Alternative pairs-trading strategies. *Financial Markets and Portfolio Management*, 31 1–26.