Pairs Trading with Copulas

May 3, 2014

Wenjun Xie
Nanyang Business School
Nanyang Technological University, Singapore
xiew0008@e.ntu.edu.sg (+65) 82280370

Rong Qi Liew
School of Physical and Mathematical Science
Nanyang Technological University, Singapore
RQLIEW1@e.ntu.edu.sg (+65) 81129976

Yuan Wu
Nanyang Business School
Nanyang Technological University, Singapore
AYWU@ntu.edu.sg (+65) 97408898

Xi Zou
School of Physical and Mathematical Science
Nanyang Technological University, Singapore
zoux0007@e.ntu.edu.sg (+65) 90825582

Pairs Trading with Copulas

May 3, 2014

ABSTRACT

Pairs trading is a well-acknowledged speculative investment strategy, with the distance method the most commonly implemented such strategy. However, this approach, is able to fully describe the dependency structure between stocks only under the assumption of multivariate normal returns. In this research, we propose a new pairs trading strategy to generalize the conventional pairs trading strategy by using the copula modeling technique. Copulas allow separate estimation of the marginal distributions of stock returns and their joint dependency structure, and thus can provide sound estimation of the true joint distribution between stock returns. The overall empirical results verify the proposed strategy's ability to generate higher profits compared with the conventional distance method.

Keywords: Pairs Trading; Copulas; Dependency Structure; Correlation.

EFM Classifications: 350, 360, 380.

1. Introduction

Pairs trading has been an important speculative investment strategy for decades, and is widely applied in hedge funds and proprietary trading desks. The idea is to identify a pair of stocks whose prices have moved together historically, and subsequently, construct long/short positions when the two prices diverge.[1] The high profitability of this simple trading strategy mainly rests on two factors: the correct identification of high-quality pairs and the optimal modeling of the associations between the two stocks within the pair. The former ensures that this simple strategy is market-risk free, and the latter helps to decide the optimal trading positions.

In practice, the distance method is the most commonly used pairs trading strategy. It uses the distance between normalized prices as the criterion to judge the degree of mispricing between stocks. In terms of modeling the associations between the stocks, this distance measurement describes the linear dependence between the two stocks, but may overlook important non-linear associations. In the early finance literature, researchers typically adopted the assumption that stock returns are multivariate normal distributed, and thus emphasized the linear association between stock returns. Under this assumption, the distance method can without doubt fully describe the associations between the two stocks. However, the current consensus is that stocks returns are rarely joint normal (Cont, 2001), and thus non-linear associations, such as tail dependence, also play an important role in the modeling of stock returns (Ane and Kharoubi, 2003). Given this concern, the distance method no longer appears optimal, and may cause traders to miss important trading opportunities or engage in trades at non-optimal positions due to a loss of dependency information. In this research, we thus propose a new pairs trading strategy to

[1] A brief history and discussion of pairs trading can be found in Vidyamurthy (2004).

generalize the conventional distance method that uses the copula modeling technique to account for both the linear and non-linear associations between stocks, thereby providing more trading opportunities and higher profits.

In the academic arena, Gatev et al. (2006) were pioneers in investigating this pairs trading strategy, documenting significant excess returns in large-sample analysis. In their sample period, the annualized excess return was found to be as high as 11% for self-financing portfolios of pairs. These researchers also demonstrated that this excess return differs from previously documented reversal profits. Do and Faff (2010) subsequently tested the profitability of pairs trading using a more recent sample period, from 1962 to 2009, and identified a decreasing trend (0.33% mean excess return per month for 2003-09 versus 1.24% mean excess return per month for 1962-88). They put forward two possible explanations for the observed decrease in profitability: the efficient market hypothesis and arbitrage risk hypothesis. The simple pairs trading strategy has been public knowledge for decades, and its implementation procedure is easy. Hence, the efficient market hypothesis posits that the market has become so efficient that it can no longer produce excess returns. Arbitrage risk, in contrast, refers to deterrence from participating in arbitrage activities owing to the risk of further divergence (De Long et al., 1999). According to Do and Faff (2010), it is arbitrage risk that explains most of the decrease in profitability observed in their sample period. Accordingly, they provided alternative methods for improving pair quality, such as choosing pairs within more confined industries.

The pairs trading strategy analyzed in both of the aforementioned papers, namely, the distance method, is the most commonly applied. This simple relative-value strategy has the advantage of ease of implementation, and its effectiveness has been documented in different time periods and markets (Andrada et al., 2005; Perlin, 2009; Pizzutilo, 2013). Although the distance

method is generally recognized as a model-free approach, as it involves no explicit assumption concerning the distribution of stock returns, in this paper we argue that its intrinsic set-up may be invalid without a rigid assumption of multivariate normal returns, an assumption that may not be aligned with the current consensus.

As is widely known, modern portfolio theory is largely built on the assumption of multivariate normal returns. Two main aspects of this assumption account for its popularity. First, assuming a joint normal distribution implies that both individual margins and their linear combinations are normally distributed. Second, it requires only one correlation measurement to fully describe the association between any two variables, which greatly simplifies the models applied in risk management, portfolio diversification and hedging.

Although the multivariate normal nature of the distance method is not made explicit, the method enjoys the simplicities that the multivariate normal assumption confers, two in particular. First, it assumes a symmetric distribution of the spread between the normalized prices[2] of the two stocks within a pair. Second, it uses a single distance measurement, which can be seen as an alternative measurement of linear association, to describe all of the associations between the two stocks. These two simplicities, particularly the latter, may be invalid unless we also assume that the stock returns are jointly normally distributed.

However, it is now widely acknowledged that stock returns are rarely multivariate normal. For example, in a study of stock return distributions, Cont (2001) identified excess skewness and

[2] We refer to Gatev et al. (2006) for our definition of normalized prices, which can essentially be seen as a measurement of cumulative returns. The normalized prices for the first day of the formation period and trading period are set to one.

leptokurtosis in individual asset returns. In addition, he also found strong non-linear dependence between stock returns, which a single correlation measurement cannot describe. Ane and Kharoubi's (2003) findings support those of Cont (2001). They investigated the behavior of stock indices, and found non-linear associations such as tail dependence between them. It is thus clear that the original multivariate normal assumption is no longer appropriate in describing either marginal distributions or the joint dependency structure of stock returns. Therefore, the distance method, which enjoys the simplicities of the multivariate normal assumption, may fail to capture the real dependence structure between the two stocks, and thus miss important trading opportunities or enter trades at non-optimal positions.

Considering these limitations inherent in the conventional distance method and its rigid assumptions, we propose a copula-based approach to generalize the conventional distance method. The copula technique is an effective tool in modeling the joint distribution of random variables.[3] Sklar's well-known theorem (1959) laid the foundation for copulas, and provides the connection between individual marginal distributions and their joint distributions. In recent years, copulas have gradually begun to receive attention in the finance literature, particularly with regard to topics like modeling of stock returns. For example, Ane and Kharoubi (2003) use parametric copula to model the stock returns and Low et al. (2013) utilize a non-parametric copula approach to deal with portfolio management. The copula approach has two main advantages. First, it allows estimations of the marginal distributions of individual variables to be performed separately, which removes the assumption of normal margins from the beginning of the process. Second, after estimating the marginal distributions, it allows estimations of different dependency structures such as tail dependency, which may differ from the normal structure in

[3] Readers are referred to Nelson (2006) for a detailed definition and the properties of copulas.

many respects. It is clear that by combining these two advantages, a copula is a very effective tool in modeling both marginal distributions and the dependency structure between several variables. Therefore, in this research, we use it to capture the optimal dependence structure between the stocks, and trade accordingly.

The literature contains various studies that conduct preliminary trials of the application of copulas to pairs trading (Ferreira, 2008; Liew and Wu, 2013). However, their use of copulas for the direct modeling of non-stationary time-series stock prices lacks theoretical support, and are thus restricted to particular pairs and certain data structures. Besides, these studies do not provide large-sample results, therefore their given examples suffer from data snooping criticisms. Capitalizing on the inspiring ideas in these studies, this research introduces the first generalized copula-based approach that can be used for any selected pair in the market, and demonstrates its effectiveness using a large-sample analysis.

Specifically, this paper makes two main contributions to the literature. First, it proposes a new pairs trading strategy that uses the copula modeling technique. We use a copula to model the joint distribution of daily stock returns, and utilize the concept of a discrete step with a continuous state stochastic process for the trading strategy. Second, the paper reports the results of the first large-sample analysis, using utility industry data, to verify the advantages of copula use. This analysis is based on and strictly follows that of Gatev et al. (2006) to allow meaningful comparison. In general, the results show the proposed method to perform better than the conventional distance method. For example, a portfolio comprising the top five pairs generates up to a 9.36% annualized excess return with the proposed method, whereas the distance method generates insignificant excess returns.

The remainder of the paper is organized as follows. Section 2 provides the theoretical background for the copula framework. Section 3 describes our proposed copula-based pairs trading strategy. Section 4 then considers a one-pair-one-cycle example and reports the results of large-sample analysis using utility industry data from 2003 to 2012. Finally, Section 5 gives concluding remarks and suggests directions for future research.

2. Statistical Background

The recent downward trend in profitability exhibited by the conventional distance method raises the question of how we can improve the profitability of this simple relative-value strategy. On the one hand, it is clearly crucial to select high-quality pairs. Do and Faff (2010) proposed the selection of pairs within more confined industries or pairs with more zero crossings during the formation period, both of which can significantly improve the profitability of pairs trading. On the other hand, it is also worth thinking about how trades can be executed more accurately. In other words, it is important to think about how to describe the associations between the two stocks, and trade accordingly.

2.1 Distance Method

Gatev et al. (2006) and Do and Faff (2010) both adopted the distance method to test the profitability of pairs trading, and subsequently confirmed its effectiveness. The method's advantages are obvious. Using the spread between normalized prices is simple, and the entire procedure is easy to implement. However, as noted in the introduction of this paper, this simple strategy, may be insufficient to fully describe the associations between stocks, particularly in the case of non-linear associations. This loss of information on the dependency structure may result

in a reduced number of trading opportunities or trade entry at non-optimal positions, as further elaborated in this subsection.

We refer to Gatev et al. (2006) for the procedure used to implement the distance method. If stock X and stock Y are selected for pairs trading, and NP_t^X and NP_t^Y represent the normalized prices (cumulative returns) of X and Y on day t, then the spread Δ_t, calculated as $NP_t^X - NP_t^Y$, is used as the measurement to characterize mispricing in the distance method. The intrinsic assumption in pairs trading is that the two stocks share the same risk exposure. Thus, according to the law of one price,

$$E(\Delta_t) = E(NP_t^X) - E(NP_t^Y) = 0.$$

Practitioners usually use two standard deviations of Δ_t during the formation period as the trigger point for the trading period.

Modern portfolio theory is largely built on the assumption of multivariate normal returns. If we assume that NP_t^X and NP_t^Y follow a multivariate normal distribution, then Δ_t should also follow a normal distribution. The distribution of Δ_t should then be symmetrical with respect to zero and invariant conditional on different price levels of NP_t^X or NP_t^Y, and the commonly used trigger point of two standard deviations approximately reflects the 95% confidence interval of Δ_t. Thus, the distance method works perfectly well under the multivariate normal assumption.

However, owing to the existence of skewness and leptokurtosis, it is now widely acknowledged that stock returns rarely follow a multivariate normal distribution (Cont, 2001), which raises two concerns. First, the marginal distribution of a stock return is unlikely to be normal, and, second, the joint dependency structure can have various non-linear associations that a single distance measurement cannot describe. Accordingly, the distribution of Δ_t is unknown.

More accurately, it is most likely to be asymmetrical around zero, and its distribution conditional on normalized prices may no longer be robust. This non-stable conditional distribution may imply that Δ_t can represent different degrees of mispricing at different price levels even with the same numerical values. To better understand this point, consider the two following scenarios. In the first, the return of stock X is 0% and that of stock Y is 1%, whereas in the second the returns are 999% and 1000%, respectively. In both scenarios, the spreads are 1%, but they represent different degrees of mispricing, as it is natural to assume that 1% constitutes a much more significant difference in the first scenario than in the second. Therefore, it is essential to understand that the conditional distribution differs for every spread even when the numerical value of two spreads is identical, as they may occur at different levels of return.

Considering these issues, the distance method faces a number of difficulties under the non-normal data structure. First, the asymmetrical distribution of Δ_t violates the symmetrical assumption implied by the distance method. Second, it may be inappropriate to use fixed trigger points because the conditional standard deviation differs at different price levels. It would be better to trade according to their conditional standard deviation rather than their fixed average. The distance method cannot be implemented on the basis of the conditional standard deviation because it cannot fully capture the information on the dependency structure between the two stocks. To be more exact, the conditional standard deviation is invariant for multivariate normal distributions, and thus a single measurement is adequate. However, in the case of non-normal data, non-linear associations such as tail dependence should also be accounted for, and the resulting conditional distribution for the spread is constantly changing. Thus, fixed trigger points are not desirable.

2.2 Copula Framework

In resolving the difficulties that the distance method encounters, the main priority is to accurately capture both the marginal distributions and dependency structure between stocks, or the joint distribution of stock returns, and then decide the entry positions accordingly. Thus, we utilize a copula as a tool to model the joint distribution of random variables in developing our new measurements of mispricing.

Assume that stocks X and Y are candidates for pairs trading. Their daily closing prices and daily returns are recorded as P_t^X and P_t^Y, and R_t^X and R_t^Y, respectively.[4]

Sklar's theorem (1958) states that if $F(.)$ is an n-dimensional cumulative distribution function for random variables $X_1, X_2, \dots, X_n$ with continuous margins $F_1, F_2, \dots, F_n$, then there exists a copula function C such that

$$F(x_1,...,x_n) = C(F_1(x_1),...,F_n(x_n)).$$

Accordingly, we denote F_X and F_Y as the marginal distribution functions of R_t^X and R_t^Y, and H as their joint distribution function. Then, according to Sklar's theorem, there must exist a copula function C such that

$$H(r_t^X, r_t^Y) = C\left(F_X(r_t^X), F_Y(r_t^Y)\right).$$

[4] For comparability with Gatev et al. (2006), we follow them in using simple returns. Hence,

$$R_t^X = (P_t^X - P_{t-1}^X)/P_{t-1}^X$$
$$R_t^Y = (P_t^Y - P_{t-1}^Y)/P_{t-1}^Y. \quad (2.1)$$

After obtaining the joint distribution of daily returns for stock X and stock Y, we construct new measurements to denote the degree of mispricing utilizing the idea of conditional probabilities.

Definition If R_t^X and R_t^Y represent the random variables of the daily returns of stocks X and Y on day t, and the realizations of those returns on day t are r_t^X and r_t^Y, we have

$$\mathbf{MI_t^{X|Y}} = P(R_t^X < r_t^X | R_t^Y = r_t^Y)$$
$$\mathbf{MI_t^{Y|X}} = P(R_t^Y < r_t^Y | R_t^X = r_t^X). \qquad (2.2)$$

$\mathrm{MI}_{\mathrm{t}}^{\mathrm{X|Y}}$ and $\mathrm{MI}_{\mathrm{t}}^{\mathrm{Y|X}}$ are our defined mispricing indexes. There are several reasons to use conditional probability to characterize the degree of mispricing. First, it is a comparable measurement, which means that its numerical values are ordered and comparable. Second, it is a consistent measurement. As noted earlier, the spread measurement of the distance method can be of identical numerical values, even though the degree or significance of mispricing differs. In contrast, the conditional probabilities used here, which are based on the estimated optimal marginal distributions and dependency structure, are consistent over different price levels. In other words, only the same degree or level of mispricing is reflected by the same numerical values. Third, this measurement is easy to calculate and implement under the copula framework. The conditional probabilities can be easily obtained in one step by taking the first derivative of the copula function.

The mispricing indexes thus far defined have the following properties.

If $\mathrm{MI}_{\mathrm{X|Y}}^{\mathrm{t}} > 0.5$, then stock X is relatively overpriced with respect to stock Y on day t.

If $\mathrm{MI}_{\mathrm{X|Y}}^{\mathrm{t}} = 0.5$, then stock X is relatively fairly priced with respect to stock Y on day t.

If $\mathrm{MI}^{t}_{X|Y} < 0.5$, then stock X is relatively underpriced with respect to stock Y on day t.

If $\mathrm{MI}^{t}_{Y|X} > 0.5$, then stock Y is relatively overpriced with respect to stock X on day t.

If $\mathrm{MI}^{t}_{Y|X} = 0.5$, then stock Y is relatively fairly priced with respect to stock X on day t.

If $\mathrm{MI}^{t}_{Y|X} < 0.5$, then stock Y is relatively underpriced with respect to stock X on day t.

Conditional probability $\mathrm{MI}^{t}_{X|Y}$ indicates whether the return of X is considered high or low at time t, given the information on the return of Y on the same day and the historical relation between the two stocks' returns. If the value of $\mathrm{MI}^{t}_{X|Y}$ is equal to 0.5, r_t^X is neither too high nor too low given r_t^Y and their historical relation. Put simply, the historical data indicate that, on average, there are an equal number of observations of the return of X being larger or smaller than r_t^X if the return of stock Y is equal to r_t^Y. In this case, we can say that stock X is fairly priced relative to stock Y on day t. The concept of relative overpricing and underpricing can be similarly defined.

Given current realizations r_t^X and r_t^Y, if F_X and F_Y are the marginal distribution functions of R_t^X and R_t^Y and C is the copula connecting F_X and F_Y, we define $u = F_X(r_t^X)$ and $v = F_Y(r_t^Y)$, and have

$$\mathrm{MI}_t^{X|Y} = \frac{\partial C(u,v)}{\partial v} \text{ and } \mathrm{MI}_t^{Y|X} = \frac{\partial C(u,v)}{\partial u}.^{5} \qquad (2.3)$$

Defining the measurements of the degree of mispricing, $\mathrm{MI}_t^{X|Y}$ and $\mathrm{MI}_t^{Y|X}$, in this way helps to overcome the difficulties faced by the distance method. First, doing so has the advantage of estimating the marginal distributions and dependency structure separately, which eliminates

[5] The formal proof can be found in Nelson (2006).

concern over the method's inability to capture the asymmetrical distribution of the spread. Second, the newly defined measurements utilize the idea of conditional probabilities based on the optimal marginal distributions and estimated dependency structure. Hence, they contain information on both the linear and non-linear associations. Because of the greater precision of the joint distribution of two stocks, we strongly believe that our newly defined measurements are more consistent and robust than those in the conventional distance method. Therefore, our proposed trading strategy based on these measurements is able to capture more trading opportunities and greater profits, as demonstrated in Section 4.

3. Copula-Based Trading Strategy

3.1 Trading Strategy

Our proposed trading strategy comprises the following steps.

Formation Period

Daily return series R_t^X and R_t^Y are calculated during the formation period. Then, the best fitted marginal distributions of R_t^X and R_t^Y are estimated, after which the optimal dependency structure is estimated using different categories of copulas (Gumbel, Frank, Clayton, Normal and Student-T).[6] That with the highest likelihood value is chosen.

[6] We acknowledge that only a limited number of copulas are considered in this paper. Owing to the computational complexity involved, only the five most commonly used copulas were chosen for demonstration purposes. We believe that these five cover the majority of the dependency structures commonly encountered in practice. To improve accuracy, practitioners can incorporate more copulas in the fitting process if they wish.

Trading Period

The daily returns of stock X and stock Y are calculated during the trading period. $\mathrm{MI}_t^{\mathrm{X|Y}}$ and $\mathrm{MI}_t^{\mathrm{Y|X}}$ are also calculated using the estimated copulas from the formation period. Trading indicators are defined as $FlagX$ and $FlagY$ and are set to zero before commencement of the trading period. During the trading period, $(\mathrm{MI}_t^{\mathrm{X|Y}} - 0.5)$ and $(\mathrm{MI}_t^{\mathrm{Y|X}} - 0.5)$ are added to $FlagX$ and $FlagY$, respectively, on a daily basis. In addition, *D* is defined as the trigger point and *S* as the stop-loss position. The following are the four possible cases for an open position (assuming that no trades are open).

When FlagX reaches D, we short-sell stock X and buy stock Y in equal amounts.

When FlagX reaches -D, we short-sell stock Y and buy stock X in equal amounts.

When FlagY reaches D, we short-sell stock Y and buy stock X in equal amounts.

When FlagY reaches -D, we short-sell stock X and buy stock Y in equal amounts.

If trades are opened based on *FlagX*, then they are closed if *FlagX* returns to zero or reaches stop-loss position *S* or -*S*. If they are opened based on *FlagY*, then they are closed if *FlagY* returns to zero or reaches stop-loss position *S* or -*S*. After trades are closed, both *FlagX* and *FlagY* are reset to zero, and all opening trades are closed at the end of the trading period regardless of the values of *FlagX* and *FlagY*.

Note that *D* and *S* are pre-specified values, of which there are endless combinations. Practitioners can perform back-testing to choose the optimal trigger points. For illustration purposes, in this paper we choose to conduct analysis using $D = 0.6$ and $S = 2$. To avoid data snooping criticisms, we also perform additional analysis that is further explained in Section 4.

3.2 How Does the Trading Strategy Work?

Looking at the proposed trading strategy in closer detail, it utilizes the concept of a discrete step with a continuous state stochastic process. The cumulative distribution $F(X)$ of any continuous random variable X can be seen as a uniform random variable ranging from zero to one. Because $\mathrm{MI}_t^{\mathrm{X|Y}}$ and $\mathrm{MI}_t^{\mathrm{Y|X}}$ are conditional cumulative distributions, they also follow a uniform (0, 1). Then, $(\mathrm{MI}_t^{\mathrm{X|Y}} - 0.5)$ and $(\mathrm{MI}_t^{\mathrm{Y|X}} - 0.5)$ follow a uniform (-0.5, 0.5) distribution. In other words, our trading signals, $FlagX$ and $FlagY$, which are accumulations of the daily $(\mathrm{MI}_t^{\mathrm{X|Y}} - 0.5)$ and $(\mathrm{MI}_t^{\mathrm{Y|X}} - 0.5)$, are the sums of a series of uniform random variables between -0.5 and 0.5 provided that we do not consider the correlations between them and view the situation as akin to a pure random walk.

Mathematically, if we define a $Flag$ series, where $Flag_t = Flag_{t-1} + e_t$, then e_t follows an i.i.d uniform distribution from -0.5 to 0.5, and $Flag_0 = 0$. In this paper, both $FlagX$ and $FlagY$ have the same characteristics as a $Flag$ series within a trading cycle[7] because of the assumption that R_i^X are independent of R_j^X and R_i^Y are independent of R_j^Y for $i \neq j$. This sum of conditional probabilities minus its mean values for every trading day can be seen as a similar measurement to the degree of cumulative mispricing. It adds up the minor effect from each day and provides a cumulative indicator. The advantages of using probability measurements under the copula framework have already been discussed. The most important advantage is that doing so allows more robust and consistent measurements over time.

[7] The trading cycle refers to the period from the first day of the trading period or closing date of the last trade to the completion of the next trade.

We now investigate the properties of this $Flag$ time series. If we assume that e_t follows a uniform distribution from -0.5 to 0.5 without time-series correlation, the $Flag$ series becomes equivalent to a pure random walk. In this case, there is no arbitrage opportunity because there is an equal chance of further divergence or convergence. Although the expected value of $Flag_t$ is always zero, it is not a stationary time series. It can either move strictly up or down or fluctuate within a certain range. However, pairs trading strategies are built on the assumption of a mean-reverting property, which means that $Flag_t$ has a tendency to converge to zero when it is far from zero. In other words, e_t is not uncorrelated. In contrast to the mean-reverting property, Do and Faff (2010) also mention arbitrage risk, which refers to the tendency toward further divergence when $Flag_t$ is far from zero in this context. Again, e_t is also not uncorrelated in this case. Accordingly, we identify the three following mechanisms that are relevant to the $Flag_t$ series.

Baseline Mechanism (Random Walk): with no time-series correlation on the residual term e_t with lagged $Flag_{t-1}$, $Flag_t$ should be a purely random walk. There is no arbitrage opportunity in this case. On average, no pairs trading strategy, whether the distance method or the proposed copula method, will work in this case.

Convergent Mechanism (Mean-Reverting): in this case, there is a negative correlation between residual term e_t and $Flag_{t-1}$, which ensures that $Flag_t$ has the tendency to converge when it is away from zero. Pairs trading strategies generally make profits in this case.

Divergent Mechanism (Arbitrage Risk): in this case, there is a positive correlation between residual term e_t and $Flag_{t-1}$, which ensures that $Flag_t$ has the tendency to

diverge when it is away from zero. Pairs trading strategies generally result in a loss in this case.

In the real market, $Flag_t$ alternates among the three mechanisms, resulting in both profits and losses in pairs trading. However, as long as the convergent mechanism dominates the divergent mechanism, pairs trading strategies work well and generate profits. In the case of the divergent mechanism, it is important to impose a stop-loss position and then recount.

4. Empirical Analysis

This section provides empirical evidence of the proposed copula method's performance relative to that of the conventional distance method. It is divided into two subsections. In the first, we present a particular stock pair as an illustrative example and provide detailed analysis of this pair over one formation and one trading period. The purpose is to distinguish between the proposed copula and conventional distance methods, and illustrate the advantages of the former. The second subsection then presents the results of cross-sectional large-sample analysis spread across 17 time periods and based on the entire utility industry dataset. This subsection also provides a comparison between the results obtained from the two methods and addresses various concerns related to transaction costs and data snooping.

4.1 One-Pair-One-Cycle Illustration

The stock pair considered in this subsection is Brookdale Senior Living Inc. and Emeritus Corporation (BKD-ESC). This highly correlated stock pair is listed in the healthcare sector of www.pairslog.com, and is also cited as an example in Liew and Wu (2013). Both companies are listed on the New York Stock Exchange (NYSE) and are very similar in terms of their business

operations. Both are in the long-term care facilities industry of the healthcare sector. The time periods considered in this example are January 2 to December 30, 2008 (formation period) and December 31, 2008 to July 1, 2009 (trading period). The stock pair is verified to be highly correlated, as seen in Figure 1(a). The stocks' normalized prices are very similar and share close upward and downward movements during the formation period. Table 2 shows the pair to have a high correlation coefficient value of 0.942.

Because of the close relation between the normalized prices of BKD and ESC, the spread during the formation period (shown in Figure 1(b)) fluctuates around zero most of the time and demonstrates a tendency to revert to zero. However, the relation changes dramatically during the trading period because the normalized prices drift farther apart over the period and the spread moves away from its original mean value of zero. This shift suggests a potential change in the structure of the stock pair or the conventional distance strategy's inability to fully capture that structure during the formation period.

We carry out a comparative analysis using the copula approach. Figure 2(a) illustrates the dataset from the formation period, with the copulas fitted to it. Panel A of Table 1 displays the Schwarz information criterion (SIC), Akaike information criterion (AIC) and Hannan-Quinn information criterion (HQIC) test values, all common statistical tools applied to measure goodness-of-fit. Figure 2(a) shows six scatter plots that compare the formation period data with the five fitted copulas. The Student-t copula best captures the general structure, particularly the upper and lower tails of the formation period data, which is further verified by the test values reported in Table 1. Figure 2(b) and Panel B of Table 1 are similar to Figure 2(a) and Panel A, except that the dataset considered is taken from the trading period. Close observation of Figure 2(b) and the test values in Panel B of Table 1 suggests that the Student-t remains the best fitting

copula for the dataset. Hence, the dependency structure of the stock returns does not change between the formation period and the trading period. However, the distance method cannot capture this optimal dependency structure, and cannot identify the optimal trading opportunities.

Overall, the results in Table 2 show that the copula approach undisputedly performs much better than the distance strategy, regardless of the one-day wait.[8] Assuming $10,000 in initial capital, the distance method results in a -$592 loss, whereas the proposed copula method results in an $847 gain in the case without a one-day wait. In the one-day-wait case, the distance method generates a loss of up to -$1526, whereas the copula method produces a gain of $1060. In this example, the copula approach is clearly more profitable and generates greater excess returns. Its consistent superiority over the distance approach is further verified in an industry-wide dataset in the next subsection.

4.2 Utility Industry Data Analysis

4.2.1 Data

Our daily stock price data are extracted from the Center for Research in Security Prices (CRSP) database. The sample period is January 1, 2003 to December 31, 2012. All of the stocks selected are publicly traded on the NYSE, AMEX or NASDAQ. To form the initial dataset, we filtered the stocks to identify those with a Standard Industrial Classification (SIC) code beginning with 49 (Utility). We subsequently dropped penny stocks and stocks with missing values, to give a final sample comprising 89 stocks. We chose the utility industry because Do

[8] Gatev et al. (2006) recommended investigating the one-day-wait case to deal with the bid-ask spread bounce, a point that is further illustrated in Section 4.2.

and Faff (2010) raised the point that choosing pairs within a more confined industry increases the profitability of pairs trading, and stocks in the utility industry have better co-movements.

4.2.2 Excess Return Computation

As previously noted, Gatev et al. (2006) identify two return measurements: the return on committed capital and the fully invested return. The former considers the return on actual employed capital, which means that all of the money prepared for potential trades is considered in the principal amount even if the position for the particular stock pair is not open. In contrast, the fully invested return considers only the money currently being traded as the principal amount to calculate the return. The first measurement is clearly much more conservative and practical because the opportunity cost of the money set aside for potential trades is taken into account. As the use of such a conservative measurement increases the credibility of our results, we adopt the return on committed capital in this paper. The formula for calculating the excess return is exactly the same as that defined in Gatev et al. (2006):

$$r_{P,t} = \frac{\sum_{i \in P} w_{i,t} r_{i,t}}{\sum_{i \in P} w_{i,t}} \tag{4.1}$$

$$w_{i,t} = w_{i,t-1}(1 + r_{i.t-1}) = (1 + r_{i.1}) \dots (1 + r_{i.t-1}), \tag{4.2}$$

where r defines returns and w defines weights, and daily returns are compounded to obtain monthly returns.

4.2.3 Main Results

We follow the analysis conducted by Gatev et al. (2006) to provide a comparable evaluation. We also produce the same tables to allow direct comparison between the proposed

copula method and the distance method. A number of issues, including the sources of gains and the risks of pairs trading, which are discussed in Gatev et al. (2006), are also applicable to this paper and are thus omitted from discussion here. Our main objective is to compare the distance trading strategy with our proposed copula trading strategy in terms of profitability.

After cleaning up the data as described in 4.2.1, we ran them across 17 time periods, both formation and trading periods. A formation period is defined as a 12-month period, taken to be 252 days.[9] Historical data from the formation period are used for pairs formation and estimation of the distributions and parameters. A trading period is defined as a six-months period, taken to be 126 days. The profits and returns generated by the two strategies during the trading period are illustrated later in this section.

During each formation period, every possible combination of the 89 stocks is considered, and the sum of squared deviations between the two normalized price series is calculated. The top 5 stock pairs chosen are those with the lowest sum of squared deviations between the two series. The same also applies to the top 20 and top 101-120 pairs.

This pair-formation approach is in accordance with that adopted in Gatev et al. (2006). Applying the minimum-distance criterion ensures that the stock pairs chosen have relatively smaller values for their sum of squared deviations than the other possible combinations. In other words, the stock pairs selected share a close relationship during the formation period, with contemporaneously similar upward and downward movements. This approach is found to best approximate the way in which actual traders choose pairs (Gatev et al., 2006).

[9] Although the actual number of trading days for each month varies, we assume 21 trading days for each. Hence, "12-month formation period" and "six-month trading period" may not be strictly accurate.

Table 3 presents the excess returns of the pair portfolios generated by the distance and copula strategies. Panel A shows the results of the two trading strategies when the positions are opened at the end of the day on which prices diverge and closed at the end of the day on which they converge. The values of average excess returns, t-statistics and lower percentage of observations with negative excess returns clearly show the copula strategy to outperform the conventional distance strategy. The excess returns produced by this strategy are consistently higher than those generated by the distance strategy for the top 5, top 20 and 101-120 stock pairs. Moreover, the values of the t-statistics indicate that the excess returns are positive and significant for the proposed copula method but not for the distance method (for all three sets of portfolios tested). In addition, the percentage of observations with negative excess returns is smaller for the copula than distance strategy across-the-board, that is, for the top 5, top 20 and top 101-120 stock pairs.

Despite the positive results shown in Panel A, we acknowledge the concern regarding the bid-ask spread (Jegadeesh, 1990; Jegadeesh and Titman, 1995, Conrad and Kaul, 1989), which is mentioned in Gatev et al. (2006) as a practical issue faced in trading. To alleviate this concern, we also implemented one-day waiting, as recommended by Gatev et al. (2006). Panel B of Table 3 reports the trading strategy results when the positions of each stock pair are opened on the day following price divergence and closed on the day following convergence. Although the average excess returns are affected by the one-day wait for both strategies, those generated by the copula strategy are still generally better than those produced by the distance strategy, thus further strengthening our confidence in the proposed strategy.

Table 4 summarizes the trading statistics and composition of the pair portfolios for both the distance and copula strategies. From the average number of pairs traded per six-month period,

we observe that the copula strategy identifies trades for all of the top pairs selected. In addition, the average number of round-trip trades per pair is higher across the board for this strategy. Although the proposed strategy is clearly able to identify more trading opportunities than the traditional distance strategy, it requires the pairs to be open for a greater number of months. However, the proposed copula strategy also brings about a smaller standard deviation in the time open per pair in months. Accordingly, the strategy affords greater certainty about the amount of time that traders will be in an open position.

The foregoing cross-sectional analysis of U.S. utility industry stocks using both strategies provides clear evidence of the proposed copula strategy's superior performance relative to the conventional distance strategy. The copula strategy exhibits very satisfactory excess returns even after the application of one-day waiting. In addition, we have also verified that this strategy is able to identify more trading opportunities, an extremely important finding for the market, as every potential market signal constitutes an opportunity to profit.

For robustness checks, we also conduct two more tests. First, we altered the trigger points for the distance method in search of the optimal profits. The results of this optimal analysis differed little from the results reported in Table 3, and we thus concluded that the results of the proposed copula method with trigger points (0.6/2) were better than the optimal result from the distance method. It should also be acknowledged that the optimal trigger points for each stock pair may differ, and they may also differ in different time periods. It is possible that other sets of trigger points would produce better results for the copula method. The second test is conducted to avoid concern that this set of trigger points was good only for a certain period. We conducted subsample analysis, and observed consistent results over time. Figure 3 shows the cumulative returns along the 17 trading periods for the top five pairs portfolios using the copula strategy. We

observe a steady increase in the cumulative returns, with especially higher returns during the financial crisis period. This is reasonable, as the market is more volatile during crisis periods. Overall, we believe that the results of these two additional analyses should alleviate data snooping concerns about the proposed method.

4.2.4 Transactions Costs

There are two main sources of transaction costs: bid-ask spread costs and short-selling costs. Getczy et al. (2002) noted that the short-selling costs of the rebate rate on short sales are low for large traders, just 4-15 basis points (bp) per year. Therefore, our main focus for transaction costs is bid-ask spread costs. Keim and Madhaven (1997) found the average effective spread for stocks in the CRSP database in 1991 to be 37 bp, and more recent NYSE research showed the average effective bid-ask spread in 2001 to range from 14 to 18 bp for NYSE and NASDAQ common stocks. This finding seems reasonable, as the liquidity of the stock market has increased over the years, and thus the current average bid-ask spread may be even lower. Our six-month excess return, with one-day waiting for the top 5 pairs, is 264 bp (44 bp per month). Considering the average of 6.8 transactions per six-month period stated in Table 4, the transactions costs would be around 110 bp. Hence, the excess returns obtained using the proposed copula strategy, and assuming one-day waiting, remain positive and significant. However, for other portfolios, and particularly for the results produced by the distance strategy, the excess returns generated may be insufficient to cover the transaction costs.

The effectiveness of pairs trading has been plagued by concerns over transaction costs, particularly in the case of the commonly applied distance method, as the excess returns it generates may be insufficient to cover those costs. The discussion in this subsection thus once

again confirms the effectiveness, and practical importance, of the proposed copula strategy, which generates positive and significant excess returns for the top 5 pairs even after considering transaction costs.

5. Conclusion

Pairs trading is a well-acknowledged technique in the financial industry, and its effectiveness has been consistently documented over time and across markets. However, a marked decline in the profitability of this simple strategy has been noted in recent years. In this paper, we thus propose a copula framework for improving the profitability of pair trading, and demonstrate its effectiveness.

The conventional distance method considers only one simple spread measurement between normalized prices, whose distribution changes over time if the stock prices do not follow a bivariate normal structure. Thus, it is essential to generalize the pairs trading framework to accommodate a more diverse data structure and, in turn, to discover new tools for accurately capturing the dependency structure of data. The intrinsic dependency structure is more robust and less subject to change, and measurements built upon it are thus more reliable and consistent over time. Accordingly, we use copulas to estimate both the marginal distributions and dependency structure between stocks, and construct mispricing indexes based on conditional probability measurements.

The empirical analysis in this paper comprises both a one-pair-one-cycle example and large-sample analysis of 10 years of utility industry data. The results demonstrate the superiority of the proposed copula framework over the conventional distance method. In the one-pair-one-cycle example, the proposed framework is more proficient in capturing the dependency structure,

as few changes are recorded between the formation and trading periods, which also supports our claim that this structure is more robust and less susceptible to change over time. The large-sample analysis further confirms the superiority of the proposed method. Over the long data period, the top 5 pairs identified indicate that the distance method produces insignificant excess returns, in contrast to the proposed copula method, which produces a 3.6% annualized excess return even after accounting for a one-day wait. In addition, the proposed method generally also exhibits better performance than the distance method for the top 20 and 101-120 stock pairs even though some returns are insignificant in the one-day-waiting scenario. Overall, we have demonstrated that the proposed copula method better captures the dependency structure and provides more trading opportunities with higher excess returns and profits than the traditional approach.

Despite the superiority of the results obtained from copulas, there are certainly areas that could be further improved and developed. For example, the stock pairs selected for both methods in this paper are identified according to minimum distance criteria, an approach that focuses only on the linear association between two stocks. Accordingly, a number of good candidates with strong non-linear associations may have been neglected. Because copulas are able to capture both linear and non-linear relations between random variables, it is possible that the proposed method might have performed even better if we had selected the stock pairs by incorporating non-linear association measurements such as Kendall's tau and Spearman's rho. Moreover, copulas can capture the dependency structure of more than two stocks. Hence, it should be possible to construct a multi-dimensional pairs trading framework. By incorporating more information from additional stocks, such a framework would likely be more accurate in

determining the relatively undervalued or overvalued positions of stocks, thereby yielding higher profits.

Finally, it should be noted that the proposed trading strategy could also be applied to high-frequency trading. Although this paper considers daily trading for illustration purposes, the overall procedure is not dependent on the time frame used, and the strategy is also applicable to a smaller time interval, such as an hour. However, owing to data constraints, we were unable to test the performance of the proposed strategy in this scenario. This and the other aforementioned unresolved issues constitute interesting research topics for further investigation. We believe that investigation of these issues will help to improve the proposed strategy and provide a broader perspective on the use of copulas in pairs trading.

References

Andrade, S., V. Di Pietro, and M. Seasholes (2005). Understanding the profitability of pairs trading. UC Berkeley Working Paper.

Ane, T., and C. Kharoubi (2003). Dependence structure and risk measure. *Journal of Business* 76(3): 411-438.

Conrad, J., and G. Kaul (1989). Mean reversion in short-horizon expected returns. *Review of Financial Studies* 2: 225-240.

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance* 1: 223-236.

De Long, J., A. Shleifer, L. Summers, and R. Waldmann (1990). Noise trader risk in financial markets. *Journal of Political Economy* 98(4):703-738.

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

Ferreira, L. (2008). New tools for spread trading. *Futures* 37(12): 38-41.

Gatev, E., W.N. Goetzmann, and K.G. Rouwenhorst (2006). Pairs trading: Performance of a relative-value arbitrage rule. *Review of Financial Studies* 19(3):797-827.

Getczy, C., D. Musto, and A. Reed (2002). Stocks are special too: An analysis of the equity lending market. *Journal of Financial Economics* 66: 241-269.

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *The Journal of Finance* 45: 881-898.

Jegadeesh, N., and S. Titman (1995). Overreaction, delayed reaction, and contrarian profits. *Review of Financial Studies* 8:e 973-993.

Keim, D., and A. Madhaven (1997). Transaction costs and investment style: An inter-exchange analysis of institutional equity trades. *Journal of Financial Economics* 46: 265-292.

Liew, R.Q., and Y. Wu (2013). Pairs trading: A copula approach. *Journal of Derivatives & Hedge Funds* 19: 12-30.

Low, R., J. Alcock, T. Brailsford, and R. Faff (2013). Canonical Vine Copulas in the context of Modern Portfolio Management: Are they Worth it? *Journal of Banking and Finance* 37 (8): 3085-3099.

Nelson, R.B. (2006). *An Introduction to Copulas* (2nd Ed.). New York: Springer.

Perlin, M.S. (2009). Evaluation of pairs-trading strategy at the Brazilian financial market. *Journal of Derivatives & Hedge Funds* 15(2): 122-136.

Pizzutilo, F. (2013). A note on the effectiveness of pairs trading for individual investors. *International Journal of Economics & Financial Issues* 3(3): 764-771.

Sklar, A. (1959). Fonctions de répartition á n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Universite de Paris* 8:229-231.

Vidyamurthy, G. (2004). *Pairs Trading: Quantitative Methods and Analysis*. Hoboken, NJ: John Wiley & Sons.

Figure 1

Distance strategy vs. Copula Strategy

Pair: Brookdale Senior Living Inc. and Emeritus Corporation (BKD-ESC)

Formation Period: January 2 to December 30, 2008

Trading Period: December 31, 2008 to July 1, 2009

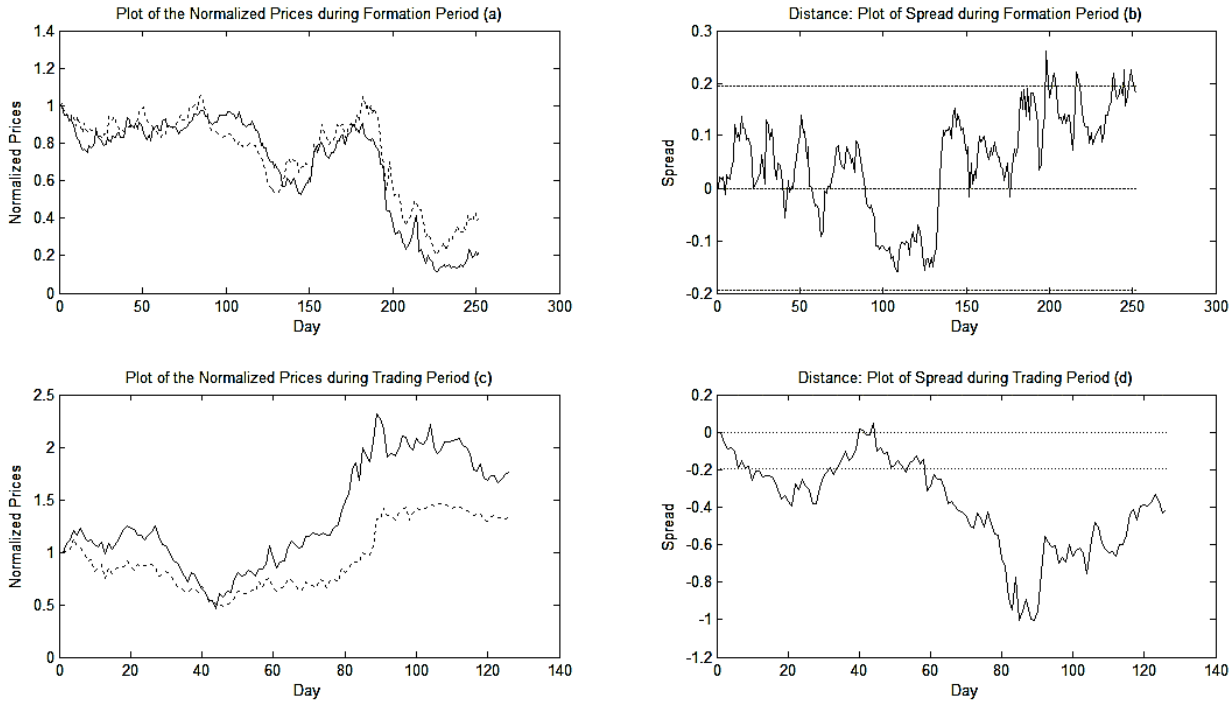


Figure 2

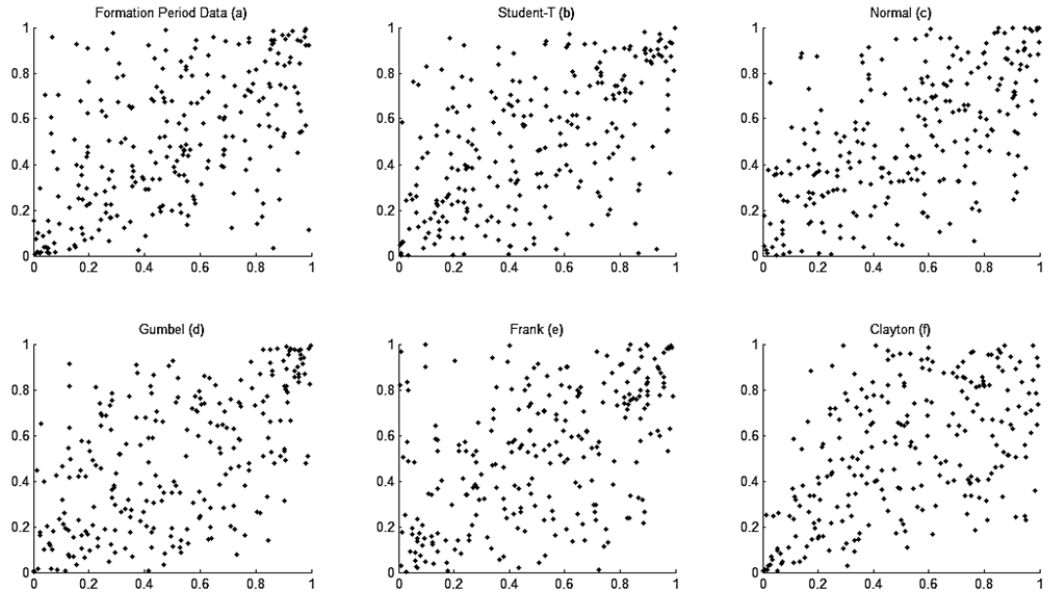
Comparison of dataset against fitted copulas

Pair: Brookdale Senior Living Inc. and Emeritus Corporation (BKD-ESC)

Formation Period: January 2 to December 30, 2008

Trading Period: December 31, 2008 to July 1, 2009

(a)



(b)

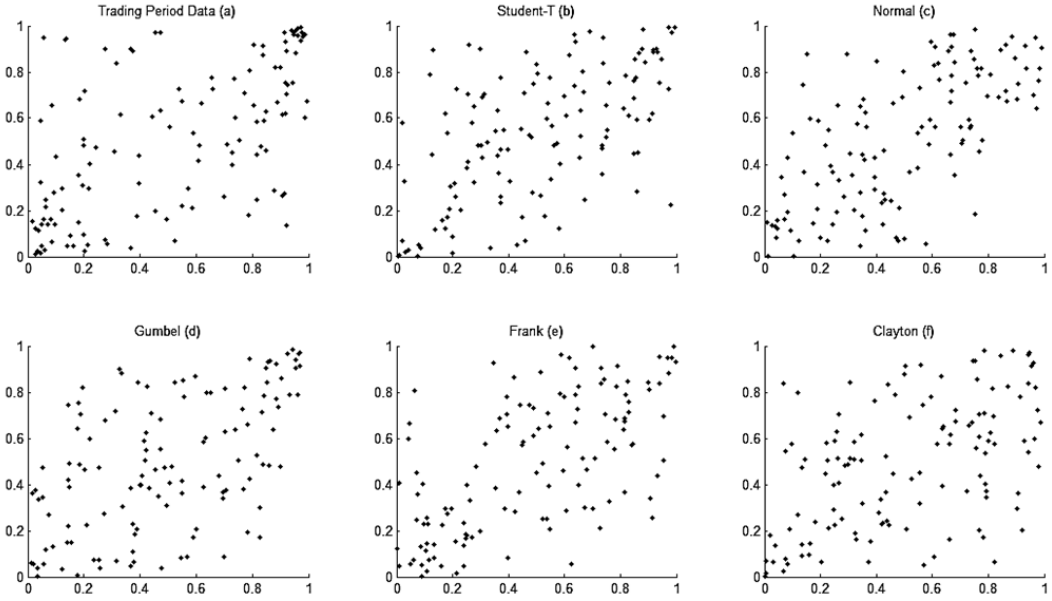


Figure 3

Cumulative returns for the top five pairs portfolios (without one-day waiting)

Sample Period: 2003-2012

Number of Trading Periods: 17

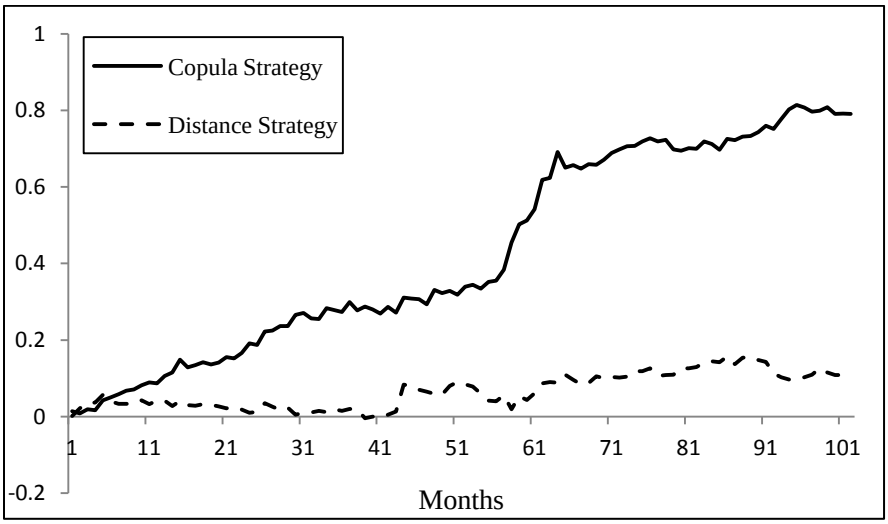


Table 1

SIC, AIC and HQIC test values of copulas during formation and trading periods

Pair: Brookdale Senior Living Inc. and Emeritus Corporation (BKD-ESC)

Formation Period: January 2 to December 30, 2008

Trading Period: December 31, 2008 to July 1, 2009

SIC = Schwarz information criterion; AIC = Akaike information criterion; HQIC = Hannan-Quinn information criterion.

	SIC (-)	***AIC (-)***	***HQIC (-)***
Panel A: Formation Period			
Student-T	103.730	110.732	107.943
Clayton	98.906	102.415	101.012
Normal	93.926	97.435	96.032
Frank	87.236	90.745	89.342
Gumbel	85.985	89.495	88.092
Panel B: Trading Period			
Student-T	62.821	68.396	66.189
Gumbel	64.041	66.845	65.725
Normal	57.436	60.240	59.120
Frank	53.658	56.462	55.342
Clayton	50.397	53.201	52.081

Table 2

Results of pairs trading strategies on stock pair BKD-ESC

This table summarizes the trading statistics of BKD-ESC in the December 2008 to July 2009 trading cycle. We followed Gatev et al. (2006) in applying the distance method and used two historical standard deviations of price differences as the trigger points. We subsequently implemented the copula method using the procedure proposed in this paper. Panel A presents the results with no waiting time, and Panel B the results with one-day waiting.

	Distance Strategy	**Copula Strategy**
Panel A: Trading strategies implemented as intended		
Total Profit (Capital: $10,000)	$-592.99	$847.27
Total Number of Transactions	3	8
Average Monthly Excess Returns	-0.009	0.014
Panel B: Trading strategies implemented with one day waiting		
Total Profit (Capital: $10,000)	$-1526.23	$1060.01
Total Number of Transactions	3	7
Average Monthly Excess Returns	-0.011	0.017

*The correlation between BKD and ESC during the formation period is 0.942.

Table 3

Excess returns of pairs trading strategies (distance method and copula method)

This table reports the summary statistics of monthly excess returns for the two pairs trading strategies. The sample consists of all stocks other than penny stocks belonging to the utility industry (SIC: 49) and with non-missing data from January 1, 2003 to December 31, 2012, for a total of 89 stocks. We follow Gatev et al. (2006) in applying the distance method, and use two historical standard deviations of price differences as the trigger points. The copula method was implemented in accordance with the procedure proposed in this paper. Panel A presents the results with no waiting, and Panel B with one-day waiting (*represents significance at the 5% level).

Table 3 - CONTINUED

	Distance Strategy			Copula Strategy		
Pair portfolio	Top 5	Top 20	101 - 120	Top 5	Top 20	101 - 120
A. Excess return distribution (no waiting)						
Average excess return (committed)	0.0011	0.0008	0.0006	0.0078*	0.0022*	0.0022*
t-Statistic	0.775	1.100	0.846	4.184	2.497	1.870
Excess return distribution						
Median	-0.0000	0.0002	0.0007	0.0062	0.0024	0.0015
Standard deviation	0.0138	0.0073	0.0077	0.0187	0.0089	0.0116
Skewness	1.1047	0.0309	0.1671	1.082	-0.0991	0.3133
Kurtosis	8.3501	3.3476	4.0179	5.7403	3.8341	4.6376
Minimum	-0.0372	-0.0187	-0.0233	-0.0415	-0.0248	-0.0355
Maximum	0.0700	0.0214	0.0255	0.0769	0.0293	0.0419
Observations with excess returns < 0 (%)	50%	49%	46%	35%	39%	44%
B. Excess return distribution (one-day waiting)						
Average excess return (committed)	0.0008	0.0004	0.0006	0.0044*	0.0006	0.0001
t-Statistic	0.5877	0.5822	0.7758	3.0500	0.6640	0.1082
Excess return distribution						
Median	0.0000	-0.0006	0.0006	0.0045	0.0004	0.0005
Standard deviation	0.0144	0.0063	0.0076	0.0147	0.0088	0.0123
Skewness	2.0471	0.2664	0.0009	-0.3860	-0.2221	0.4018
Kurtosis	12.850	3.152	4.392	6.501	5.293	4.996
Minimum	-0.0342	-0.0156	-0.0223	-0.0599	-0.0264	-0.0342
Maximum	0.0814	0.0187	0.0260	0.0542	0.0339	0.0449
Observations with excess returns < 0 (%)	48%	52%	45%	34%	47%	49%

Table 4

Trading statistics and composition of pair portfolio

This table reports the summary statistics of monthly excess returns for the two pairs trading strategies, i.e., the distance method and copula method. The sample comprises all stocks other than penny stocks belonging to the utility industry (SIC: 49) and with non-missing data from January 1, 2003 to December 31, 2012, for a total of 89 stocks. We followed Gatev et al. (2006) in applying the distance method and used two historical standard deviations of price differences as the trigger points. We implemented the copula method in accordance with the procedure proposed in this paper. Panel A presents the results with no waiting, and Panel B with one-day waiting.

	Distance Strategy			**Copula Strategy**		
Pair portfolio	Top 5	Top 20	101 - 120	Top 5	Top 20	101 - 120
A. Trading statistics (no waiting)						
Average price deviation trigger for opening pairs	0.0416	0.0495	0.0740			
Average number of pairs traded per six-month period	4.82	18.52	17.70	5.00	20.00	20.00
Average number of round-trip trades per pair	1.74	1.54	1.26	6.83	6.60	6.67
Standard deviation of number of rounds trips per pair	1.13	1.03	0.82	2.27	2.11	2.14
Average time pairs are open in months	2.14	1.95	2.00	4.40	4.45	4.51
Standard deviation of time open, per pair, in months	1.49	1.38	1.41	0.52	0.49	0.50
B. Trading statistics (one-day waiting)						
Average price deviation trigger for opening pairs	0.0416	0.0495	0.0740			
Average number of pairs traded per six-month period	4.82	18.52	17.70	5.00	20.00	20.00
Average number of round-trip trades per pair	1.74	1.54	1.26	6.80	6.55	6.62
Standard deviation of number of rounds trips per pair	1.13	1.03	0.82	2.24	2.08	2.13
Average time pairs are open in months	2.12	1.93	1.99	4.36	4.41	4.47
Standard deviation of time open, per pair, in months	1.48	1.37	1.40	0.51	0.49	0.50