# ACADEMIC RESEARCH REVIEW

**Statistical Arbitrage in US Equities Market**
Avellaneda, Marco and Lee, Jeong-Hyun
*Quantitative Finance*
*Volume 10, Issue 7, 2010*

NUS INVESTMENT SOCIETY

***Research Analyst: Sylvanus Quek***

**Introduction**

Statistical arbitrage refers to a family of market-neutral strategies that employs systematic trading signals based on statistics rather than fundamentals. By making many such bets with positive expected returns across stocks, the expected returns are accompanied by lower risk due to diversification.

**Model**

Stock returns, $R_i$ can be decomposed into their systematic ($\sum_{j=1}^{m} \beta_{ij} F_j$ with $m$ factors) and idiosyncratic components ($\tilde{R}_i$); and statistically modelling the latter[1]:

$$R_i = \sum_{j=1}^{m} \beta_{ij} F_j + \tilde{R}_i = \sum_{j=1}^{m} \beta_{ij} F_j + \alpha \mathrm{d}t + \mathrm{d}X_t$$

The idiosyncratic component or residual $X_t$ is further parameterised by the Ornstein-Uhlenbeck process[2]. For a market-neutral portfolio of $N$ stocks, the systematic component disappears by allocating dollar amounts $Q_i$ such that $\sum_{i=1}^{N} \beta_{ij} Q_i = 0, j = 1,2 \ldots, m$.

**Methodology**

The authors investigated market-neutrality using Principal Component Analysis (PCA) and industry-sector Exchange Traded Funds (ETFs). PCA is able to extract uncorrelated risk factors of economic significance from the correlation matrix as they can be viewed as long-short portfolio of industry sectors, i.e. eigenportfolios. In contrast, ETFs may be biased towards large-capitalization stocks due to the capitalization-weighted nature of ETFs. In the ETF method, the returns of sector ETFs are the factors.

*Trade Signal generation*

Essentially, the trading signal is constructed based on the residual. The mean-reverting s-score without drift is given by $s_i = [X_i(t) - E(X_i(t))]/SD(X_i(t))$ while the modified s-score incorporating drift is further dependent on the speed of mean-reversion.

Since an s-score near zero is when the residual is at equilibrium, cutoff values denoting disequilibrium can be determined empirically. When the s-score is negative below a certain threshold, it indicates a buy signal; that position is closed when the s-score is above another larger but still negative value. The converse applies for a sell signal.

**Back-testing Results**

The parameters of the residual model, including whether the drift is statistically significant, are estimated using a 60-day rolling window. All trades are done at closing prices. A round-trip transaction cost of 10 basis points was imposed to make the simulations more realistic. The trading strategies were back-tested over a period from 1997 to 2007. If the ETFs were not launched during the timeframe of back-testing, synthetic ETFs were constructed as proxy. Market neutrality was ensured by hedging using the S&P 500 tracker. It is evident from figure 1 that the PCA-based approach is superior to the ETF-based

[1] $\alpha$ is the drift which is assumed to be small, $\beta_{ij}$ is the beta coefficient and $X_t$ is a stationary process, termed "residual".
[2] Well-known mean-reverting stochastic process.

approach. However, in 2007, the statistical arbitrage strategy suffered a large drawdown similar to the experience of many quantitative hedge funds during that time. That being said, the PCA-based approach was still more resilient than its ETF counterpart in 2007.

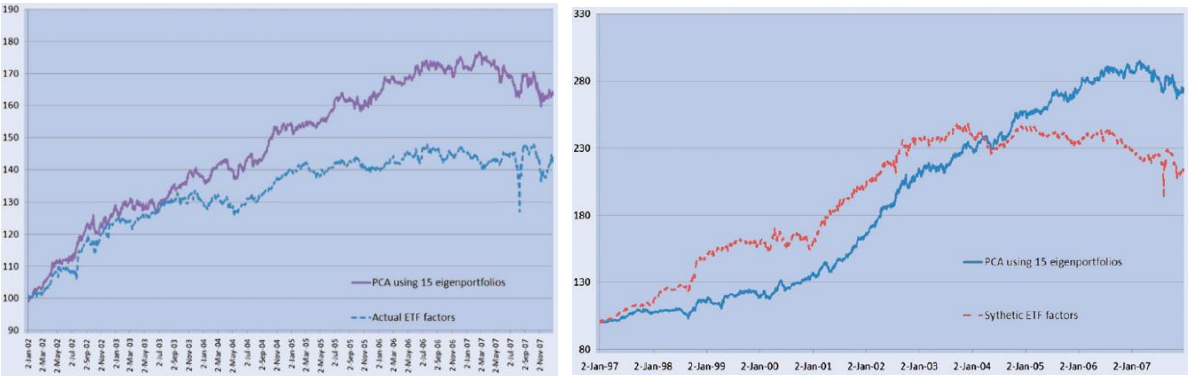


Figure 1: Comparison of Strategies. Left: PCA using 15 eigenportfolios (purple) against actual ETF factors (blue) from 2002 to 2007. Right: PCA using 15 eigenportfolios (blue) against synthetic ETF factors (red) for 1997 to 2007.

The PCA-based risk factors were observed to vary over time with the business cycle unlike the ETF method. Hence, the authors attempted to find the number of eigenportfolios needed to explain a given level of variance in stock returns. Interestingly, a low number of eigenportfolios is needed when equity risk premium is high like the aftermath of crashes. During these periods, the variance is mainly attributable to a few eigenportfolios.

The performance of strategies with variable number of eigenportfolios explaining different levels of variance (45%, 55%, 65% and 75%) were also compared. The one that explained 55% of variance was determined to be the best performing.

*Volume Data Integration*

The role of volume was studied by modifying the returns to include volume information. If a rally is underway with high volume, modified returns will be lower and therefore would not necessarily trigger a sell signal. The back-test results demonstrated that volume integration increased profits and Sharpe ratios for ETF-generated signals, but not PCA-based strategies.

**Conclusion**

A simple statistical arbitrage trading strategy based on mean-reversion process was found to be feasible over the period from 1997 to 2007. However, it should be cautioned that large drawdowns are still possible during financial crashes. The strategies studied worked better when trading volume is accounted for and the number of explanatory factors or eigenportfolios is relatively small.

---