Pairs Trading with Commodity Futures: Evidence from the Chinese market*

Yurun Yang [†], Ahmet Göncü [‡], Athanasios A. Pantelous [§]

August 22, 2016

Abstract

In this study, the profitability of different pairs selection and spread trading methods are compared using the complete dataset of commodity futures from Dalian Commodity Exchange (DCE), Shanghai Futures Exchange (SHFE) and Zhengzhou Commodity Exchange (CZCE). Pairs trading methods that are already known in the literature are compared in terms of the risk-adjusted returns via *in-sample* and *out-of-sample* backtesting and bootstrapping for robustness. The empirical results show that pairs trading in the Chinese commodity futures market offers high returns, whereas, the profitability of these strategies primarily depends on the identification of suitable pairs. The observed high returns are a compensation for the spread divergence risk during the potentially longer holding periods, which implies that the maximum drawdown is more crucial compared to other risk-adjusted return measures such as the Sharpe ratio. Complementary to the existing literature, for our market, it is shown that if shorter maximum holding periods are introduced for the spread positions, then the pairs trading profits decrease. Therefore, the

*The authors would like to acknowledge also the gracious supports of this work by the EPSRC and ESRC Centre for Doctoral Training on Quantification and Management of Risk & Uncertainty in Complex Systems & Environments (EP/L015927/1) and the Research Institute of Quantitative Finance at Xian Jiaotong–Liverpool University. We would like to thank participants at the 1^{st} Annual China Derivatives Markets Conference (CDMC), the 2^{nd} Workshop on Quantitative Finance organized by the Research Institute of Quantitative Finance, and seminar talks at the University of Liverpool and Xian Jiaotong–Liverpool University, as well as Dr. Hongsong Chou and Dr. Honghai Zhu for helpful comments. Remaining errors are ours.

[†]Department of Mathematical Sciences, Xian Jiaotong Liverpool University, China

[‡]Research Institute of Quantitative Finance, Xian Jiaotong Liverpool University, Suzhou, China, Corresponding Author Email: Ahmet.Goncu@xjtlu.edu.cn

[§]Department of Mathematical Sciences and Institute for Risk and Uncertainty, University of Liverpool, United Kingdom.

returns do *not* necessarily imply market inefficiency when the higher maximum drawdown associated with the holding period of the spread position is taken into account.

JEL classification: C12; G13; G14; C52

Keywords: Commodity Futures; Pairs Trading; Statistical Arbitrage; Market Efficiency

1 Introduction

Commodity futures markets are increasingly in the focus of investors, in particular of the *hedge funds*, and China is no exception in this trend.[1] With its rapidly growing economy, China has some of the world's most highly traded commodity futures including the contracts on *copper, gold, iron ore, palm oil, white sugar*, which are some of the commodities that are considered in this paper.[2]

In the aftermath of the Chinese stock market slump in the summer of 2015, the volume of commodity futures trading spiked due to the following three main reasons: first, the short selling restrictions in the Chinese stock market led the hedge funds to consider commodity futures as a main alternative, secondly the falling of commodity prices globally and the slowing down of the Chinese economy led the speculative investors to take short positions in the commodity futures betting that the slowdown of the Chinese economy might continue coupled with the slow global growth, and finally, since commodity futures prices have low correlation with the stock markets, while the Chinese stock market slumped, they offered important diversification benefits for the investors.

The increasing presence of hedge funds and investment banks in the commodity futures markets has the effect of increased use of quantitative trading strategies to generate statistical arbitrage profits. In the present paper, the performance of

[1]See Financial Times article reported by Yang Yuan (Zhengzhou), Christian Shepherd, Wan Li and Lucy Hornby (Beijing) and written by Lucy Hornby and Neil Hume entitled: "*Chinese retail investors throw global commodities into a tailspin*" (published, May 6, 2016 5:44 pm)

[2]See 2015 WFE/IOMA Derivatives Market Survey reported by World Federation of Exchanges (WFE) and IOMA,"*the commodity options and futures traded in Shanghai and Dalian accounting for 50% of the volume traded in 2015 in terms of number of contracts*" (published, April 2016)

pairs trading is analysed and by considering its maximum drawdown with different maximum holding periods for the spread position, our main contribution is to give evidence that at the shorter maximum holding periods, the profitability of pairs trading decreases in the Chinese commodity futures markets. The maximum drawdown, which is widely used in the hedge funds, is a function of the duration of the spread position. What is more, the relationship between the profitability and the maximum holding period for spreads appears to be robust both in time and across the different pairs. The intuition behind this conclusion is that if one does not have stop-loss barriers and can hold the spread positions for longer periods of time, then a higher premium can be received from pairs trading, which of course comes with the risk of a larger potential drawdown during this waiting time. Furthermore, using the complete dataset of Chinese commodity futures prices from 2005 to 2016, the profitability of the main pairs trading models used in the literature is verified and compared.

The plan of the paper is the following. In Section 2, a brief review on pairs trading and statistical arbitrage is presented. In Section 3, the dataset utilized in this study is discussed, whereas the spread model is introduced together with the identification methods of potential pairs. Empirical results are presented and analysed in Section 4, and finally Section 5 concludes the whole discussion.

2 Pairs Trading and Statistical Arbitrage

The intuition behind *statistical arbitrage* is related to the spread of expected returns of large portfolios and asset classes. Typically, statistical arbitrage strategies go *long* in the portfolio of assets with the *highest* expected return and *short* of those with the *lowest* expected return. Beating the market returns via those strategies have been in the focus of asset management and hedge fund industry since the use of sophisticated statistical methods to develop high-tech pairs trading programs at Morgan Stanley in mid-1980s by the team of the Wall Street quant Nunzio Tartaglia. Since then, *pairs trading* has become increasingly popular and it has been widely applied across markets and different asset classes. In contrast to the popularity of statistical arbitrage strategies in the financial industry, academic literature has been slow to lay the theoretical foundations, and in particular, its definition.

As a result of this long lasting gap in the financial literature, Bondarenko (2003) introduces the concept of a *statistical arbitrage opportunity* in a finite time economy. However, his definition assumes technical requirements on the pricing kernels, whereas Hogan et al. (2004) define the statistical arbitrage in a general probability space and in an infinite time horizon economy instead. Under this new framework, Hogan et al. (2004) test the efficient market hypothesis. Their test methodology is slightly extended by Jarrow et al. (2012) by avoiding to penalize incremental profits with positive deviations. Among them, up-to-date, Hogan et al. (2004) remains as the most popular mathematical definition for a statistical arbitrage opportunity that considers the asymptotic behaviour of a trading strategy.

To be more detailed, *pairs trading* is a market-neutral strategy to elaborate statistical arbitrage profits. Actually, it exploits the temporary deviations of a pair of asset prices from their long-term equilibrium level. However, in practice, even if perfect information exists on the model parameters of the spread, there is the risk that mean reversion to the long-term equilibrium level might take *too* long or *never* happen. The main risk in pairs trading is the fact that the spread might diverge relative to its long-term mean level after a position is opened (Gatev et al., 2006). For example, in the equity markets there are important limitations in terms of the possible duration of short positions. If the spread between the pairs has widened the trader might be forced to liquidate the spread position and realize a loss.

There are two steps in implementing successfully a pairs trading strategy. The first step is the identification of suitable pairs and the second one is to determine the appropriate strategy deciding which asset takes the long-short position and when. In this paper, regarding the first step, the *profitability index* (based on cointegration), the *minimum distance*, and the *correlation* methods are considered. Initially, Alexander (1999) and Alexander and Dimitriu (2005a,b) propose the framework of cointegration for identifying optimal pairs for trading. Later, Zeng and Lee (2014) propose a profitability index based on the cointegration relationship for the spread of two correlated assets. Since the minimum distance approach is mostly used by practitioners, Gatev et al. (2006) and Perlin (2009) construct the pairs portfolio by using it and they calculate the sum of squared deviations of two price time series, whereas Huck and Afawubo (2015) explore the performance of pairs trading strategy based on several pairs selection criteria, using the components of the S&P 500 index. Although, it is difficult to come up with a universally superior selection method, our numerical experiments show that a better approach is to *rank* suitable pairs by applying these methods simultaneously.

The *existence* of statistical arbitrage opportunities with respect to Hogan et al. (2004)'s definition for the Black-Scholes and mean-reverting stochastic spread models is provided by Göncü (2015) and Göncü and Akyıldırım (2016), respectively. Amongst others, profitability of pairs trading strategies is shown in Gatev et al. (2006), Baronyan et al. (2010), Cummins and Bucca (2012), Do and Faff (2010, 2012), Bowen and Hutchinson (2014), Jacobs and Weber (2015) and Focardi et al. (2016). Particularity, Gatev et al. (2006) provide a detailed analysis of the performance of pairs trading strategies and find evidence of profit. Do and Faff (2010) demonstrate that pairs trading performed strongly during periods of economic depression, including the 2008-2009 financial crisis. Baronyan et al. (2010) investigate a number of market-neutral trading strategies and display that pairs trading performs strongly. For the US equity market over the period 1963-2009, Do and Faff (2012) confirm that pairs trading remains profitable after controlling for commissions, market impact and short selling fees. Bowen and Hutchinson (2014) provide a comprehensive UK evidence on the profitability of pairs trading strategies. Jacobs and Weber (2015) analyse the U.S. as well as 34 international stock markets, and they reveal that trading pairs solely constructed from information about past prices turns out to be persistently profitable. Recently, Focardi et al. (2016) provide empirical evidence from S&P 500 from 1989-2011 that profitability can also be constructed by using their dynamic factor models of prices.

As Gatev et al. (2006) mention, the risk-adjusted returns seems to be a compensation for a latent or dormant risk factor that has not been considered. The vast majority of pairs trading studies focus on equity markets, where the risk-adjustment of returns are done with widely used factor models. However, in the present paper, the *commodity futures* market is considered, where the dynamics are different, and furthermore, most of the factor models that have already been used in the equity markets in trying to explain the pairs trading returns are not feasible here (Focardi et al., 2016). Therefore, the Chinese future market offers a promising testing environment for the profitability of pairs trading strategies.

Our objective in this paper is to give evidence that the profitability of pairs trading in the commodity futures markets does *not* necessarily imply a market *anomaly* or *inefficiency*. The rationale behind our argument is two-folds: first, pairs trading requires short positions in the relatively over-priced asset, which is often assumed to be arbitrarily long in the literature. Actually, it is ignored the fact that high maximum drawdown might lead to the liquidation of spread position with big losses, and thus, proper risk adjustment can only be done via the average or maximum drawdown calculated during the holding period of each spread position;[3] Second, by considering different maximum holding periods for our spread positions, it can be shown that the profitability of pairs trading might change drastically. Risk-adjusted returns decrease significantly as shorter maximum holding periods are enforced for each spread position. This is an early indication that the market participants might associate the duration of spread positions with higher drawdown risk.

Essentially, the literature on pairs trading in equity markets attempts to find risk factors to explain the profitability instead of calculating the simple maximum drawdown in different trade durations. In our case, the maximum drawdown is documented in order to help us to explain why some profit opportunities might not be practically attainable in the existence of risk-controls, such as stop-loss barriers which are used extensively in the hedge fund industry. Therefore, the market inefficiency appears not to be necessarily the case even with the existence of consistently high returns and Sharpe ratios from pairs trading. In that direction, 3 different pairs selection criteria are considered, namely, empirical distance, prof-

[3] Based on various discussions we had with senior managers on hedge funds operating in the commodity futures markets in China, it is clear that the industry considers the average or maximum drawdown measures as the norm. Furthermore, longer duration to hold the spread position and high maximum drawdown imply that the spread position will be liquidated before the pairs trading profit is realized due to stop-loss barriers before the pairs that are commonly used to control drawdown risk.

itability index and correlation methods utilizing the dataset of Chinese commodity futures traded in Shanghai, Dalian, and Zhengzhou Commodity Exchanges. After identifying the best available pairs, a pairs trading strategy with different spread thresholds are considered given by Gatev et al. (2006), Zeng and Lee (2014), Göncü and Akyıldırım (2016), and the standard deviation rules with and without the Kalman-filtering technique given by Elliot et al. (2005).

3 Data and Models

Let us now discuss the data and the models used to compare the profitability of different pairs selection and spread trading using the Chinese commodity futures market.

3.1 Data

Our dataset includes the *whole* history of the daily close prices of commodity futures traded in *Dalian Commodity Exchange* (DCE), *Shanghai Futures Exchange* (SHFE), and *Zhengzhou Commodity Exchange* (CZCE), and covers the period from January 1^{st}, 2005 to June 1^{st}, 2016.[4] In our empirical analysis, the commodity futures with more than 1,000 observations are considered only, which gives us 9 different commodities in DCE, 9 in SHFE and 7 in CZCE. Table 1 summarises the names and maturity dates of the futures contracts in these 3 Exchanges. Since, the trading of the 25 commodity futures did not launch simultaneously, the data lengths are not the same for all the contracts. For each commodity there are 6 different maturities all with 1 year tenor, which results in 150 underlying contracts for pairs trading.

3.2 Models

The trading is implemented by identifying the best potential pairs, utilizing the training sample, and then trading these pairs with 1-year length of out-of-sample periods. The choice of 1 year for trading intervals is consistent with the tenor of

[4]The data is obtained from WIND Information Co. Ltd. financial terminal.

futures contracts in our dataset, and in this way the problem of rolling over the next contract is avoided since the spread position is already closed at least 5 days prior to the maturity day of each futures contract.

3.2.1 Modelling the Spread Process

In this paper, the difference in *log-prices* is considered for modelling the stochastic spread process between the two assets. This has important advantages, since: (i) it leads to stochastic models that are consistent with the use of geometric Brownian motion process for the asset prices; (ii) it allows the model to focus on the returns and avoids the scaling issues; (iii) empirically, it helps to detect more easily the mean reversion property. Therefore, following Elliot et al. (2005), Avellanda and Lee (2010), Zeng and Lee (2014) and Göncü and Akyıldırım (2016), the *spread* of two *assets* is defined by the cointegration equation

$$\ln(S_t^i) = \alpha + \gamma \ln(S_t^j) + \epsilon_t, \quad \epsilon_t \text{ i.i.d. } \sim (0, \sigma_\epsilon^2), \tag{1}$$

where the estimate of $\hat{\gamma}$ is used to construct the spread between assets i and j, and it is given by

$$X_t = \ln(S_t^i) - \hat{\gamma} \ln(S_t^j). \tag{2}$$

Additionally, the *dynamics* of the *spread* is often assumed to follow a mean reverting Ornstein-Uhlenbeck (OU) process (Elliot et al., 2005; Avellanda and Lee, 2010; Bertram, 2010; Bogomolov, 2013; Zeng and Lee, 2014; Göncü and Akyıldırım, 2016), and it is given by

$$dX_t = -\rho(X_t - \mu)dt + \sigma dW_t, \tag{3}$$

where ρ is the speed of mean reversion, W_t is a standard Brownian motion (on some probability space), and μ is the long term equilibrium level of the spread. The solution of Eq. 3 is provided by

$$X_t = X_0 e^{-\rho t} + \mu(1 - e^{-\rho t}) + \sigma \int_0^t e^{-\rho(t-s)} dW_s, \tag{4}$$

where X_t is normally distributed with $E[X_t] = X_0 e^{-\rho t} + \mu(1 - e^{-\rho t})$ and $V(X_t) = \frac{\sigma^2}{2\rho}(1 - e^{-2\rho t})$, respectively. The stationary mean and variance are given as μ and $\sigma^2/2\rho$ as $t \to \infty$, respectively.

It is known that the parameters of the OU process can be estimated by an AR(1) representation for the discretization of the Eq. 3, where the discretization error of the stochastic differential equation is $O(\Delta t)$. Let us denote the mean subtracted process $\tilde{X}_t := X_t - \mu$, and thus, the AR(1) process is written by

$$\tilde{X}_{t+1} = (1 - \rho\Delta t)\tilde{X}_t + \epsilon_t, \quad \epsilon_t \text{ i.i.d. } \sim (0, \sigma_\epsilon^2 \Delta t), \quad \rho > 0. \tag{5}$$

Our dataset for the spread of futures contracts consists of daily observations, thus by implementing Eq. 5, the daily parameter values of the model are obtained. Additionally, the partial autocorrelation function for the spreads can be verified, as the AR(1) effect is reported. For instance, in Fig. 1, the plots of the partial autocorrelation function for the spread of WR-RB and PB-FU[5] pairs with different maturities are provided.[6] See also the plot of the spreads in Fig. 2 for the above mentioned pairs as they are derived by Eq. 2. From this small sample of pairs that is demonstrated in Fig. 1 and 2, it is clear that the spreads have historically strong mean reversion around the long-term mean. Therefore, they can be considered as potentially good candidates for the pairs trading.

3.2.2 Formation of Pairs

Our universe is consisted of 25 different kinds of commodities with more than 1,000 observations from the 3 commodity futures Exchanges in the Chinese market. The challenge here is that each Exchange specializes in different types of commodities with diverse rules and non-synchronized maturity dates (see Table 2). In this paper, we only form pairs between contracts from the same Exchange, but there is also flexibility to create portfolios with pairs that are contained in different Exchanges. Perhaps, our approach might be considered somehow as conservative, since the po-

[5]WR - RB stands for Wire Material and Screw Steel; PB - FU stands for Lead and Fuel Oil, see Table 1.

[6]The maturity for WR-RB pair is on January, May and September, and PB-FU pair is on March, July, November

tential spread formation space is restricted within pairs from the same Exchanges, however by relaxing this assumption the potential profitability increases further.[7] In the backtesting analysis, each candidate pair is ranked with respect to the values of the selection criteria by utilizing the longest available common history going backwards. Although, the best pairs rankings take into account the whole historical dataset, out-of-sample backtesting does not use any future information in the formation of pairs.

As selection criteria, we consider: (i) the correlation coefficient between log-prices, (ii) a profitability index calculated in terms of the volatility and mean reversion of the spread, and (iii) the sum of squared deviations of two futures price series. Additional to the simple historical correlation coefficient, the *sum of squared differences* (SSD) is calculated, which is used to pick potential pairs of assets as given by Gatev et al. (2006), Huck (2013), and Huck and Afawubo (2015). In order to rank the different pairs, the average SSD is calculated with respect to the normalized prices given by

$$SSD_{i,j} = \frac{\sum_{t=1}^{T} (S_t^i/S_0^i - S_t^j/S_0^j)^2}{T}, \tag{6}$$

with S_t^i and S_t^j are the prices of the two assets. Finally, the profitability index measure is calculated, see Zeng and Lee (2014), which is given in terms of the volatility σ and the mean-reversion parameter ρ by

$$\text{Profitability index } = \sigma\sqrt{\rho/2}. \tag{7}$$

The profitability of a spread is improved if the spread of prices show high volatility and mean reversion at the same time.

3.2.3 Maturity Months and Trading Volume

In the hedge fund industry, intraday trading of commodity futures is often based on the contract with the highest volume. Thus, if the investor is trading a particular

[7] Indeed, mathematically speaking, by expanding the set of possible pairs, it is more likely to have pairs that are more profitable than those that have already been considered in our more conservative approach.

pair spread, the liquidity risk is minimized by utilizing the most actively traded maturity month for the same pair. Therefore, in our case, since lower frequency trading exists, extra cost for rolling the futures over different contracts and maturity mismatches is generated as the spread is traded over longer time horizons.

Therefore, 1-year fixed maturity contracts are preferred, and thus the spread position can be held until the end of the maturity without rolling over a different maturity contract. Furthermore, in order to avoid data selection biases, we also present the results obtained for all the 6 different maturity months traded in all the Exchanges. The trading volume of these contracts fluctuates depending on the underlying commodity. For example, Soybean futures with maturities in January, May and September are more popular than those with maturities in March, July and November. Normally, the trading volume goes down prior to the last trading day. To avoid liquidity and physical delivery issues, it should be enforced that the spread position is closed at least 5 trading days before the maturity date.

3.2.4 Data snooping

In order to minimize potential data snooping biases, we mainly utilize out-of-sample backtesting, and form the pairs only with the used training sample. Furthermore, out-of-sample backtesting is done with different model training and trading periods, and focus on the model free approach of two-standard deviation (2-stdev) rule without modifying our trading rules (Gatev et al., 2006), which is also commonly used in the hedge fund industry. In this trading rule, whenever the spread process deviates by two historical standard deviations away from its long-term mean level, the trader takes a short or long position in the spread portfolio, and he closes the open position. Huck (2009), Do and Faff (2012), Huck (2013) and Huck and Afawubo (2015) apply and compare the performance of empirical standard deviation (sigma) rules.

3.2.5 Thresholds for Pairs Trading

For completeness and robustness, additional to the 2-stdev rule which is widely used by the practitioner (Gatev et al., 2006), alternative pairs selection and trading models are considered and all the results are presented. Thus, two known pairs trading thresholds (triggers) existing in the literature are implemented with and without the Kalman filtering technique of Elliot et al. (2005). Next, the main features of each trading method are discussed.

Elliot, Van Der Hoek, Malcolm's Kalman-filter (KF) method: Elliot et al. (2005) consider a mean-reverting Gaussian Markov chain model for the spread which is observed in Gaussian noise. Therefore, the Kalman Filter technique is used to filter the noise and tried to obtain better estimates of the true mean-reverting OU process. Elliot et al. (2005) employ two EM-Algorithms for implementing the Kalman filter, which are given by Shumway and Stoffer (1982) and Elliot and Krishnamurthy (1999), respectively. In our empirical examples the Shumway and Stoffer (1982) EM-algorithm is implemented.

Zeng and Lee (ZL)-method: Zeng and Lee (2014) propose the optimal trading thresholds as functions of the parameters of the OU process, and the transaction cost. They derive a polynomial expression for the expectation of the first-passage time of the OU process with two-side boundary, and obtain the analytic formula for optimal trading thresholds as the solution. In Zeng and Lee (2014) model, the trader needs to choose the best threshold level that maximizes the expected profit per unit time in order to initiate the pairs trading. If the trading threshold has narrow bands around the mean level, then the time it takes to return to the long-term mean is small, and thus is the profit per trade. On the other hand, if the threshold is far away from the long-term mean, the profit in each trade is larger, and thus on average it takes longer to realize it. Since the transaction cost should be greater than zero in real trading, we only consider the second case.

Göncü and Akyıldırım (2016) (GA) method: Finally, Göncü and Akyıldırım (2016) derive an optimal threshold level that maximizes the probability of successful termination (mean-reversion probability) of the spread portfolio for 1-year investment horizon.

Note that there is an important difference between the trading rules of the 2-

stdev, Zeng and Lee (2014), and Göncü and Akyıldırım (2016) methods. In Zeng and Lee (2014) the trade cycles are longer since a trader that opens a spread position at the upper (lower) threshold closes it at the lower (upper) threshold instead of at the long-term mean level. In the 2-stdev and Göncü and Akyıldırım (2016) methods, the trader closes the position whenever the long-term mean level is reached (which can be considered as the classical way of implementing pairs trading).

4 Empirical Results

4.1 Preliminary Analysis

In this section the profitability of different pairs selection and spread trading methods are compared using the complete dataset for 25 commodities from the 3 Chinese commodity futures Exchanges. Both in-sample and out-of-sample backtesting are applied to verify the performance of the pairs trading methods. However, in order to avoid *physical delivery* or *low liquidity* problem towards the maturity of the future contracts, the liquidation of the spread position is enforced at least 5 trading days before the maturity date. In that way the problem of rolling over different futures contracts is avoided.

To test the robustness of our strategies to parameter changes, subperiods are considered that are formed by selecting different *out-of-sample* trading periods and the pairs selection is decided at the beginning of each subperiod. Thus, the Subperiod 1 contains the last 1 year as the trading period utilized, and all the previous prices as the training period, whereas the subperiod 2 starts the out-of-sample trading period by being shifted backwardly by 75 trading days, and thus, the interval for training period contains 75 less observations, and so on and so forth. By shifting the out-of-sample starting dates, 9 different annual out-of-sample trading periods are obtained. What is more, because the expanding window is used for the training period, the size of the training period is largest for the subperiod 1. The rationale behind the choice of 75 trading days is related to the fact that a larger number leads to smaller number of subperiods since some pairs in the Chinese futures market do not have long history of trade and a smaller number increases the overlap between different out-of-sample intervals which might cause higher autocorrelation between the out-of-sample returns obtained.

During the implementation of pairs trading the model parameters are updated by utilizing an expanding window of daily observations. Every time that a spread position is closed for any given pair in the portfolio, the parameters of that spread are re-estimated with the available, up-to-date information. For the subperiods, as they have been designed and we present them above, it is implied that the subperiod 1, which has the longest training period of forming pairs, utilizes a larger dataset both for the formation of pairs as well as the estimation of the model parameters.

For the formulation of portfolios of pairs it is assumed that each pair has a committed capital proportional to the weight of that pair within the portfolio of pairs. Then, when the portfolios of pairs are formulated, two alternative return calculations can be implemented. Thus, as it is also considered by Gatev et al. (2006), the committed capital or the fully invested return is calculated. However, in our case, the most conservative approach is followed and the committed capital return is used. Thus, it is assumed that even if a pair is not traded for the whole period of trading time, an equal amount of capital committed to this pair is still allocated, and the equally weighted average is calculated with the number of pairs. In this way the opportunity cost of the investor to commit capital to each pair in the portfolio is taken into account even though all the pairs are not traded.

Table 2 presents the calculated returns of the pairs trading for the 9 out-of-sample trading subperiods with different starting dates using the 2-stdev trading rule as a benchmark. Panel A reports the annual returns obtained from the pairs trading when the spread positions are opened at the end of the day when the prices diverge and closed at the end of the day when the price converge. It is observed that in 5 out of 9 subperiods, the returns are very high, which shows that potentially pairs trading might be very profitable. Lower returns are observed in subperiods 7, 8 and 9 which is likely due to the smaller size of the training sample as an expanding window for the subperiod analysis is used. Alternatively, in Fig. 3, the growth of 1 money-unit allocated to the pairs trading with a portfolio of 1, 3 and 6 best pairs in

2013–2016 trading period is plotted. Different from the analysis in Table 2, in Fig. 3, it is assumed that the 1 money-unit position is traded annually over the 3-years period, and the best pairs are re-identified at the end of each annual trading cycle.

4.2 Transaction Costs

Due to the contrarian nature of pairs trading strategies, the returns might be biased upward because of the bid-ask bounce (Conrad and Kaul, 1989; Jegadeesh, 1990; Jegadeesh and Titman, 1995). In particular, we sell assets that have done relatively well and buy the ones that have done poorly. As discussed in Gatev et al. (2006), the winner's price is more likely to be an ask quote and loser's price is a bid quote, whereas the opposite is true at the second crossing when the spread converges to its long-term mean level. To address this issue, following closely Gatev et al. (2006), in Table 2 Panel B, pairs trading is presented on the day following the divergence and liquidate on the day following the crossing. As it is observed in Table 2 Panel C, the average return can drop by $24.03 - 20.75 = 3.28\%$ with 4 average number of pairs trading cycles per year. Therefore, as a rather conservative estimate, we assume 2% round trip transaction cost per pairs trade, which is expected to be sufficient to cover market frictions, liquidity costs, commission fees, etc.[8]

4.3 Testing alternative pairs trading models

In Table 2, it is reported that the pairs trading generates high returns even after transaction cost is removed. The question as to whether an alternative pairs selection or trading models can change the results significantly has not received much attention in the literature. Actually, a comparison among the alternative theoretical approaches proposed by Zeng and Lee (2014) and Göncü and Akyıldırım (2016) as well as with the practical, 2-stdev, approach used by Gatev et al. (2006) (and practitioners) with and without the Kalman filtering technique of Elliot et al. (2005) has not been implemented so far.[9] The argument is the following. Consider

[8]Note that the difference in average returns is 3.28% with 4 average number of pairs. Roughly, for each trade the estimated cost is given by $3.28/4 = 0.82\%$, i.e. 1.64% for round trip transaction cost. With a conservative approach the transaction cost is rounded up as 2%.

[9]For instance, ZL-EL and ZL are the Zeng and Lee (2014) approach with or without the Kalman filtering technique of Elliot et al. (2005) which is based on the EM-algorithm (Shumway and Stoffer,

the backtesting results derived by the 3 alternative approaches with and without the Kalman filtering technique of Elliot et al. (2005), how robust the pairs trading methods are under the different pairs selection and trading models.

In-sample backtesting

For the in-sample comparison, the best pairs are identified with respect to (i)–(iii) criteria presented in Section 3.2.2 using the whole dataset. The cointegration equation for the spread is also estimated from the whole sample until the end of the trading period. Using the last 252 observations for the formulation of the trading period, the annualized returns (after subtracting the transaction costs) are calculated for each pairs trading that is derived with the different methods.

The results of the in-sample backtesting are reported for the trading periods from May 22^{nd}, 2015 to June 1^{st}, 2016 and May 13^{th}, 2014 to May 22^{nd}, 2015 in Tables 3 and 4, respectively. Thus, the following observations are made which are consistent for both Tables 3 and 4. The profitability index and the combination methods identify pairs that are more likely to produce high returns at least for the GA and 2-stdev rules. Furthermore, Kalman-filtering fails to improve the obtained returns. The reason why it fails is that the commodity futures prices exhibit large fluctuations, where the Kalman-filter tends to treat these fluctuations as the noise term, which in turn cause lower generated thresholds and potentially lowering the profitability. What it more, astonishingly, 2-stdev rule, which is the simplest one, is applied by practitioners and is free from model specification errors, seems to do quite well comparing with the all the other theoretical methods with or without Kalman-filtering.

Out-of-sample backtesting

The in-sample backtesting gave us the idea about the relative performance of different pairs selection and trading models. Now, the out-of-sample backtesting will show us whether or not these methods can work in reality, when the pairs and model parameters are obtained without any future information. For the out-of-

1982).

sample backtesting, first the dataset is split into two parts named as the *training* and *out-of-sample backtesting* periods.

Using the training period the pairs are identified and the initial estimates of the model parameters are obtained. An expanding window of observations is used to estimate the parameters of the models (Elliot et al., 2005; Zeng and Lee, 2014; Göncü and Akyıldırım, 2016). The model free 2-stdev method (Gatev et al., 2006) simply utilizes the sample means and standard deviations obtained from the expanding window of observations.

Using the least squares regression, the spread parameter $\hat{\gamma}$ is estimated and updated regularly every time a spread position is closed. Backtesting is performed from the last data point in the training sub-sample and update the thresholds to include new information in every trading day repeatedly. Moreover, by assuming that during the 1 year trading period the spread position is closed n-times, the annual return is calculated as: $(1+r_1)\cdots(1+r_n)$, where each return from the pairs trading r_i is subtracted by the transaction cost. At the end of the out-of-sample period, we enforce to close the position and realize the profit or loss at the last trading day. In both in-sample and out-of-sample experiments, the GA method is implemented with the investment horizon of $T = 1$ year, which is consistent with the maturity of futures contracts. When the Elliot et al. (2005) method is implemented, the trading threshold is $\mu \pm \sigma/\sqrt{2\rho}$.

The out-of-sample backtesting results are provided in Tables 5 and 6. On a few occasions pairs with no trade opportunity can be observed and thus a return of zero is recorded. Out-of-sample backtesting results are overall in line with the in-sample results in the sense that 2-stdev rule offers consistently good performance and the Kalman filter and ZL methods are not able to improve the profits considerably. Therefore, both results confirm that 2-stdev and GA methods perform well, however, given the simplicity and the model free nature of the 2-stdev rule, it is not now a surprise why it does remain as the first choice among practitioners.

4.4 Profitability and the maximum holding period for the spread positions

The in-sample and out-of-sample backtesting results show that pairs trading generates high profits which are obtained for various contracts and commodity pairs.[10] Non-surprisingly, the simplest and model free approach of 2-stdev trading method can be considered safely as a very good first choice for the traders. The out-of-sample backtesting is repeated with 9 blocks of sub-samples with 1 year trading periods.

We can also appreciate that the relationship between the duration of the spread positions and the profitability of pairs trading is rather crucial. Thus, in the presence of risk-controls, which is often the case in practice, the spread positions do not last for the whole maturity of futures contracts. For instance, if the pairs trading of 2 commodities has an average drawdown of 5%, this implies that on average for every spread position, the spread diverges 5% above the threshold that the position was opened. In case that a stop-loss barrier of 3% exists, then most of the time the stop-loss barrier is hit before realizing the mean-reversion and the profit. This drastically changes the profitability of pairs trading in practice. Furthermore, if the hedge funds are expected to act without risk-controls and they can exploit the so called *market inefficiency*, this might not be a feasible expectation since a high leverage exists in the hedge fund industry and the highly possible maximum drawdown that they experience from the pairs trading might lead them to bankruptcy. Therefore, in order to test that argument, the out-of-backtesting is re-run with 6 different constraints on the maximum holding periods of each spread position with 10, 22, 44, 66, 126, and 252 trading days. By repeating the out-of-sample backtesting for the last trading year, the Sharpe ratio and the excess return/maximum drawdown are calculated.

Figure 4 shows that the Sharpe ratio and the maximum drawdown normalized average returns obtained from pairs trading with 1, 2 and 3 best spread positions selected from the out-of-sample backtesting. It is observed that the Sharpe ratio

[10] We should also note that in our backtesting experiments we do not assume any leveraged position in line with the conservative approach followed in this study.

and maximum drawdown normalized returns decrease as the maximum holding periods is getting shorter. This confirms that if the spread position cannot be held for long-periods of time, then the profitability decreases rapidly. Therefore, the high-profits that can be observed in the Chinese commodity futures market does not necessarily imply the inefficiency of the market.

As a robustness check, we verify the effect of different maximum holding positions for the spread trading. In Table 7, the out-of-sample backtesting results for the last trading year is repeated for all the best 6 pairs from Table 5 with different maximum holding periods. Then, the annual returns, the largest drawdown during each spread position, and the maximum 1 year drawdown is calculated. Our results in Table 7 confirm that as the maximum holding period of each spread position gets shorter, this reduces the profitability and annual return/average drawdown ratio. Overall, the same phenomenon is observed to occur in all the identified best pairs. Therefore, the relationship between the profitability and the maximum holding period for spreads is robust in time and across the different pairs that are considered.

Finally, we construct six different portfolios that consist of equally weighted investments in 1 to 6 different pairs that are identified from the combined criteria of correlation, SSD, and profit index ranking. The results in Table 8 can be summarized as follows. The pairs that are identified as the best ones from the combined criteria of correlation, SSD, and profit index, indeed produce higher returns. In case that a shorter maximum duration is observed for each spread position, the profitability for pairs trading portfolios with different number of pairs appears to have the same behaviour in all cases.

5 Conclusion

The results which are presented in this paper verify the profitability of different pairs selection and pairs trading models for the Chinese commodity futures market. All the available pairs trading models which have been suggested in the literature are compared in terms of the returns and risk-adjusted returns generated via in-sample and out-of-sample backtesting. There is evidence that the profitability of pairs trading decreases as shorter maximum holding periods are imposed for each spread position. Robustness of results are verified for different subperiods and for considering a variety of portfolios of pairs. Pairs trading offers consistently high profits, however, at the same time high drawdown risk. Therefore, high pairs trading profits at least should not be taken for granted as market anomaly or inefficiency. There exists a clear risk-return relationship when one accounts for the short positions in spread trading and the potential drawdown during the holding period of the spread. If the trader is able to hold the spread position for very long periods of time withstanding high-drawdown, then s/he can realize the pairs trading profits. However, if the decrease in the value of the spread position during the long holding periods cannot be tolerated (for example, due to the stop-loss barriers), then the profitability of pairs trading decreases drastically, invalidating the market inefficiency arguments.

Furthermore, our empirical comparison of pairs formation and trading models shows that the 2-stdev rule performs quite well compared to the more sophisticated methods which are model-driven pairs trading. The identification of the best pairs is the primary determinant for the profitability of pairs trading, whereas the choice of the trade thresholds (triggers) can be based on simpler rules based on the historical standard deviation of the spread process.

References

Alexander, C. (1999), "Optimal hedging using cointegration", *Philosophical Transactions of the Royal Society Series A*, Vol. 357, pp. 2039–2058.

Alexander, C. and Dimitriu, A. (2005a), "Indexing and statistical arbitrage: tracking error or cointegration?", *Journal of Portfolio Management*, Vol. 31 No. 2, pp. 50–63.

Alexander, C. and Dimitriu, A. (2005b), "Indexing, cointegration and equity market regimes", *International Journal of Finance and Economics*, Vol. 10, pp. 213–231.

Avellaneda, M. and Lee, J. (2010), "Statistical arbitrage in the US equities market", *Quantitative Finance*, Vol. 10 No. 7, pp. 761–782.

Baronyan, S., Boduroglu, l. and Sener, E. (2010), "Investigation of stochastic pairs trading strategies under different volatility regimes", *The Manchester School*, Vol. 78 No. s1, pp. 114–134.

Bertram, W. (2010), "Analytical solutions for optimal statistical arbitrage trading", *Physica A: Statistical Mechanics and its Applications*, Vol. 389 No. 11, pp. 2234–2243.

Bondarenko, O. (2003), "Statistical arbitrage and securities prices", *Review of Financial Studies*, Vol. 16 No. 3, pp. 875–919.

Bogomolov, T. (2013), "Pairs trading based on statistical variability of the spread process", *Quantitative Finance*, Vol. 13 No. 9, pp. 1411–1430.

Bowen, D. and Hutchinson, M. (2014), "Pairs trading in the UK equity market: risk and return", *The European Journal of Finance*, pp. 1–25 (forthcoming).

Conrad, J. and Kaul, G. (1989), "Mean Reversion in Short-horizon Expected Returns", *Review of Financial Studies*, Vol. 2, pp. 225–240.

Cummins, M. and Bucca, A. (2012), "Quantitative spread trading on crude oil and refined products markets", *Quantitative Finance*, Vol. 12 No. 12, pp. 1857–1875.

Do, B. and Faff, R. (2010), “Does Simple Pairs Trading Still Work?”, *Financial Analysts Journal*, Vol. 66 No. 4, pp. 83–95.

Do, B. and Faff, R. (2012), “Are Pairs Trading Profits Robust to Trading Costs?”, *The Journal of Financial Research*, Vol. 35 No. 2, pp. 261–287.

Elliott, R. and Krishnamurthy, V. (1999), “New finite-dimensional filters for parameter estimation of discrete-time linear Gaussian models”, *IEEE Transactions on Automatic Control*, Vol. 44 No. 5, pp. 938–951.

Elliott, R., Van Der Hoek, J. and Malcolm, W. (2005), “Pairs trading”, *Quantitative Finance*, Vol. 5 No. 3, pp. 271–276.

Focardi, S.M., Fabozzi, F.J. and Mitov, I.K. (2016), “A new approach to statistical arbitrage: Strategies based on dynamic factor models of prices and their performance”, *Journal of Banking & Finance*, Vol. 65, pp. 134–155.

Gatev, E., Goetzmann, W. and Rouwenhorst, K. (2006), “Pairs trading: Performance of a relative-value arbitrage rule”, *Review of Financial Studies*, Vol. 19 No. 3, pp. 797–827.

Göncü, A. (2015), “Statistical arbitrage in the Black - Scholes framework”, *Quantitative Finance*, Vol. 15 No. 9, pp. 1489–1499.

Göncü, A. and Akyıldırım, E. (2016), “Statistical Arbitrage with Pairs Trading”, *International Review of Finance*, Vol. 16 No. 2, pp. 307–319.

Hogan, S., Jarrow, R., Teo, M. and Warachka, M. (2004), “Testing market efficiency using statistical arbitrage with applications to momentum and value strategies”, *Journal of Financial Economics*, Vol. 73, pp. 525–565.

Huck, N. (2009), “Pairs selection and outranking: An application to the S&P 100 index”, *European Journal of Operational Research*, Vol. 196 No. 2, pp. 819–825.

Huck, N. (2013), “The high sensitivity of pairs trading returns”, *Applied Economics Letters*, Vol. 20 No. 14, pp. 1301–1304.

Huck, N. and Afawubo, K. (2015), “Pairs trading and selection methods: is cointegration superior?”, *Applied Economics*, Vol. 47 No. 6, pp. 599–613.

Jacobs, H. and Weber, M. (2015), “On the determinants of pairs trading profitability”, *Journal of Financial Markets*, Vol. 23, pp. 75–97.

Jegadeesh, N. (1990), “Evidence of Predictable Behavior of Security Returns”, *Journal of Finance*, Vol. 45, pp. 881–898.

Jegadeesh, N., Titman, S. (1993), “Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency”, *Journal of Finance*, Vol. 48, pp. 65–91.

Jegadeesh, N., Titman, S. (1995), “Overreaction, Delayed Reaction, and Contrarian Profits”, *Review of Financial Studies*, Vol. 8, pp. 973–993.

Jarrow, R., Teo, M., Tse, Y. K., and Warachka, M. (2012), “An improved test for statistical arbitrage”, *Journal of Financial Markets*, Vol. 15 No. 1, pp. 47–80.

Perlin, M. (2009), “Evaluation of pairs-trading strategy at the Brazilian financial market”, *Journal of Derivatives & Hedge Funds*, Vol. 15 No. 2, pp. 122–136.

Shumway, R. and Stoffer, D. (1982), “An approach to time series smoothing and forecasting using the EM algorithm”, *Journal of Time Series Analysis*, Vol. 3 Np. 4, pp. 253–264.

Zeng, Z. and Lee, C. (2014), “Pairs trading: optimal thresholds and profitability”, *Quantitative Finance*, Vol. 14 No. 11, pp. 1881–1893.

Table 1: Contracts traded in the Shanghai (SHFE), Dalian (DCE), and Zhengzhou (CZCE) commodity futures markets with more than 1,000 observations.

Commodity	SHFE	Wind Ticker	Start Date	Maturity Months
Aluminum	SHFE	AL	May 1992	FGHJKMNQUVXZ
Gold	SHFE	AU	January 2008	FGHJKMNQUVXZ
Copper	SHFE	CU	May 1992	FGHJKMNQUVXZ
Fuel Oil	SHFE	FU	August 2004	FGHJKMNQUVXZ
Lead	SHFE	PB	March 2011	FGHJKMNQUVXZ
Screw Steel	SHFE	RB	March 2009	FGHJKMNQUVXZ
Zinc	SHFE	ZN	March 2007	FGHJKMNQUVXZ
Natural Rubber	SHFE	RU	January 1999	FGHJKMNQUVXZ
Wire Material	SHFE	WR	March 2009	FGHJKMNQUVXZ
Soybean 1	DCE	A	March 2002	FHKNUX
Soybean 2	DCE	B	December 2004	FHKNUX
Bean Dreg	DCE	M	July 2000	FHKNQUXZ
Bean Oil	DCE	Y	Jaunary 2006	FHKNQUXZ
Corn	DCE	C	September 2004	FHKNUX
Paim Oil	DCE	P	Octomber 2007	FGHJKMNQUVXZ
Polythene	DCE	L	July 2007	FGHJKMNQUVXZ
Ployvinyl Chloride	DCE	V	May 2009	FGHJKMNQUVXZ
Coke	DCE	J	April 2011	FGHJKMNQUVXZ
Cotton	CZCE	CF	June 2004	FHKNUX
Methyl Alcohol	CZCE	MA	Octomber 2011	FGHJKMNQUVXZ
Strong Wheat	CZCE	WH	March 2003	FHKNUX
White Sugar	CZCE	SR	January 2006	FHKNUX
Pure terephthalic Acid	CZCE	TA	December 2006	FGHJKMNQUVXZ
Colza Oil	CZCE	OI	June 2007	FHKNUX
Indica Rice	CZCE	RI	April 2009	FHKNUX

Note: The letter codes are F (January), G (February), H (March), J (April), K (May), M (June), N (July), Q (August), U (September), V (October), X (November) and Z (December).

Table 2: Pairs trading returns with different number of pairs in a portfolio.

Periods/Portfolios	**1 Pair**	**2 Pairs**	**3 Pairs**	**4 Pairs**	**5 Pairs**	**6 Pairs**
Panel A: Returns (%) (no waiting)						
Subperiod 1	31.39	24.24	33.48	29.62	28.61	29.95
Subperiod 2	44.29	48.24	45.08	41.58	32.87	29.49
Subperiod 3	38.30	43.25	37.90	44.24	44.02	41.28
Subperiod 4	13.98	28.58	21.16	21.21	22.90	19.46
Subperiod 5	43.45	24.65	14.96	13.12	10.95	9.06
Subperiod 6	24.83	4.20	0.52	-1.23	3.47	3.28
Subperiod 7	4.42	5.80	4.95	4.63	5.85	3.78
Subperiod 8	10.50	9.59	6.94	3.96	3.88	4.18
Subperiod 9	5.26	9.63	6.64	5.04	5.67	6.59
Average Return	24.03	22.02	19.07	18.02	17.58	16.34
Stdev.	16.06	16.13	16.22	17.05	14.97	14.21
Panel B: Returns (%) (one day waiting)						
Subperiod 1	31.16	28.14	33.74	31.75	30.41	32.67
Subperiod 2	37.67	51.79	48.86	43.93	34.22	30.00
Subperiod 3	32.86	47.49	39.35	43.01	43.04	40.12
Subperiod 4	12.00	27.94	18.35	16.88	20.37	16.91
Subperiod 5	38.38	20.78	12.86	9.80	7.76	6.34
Subperiod 6	17.20	-0.07	-2.80	-2.50	2.38	1.91
Subperiod 7	4.43	3.82	3.14	2.71	3.44	1.71
Subperiod 8	12.58	9.20	4.99	2.54	2.32	2.28
Subperiod 9	0.47	5.61	4.06	2.80	3.42	3.90
Average Return	20.75	21.63	18.06	16.77	16.37	15.09
Stdev.	14.52	18.91	18.37	18.26	15.99	15.33
Panel C: Trading Statistics						
Avg Number of Trades/Year	4	4.3	4.1	3.9	3.6	3.4
Average Trade Duration	35.2	37.8	39.1	38	43.3	44.1
Percentage of Negative Returns	0	6%	11%	14%	13%	15%

Best pairs are identified using the training sample by first filtering the pairs that have correlation coefficient larger than 0.5 and choosing the top eight minimum SSD and largest profit index values. Annual returns are obtained with the annual trading period. We trade according to the rule that opens a position in a pair at the end of the day when the divergence of the spread exceeds the historical two standard deviations around the mean (Panel A). The results in Panel B correspond to a strategy that delays opening of the spread position by one day. Panel C presents the trading statistics.

Table 3: Backtesting results (transaction costs included): annualized returns achieved from the in-sample data for the considered methods: Göncü and Akyildirim (GA), Zeng and Lee (ZL), and Kalman-Filter (KF) methods. Average number of days per spread position is given in parenthesis. Trading period is from May 22^{nd}, 2015 to June 1^{st}, 2016, consisting of totally 252 trading days.

Criteria	Maturity	Pair	GA	GA-KF	ZL	ZL-KF	2-σ	2-σ-KF
Correlation	JAN	(WR-RB)	22.71%	43.01%	27.80%	20.78%	49.57%	20.53%
		SHFE	(16.8)	(20.3)	(33.7)	(33.9)	(19.9)	(19.9)
	MAR	(WR-RB)	23.38%	32.83%	37.89%	44.74%	32.95%	30.12%
		SHFE	(14.3)	(20.2)	(27.9)	(22.8)	(19.7)	(22.3)
	SEP	(WR-RB)	36.74%	36.98%	49.96%	39.30%	31.52%	32.60%
		SHFE	(32)	(32.3)	(31.4)	(31.4)	(34.7)	(30.3)
	Average	**Return**	**27.61%**	**37.61%**	**38.55%**	**34.94%**	**38.01%**	**27.75%**
		Hitting Days	(21)	(24.2)	(31)	(29.4)	(24.7)	(24.2)
Profitability Index	SEP	(WR-RB)	36.74%	36.98%	49.96%	39.30%	31.52%	32.60%
		SHFE	(32)	(32.3)	(31.4)	(31.4)	(34.7)	(30.3)
	JULY	(PB-FU)	66.23%	-31.10%	14.81%	-38.62%	48.50%	-31.10%
		SHFE	(43.5)	(122.5)	(76.7)	(83.7)	(51.5)	(122.5)
	NOV	(V-J)	0%	-33.24%	45.39%	-76.61%	0%	-33.24%
		DCE	(0)	(122.5)	(35.9)	(83.7)	(0)	(122.5)
	Average	**Return**	**33.67%**	**-9.12%**	**36.72%**	**-25.31%**	**26.00%**	**-10.58%**
		Hitting Days	(25.2)	(92.4)	(48)	(66.2)	(31.7)	(91.8)
SSD	MAY	(PB-AL)	17.87%	12.26%	17.47%	17.47%	13.52%	18.15%
		SHFE	(127)	(132)	(62.5)	(62.5)	(130)	(89)
	MAY	(WR-RB)	8.87%	8.87%	-4.73%	-11.39%	13.57%	-5.20%
		SHFE	(23)	(23)	(83.7)	(83.7)	(28.3)	(34.8)
	JULY	(PB-AL)	14.94%	12.42%	11.53%	9.93%	12.42%	16.98%
		SHFE	(127)	(130)	(105.5)	(105.5)	(130)	(72.5)
	Average	**Return**	**13.89%**	**11.18%**	**8.09%**	**6.34%**	**13.17%**	**9.98%**
		Hitting Days	(92.3)	(95)	(83.9)	(83.9)	(96.1)	(65.4)
Combination	SEP	(WR-RB)	36.74%	36.98%	49.96%	39.30%	31.52%	32.60%
		SHFE	(32)	(32.3)	(31.4)	(31.4)	(34.7)	(30.3)
	JULY	(PB-FU)	66.23%	-31.10%	14.81%	-38.62%	48.50%	-31.10%
		SHFE	(43.5)	(122.5)	(76.7)	(83.7)	(51.5)	(122.5)
	MAR	(PB-FU)	13.86%	-32.14%	11.27%	-28.48%	7.06%	-32.14%
		SHFE	(129)	(122.5)	(58.5)	(83.7)	(130)	(122.5)
	Average	**Return**	**38.95%**	**-8.75%**	**25.35%**	**-9.27%**	**29.02%**	**-10.21%**
		Hitting Days	(68.2)	(92.4)	(55.5)	(66.2)	(72.1)	(91.8)

Table 4: Backtesting results (transaction costs included): annualized returns achieved from the in-sample data for the considered methods: Göncü and Akyildirim (GA), Zeng and Lee (ZL), and Kalman-Filter (FL) methods. Average number of days per spread position is given in parenthesis. Trading period is from May 13^{th}, 2014 to May 22^{nd}, 2015, totally 252 trading days.

Criteria	Maturity	Pair	GA	GA-KF	ZL	ZL-KF	2-σ	2-σ-KF
Correlation	MAY	(WR-RB)	37.01%	26.89%	34.64%	50.57%	29.32%	40.76%
		SHFE	(10.2)	(14.7)	(24.3)	(16.9)	(12.5)	(11.1)
	JAN	(WR-RB)	42.03%	37.99%	32.85%	29.01%	34.07%	44.53%
		SHFE	(16.6)	(22.2)	(34.1)	(35)	(17)	(16.7)
	MAR	(WR-RB)	11.49%	11.98%	44.00%	40.97%	12.85%	27.70%
		SHFE	(8.3)	(13.5)	(26.4)	(27.2)	(13)	(14.9)
	Average	**Return**	**30.18%**	**25.62%**	**37.17%**	**40.18%**	**25.31%**	**37.66%**
		Hitting Days	(11.7)	(16.8)	(28.3)	(26.3)	(14.2)	(14.2)
Profitability Index	SEP	(WR-RB)	18.70%	25.39%	8.40%	23.61%	14.20%	20.06%
		SHFE	(22)	(20.3)	(56.8)	(35.4)	(27.5)	(21)
	MAY	(WR-RB)	37.01%	26.89%	34.64%	50.57%	29.32%	40.76%
		SHFE	(10.2)	(14.7)	(24.3)	(16.9)	(12.58)	(11.1)
	JAN	(WR-RB)	42.03%	37.99%	32.85%	29.01%	34.07%	44.53%
		SHFE	(16.6)	(22.2)	(34.1)	(35)	(17)	(16.7)
	Average	**Return**	**32.58%**	**30.09%**	**25.30%**	**34.39%**	**25.86%**	**35.12%**
		Hitting Days	(16.3)	(19.1)	(38.4)	(29.1)	(19)	(16.3)
SSD	MAY	(PB-AL)	7.72%	10.92%	6.93%	6.15%	7.22%	7.64%
		SHFE	(25)	(14)	(41.8)	(41.8)	(13.5)	(18.8)
	MAY	(WR-RB)	37.01%	26.89%	34.64%	50.57%	29.32%	40.76%
		SHFE	(10.2)	(14.7)	(24.3)	(16.9)	(12.5)	(11.1)
	JULY	(WR-RB)	45.08%	43.80%	35.68%	32.93%	44.72%	44.34%
		SHFE	(17.8)	(21.2)	(35.7)	(31.3)	(21.2)	(22.3)
	Average	**Return**	**29.94%**	**27.20%**	**25.75%**	**29.88%**	**27.09%**	**30.91%**
		Hitting Days	(17.6)	(16.6)	(34)	(30)	(15.7)	(17.4)
Combination	SEP	(WR-RB)	18.70%	25.39%	8.40%	23.61%	14.20%	20.06%
		SHFE	(22)	(20.3)	(56.8)	(35.4)	(27.5)	(21)
	MAY	(WR-RB)	37.01%	26.89%	34.64%	50.57%	29.32%	40.76%
		SHFE	(10.2)	(14.7)	(24.3)	(16.9)	(12.5)	(11.1)
	JAN	(WR-RB)	42.03%	37.99%	32.85%	29.01%	34.07%	44.53%
		SHFE	(16.6)	(22.2)	(34.1)	(35)	(17)	(16.7)
	Average	**Return**	**32.58%**	**30.09%**	**25.30%**	**34.39%**	**25.86%**	**35.12%**
		Hitting Days	(16.3)	(19.1)	(38.4)	(29.1)	(19)	(16.37)

Table 5: Backtesting results (transaction costs included): annualized returns achieved in the out-of-sample backtesting with daily updated parameters for the considered methods: Goncu and Akyildirim (GA), Zeng and Lee (ZL), and Kalman-Filter (KF) methods. Average number of days per spread position is given in parenthesis. The out-of-sample period is from May 22^{nd}, 2015 to June 1^{st}, 2016, totally 252 trading days.

Criteria	Maturity	Pair	GA	GA-KF	ZL	ZL-KF	2-σ	2-σ-KF
Correlation	MAY	(WR-RB)	13.11%	3.02%	1.70%	1.70%	5.09%	2.53%
		SHFE	(31)	(35.3)	(50.2)	(50.2)	(33.8)	(36)
	JAN	(WR-RB)	46.71%	38.58%	15.48%	15.48%	37.96%	17.87%
		SHFE	(17.7)	(21.4)	(33.9)	(33.9)	(22.4)	(20.2)
	MAR	(WR-RB)	29.55%	24.64%	39.22%	30.61%	24.64%	14.51%
		SHFE	(17)	(22.7)	(27.9)	(27.9)	(22.7)	(23.3)
	Average	**Return**	**29.79%**	**22.08%**	**18.80%**	**15.93%**	**22.56%**	**11.64%**
		Hitting Days	(21.9)	(26.5)	(37.3)	(37.3)	(26.3)	(26.5)
Profitability Index	SEP	(WR-RB)	32.26%	32.77%	8.43%	18.12%	25.39%	40.27%
		SHFE	(37)	(31.8)	(83.7)	(50.2)	(42)	(27.8)
	MAY	(WR-RB)	13.11%	3.02%	1.70%	1.70%	5.09%	2.53%
		SHFE	(31)	(35.3)	(50.2)	(50.2)	(33.8)	(36)
	JAN	(WR-RB)	46.71%	38.58%	15.48%	15.48%	37.96%	17.87%
		SHFE	(17.7)	(31.4)	(33.9)	(33.9)	(22.4)	(20.2)
	Average	**Return**	**30.69%**	**24.79%**	**8.54%**	**11.87%**	**22.81%**	**20.22%**
		Hitting Days	(28.6)	(32.8)	(55.9)	(44.8)	(32.7)	(28)
SSD	MAY	(PB-AL)	7.94%	4.82%	-4.71%	-6.34%	6.38%	4.82%
		SHFE	(140)	(173)	(72.7)	(72.7)	(163)	(173)
	MAY	(WR-RB)	13.11%	3.02%	1.70%	1.70%	5.09%	2.53%
		SHFE	(31)	(35.3)	(50.2)	(50.2)	(33.8)	(36)
	JULY	(WR-RB)	9.69%	6.09%	4.00%	3.18%	10.03%	7.49%
		SHFE	(36.2)	(39)	(62.3)	(62.8)	(35.3)	(42)
	Average	**Return**	**10.25%**	**4.64%**	**0.33%**	**-0.49%**	**7.17%**	**4.95%**
		Hitting Days	(69.1)	(82.4)	(61.7)	(61.9)	(77.4)	(83.7)
Combination	SEP	(WR-RB)	32.26%	32.77%	8.43%	18.12%	25.39%	36.31%
		SHFE	(37)	(31.8)	(83.7)	(50.2)	(42)	(33)
	MAY	(WR-RB)	13.11%	3.02%	1.70%	1.70%	5.09%	2.53%
		SHFE	(31)	(35.3)	(50.2)	(50.2)	(33.8)	(36)
	JAN	(WR-RB)	46.71%	38.58%	15.48%	15.48%	37.96%	17.87%
		SHFE	(17.7)	(21.4)	(33.9)	(33.9)	(22.4)	(20.2)
	Average	**Return**	**30.69%**	**24.79%**	**8.54%**	**11.87%**	**22.81%**	**18.90%**
		Hitting Days	(28.6)	(32.8)	(55.9)	(44.8)	(32.7)	(29.7)

Table 6: Backtesting results (transaction costs included): annualized returns achieved in the out-of-sample backtesting with daily updated parameters for the considered methods: Goncu and Akyildirim (GA), Zeng and Lee (ZL), and Kalman-Filter (KF) methods. Average number of days per spread position is given in parenthesis. The out-of-sample period is from May 13^{th}, 2014 to May 22^{nd}, 2015, totally 252 trading days.

Criteria	Maturity	Pair	GA	GA-KF	ZL	ZL-KF	2-σ	2-σ-KF
Correlation	JAN	(WR-RB)	19.76%	20.13%	28.15%	23.64%	15.31%	19.81%
		SHFE	(60.3)	(48.5)	(35.7)	(35.7)	(64.7)	(37)
	MAY	(WR-RB)	17.85%	39.86%	18.33%	17.49%	34.28%	29.23%
		SHFE	(22)	(9.9)	(27.9)	(27.9)	(12.6)	(12)
	MAY	(Y-P)	0%	6.03%	-11.32%	-11.79%	6.03%	-4.52%
		DCE	(0)	(5)	(124)	(125.5)	(5)	(122.5)
	Average	**Return**	**12.54%**	**22.00%**	**11.72%**	**9.78%**	**18.54%**	**14.84%**
		Hitting Days	(27.4)	(21.1)	(62.5)	(63.0)	(27.4)	(57.2)
Profitability Index	SEP	(WR-RB)	5.41%	8.24%	-8.75%	-8.75%	1.05%	16.03%
		SHFE	(43)	(45)	(62.5)	(62.5)	(59.3)	(29.1)
	MAY	(WR-RB)	17.85%	39.86%	18.33%	17.49%	34.28%	29.23%
		SHFE	(22)	(9.9)	(27.9)	(27.9)	(12.6)	(12)
	NOV	(WR-RB)	10.14%	21.77%	19.03%	19.03%	15.59%	27.03%
		SHFE	(67)	(34)	(47.8)	(47.8)	(48.7)	(32)
	Average	**Return**	**11.13%**	**23.29%**	**9.54%**	**9.25%**	**16.97%**	**24.09%**
		Hitting Days	(44)	(29.6)	(46.1)	(46.1)	(40.2)	(24.3)
SSD	MAY	(WR-RB)	17.85%	39.86%	18.33%	17.49%	34.28%	29.23%
		SHFE	(22)	(9.9)	(27.9)	(27.9)	(12.6)	(12)
	JAN	(ZN-PB)	-12.99%	-12.56%	-17.16%	-17.16%	-13.33%	-15.01%
		SHFE	(138)	(109.5)	(83.7)	(83.7)	(113)	(122.5)
	NOV	(ZN-PB)	-12.97%	-24.05%	-27.33%	-27.33%	-22.60%	-24.95%
		SHFE	(93)	(118.5)	(83.7)	(83.7)	(113.5)	(122.5)
	Average	**Return**	**-2.70%**	**1.08%**	**-8.72%**	**-9.00%**	**-0.55%**	**-3.58%**
		Hitting Days	(84.3)	(79.3)	(65.1)	(65.1)	(79.7)	(85.7)
Combination	SEP	(WR-RB)	5.41%	8.24%	-8.75%	-8.75%	1.05%	16.03%
		SHFE	(43)	(45)	(62.5)	(62.5)	(59.3)	(29.1)
	MAY	(WR-RB)	17.85%	39.86%	18.33%	17.49%	35.71%	30.62%
		SHFE	(22)	(9.9)	(27.9)	(27.9)	(12.6)	(12)
	NOV	(WR-RB)	10.14%	21.77%	19.03%	19.03%	15.59%	27.03%
		SHFE	(67)	(34)	(47.8)	(47.8)	(48.7)	(32)
	Average	**Return**	**11.13%**	**23.29%**	**9.54%**	**9.25%**	**16.97%**	**24.09%**
		Hitting Days	(44)	(29.6)	(46.1)	(46.1)	(40.2)	(24.4)

Table 7: Out-of-sample backtesting with the different contracts using the 2-stdev rule in the last one year as the out-of-sample period. Annual returns and average/maximum drawdown are calculated with different maximum holding periods for the spread positions. Annual returns divided by average drawdown are also given.

Pair	Measures	22-days	44-days	66-days	126-days	252-days
(WR-RB)	Annual Return	38.09%	42.34%	37.96%	37.96%	37.96%
JAN	Avg. Drawdown	5.31%	6.61%	7.24%	7.24%	7.24%
SHFE	Max. Drawdown	21.43%	21.43%	21.43%	21.43%	21.43%
	Annual Return / Avg. Drawdown	7.17	6.40	5.24	5.24	5.24
(WR-RB)	Annual Return	10.94%	18.04%	24.64%	24.64%	24.64%
MAR	Avg. Drawdown	5.91%	6.14%	6.14%	6.14%	6.14%
SHFE	Max. Drawdown	12.84%	12.84%	12.84%	12.84%	12.84%
	Annual Return / Avg. Drawdown	1.85	2.94	4.01	4.01	4.01
(WR-RB)	Annual Return	-10.37%	2.10%	3.96%	5.09%	5.09%
MAY	Avg. Drawdown	5.99%	8.85%	9.34%	10.07%	10.07%
SHFE	Max. Drawdown	23.20%	23.80%	45.10%	45.10%	45.10%
	Annual Return / Avg. Drawdown	-1.73	0.24	0.42	0.51	0.51
(WR-RB)	Annual Return	-9.78%	8.46%	9.48%	10.03%	10.03%
JULY	Avg. Drawdown	8.74%	8.99%	11.24%	11.24%	11.24%
SHFE	Max. Drawdown	17.11%	24.50%	24.50%	24.50%	24.50%
	Annual Return / Avg. Drawdown	-1.12	0.94	0.84	0.89	0.89
(WR-RB)	Annual Return	9.49%	20.19%	25.39%	25.39%	25.39%
SEP	Avg. Drawdown	10.07%	9.31%	12.09%	12.09%	12.09%
SHFE	Max. Drawdown	17.54%	21.64%	21.64%	21.64%	21.64%
	Annual Return / Avg. Drawdown	0.94	2.17	2.10	2.10	2.10
(PB-AL)	Annual Return	-5.26%	-3.15%	-4.68%	2.51%	6.38%
MAY	Avg. Drawdown	3.67%	6.36%	7.16%	9.07%	16.51%
SHFE	Max. Drawdown	15.09%	16.51%	16.51%	16.51%	16.51%
	Annual Return / Avg. Drawdown	-1.43	-0.49	-0.65	0.28	0.39

Table 8: Average annualized returns, standard deviations of returns, and the Sharpe ratios calculated with the risk-free rate assumed as 3%. The t-statistic of the mean is computed using Newey-West standard errors with two lags. The result is calculated by repeating the out-of-sample backtesting for the 9 trading subperiods with respect to different maximum holding period.

Profitability	1 pair	2 pairs	3 pairs	4 pairs	5 pairs	6 pairs
Panel A: Termination T=22						
Average Return (%)	8.94	8.13	6.95	5.79	5.28	4.44
Standard Deviation (%)	13.62	10.79	10.93	11.46	10.19	9.19
Standard Error (Newey-West) (%)	6.94	6.78	6.63	6.42	5.79	5.07
Sharpe Ratio	0.44	0.48	0.36	0.24	0.22	0.16
Return /Avg Max.Drawdown	0.99	0.92	0.74	0.62	0.61	0.53
t-Statistics	1.29	1.20	1.05	0.90	0.91	0.88
p-Value	0.12	0.13	0.16	0.20	0.19	0.20
Median (%)	9.49	1.33	8.72	6.86	2.72	2.51
Skewness	0.28	0.43	0.31	0.43	0.53	0.88
Kurtosis	1.82	1.43	1.94	1.86	1.84	2.63
Minimum (%)	-6.82	-2.79	-5.23	-6.59	-5.04	-3.96
Maximum (%)	31.91	23.04	25.83	22.91	22.31	22.79
Negative Return	33%	33%	33%	44%	44%	44%
Panel B: Termination $T4$=66						
Average Return (%)	15.43	13.04	10.17	9.44	9.45	8.49
Standard Deviation (%)	14.53	13.91	13.69	14.52	12.49	11.36
Standard Error (Newey-West) (%)	11.00	10.09	8.75	8.82	8.18	7.35
Sharpe Ratio	0.86	0.72	0.52	0.44	0.52	0.48
Return /Avg Max.Drawdown	1.54	1.23	0.94	0.85	0.94	0.87
t-Statistics	1.40	1.29	1.16	1.07	1.16	1.15
p-Value	0.10	0.12	0.14	0.16	0.14	0.14
Median (%)	11.17	14.67	8.64	7.42	6.00	3.01
Skewness	0.14	0.50	0.52	0.53	0.62	0.63
Kurtosis	1.44	1.95	1.92	1.83	1.99	1.89
Minimum (%)	-2.98	-2.98	-6.27	-7.82	-4.00	-3.28
Maximum (%)	36.19	36.24	33.75	31.08	31.07	28.24
Negative Return	11%	11%	33%	44%	33%	33%
Panel C: Termination T=126						
Average Return (%)	16.03	13.35	10.92	10.24	10.56	9.50
Standard Deviation (%)	13.85	13.50	13.77	14.67	12.52	11.79
Standard Error (Newey-West) (%)	11.05	10.05	9.11	9.22	8.63	7.87
Sharpe Ratio	0.94	0.77	0.58	0.49	0.60	0.55
Return /Avg Max.Drawdown	1.60	1.23	0.99	0.91	1.03	0.95
t-Statistics	1.45	1.33	1.20	1.11	1.22	1.21
p-Value	0.09	0.11	0.13	0.15	0.13	0.13
Median (%)	12.83	14.65	6.96	6.12	4.95	2.48
Skewness	0.19	0.52	0.46	0.48	0.57	0.62
Kurtosis	1.40	2.10	1.87	1.84	1.90	1.88
Minimum (%)	0.42	-3.80	-6.82	-8.23	2.93	2.39
Maximum (%)	36.19	36.24	33.75	32.74	32.42	30.28
Negative Return	0%	11%	11%	33%	11%	11%

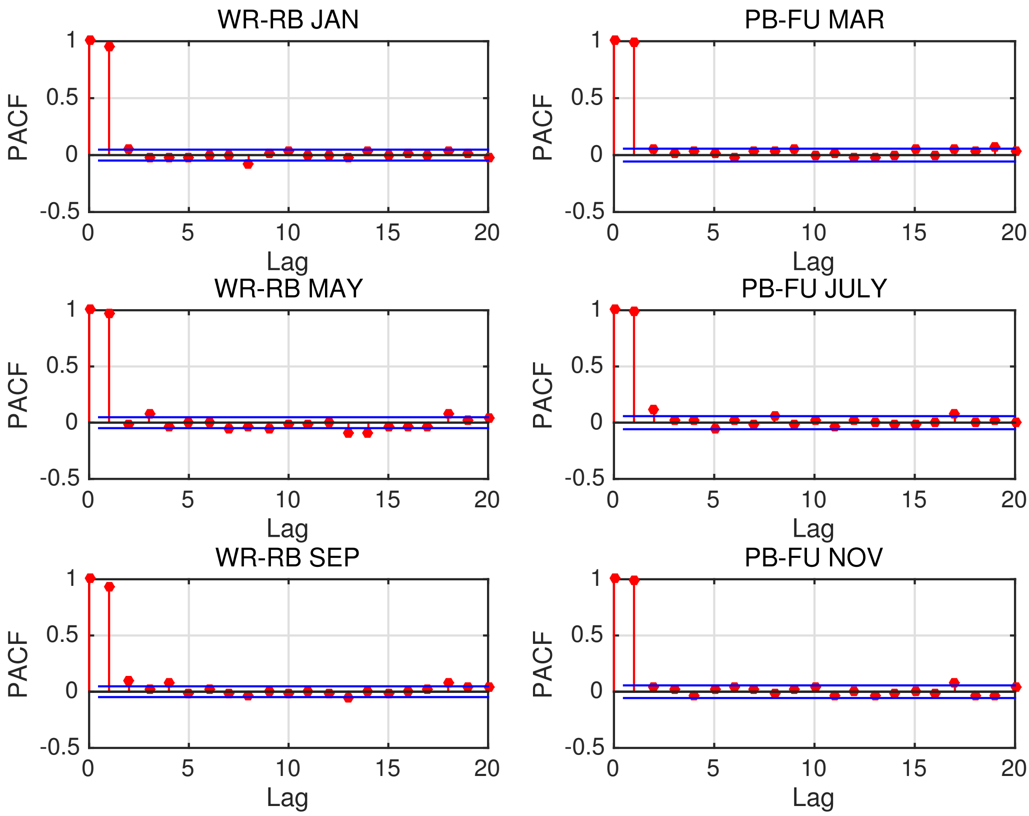

Figure 1: Partial autocorrelation functions for the spreads of the given pairs of futures contracts, which exhibit AR(1).

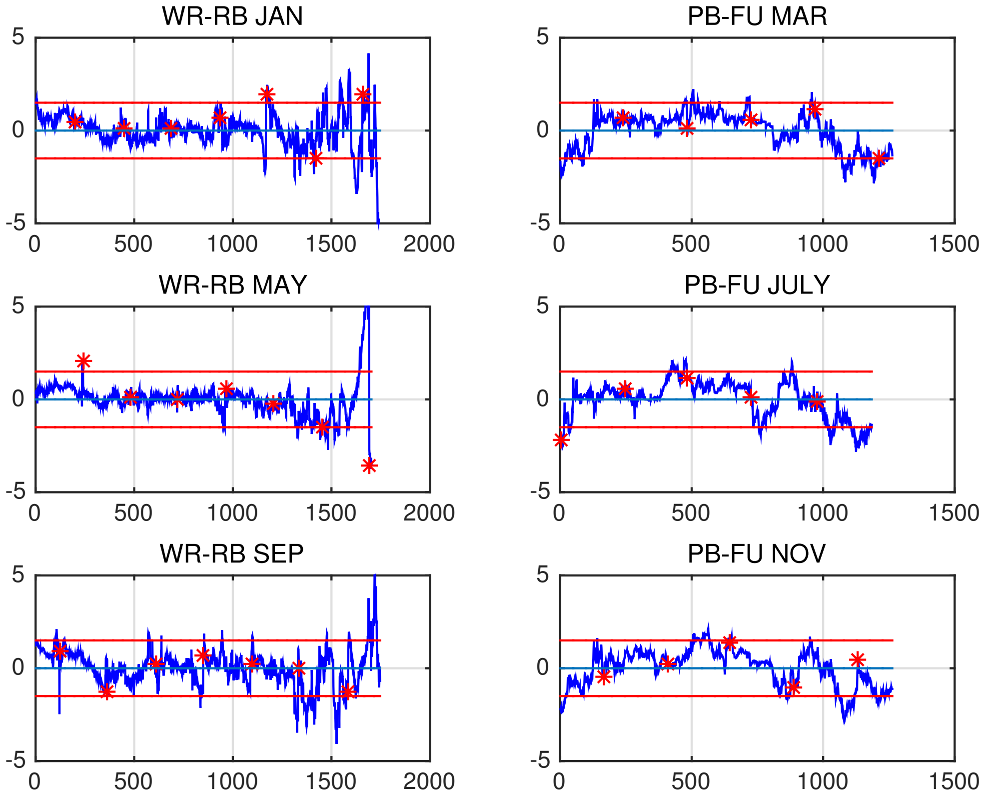

Figure 2: Historical spreads of the given pairs of futures contracts.

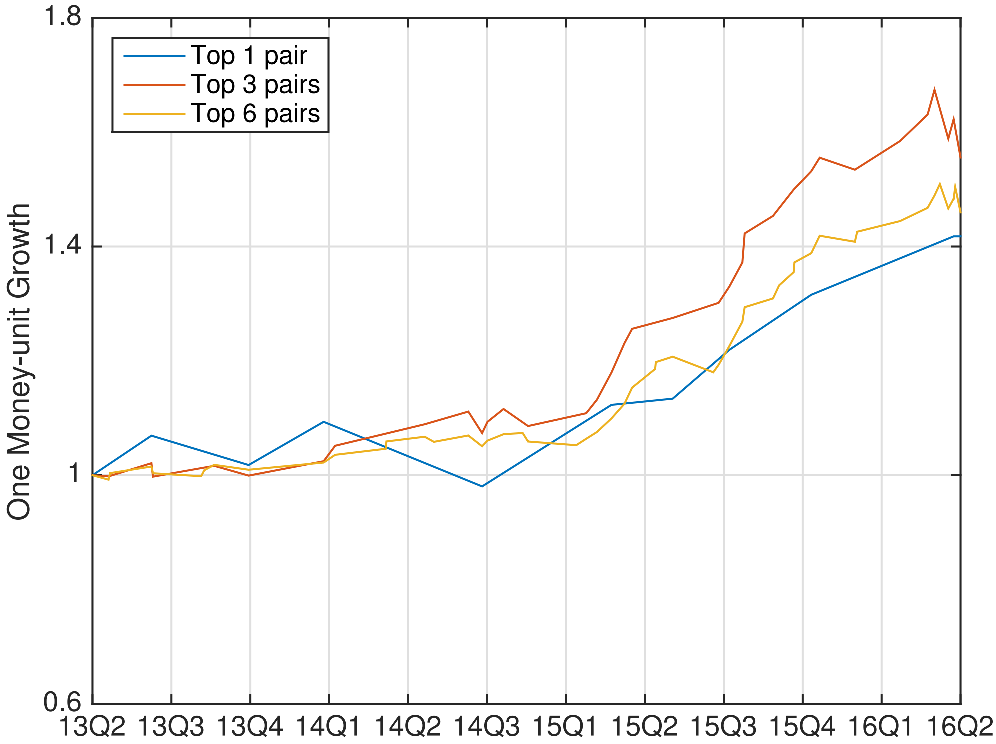

Figure 3: 1 money unit growths with top 1, 3 and 6 pairs in three year trading period.

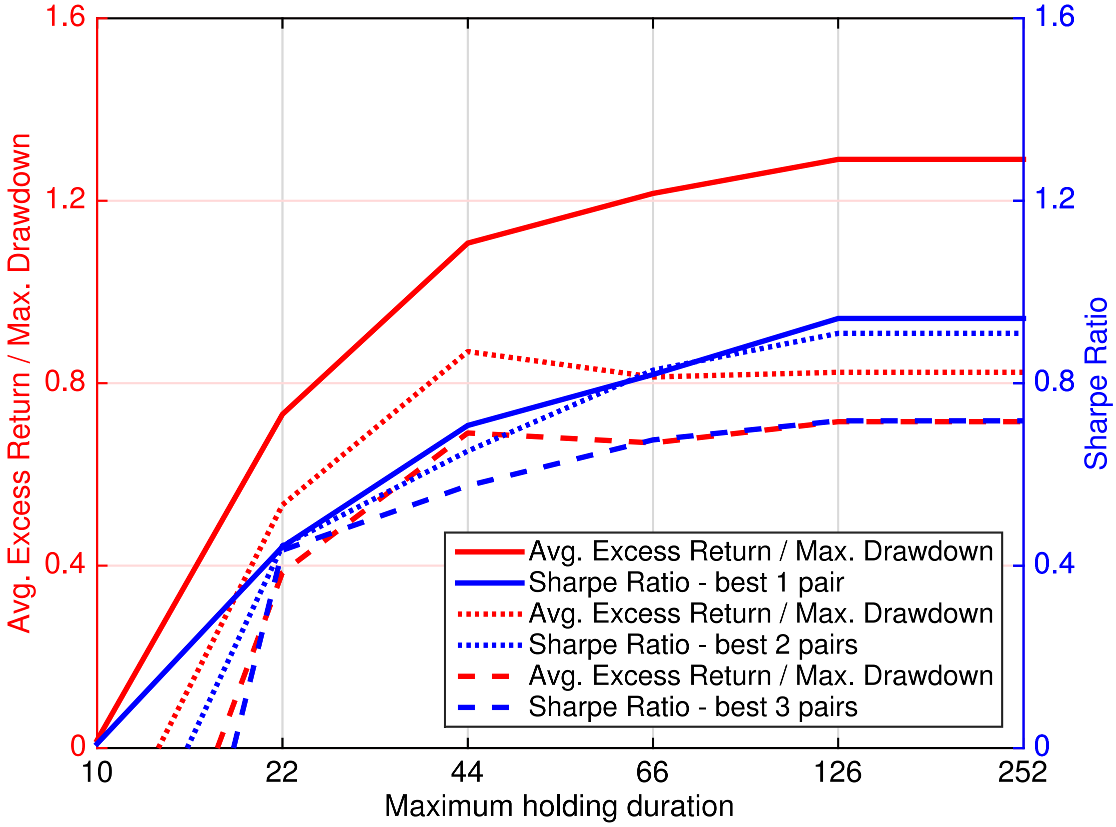

Figure 4: Risk adjusted returns (assuming the risk-free rate as 3%): Sharpe ratio and the average excess returns/Maximum drawdown with respect to different numbers of pairs in the portfolio and maximum holding durations.