ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 237-244

Input Data Reduction for the Prediction of Financial Time Series

A. Lendasse[1], J. Lee[2], E. de Bodt[3], V. Wertz[1], M. Verleysen[2]

[1]Université catholique de Louvain, CESAME, 4 av. G. Lemaître
B-1348 Louvain-la-Neuve, Belgium, {lendasse, wertz}@auto.ucl.ac.be.
[2]Université catholique de Louvain, Electricity Dept., 3 pl. du Levant,
B-1348 Louvain-la-Neuve, Belgium, {lee, verleysen}@dice.ucl.ac.be.
[3]Université catholique de Louvain, IAG, 1 pl. des Doyens,
B-1348 Louvain-la-Neuve, Belgium, debodt@fin.ucl.ac.be.

Abstract. Prediction of financial time series using artificial neural networks has been the subject of many publications, even if the predictability of financial series remains a subject of scientific debate in the financial literature. Facing this difficulty, analysts often consider a large number of exogenous indicators, which makes the fitting of neural networks extremely difficult. In this paper, we analyze how to aggregate a large number of indicators in a smaller number using -possibly nonlinear- projection methods. Nonlinear projection methods are shown to be equivalent to the linear Principal Component Analysis when the prediction tool used on the new variables is linear. The methodology developed in the paper is validated on data from the BEL20 market index.

1. Introduction

Since the beginning of this century, the question of the predictability of financial series (at least of stock market prices) has been the subject of a highly controversial debate in finance [1,2,3]. Many empirical works, mainly based on linear statistical tests, have conducted to the same conclusion in the years sixties and seventies, despite the heavy use of charts and technical indicators by the professional community. On the basis of all empirical evidences, we will consider that there is some interest in trying to predict the evolution of financial asset prices, as do Refenes, Burgess and Bentz [4] in their introduction to the methods used in financial engineering. When time series prediction is viewed as a regression problem, the inputs being past values of the series and exogenous variables, one may expect useful information (for the prediction of the series) to be contained in these inputs. Nevertheless, it is difficult to know if the information content of specific inputs is relevant, redundant or useless. Furthermore, it is well known that any regression method (in particular non-linear ones), is difficult to use when the number of inputs is large. There is thus a strong interest in reducing the number of inputs; the question is how to reduce the number of inputs without loosing relevant information. In this paper, we will focus on this question and we will compare the classical linear data compression approach with a

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 237-244

new non-linear one. The objective will be to keep as much as possible the information contained in the initial inputs, while reducing as much as possible the number of new "constructed" variables (or the dimension of the projection space). The new "constructed" variables are then used as input to the prediction algorithm, as illustrated in Fig. 1:

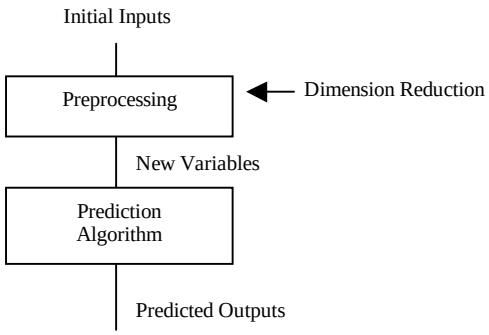

Fig. 1. The two steps of the methodology.

To perform the transformation between the initial inputs and the new variables, we may choose to use a linear method or a non-linear one. In this paper, we will use Principal Component Analysis as a linear transformation and Curvilinear Component Analysis [5] as a non-linear one.

2. Dimension reduction

2.1. Intrinsic Dimension

First, it is important to evaluate the projection dimension, i.e. the dimension of the space of new variables. If the estimation of this dimension is too small, information will be lost in the projection. If it is too large, the usefulness of the method is lost. To evaluate this dimension, the concept of intrinsic dimension is used. The intrinsic dimension is the effective number of degrees of freedom of a set [6]. This concept is presented here with the well-known horseshoe distribution (Fig.2): for this data set, the intrinsic dimension is equal to two as two degrees of freedom are sufficient to uniquely determine any data in the set, although the data live in R^3. A possible way of computing this intrinsic dimension is explained in [7], but its determination remains very difficult to apply, not to say approximate, for high dimensional data sets.

2.2. Curvilinear Component Analysis

This nonlinear extension of the Principal Component Analysis [5] spreads out the manifold that contains the data and projects it from a high dimensional space to a smaller dimensional one. The projection of the horseshoe distribution carried out by Curvilinear Component Analysis (CCA) is illustrated in Fig. 3.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 237-244

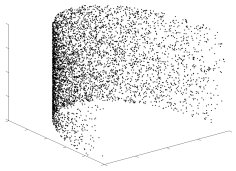

Fig. 2. Horseshoe distribution.

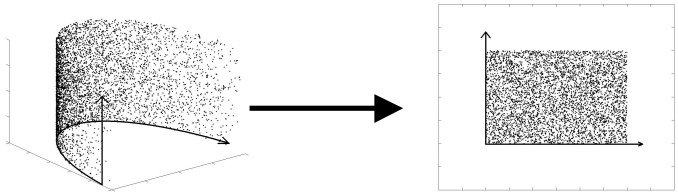

Fig. 3. Projection carried out by CCA from R^3 to R^2.

2.3. Evaluation Criteria

The goal of the projection is the reduction of the number of inputs (for the prediction method) and the preservation of the initial information. Two methods can be used to evaluate how the projection is successful (with respect to the projection dimension). Looking to Fig. 1 again, one can try to measure if the new variables contain the same information as the initial inputs, without taking care of the subsequent prediction algorithm. The quality of the projection can also be evaluated by looking at the end of the chain, i.e. the predicted outputs. These two methods are illustrated below through a first criterion and two next ones, respectively.

2.3.1. Mean Square Reconstruction Error

The projection by PCA or CCA is a reversible operation; nevertheless the result of an inverse projection will not correspond to the initial inputs. The mean distance between the initial inputs and the results of the inverse projection is called the Mean Square Reconstruction Error (MSRE). If the dimension of the projection space is decreased below the intrinsic dimension, then the MSRE will increase; this will validate or invalidate the rough intrinsic dimension estimation. The main advantage of this criterion is that it is totally independent of the prediction method used afterwards. Unfortunately, this advantage has a counterpart: there is no guarantee that a projection dimension chosen according to this measure will be optimal when the best prediction outputs are looked for.

2.3.2. Mean Square Error

The Mean Square Error (MSE) after prediction is defined by

$$MSE = \sqrt{\frac{\sum_t (y_t - \hat{y}_t)^2}{\sum_t 1}} . \quad (1)$$

The MSE must be made as small as possible for a successful prediction. It is worth noticing that typical MSE (on normalized data) values in financial application are much higher than in other domain, because of the inherent difficulty to predict such time series accurately.

2.3.3. Percentage of Correct Approximations of Sign

As the amplitude of financial time series is always difficult to predict, a less difficult task is to predict the sign of a series (for example a return on financial index), leading to the Percentage of Correct Approximations of Sign (PCAS) criterion defined by

$$\mathrm{PCAS} = \frac{\sum_t \left(\mathrm{sign}(y_t)\,\mathrm{sign}(\hat{y}_t) + 1\right)}{2\sum_t 1} 100 . \quad (2)$$

2.4. Methodology

The procedure used to project the initial inputs on new relevant variables is described here, together with the tests that determine if the projection improves the results of prediction, and if a nonlinear projection must be used instead of classical linear one. In the following, we denote by:

- u_t^i, $1 \le i \le n$, the technical indicators or exogenous variables used;
- z_t the input vector containing the values of the series;
- T the total number of samples available in the series ($1 \le t \le T$);
- D the intrinsic dimension of the learning set.

2.4.1. Mixture

The input vectors z_t are randomly mixed according to a draw without replacement.

2.4.2. Learning and Validation sets

The data must be divided in learning and validation sets. The validation set is used to check the results obtained on the learning set and justify that no overfitting has occurred. The division is made between the learning set z^L and the validation set z^V. These sets are normalized.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 237-244

2.4.3. Intrinsic Dimension

The intrinsic dimension d of the technical indicators of the learning set is computed and will be used as a first rough approximation of the projection dimension.

2.4.4. Preliminary Principal Component Analysis

A Principal Component Analysis of the technical indicators in the learning set is performed. The dimension d_1 is chosen as the minimal dimension for which the loss of information after PCA is negligible. The technical indicators of the learning and validation sets are projected from a n-dimensional space to a d_1-dimensional space according to the parameters defined by the PCA on the learning set.

2.4.5. Preliminary Curvilinear Component Analysis

A first Curvilinear Component Analysis is performed on the n-dimensional space, reducing it to a d_2-dimensional space. At this stage, dimension d_2 is chosen equal to the intrinsic dimension calculated above.

2.4.6. Error Criteria on the validation set

The d_1 and d_2 dimensions are now varied; new PCA and CCA projections are respectively performed. In each case, the MSRE error on the validation set is computed. Then, the MSE and the PCAS are evaluated on the validation set. For the purpose of computing these two last criteria, we need the predicted outputs $\hat{y}_t$; any prediction algorithm can be used to compute these values. Note that in practise we use extended cross-validation instead of a single validation set [8].

3. Application to the bel20 market index

The methodology is tested on the Bel20 market index from December 1st, 1987 to February 14th, 1998 (2663 daily observations). 42 technical indicators have been used.

As represented in Fig.4, the daily return y_t of this market is calculated according to

$$y_t = \log\left(\frac{BEL20_t}{BEL20_{t-1}}\right) \qquad (1)$$

The methodology described in the previous section is then applied.

- Stages 2.4.1 and 2.4.2 are performed.
- Stage 2.4.3: the intrinsic dimension of the technical indicators is 7.
- Stage 2.4.4: The result of the Principal Component Analysis is shown in Fig.5. This result is the percentage of information kept by a Principal Component Analysis versus the projection dimension. It is equal to one minus the normalized MSRE.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 237-244

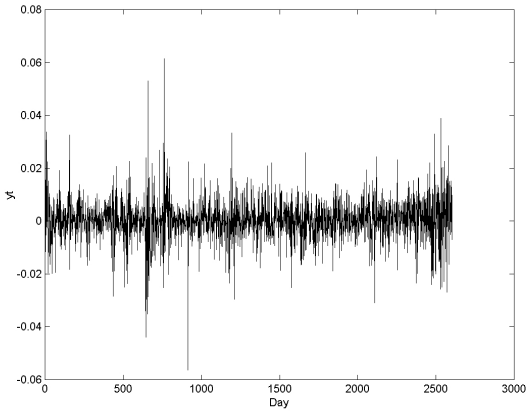

Fig. 4. Daily return of Bel20 Market Index from December 1987 to February 1998.

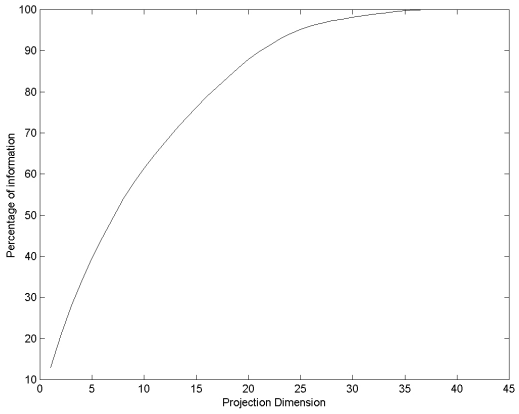

Fig. 5. Percentage of information kept by a PCA versus the projection dimension.

- Stage 2.4.5: CCA from a 42-dimensional space to a 7-dimensional space.
- Stages 2.4.6: a new set of projections is performed from the 42-dimensional space to the d-dimensional space when varying d around 7. Note that, for the sake of validation of our methodology, the computations have been done for d = 2 to 24. A linear prediction model of y_t based on the projected data is estimated on the learning set. Then, the mean square error and the percentage of correct approximations of sign are calculated on the validation set; theses two criteria are shown in Fig. 6 and 7 respectively.

4. Conclusions

It can be noticed that the PCA method gives slightly better results that the CCA method. We might have expected the opposite, or at least equivalent results.

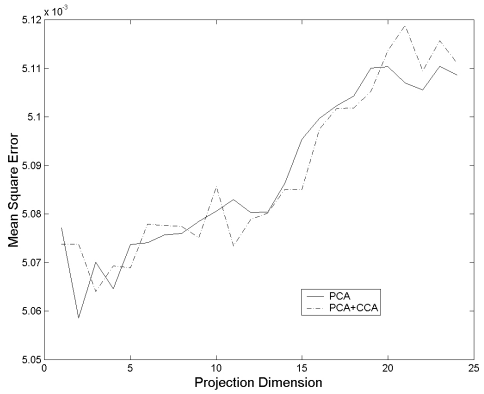

Fig. 6. Mean square error on the validation set versus the projection dimension.

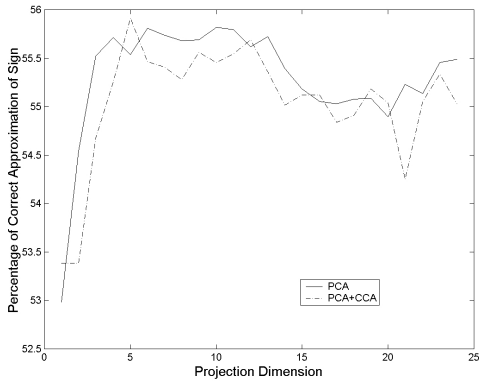

Fig. 7. Percentage of correct approximations of sign on the validation set versus the projection dimension

In fact, this phenomenon may be explained by a too small number of data and by the important noise contained in the series. Another explanation is the use of a linear model for prediction. One may expect that a nonlinear prediction method would give improved results, especially when the initial inputs are preprocessed in a nonlinear way; some results obtained with a nonlinear model may be found in [9]. Nevertheless, it must be mentioned that problems related to a difficult convergence of nonlinear models add to the difficulty of choosing the parameters in our methodology, making the results more difficult to illustrate. The best projection dimension depends on the criterion: 2 with the mean square error and 4 to 6 with the percentage of correct approximations of sign; this last one is the only criterion for which a nonlinear projection gives improved results. Referring to (1), it must be mentioned that this series increases about 53% of the time. It must be stressed that obtaining results on the forecasting of the sign of daily returns time series is particularly important. The anticipation of the orientation of the market is at the basis of any market timing strategies, which justify the technical analysis approach. Furthermore, as mentioned

above, a non-linear prediction algorithm [9] would still improve the results (the level of improvement remains limited here because a simple linear prediction model is used).

Even if it is considered that there is some interest in trying to predict the evolution of financial asset prices, the use of a large number of technical indicators remains difficult with any prediction tool. This paper shows how to use –possibly nonlinear- data compression techniques to reduce the number of technical indicators used for the prediction. The application of this methodology on the BEL20 Market Index shows that comparable results are obtained when using a linear projection method or a non-linear one, when a subsequent linear prediction tool is used. The advantage of the methodology presented here is that it automatically evaluates the number of new variables that must be kept after projection, in order to keep the necessary and relevant information needed for the prediction.

Acknowledgement

Michel Verleysen is Research Associate of the Belgian National Fund for Scientific Research (FNRS). Part of this work has been founded by the "Ministère de la Région Wallonne", under the "Programme de Formation et d'Impulsion à la recherche scientifique et technologique".

References

[1] Fama E., "*The Behavior of Stock Market Prices*", Journal of Business, 38, pp. 34-105, 1965.

[2] Bachelier L., *Théorie de la spéculation*, Gauthier-Villars, Paris,1900.

[3] Campbell J., Lo A. and MacKinlay A. C., "*The Econometrics of Financial Markets*", Princeton University Press, 1997.

[4] Refenes A.P., Burgess A.N. and Bentz Y., "*Neural Networks in Financial Engineering: A Study in Methodology*", IEEE Transactions on Neural Networks, vol. 8, n°6, pp. 1222-1267, November 1997.

[5] Demartines P., Hérault J., "*Curvilinear Component Analysis: A self-organizing neural network for nonlinear mapping of data sets*", IEEE Trans. on Neural Networks, 8(1), pp. 148-154, 1997.

[6] Takens F., "*On the numerical Determination of the dimension of an attractor*", Lecture Notes in Mathematics, Vol. 1125, Springer-Verlag, pp. 99-106, 1985.

[7] Grassberger P. and Procaccia I., "*Measuring the Strangeness of Strange Attractors,*" Physica *D*, 56, pp. 189-208, 1983.

[8] Bishop C.M., "*Neural Networks for Pattern Recognition*", Glarendon Press, Oxford 1995.

[9] Lendasse A., de Bodt E., Wertz V. and Verleysen M., "*Nonlinear Financial Time Series Forecasting – Application to Bel20 Stock Market Index*", European Journal of Economic and Social Systems, 14, N°1, pp 81-91, 2000.