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

Nick Vintila

July 1st, 2015

Contents

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

- The extent to which a pair of stocks in EuroStoxx 50 is cointegrated at daily frequencies is a good indicator for profitability at higher frequencies if the pair is selected by a combination of criteria involving cointegration.
- using high-frequency data leads to higher potentially achievable information ratio compared to the use of daily closing prices

Issue: implies causality but only shows "association".

Main techniques

1. Group stocks in the Eurostoxx 50 index in industry groups Q: How to reproduce the groups?
2. Within each industry group create pairs of stocks Q: Does it matter how pairs are made or create all the combinations?
3. 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.

4. Trade the following 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

5. Calculate the information ratio of the series (what frequency?)
6. Obtain the t-stat of the ADF test of the series sample at daily interval
7. 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.

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

. . . ?

9. Test for a positive correlation between the t-stat of the ADF test on the OLS residuals and the out of sample "information ratio" (uses bootstrap); 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.

10. Use the 2 above indicators to select a subset of 5 pairs

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.