Statistical Arbitrage and Robust Tests for Cointegration

Thomas A. Hanson

Kent State University

Joshua R. Hall

Kent State University

September 25, 2012

Correspondence concerning this article should be addressed to Thomas A. Hanson, Kent State University, College of Business Administration, Department of Finance, P.O. Box 5190, Kent, Ohio 44242. Phone: (330) 672 – 1129. Fax: (330) 672 – 9806. E-mail: thanson3@kent.edu

Statistical Arbitrage Using Robust Tests for Cointegration

Abstract

One application of cointegration tests is screening candidate stocks for the investment strategy known as statistical arbitrage. This paper develops two robust tests for cointegration by using rank-based and least absolute deviation regression to modify the seminal Engle-Granger test. Critical values are generated and power calculated for various error distributions. Finally, the tests are utilized in a simple pairs trading strategy and backtested on daily data from 2001 to 2010. The rank-based cointegration test has superior qualities in that context, suggesting one application of this new statistical test.

Keywords: Statistical arbitrage, pairs trading, cointegration, robust statistics, rank-based statistics, weighted Wilcoxon regression, least absolute deviation

Statistical Arbitrage Using Robust Tests for Cointegration

1.0 Introduction

Statistical arbitrage is a heavily quantitative strategy that remains popular with traders today as a way to make money on relative market moves, rather than absolute price bets through fundamental analysis (Gatev, Goetzmann, & Rouwenhorst, 2006; Vidyamurthy, 2004). Determining candidate stocks for implementation of the strategy requires some statistical measure to identify profitable opportunities. The statistical concept of cointegration (Engle & Granger, 1987) provides one such tool. However, extant cointegration tests are quite sensitive to the assumption of normality of the time series and have notoriously low power. This paper seeks to develop robust cointegration tests that overcomes these problems and to apply those tests to a simple pairs trading strategy.

The roots of the pairs trading strategy can be traced to the long-short mutual fund investment strategies of A. Winslow Jones in the 1950s. His idea was to create a hedged portfolio of long and short positions to eliminate market risk; the return on the portfolio stemmed from the relative, not absolute, performance of the positions. By purchasing undervalued stocks and selling short overvalued stocks, this general strategy is a form of contrarian trading that amounts to betting that prices will converge, or to put the same thing another way, that the spread between the two assets will shrink.

While the origins of the strategy involved fundamental analysis to identify over- and under-valued companies, statistical arbitrage took shape in the mid-1980s when Nunzio Tartaglia assembled a group of traders to create computer models that could efficiently access and analyze more data. Increases in computing power allowed pairs trading to become based primarily on statistics and technical factors, rather than fundamental analysis. (The latter still exists, and

many strategies of that sort fall into the category of risk arbitrage, which lie beyond the scope of this investigation.)

This paper offers two new cointegration tests, which are more robust than the Engle-Granger procedure regarding the distribution underlying the price process. The robust qualities are achieved through the use of weighted-Wilcoxon (rank-based) and least absolute deviation regression techniques. The development of these tests, as well as Monte Carlo simulations to calculate critical values and power for these tests is a primary contribution of this paper.

Among the potential applications of these robust tests, they could be particularly useful for use with stock prices, which are well known to violate normality assumptions (Mandelbrot, 1963; Fama, 1965). Pairs trading is the application presented here and is a second contribution of this paper. The robust tests are shown to create more profitable trading opportunities with a superior Sharpe ratio.

The next section discusses the development of the new cointegration tests. Section three describes a simulation study to test these cointegration tests, and results of those simulations are presented in section four. Section five outlines a simple pairs trading strategy that employs these tests, and section six provides backtest results of that trading strategy from 2001 to 2010. Section seven concludes.

2.0 Robust Cointegration Test Description

Cointegration can be conceptually described as a long-run equilibrium relationship between two or more nonstationary time series[1]. In other words, when considered independently, the time series appear to wander in an unpredictable random walk, but some

[1] In this study, all nonstationary time series will be assumed to be I(1), which implies that taking a first difference results in a stationary, invertible ARMA representation. This has been a common assumption in finance and economics since the work of Nelson and Plosser (1983) and also follows from the assumption that stock prices are random walks in an efficient market.

linear combination of those same variables exists that is stationary. A properly identified linear combination therefore follows a pattern of mean-reversion.

To help fix this idea, consider the graphs in Figure 1. The top row consists of time series plots of two variables constructed by the cumulative sum of 100 independent standard normal random variables. Both series were transposed vertically so that the first term is zero. Therefore, while starting at the same point, these variables are otherwise completely independent of each other and are nonstationary. The third plot, a scatterplot of *y* versus *x*, makes the random and independent nature of their relationship visually clear. There is no evident relationship between the two time series. Finally, the fourth plot is a time series plot of the difference of these two variables. If the two series were cointegrated, this difference would be stationary and demonstrate mean reversion, which would manifest itself in multiple crossings of the x-axis. In this case, however, the difference series appears qualitatively similar to the random walks of each time series.

Figure 1 also suggests the relationship between cointegration and spurious regression, which can occur when dealing with nonstationary time series. Hendry (1986) provides the long historical background of nonsense regressions, of which one example is Hooker (1901), who demonstrated a relationship between trade and the marriage rate. Granger and Newbold (1974) provide a rule of thumb for identifying such meaningless relationships, suggesting that any R^2 greater than the associated Durbin-Watson statistic should raise suspicions. Such problematic time series require first differencing of the variables to induce stationarity, unless the two series under consideration are cointegrated.

By contrast, Figure 2 shows the same four plots for two cointegrated series, *c* and *d*. The data for time series *c* were again generated by the cumulative sum of 100 independent standard

random normal variables. However, in this case, the second series was set equal to the first, plus a random disturbance term of the same form. The two series were then transposed vertically to begin at the origin.

The cointegrated relationship between *c* and *d* is clear in all four graphs. The first two plots demonstrate that the series tend to vary directly with each other; their local minima and maxima occur at approximately equal points. The scatterplot shows a strong linear relationship between the two variables, in stark contrast to the random scatterplot of Figure 1. Finally, the difference series appears to be stationary. That is, the mean and variance are both approximately constant. Finally, the fourth graph exhibits a strong mean-reversion with its multiple crossings of the x-axis.

The goal of cointegration testing is to identify time series that exhibit the same kind of stable relationship as in Figure 2. As implied by those graphs and the discussion above regarding spurious regression, one basic approach for cointegration testing involves a regression of one time series on the other. This seminal test for cointegration was developed by Engle and Granger (1987). Their definition states that two series are cointegrated if they are both I(1) and there exists a cointegrating coefficient β such that $z_t = y_t - \beta x_t$ is I(0).

The test for cointegration is a two-step procedure. First, the following regression equation is estimated: $y_t = \alpha + \beta x_t + z_t$. The series of residuals $\hat{z}_t$ is then tested for stationarity[2]. If the residuals are stationary (do not contain a unit root), the original series are cointegrated. This second-step requires a unit root test; in their consideration of seven contenders, Engle and Granger (1987) favor the Dickey-Fuller test as the most powerful.

[2] The estimate $\hat{\beta}$ is termed superconsistent because it converges to its true value at a rate of *T*, rather than the usual convergence rate of $T^{1/2}$ (Stock, 1987).

The Dickey-Fuller test (Dickey & Fuller, 1979) and Augmented Dickey-Fuller Test (ADF; Dickey & Fuller, 1981) remain the most popular and commonly utilized unit root tests, thanks largely to their implementation in statistical software packages. The basic version is based on the equation $x_t = \mu + \rho x_{t-1} + \varepsilon_t$, which is reparameterized to the following form: $\Delta x_t = \mu + (\rho - 1)x_{t-1} + \varepsilon_t$. This equation is estimated with a least-squares regression, and the appropriate test statistic is used for the null hypothesis that $\rho = 1$. While the test statistic is a standardized version of the mean, it does not follow a standard *t*-distribution. Mistakenly doing so can result in significant over-rejection of the null (Maddala & Kim, 1998). Therefore, Dickey and Fuller (1981) calculated the appropriate critical values through numerical analysis.

One shortcoming of the original Dickey-Fuller test is that it includes the implicit assumption that the error term is not serially correlated. The ADF test overcomes this problem by expanding the previous equation as follows:

$$\Delta x_t = \mu + \beta t + (\rho - 1)x_{t-1} + \sum_{i=1}^{m} \lambda_i \Delta x_{t-1} + \varepsilon_t$$

The lag length *m* must be determined by the researcher. The number can be assumed *a priori* or determined from the data. Once the lag length has been chosen, a least squares regression leads to a ratio statistic to test the null hypothesis that $\beta = 0$ and $\rho = 1$. (If no time trend is included, only the latter parameter is tested.) As before, the critical values are non-standard due to the non-stationary nature of the series under the null hypothesis.

This extended description of the Engle-Granger two-step testing procedure for cointegration is meant to highlight the fact that it demands the use of two least-squares regressions. This approach results in maximum likelihood estimators when the experimental errors are normally distributed. However, the estimates can be suboptimal in the presence of

outliers, which may stem directly from a violation of the normality assumption. Therefore, each regression will be replaced with robust regression techniques in the next section's development of more robust cointegration tests.

As a final note, it should be noted that the standard Dickey-Fuller test critical values do not apply in the second step of the Engle-Granger cointegration test. The reason is that the series of residuals is itself based upon the first step regression estimate. For this reason, Engle and Granger (1987) employed Monte Carlo simulation to generate the appropriate critical values, an approach that is followed in the present investigation. This allows a direct comparison both to the original study and to the robust cointegration tests developed next.

2.1 Robust Regression

Least squares regression is conducted by minimizing a dispersion function, equal to the sum of the squared residuals, with respect to the parameter *β:*

$$D_{LS}(\beta) = \sum_{i=1}^{n} \big(\varepsilon_i(\beta)\big)^2$$

Here, the expression $\varepsilon_i(\beta)$ represents the regression residuals $y_t - \beta_0 - x_t\beta_1$ while making their dependence on the parameter explicit in the notation. Problems with this estimate are outlined by Wilcox (1997), and they include low power, poor coverage, and distortions that arise from incorrect variance estimates when the error distribution is fat tailed. This has motivated a range of alternative regression techniques, two of which are employed in this study.

The first is weighted-Wilcoxon (WW) regression, also referred to as rank-based regression. Terpstra and McKean (2005) describe how an estimate is found through minimizing the following linear model, in which b_{ij} denotes a weight used in the (i, j)th comparison:

$$D_{WW}(\beta) = \sum_{1 \le i < j \le n} b_{ij} \big|\varepsilon_j(\beta) - \varepsilon_i(\beta)\big|$$

When $b_{ij} = 1$ for all $i \neq j$ and zero otherwise (known as Wilcoxon weights), Hettmansperger (1984) demonstrated that this dispersion function reduces to a simpler form:

$$D_{WW}(\beta) = 2\sum_{i=1}^{n}\left(R\big(\varepsilon_i(\beta)\big) - \frac{n+1}{2}\right)\varepsilon_i(\beta)$$

Here, $R\big(\varepsilon_i(\beta)\big)$ denotes the rank of the residuals, which demonstrates why WW regression can be considered a rank-based procedure. Terpstra and McKean (2005) further noted that this function corresponds to Jaeckel's (1972) dispersion function with Wilcoxon scores. Since this function is invariant to location, a reasonable and typical estimate of the intercept is given by $\hat{\beta}_0 = median\left(\varepsilon_i\big(\hat{\beta}\big)\right)$ (Hettmansperger & McKean, 1998). Finally, the robust nature of these estimates is evident from the linear effect of outliers as opposed to the quadratic influence in the least squares method (Hettmansperger, 1984).

The rank-based regression estimates for this study are conducted with R's contributed package Rfit (Kloke & McKean, 2012). This statistical code provides point estimates as well as standard errors, and the reader is referred to Hettmansperger and McKean (2011) and the associated R package for further details.

The second alternative regression technique employed in creating robust cointegration tests is least absolute deviation (LAD) regression. Historically, this method actually predates the development of least squares by 50 years, and its potential efficiency advantage over least squares was known by Laplace (Wilcox, 1997). The LAD estimator minimizes the sum of the absolute values of the residuals, giving rise to the following dispersion function:

$$D_{LAD}(\beta) = \sum_{i=1}^{n}|\varepsilon_i(\beta)|$$

This function further limits the influence of outliers and produces better estimates for leptokurtic distributions.

The coefficients are standardized to calculate a test statistic for inference. These estimates are generally inferior for normally distributed errors but superior for a wide range of other error distributions (Koenker & Bassett, 1978). The numerical estimations used here come from the R package quantreg (Koenker, 2012).

2.2 Robust Cointegration Test Methodologies

The robust cointegration tests proposed here follow the same two-step framework as the Engle-Granger procedure. They begin with a cointegrating regression. Rather than using ordinarly least squares, however, the rank-based test utilizes WW regression and the LAD-based test employs LAD regression, both as described in the previous section. By using different norms in their dispersion functions, these regression techniques minimize the influence of outliers and yield slightly different parameters and residuals.

The second step, again following Engle-Granger, is a unit root test on the residuals from the cointegrating regression. The tests are identical in spirit to the ADF test, but again replace the OLS regressions with WW and LAD regressions. Once again, the goal is to overcome any undue influence from outlying observations or violations of the normality assumption that underlies the Engle-Granger test. As an example of this property, the Wilcoxon test for the location problem of locating the median is known to have 95% efficiency for the normal distribution. Hettmansperger and McKean (1998) show further that rank-based procedures are superior to least-squares for even 1% contamination of the normal distribution and that LAD estimation also surpasses least squares, with a crossover at 10% contamination. Thus, error

distributions with fatter tails (i.e., greater variance) than a normal distribution will be better modeled and tested with rank- and LAD-based methods.

The robust cointegration tests are not solvable in closed form and do not follow a standard distribution (similar to the ADF and Engle-Granger tests themselves). Therefore, the critical values are generated through Monte Carlo simulation with 10,000 repetitions, as described more fully in section three.

2.3 Alternative Robust Tests

Before describing the Monte Carlo simulation, we briefly consider other alternative cointegration tests that have been developed. In the past thirty years, a wide range of modifications and alternative procedures have been proposed for more robust unit root and cointegration tests. Some of these tests still contain within them a least squares estimation step; for example, the Said-Dickey test (1984) and Phillips-Perron test (1987) for a unit root. The underlying least squares regression limits their robustness. Furthermore, both suffer from significant size distortion (Phillips & Perron, 1988), so these tests offer only marginal improvements over the ADF test.

Maddala and Kim (1998) discuss a broad assortment of other unit root tests. Among them are tests based on the Durbin-Watson statistic (Sargan & Bhargava, 1983), a variance ratio test (Cochrane, 1988; Lo & MacKinlay, 1988), and instrumental variable regression (Hall, 1989; Choi, 1992). One test that reverses the hypotheses and uses stationarity as the null is the KPSS test (Kwiatkowski, Phillips, Schmidt, & Shin, 1992). Each of these tests offers some improvement at the cost of complexity, but none has been found to be universally better in simulation studies. Maddala and Kim (1998) end their extensive review by urging simple but

robust replacements for the ADF test. The rank- and LAD-based tests proposed here are offered as conceptually straightforward statistics with superior qualities in practice.

Alternative cointegration tests have also flourished in recent years. The most prominent is the Johansen test (1988). However, both the Engle-Granger and the Johansen tests perform poorly when faced with nonlinearities, homoscedasticity, or structural breaks. Some proposed alternatives that attempt to overcome these problems include induced order statistics (Escribano, Santos, & Sipols, 2008), record counting correction (Escribano, Sipols, & Aparicio, 2006), and a nonparametric version of the Johansen test (Bierens, 1997). Despite the ongoing research in this area, few of the alternative cointegration tests have been used in applications or compared in published simulations. One advantage of the present investigation is the direct comparison to the Engle-Granger procedure and the direct application to pairs trading.

3.0 Simulation Study Parameters

This section provides details on the Monte Carlo simulations that allow for comparison of the power of the proposed robust cointegration tests with the Engle-Granger procedure. All simulations employ the following data-generation process for the two time series x_t and y_t:

$$x_t + y_t = u_t$$

$$x_t + \beta y_t = e_t$$

$$u_t = u_{t-1} + \varepsilon_{1t}$$

$$e_t = \rho e_{t-1} + \varepsilon_{2t}, |\rho| \leq 1$$

In general, the cointegrating constant β can take on any real value, but it is convenient for the purpose of testing and comparison to limit the data-generation to one particular value. In this study, it is fixed at $\beta = 2$, which is a common assumption in previous studies (e.g., Phillips & Hansen, 1990). The error vector $(\varepsilon_{1t}, \varepsilon_{2t})'$ consists of two independent and identically

distributed stable random variables (as formally defined below), each with an expected value of zero. This basic form of the data generating process has been widely employed (e.g., Banerjee, Dolado, Hendry, & Smith, 1986; Engle & Granger, 1987; Phillips & Loretan, 1991; Kremers, Ericsson, & Dolado, 1992). Solving for *x* and *y* gives the following equations, used to generate the simulated time series:

$$x_t = \beta(\beta - 1)^{-1}u_t - (\beta - 1)^{-1}e_t$$

$$y_t = -(\beta - 1)^{-1}u_t + (\beta - 1)^{-1}e_t$$

When $|\rho| = 1$, both time series are I(1), and the two series are not cointegrated. In that case, the null hypothesis is false; this allows for estimation of the tests' size. When $|\rho| < 1$, both variables are still I(1) but they are cointegrated with cointegrating coefficient $\beta = -2$. Thus, varying the value of ρ allows for estimation of power. It is more difficult for the tests to differentiate the near-cointegrated cases when $|\rho|$ is close to one; that is, power varies inversely with ρ. The simulations include a range of values for comparison: ρ = (0.8, 0.9, 0.95, 1).

3.1 Time series length

Much of the theory of cointegration tests is based on the validity of asymptotic distributions. However, practitioners frequently work with time series of limited length. This simulation focuses on time series of lengths t = (50, 100, 250). The largest of these values corresponds approximately to the number of trading days on the New York Stock Exchange in a one year period. This calibration is thus ideal for pairs trading strategies that use a one-year formation period, as explained in section 4.

In the simulations, time series of length t + 20 are generated, and the first 20 values are discarded. This methodology minimizes start-up effects. Furthermore, it allows for the valid inclusion of a constant term in all regressions and tests.

3.2 Error distributions

The data-generating process is quite general with regard to the distribution of the error terms. Engle and Granger (1987) considered only normally distributed errors in the development of their cointegration test. Most papers since then have made similarly strong restrictions by considering a narrow range of error distributions. The purpose of the proposed Rank- and LAD-based methods is to demonstrate more robust tests that apply to a broader range of error distributions.

Stable distributions provide a family of probability distributions that can include the properties of both heavy tails and skewness. One theoretical reason for their use is that the normal distribution is a special case within this family. In fact, the normal, Cauchy, and Levy distributions are the only three special cases within this family for which a density function exists in closed form. In general, stable distributions are those distributions that maintain their shape under addition, leading to Nolan's (2003) definition of the stable law: if $X, X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) stable random variables, then for all n, $X_1 + X_2 + \cdots + X_n \xrightarrow{d} c_n X + d_n$ for some constants $c_n > 0$ and d_n.

There are at least three reasons that stable distributions meeting this criterion may be chosen as a model for a given data set, as outlined in Nolan (2003). First, there may be theoretical reasons for expecting such a distribution in a physical process. Second, empirical investigations may demonstrate that a data set exhibits heavy tails and skewness. This motivation has a long history in finance, with examples from Mandelbrot (1963) and Fama (1965) to Rachev and Mittnik (2000). Stock prices, in particular, are well known to exhibit leptokurtic, left-skewed distributions.

The final motivation for modeling with stable distributions is the Generalized Central Limit Theorem (GCLT). The classical Central Limit Theorem is based on a random sample with non-infinite variance and proves convergence in distribution to the standard normal distribution. The GCLT drops the assumption of finite variance and shows that the limiting distribution must be stable. That is, if $X, X_1, X_2, \ldots, X_n$ are i.i.d. random variables, there exist constants $a_n > 0$, b_n, and a non-degenerate random variable Z such that $a_n(X_1 + \cdots + X_n) - b_n \xrightarrow{d} Z$ if and only if Z is stable (Nolan, 2003). Therefore, if an observed data series can be thought of as the sum of many small terms, it may be modeled adequately by a stable distribution. This coincides well with the notion of efficient markets and random walks in price series.

In general, stable distributions are characterized by four parameters: a characteristic exponent $\alpha \in (0, 2]$, a skewness parameter $\beta \in [-1, 1]$, a scale parameter $\gamma \geq 0$, and a location parameter $\delta \in \Re$ (Nolan, 2003). In this investigation, the normal distribution ($\alpha = 2, \beta = 0$) is included as a baseline for comparison to the heavy-tailed distributions. The other distributions considered are chosen for their leptokurtic shapes or fat tails.

In the Monte Carlo simulations, the particular value of $\alpha = 1.7$ is selected based on Mandelbrot's (1963) estimate of the parameter from the time series of cotton prices. In that same work, Mandelbrot notes that the data suggest the series are not symmetric. Instead, β "takes a small negative value" (p. 405). This motivates consideration of a skewness parameter of –0.25 in this study. The power of the Rank- and LAD-based tests should increase as α decreases, and the power of all tests should suffer from skewness in the error distributions.

The Engle-Granger data-generating process originally required an expected value of zero. Stable distributions, however, do not always have well-defined expectations. Specifically, the mean does not exist whenever $\alpha \leq 1$. In that case, the requirement could be more broadly

interpreted to mean a location parameter (δ) equal to zero. This is equivalent to the original assumption of a zero mean, when the mean is well-defined.

The final parameter is the scale parameter γ, which corresponds to variance for the normal distribution. Unlike the other three parameters, which are equal in both time series within a given simulation, the ratio of scale parameters is varied so that it takes on the following values: $\gamma_1/\gamma_2 = (4, 2, 1)$. This can be thought of as a signal-to-noise ratio, as in Hansen and Phillips (1990). Using the normal distribution in a Monte Carlo simulation study of cointegration, Banerjee, et al. (1993) allowed the ratio of variances to cover a similarly wide range. The authors measured the bias in estimating the cointegrating coefficient and found that it varied inversely with the ratio of variances. This logically implies that the power should vary directly with the ratio of the spread parameter used to generate the error terms.

3.3 Lag length

The size and power of the cointegration tests are influenced by the number of lagged terms included in the second-step unit root test (Schwert, 1989). This necessitates a choice from several potential guidelines for the number of lagged terms: an arbitrary fixed level, a function of the length of the time series, the Akaike or Schwart Information Criteria, of a sequential rule. Examples of the last method are discussed in Hall (1994), in which a general-to-specific methodology is described. The researcher begins with a large value, tests the significance of the last term, and then decreases the lag length until a significant statistic is found.

In a comparison study of these choices, Ng and Perron (1995) conclude that the general-to-specific method consistently chooses larger values for the appropriate lag length. This inclusion of more lags tends to decrease power slightly, but it lessens size distortions (DeJong, Nankervis, Savin, & Whiteman, 1992).

In a computational study, any search procedure is expensive in terms of processing time. Therefore, the Monte Carlo simulations employ a fixed number of lags as a function of the time series length. The particular function for the number of lags is given as follows (Banerjee, Dolado, Galbraith, & Hendry, 1993), $k = \lfloor (length(x) - 1)^{1/3} \rfloor$.

3.4 Critical Value Generation

In making comparisons among Monte Carlo simulation results, it is imperative to ensure approximately equal size for each test. Tests that are liberal or conservative in comparison to their nominal size also have distorted power. Naïve power comparisons that ignore this issue of size distortion can lead to incorrect conclusions. Lloyd (2005a, 2005b) provides several suggestions to ensure valid comparisons, including *post hoc* methods to calculate size-adjusted power. He argues most forcefully that no matter what method is chosen, steps must be taken to ensure the validity of power comparisons among proposed tests.

This investigation avoids the necessity for size-adjusted power by generating the critical values in the same manner as Engle and Granger (1987). The distributions of the test statistics are built through 10,000 repetitions of the test under the null hypothesis, and the 5% critical values are determined from the empirical distribution of the test statistic. These critical values are then used in the simulations to calculate size and power. In this way, all of the tests have approximately equal empirical size of 5% and not just a nominal size that can vary widely in application.

There are two advantages of this approach. First, it eliminates the need for *post hoc* size adjustments before comparing the tests. Second, the least squares results are analogous to the original results of Engle and Granger (1987). Similar results in that case provide assurance that the simulation code is correctly specified.

3.5 Size and Power Estimation

Once the critical values have been generated, the simulations are run again under the null hypothesis of no cointegration. In that case, the time series are built from independent random errors, and the tests are based upon the generated critical values. Theoretically, all tests should have an empirical size of 5% with a standard error of $\sqrt{\frac{\hat{\alpha}(1-\hat{\alpha})}{10{,}000}}$, since the simulations are all run with 10,000 repetitions (Lloyd, 2005a). This standard error is dependent upon the calculated size, but for an expected size of 5% the standard error would be approximately just 0.2179%.

Power calculations are made by generating cointegrated series and applying the tests with the estimated critical values. As mentioned in the discussion of the data generating process, power is calculated for three values of ρ. Comparison of the power of the three different tests is a primary contribution of this study.

3.6 Summary

The Monte Carlo simulations employ 10,000 replications on the parameter space implied by the following specification: $t = (50, 100, 250), \gamma_1/\gamma_2 = (4, 2, 1),$ and $\rho = (0.8, 0.9, 0.95, 1)$. Furthermore, three different error distributions are considered in the data generating process. In addition to the normal distribution, errors are generated using two stable distributions, one with $\alpha = 1.7$ and $\beta = 0$, the other with $\alpha = 1.7$ and $\beta = -0.25$. Finally, each of the three cointegration tests (OLS-, Rank-, and LAD-based) is applied to allow for comparison of the tests' size and power. This gives a total of 324 experiments.

4.0 Cointegration Test Results

Results of the Monte Carlo simulations are presented in Tables 1, 2, and 3. In Table 1, the error distributions are generated with a dispersion ratio of one ($\gamma_1/\gamma_2 = 1$), while Tables 2 and 3 represent dispersion ratios of two and four, respectively. In each table, Panel A employs

normally distributed errors in the data generating process, Panel B uses a stable distribution with fatter tails ($\alpha = 1.7$) and Panel C additionally includes a negative skew in the error distribution ($\beta = -0.25$).

In the tables, the three tests are labeled as follows: least squares Engle-Granger procedure (OLS), weighted-Wilcoxon rank-based (Rank), and least absolute deviation-based (LAD). All reported values are the means over 10,000 replications, so standard errors of the reported figures are given by the formula $\sqrt{\frac{\hat{\alpha}(1-\hat{\alpha})}{10{,}000}}$. Therefore, the maximum half-width of a 95% confidence interval is 0.0098; any reported figures for power that differ by more than 0.0196 between experiments would be guaranteed to differ significantly at the 5% level. This is the most conservative estimate. For reported power of 0.8, for example, the half-width would shrink to 0.004. Throughout the tables, the results display the superiority of the proposed rank- and LAD-based test statistics over the original Engle-Granger procedure[3].

First, we observe that the power increases in all cases with the length of the time series being tested, as expected. Performance for $t = 50$ is unacceptable in any case, with power never surpassing 50% for any of the test procedures. However, results for $t = 250$ are adequate for practical use in many cases, depending on the other parameters.

In all cases, the power of the tests varies inversely with ρ, again matching expectations. Power suffers significantly for all tests in the near cointegrated cases, where $\rho = 0.95$. This is one parameter that demonstrates the superiority of the robust tests. For example, in Table 1, Panel C, the rank-based statistic has reported power of 70.92% for $t = 250$ and $\rho = 0.95$, while

[3] In unreported results, the newly developed statistics were also compared to the Johansen test (Johansen, 1988; Johansen and Juselius, 1990). The Johansen procedure was much inferior, in terms of power, to all three tests under consideration here.

the OLS test has power of only 24.94%. In general, the rank-based test maintains its power better as ρ increases, which is one important aspect of its overall performance.

To compare the effect of the error distribution, we can examine the corresponding numbers from Panels A and B within each table. From these comparisons, it is evident that the LAD- and rank-based tests perform better in the presence of fat-tailed distributions, which is precisely the situation in which the OLS test has declining power. This is a primary advantage of the robust tests over the standard Engle-Granger procedure. As an example, in Table 1 for t = 250 and ρ = 0.95, the power of the rank-based test increases from 27.22% in Panel A to 70.34% in Panel B. For the same parameters, the power of the OLS-based test decreases from 27.74% to 24.14%.

The skewness parameter introduced in Panel C of all three tables causes little change in the performance of the tests. In some cases, there is actually moderate improvement in the tests' power, particularly for lower values of ρ. To give one example, for t = 100 and ρ = 0.8, the power of the rank-based test is 87.31% in Panel B of Table 1, but it improves to 89.42% in Panel B. For the same parameters, the OLS test improves in power from 59.07% to 63.67%. From this, we conclude that moderate skewness does not damage performance of the cointegration tests significantly.

To consider the effect of the dispersion parameter ratio, comparisons are made across the three tables. There is little difference in power due to variation in this parameter; however, the general trend is for power to increase as the ratio increases. This trend can be seen in the slightly higher reported power in Table 3. For instance, in Panel B, for the rank-based test with t = 250 and ρ = 0.95, Table 1 reports power of 70.34%, which increases to 74.91% in Table 2 and 76.38% in Table 3.

Overall, the results show that in the case of normally distributed errors (Panel A of all tables), the most power test is the Engle-Granger methodology. This is unsurprising, since normally distributed errors are a key assumption of the theoretical validity of that procedure. However, the rank-based procedure sacrifices little power, even in this baseline case. This minimal loss of power suggests there is little misspecification risk from using the rank-based test when errors are, in fact, normally distributed.

For non-normally distributed errors, the rank- and LAD-based tests quickly gain power and overtake the Engle-Granger estimation method. In fact, the robust tests perform better as the tails of the error distribution get fatter, even for large values of ρ. This is especially powerful evidence in favor of the robust tests, because the Engle-Granger (and other) early testing procedures suffered from notoriously low power. For time series of lengths 50 or 100, the power of the OLS method barely exceeds the test's size. Clearly, this is a substandard test for drawing valid conclusions, because the risk of Type II error is so large.

Finally, note that the critical values of the robust test statistics are relatively stable over all simulations. The critical value of the rank-based test ranges from -3.36 for a time series of length 50 and normally distributed errors to just -3.15 for a time series of length 250 in Table 3 Panel C. This stability is a desirable feature because it demonstrates that slight misspecifications will not dramatically alter the test's conclusions. It also provides a heuristic rule-of-thumb for the test, with -3.25 serving as a useful benchmark for a general rank-based test.

Overall, the superiority of the robust cointegration tests developed in this paper is evident. The new tests possess greater power for fat-tailed distributions, while sacrificing little in the baseline case of normal errors. They have stable critical values, with a nominal and

empirical size of approximately 5%. They perform well for near-cointegrated series and are conceptually simple and easy to implement with existing statistical packages.

5.0 Pairs Trading Strategy

Pairs trading is one direct application of the robust cointegration tests developed in the previous section. Though the algorithm or model employed in a particular statistical arbitrage strategy will be unique to the individual or firm, they all follow the same basic format. First, an initial universe of candidate stocks is identified. This preliminary step allows a trader to focus on a particular industry or other subset that might be presumed to have a high degree of comovement, if desired.

Second, candidate pairs are tested for suitability. In this study, we use a cointegration test to determine if a pair of stocks is likely to be profitable in a mean-reverting trade. Previous papers have also used a minimum squared distance requirement, most notably the work of Gatev, Goetzmann, & Rouwenhorst (2006). This step is conducted over a training period, and only stocks that reach a predetermined threshold pass on to the trading that takes place in step three. This study utilizes a one-year training period. Stocks are tested for cointegration based on 250-day time series, using the Engle-Granger procedure as well as the rank- and LAD-based cointegration tests.

Third, the spread between the stocks is calculated. This can be a difference, a ratio, or regression based. This study, like most that use a cointegration test, uses the third approach. The regression parameters from the training period are assumed stable for the subsequent 125 trading days, and the residual is calculated for the pair over that period. The spread is then traded, according to the trader's risk management scheme.

In this study, a trade is put on whenever the spread moves more than two standard deviations (in either direction) beyond its historical mean from the training period. The spread is then considered a synthetic instrument for trading, though it requires two positions, a long and a short position in the two stocks. The trade is closed at a profit when the pair reverts to its historical mean. The trade is also closed at a loss if the spread widens to three standard deviations, as a basic loss prevention technique. The trade is also closed, regardless of profitability, at the end of the 125-day trading period.

One example of a pair that is traded is presented in Figure 3. In the top panel, the spread of the two stocks is plotted for the trading period. The solid horizontal line represents the mean from the training period, and the dotted horizontal lines represent the two- and three-standard deviation marks above the mean. This pair opens on day 8 and closes on day 39 at a loss, because the stop loss limit is reached when the spread widens beyond three standard deviations. On day 73, the spread crosses the second standard deviation from below again, so the trade is reentered. It terminates profitably at day 122 and is not open at the end of the period. The middle panel presents a bar graph of when a position is held in the pair, and the lower panel is a graph of cumulative profitability.

The data used for the backtest are daily prices for the stocks that constitute the S&P100 within each year, with data collected from CRSP. Because these are large, American companies, there is a reasonable *a priori* expectation that prices are more likely to be cointegrated than random pairings in the broader market. We do not attempt to limit the candidate pool further by selecting stocks in the same industry or any other similar requirement. Doing so would likely increase profitability and be a more sophisticated trading strategy.

For the S&P 100 stocks, there are 4,950 possible pairs to test for cointegration each year. All stocks that pass the cointegration test become candidates for inclusion in the trading portfolio, and cumulative log returns are calculated for all open positions. The pairs are equally weighted, as if $1 were invested in each synthetic spread position, and the portfolio return is considered for comparison among the various testing procedures, as well as the Sharpe ratio and excess returns in the form of Jensen's alpha and Fama-French 3-factor regressions.

The general strategy is chosen because it compares well with previous work (e.g., Do & Faff, 2009). It is not presented as a realistic trading algorithm but rather as a straightforward rubric by which to compare the usefulness of the three cointegration tests. We stress that the main interest here is not profitability *per se,* but rather the difference in profitability due to the various cointegration tests. The next section provides results from the trading.

6.0 Pairs Trading Profitability

Table 4 displays the number of candidate pairs that are identified as cointegrated for each year of the sample. Recall that for 100 stocks, there are 4,950 pairs that are tested for cointegration. The fluctuation in the number of cointegrated pairs is dramatic for just a ten year sample. As the extreme cases, the rank-based test identifies 289 pairs (5.8%) as cointegrated in 2006, while in 2009 fully 1,489 (30.1%) are similarly identified.

The results in Table 4 demonstrate that the rank-based cointegration test generally identifies more candidate pairs than the Engle-Granger (OLS) test, while the LAD test typically offers the fewest pairs for trading. These results suggest that the stock price series are not fat-tailed enough for the LAD test to be superior. In the remainder of the results, we focus primarily on the OLS and rank-based results.

In particular, Table 4 also presents the degree to which these two tests identify the same pairs. In each year, the two tests each identify a certain subset that the other does not, while there is also a subset of stocks that are identified by both screening tests. The overlap percent ranges from 38.1% in 2001 to 57.5% in 2004. This creates multiple portfolios for consideration and comparison in terms of profitability.

We present five portfolios for comparison; three of them are based on all pairs identified by the OLS, rank, and LAD procedures. The fourth portfolio is all pairs that the OLS test recognizes as cointegrated while the rank test does not. Similarly, the fifth portfolio is all stocks that are identified as cointegrated by the rank test, but not so identified by the OLS test. Profitability of these five portfolios appears in Table 5.

In six of the ten years in the sample period, the rank-based trading procedure outperforms the OLS version. Over the 10-year period, the average outperformance is 0.37%. The LAD test outperforms OLS in five years and outperforms the rank version in five years. The "rank only" portfolio consists of the pairs that are identified only by the rank cointegration test. This portfolio outperforms the OLS procedure in seven of the ten years and with a 10-year average of 0.77%, more than double the whole rank-based portfolio. These results suggest that the pairs identified solely by the rank procedure are uniquely profitable for this sample.

The above results focus on year-by-year comparisons, but it is in the longer term returns that the superiority of the rank test can truly be seen. Over the entire 10-year period, the OLS test yields a total return of 3.62% while the rank test returns 7.32%, more than double. It does so on almost identical standard deviation (0.023 for OLS, 0.024 for rank). The Sharpe ratio of the rank-based test is a high 1.81, compared to just 0.26 for the OLS version.

We present excess returns based on Jensen's alpha and Fama-French three-factor regressions in Table 6 for the rank-based profits. The alphas are generally positive but not significant, with the exception of 2002. This is in line with expectations from previous research (e.g., Do & Faff, 2009). It would appear that the profits of this simple trading strategy do not exceed what could be expected due to known risk factors. We conjecture that there are too many losing trades in the portfolios because the stock pairs are not limited to be beta neutral, within the same industry, or any additional screening factors that might further limit divergence risk. This bears further scrutiny in future research.

Finally, Table 7 begins to explore the source of the superiority of the rank-based methodology. The table presents the percent of identified pairs that lose money during the trading period. In eight of the ten years, the robust cointegration tests results in a small percentage of losing pairs. In other words, pairs identified by the OLS procedure are more likely to diverge beyond the three standard deviation stop-loss rule, compared to the rank-based test. This is one source of the test's superiority, which again will be explored in further applications in the future.

7.0 Conclusion

Pairs trading and various incarnations of convergence trading remain popular on Wall Street. These strategies will always demand some statistical measure to identify portfolios of tradable opportunities. This paper proposes two new tests as alternative to the Engle-Granger cointegration test. These tests are more robust to outliers and violations of the assumption of normally distributed errors in stock returns. Robustness is accomplished through weighted-Wilcoxon (rank-based) and LAD regression techniques. The two new tests have superior power

in many situations of interest. The critical values generated through Monte Carlo simulation and presented here allow for the use of these tests in more situations in future research.

To demonstrate one such application, pairs trading results for the S&P100 from 2001 to 2010 were calculated using a simple pairs trading strategy. The rank-based test provided stock pairs with superior traits. Profitability was generally better, particularly in the long term, and the Sharpe ratio was considerably improved over OLS testing.

8.0 References

Banerjee, A., Dolado, J., Galbraith, J., & Hendry, D. (1993). *Cointegration, Error Correction, and the Econometric Analysis of Non-Stationary Data*, Oxford: Oxford University Press.

Banerjee, A., Dolado, J., Hendry, D., & Smith, G. (1986). Exploring equilibrium relationships in econometrics through static models: Some Monte Carlo evidence. *Oxford Bulletin of Economics and Statistics, 48,* 253 – 278.

Bierens, H. (1997). Nonparametric cointegration analysis. *Journal of Econometrics, 77,* 379 – 404.

Choi, I. (1992). Effects of data aggregation on the power of tests for a unit root: A simulation study. *Economics Letters, 40,* 397 – 401.

Cochrane, J. (1988). How big is the random walk in GNP? *Journal of Political Economy, 96,* 893 – 920.

DeJong, D., Nankervis, J., Savin, N., & Whiteman, C. (1992). The power problems of unit root tests for time series with autoregressive errors. *Journal of Econometrics, 53,* 323 – 343.

Dickey, D., & Fuller, W. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74,* 427 – 431.

Dickey, D., & Fuller, W. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica, 49,* 1057 – 1072.

Do, B., & Faff, R. (2009). Does simple pairs trading still work? *Financial Analysts Journal, 66,* 83 – 95.

Engle, R., & Granger, C. (1987). Cointegration and error correction: Representations, estimation, and testing. *Econometrica, 55,* 252 – 276.

Escribano, A., Santos, M., & Sipols, A. (2008). Testing for cointegration using induced-order statistics. *Computational Statistics, 23,* 131 – 151.

Escribano, A., Sipols, A., & Aparicio, F. (2006). Nonlinear cointegration and nonlinear error correction: Record counting cointegration tests. *Communications in Statistics: Simulation & Computation, 35,* 939 – 956.

Fama, E. (1965). The behavior of stock prices. *Journal of Business, 38,* 34 – 105.

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

Granger, C., & Newbold, P. (1974). Spurious regression in econometrics. *Journal of Econometrics, 2,* 111 – 120.

Hall, A. (1989). Testing for a unit root in the presence of moving average errors. *Biometrica, 76,* 49 – 56.

Hall, A. (1994). Testing for a unit root in time series with pretest data-based model selection. *Journal of Business and Economic Statistics, 12,* 461 – 470.

Hendry, D. (1986). Econometric modeling with cointegrated variables: An overview. *Oxford Bulletin of Economics and Statistics, 48,* 201 – 212.

Hettmansperger, T. (1984). *Statistical Inference Based on Ranks.* New York: Wiley & Sons, Inc.

Hettmansperger, T., & McKean, J. (2011). *Robust Nonparametric Statistical Methods, 2nd ed.* New York: Chapman Hall.

Hooker, R. (1901). Correlation of the marriage rate with trade. *Journal of the Royal Statistical Society.*

Jaeckel, L. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Annals of Mathematical Statistics, 43,* 1449 – 1458.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control, 12,* 231 – 254.

Johansen, S., & Juselius, K. (1990). Maximum likelihood estimation and inference on cointegration with applications to the demand for money. *Oxford Bulletin of Economics and Statistics, 52,* 169 – 210.

Koenker, R. (2012). quantreg: Quantile Regression. R package version 4.90.

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46,* 33 – 50.

Kloke, J., & McKean, J. (2011). Rfit: Rank estimation for linear models. R package version 0.14.

Kremers, J., Ericsson, N., & Dolado, J. (1992). The power of cointegration tests. *Oxford Bulletin of Economics and Statistics, 54,* 325 – 348.

Kwiatkowski, D., Phillips, P., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics, 54,* 159 – 178.

Lloyd, C. (2005a). Estimating test power adjusted for size. *Journal of Statistical Computation and Simulation, 75,* 921 – 934.

Lloyd, C. (2005b). On comparing the accuracy of competing tests of the same hypotheses from simulation data. *Journal of Statistical Planning and Inference, 128,* 497 – 508.

Lo, A., & MacKinlay, A. (1988). Stock market prices do not follow random walks: Evidence from a simple specification test. *Review of Financial Studies, 1,* 41 – 66.

Maddala, G., & Kim, I. (1998). *Unit Roots, Cointegration, and Structural Change.* Cambridge: Cambridge University Press.

Mandelbrot, B. (1963). The variation of certain speculative prices. *Journal of Business, 36,* 394 – 419.

Nelson, C., & Plosser, C. (1982). Trends and random walks in macroeconomic time series. *Journal of Monetary Economics, 10,* 139 – 162.

Ng, S., & Perron, P. (1995). Unit root tests in ARMA models with data-dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association, 90,* 268 – 281.

Nolan, J. (2003). Modeling financial data with stable distributions. *Handbook of Heavy Tailed Distributions in Finance,* ed. Rachev, S., Amsterdam: Elsevier Science, B.V.

Phillips, P., & Hansen, B. (1990). Statistical inference in instrumental variables regression with I(1) processes. *Review of Economic Studies, 57,* 99 – 125.

Phillips, P., & Loretan, M. (1991). Estimating long-run economic equilibria. *Review of Economic Studies, 58,* 407 – 436.

Phillips, P., & Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika, 75,* 335 – 346.

Rachev, S., & Mittnik, S. (2000). *Stable Paretian Models in Finance.* New York City: Wiley.

Said, S., & Dickey, D. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrica, 71,* 599 – 607.

Sargan, J., & Bhargava, A. (1983). Testing residuals from least squares regression for being generated by the Gaussian random walk. *Econometrica, 51,* 153 – 174.

Schwert, G. (1989). Tests for unit roots: A Monte Carlo investigation. *Journal of Business and Economic Statistics, 7,* 147 – 159.

Stock, J. (1987). Asymptotic properties of least squares estimators of cointegrating vectors. *Econometrica, 55,* 1035 – 1056.

Terpstra, J., & McKean, J. (2005). Rank-based analyses of linear models using R. *Journal of Statistical Software, 14,* 1 – 26.

Vidyamurth, G. (2004). *Pairs Trading, Quantitative Methods and Analysis.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Wilcox, R. (1997). *Introduction to Robust Estimation and Hypothesis Testing.* San Diego: Academic Press.

Figure 1. Example of non-cointegrated time series

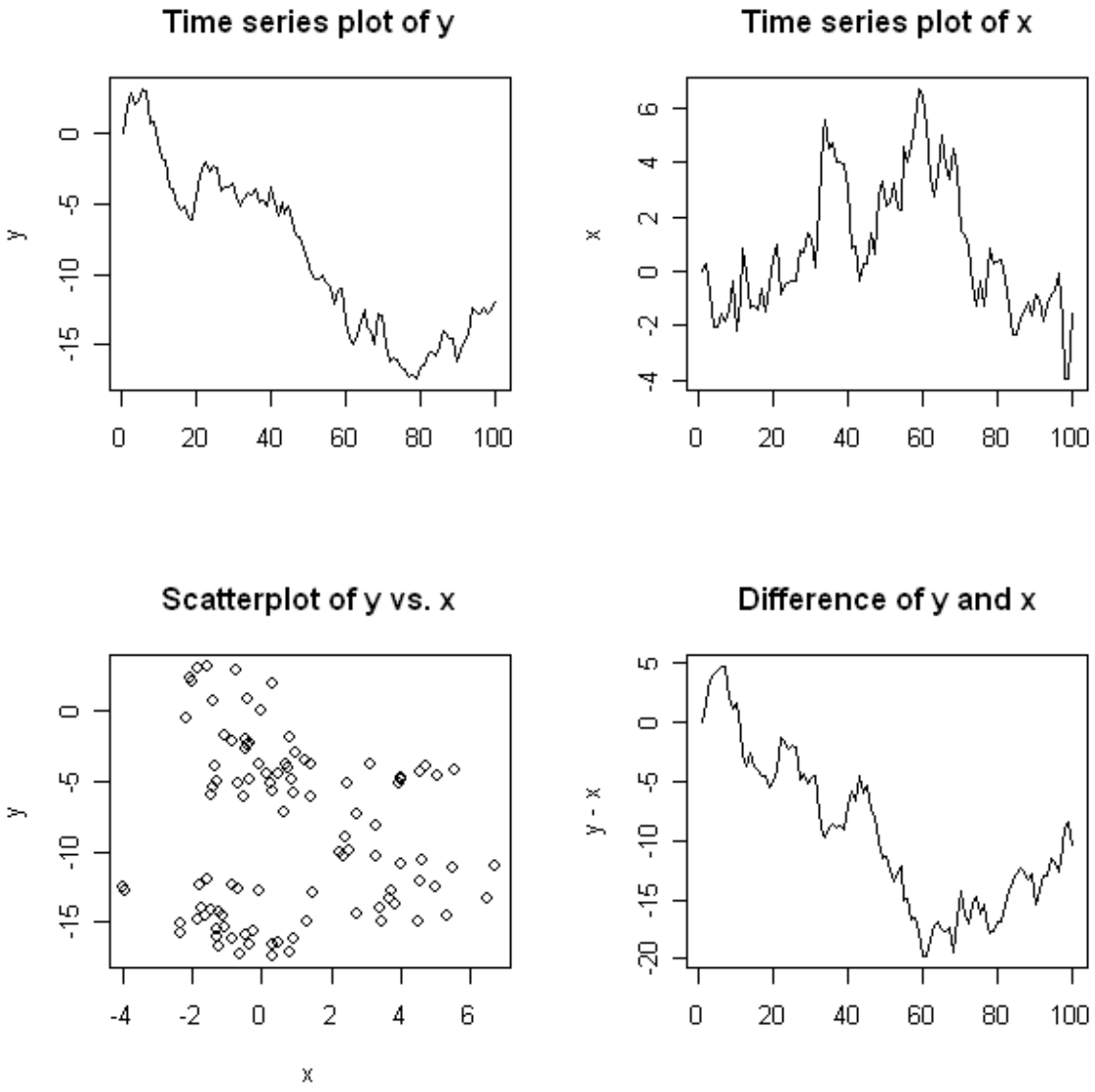

Figure 2. Example of cointegrated time series

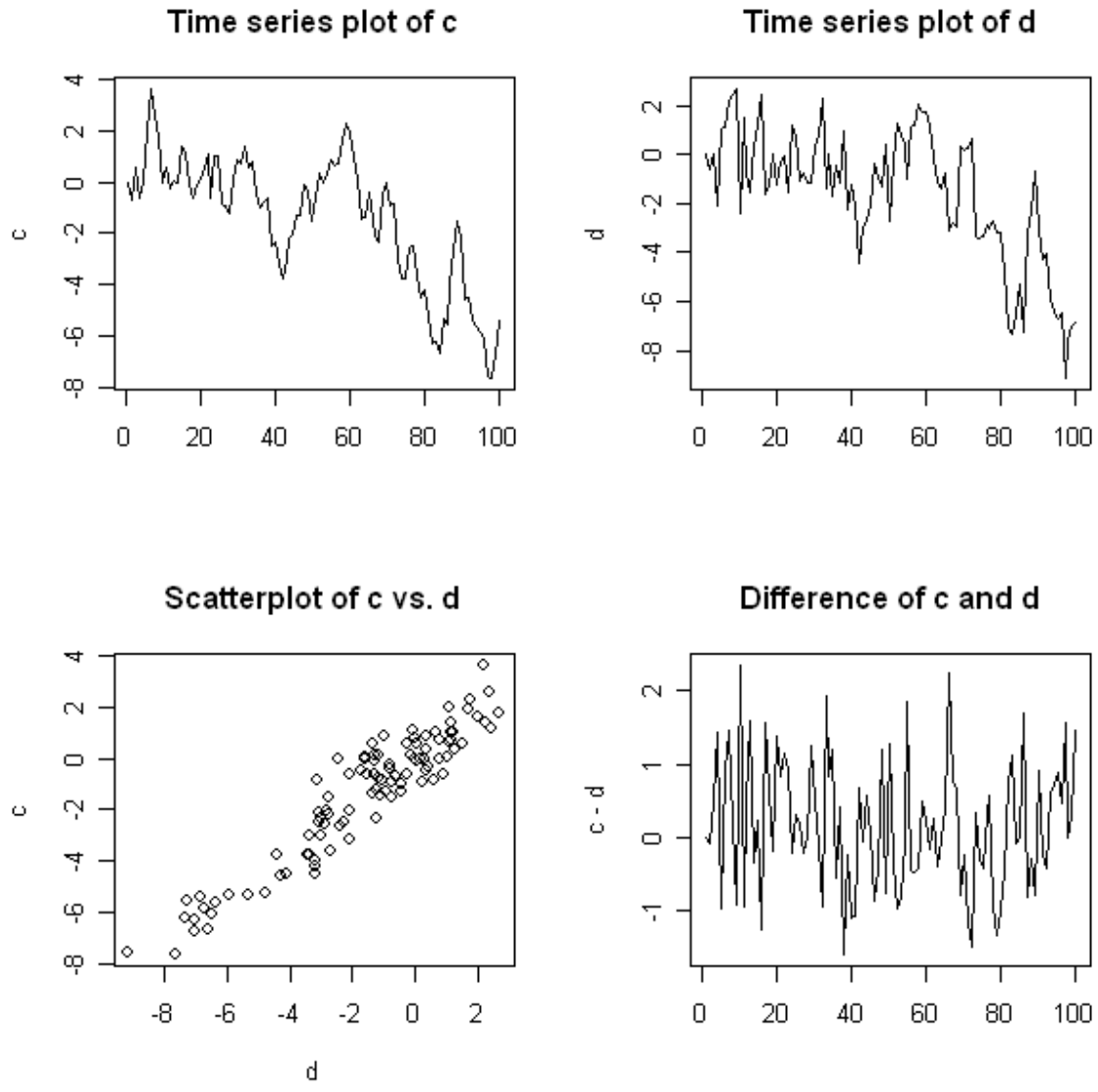

Figure 3. Pairs trading example

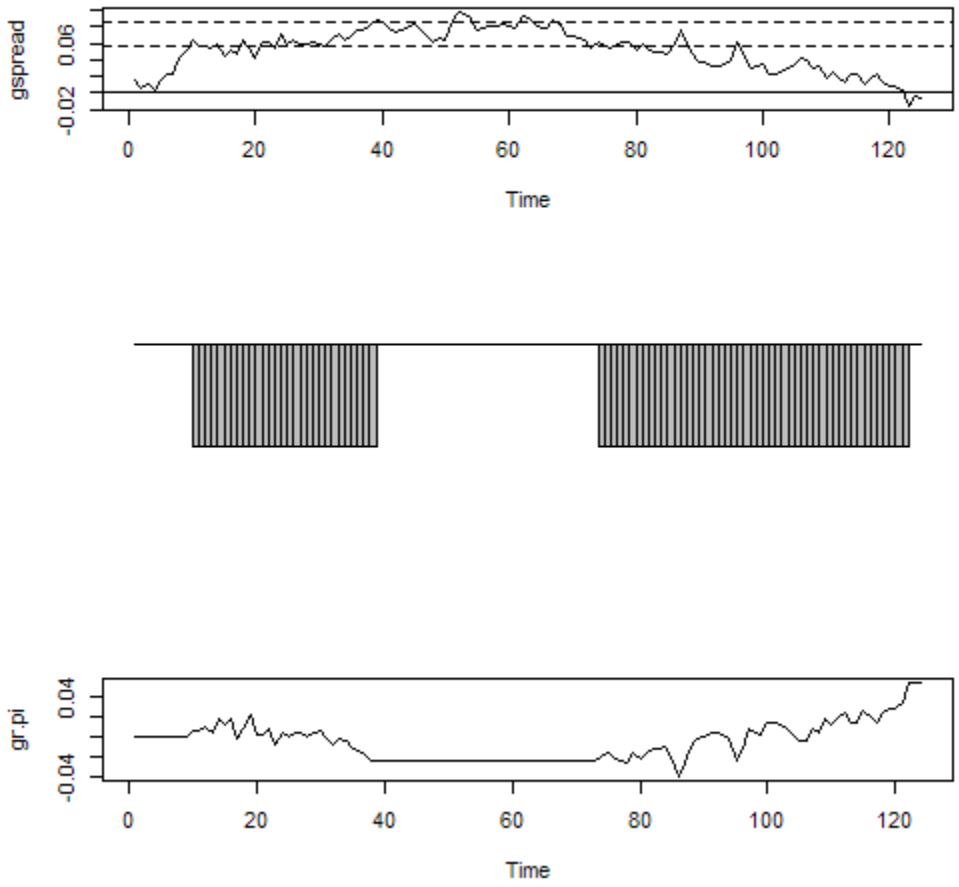

Table 1. Simulation results for $\gamma_1/\gamma_2 = 1$

Panel A						
Error Distribution: Normal						
Method	t	CV	Size	0.8	0.9	0.95
OLS	50	-3.4451	0.0464	0.1970	0.0846	0.0583
	100	-3.3570	0.0525	0.6744	0.2009	0.0926
	250	-3.3595	0.0465	1.0000	0.8424	0.2774
Rank	50	-3.3637	0.0446	0.1710	0.0822	0.0542
	100	-3.3008	0.0549	0.6008	0.1852	0.0890
	250	-3.3049	0.0500	0.9999	0.7917	0.2722
LAD	50	-3.7573	0.0468	0.1278	0.0786	0.0600
	100	-3.5673	0.0529	0.2971	0.1171	0.0738
	250	-3.2925	0.0565	0.8746	0.4409	0.1804
Panel B						
Error Distribution: Stable ($\alpha = 1.7$)						
Method	t	CV	Size	0.8	0.9	0.95
OLS	50	-3.5084	0.0463	0.1648	0.0706	0.0509
	100	-3.4543	0.0482	0.5907	0.1568	0.0719
	250	-3.3896	0.0498	0.9944	0.8009	0.2414
Rank	50	-3.3453	0.0497	0.3364	0.1229	0.0705
	100	-3.2876	0.0464	0.8731	0.4213	0.1557
	250	-3.1685	0.0483	0.9990	0.9844	0.7034
LAD	50	-3.7025	0.0534	0.2423	0.1111	0.0699
	100	-3.4337	0.0471	0.6320	0.3127	0.1304
	250	-3.0901	0.0498	0.9950	0.8765	0.5325
Panel C						
Error Distribution: Skewed stable ($\alpha = 1.7$, $\beta = -0.25$)						
Method	t	CV	Size	0.8	0.9	0.95
OLS	50	-3.4908	0.0483	0.1707	0.0749	0.0545
	100	-3.3929	0.0553	0.6367	0.1823	0.0757
	250	-3.3895	0.0524	0.9953	0.8187	0.2494
Rank	50	-3.3656	0.0456	0.3334	0.1219	0.0668
	100	-3.2249	0.0553	0.8942	0.4501	0.1688
	250	-3.1552	0.0484	0.9990	0.9861	0.7092
LAD	50	-3.6807	0.0518	0.2503	0.1138	0.0731
	100	-3.3901	0.0531	0.6487	0.3159	0.1417
	250	-3.1056	0.0472	0.9956	0.8756	0.5287

Table 2. Simulation results for $\gamma_1/\gamma_2 = 2$

Panel A						
Error Distribution: Normal						
Method	t	CV	Size	0.8	0.9	0.95
OLS	50	-3.4125	0.0520	0.2251	0.0975	0.0610
	100	-3.3722	0.0523	0.6991	0.2170	0.0988
	250	-3.3764	0.0451	1.0000	0.8542	0.2967
Rank	50	-3.3135	0.0516	0.2060	0.0918	0.0610
	100	-3.3197	0.0524	0.6251	0.2007	0.0937
	250	-3.3613	0.0489	1.0000	0.8044	0.2849
LAD	50	-3.7356	0.0508	0.1331	0.0766	0.0616
	100	-3.5960	0.0495	0.2980	0.1246	0.0796
	250	-3.3808	0.0491	0.8659	0.4184	0.1675
Panel B						
Error Distribution: Stable ($\alpha = 1.7$)						
Method	t	CV	Size	0.8	0.9	0.95
OLS	50	-3.4863	0.0466	0.1872	0.0792	0.0549
	100	-3.3956	0.0491	0.6753	0.1918	0.0850
	250	-3.3872	0.0559	0.9977	0.8548	0.2680
Rank	50	-3.3655	0.0457	0.3747	0.1518	0.0733
	100	-3.2332	0.0514	0.9086	0.4913	0.1994
	250	-3.145	0.0512	0.9998	0.9916	0.7491
LAD	50	-3.7836	0.0477	0.2628	0.1206	0.0705
	100	-3.3833	0.0505	0.6692	0.3492	0.1667
	250	-3.0873	0.0495	0.9964	0.8938	0.5585
Panel C						
Error Distribution: Skewed stable ($\alpha = 1.7$, $\beta = -0.25$)						
Method	t	CV	Size	0.8	0.9	0.95
OLS	50	-3.4606	0.0505	0.1983	0.0820	0.0542
	100	-3.4496	0.0462	0.6496	0.1801	0.0790
	250	-3.3655	0.0543	0.9977	0.8685	0.2737
Rank	50	-3.304	0.0559	0.4091	0.1559	0.0800
	100	-3.2002	0.0567	0.9116	0.5174	0.2110
	250	-3.1415	0.0510	0.9998	0.9914	0.7553
LAD	50	-3.6893	0.0543	0.2832	0.1268	0.0741
	100	-3.3937	0.0516	0.6650	0.3515	0.1570
	250	-3.1219	0.0489	0.9964	0.8785	0.5575

Table 3. Simulation results for $\gamma_1/\gamma_2 = 4$

Panel A						
Error Distribution: Normal						
Method	t	CV	Size	0.8	0.9	0.95
OLS	50	-3.4352	0.0494	0.2287	0.0955	0.0685
	100	-3.3874	0.0462	0.6993	0.2234	0.0963
	250	-3.3462	0.0503	1.0000	0.8747	0.3157
Rank	50	-3.3361	0.0513	0.2084	0.0929	0.0668
	100	-3.3513	0.0449	0.6134	0.1957	0.0896
	250	-3.3115	0.0506	1.0000	0.8253	0.2937
LAD	50	-3.8014	0.0473	0.1307	0.0753	0.0611
	100	-3.5844	0.0469	0.3063	0.1255	0.0751
	250	-3.3837	0.0466	0.8696	0.4285	0.1762
Panel B						
Error Distribution: Stable ($\alpha = 1.7$)						
Method	t	CV	Size	0.8	0.9	0.95
OLS	50	-3.4652	0.0519	0.1999	0.0874	0.0580
	100	-3.4146	0.0522	0.6849	0.1963	0.0789
	250	-3.3885	0.0532	0.9991	0.8651	0.2740
Rank	50	-3.3428	0.0503	0.4114	0.1724	0.0818
	100	-3.2155	0.0536	0.9154	0.5184	0.2168
	250	-3.1421	0.0541	0.9999	0.9922	0.7638
LAD	50	-3.6671	0.0529	0.2893	0.1385	0.0780
	100	-3.4183	0.0491	0.6668	0.3555	0.1677
	250	-3.0954	0.0490	0.9969	0.8873	0.5729
Panel C						
Error Distribution: Skewed stable ($\alpha = 1.7$, $\beta = -0.25$)						
Method	t	CV	Size	0.8	0.9	0.95
OLS	50	-3.4661	0.0495	0.2008	0.0849	0.0608
	100	-3.4563	0.0477	0.6604	0.1759	0.0792
	250	-3.4412	0.0475	0.9992	0.8503	0.2536
Rank	50	-3.3232	0.0515	0.4103	0.1719	0.0863
	100	-3.2654	0.0487	0.9031	0.4934	0.2137
	250	-3.1514	0.0489	1.0000	0.9923	0.7595
LAD	50	-3.7416	0.0469	0.2760	0.1350	0.0741
	100	-3.435	0.0469	0.6558	0.3446	0.1671
	250	-3.1023	0.0465	0.9970	0.8846	0.5736

Table 4. Number of cointegrated pairs identified

	OLS	Rank	LAD	OLS/Rank Overlap	Percent Overlap
2001	555	742	685	358	38.13
2002	719	854	626	514	48.54
2003	798	570	392	450	49.02
2004	798	618	439	517	57.51
2005	623	461	372	371	52.03
2006	396	289	240	207	43.31
2007	356	382	411	222	43.02
2008	646	825	605	450	44.07
2009	1388	1489	1210	1038	56.44
2010	429	459	399	275	44.86

Table 5. Average profitability

	OLS	Rank	LAD	OLS only	Rank only
2001	-2.98%	-2.41%	-3.22%	-3.46%	-1.89%
2002	4.59%	4.29%	2.63%	4.14%	2.94%
2003	0.74%	1.38%	1.25%	0.39%	1.89%
2004	-0.49%	0.44%	1.94%	-2.16%	2.32%
2005	0.22%	1.03%	1.25%	0.16%	2.12%
2006	0.60%	2.65%	0.96%	-1.13%	3.70%
2007	-3.13%	-3.30%	-3.11%	-1.87%	-2.11%
2008	2.15%	2.70%	1.92%	3.56%	3.42%
2009	2.41%	1.74%	2.17%	2.31%	1.01%
2010	-0.48%	-1.18%	-1.93%	-0.56%	-2.08%

Table 6. Regression results for portfolio returns, rank-based testing procedure

	Jensen's Alpha		Fama-French 3-Factor	
	Alpha	Prob	Alpha	Prob
2001	-0.074	0.219	-0.066	0.296
2002	0.072	0.014*	0.082	0.006*
2003	0.015	0.543	0.017	0.489
2004	0.009	0.347	0.007	0.511
2005	-0.003	0.921	0.001	0.976
2006	0.011	0.672	0.009	0.702
2007	-0.016	0.613	-0.016	0.596
2008	0.103	0.419	0.089	0.486
2009	0.035	0.451	0.049	0.288
2010	-0.029	0.404	-0.026	0.466

Table 7. Percent of losing pairs

	OLS	Rank
2001	19.28%	30.05%
2002	9.04%	8.78%
2003	15.41%	8.42%
2004	21.43%	6.31%
2005	20.22%	10.63%
2006	24.49%	12.80%
2007	21.35%	21.20%
2008	15.94%	15.88%
2009	8.65%	8.46%
2010	17.95%	24.18%