2013 IEEE 25th International Conference on Tools with Artificial Intelligence

Improving the Statistical Arbitrage Strategy in Intraday Trading by Combining Extreme Learning Machine and Support Vector Regression with Linear Regression Models

Jarley P. Nóbrega
Centro de Informática
Universidade Federal de Pernambuco
Recife (PE), Brazil
jpn@cin.ufpe.br

Adriano L. I. Oliveira
Centro de Informática
Universidade Federal de Pernambuco
Recife (PE), Brazil
alio@cin.ufpe.br

***Abstract*— In this paper we investigate the statistical and economic performance for statistical arbitrage strategy using Extreme Learning Machine (ELM) and Support Vector Regression (SVR) models, and their forecast combination through four linear combination models. The application of the traditional Kalman Filter for the statistical arbitrage strategy improves the statistical performance of ELM and SVR individual forecasts. It is presented evidence that the financial performance for most of cointegrated pairs can be improved by at least one linear combination technique.**

Keywords- Statistical Arbitrage, Pair Trading, Extreme Learning Machine, Support Vector Regression, Kalman Filter, Forecast Combinations.

I. INTRODUCTION

There has been growing interest in financial time series forecasting in recent years and accurate predictive models are taking a central place for decision making. In this context, neural networks have become one of the most useful tools for applications in financial time series analysis and forecasting [1].

Neural networks can approximate complex nonlinear functions from the input samples without knowing the internal structure of the time series. However, it is known that neural networks have the issue of heavy computation, since all parameters need to be tuned during the training phase. Such iterative approach will take a long time for tuning and the learning process can be inefficient [2].

To overcome the weakness of traditional neural networks, new classes of learning methods were introduced. For instance, Extreme Learning Machine (ELM) was introduced by Huang et al. in [3]. One advantage of ELM over traditional neural networks is that it is not necessary to adjust parameters iteratively [4].

Support Vector Regression (SVR) is another class of learning algorithms that has also been receiving increasing attention to solve nonlinear estimation problems, including financial time series forecasting [5]. According to the model described by Vapnik et al. [6], it presents a global minimum solution for regression problems and faster training speed when compared with traditional neural networks architectures [7].

In financial time series forecasting there has been an increasing interest in the application of state space modeling, such as Kalman Filter [9]-[11]. In particular, Kalman Filter is largely applied to build forecasting models as a noisy observation of some mean-reverting state process [10].

In addition to the forecasting methods mentioned before, some new approaches based on the combination of them have been successfully applied in order to outperform the accuracy of individual forecasts. Experiments involving the combination of time-varying financial forecasts with neural networks and a Kalman Filter regression model were conducted with encouraging results [11]-[13].

In this paper, we propose to apply a similar method presented in [12] in order to investigate the statistical and trading performance of the combination of ELM and SVR with four linear combination methods: the Bayesian Average, Granger and Ramanathan Regression (GRR), Least Absolute Shrinkage and Selector Operator (LASSO) and Kalman Filter. The statistical performance of the models described in this paper was estimated using the statistical arbitrage trading strategy [17]. We also evaluated the trading risk for the series described in this paper, using the Sharpe ratio [18] as a risk measure.

The novelty of this research is the proposed combination framework: we put together two different classes of learning methods, ELM and SVR, and combine them using a state space model. We also compare the forecast performance with a set of linear regression combination methods.

The remainder of this paper is organized as follows. Section II discusses the related work. In Section III, we briefly present the statistical arbitrage concepts. Section IV describes the forecasting models. Section V addresses the detailed description of the sample data. Sections VI and VII present the statistical and trading performance, respectively. Section VIII concludes this work and discusses its results.

II. RELATED WORK

Despite the promising use of ELM to forecast financial instruments, its application within statistical arbitrage strategies was not reported in the literature.

The related literature discusses several works about the application of nonlinear techniques for statistical arbitrage models. Some relevant proposals include constructing combinations of financial time series using the concept of cointegration and a posterior building of predictive models that capture the price anomalies [20]. Another relevant

1082-3409/13 $31.00 © 2013 IEEE
DOI 10.1109/ICTAI.2013.36

approach is to model the dynamics of the time series as a mean-reverting Gaussian Markov chain model [10]. In [19], the risk-factors are extracted using Principal Component Analysis (PCA) and the trading signals are generated by the significant deviations of the residuals from the estimated mean.

Some recent works address the use of hybrid models for describing the dynamics of the correction for relative prices [20]. This approach attempts to detect nonlinearities both in the mean and the volatility using a combined neural network-GARCH volatility model. Another category of model considers genetic programming for statistical arbitrage, since an evolutionary computation technique can be used to evolve strategies that generate positive out-of-sample returns [21].

Financial data can also be described in terms of a time-varying structure. In this context, the use of state space modeling for representing dynamic systems with unobserved variables can be integrated within an observable model [8], [11], [22]. This approach is suitable for processing complex nonlinear and nonstationary signals with strong component correlations.

III. Statistical Arbitrage

In the world of finance, the term statistical arbitrage encompasses a variety of strategies that attempt to profit from pricing discrepancies that appear in a group of assets. One of the techniques used for statistical arbitrage involves trading securities in pairs [17]. This process identifies pairs of securities whose prices tended to move together. In Fig. 1 is presented an example of pair with correlated price movement. In this example, the time series for AMBV3 and AMBV4 presents similar movement for the period of January, 2013 to March, 2013. This strategy was designed to exploit short-term deviations from a long-run equilibrium pricing relationship between two assets, based on cointegration, correlation and other nonparametric decision rules [17].

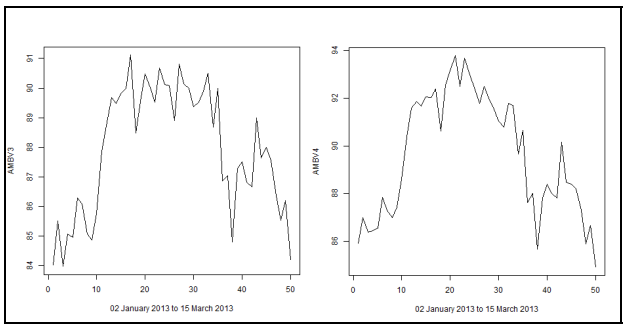

Figure 1. AMBV3 and AMBV4 daily price series.

The detection of price anomalies is based upon the identification of a linear combination of assets, whose time series is mean-reverting and has finite variability. This model will allow us to make predictions for the relative difference of a pair of financial assets, named *pair spread*. If observations are larger or smaller than the predicted value (by some threshold value) we take a long or short position in the portfolio and we unwind the position and make a profit when the spread reverts. A history and discussion about statistical arbitrage and pairs trading can be found in [28].

IV. Forecasting Models

A. Extreme Learning Machine

Unlike the traditional learning algorithms for neural networks, the main characteristic of ELM is learning without iterative training, as proposed by Huang et al [3]. Let the training set be $\{(x_i, t_i) \mid x_i \in R^n, t_i \in R^m, i = 1,..., N\}$, where x_i is an $n \times 1$ input vector and t_i is a $m \times 1$ target vector. The training process is briefly described as follows.

Step 1: Randomly assign values to the inputs weights and the hidden neuron biases.

Step 2: The output weights are analytically determined through the generalized inverse operation of the hidden layer matrices, according to the following equation:

$$\sum_{i=1}^{L} \beta_i G(a_i, b_i, x_j) = t_j,\ j=1,...,N \tag{1}$$

where a_i is the input weights, b_i is the hidden layer biases, β_i is the output weight that connects the ith hidden node and output node, and G is the activation function. L is the number of hidden neurons. N is the number of distinct input or output data. This is equivalent to $H\beta = T$, where

$$H = \begin{bmatrix} G(a_1,b_1,x_1) & \cdots & G(a_L,b_L,x_L) \\ \vdots & \cdots & \vdots \\ G(a_1,b_1,x_N) & \cdots & G(a_L,b_L,x_N) \end{bmatrix}_{N \times L} \tag{2}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_L^T \end{bmatrix}_{L \times m} \text{ and } T = \begin{bmatrix} t_1^T \\ \vdots \\ t_L^T \end{bmatrix}_{N \times m}$$

H is the hidden layer output matrix. The activation function *G(x)* should be assigned before training is carried out. For this paper, we used the sigmoid function. It is given by

$$G(a_i, b_i, x_i) = \left(1 + e^{-(a_i . x_i + b_i)}\right)^{-1},\quad i = 1, ..., N \tag{3}$$

Step 3: Calculate the output weight by $\beta = H^+ T$, where H^+ is the Moore-Penrose generalized inverse of *H*.

B. Support Vector Regression

Support Vector Machines (SVM) use an implicit mapping Φ of the input data into a high-dimensional feature space defined by a kernel function, i.e., a function returning the inner product $\langle \Phi(x_i) \Phi(x) \rangle$ between the images of two data points x_i, x in the feature space [6]. If a projection $\Phi : X \to H$ is adopted, the dot product

$\langle \Phi(x_i), \Phi(x) \rangle$ can be represented by the following kernel function k

$$k(x_i, x) = \langle \Phi(x_i) \Phi(x) \rangle \quad (4)$$

If applied for classification problems, SVM separate the different classes of data by a hyper-plane $\langle w, \Phi(x) \rangle + b = 0$, corresponding to the decision function [6]

$$f(x) = sign(\langle w, \Phi(x) \rangle + b)$$

where $w = \sum_i \alpha_i \Phi(x_i)$, α_i are the coefficients and b is a constant. The optimal hyper-plane is the one with the maximal margin of separation between the two classes, with the optimization problem solved by a constrained quadratic method for a subset of training patterns. These patterns are called support vectors.

For the regression task, let the training data $\{(x_1, y_1), ..., (x_n, y_n)\} \subset \chi \times \Re$, where χ denotes the space of input patterns. We have to find a function *f(x)* that has at most ε deviation from targets y_i for all training data and is as flat as possible [30]. The formulation of support vector regression stated in [6] implies in the minimization of the following expression:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \quad (5)$$

subject to

$$y_i - \langle w, x_i \rangle - b \le \varepsilon + \xi_i$$

$$\langle w, x_i \rangle + b - y_i \le \varepsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \ge 0$$

where the constant $C > 0$ defines the trade-off between the flatness of f and the level of tolerance for ε. ξ_i, ξ_i^* are two positive slack variables which can be used to measure the deviation from the boundaries of the ε-sensitive zone. The construction of a Lagrange function for Eq. (5) generates the following regression:

$$f(x) = (\alpha_i - \alpha_i^*)\langle x_i, x \rangle + b \quad (6)$$

where α_i, α_i^* are Lagrange multipliers. The expansion of Eq. (6) for the nonlinear case can be written as

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_i, x) + b \quad (7)$$

For this paper, we are adopting the RBF kernel

$$K(x_i, x) = \exp(-\gamma\|x_i - x\|^2), \ \gamma > 0 \quad (8)$$

where γ is the parameter of the kernel.

C. Bayesian Model Averaging

Let f_{ELM_t} and f_{SVR_t} be the forecasts at time t for the ELM and SVR models, respectively. This strategy performs a weighting over the forecasts in order to achieve the optimal weights for the combination based on the Akaike Information Criterion (AIC) [14]. AIC measures the relative goodness of fit of a statistical model, as introduced by Akaike [23] and can be computed as

$$AIC = N\log(s^2) + 2k \quad (9)$$

where N is the sample size, s^2 the maximum likelihood estimate of the error variance and k is the total number of parameters of the model. The Bayesian weights can be calculated as

$$W_{AIC_{ELM}} = \frac{\exp(-0.5\Delta AIC_{ELM})}{\exp(-0.5\Delta AIC_{ELM}) + \exp(-0.5\Delta AIC_{SVR})} \quad (10)$$

$$W_{AIC_{SVR}} = \frac{\exp(-0.5\Delta AIC_{SVR})}{\exp(-0.5\Delta AIC_{ELM}) + \exp(-0.5\Delta AIC_{SVR})} \quad (11)$$

where $\Delta AIC_i = AIC_i - AIC_{i,\min}$ and $i=1,2$ for f_{ELM_t} and f_{SVR_t}, respectively. The combination forecast at time t is given by the following expression:

$$f_t = \frac{W_{AIC_{ELM}} f_{ELM_t} + W_{AIC_{SVR}} f_{SVR_t}}{2} \quad (12)$$

D. Least Absolute Shrinkage and Selector Operator (LASSO)

A LASSO regression is a shrinkage and selection method for linear regression which minimizes the sum of squared errors, with a bound on the sum of the absolute values of the coefficients [16]. The constrained form of LASSO solves this problem with L1 regularization. Given the vectors of independent and dependent variables

$$\begin{bmatrix} X_1^T \\ \vdots \\ X_N^T \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1N} \\ \vdots & \cdots & \vdots \\ x_{N1} & \cdots & x_{NN} \end{bmatrix}, \ Y = (y_1, ..., y_N)^T \quad (13)$$

and the training data $\{(x_1, y_1), ..., (x_N, y_N)\}$, the regression coefficients are estimated as:

$$\hat{\beta}_i = \min_{\beta}(\|x\beta - y\|_2^2) \quad (14)$$

subject to

$$\|\beta\|_1 \le s, \ s>0$$

where s is the penalization parameter for the constrained form of LASSO and $i=1,2$ for ELM and SVR forecasts. This constraint makes the model adaptive, since there is a penalization balance on each estimate. For this paper, we optimized the value of s for the in-sample dataset for each time series. According to Eq. (14), the combination forecast can be determined by the following:

$$f_t = \beta_{ELM} f_{ELM_t} + \beta_{SVR} f_{SVR_t} + \varepsilon_t \quad (15)$$

subject to $|\beta_{ELM}| + |\beta_{SVR}| \le s$.

E. Granger and Ramanathan Regression (GRR)

The GRR approach combines a set of forecasts inputs using a single linear regression model in order to outperform the individual forecasts benchmarks, as defined by Bates and Granger [15]. The combination of ELM and SVR forecasts can be described by the following:

$$f_t = \alpha_{ELM_t} f_{ELM_t} + \alpha_{SVR_t} f_{SVR_t} + \varepsilon_t \quad (16)$$

where α_{ELM_t} and α_{SVR_t} are the regression coefficients of f_{ELM_t} and f_{SVR_t}, and ε_t is the error term of the regression model.

F. Kalman Filter Regression Model

The Kalman Filter can be described as a recursive method to estimate the state of a dynamic system from a series of incomplete and noisy measurements [24]. The state representation of the dynamics of the time-varying regression coefficients is given by the following system of equations:

$$P_{Y_t} = P_{X_t}\beta_t + \varepsilon_t, \quad \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right) \quad (17)$$

$$\beta_t = \beta_{t-1} + \eta_t, \quad \eta_t \sim N\left(0, \sigma_\eta^2\right) \quad (18)$$

where P_{Y_t} is the dependent variable, β_t is a time-varying regression coefficient, and P_{X_t} is the independent variable at time *t*, respectively. ε_t and η_t are independent uncorrelated error terms with standard variances σ_ε^2 and σ_η^2, respectively. Eq. (17) is also named the measurement equation and Eq. (18) is named the state equation, which defines the regression coefficient as a simple random walk. Here, the variances of the noise process and other unknown parameters have to be estimated. This is accomplished by maximizing the following likelihood function:

$$\log L = -\frac{NT}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log F_t - \frac{1}{2}\frac{\nu_t' \nu_t}{F_t} \quad (19)$$

where ν_t is the one-step ahead residual and F_t its variance. *N* and *T* are the number of columns of P_{X_t} and the number of elements of the time series P_{Y_t}, respectively. When we apply the Kalman Filter, the β_t coefficients are estimated by maximizing Eq. (19) with a numerical algorithm based on σ_η^2. The coefficients are estimated at time *t* based on the new observations and the states estimates are propagated in time *t+1*. The regression model is defined by the following equations:

$$f_t = \alpha_{ELM_t} f_{ELM_t} + \alpha_{SVR_t} f_{SVR_t} + \varepsilon_t \ , \ \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right) \quad (20)$$

$$\alpha_{ELM_t} = \alpha_{ELM_{t-1}} + \eta_{ELM_t} \ , \ \eta_{ELM_t} \sim N\left(0, \sigma_{\eta_{ELM}}^2\right) \quad (21)$$

$$\alpha_{SVR_t} = \alpha_{SVR_{t-1}} + \eta_{SVR_t} \ , \ \eta_{SVR_t} \sim N\left(0, \sigma_{\eta_{SVR}}^2\right) \quad (22)$$

where, α_{ELM_t} and α_{SVR_t} are time-varying regression coefficients at time *t* for ELM and SVR. Eq. (20) is the measurement equation and Eq. (21) and Eq. (22) describe the states equations. Here, η_{ELM_t} and η_{SVR_t} are the error terms with variances $\sigma_{\eta_{ELM}}^2$ and $\sigma_{\eta_{SVR}}^2$, respectively.

V. Sample Data

We selected a set of tradable financial assets that are listed on BM&FBovespa Exchange[1]. For this task, we captured all asset quotes between the period of January, 2013 and March, 2013. The combination of assets results in five pair spreads generated from the cointegration approach. The pairs and their underlying assets are described in Table I.

The cointegration approach was defined by Engle and Granger [25] as a statistical feature whereby two time series that are integrated of order 1, *I(1)*, can be linearly combined to produce one stationary time series, denoted as *I(0)*. Let $X_{1,t}, X_{2,t}, ..., X_{k,t}$ a sequence of time series *I(1)*. If exist non-zero real numbers $\beta_1, \beta_2, ..., \beta_k$, and $\beta_1 X_{1,t}, \beta_2 X_{2,t}, ..., \beta_k X_{k,t}$ is a time series *I(0)*, then $X_{1,t}, X_{2,t}, ..., X_{k,t}$ is a sequence of cointegrated time series. $\langle \beta_1, \beta_2, ..., \beta_k \rangle$ is called the cointegrated vector. In order to train the ELM and SVR forecasting models we further divided our in-sample dataset into three sub-periods, as presented in Table II. Since all pairs in this study were generated by cointegration, their stationary property is confirmed at the 1% confidence level and its summary statistics are shown in Table III. The Jarque-Bera statistics confirms that the spread series are non-normal at the 99% confidence interval. By applying the Phillips-Ouliaris test for cointegration [26], we confirm at the 1% confidence level that all pairs have at least one cointegration vector.

In our application, we conduct some ELM and SVR experiments in a set of potential inputs in the training dataset in order to achieve the best trading performance. These set of inputs are presented in Table IV below.

TABLE I. Selected Cointegrated Pairs

Pair	Description
AMBV3-AMBV4	Companhia de Bebidas Das Americas ON/PN
GOAU4-GGBR4	Gerdau Metalurgica ON/PN
ITUB4-ITSA4	ItauUnibanco / Itau SA
USIM3-USIM5	Usiminas ON/PN
VALE3-VALE5	Vale ON/PN

TABLE II. Training, Validation and Testing Datasets

	Trading Days	Start Date	End Date
Total dataset	50	02/01/2013	15/03/2013
Training dataset	25	02/01/2013	26/02/2013
Validation dataset	12	07/02/2013	26/02/2013
Testing dataset	13	27/02/2013	15/03/2013

[1] http://www.bmfbovespa.com.br

TABLE III. STATISTICS SUMMARY

	AMBV3-AMBV4	GOAU4-GGBR4	ITUB4-ITSA4	USIM3-USIM5	VALE3-VALE5
Sample Size	14830	20550	22063	15030	21809
Pearson Correlation	0.958	0.996	0.993	0.990	0.998
Mean	2.15E-5	3.65E-5	7.19E-5	-6.34E-5	1.04E-5
Std. Dev.	0.006	0.005	0.009	0.013	0.003
Median	0.001	5.97E-5	4.12E-4	-0.001	-4.60E-5
Maximum	0.016	0.018	0.037	0.056	0.011
Minimum	-0.021	-0.017	-0.036	-0.044	-0.011
Skewness	-0.671	-0.280	0.194	0.529	0.207
Kurtosis	0.283	-0.235	-0.401	0.867	0.153
ADF Test	0.01 (-10.47)	0.01 (-15.46)	0.01 (-10.20)	0.01 (-10.76)	0.01 (-15.96)
Phillips–Ouliaris	0.01 (-141.07)	0.01 (-1233.46)	0.01 (-7243.57)	0.01 (-169.38)	0.01 (-601.86)
Jarque-Bera	0.0 (1164.21)	0.0 (316.50)	0.0 (288.11)	0.0 (1171.41)	0.0 (178.58)

(*) Values in brackets represent the test value.

TABLE IV. DATASET ATTRIBUTES

Number	Atribute	Lag(*)
1	$Pair_1$ spread	1
2	$Pair_2$ spread	2
3	$Pair_3$ spread	3
4	$Pair_4$ spread	4
5	$Pair_5$ spread	5
6	$Pair_6$ spread	6
7	$Pair_7$ spread	7
8	$Pair_8$ spread	8
9	$Pair_9$ spread	9
10	$Pair_{10}$ spread	10
11	Historical volatility	1
12	Half-Life	1
13	Alpha (fromVECM)	1
14	Mean	1

(*) Lag 1 implies in forecasting tomorrow close price using today's close price.

VI. STATISTICAL PERFORMANCE

In order to evaluate the statistical performance of the ELM and SVR forecasts, as well the performance of the combination among them, we have executed 35 experiments with 200 repetitions for each individual experiment.

The values used for the size of ELM hidden layer nodes were defined empirically considering the range [50,100], adjusted after each repetition to obtain the best network. To optimize the values of the SVR parameters, γ and C, we used the Caret R package [27] in the in-sample dataset, according to Eq. (5) and Eq. (8).

We compute the RMSE and Theil-U statistics to support our analysis in terms of the accuracy of the models. The RMSE is the difference between values predicted by a model or an estimator and the values actually observed. The Theil-U measure lies between zero and 1, with zero indicating a perfect fit. In Tables V and VI we present a summary of the statistical performance for the in-sample and out-of-sample datasets respectively. When comparing the individual results, the ELM forecast outperforms the SVR accuracy in both in-sample and out-of-sample datasets. Since RMSE is a scale dependent measure, we can confirm the ELM superior performance with Theil-U statistics which are independent of the scales. When we apply the combinations techniques, the overall performance is quite similar to all of them, except for the Kalman Filter which outperforms all models studied here.

The iterative approach of the Kalman Filter can explain these findings, since it can be considered an optimal forecast estimator for the time-varying series presented in this study.

TABLE V. IN-SAMPLE STATISTICAL PERFORMANCE

		ELM	SVR	Bayesian Average	GRR	LASSO	KF
AMBV3-AMBV4	RMSE	0.0009	0.0010	0.0009	0.0009	0.0009	0.0002
	Theil-U	0.15	0.17	0.15	0.15	0.16	0.03
GOAU4-GGBR4	RMSE	0.0010	0.0010	0.0010	0.0009	0.0010	0.0002
	Theil-U	0.17	0.19	0.18	0.17	0.19	0.03
ITUB4-ITSA4	RMSE	0.0013	0.0017	0.0013	0.0013	0.0013	0.0007
	Theil-U	0.14	0.17	0.14	0.14	0.14	0.07
USIM3-USIM5	RMSE	0.0024	0.0028	0.0023	0.0023	0.0023	0.0017
	Theil-U	0.18	0.22	0.18	0.18	0.17	0.13
VALE3-VALE5	RMSE	0.0006	0.0008	0.0006	0.0006	0.0006	0.0001
	Theil-U	0.19	0.23	0.19	0.19	0.19	0.03

TABLE VI. OUT-OF -SAMPLE STATISTICAL PERFORMANCE

		ELM	SVR	Bayesian Average	GRR	LASSO	KF
AMBV3-AMBV4	RMSE	0.0012	0.0013	0.0012	0.0012	0.0012	0.0007
	Theil-U	0.20	0.20	0.20	0.20	0.20	0.12
GOAU4-GGBR4	RMSE	0.0013	0.0015	0.0013	0.0013	0.0013	0.0007
	Theil-U	0.24	0.27	0.24	0.24	0.23	0.14
ITUB4-ITSA4	RMSE	0.0015	0.0016	0.0015	0.0015	0.0015	0.0009
	Theil-U	0.16	0.17	0.16	0.16	0.16	0.10
USIM3-USIM5	RMSE	0.0029	0.0039	0.0030	0.0029	0.0030	0.0025
	Theil-U	0.22	0.29	0.23	0.22	0.23	0.19
VALE3-VALE5	RMSE	0.0009	0.0011	0.0009	0.0009	0.0009	0.0001
	Theil-U	0.26	0.32	0.26	0.26	0.26	0.04

VII. ECONOMETRIC PERFORMANCE

In the second part of our study, we evaluated the econometric performance of all models described here.

In general, pairs trading are a bet on the mean reversion property of the linear combination of cointegrated time series. As presented in [25], this property can be defined by two quantities: the cointegration coefficient and the equilibrium value. Moreover, the equilibrium adjustment can be estimated by modeling the mean-reverting process as an Ornstein-Uhlenbeck (O-U) model [29].

Based on the model above, a forecast model is built for the horizon determined by the O-U model. The trading rule implies in opening a long or short position for the pair depending the following conditions:

- Enter a long position, buying the first part of the pair and selling the second one when $Z_t < \hat{Z}_{t+h}^{L,\alpha}$ and unwind the operation when Z_t converges to its mean.
- Enter a short position, selling the first part of the pair and buying the second one when $Z_t > \hat{Z}_{t+h}^{H,\alpha}$ and unwind the operation when Z_t converges to its mean.

where Z_t is the value of the pair spread at time *t*, $\hat{Z}_{t+h}^{L,\alpha}$ and $\hat{Z}_{t+h}^{H,\alpha}$ denote the *(1-α)* low and high confidence bound on the forecasted value of the spread *h* minutes ahead, and *h* is the estimate of mean reversion speed.

In this paper, the cost for opening and closing positions for a selected pair was estimated by market benchmarks, including the brokerage and fee values. Another source of cost for trading transaction is the *slippage*, the difference between ask and bid values, which leads to a small loss when we buy or sell an asset. The slippage values for each pair are calculated at the first level of the book orders. As a single risk management policy, we established the maximum percentage level of loss for an opened trade position. If the current net operation exceeds 1% of loss, the position is immediately closed.

The evaluation of trading performance was made in terms of the annualized return, the annualized volatility and the Sharpe ratio. Annualized return describes the average amount of money earned by an investment each year over a given time period and it can be calculated as follows:

$$AR = \frac{252}{n}\left(\sum_{t=1}^{n} R_t\right) \tag{23}$$

where R_t is the daily return of the strategy. The annualized volatility is determined by the standard deviation of the return, according to the following equation:

$$AV = \sqrt{252}\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(R_t - \overline{R}\right)^2} \tag{24}$$

where $\overline{R}$ is the return average. Finally, the Sharpe ratio measures the excess return or risk premium per unit deviation in a trading strategy [18]. It can be calculated according the following:

$$SR = \frac{\left(AR - R_f\right)}{AV} \tag{25}$$

where R_f is the rate of return of a free risk asset (e.g. government bonds). For this paper, the Sharpe ratio characterizes how well the return of the statistical arbitrage strategy compensates the investor for the risk taken.

In Tables VII and VII we present a summary of the trading performance for all models described in this study. The criteria for performance here is the highest annualized return with low volatility and high Sharpe ratio. We note that despite the poor performance of ITUB4-ITSA4 for both in-sample and out-of-sample datasets, the remaining pairs presented positives annualized returns and corresponding positive annualized Sharpe ratios. The best pairs in terms of trading performance were GOAU4-GGBR4 and USIM3-USIM5. For the in-sample dataset GOAU4-GGBR4 pair achieved 89.27% of annualized return when combining the ELM and SVR forecasts with the Kalman Filter model. The trading performance for USIM3-USIM5 in the out-of-sample dataset achieved 112.81% of annualized return. For this specific pair, we conclude that the iterative nature of Kalman Filter model gives it some advantage when compared to the remaining models, since it computes forecasts propagating the states estimates when new observations are feeding the model.

For all pairs, at least one combination technique outperforms the individual ELM and SVR trading returns for both in-sample and out-of-sample, which confirms our research hypothesis. In opposition of [12], we have not found an evidence of the superior performance of Kalman Filter method for all pairs.

TABLE VII. IN-SAMPLE TRADING PERFORMANCE

		ELM	SVR	Bayesian Average	GRR	LASSO	KF
AMBV3-AMBV4	Return	25.10%	18.67%	16.78%	26.51%	42.37%	23.97%
	Volatility	5.58%	4.97%	5.24%	5.65%	3.99%	3.14%
	Sharpe	4.48	3.75	3.19	4.68	10.59	7.61
GOAU4-GGBR4	Return	83.74%	82.25%	85.66%	83.18%	82.11%	89.27%
	Volatility	6.88%	7.03%	6.82%	6.81%	6.76%	8.33%
	Sharpe	12.16	11.69	12.54	12.21	12.12	10.70
ITUB4-ITSA4	Return	5.44%	5.93%	-4.19%	5.44%	-1.58%	1.23%
	Volatility	2.01%	2.08%	1.71%	2.01%	1.71%	1.08%
	Sharpe	2.69	2.83	-2.46	2.69	-0.93	1.08
USIM3-USIM5	Return	48.33%	50.12%	78.20%	48.41%	39.77%	71.90%
	Volatility	12.87%	13.39%	14.64%	12.87%	11.08%	14.38%
	Sharpe	3.75	3.74	5.33	3.76	3.58	4.99
VALE3-VALE5	Return	0.55%	5.77%	-0.33%	3.29%	3.13%	-4.91%
	Volatility	1.93%	2.17%	2.98%	1.87%	1.79%	2.42%
	Sharpe	0.27	2.64	-0.11	1.75	1.74	-2.03

TABLE VIII. OUT-OF-SAMPLE TRADING PERFORMANCE

		ELM	SVR	Bayesian Average	GRR	LASSO	KF
AMBV3-AMBV4	Return	14.97%	10.87%	18.80%	16.32%	6.08%	11.88%
	Volatility	3.70%	3.59%	3.82%	3.79%	3.79%	4.22%
	Sharpe	4.03	3.02	4.91	4.30	1.59	2.80
GOAU4-GGBR4	Return	57.56%	48.36%	22.63%	49.83%	32.73%	25.95%
	Volatility	5.54%	5.09%	4.82%	6.68%	5.54%	4.33%
	Sharpe	10.38	9.48	4.69	7.45	5.90	5.98
ITUB4-ITSA4	Return	-15.63%	-16.49%	11.93%	-15.63%	-12.39%	-25.99%
	Volatility	1.85%	2.13%	6.36%	1.85%	1.89%	2.63%
	Sharpe	-8.44	-7.71	1.87	-8.44	-6.56	-9.87
USIM3-USIM5	Return	62.76%	102.00%	150.79%	56.75%	184.81%	112.81%
	Volatility	13.67%	15.64%	11.73%	13.92%	10.70%	14.53%
	Sharpe	4.58	6.52	12.84	4.07	17.26	7.76
VALE3-VALE5	Return	-0.29%	6.10%	0.11%	1.28%	-0.16%	19.20%
	Volatility	3.59%	3.22%	3.24%	3.42%	2.80%	3.76%
	Sharpe	-0.08	1.89	0.02	0.37	-0.06	5.09

VIII. CONCLUSIONS

In this paper we have investigated the statistical and economic performance for pairs trading strategy using ELM and SVR models, and their forecast combination through four linear regression models. We first selected a set of pairs, which had their cointegration and stationary properties confirmed by ADF and Phillips-Ouliaris tests. Next, we modeled the pair spreads dynamics for each pair using the Ornstein-Uhlenbeck process in order to estimate the speed of mean reversion. The statistical arbitrage model feeds both ELM and SVR models that aimed to produce forecasts for the price spread direction. We also combined the ELM and SVR forecasts with four linear regression models in order to compare the overall performance.

The results have shown that Kalman Filter outperforms all models studied in terms of statistical performance. On the other hand, the financial evaluation of the models indicates that there is no single combination method that

outperforms the remaining. Although the annualized returns and Sharpe ratio for most of pairs were positive, there is no evidence that the statistical superiority of one combination method can improve the profitability level for pairs trading. However, we found evidences that combining forecast entries through linear regression models can outperform the individual forecast accuracy.

We have the intention to test other forecast approaches for financial time series, such as GARCH modeling, for evaluating the impact of conditional heteroskedasticity on the spread dynamics. Finally, we plan to incorporate into the trading design phase a portfolio management model in order to maximize the expected return for the set of pairs.

ACKNOWLEDGMENT

This work was supported by the National Institute of Science and Technology for Software Engineering (INES) funded by CNPq and FACEPE, grants 573964/2008-4 and APQ-1037-1.03/08.

REFERENCES

[1] M. Deutsch, C.W.J. Granger, T. Teräsvirta, "The combination of forecasts using changing weights," International Journal of Forecasting 10 (1) (1994) 47–57.

[2] G.-B. Huang, "Learning capability and storage capacity of two hiddenlayer feedforward networks," IEEE Trans. Neural Networks, Vol. 14, no. 2, pp. 274-281, 2003.

[3] G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, "Extreme learning machine: Theory and applications", Neurocomputing, Vol.70, no.1-3, pp. 489–501, Dec. 2006.

[4] N.-Y. Liang, G.-B. Huang, P. Saratchandran, and N. Sundararajan, "A fast and accurate online sequential learning algorithm for feedforward networks," IEEE TRANS. on Neural Networks, Vol. 17, no. 6, pp. 1411-1423, NOV. 2006.

[5] C.-J. Lu, T.-S Lee, C.-C. Chiu, "Financial time series forecasting using independent component analysis and support vector regression," Decision Support Systems, 47(2), 2009,115–125. doi:10.1016/j.dss.2009.02.001

[6] V. Vapnik, S. Golowich, A. Smola, "Support vector machine for function approximation, regression estimation, and signal processing," Advances in Neural Information Processing Systems, vol. 9, p. 281–287, 1996.

[7] F.E.H. Tay, L.J. Cao, "Application of support vector machines in financial time series forecasting," Omega 29 (2001), 309–317.

[8] C.L. Dunis, G. Shannon, "Emerging markets of South-East and Central Asia: do they still offer a diversification benefit?,"Journal of Asset Management 6 (3) (2005) 168–190.

[9] R. Theoret, F. Racicot, "Forecasting stochastic Volatitlity using the Kalman filter: an application to Canadian Interest Rates and Price-Earnings Ratio", Published in Aestimatio. The IEB International Journal of Finance No. 1, 2010,pp. 1-20

[10] R. J. Elliot, J. V. D. Hoek, W. P. Malcolm, "Pairs trading," Quantitative Finance, 5(3), 271–276, 2005.

[11] S.L. Goh, D.P. Mandic, "An augmented extended Kalman Filter algorithm for complex-valued recurrent neural networks," Neural Computation 19 (4) (2007) 1039–1055.

[12] G. Sermpinis, C. Dunis, J. Laws, C. Stasinakis, "Forecasting and Trading the EUR/USD Exchange Rate with Stochastic Neural Network Combination and Time-varying Leverage," Decision Support Systems, ISSN 0167-9236 (2013)

[13] N. Terui, H.K. Van Dijk, "Combined forecasts from linear and nonlinear time series models," International Journal of Forecasting 18 (3) (2002) 421–438.

[14] S.T. Buckland, K.P. Burnham, N.H. Augustin, "Model selection: an integral part of inference", Biometrics 53 (2) (1997) 603–618.

[15] C.W.J. Granger, R. Ramanathan, "Improved Methods of Combining Forecasts," Journal of Forecasting 3 (2), 197–-204 (1984)

[16] H. Wang, G. Li, G. Jiang, "Robust Regression Shrinkage and Consistent Variable Selection through the LAD–Lasso," Journal of Business and Economic Statistics 25 (3), 347–-355 (2007)

[17] G. Vidyamurthy, "Pairs Trading, Quantitative Methods and Analysis," John Wiley & Sons, (2004)

[18] W.F. Sharpe, "The Sharpe Ratio," The Journal of Portfolio Management 21 (1), 2010, 49–58, doi:10.3905/jpm.1994.409501

[19] M. Avellaneda, J. Lee, "Statistical arbitrage in the US equities market," Quantitative Finance, Taylor and Francis Journals, vol. 10(7), 761-782, 2010.

[20] N.S. Thomaidis, N. Kondakis, "Detecting statistical arbitrage opportunities using a combined neural network - GARCH model," Advances in Artificial Intelligence, Lecture Notes in Computer Science, Volume 3955/2006, pp. 596-599, 2012.

[21] P. Saks, D. Maringer, "Genetic Programming in Statistical Arbitrage," EvoWorkshops, 2008, LNCS 4974, pp. 73–82.

[22] D. Jiani, L. Zhitao, C. Can, W. Youyi, "Li-ion Battery SOC Estimation Using EKF Based on a Model Proposed by Extreme Learning Machine," IEEE Conference on Industrial Electronics and Applications, 1648--1653 (2012)

[23] H. Akaike, "A new look at the statistical model identification," IEEE Transactions on Automatic Control, 19 (6) (1974) 716–723.

[24] A.C. Harvey, "Forecasting, Structural Time Series Models and the Kalman Filter," Cambridge University Press, Cambridge, U.K., 1990

[25] R.F. Engle, C.W.J. Granger, "Cointegration and error-correction: representation, estimation and testing," Econometrica, 55, 1987, 251-276.

[26] P. C. B. Phillips, S. Ouliaris, "Asymptotic Properties of Residual Based Tests for Cointegration," Econometrica 58, 1990, 165–193.

[27] "Caret R Package", June, 2013 [online], Available: http://cran.r-project.org/web/packages/caret/caret.pdf

[28] A. Pole, "Statistical Arbitrage: Algorithmic Trading Insights and Techniques," Wiley Finance, 2007

[29] E. Bibbona, G. Panfilo and P. Tavella: "The Ornstein-Uhlenbeck process as a model of a low pass filtered whitenoise," *Metrologia*, 45, 2008, S117-S126.

[30] A. L. I. Oliveira, "Estimation of software project effort with support vector regression," Neurocomputing 69(13-15): 1749-1753 (2006).