Genetic Programming in Statistical Arbitrage

Philip Saks and Dietmar Maringer

Centre for Computational Finance and Economic Agents, University of Essex

Abstract. This paper employs genetic programming to discover statistical arbitrage strategies on the banking sector in the Euro Stoxx universe. Binary decision rules are evolved using two different representations. The first is the classical single tree approach, while the second is a dual tree structure where evaluation is contingent on the current market position. Hence, buy and sell rules are co-evolved. Both methods are capable of discovering significant statistical arbitrage strategies.

1 Introduction

During the last decades, the *Efficient Market Hypothesis* (EMH) has been put to trial, especially with the emergence of behavioral finance and agent-based computational economics. The rise of the above mentioned fields provides a theoretical justification for attacking the EMH, which in turn stimulates the empirical forecasting literature.

A basic premise for efficiency is the existence of *homo economicus*, that the markets consists of *homogeneous rational* agents, driven by utility maximization. However, cognitive psychology has revealed that people are far from rational, instead they rely on *heuristics* in decision making to simplify a given problem [15]. This is both useful and necessary in everyday life, but in certain situations it can lead to biases such as *overconfidence*, *base-rate neglect*, *sample-size neglect*, *gamblers fallacy*, *conservatism* and *aversion to ambiguity* [3]. However, what is more important is that these biases manifest themselves on an aggregate level in the markets as momentum and mean-reversion effects [12, 7]. Accepting the existence of *heterogeneous* agents, have pronounced effects on a theoretical level. In such a scenario the market clearing price cannot be determined formally, since agents need to form expectations about other agents' expectations etc. This leads to an "infinite regress in subjectivity", where no agent irrespective of reasoning powers, can form expectations by deductive means, thus perfect rationality is not well-defined. Instead, investors are forced to hypothesize expectational models, where the only means of verification is to observe the models' performance in practice [2]. In such a world it is indeed sensible to develop expectational models beyond traditional equilibrium analysis which we seek to accomplish using *genetic programming* (GP).

The majority of existing applications of GP in financial forecasting have for some reason focused on foreign exchange. Here, the general consensus is that GP can discover profitable trading rules at high frequencies in presence of transaction costs [14, 8, 6]. For the stock market results are mixed. [1] do not outperform

M. Giacobini et al. (Eds.): EvoWorkshops 2008, LNCS 4974, pp. 73–82, 2008.

the buy-and-hold strategy on daily S&P500 data, while [5] do on a monthly frequency. Besides changing the frequency, they reduce the grammar and consider a cooperative co-evolution scheme, where buy and sell rules are evaluated separately.

In this paper, we consider genetic programming for statistical arbitrage. Arbitrage in the traditional sense is concerned with identifying situations where a self-funding is generated that will provide only non-negative cash flows at any point in time. Obviously, such portfolios are possible only in out-of-equilibrium situations. Statistical arbitrage is a wider concept where, again, self-funding portfolios are sought where one can expect non-negative pay outs at any point in time. Here, however, one accepts negative pay-outs with a small probability as long as the expected positive payouts are high enough and the probability of losses is small enough; ideally this shortfall probability converges to zero. In practice, such a situation can occur when price processes are closely linked. In the classical story of Royal Dutch and Shell [3], the pair of stocks are cointegrated since they are fundamentally linked via their merger in 1907. In most cases, however, such links are not as obvious, but that does not eliminate the possibility that such relationships might exist and can be detected by statistical analysis. Since it can be argued that these stocks are exposed to many of the same risk factors and should therefore have similar behavior, this paper considers stocks within the same industry sector.

We shall construct so-called arbitrage portfolios, where the proceedings from short selling some stocks are used to initiate long positions in other stocks. This scheme has several advantages, firstly, it is self-financing and second the profits made from this strategy are in excess of the risk-free rate and virtually uncorrelated with the market index. Furthermore, by modeling the relationships between stocks, we are not trying to predict the future, rather we focus the attention in a direction where more stable patterns should exist.

The rest of the paper is organized as follows. Section 2 provides evidence of significant clustering between sectors within the Euro Stoxx universe. Section 3 introduces the data, model and framework. Section 4 presents results under the assumptions of frictionless trading. Finally, conclusions are drawn and pointers are given to future research in Section 5.

2 Clustering of Financial Data

Previously we hypothesized that stocks within the same industry sector are exposed to many of the same risk factors and should therefore have similar behavior. In order to clarify this, we investigate the majority of stocks in the Euro Stoxx 600 index. The data is gathered from Bloomberg and includes information such as company name, ticker symbol, industry sector and industry group. In addition hereto, we obtain the adjusted closing prices in the period from 21-Jan-2002 to 26-Jun-2007. Since the index composition is changing over time, we only consider stocks where data exists for the last two years for both price and volume series. Taking this into account, the universe comprises of a total of 477 stocks.

The notion that stocks have similar behavior needs to be specified in order to conduct a proper analysis. An obvious measure for price data is the correlation of returns, where a higher correlation implies stronger similarity. To test the hypothesis that stocks within the same sector tend to be clustered together we employ the *k-means algorithm* to construct statistical clusters. Thus, if the statistical clustering is independent of the fundamental clustering, dictated by the industry sectors, then we must reject our hypothesis. Let, S and F be two stochastic variables, which describe the statistical and fundamental clusters.[1] The maximum number of clusters is denoted by the integers k_s and k_f, respectively. Define s_i (f_i) as the statistical (fundamental) cluster asset i belongs to, and $I_{\{s_i=j\}}$ ($I_{\{f_i=m\}}$) be a binary indicator which is 1 if this statistical (fundamental) cluster is equal to j (m) and 0 otherwise. Then, for a universe of N stocks,

$$V_{j,m} = \sum_{i=1}^{N} I_{\{s_i=j\}} \cdot I_{\{f_i=m\}} \quad \forall \quad j = 1, 2, \ldots, k_s \quad m = 1, 2, \ldots, k_f \qquad (1)$$

is the number of stocks that, at the same time, belong both to statistical cluster j and the fundamental factor m. The hypothesis of independence can be tested via a χ^2-statistic for contingency tables. Setting $k_s = k_f = 10$, we obtain a test statistic $\chi^2 = 1314.1$ where the critical value is $\chi^2_{0.05}(81) = 103.0$, thus strongly rejecting the null hypothesis of independence.

The analysis above proves that there are significant clustering within the sectors, and in the following we shall focus on the actual statistical arbitrage application.

3 The Framework

As mentioned previously, the objective is to develop a trading strategy for statistical arbitrage based on price and volume information, and in the following we elaborate on data, preprocessing and model construction.

3.1 Data

The data comprises Volume-Weighted Average Prices (VWAP) and volume, sampled on an hourly frequency for the banks in the Euro Stoxx 600 index. It covers the time period from 01-Apr-2003 to 29-Jun-2007, corresponding to a total of 8648 observations. Again, we only consider stocks for which we have enough data, which limits the portfolio to 30 assets.

When analyzing high frequency data it is important to take intraday effects into account, e.g., the intraday volume is higher after open and before close than during the middle of the day [11]. In the context of trading rule induction it is important to remove this bias, which is basically a proxy for the time of day, and prohibits sensible conditioning on intraday volume.

[1] The statistical clusters are obtained by using k-means on the correlation matrix, i.e., the data comprises of N observations in N dimensional space.

3.2 Preprocessing

The return series for each stock is standardized with respect to its volatility, estimated using simple exponential smoothing. Likewise, a volume indicator is constructed that removes the intraday bias, and measures the extent to which the level is lower or higher than expected. Specifically, we take the logarithm of the ratio between the realized and expected volume, where this ratio has been hard limited in the range between 0.2 and 5.

Since we are interested in cross-sectional relationships between stocks, rather than their direction, we subtract the cross-sectional average from the normalized returns and volume series for each stock. Based on these series, we calculate the moving averages over the last 8, 40 and 80 periods, corresponding to one day, a week and two weeks on the hourly frequency. All indicators are scaled in the range between 0 and 1.

3.3 Model

There are two approaches for modeling trading rules, either as decision trees where market positions or actions are represented in the terminal nodes [16], or as a single rule where the conditioning is exogenous to the program [6].

We consider the latter approach in the context of a binary decision problem, which corresponds to long and short positions. As mentioned previously, we are interested in arbitrage portfolios, where the purchase of stocks is financed by short selling others. Naturally, a precondition for this to be achieved is that not all the forecasts across the 30 stocks are the same, e.g., if the trading rule take a bullish view across the board, then short-selling opportunities have not been identified and proper arbitrage portfolios cannot be constructed. In this case we do not hold any stocks. However, when forecasts facilitate portfolio construction, this is done on a volatility adjusted basis. Let $o_t^i \in \{-1, 1\}$ denote the forecast on stock i at time t, then the holding is given,

$$h_t^i = o_t^i \frac{\frac{1}{\sigma_t^i}}{\sum_{j=1}^{n} \frac{1}{\sigma_t^j} I_{\{o_t^j = o_t^i\}}} \tag{2}$$

where σ_t is the volatility, n is the number of stocks in the universe, and $I_{\{o_t^j = o_t^i\}}$ is an indicator variable that ensures that forecasts are normalized correctly, i.e., it discriminates between long and short positions.

We employ two different methods for solving the binary decision problem. The first uses a standard single tree structure, while the second follows [4] and considers a dual structure in conjunction with cooperative co-evolution. In both methods, the trees return boolean values. For the dual structure, program evaluation is contingent on the current market position for that particular stock, i.e., the first tree dictates the long entry, while the second enters a short position. More formally, let $b_t^{j,i} \in \{0, 1\}$ be the the truth value for tree j on stock i at time t, then the forecast is given as, **if** $o_{t-1}^i < 0$ **then** $o_t^i = 2 \cdot b_t^{1,i} - 1$ **else** $o_t^i = -2 \cdot b_t^{2,i} + 1$ **end**.

In both settings, the programs are constructed from the same grammar which is fairly restricted. It consists of numeric comparators ($<$, $>$), boolean operators (`AND`, `OR`, `XOR`, `NOT`) and *if-then-else* statements (`ITE`). Furthermore, we have introduced a special function `BTWN`, that takes three arguments and evaluates if the first is between the second and third. The terminals comprises the six indicators and numerical constants ranging from 0 to 1. The parsimonious grammar reduces the risk of overfitting, and enhances interpretability of the evolved solutions.

3.4 Objective Function

The choice of objective function is essential in evolutionary computation. It has previously been found that a risk-adjusted measure improves out-of-sample performance, relative to an absolute return measure [6]. In this context the Sharpe Ratio is an obvious candidate. However, using this measure it is possible to evolve strategies that do extremely well only on a subset of the in-sample data and mediocre on the remainder.

Instead we employ the *t-statistic* of the linear fit between cumulated returns and time, since it maximizes the slope while minimizing the deviation from the ideal straight line performance graph.

3.5 Parameter Settings

In the following experiments we consider a population size of 250 individuals, initialized using the *ramped half-and-half* method. It evolves for a maximum of 51 generations, but is stopped after 15 generations if no new *best-so-far* individual has been found. We use normal tournament selection with a size of 5, and the crossover and mutation probabilities are 0.9 and 0.1, respectively. Moreover, the probability of selecting a function node during reproduction is 0.9, and the programs are constrained to a maximum complexity of 50 nodes.

The data is split into a training and test set. The former contains 6000 samples and covers the period from 01-Apr-2003 to 10-Mar-2006, and the latter has 2647 samples in the period from 13-Mar-2006 to 29-Jun-2007.

4 Empirical Results

In this section we assume the absence of market impact, i.e., that it is possible to execute on the realized VWAP. Trading on VWAP differs from a traditional market order, where a trade is executed at the current observed price. Contrary, the VWAP is a backward looking measure, and it is therefore not possible to trade on the observed VWAP at time t. Instead the execution occurs gradually between t and $t+1$, resulting in the VWAP at $t+1$. In summary, a trading decision is formed based on the VWAP at time t, the entry price is observed at time $t+1$ and the one period return is evaluated at $t+2$.

We perform 10 experiments using both the single and dual tree method, according to the settings outlined in Section 3. For each experiment, the *best-so-far* individual is evaluated on the training and test set. Figure 1 shows the

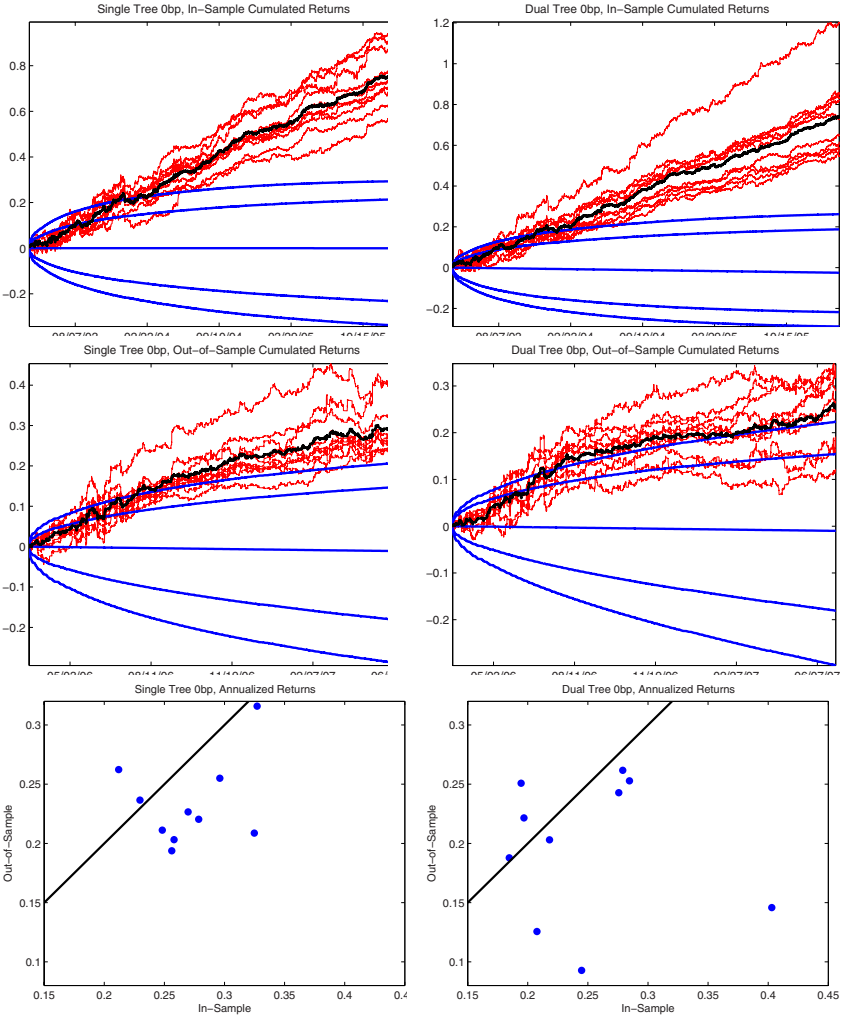

Fig. 1. In-sample (top) and out-of-sample cumulated returns (center) assuming frictionless trading. The thick lines are the average performances and the 95% and 99% confidence intervals are constructed using the stationary bootstrap procedure. Annualized in-sample versus out-of-sample returns with the 45°-line (bottom). The left and right column are the single and dual tree results, respectively.

performance for the single and dual tree method. Judging from the in-sample results, the *t-statistic* measure works as intended, and all experiments have steady increasing cumulative returns. The average annualized in-sample returns are 27.0% and 24.9%, for the single and dual tree method, respectively. As a proxy for generalization, it is instructive to consider the *shrinkage* which is defined,

$$\psi = \frac{X_{\text{train}} - X_{\text{test}}}{X_{\text{train}}} \tag{3}$$

where X is an arbitrary performance measure [6]. Both methods generalize extremely well out-of-sample, and the shrinkage is only 0.12 and 0.16, corresponding to annualized returns of 23.3% and 19.9%.

In order to asses the significance of these results, confidence intervals are constructed using the *stationary bootstrap* method, which is a superior alternative to well known *block bootstrap* procedure [17]. Instead of using a fixed block size, it varies probabilistically according to a geometric distribution.[2] Thus, sampling with replacement is performed from the holdings, and statistics are gathered from 500 bootstrap runs. The 99% upper confidence limits are 16.1% and 17.7% for the single and dual tree method, respectively. For the single tree method all ten experiments exceed the limit, while for the dual tree method it is only 7. From a risk adjusted perspective, there is also significant performance. The average annualized out-of-sample Sharpe Ratio is 2.83 and 2.60, where the 99% confidence limits are 2.17 and 2.41, for the single and dual tree method, respectively.

Positive out-of-sample returns need not imply market inefficiency. Traditionally, this is investigated by comparing the trading strategy to the buy-and-hold strategy. This, however, is not a suitable benchmark for statistical arbitrage strategies. Firstly, because a statistical arbitrage is self-financing and the buy-and-hold is not. This could be addressed by considering the returns of the buy-and-hold in excess of the risk-free rate, but this is a naive approach contingent on a specific equilibrium model.[3] Instead, we employ a special statistical test for statistical arbitrage strategies where this is not the case [10]. Hence, it circumvents the joint hypothesis problem, that abnormal returns need not imply market inefficiency, but can be due to misspecification of a given equilibrium model [9]. The constant mean version of the test assumes that the discounted[4] incremental profits satisfy,

$$\Delta v_i = \mu + \sigma i^{\lambda} z_i \qquad i = 1, 2, \ldots, n \tag{4}$$

where $z_i \sim N(0,1)$. The joint hypothesis, H1 : $\mu > 0$ and H2 : $\lambda < 0$ determines the presence of statistical arbitrage.

In order to make some general inferences, we consider an aggregation or bagging of the evolved trading strategies in the following.[5] Table 1 contains the test

[2] With the probability parameter $p = 0.01$, blocks with an expected length of 100 samples are generated.

[3] For the interested reader, the benchmark of an equally weighted portfolio of the banking stocks in excess of the risk free rate obtains an annualized return of 4.47% and has a Sharpe Ratio of 0.35 in the out-of-sample period.

[4] As a discount rate we employ the 1-month LIBOR rate for the Eurozone.

[5] All 10 evolved strategies are aggregated, but one could employ various schemes to improve out-of-sample performance. Generally, this requires the use of additional validation sets, but since data is limited this is problematic. Moreover, aggregating all strategies is clearly the conservative approach and is therefore preferred.

Table 1. Statistical arbitrage test results

	μ ($\cdot 10^{-4}$)	σ	λ	H1	H2	H1+H2 ($\cdot 10^{-6}$)
Single Tree	1.1321	0.0018	-0.0471	0.0000	0.0000	3.75
Dual Tree	0.9256	0.0020	-0.0922	0.0000	0.0000	4.78

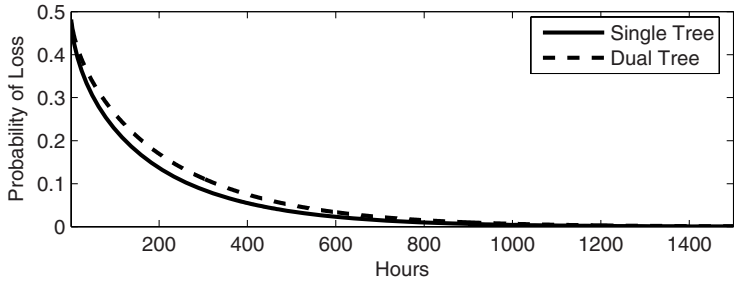

Fig. 2. Probability of loss for the bagged statistical arbitrage strategies as a function of trading time

results for the bagged models, where p-values for the joint hypothesis is obtained via the *Bonferroni* inequality. Both methods discover highly significant statistical arbitrage dynamics at all usual levels of significance.[6] Figure 2 depicts the probability of a loss as a function of trading time, and after 422 and 502 hours it is less than 0.05 for the single and dual tree, respectively. This demonstrates the essence of statistical arbitrage, that a riskless profit is earned in the limit.

5 Conclusion

In this paper genetic programming is employed to evolve trading strategies for statistical arbitrage. This is motivated by the fact that stocks within the same industry sector should be exposed to the same risk factors and should therefore have similar behavior. This certainly applies to the Euro Stoxx universe, where we find evidence of significant clustering.

Traditionally there has been a gap between financial academia and the industry. This also applies to statistical arbitrage, an increasingly popular investment style in practice, but to the authors' knowledge little formal research has been undertaken within this field. This paper addresses this imbalance and aims to narrow this gap. We consider two different representations for the trading rules. The first is a traditional single tree structure, while the second is a dual tree structure in which evaluation is contingent on the current market position. Hence, buy and sell rules are co-evolved. Both methods have substantial market timing and discovers significant statistical arbitrage strategies. However, in a

[6] For the individual strategies, 9/10 and 7/10 have significant performance on a 0.01 level of significance, for the single and dual tree method, respectively.

frictionless environment we cannot conclude that this violates the efficient market hypothesis, since "prices reflect information to the point where the marginal benefits of acting on information (the profits to be made) do not exceed the marginal costs" [13].

Having assumed a frictionless environment it is important to asses the implications of such an assumption. Naturally, were these strategies to be implemented in practice a cost would be associated with trading, but it is not a serious as one might think. When closing prices or point estimates of the price are considered, trading is associated with crossing the bid-ask spread, and thereby incurring a cost ranging from 2bp to as much as 40bp depending on the liquidity and the price level of the stock. In addition to the bid-ask spread, *slippage* occurs if the order size is large relative to the depth of the market. This paper, however, does not use closing prices, but instead proprietary VWAP's. The advantage of using these series is that algorithmic trading actually permits these prices to be obtained within 1bp for moderate order sizes.[7] Naturally, the assumption of a complete frictionless environment is an idealization. However, preliminary research suggest that positive returns can be generated under realistic market impact, but more work is needed.

References

[1] Allen, F., Karjalainen, R.: Using genetic algorithms to find technical trading rules. Journal of Financial Economics 51, 245–271 (1999)
[2] Arthur, B., Holland, J.H., LeBaron, B., Palmer, R., Tayler, P.: Asset pricing under endogenous expectations in an artificial stock market, Technical report, Santa Fe Institute (1996)
[3] Barberis, N., Thaler, R.: Handbook of the Economics of Finance, pp. 1052–1090. Elsevier Science, Amsterdam (2003)
[4] Becker, L.A., Seshadri, M.: Cooperative coevolution of technical trading rules, Technical report, Department of Computer Science, Worcester Polytechnic Institute (2003a)
[5] Becker, L.A., Seshadri, M.: GP–evolved technical trading rules can outperform buy and hold, Technical report, Department of Computer Science, Worcester Polytechnic Institute (2003b)
[6] Bhattacharyya, S., Pictet, O.V., Zumbach, G.: Knowledge–intensive genetic discovery in foreign exchange markets. IEEE Transactions on Evolutionary Computation 6(2), 169–181 (2002)
[7] Bondt, W.F.M.D., Thaler, R.: Does the stock market overreact. The Journal of Finance 40(3), 793–805 (1985)
[8] Dempster, M.A.H., Jones, C.M.: A real-time adaptive trading system using genetic programming. Quantitative Finance 1, 397–413 (2001)
[9] Fama, E.F.: Market efficiency, long-term returns, and behavioral finance. Journal of Financial Economics 49, 283–306 (1998)
[10] Hogan, S., Jarrow, R., Teo, M., Warachaka, M.: Testing market efficiency using statistical arbitrage with applications to momentum and value strategies. Journal of Financial Economics 73, 525–565 (2004)

[7] Lehman Brothers Equity Quantitative Analytics, London.

[11] Jain, P.C., Joh, G.-H.: The dependence between hourly prices and trading volume. Journal of Financial Economics 23(3), 269–283 (1988)
[12] Jegadeesh, N., Titman, S.: Returns to buying winners and selling losers: Implications for stock market efficiency. Journal of Finance 48(1), 65–91 (1993)
[13] Jensen, M.C.: Some anomalous evidence regarding market efficiency. Journal of Financial Economics 6, 95–101 (1978)
[14] Jonsson, H., Madjidi, P., Nordahl, M.G.: Evolution of trading rules for the FX market or how to make money out of GP, Technical report, Institute of Theoretical Physics, Chalmers University of Technology (1997)
[15] Kahneman, D., Tversky, A.: Judgment under uncertainty: Heuristics and biases. Science 185(4157), 1124–1131 (1974)
[16] Li, J.: FGP: a genetic programming based tool for financial forecasting, PhD thesis, University of Essex (2001)
[17] Politis, D.N., Romano, J.P.: The stationary bootstrap. Jounal of the American Statistical Association 89(428), 1303–1313 (1994)