Discovery of multi-spread portfolio strategies for weakly-cointegrated instruments using boosting-based optimization

Valeriy V. Gavrishchaka[1]

[1]Head of Quantitative Research, Alexandra Investment Management, New York, 10017

Abstract

Increasing complexity of the modern market dynamics requires new quantitative frameworks for the discovery of stable portfolio strategies. Important requirements include the ability of the coupled and self-consistent optimization of the dynamic strategies and asset allocations as well as robust built-in mechanisms for the strategy complexity control to ensure acceptable out-of-sample performance. Recently introduced boosting-based optimization naturally incorporates all these features. Originally, the framework was described as a generic tool for the discovery of compact portfolio strategies from a given pool of existing financial instruments and base trading strategies. Here I outline the important generalization of this framework that allows simultaneous discovery of new synthetic instruments represented as generalized spreads of existing financial instruments and dynamic trading strategies for each such spread. Detailed arguments and real-market example clarify the essence of this new framework as a powerful generalization of the exiting pairs trading strategies and cointegration-based techniques.

Keywords: Boosting, Ensemble Learning, Cointegration, Pairs/Spread Trading, Market-Neutral Strategies, Portfolio Optimization.

1. Introduction

Increasing complexity of the modern market dynamics and financial instruments requires new quantitative frameworks for the discovery of the robust portfolio strategies. Important requirements for such a system include the ability of the coupled and self-consistent optimization of the dynamic strategies and asset allocations as well as interpretability of the obtained portfolio strategy. In contrast to the classical trading strategy and portfolio optimization frameworks, it should also provide robust built-in mechanisms for the strategy complexity control that ensures acceptable out-of-sample performance.

Recently introduced boosting-based optimization framework naturally incorporates all these features [5-7]. Preliminary results of its application to the real market data confirmed the practical value of the proposed framework. Originally, the framework was described as a generic tool for the automated discovery of the compact portfolio strategy from a given pool of existing financial instruments and base trading strategies.

Increasing number of the intelligent market participants improves market efficiency and makes it more difficult to discover stable portfolio strategies that operate with existing financial instruments. Novel cointegration-based and related approaches for the discovery of the long/short market-neutral portfolios and "pairs" trading strategies could offer more stable solutions [1-3]. These tools help to discover new synthetic instruments (expressed as spreads of existing instruments) that have more predictable time series with a typical mean-reverting feature and not directly observed by the majority of other market participants.

However, as these techniques become more widespread, typical pairs or sets of cointegrated instruments become well-known. This makes it much more difficult to exploit deviations (spreads) from such co-dependencies on a regular basis. Moreover, rigorous cointegration relations can be often violated making strategies relaying on the simple mean-reverting dynamics of the particular spread unstable. Cointegration-based techniques alone do not provide any generic solutions for spread trading in the more common and less ideal situations where only weak/partial or complex time-varying co-dependencies exist.

The pool of such weakly cointegrated instruments could be very large. This could offer a lot of unexplored potential for the discovery of stable market-neutral portfolio strategies. Indeed, any pair/set of instruments that have one or more common economic/market factors in their approximate factor model can be a candidate for such a pool. However, to discover stable portfolio strategies based on spreads of weakly cointegrated instruments one should exploit not only spreads (synthetic instruments) with simple mean-reverting features but also with more general characteristics including multi-scale trends and other complex dynamics. The generic framework should also be able to discover dynamic strategies for each of these multiple spreads and optimally combine them.

In this work, I outline the important generalization of the original boosting framework that allows simultaneous discovery of new synthetic instruments represented as generalized spreads of existing financial instruments and dynamic trading strategies for each such spread. This framework can be considered as a powerful generalization of the exiting techniques for the pairs trading including cointegration-based tools [1-3]. Real market example illustrating operational details of the new framework and its potential practical value is also presented.

2. Market-neutral portfolio selection and pairs trading: Limitations of the existing techniques and multi-spread generalization

Traditional techniques for "pairs trading" and hedging are based on a relative value analysis of asset prices [3]. The two or more assets can be selected on the basis of intuition, fundamentals, long-term correlations, past experience or other empirical knowledge. Obvious limitation of these approaches is the absence of the systematic statistical or fundamental framework that can be generically applied across the universe of financial instruments.

More sophisticated and systematic techniques for hedging and construction of the market-neutral portfolios are based on factor models that can be built using various statistical or fundamental approaches [2,3]. However, in many cases, it is difficult to identify a minimum set of measurable risk factors that are required for a factor model with acceptable accuracy.

Novel cointegration techniques [1-4] provide a powerful generalization of the traditional approaches by introducing a systematic statistical framework for the construction of synthetic pairs/sets in the form of appropriate long/short combinations of two or more assets. Cointegration is essentially an econometric tool to identify possible stable relationships between a set of time series [3,4].

In its original form cointegration can be used to identify potential hedges for a given position or to construct market-neutral buy&hold portfolios [1-3]. Cointegration can also be used to measure the short-term deviations from the equilibrium [3]. This is interesting as a potential source of statistical arbitrage strategies. Deviations from the long-term "fair price" relationship, identified by cointegration analysis, can be considered as statistical "mispricings" that should always revert towards the longer term equilibrium [3].

In the case of statistical arbitrage, cointegration can be considered as an extension of the relative value strategies such as "pairs trading". In the case of hedging and market-neutral portfolio construction cointegration can be considered as extending factor-model hedging to include situations where the underlying risk factors are not measurable directly, but are instead manifested implicitly through their effect on asset prices [3].

The most common method of testing for cointegration [4] is based upon the concept of a "cointegrating regression". In this approach a particular time series ("target series") $S_{0,t}$ is regressed upon the remainder of the set of time series ("cointegrating series") $S_{1,t} \dots S_{n,t}$:

$$S_{0,t} = \alpha + \sum_{j=1}^{n} \beta_j S_{j,t} + s_t \tag{1}$$

If series are cointegrated then statistical tests will indicate that s_t is stationary and the parameter vector $(1,-\alpha, -\beta_1, \dots, -\beta_n)$ determines the cointegration relationship. Two standard statistical tests are recommended by Engle and Granger [4].

The relevance of cointegration to hedging is based upon the recognition that much of the "risk" or stochastic component in asset returns is caused by variations in market and/or economic factors which have a common effect on many assets. This viewpoint forms the basis of traditional asset pricing models such as the CAPM (Capital Asset Pricing Model) of Sharpe and the APT (Arbitrage Pricing Theory) of Ross [2]. Essentially these pricing models can be formulated as

$$\Delta S_{i,t} = \alpha_i + \sum_{j=1}^{n} \beta_{i,j} \Delta f_{j,t} + \varepsilon_{i,t} \tag{2}$$

This general formulation relates changes in asset prices $\Delta S_{i,t}$ to sources of systematic risk expressed as changes in market and/or economic factors $\Delta f_{j,t}$ together with an idiosyncratic asset-specific component $\varepsilon_{i,t}$.

The presence of market-wide risk factors creates the possibility of hedging through the construction of appropriate combination of assets. For example, if asset price dynamics can be described by (2) then the combined return of the portfolio with long position in asset S_1 and short position in asset S_2 is given by

$$\Delta S_{1,t} - \Delta S_{2,t} = (\alpha_1 - \alpha_2) + \sum_{j=1}^{n} (\beta_{1,j} - \beta_{2,j}) \Delta f_{j,t} \\ + (\varepsilon_{1,t} - \varepsilon_{2,t}) \tag{3}$$

From (3) it is clear that the proportion of variance caused by market-wide factors will be significantly reduced if all corresponding factor exposures are similar, i.e.

$$\forall j : \beta_{1,j} \approx \beta_{2,j} \qquad (4)$$

A common approach to hedging is to assume that one can explicitly construct a reasonable factor model (2) using existing fundamental or statistical approaches [2]. Using these factor models one can try to select an optimal long/short portfolio with minimal exposures to these factors (i.e., market-neutral).

However, when it is not possible to identify a reasonable set of explicit factors, cointegration provides an alternative method for implicit hedging the common underlying sources of risk. Given a particular "target asset" S_0, a cointegration regression (1) is used to create a synthetic asset as a linear combination of assets S_j (j=1...n) which exhibit the maximum possible long-term correlation with the target asset [1-4]. The coefficients β_j of this linear combination can be found using standard ordinary least squares (OLS) regression.

The obtained synthetic asset provides an optimal statistical hedge for the target asset. Coefficients β_j specify capital allocations for the hedging assets (in units of the target asset capital), and position types: short for positive β and long otherwise (target asset is long). A linear combination of assets acts as a proxy for the unobserved (implicit) common risk factors.

Although cointegrating regression can be used to estimate the "fair price" relationship between a set of assets, it does not provide any tools to detect deterministic components in the mispricing dynamics for a possible statistical arbitrage strategy. Only for simple mean-reverting dynamics of such mispricings, additional intelligent techniques are not required.

In the real applications to stock/index data, cointegration relationship between logarithms of time series is considered, i.e., all S_j should replaced by ln(S_j) in (1). The essence of the pairs trading can be described in terms of the spread s recovered from (1):

$$s_t = \ln(S_{0,t}) - \sum_{j=1}^{n} \beta_j \ln(S_{j,t}) \qquad (5)$$

Without loss of generality we will consider α=0 from now on. The main goal of the traditional pairs trading is to find a set of β_j that produces spread (5) with stable mean-reverting properties. This search can be based on cointegration or other relevant techniques.

A strict requirement for a stable mean-reverting dynamics of the spread makes pairs trading simple once such pair or group of instruments is found. For example, one can take long or short position (depending on the spread sign) in the synthetic instrument described by (5), i.e. long/short positions in the underlying instruments S_j with capital allocations specified by β_j, when spread $|s|$ becomes larger than some critical value and exit the position when s becomes close to zero.

However, the number of undiscovered instrument pairs/sets with simple mean-reverting spread dynamics decreases as cointegration and related techniques become widespread. This leads to more efficiency with respect to this type of arbitrage and spread dynamics becomes more complex and less stable. Also, even the best pairs/sets found by cointegration on historical data can significantly change their co-dependency in the future. Therefore, strategies, tuned to the single regime of the mean-reverting spread, could abruptly break down.

To resolve these limitations, the more generic spread trading strategies, that can work with different types of complex spread time series, should be considered. Besides spreads with more complex mean-reverting behavior, one can also use a large class of spreads with multi-scale trends that can often be exploited by the portfolio of dynamic strategies [5,6].

According to (3), this means that the more generic goal of synthetic instrument construction is not to remove dependence on all common factors specified by condition (4), but only on some least deterministic factors. In this way, it could be possible to discover spreads with cleaner and more predictable trends and other patterns than in the original time series of the underlying instruments.

The relaxation of requirements for the spread dynamics will significantly expand the universe of existing instruments that can be potential constituents of the long/short synthetic instruments suitable for the stable trading. Indeed, many different weakly-cointegrated instruments, that have only some similar exposures to the common risk factors, could be included. In addition, the more complex spread dynamics will ensure that several different regimes are covered at the training stage. This could significantly improve out-of-sample stability of the strategies compared to the traditional pairs strategies tuned for the strict mean-reverting historical behavior.

To illustrate potential spread-dynamics diversity that can be found even for two underlying time series, we use historical data for S&P mid-cap (MID) and S&P500 (SPX) indexes. Mid-cap index is considered as a target time series S_0. These two indexes demonstrate quite significant cointegration with typical mean-reverting feature for the last few years. However, in the longer run (> 6-7 years), their overall codependence is less stable and more diverse. This suggests that varying only one parameter β_1, many different types of spread dynamics can be observed.

In the following we will use argument of the logarithm in (5) as a spread measure:

$$s_t^{*} = S_{0,t} \times \prod_{j=1}^{n} S_{j,t}^{\;-\beta_j} \tag{7}$$

In the limiting case of β_j=0, the original target time series S_0 is recovered from s^*. This simplifies direct comparisons between spread and underlying time series.

In fig.1, the original mid-cap time series and MID-SPX spread for different values of β_1 are plotted. It is clear that by changing β_1, different types of spread dynamics can be obtained: from complex multi-scale trends to the low-noise long-term trends and mean-reverting behavior. More detailed analysis of the types and predictability of spread dynamics can be performed using variance ratio tests [3].

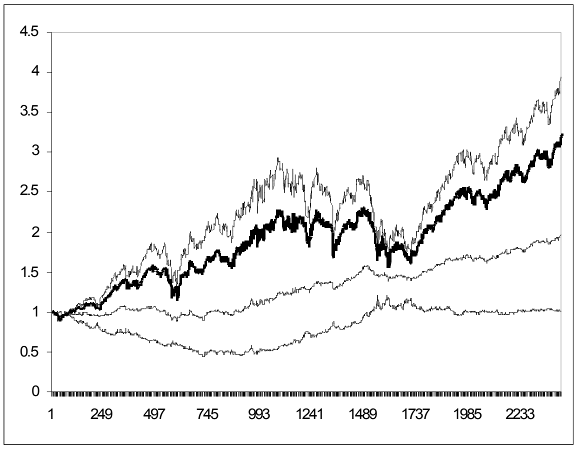

Fig.1. MID time series (solid line) and MID-SPX spread time series for β_1=1.75 (dashed line, first from the bottom), β_1=0.75 (dashed-dotted line), and β_1=-0.3 (dotted line, last from the bottom).

One of the possible approaches to the generic spread trading is to choose a single set of β_j that creates spread with desired dynamics and to search for the technical strategies that are capable to exploit possible deterministic patterns in the spread time series. As shown in previous works [5,6], boosting-based optimization can be effectively used for the discovery of stable portfolio of strategies with adaptive capital for a single instrument. The same framework can be used to exploit multi-scale trends and/or oscillatory patterns in spread time series.

However, diversity of the spread dynamics suggests that more generic and more powerful approach should include combination of dynamic trading strategies applied to multiple spreads specified by different sets of β_j (i.e., multiple synthetic instruments). The final portfolio strategy would still be specified as a long/short variable-capital strategy applied to the initially chosen underlying instruments. As described in the next section, boosting-based optimization could be easily generalized to allow simultaneous self-consistent discovery of trading strategies and new synthetic instruments.

3. Boosting-based framework for the discovery and optimization of the multi-spread portfolio strategy

As described in my previous works [5-7], boosting for optimization could be based on different boosting frameworks. However, the generalized AdaBoost algorithm for classification [8,12,13] could be a preferred choice in many applications due to its simplicity, comprehensive theoretical foundation, and proven robust performance in a large number of realistic classification problems.

For our purposes it is sufficient to describe boosting algorithm only for two-class classification problem, where classifier outputs either +1 or -1. Generalized AdaBoost for two-class classification consists of the following steps [12]:

$$w_n^{(1)} = 1/N \tag{8.1}$$

$$\varepsilon_t = \sum_{n=1}^{N} \left(w_n^{(t)} I(-y_n h_t(x_n)) \right) \tag{8.2}$$

$$\gamma_t = \sum_{n=1}^{N} \left(w_n^{(t)} y_n h_t(x_n) \right) \tag{8.3}$$

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1+\gamma_t}{1-\gamma_t}\right) - \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \tag{8.4}$$

$$w_n^{(t+1)} = w_n^{(t)} \exp\left(-\alpha_t y_n h_t(x_n)\right) / Z_t \tag{8.5}$$

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x) / \sum_{t=1}^{T} \alpha_t \tag{8.6}$$

Here N is a number of training data points, x_n is a model/classifier input set of the n-th data point and y_n is the corresponding class label (i.e., -1 or +1), $I(z) = 0$ for z<0 and $I(z)$=1 otherwise, T is a number of boosting iterations, $w_n^{(t)}$ is a weight of the n-th data point at t-th iteration, Z_t is weight normalization constant at t-th iteration, $h_t(x_n)$->[-1; +1] is the best base hypothesis/model at t-th iteration, ρ is a margin control parameter, and $f(x)$ is a final weighted linear combination of the base hypotheses.

Boosting starts with equal and normalized weights for all training data (step (8.1)). A base classifier $h_t(x)$ is trained using weighted error function ε_t (step (8.2)). If a pool of several types of base classifiers is used, then each of them is trained and the best one (according to error function) is chosen at the current iteration. The training data weights for the next iteration are computed in steps (8.3)-(8.5).

According to (8.5), at each boosting iteration, data points misclassified by the current best model (i.e., $y_n h_t(x_n) < 0$) are penalized by the weight increase for the next iteration. In subsequent iterations, AdaBoost constructs progressively more difficult learning problems that are focused on hard-to-classify patterns. This process is controlled by the weighted error function (8.2).

Steps (8.2)-(8.5) are repeated at each iteration until stop criteria $\gamma_t < \rho$ (i.e., $\varepsilon_t >= 1/2(1-\rho)$) or $\gamma_t = 1$ (i.e., $\varepsilon_t = 0$) occurs. Step (8.6) represents the final combined (boosted) model that is ready to use. The model classifies unknown sample as class +1 when $f(x) > 0$ and as -1 otherwise. Details of more general versions of boosting algorithms are given in [12].

Boosting also offers a flexible framework for the incorporation of other ensemble learning (model combination) techniques. Instead of choosing the single best model at each boosting iteration, one can choose mini-ensemble of models that is optimal according to other ensemble learning techniques. For example, it is often useful to form an equal weight ensemble of several comparable best models at each iteration. In many cases, this generalization improves out-of-sample performance compared to the standard boosted ensemble.

Portfolio strategy discovery is a direct optimization rather than classification problem. However, it was argued [5-7] that for a large class of objective functions, boosting for classification (8.1)-(8.6) can be efficiently used as a basis for the framework that could be labeled as "boosting for optimization" or "boosting-based optimization".

One of the natural and robust objectives for the trading strategy optimization consistent with market neutrality is to require returns (r) generated by the strategy on a chosen time horizon (τ) to be above certain threshold (r_c). By calculating strategy returns on a series of intervals of length τ shifted with a step $\Delta\tau$ and encoding them as +1 (for $r >= r_c$) and -1 (for $r < r_c$), one obtains symbolically encoded time series (distribution) of strategy returns.

Contrary to the classification problems, here the purpose is not to correctly classify (between +1 and -1), but rather to increase the number of +1 samples (i.e., the number of cases with supercritical returns). This can still be considered as classification problem with potentially uneven sample number between two classes. The objective to have maximum number of samples in +1 class (i.e., $r >= r_c$) can be incorporated into the boosting operation (8.1)-(8.6) by considering output -1 as misclassification, i.e., y_n=+1 for all n. In such setting, boosting (8.1)-(8.6) provides a framework for optimization, where maximization objection function is a "hit rate", i.e., number of +1 samples divided by the total number of samples.

Symbolic encoding and corresponding objective function can be based on any complex condition that combines different measures of profit maximization and risk minimization specified by the utility function of interest. This can be easily achieved with combination of several simple conditions.

In the case of trading strategy optimization, the final usage of boosting output is different from the classical case of boosting for classification. Instead of using weighted linear combination (8.6) of the base models as a final model for classification, one uses boosting weights to construct portfolio of strategies. The initial capital is distributed among different base strategies in amounts according to the weights ($\alpha_t/\Sigma\alpha_t$) obtained from boosting which are already normalized.

As discussed in [5-7], boosting can be used to discover the optimal combination of different dynamic trading strategies for a single financial instrument as well as the simultaneous combination of trading strategies and different instruments. In both cases, boosting steps (8.1)-(8.6) are applied to a pool of base strategies $\{BS_i(\boldsymbol{p}_i)\}$, where $\boldsymbol{p}_i$ is a vector of adjustable parameters for strategy BS_i. However, in the first case all base strategies are applied to a time series of a single instrument S_0, while set $\{BS_i\}$x$\{S_j\}$ of all possible pairs of strategies BS_i and instruments S_j should be used in the latter case.

According to error function (8.2), if the objective is to maximize the number of supercritical returns on the shifted intervals, the following optimization problems are solved for all base strategies $BS_i(\boldsymbol{p}_i)$ and instruments S_j at each boosting iteration,:

$$\min_{p_i}\left[\sum_{n=1}^{N} w_n^{(t)} I\left(r_c - r_n^{\tau}\left(BS_i(p_i), S_j\right)\right)\right] \qquad (9)$$

Here, r_n^{τ} is a return produced by the strategy $BS_i(p_i)$ applied to the instrument S_j over n-th shifted interval of length τ and r_c is a chosen threshold value. Based on the results of these minimization procedures for all (i,j) pairs, the best pair "strategy-instrument" of the current iteration is added to the portfolio.

A possible generalization is to consider synthetic instruments (defined by the fixed β sets) instead of original instruments S_j and to apply base strategies to the corresponding spread time series (7). However, the more generic and flexible approach should include simultaneous optimization of the spread time series itself. This can be interpreted as an adaptive discovery of new synthetic instruments. Thus, at each boosting iteration, an optimal vector of β coefficients is found by solving the following coupled minimization problem:

$$\min_{p_i,\beta}\left[\sum_{n=1}^{N} w_n^{(t)}\, I\Big(r_c - r_n^{\tau}\big(BS_i(p_i), s^*(\beta)\big)\Big)\right] \quad (10)$$

In this generalized framework, the final output of the boosting optimization is a portfolio of different synthetic instruments defined by their β vectors (i.e., multiple spreads) with corresponding optimal strategies for each of them.

In general, performance of this framework could be very different from that relying on the same set of base strategies and underlying instruments but treating these instruments independently. Indeed, search for the stable strategies for each individual instrument (specified by (9)) could encounter all the problems associated with constantly improving efficiency of the existing instrument dynamics due to their direct availability to the increasing number of intelligent market participants. In contrast, the generalized framework (10) adaptively creates new synthetic instruments that are not directly exposed to other market players and featuring time series properties exploitable by the technical trading strategies.

4. Application example

The proposed framework for the discovery of multi-spread portfolio strategies can be used with a wide range of underlying financial instruments and base trading strategies. For the illustration of a typical application, two well-known indexes, S&P500 and S&P mid-cap, will be used.

As discussed in section 2, the long-term co-dependency of these two time series is quite complex, so that stable portfolio strategy cannot be easily discovered by co-integration and related tools alone. It is also one of many examples where spreads with significantly different dynamics can be generated by changing just one β coefficient.

In practical settings, different types of base technical strategies (trend-following, oscillator-type, etc.) can be used. However, to stress out flexibility of the boosting framework in comparison to the existing approaches dealing solely with simple mean-reverting spreads, only trend-following (momentum) strategies will be included in the base pool. This will be the most obvious illustration of difference with simple oscillator-type strategies used in the case of stable mean-reverting dynamics.

For clarity, a base strategy pool is restricted to a set of simple trend-following strategies [9]: the exponential moving average of the daily closing prices EMA(n,a) for entry combined with adaptive trailing stops ATS(m,α) based on different volatility measures for exit. Entry into long (short) position on spread occurs when $\mathrm{EMA}_t(s^*)>\mathrm{EMA}_{t-1}(s^*)$ ($\mathrm{EMA}_t(s^*)< \mathrm{EMA}_{t-1}(s^*)$), i.e., the current value of EMA is greater (smaller) than the previous day value. Here s^* is spread given by (7), n is a number of points to be averaged, and a is a smoothing constant. Position entry occurs at the next trading day. More sophisticated low-pass filters [9] could be added in addition to or instead of EMA in real applications.

As mentioned in section 2, entrance into long spread position means taking simultaneous long and short positions in the underlying instruments according to their β values. Position side for a particular instrument (i.e., long or short) is determined by the sign of its β_j. The capital allocation C_j for each instrument is specified by the absolute value of its β_j:

$$C_j = |\beta_j| \frac{C_{\max}}{1+\sum_{j=1}^{N}|\beta_j|} \quad (11)$$

Here N is the total number of underlying instruments and C_{max} is the maximum total capital exposure (both long and short). In the short spread position the side of each underlying instrument is just the opposite.

Long (short) spread position is closed when intraday spread fells below $sc_{max}[1-\alpha\sigma(m)]$ (jumps above $sc_{min}[1+\alpha\sigma(m)]$). Here $sc_{max}(sc_{min})$ is a maximum (minimum) of the spread (computed using closing prices) with respect to present day, and $\sigma(m)$ is a daily return standard deviation of the spread or other volatility measure computed using m last points. Trailing stops based on global and local (i.e., the most recent) maximum (minimum) calculations used for the exit spread level can demonstrate complimentary performance. In this example, both exit types have been used in the base strategy pool.

It is often useful to include multiple trailing stops with different volatility measures into a base strategy pool. For example, volatility measures based on the extreme (bar) values (i.e., open, high, low, and close prices) [10 and references therein] usually provide significantly more accurate estimation of the current volatility, especially when time interval m is small (important in the regimes with fast volatility changes). In the presented example, four different volatility measures have been used: simple standard deviation based on close prices as well as Parkinson, Garman & Klass, and Rogers & Satchell estimators [20].

For each type of volatility measure (estimator) and exit level (global/local), two base strategies [EMA(n,a), ATS(m,α)] are included in the pool. One is long-only (ignores short entry signals) and the other is short-only (ignores long entry signals). Thus, the pool of total 16 base strategies has been used in the presented example.

Daily strategies considered here are very tolerant to a wide range of transaction costs that depend on the type of trading vehicles used (e.g., ETFs, futures, or mixture). For simplicity of presentation, in the reported results such costs are ignored. The min/max values of intraday spread used for testing of stop-loss execution are approximated based on daily bar data. Unlike single instrument trading, a rigorous modeling of "pairs" trading requires intraday data.

Adaptive boosting (8.1)-(8.6), (10) with ρ=0.1 and T=7 has been applied to the described base strategy pool and two underlying instruments: MID as target and SPX with variable β. Binary hit rate objective function with r_c=5% (annualized) for the horizon of τ=63 days and $\Delta\tau$=20 days is used. Training data period is 1997/05/15-2005/05/15. At every boosting iteration, each strategy [EMA(n,a), ATS(m,α)] from the base strategy pool is optimized with respect to (n,a,m,α,β) parameter set using global optimization algorithm based on simulated annealing combined with downhill simplex method [11]. At each boosting iteration, in addition to the best model, other models that are inferior to the best model by no more than 2% are also retained with equal weights.

The result of boosting-based optimization is portfolio of 17 complimentary multi-scale strategies [EMA(n,a), ATS(m,α)] operating on 17 different spread time series calculated from the underlying MID and SPX data with 17 different β values. The ranges of parameters are the following: n=11..130, a=0.96..0.99, m=4..87, α=0.20..0.55, and β=-0.34..0.43.

Distribution of annualized returns of the boosted portfolio of MID-SPX multi-spread trading strategies for the horizon of τ=63 business days is plotted in fig.2 (solid line). Here, a period from 1997/01/01 to 2006/01/10 is covered. Fig.2 presents both in-sample (training) data and up to 1.5 years of out-of-sample data. Stability of the obtained portfolio strategy even for such short-term horizon is obvious. Return distribution improves for larger time horizons.

Complexity of the underlying indexes is evident from the distribution of the returns produced by the buy&hold strategies: MID-long (dashed line) and ½ MID-long, ½ SPX-short (dotted line). In contrast to the stable distribution of only positive returns generated by the boosted portfolio strategy, buy&hold strategies have comparable number of negative and positive returns.

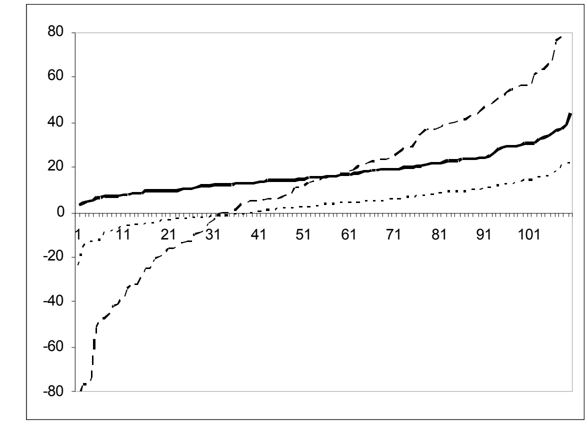

Fig.2. Distribution of the annualized returns (%) for 63 days horizon: MID-SPX multi-spread portfolio strategy (solid line), MID long buy&hold (dashed line), ½ MID long and ½ SPX short buy&hold (dotted line).

Finally, we illustrate the mechanism responsible for stability and robustness of the boosted portfolio strategy. As obvious from the very boosting operation, obtained single strategies are locally uncorrelated or negatively correlated and capable to support each other through many different market regimes. Often the global performance (i.e., an average over all market regimes) of single members of the boosted portfolio is not impressive. However, since they are locally complementary to each other, the global performance of the final portfolio could be very stable and significantly superior to the best single strategy.

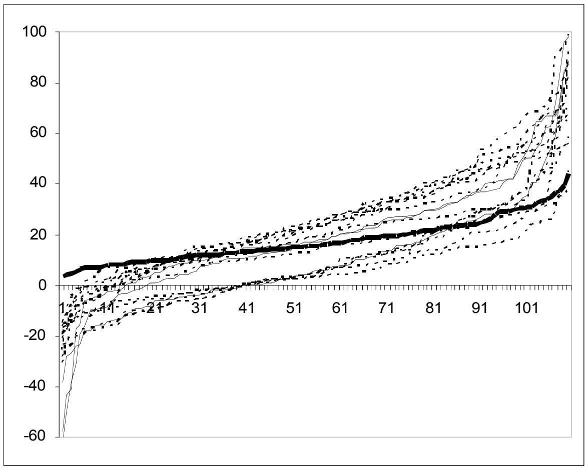

Fig.3. Distribution of the annualized returns (%) for 63 days horizon: MID-SPX multi-spread portfolio strategy (solid line) and each individual strategy from the portfolio (dotted lines).

This is illustrated in fig.3, where distributions of the annualized returns on the 63 day period for all 17 strategies from the boosted portfolio are shown by dotted lines. The corresponding distribution of returns produced by the boosted portfolio is shown by solid line. The qualitative jump from the sub-optimal performance of single strategies to the impressive stability of their combination is very clear.

Returns of three strategies with comparable weights from the current portfolio are plotted in fig.4 chronologically. These strategies exploit spread trends on different time scales (*n*=70, 11, 50) and operate on different spread time series (β=0.18, -0.11, -0.34) (solid, dashed, and dotted lines). It is clear from fig.4 that these strategies are complimentary to each other. This is the underlying source of the boosted portfolio robustness and stability.

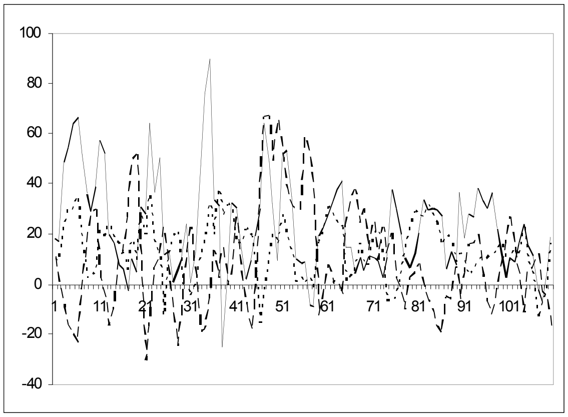

Fig.4. Shifted annualized returns (%) of 3 individual strategies from the MID-SPX multi-spread boosted portfolio (in chronological order).

5. Conclusions

Limitations of the modern techniques for the identification of the market-neutral portfolios and dynamic spread ("pairs") trading strategies have been discussed. A generic boosting-based framework for the discovery of the stable multi-spread portfolio strategy, that can address many unresolved issues, has been proposed. The framework can be efficiently used for the simultaneous discovery of new synthetic instruments specified as spreads of existing instruments and optimal trading strategies for each such spread.

Presented arguments and real-market example clarify the essence of the new framework as a powerful generalization of the exiting pairs trading techniques and cointegration tools. Existing approaches are applicable only to a small and constantly decreasing subset of existing instruments with a simple mean-reverting spread dynamics. In contrast, the proposed framework can be used with much larger set of weakly-cointegrated instruments featuring complex spread dynamics.

6. References

[1] C. Alexander, and A. Dimitriu, "The Cointegration Alpha: Enhanced Index Tracking and Long-Short Equity Market Neutral Strategies", *ISMA Finance Discussion Paper* No. 2002-08, 2002

[2] C. Alexander, "Market Models: A Guide to Financial Data Analysis", *Wiley*, 2001

[3] A.N. Burgess, "Using cointegration to hedge and trade international equities", in "*Applied quantitative methods for trading and investment*", edited by C. Dunis, J. Laws, and P. Naim, *Wiley*, 2003

[4] R.F. Engle, and C.W.J. Granger, "Cointegration and error-correction: Representation, Estimation and Testing", *Econometrica*, 55, 251, 1987

[5] V.V. Gavrishchaka, "Boosting frameworks in financial applications: From volatility forecasting to portfolio strategy optimization", *CIEF*, 2005

[6] V.V. Gavrishchaka, "Boosting-based framework for portfolio strategy discovery and optimization", *New Mathematics and Natural Computation*, vol. 2, 2006 (to appear)

[7] V.V. Gavrishchaka, "Boosting-Based Frameworks in Financial Modeling: Application to Symbolic Volatility Forecasting", *Advances in Econometrics*, 20B, 123, 2006

[8] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", *Springer*, 2001

[9] J.O. Katz, and D.L. McCormick, "The encyclopedia of trading strategies", *McGraw-Hill*, 2000

[10] K. Li and D. Weinbaum, "The empirical performance of the alternative extreme value volatility estimators", *Working Paper, Stern School of Business, New York,* 2000

[11] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, "Numerical Recipes in C: The Art of Scientific Computing", *Cambridge University Press*, 1992

[12] G. Ratsch, "Robust boosting via convex optimization: Theory and applications", *PhD thesis*, *Potsdam University*, 2001

[13] R.E. Schapire, "The design and analysis of efficient learning algorithms", *PhD thesis*, *MIT Press*, 1992