

Replication of : Statistical Arbitrage and High-Frequency Data with an Application to Eurostoxx 50 Equities

Nick Vintila

July 3rd, 2015

Contents

Data Summary	1
Paper Summary	2
Literature Review	5

Data Summary

As used in paper

The authors used Euro Stoxx 50 data at both daily and intraday frequencies

- Daily for from 01 Jan 2000 to 17 Nov 2009
- Intraday from 03 Jul 2009 03-07-2009 to 17 Nov 2009 at 5 min, 10 min, 20 min, 30 min, 60 min frequencies

Data was adjusted for dividend payments and splits.

The authors used the intraday data as ‘out of sample’ and the daily for “in sample”.

As obtained by me

From studying the paper it was quite clear that * there is too little data for “out of sample” verification (1 to 10 ratio) * there is an opportunity to further improve the strategy by checking the real demand at the time the signals are triggered (bid/ask, volume, and, in the long run, the limit order book)

As such, I decided to get * minute data for all 50 stocks for 10 years * trades and quotes data for 1-2 years for the 6 pairs of stocks the PCA analysis will select to be accounting for 97% of the variance.

I also excluded “UNIBAIL-RODAMCO” since it is the only one in it’s category so it cannot form a pair as per the paper approach.

For the same reason, included “ESSILOR INTERNATIONAL” even though it was not used by the authors (but it is currently in the index).

In terms of provider I chose TickData.com because they provide tools to convert data for corporate actions (but also because QuantGo.com did not offer access in time).

Potential issues

- Currency: different stocks might be priced in currencies different than EUR : must check individually
 - Not clear if the authors use actual bid/ask spreads or not : on one hand they say “we do not dispose of bid and ask prices” and on the other, they use “minute data” (the tickdata.com definition of “minute data” may not contain bid/asks but just trades only)
-

Tactical purchase of data

- the high cost of the data
- the above unknowns
- the fact that the rationale above is not yet verified by Brian

I chose to acquire part of the data for now and acquire more later if indeed Brian considers the approach valid and the potential worthwhile.

If indeed the daily data is enough then may choose to use free daily data from elsewhere.

Overview of data collection approach

- Compare current EURO STOXX 50 index composition and compare with the composition at the time the paper was written
 - Document differences in index composition
 - Decide the selection of stocks to obtain in order to make paper replication as faithful as possible
 - Locate the data (daily and intraday/minute) for the stocks to use
 - Ensure stock is properly identified using Quant Finance and stoxx.com (disambiguate between different symbols used by different exchanges; ended up using ISIN numbers)
 - Check which historical data providers : tickdata.com, quantquote.com, quantgo.com, eoddata.com.
 - Download and verify data
 - If stock data is obtained from different vendors or from different exchanges then convert to EUR (how (question))
 - If data is obtained from different vendors then ensure it is consistent across the various frequencies.
-

Paper Summary

Introduction

Argue for the presence of arbitrage profitability in a selection of EuroStoxx 50 equities when using a novel approach combining cointegration at HFT and daily levels, time adaptable beta and a test to select pairs that account for most of the variability in the set of stocks.

Main points

- Pairs grouped by industry sector are sometimes cointegrated and can be traded with pair-trading strategies
- Using a Kalman filter to determine a time adaptable beta is superior to using OLS or DESP
- Using sampling frequencies higher than daily for cointegration tests offers investment characteristics that are more attractive than when using daily data
- Using PCA to subselect pairs which account for most of the variability improves returns

Conclusion Statistical arbitrage at daily frequencies in EuroStoxx 50 is not attractive but testing cointegration at high frequencies, using a time adaptive beta and filtering stocks with maximum variability makes strategies attractive.

Main techniques

1. Group stocks in the Eurostoxx 50 index by industry groups. Within each industry group create pairs of stocks
2. For each pair calculate time adaptive/varying betas using Kalman filters

Note: OLS (ordinary least squares) and DESP (double exponential smoothing prediction) are dropped in favor of Kalman filters.

3. The traditional simple strategy :
 - enter both positions when spread > 2 standard deviations
 - exit both positions when spread returned < 0.5 standard deviations of long term mean

Key finding Results are not attractive

4. Calculate the information ratio of the series

$$AnnualizedInformationRatio = (R/\sigma)\sqrt{Hourstradedperday * 252}$$

5. Obtain the t-stat of the ADF test of the series sampled at daily interval
6. Build a diversified portfolio of 5 pairs with the best indicator value

Key finding An indicator made from a combination of the two above produces information ratios of

- over 3 for high frequencies
- 1.3 for daily frequency

Compares favorably with the performance of Eurostoxx 50 index and the index of Market Neutral Hedge Funds.

7. Calculate the cointegration of pairs using Engle and Granger (chosen vs. the Johansen test due to simplicity and lower variance)

$$Y_t = \beta X_t + \epsilon_t$$

8. Test the hypothesis of positive correlation between the t-stat of the ADF test on the OLS residuals and the out of sample “information ratio” (uses bootstrap); the test is a PCA on a normalized matrix of t-stats from the ADF test for 176 pairs x (5, 10, 20, 30, 60) frequencies

Key finding In sample t-stat at 5min and 10 min frequencies have certain predictive power for the out of sample information ratio.

The first principal component explains 97% of the variation in the data for the overall set of pairs in the index.

9. Use the above indicators to trade the enhanced strategy on the subset of 5 pairs filtered through the PCA test.

Key finding The extra indicators heavily improved the attractiveness of the results for the pairs sampled at the high-frequency intervals.

- Almost all the information ratios are above 2 when traded at high frequencies.
- Average maximum drawdown at high frequencies is 1.58% and 4.26% at daily frequency.
- Average maximum drawdown duration at high frequencies is 22 days at high-frequencies, and 55 days for daily frequency.

Issues and possible improvements

“For daily data, the in-sample period is much longer than the out-of-sample period.” Too little out of sample data.

“For high-frequency data the in- and out-of-sample periods have the same lengths” Data is split just in two sets. Needs more sophisticated cross validation, k-fold validation, etc.

The HFT data used in the paper spans 4 months for Summer and Fall. This makes the results to seasonality of the spread. A solution would be to test with larger out of sample HFT data and seek patterns of seasonality. (“seasonal version of the Ornstein-Uhlenbeck formula”: <https://www.evernote.com/shard/s175/nl/19543423/4ad339f8-c347-43a4-a9c5-04bedd77c3e8>) Additionally, add a momentum component to the strategy to go around the assumption of mean reversion.

Misc Formulae

Spread

$$z_t = P_{Y_t} - \beta_t P_{X_t}$$

Halflife

$$Halflife = -\ln(2)/k$$

O-U equation

$$dz_t = k(\mu - z_t)dt + \sigma dW_t$$

Returns

$$Ret_t = \ln(P_{X_t}/P_{X_{t-1}}) - \ln(P_{Y_t}/P_{Y_{t-1}})$$

Literature Review

(Alexander and Dumitriu) makes the case for use of cointegration over correlation. Optimization models for benchmark replication focus on minimizing the variance of the tracking error.

There are a number of drawbacks to optimization models based on tracking error or on correlation measures.

1. Minimizing tracking error with respect to an index that is a linear combination of stock prices may result in a portfolio that is very sample-specific and unstable under volatile market circumstances.
2. Additional limitations relate to the very nature of correlation as a measure of dependence; it is applicable only to stationary variables such as stock returns, so the method requires prior detrending of level variables (i.e., stock prices) and has the disadvantage of losing valuable information (i.e. the common trends in prices). Correlation is thus only a short-term statistic, and it lacks stability.
3. Finally, depending on the model used to estimate it, correlation can be very sensitive to the presence of outliers, non-stationarity, or volatility clustering, which limit the use of a long data history

Cointegration has gained far wider acceptance than correlation. According to Lucas's results, the presence of cointegration relations also has important consequences for the short-term predictability of asset prices.

(Dunis and Ho 2005) Over the last ten years, quantitative portfolio managers have made increasing use of conditional correlations. This paper uses the concept of cointegration which relies on the long-term relationship between time series, and thus assets, to devise quantitative European equities portfolios.

The results show that the designed portfolios are strongly cointegrated with the benchmark and indeed demonstrate good tracking performance. In the same vein, the long/short market neutral strategy generates steady returns under adverse market circumstances but, contrary to expectations, does not minimise volatility.

(Miao)

The proposed pairs trading system was applied to equity trading in U.S. equity markets in any type of market cycle condition to capture statistical mispricing between the prices of each stock pair based on its residuals and to model the stock pairs naturally as a mean-reversion process.

The trading strategy yields cumulative returns up to 56.58% for portfolios of stock pairs, well exceeding the S&P 500 index performance by 34.35% over a 12-month trading period.

The trading system might be especially more profitable at times when the U.S. stock market performed poorly.

The cumulative net profit of the proposed pairs trading system during the 12 months of the out-of-sample test periods was 56.58%, with an average monthly return of 3.82% and a standard deviation of 1.40%.

(Triantafyllopoulos and Montana) Gaussian linear state-space processes have recently been proposed as a model for such spreads under the assumption that the observed process is a noisy realization of some hidden states.

Real-time estimation of the unobserved spread process can reveal temporary market inefficiencies which can then be exploited to generate excess returns. This work employs the state-space framework for modeling spread processes and extend this methodology along three different directions.

1. Introduce time-dependency in the model parameters, which allows for quick adaptation to changes in the data generating process.
2. Provide an on-line estimation algorithm that can be constantly run in real-time. The algorithm is particularly suitable for building aggressive trading strategies based on high-frequency data and may be used as a monitoring device for mean-reversion.
3. The framework provides informative uncertainty measures of all the estimated parameters. Experimental results based on Monte Carlo simulations and historical equity data are discussed, including a cointegration relationship involving two exchange-traded funds.

References

Alexander, C., and A Dumitriu. “Indexing and Statistical Arbitrage.”

Dunis, C. L., and R. Ho. 2005. “Cointegration Portfolios of European Equities for Index Tracking and Market Neutral Strategies.” *Journal of Asset Management* 6 (1): 33–52.

Miao, George. “High Frequency and Dynamic Pairs Trading Based on Statistical Arbitrage Using a Two-Stage Correlation and Cointegration Approach.”

Triantafyllopoulos, K., and G. Montana. “Dynamic Modeling of Mean-Reverting Spreads for Statistical Arbitrage.”