5 F performance indicator: A robust metric for trading systems evaluation?

JIŘÍ SVOBODA
Finance department
Tomas Bata University in Zlín
Náměstí T.G. Masaryka 5555, 760 01 Zlín
CZECH REPUBLIC
j2svoboda@fame.utb.cz   http://www.fame.utb.cz

Abstract: The article is focused on capital market trading systems evaluation. An aggregate performance index is put forward to ensure a complex measurement of different dimensions of trading system performance. The model is developed in several modifications. Dynamic version of the index is put under particular scrutiny and through logistic regression is compared with three existing performance indicators. Binomial type logistic regression shows the aggregate index has higher accuracy in filtering profitable and unprofitable out-of-sample returns than the three compared indicators.

Key-Words: currency market, performance indicator, yield, logistic regression, drawdown, out-of-sample, Sortino ratio.

1 Introduction

Technical analysis, as a tool for providing recurrent trading signals, has been under academic scrutiny from several perspectives. Most part of academic work in past decade has been focused on providing empirical proof value adding capability of technical analysis. Great deal of academic work is focused on investment funds and trading systems performance evaluation. The following content is aiming to develop a metric, which encompasses different perspectives of trading system performance (quality).

2 Literature research

Allen and Karjalainen (1999) applied through genetic algorithms trading rules which accounted for gross profits. However, after inclusion of transaction costs, the utilized trading system proved to be unprofitable.

To measure performance of technical analysis rules, numerous metrics are utilized. Academic focus is put on evaluating hedge and investment funds performance. Several works provide evidence of hedge funds assymetrical returns distribution (Brooks and Kat 2002, Eling and Schumacher 2007, Eling, 2008).

Traditional performance metrics measure excess returns adjusted by their volatility (standard deviation). Sharpe ratio (1966) is considered as the pivotal performance indicator.

$$E = \frac{R - R_f}{\sigma} \qquad (1)$$

However, several authors put forward the inconsistency of this performance measure when evaluating assymetrical return distributions (Eling and Schumacher 2007, Mamoghli and Daboussi, 2010). Farinelli et al. (2008) prove Sharpe ratio never outperformed assymetrical performance metrics. They state Sharpe ratio doesn’t comply with fat-tail distributed returns.

Several other risk-adjusted or probability-based metrics are suggested to be more feasible. Agarwal and Naik (2004) and Capocci (2004) show that traditional linear based performance models cannot capture dynamics of different markets.

Risk-adjusted performance metrics can be divided into several categories. Beta-based performance models capture riskless return aggregated with beta-adjusted risk premiums. Among these performance metrics stands Jensens Alpha (1968) and Treynor ratio. Jensens Alpha is compared with modified beta-based factors by Mamoghli and Daboussi (2010). Wide group of performance metrics is built on Sharpe ratio rationale. For example Treynor ratio replaces standard deviation denominator with systematic risk denoted by beta.

$$T = \frac{R - R_f}{\beta} \qquad (2)$$

Sortino ratio, developed on the basis of Sharpe, takes into account only negative deviations of returns, therefore not penalizing positive capital fluctuations. Positive fluctuations are considered desirable for investors. Ellen and Schumacher (2007) define negative deviations as “lower partial moments” or LPM. Lower partial moments are integrated in Omega and Stutzer indexes as well. For capital asset *i* of order *n*, LPM formula is written as follows:

$$LPM_{ni}(\tau) = \frac{1}{T}\sum_{t=1}^{T} \max\left[t - r_{it}, 0\right]^{n} \qquad (3)$$

Average return of capital asset *i* is denoted as r_{it} and minimal acceptable return as *T*. Bacmann and Scholz (2003) state Omega and Stutzer indexes are most feasible performance metrics for investment funds evaluation, while Chaudhry and Johnson (2008) present results of benchmarking performance analysis with conclusion stating Sortino ratio are the most viable metric for evaluating assymetrical distributions.

Other metrics used for evaluation of testing and trade results involve drawdown calculation (Calmar and Sterling ratio), value-at-risk based metrics, sensitivity and cluster analysis.

Trading systems and funds’ performance metrics focus mainly on evaluating the relationship between excess returns and risk. However, to ensure complex assessment of testing and trading outputs, there have to be several performance metrics involved. Stability of trade results and trading system liquidity are often omitted. Performance analysis requires great deal of scrutiny and robust trade results measurement in order to archive long term profitability of trading systems portfolios.

3 Aims and methodology

The aim of the article is to develop a model to assess different aspects of trading system performance. This model is called 5 F performance index and should be multifunctional in sense that the user would be able to utilize it to filter profitable out-of-sample trading results and to assess cohesion of trading system results under different market conditions.

3.1 Data sample

Created aggregate index was tested on 20 different algorithmic trading systems based on technical analysis. These systems were created by author on training set. Aggregate index with all subindexes was calculated on the basis of historical trading results of in-sample and one out-of-sample data set. Further illustration is given by Fig. 1.

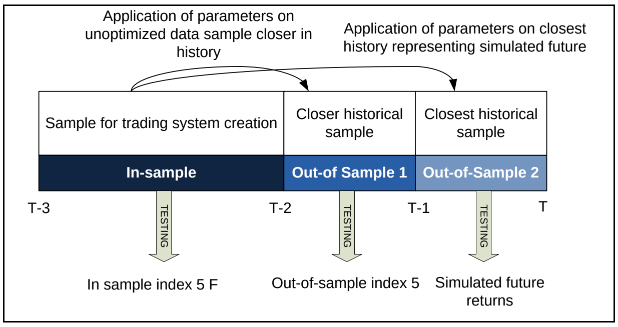
Application of parameters on unoptimized data sample closer in history
Application of parameters on closest history representing simulated future
Sample for trading system creation
Closer historical sample
Closest historical sample
In-sample
Out-of Sample 1
Out-of-Sample 2
T-3
T-2
T-1
T
TESTING
TESTING
TESTING
In sample index 5 F
Out-of-sample index 5
Simulated future returns

Fig. 1: Division of tested time interval

Trading systems were created, modeled and tested on in-sample time interval <T-3; T-2>. Then the result parameters of trading systems were also tested on out-of-sample time interval <T-2; T-1>. On this time interval, the parameters weren’t modeled or optimized. A second out-of-sample time interval <T-1; T> was utilized in order to achieve simulation effect via closest historical data sample.

First combination was in-sample time interval from 09/2001 to 09/2007, first out-of-sample from 09/2007 to 09/2009 and second out-of-sample from 09/2009 to 09/2011. Second combination was in-sample time interval from 09/2005 to 09/2009, first out-of-sample from 09/2009 to 09/2010 and second out-of-sample from 09/2010 to 09/2011. Third combination was in-sample time interval from 09/2009 to 09/2010, first out-of-sample from 09/2010 to 03/2011 and second out-of-sample from 03/2011 to 09/2011.

Currency markets were chosen for testing mainly due to data accessibility, reliability and consistency. Major currency instruments were analyzed, among the most tested EUR/USD, EUR/JPY, AUD/JPY, GBP/USD etc. Timeframes utilized to generate trade results on chosen samples were in range from 5 minute to 1 day. The combinations of market, timeframe and testing interval distribution were randomly generated.

Starting equity for testing was 10 000 USD, with fixed traded amount of 0.25 lots, which stands for 1 to 3 % or risked capital per trade depending on stop-loss implementation technique. Due to immense computational power requirements, opening prices only were used for testing. Opening prices provide lesser accuracy of testing, particularly on lower timeframes. Standard transaction costs were integrated into backtesting process. These transaction costs are presented in brokerage company regulations and refer to spread during standard liquidity and volatility market conditions.

3.2 Testing vehicle

Logistic regression estimates probability of a certain event on the basis of known variables. This regression type is used to predict the outcome of categorical criterion variable. General logistic regression model is as follows.

$$y = \frac{1}{1 + e^{-(a_0 + a_1 x)}} \qquad (4)$$

The result of logistic regression will be a percentage of successful predictions in tested data sample. The 5 F index prediction success will be compared with three existing performance metrics – Sortino ratio, Calmar ratio and Ulcer index.

4 Results

The model is represented by 5F aggregate index based on the notion of investment triangle. The fundament of the model provides aggregation of 5 subindexes with equivalent weight factors. These subindexes consist of security, liquidity, yield, probability of success and stability of returns. Each of the five attributes is quantified with accent on scale cohesion of each subindex.

$$I_{5F} = \left(I_{y,IS} + I_{l,IS} + I_{sf,IS} + I_{s,IS} + I_{fd,IS}\right) \qquad (5)$$

The above formula represents 5 F index. Each subindex is denoted abbreviation of its name (yield, liquidity, success factor, stability, financial default). Also an in-sample notation is given. It is very important to separate in-sample and out-of-sample sets, to recognize on which data sample the trading system was created (modeled or optimized).

4.1 Filtering capability evaluation

There will be subindexes defined in following subsections. These subindexes will be chosen according to their scale similarity and cohesion with different evaluated attributes of trading system performance.

4.1.1 Fertility (yield)

Several metrics have been assessed in order to pick a subindex to represent yield calculation. Sharpe ratio, as probably the most used and popular performance metric, is not suitable for fat tail asymmetric distributions. Sharpe also penalizes positive returns volatility. Omega and Stutzer indexes are more suitable for hedge and investment funds, as they are developed for benchmarking and ranking purposes. Treynor ratio and Jensens alpha involve beta calculation, which is less applicable and precise in currency market terms. Sortino ratio is suggested as suitable yield index for fat tailed distributions (Ellen and Schumacher, 2007). Therefore Sortino ratio has been chosen to represent the first subindex of 5 F model.

4.1.2 Frequency (liquidity)

For frequency of trading signals a logarithmic proxy function was used. Author of this article developed a formula based on several assumptions.

The longer the evaluated time interval is the more statistically significant are results of performance measurements. A short historical data sample causes the modeled and optimized trading system parameters to overly adjust to a certain time interval (curve fitting). Out-of-sample trading results may deteriorate significantly compared to in sample performance results. A logarithmic proxy function that follows the relevance of increasing data sample stands as follows:

$$0{,}1.\left(log\frac{1}{N}\right)^2, \qquad (6)$$

where N is the length of tested time interval (in days). This formula uses a large portion of approximation. However the aim is to define a proxy to appreciate longer data interval.

Decimal logarithm function is less prone to overestimate liquidity subindex when incrementing lesser numbers. Function solves the main problem with evaluated time interval, which lacks the maximum value (time is an ongoing value).

$$\frac{\sum_{i=1}^{n} D_u - \sum_{j=1}^{m} D_d}{N}, \qquad (7)$$

Another important factor, which influences trading system performance is its' resistance versus changing market dynamics. Therefore overlay of total time of trades in tested interval should be measured. This "coverage" represents exposure of historic trades to different market conditions. D_u represents total time coverage of all contracts in history. D_d represents the total time interval where two or more contracts are covering each other.

$$Liquidity\ subindex = I_l = \frac{0{,}1.\left(log\frac{1}{N}\right)^2 + \frac{\sum_{i=1}^{n} D_u - \sum_{j=1}^{m} D_d}{N}}{2} \qquad (8)$$

Liquidity subindex is calculated as an average value of logarithmic function for tested time interval length and a rate of tested time interval coverage with trade positions. The subindex is respectable to data.

4.1.3 Favourability (success factor)

Unlike tangible and intangible assets, financial investments provide standardization (investment horizon, number of trades, transaction costs, asset type etc.). A significant historical data sample enables implementation of trading system success factor into performance measurement. Success factor can be measured on probabilistic or historic basis. In order to be compliant with other subindexes, the success factor will be based on historical probability of success. Therefore the subindex will be measured by a simple ratio.

$$Success\ factor\ subindex = I_{sf} = \frac{n_p}{n}, \qquad (9)$$

Number of profitable trades is denoted as n_p. The number involves also breakeven trades because transaction costs are covered. Number of total trades is denoted as *n*.

4.1.4 Fixity (stability)

Trading system stability is a desirable attribute to great deal of trading subjects. Companies strive to reduce their cash flow volatility and individual traders desire to eliminate drawdown of their equity curves. Ideal equity curve is represented by straight rising line, without any drawbacks. A performance indicator which can be used for measuring stability is coefficient of determination. Coefficient of determination reflects the quality of regression model. The value of coefficient denotes the percentage of dispersion of dependent variable (equity curve), which is explained by the model (straight regression line).

$$\text{Stability subindex} = \mathrm{I_S} = R^2 = 1 - \frac{SS_{err}}{SS_{tot}}, \qquad (10)$$

The closer the subindex is to zero, the lesser stability for evaluated trading system. Coefficient of determination is often prone to multicolinearity. However there is only one dependent and one independent variable in this regression model, which simplifies the analysis.

4.1.5 Financial default (riskiness)

Certain performance measures focus on quantification of threshold which represents worst case scenario. The probabilistic version of financial default quantification uses Risk of ruin measurement. Historical version of worst case scenario is based on drawdown metric, which has been chosen as a component for financial default subindex calculation. Unlike risk of ruin, drawdown is an interval-closed metric, which is coherent with other subindexes in 5 F aggregate index.

$$\text{Financial default subindex} = \mathrm{I_{fd}} = 100\% - DD(\%) \qquad (11)$$

Drawdown represents a negative measure with maximum value of 100 %. An inverse calculation ensures ascendancy of performance metric measurement.

Capital markets reflect non-linear progression with different amount of white noise. Dynamic market nature causes the progression to deviate from various function approximations. Performance measurement of trading results should take capital market dynamics into account. One of tools used to test trading system robustness is out-of-sample analysis. Out-of-sample analysis is implemented into index 5 F as follows:

$$I_{5F} = \left(I_{y,IS} + I_{l,IS} + I_{sf,IS} + I_{s,IS} + I_{fd,IS}\right) + \sum_{i=1}^{n}\left(I_{y,OOS} + I_{l,OOS} + I_{sf,OOS} + I_{s,OOS} + I_{fd,OOS}\right)_i \qquad (12)$$

Aggregate performance index value is calculated as a sum of in-sample subindexes and *n* out-of-sample 5 F indexes. Performance measurement can be done through absolute value of the overall index. Therefore the absolute value can serve as a benchmark in trading system development process. Moreover the cohesion among subindex values in in-sample and out-of-samples can be measured, thus evaluating robustness of the trading system.

4.2 Filtering capability evaluation

Binomial type of logistic regression was used to separate profitable and unprofitable trading systems in out-of-sample. Positive returns were denoted by 1 and negative returns by 0. Value *y* belongs to interval (0, 1). As a threshold for assigning value 0 a y ≤ 0,5 condition was used. Otherwise value 0 was assigned.

Firstly a filtering capability was examined by maximum likelihood estimation function:

$$y = \frac{1}{1 + e^{-(-4{,}0150 + 0{,}8543x)}} \qquad (13)$$

The model correctly assigned 129 out of 200 observations (64.5 %).

Unlike the model 5 F, the compared metrics (Calmar ratio, Sortino ratio and Ulcer index) were less filtering-effective. Sortino ratio successfully predicted 61 %, Calmar ratio 60 % and Ulcer index 58 % of profitable trading systems.

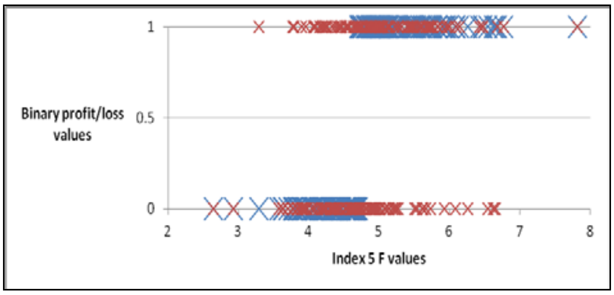

Fig. 2: Dispersion of binomic prediction accuracy of index 5 F through logistic regression

Fig. 2 represents a chart with values of 5 F index on axis x and binary values of profit and loss on axis y. Bigger crosses denote profit/loss estimation by aggregate index 5 F. Smaller crosses represent real distribution of profits and losses. The less vertical intersected profit/loss crosses are, the more effective filtering is. It is obvious that bigger crosses (model estimations) on upper and lower border are on the same vertical lines (on left of value 5).

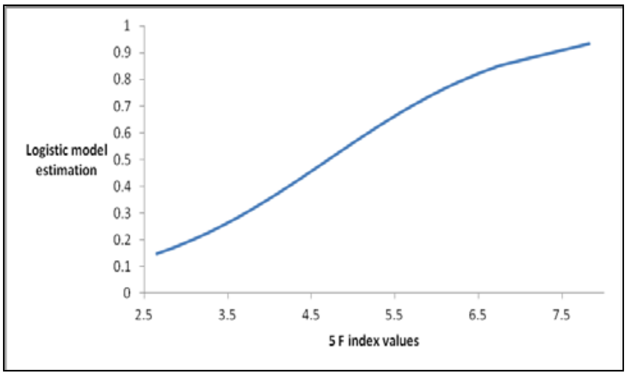

Fig. 3: Regression function estimation through index 5 F values

Fig. 3 illustrates the 5 F model quality from other perspective. Axis x encompasses values of aggregate 5 F index, while axis y represents value y (maximum likelihood estimation). The more acute is the function development near median values, the more effective the model is. Although the created model doesn't have extremely high filtering capability it beats compared metrics on the tested data sample.

4 Discussion

The created model in this work should not be perceived as a fixed version. Aggregate performance model 5 F can be augmented through discrimination analysis or by extending the measurement on several markets and timeframes. As such 5 F could serve as benchmark of robustness of evaluated trading systems. Moreover, liquidity subindex could be modified in order to capture the relevance of evaluated time horizon in relationship with out of sample sets. The length of in sample and out of sample could prove vital for predicting future returns. Moreover residuum normality test could be utilized in order to evaluate multicollinearity and heteroscedasticity of the model.

5 Conclusion

Please, follow our instructions faithfully, otherwise Aggregate index 5 F was composed with the aim to provide a complex performance vehicle, which could be further adjusted and extended. Five subindexes represent different attributes of trading system quality. Liquidity subindex was newly developed in order to evaluate length of tested time horizon and rate of open orders duration in history of trades. Aggregate index provides several possibilities in trading systems evaluation. Is serves as a benchmark for trading systems development and helps to unwind the relationship of several system quality tradeoffs. Although the model is undoubtedly rough around the edges, its goal is to provide a performance metric for further examination.

References:

[1] A. Chaudhry, and H. Johnson, The efficacy of the Sortino ratio and other benchmarked performance measures under skewed return distributions, Australian, *Journal of Management*, Vol. 32, No. 3, 2008, pp. 485-502.

[2] C. Brooks, and H. M. Kat, The Statistical Properties of Hedge Fund Index Returns and Their Implications for Investors, *Journal of Alternative Investments*, Vol. 5, No. 2, 2002, pp. 26–44.

[3] D. Capocci, and G. Hubner, An Analysis of Hedge Fund Performance, *Journal of Empirical Finance*, Vol. 11, No. 1, 2004, pp. 55-89.

[4] F. Allen, R. Karjalainen, Using genetic algorithms to find technical trading rules, *Journal of Financial Economics*, Elsevier, Vol. 51, No. 2, 1999, pp. 245-271.

[5] M. Eling, and F. Schumacher, Does the choice of performance measure influence the evaluation of hedge funds? *Journal of Banking and Finance*, Vol. 31, No. 9, 2007, pp. 2632-2647.

[6] M. Jensen, The Performance of Mutual Fund in the period 1945-1964, *Journal of Finance*, Vol. 23, No. 2, 1968, pp. 389-416.

[7] S. Farinelli, M. Ferreira, D. Rossello, M. Thoeny and L. Tibiletti, Beyond Sharpe Ratio: Optimal Asset Allocation using Different Performance Ratios, *Journal of Banking and Finance*, Vol. 32, No. 11, 2008, pp. 2057-2063.

[8] V. Agarwal, and N. Naik, Risks and portfolio decisions involving hedge funds, *Review of Financial Studies,* Vol. 17, No. 1, 2004, pp. 63-98.

[9] W. F. Sharpe, Mutual Fund Performance, *The Journal of Business*, University of Chicago Press, Vol. 39, 1965, pp. 119.

[10] M. Eling, Does the Measure matter in Mutual Fund Industry, *Financial Analysts Journal*, 2008, Vol. 64, No. 3, pp. 54-67.