Modeling the EUR/USD Index Using LS-SVM and Performing Variable Selection

Luis-Javier Herrera[1(✉)], Alberto Guillén[1], Rubén Martínez[2], Carlos García[2], Hector Pomares[1], Oresti Baños[1], and Ignacio Rojas[1]

[1] Department of Computer Architecture and Technology, Universidad de Granada, Granada, Spain
jherrera@ugr.es
[2] CoTrading S. L., Barcelona, Spain

Abstract. As machine learning becomes more popular in all fields, its use is well known in finance and economics. The growing number of people using models to predict the market's behaviour can modify the market itself so it is more predictable. In this context, the key element is to find out which variables are used to build the model in a macroeconomic environment. This paper presents an application of kernel methods to predict the EUR/USD relationship performing variable selection. The results show how after applying a proper variable selection, very accurate predictions can be achieved and smaller historical data is needed to train the model.

1 Introduction

The foreign exchange market is one of the most liquid markets together with the derivate financial market. The EUR/USD relationship is one of the most negotiated as they represent the two most world powerful economies, Europe and the United States of America. The price of the foreign exchange is the main point of interest of various Economists and experts that are wondering about those prices and whether they can be a picture of their economies or not [2]. It is important to highlight that this asset is very complex and it can be affected by several circumstances, as there are several agents on both sides, offering and demanding.

The role of the central banks, the Federal Reserve as well as from the European Central Bank, have been meaningful in the last few years for the fluctuation of foreign exchanges [7]. The importance of the changes related to interests has been decisive for their economies and for the appreciation and depreciation of the relationship EUR/USD along their history. The improvement of the currency mass has been reflected in the stock markets like in the relationship analysed in this paper. These and other aspects can be detected studying how the macroeconomic variables affect the final price EUR/USD.

This study will focus on the creation of a model that is able to check the impact that variables have on the quotation. In order to detect those facts in the

I. Rojas et al. (Eds.): IWANN 2015, Part II, LNCS 9095, pp. 165–172, 2015.
DOI: 10.1007/978-3-319-19222-2_14

fluctuation of foreign exchanges, we will choose as a reference the macroeconomic data from Europe and the USA, stock markets and raw materials like gold and oil.

2 Data Set Definition

There are plenty of reference data that could be chosen, however, this paper presents a selection based on the experience of a trading company. The variables have been shown trough the time that are meaningful and important to each economy, making posible to create a solid model. As the paper is focused in the EUR/USD relationship, the macroeconomic variables are taken from both, European and Northamerican sources.

The EUR/USD Foreign Exchange Rate is the output variable to be predicted. It is a type of negotiated change in financial markets, and it is the most liquid asset.The data used in the following section can be downloaded from http://research.stlouisfed.org/fred2/series/EXUSEU.

The subset of macroeconomic variables and their sources are listed below:

1. **Consumer Price Index for All Urban Consumers (CPIAUCNS)** The Consumer Price Index is a statistical measurement of the development, of all prices of goods and services consumed. A high Price Index Consumer means a significant loss of purchasing power. Source: http://research.stlouisfed.org/fred2/series/CPIAUCNS/downloaddata?cid=9
2. **Civilian Unemployment Rate (UNRATE)**. Unemployment rate is the percentage of the total work force that is unemployed actively. The lower the better for the currency value. Source: http://research.stlouisfed.org/fred2/series/UNRATE/downloaddata.
3. **10-Year Treasury Constant Maturity Rate (GS10)**. It represents the interest rate the U.S government would pay on top of principal to the bond holder once ten years have passed. The market value of an existing bond will move in the opposite direction of the change in market interests. The bond yields actually serve as an excellent indicator of the strength of a nations stock market, which increases the value of currencies. Source: http://research.stlouisfed.org/fred2/series/GS10.
4. **Effective Federal Funds Rate (FEDFUNDS)**. It is the interest ratio at which depositary institutions exchange money. The Reserve Federal regulating banks and other important financial institutions to ensure the safety and soundness between them. Monetary policy decisions involve setting the interest rate. Source: http://research.stlouisfed.org/fred2/series/FEDFUNDS/.
5. **Real Effective Exchange Rates Based on Manufacturing Consumer Price Index for the United States (CCRETT01USM661N)**. Wellbeing or utility depends on consumption. This indicator measures the health of the economy over consumption. Source: http://research.stlouisfed.org/fred2/series/CCRETT01USM661N/downloaddata.
6. **All Employees: Total nonfarm (PAYEMS)**. This measure provides useful insights into the current economic situation. It measures the number of U.S

workers in the economy that excludes proprietors, private household employees and unpaid volunteers. The stability currency depends on the labor market. Source: https://research.stlouisfed.org/fred2/series/PAYEMS/.

7. **M2 Money Stock (M2NS)**. It is a monetary indicator sign that shows us the amount of money that there is in a country or region. In this fact, it is the amount of money that the houses of the USA have. M2 is formed by saving deposits, the small-denomination time deposits less than 100.000$ and the balances in retail money market mutual funds in amount retail trade. A good situation of M2 makes us understand an upward trend of the currency, because saving is a fundamental matter for the good working of the country. The more savings we have, the more capacity to affront a financial crisis we will have. The currency values significantly at a bigger saving in the reference country. Source: http://research.stlouisfed.org/fred2/series/M2NS.
8. **Trade Balance: Goods and Services, Balance of Payments Basis (BOPGSTB)**. The trade balance in a country shows the exportations minus the importations. If the balance were positive, we would be talking about a positive trade balance that reflects income in the balance of payments. A negative trade balance would damage the economy each time the impact in the trade balance is more relevant because of the globalization. A strong country should have a positive trade balance and it would also have an appreciation in its currency. Source: http://research.stlouisfed.org/fred2/series/BOPGSTB.
9. **WTI Crude Oil Spot Price Cushing**. The price of oil affects significantly to the price of one currency for its relationship with other economic indices. This variable shows the price of a barrel in dollars. Oil is directly related to prices, which is one of the variables studied, the level of household savings and exports. Source: http://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=RWTC&f=D.
10. **S&P 500 Index**. The S&P500 is an American stock index based on the capitalization of five hundred large companies with shares traded on the NYSE or NASDAQ. Source: https://www.quandl.com/data/YAHOO/INDEX_GSPC-S-P-500-Index.
11. **DAX Index (Germany)**. It is the benchmark of Germany and possibly Europe. It contains thirty companies over capitalization and it is a clear indicator of reference when assessing the euro. Source: https://www.quandl.com/data/YAHOO/INDEX_GDAXI-DAX-Index-Germany.
12. **Gold**. It is one of the most traded commodities worldwide. Its price is relevant because the gold is used to protect the part of investors in times of crisis. Therefore it is a good sign of the state of the economy. Source: http://research.stlouisfed.org/fred2/series/GOLDAMGBD228NLBM.

3 Model Design

To perform the prediction of the future values, Least Squares Support Vector Machines (LS-SVMs) [10], have been used. They are kernel-based methods so they are also known as Kernel Ridge Regression method (KRR) [9].

These models are well suited for function approximation and they have some advantages over classical Support Vector Regression (SVR):

- easier mathematical resolution
- the parameter ε used in SVR is not needed
- the number of Lagrange multipliers is reduced to half.

Nonetheless, one of the main problems with LS-SVMs is that they do not generate sparse models so risk of overfitting has to be controlled.

In case we consider Gaussian kernels, σ is the width of the kernel, that together with the regularization parameter γ, are the hyper-parameters of the problem. Note that in the case in which Gaussian kernels are used, the models obtained resemble Radial Basis Function Networks (RBFN); with the particularities that there is an RBF node per data point, and that overfitting is controlled by a regularization parameter instead of by reducing the number of kernels [8].

In LS-SVM, the hyper-parameters of the model can be optimized by cross-validation. Nevertheless, in order to speed-up the optimization, a special formulation for a reduced cost evaluation of the cross-validation error of order l (l-fold CV) taken from the work [1] was used. With this formulation, the error evaluation cost of cross-derivation does not depend on the order l, but on the number of data points of the problem, since in fact the computational cost is dominated by the inversion of the kernels K activation matrix. Such inversion is performed through a Cholesky decomposition; the most efficient exact algorithm for this case is $O(N^3)$ where N is the number of samples.

In order to perform the evaluation of the performance of the stopping criteria in the forward selection strategy, it is necessary to learn a number of LS-SVMs, each one considering the eventual state of the selected subset in the iterative process X_G of the variable selection process. This requires therefore the training of a considerable number of LS-SVM, depending on the problem. In this work, this process was distributed in a computer cluster, so that each training process of a LS-SVM was sent to a different node. This way the computational time was reduced in a factor of N (considering a computer with N nodes, ignoring communication delays), supposing that every execution takes the same amount of computational time.

4 Feature Selection

In this work, a Mutual Information (MI) -based feature selection algoritm has been used with the objective of finding the most relevant factors needed to predict the EUR/USD exchange rate. Mutual information comes from Shannon's Information Theory, and can be expressed as

$$I(X,Y) = H(Y) - H(Y|X), \tag{1}$$

To estimate the mutual information, only the estimate of the joint probability density function (PDF) between X and Y is needed [8]. For continuous variables,

this estimation is complex. However in recent years, a k-nearest neighbours-based mutual information estimator technique [6] has opened the door to more robust MI estimations among groups of variables [3].

The feature selection algorithm used in this work [4] makes use of the Markov Blanket concept [5]. Markov blankets are groups of variables M_i that subsume all the information that a single variable x_i has with respect to a different variable (or group of variables) Y; in practice and for our purposes, with respect to the objective variable. The algorithm consists of a backwards variable selection method which starts with the complete set of variables, and iteratively discards those which are detected to have a Markov Blanket in the remaining set X_G of variables, i.e. those whose information with respect to Y is already present in the remaining set X_G of variables [11].

The algorithm states the following steps:

1. Calculate the MI between every pair of input variables $I(x_i, x_j)$
2. Starting from the complete set of input variables $X_G = X$, iterate:
   a) For each variable x_i, let the candidate Markov blanket M_i be the set of p variables in X_G for which $I(x_i, x_j)$ is highest.
   b) Compute for each x_i
   c) Choose the x_i for which $Loss_i$ is lowest and eliminate x_i from X_G.
3. Continue with step 2 until no variables remain.

This way, a ranking of relevance of variables (in reverse order) is obtained. Under this operation, it is to be noted that variables that have low influence with respect to the output variable (irrelevant variables) will be soon discarded, as $Loss_i$ value should tend to 0. Similarly, redundant variables will be iteratively discarded at earlier stages. Relevant variables with low redundancy will be the last ones in being chosen.

The p parameter of the algorithm (in step 2.a of the algorithm) will take the value $p = 1$, as recommended in previous works [4] [11].

5 Experiments

This section presents the dataset and how the series were defined considering different subsets of variables. Afterwards, the models are designed showing the approximation errors and a final comment on the behaviour of the models considering different variables is made.

5.1 Defining Regressors and Data Sets

Taking the data from the sources specified in previously, the data sets is built considering a monthly based. More concretely, all the measurements in the previous month, three months ago and a year ago. Output variable was differentiated so that difference between the current month and previous month is the objective to be estimated. Given $X(t) = \{x_1(t), x_2(t), x_3(t)..., x_{12}(t)\}$ as the value of

the independent variables at month t, the initial regressors considered for the modeling problem are

$$\hat{Y}(t) = y(t) - y(t-1) = F(X(t), X(t-1), X(t-3, X(t-12)) \quad (2)$$

as suggested by the trading experts, and being $y(t)$ the EUR/USD change at time step t. After arranging all the variables, the final dataset consist of 139 samples corresponding to the 11 years and 7 months of data available.

5.2 Variable Selection and Regression Results

The variable selection was performed building the corresponding LSSVM for each subset of variables obtaining several values for Root Mean Squared Error for test that are represented graphically in Figure 1. It is easy to see that the information provided by most of the variables does not improve the accuracy although there are some of them that are critical to obtain proper results.

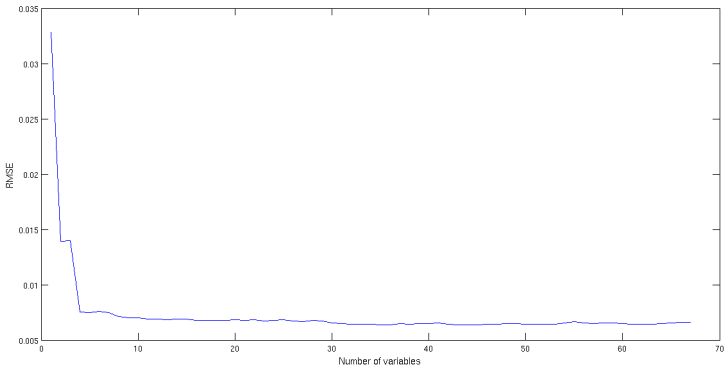

Fig. 1. Evolution of the Root Mean Squared Test Error as the number of variables increases

In Figure 2 are depicted the real output for the EUR/USD asset and the approximations obtained by the LSSVM using 3,10 and all variables. Three variables correspond to the first local minimum in the optimization process (LS-SVM attain a 0.94 of R2, meaning the model explains 94% of the variance of the output); ten variables corresponds to the second local minimum observed in the process(LS-SVM attain a 0.95% of R2, meaning the model explains 94% of the variance of the output). Choosing one or another would depend on the tradeoff between interpretability and accuracy desired. Optimal performance is obtained using in any case 10 variables. Table 1 shows the approximation errors of LSSVM and Radial Basis Function Neural Networks (RBFNNs) showing that, the variable selection is adequate and the modeling can be performed by several paradigms correctly.

Table 1. Approximation errors (using RMSE) comparison between LSSVM and RBFNN with 5 neurons

	LSSVM	RBFNN
3 variables		
Train	0.0138 (1.1e-4)	0.0313 (7.01e-3)
Test	0.0105 (9.3e-3)	0.0116(1.01e-2)
10 variables		
Train	0.0068(2.8e-4)	0.0102 (3.5e-3)
Test	0.0055(4.4e-3)	0.0067(8.4e-3)
All variables		
Train	0.0099(2.7e-4)	0.0230(3.8e-3)
Test	0.00554(3.9e-3)	0.0130(9.9e-3)

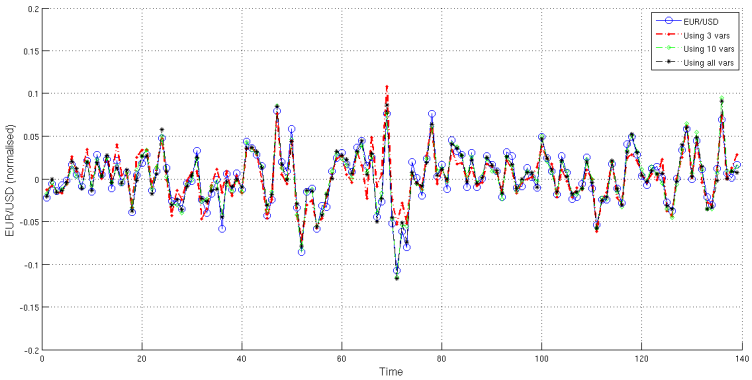

Fig. 2. This figure plots the real output (blue line) versus the approximations made by LSSVM using 3, 10 and all variables, according to equation 2

6 Conclusions and Further Work

The prediction of the relationship between the euro and the US dollar remains as a difficult task due to the macroeconomic environment and the operations on this assets that condition its own value. This paper has presented the application a ranking of macroeconomic variables based on experts' advice and numerical results. The subset of variables were feeded to an LSSVM in order to predict the final value. The reduced subset of variables were able to provide enough information to model the relationship, making the trading easier as traders could consider less variables.

Acknowledgments. This work has been supported by the GENIL-PYR-2014-12 project from the GENIL Program of the CEI BioTic, Granada, and the Junta de Andalucia Excellence Project P12-TIC-2082.

References

1. An, S., Liu, W., Venkatesh, S.: Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression. Pattern Recogn. **40**(8), 2154–2162 (2007)
2. Ettredge, M., Gerdes Jr., J., Karuga, G.G.: Using web-based search data to predict macroeconomic statistics. Commun. ACM **48**(11), 87–92 (2005)
3. François, D., Rossi, F., Wertz, V., Verleysen, M.: Resampling methods for parameter-free and robust feature selection with mutual information. CoRR, abs/0709.3640 (2007)
4. Herrera, L.J., Pomares, H., Rojas, I., Verleysen, M., Guilén, A.: Effective input variable selection for function approximation. In: Kollias, S.D., Stafylopatis, A., Duch, W., Oja, E. (eds.) ICANN 2006. LNCS, vol. 4131, pp. 41–50. Springer, Heidelberg (2006)
5. Koller, D., Sahami, M.: Toward optimal feature selection. In: Saitta, L. (ed.) Proceedings of the Thirteenth International Conference on Machine Learning (ICML), pp. 284–292. Morgan Kaufmann Publishers (1996)
6. Kraskov, A., Stogbauer, H., Grassberger, P.: Estimating mutual information. Phys. Rev. E **69**, 066138 (2004)
7. Martinsen, K., Ravazzolo, F., Wulfsberg, F.: Forecasting macroeconomic variables using disaggregate survey data. International Journal of Forecasting **30**(1), 65–77 (2014)
8. Rossi, F., Lendasse, A., Franois, D., Wertz, V., Verleysen, M.: Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. Chem. and Int. Lab. Syst. **80**, 215–226 (2006)
9. Saunders, C., Gammerman, A., Vovk, V.: Ridge regression learning algorithm in dual variables. In: Proceedings of the 15th International Conference on Machine Learning, pp. 515–521. Morgan Kaufmann (1998)
10. Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, J., Vandewalle, B.: Least Squares Support Vector Machines. World Scientific, Singapore (2002)
11. Del Mar Perez, M., Val, J., Negueruela, I., Lafuente, V., Herrera, L.J.: Firmness prediction in prunus persica calrico peaches by visible/short-wave near infrared spectroscopy and acoustic measurements using optimised linear and non-linear chemometric models. J. Sci. Food Agric., 15, September 2014