ELSEVIER

Neurocomputing 14 (1997) 123–138

NEUROCOMPUTING

Financial time series modelling with discounted least squares backpropagation [1]

A.N. Refenes *, Y. Bentz, D.W. Bunn, A.N. Burgess, A.D. Zapranis

Department of Decision Science, London Business School, Sussex Place, Regents Park, London NW1 4SA, United Kingdom

Received 29 May 1995; accepted 28 December 1995

Abstract

We propose a simple modification to the error backpropagation procedure which takes into account gradually changing input-output relations. The procedure is based on the principle of Discounted least squares whereby learning is biased towards more recent observations with long term effects experiencing exponential decay through time. This is particularly important in systems in which the structural relationship between input and response vectors changes gradually over time but certain elements of long term memory are still retained. The procedure is implemented by a simple modification of the least-squares cost function commonly used in error backpropagation. We compare the performance of the two cost functions using both a controlled simulation experiment and a non-trivial application in estimating stock returns on the basis of multiple factor exposures. We show that in both cases the DLS procedure gives significantly better results. Typically, there is an average improvement of above 30% (in MSE terms) for the stock return modelling problem.

Keywords: Neural networks; Time series analysis; Discounted least squares; Financial engineering; Stock selection

* Corresponding author. Email: prefenes@lbs.lon.ac.uk

[1] This research is supported by the Department of Trade and Industry under the NCTT programme and the corporate members of the NeuroForecasting Club. We would like to thank Barclays-BZW, Citibank International, Mars Group, Postel Investment Management, Sabre Fund Management, and Societe Generale, for their material and technical support.

0925-2312/97/$17.00 © 1997 Elsevier Science B.V. All rights reserved
PII S0925-2312(96)00005-7

1. Introduction

Neural networks have attracted much interest in financial engineering but as with the more traditional econometric techniques, the special characteristics of many financial time-series have posed a number of modelling challenges. One of the most problematic of these is concerned with structural change in the data. Not only is a single data series frequently non-stationary (in the sense of its mean, variance and covariance changing over time), but the specification of its relationship to other related data-series may also be changing. Consider for example the relationship between an asset price and its determinants. It is generally believed that there are gradual structural changes both in the characteristics of the determinant time series and in the actual relationship between independent and dependent variables. These changes tend to occur slowly over time as the economic environment evolves and although short term trends may have diversionary effects many of the underlying economic laws still apply in the long run. This creates a fundamental problem when attempting to model this type of time series. The (parameters of the) models must be re-estimated to make use of the most recent data conveying these changes. But if one merely uses an extending window with static networks, the effect of the most recent observations is *averaged* out. On the other hand, if one uses a moving window one runs the risk of "chasing a moving target". Both static networks and fully adaptive methods have problems. Static methods tend to have an averaging effect over (potentially) long periods which is undesirable. Adaptive methods on the other hand tend to concentrate on short term trends often penalising the long term tendency of economic variables to move back in line, albeit at different levels.

Structural change is evidently an issue for any methodological approach to time-series analysis, and has indeed occupied an increasing amount of research in econometric theory (e.g. [1,2,5]). Because of the theoretical emphasis upon model parsimony in traditional time-series and econometric modeling, the structural change issue manifests itself both as a problem of *model specification* (i.e. identifying the correct functional form for the model) and *time-varying parameters* (i.e. estimating the model's parameters as stochastic processes in themselves). With this distinction, model specification can be dealt with using intervention analysis or regime-switching meta-modeling (see for example the Bayesian approaches to model-switching [9]) and the time-varying parameters by estimation using adaptive techniques (e.g. structural time-series models using Kalman filtering [6]). However, in the context of more highly parameterised modelling approaches, such as neural networks, this distinction between model specification and parameter estimation is not so clear and the former is essentially subsumed into the latter. Thus, for a given topology, the weight estimation procedure effectively performs a model specification function as well.

One of the simplest time-varying approaches to model estimation in the time series literature is to is to use Discounted Least Squares (DLS) as an estimation criterion rather than the usual Ordinary Least Squares. For example, Bonini and Freeland [4] used a regression model for forecasting, in which the past observations were weighted according to an exponential decay function. The approach of DLS, however, is more general and can admit any form of decay function to the data. In this paper we adapt this idea to the neural network context. This implementation is based on a simple modification of

the cost function which biases learning towards most recent observations in a time series but without ignoring long term effects. In Section 2 we describe backpropagation with Discounted Least Squares. In Section 3 we analyse its performance on a simple benchmark with weak non-stationarity. In Section 4 we use the DLS procedure to model asset returns in a non trivial (multi-variate) context and finally in Section 5 we make some concluding remarks.

2. Backpropagation with discounted least squares

Learning by error backpropagation is an error minimisation procedure which uses gradient descent into weight error space to minimise a quadratic measure of total error. The most commonly used error measure is the Ordinary Least-Squares criterion (OLS) as shown in (1).

$$E_{\mathrm{LS}} = \frac{1}{2N} \sum_{p=1}^{N} (t_p - o_p)^2, \tag{1}$$

where N is the total number of observations in the sample, t_p is the desired response and o_p is the observed (or actual) response, with $p = N$ being the most recent observation. Least Squares measures give equal weight to all observations in the sample (training set). In time series analysis with structural changes it is often desirable to overweight more recent observations for the reasons discussed above.

This is particularly important in systems with low frequency data (e.g. economic time series) and also in other physical systems with similar non-stationary characteristics. The idea behind the Discounted Least Squares procedure proposed here is that in low frequency financial data the structural relationship between an asset price and its determinants changes gradually over time as the economic environment evolves. Thus recent observations should be weighted more heavily than older observations. The cumulative error calculated by the DLS procedure is given by

$$E_{\mathrm{DLS}} = \frac{1}{2N} \sum_{p=1}^{N} w(p)(t_p - o_p)^2, \tag{2}$$

where N is the total number of observations in the training set, t_p is the target network output, o_p is the predicted network output, and $w(p)$ is an adjustment of the contribution of observation p to the overall error. In general, there are many different ways of biasing the cost function through $w(p)$ (such as linear, exponential, etc.) to differentially weight the contribution of each observation towards the total error. In this paper we examine a simple sigmoidal decay as shown in (3) and evaluate the performance of the modified cost function against the more usual OLS.

$$w(p) = \frac{1}{1 + e^{(a - bp)}}, \tag{3}$$

where

$$b = \frac{2a}{N}. \tag{4}$$

The parameters a and b are used to scale and offset the sigmoid. The DLS cost function is asymptotically invariant with respect to the sample size (N). Since b in (4) is derived from a and N, the only control parameter is the discount rate a.

The effect of the discount rate can be seen in Fig. 1. The smoothness of the sigmoid function ensures that, for relatively low discount rates a, the weighting $w(p)$ is also smooth, reflecting a gradual discounting of past observations. When $a \to 0$ then $w(p)$ is constant for all patterns (x-axis). Actually,

$$\lim_{a \to 0} w(p) = 0.5. \tag{5}$$

In this case the DLS cost function becomes

$$\lim_{a \to 0} E_{\mathrm{DLS}} = \frac{1}{4N} \sum_{p=1}^{N} (t_p - o_p)^2 = \tfrac{1}{2} E_{\mathrm{LS}}. \tag{6}$$

When $a \to \infty$ then $w(p)$ becomes a scalar function:

$$w(p) = \begin{cases} 0 & \text{when } p < N/2, \\ 1 & \text{when } p > N/2. \end{cases} \tag{7}$$

In this case the DLS cost function becomes

$$\lim_{a \to \alpha} = \begin{cases} 0 & \text{when } p < N/2, \\ \dfrac{1}{2N} \displaystyle\sum_{p=1}^{N} (t_p - o_p)^2 = E_{\mathrm{LS}} & \text{when } p > N/2. \end{cases} \tag{8}$$

The discount rate is defined as the ratio

$$\text{d.r.} = \frac{w(N) - w(1)}{N}. \tag{9}$$

The learning (weight update) rule is derived in the usual way by repeatedly changing the weights by an amount proportional to

$$\frac{\partial E_{\mathrm{DLS}}}{\partial W} = \frac{1}{1 + e^{(a - bp)}} \frac{\partial E_{\mathrm{LS}}}{\partial W}. \tag{10}$$

Note that the discount factor $w(p)$ for an observation is a function of the recency of the observation and is independent of the actual order of pattern presentation within the learning procedure. This means that the algorithm is not affected by randomising the order in which patterns are presented and can be applied to both batch and stochastic update rules.

In the following sections we compare the performance of the two cost functions using both a benchmark (i.e. a sine wave with changing amplitude) and a non-trivial application in exposure analysis of stock returns to multiple factors.

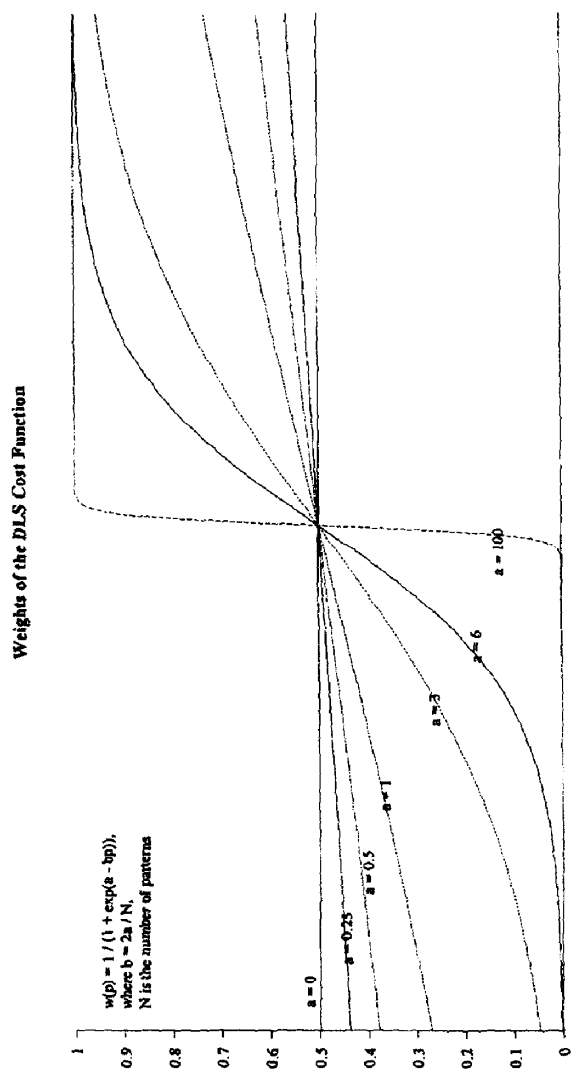


Fig. 1. Weights $w(p)$ in the DLS cost function (Y-axis) as a function of the discount rate a. In the X-axis p represents the training pattern sequence counting forwards from the least recent observation.

3. Controlled simulation analysis

3.1. Experimental set-up

The performance characteristics of the DLS procedure are analysed using a simple sinusoid in two controlled experiments. In the first experiment we focus our analysis in the case that the variance of the time series is changing through time. In this experiment we use a simple sine wave with changing amplitude as shown in Fig. 2.

The data consists of seven periods of the sine wave; each of which contains 100, consecutive and equally spaced data points. We use the first six periods (i.e. 600 data points) for training and the remaining (seventh) period for testing.

The second case of non-stationarity deals with the situation in which both the variance and mean of the time series are changing over time. In the second experiment we use a similarly generated data series with a sinusoid using seven periods as shown in Fig. 3. The variance of the series increases linearly in $(n + x)$ which could clearly be modelled directly. However, in general the form of the non-stationarity is not known a priori. For the purposes of the benchmark, we compare the effect of including indirect information in the form of the discount function about $w(p)$ to the results obtained when ignoring $w(p)$ completely. This is equivalent to comparing the assumption of weak non-stationarity on the one hand to the alternative assumption that data from the different periods is equally valid in the future.

Again we use the first 600 points for training and the remaining 100 points for out of sample testing. In both cases, we use backpropagation networks with a single hidden layer of 12 units and batch update.

3.2. Non-stationary variance

To investigate the effects of non-stationary variance on the least-squares procedure several networks were trained to obtain best fit in-sample of the function shown in Fig. 2. Typically, for the least-squares procedure, the mean square error drops off after 1000 iterations and the improvement tails off at 10000 iterations. Likewise, the DLS procedure with the parameters set to $(a = 3,\ b = [2a]/N)$ the mean square error tails off at

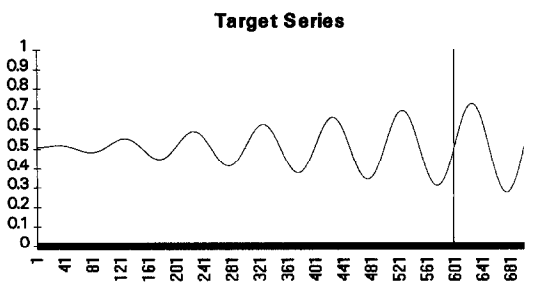


Fig. 2. Non-stationary variance; $y = a(n + x) * (\sin(2\pi x) + b,\ x \in [0, 1],\ n \in \{0, \ldots, 6\}$. The constants, a and b, are used to scale y linearly in the range of the transfer function.

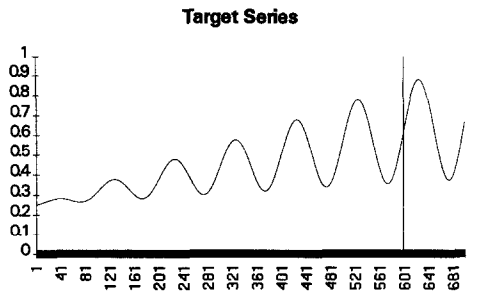


Fig. 3. Non-stationary series with changing mean and variance; $y = a(n + x) + b(n + x) * \sin(2\pi x)$.

the same number of iterations. Fig. 4 shows the estimators' fitness in- and out-of-sample.

As expected (see Fig. 4), the LS procedure fits an estimator which minimises the *average* error over all observations. The minimum MSE is around the third period of the sinusoid. The DLS procedure fits an estimator which is biased towards the more recent observations with a bias determined by the decay parameters in the sigmoid. In this case, the best fit of the DLS estimator is around the fifth period of the sigmoid. A better fit (to the more recent data) would be obtained if a higher discount rate were used. For this artificial example, this would be possible because of the regularity of the underlying function but instead a relatively low discount rate has been chosen to reflect the trade-offs which occur when modelling real time-series. The out of sample fitness is shown in Fig. 5. As expected, the DLS procedure clearly outperforms its least-squares counterpart.

The difference in prediction accuracy is better visualised in the scatterplot shown in Fig. 6 depicting target (x-axis) against predicted (y-axis). Ideally, we should expect a straight line passing through the origin at an angle of 45 degrees. The divergence from this line is a good measure of fitness.

The unbiasedness of forecasts is often assessed by means of a regression of the outcomes against the forecasts (e.g. [7]). Thus, for an unbiased forecasting method,

$$y = a + by_{\mathrm{EST}} + u \tag{11}$$

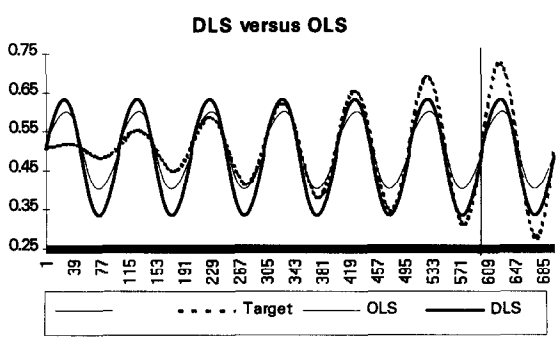


Fig. 4. Estimator fitness in- and out-of sample.

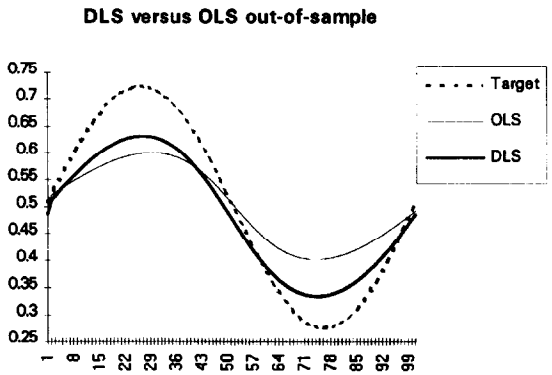


Fig. 5. LS vs DLS in out of sample predictions.

should have $a = 0$, $b = 1$ and u is white noise. The regression lines for the LS estimator is given in (12) below,

$$y = \underset{(0.009)}{-0.59} + \underset{(0.02)}{2.18} y_{\mathrm{LS}}, \tag{12}$$

where the coefficients' standard errors are in parentheses. The estimated coefficients are clearly significantly different from 0 and 1 respectively.

The regression line for the DLS estimator is significantly better (with the intercept closer to zero and the slope closer to one):

$$y = \underset{(0.004)}{-0.21} + \underset{(0.01)}{1.47} y_{\mathrm{DLS}}. \tag{13}$$

However, we still have to reject the unbiasedness hypothesis, even though it is significantly better than the OLS.

3.3. Non-stationary mean and variance

The second benchmark consists of a function which has non-stationarity both in variance and mean (as shown in Fig. 3). Fig. 7 shows the out-of sample predictions for

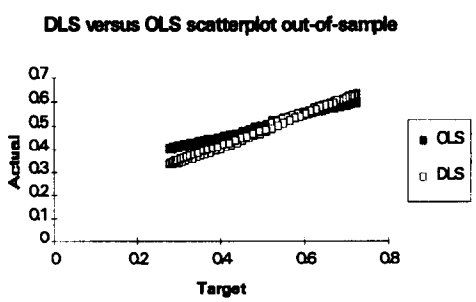


Fig. 6. Scatterplot-actual vs predicted out-of sample.

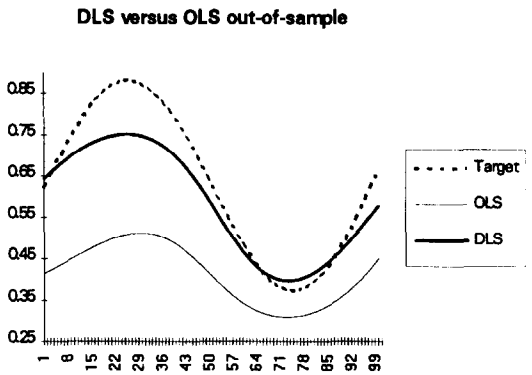


Fig. 7. DLS vs LS out-of-sample predictions with changing mean and variance.

the second controlled simulation. The procedure used is the same as that described in the previous section. As shown in Fig. 7, the least-squares procedure is estimating a lower mean and variance than its DLS counterpart.

The prediction accuracy of the two procedures is better visualised in the Scatterplot shown in Fig. 8 depicting estimated against actual values for the seventh period of the sinusoid.

The regression lines for the LS and DLS estimators from Fig. 8 are given in (14) and (15) respectively. As expected LS underestimates both the mean and variance of the time series in comparison to DLS.

$$y = \underset{(0.01)}{-0.39} + \underset{(0.03)}{2.48} y_{\mathrm{LS}}, \tag{14}$$

$$y = \underset{(0.01)}{-0.15} + \underset{(0.02)}{1.34} y_{\mathrm{DLS}} + u_1. \tag{15}$$

Again, the DLS estimate is again less biased, ($|-0.15| < |-0.34|$), but in both cases we reject the hypothesis of unbiasedness.

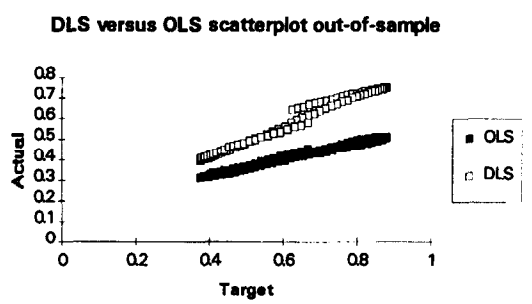


Fig. 8. Scatterplot for LS and DLS – out-of sample.

4. Modelling stock returns

Our non-trivial task is to predict the return of a stock, given the values of several indicators such as Long Term Interest Rates, Earnings per Share, etc. Specifically, we model the thirty day stock return as a function of six parameters,

$$\frac{\Delta S^{(30)}}{S} = f\left(\frac{\Delta LR^{(30)}}{LR}, \frac{\Delta SR^{(30)}}{SR}, \frac{\Delta EPS^{(30)}}{EPS}, \frac{\Delta \$^{(30)}}{\$}, \frac{\Delta S^{(-30)}}{S}, PER\right), \tag{16}$$

where

LR: long rates; thirty day change
SR: short rates, thirty day change
EPS: Earnings Per Share (IBES-estimation)
$: US$ against *FF*; thirty day change
PER: Price to Earnings ratio

The values of the inputs and outputs are daily closing prices of CAC-40 stocks in the Paris stock Exchange. The available pre-processed data (normalised and smoothed) consists of 1025 observations. Table 1 shows a part of the unscaled (raw) data. We use feedforward, multi-layered and fully connected networks.

The learning algorithm is the standard backpropagation with a momentum term. The activation function is the asymmetric sigmoid within the range (0, 1) and the parameter for the discount rate is $a = 3$. When not otherwise stated, the first 800 patterns are used for training and the remaining 225 for testing.

We experimented with a wide range of architectures having fixed all control parameters except for the cost function. The experiments not concerning the DLS procedure are discussed fully elsewhere (e.g. [8]).

Fig. 9 shows the goodness of fit of the two methods in-sample (in the best case). The solid line is the target return, the dotted line is the predicted return (in-sample) for 800 days. In Fig. 9(a) the least-squares procedure is trying to minimise the error equally across all observations and it is making significant overshoots at the most recent data (having fitted the earlier part of the curve rather well). In Fig. 9(b), the DLS procedure shows near perfect fit for the most recent observations with most of the error appearing at the beginning of the time series as expected.

Table 1.
Unscaled data. $X_1(t)$ to $X_6(t)$ are to network inputs and $Y(t)$ is the network output.

$X_1(t)$	$X_2(t)$	$X_3(t)$	$X_4(t)$	$X_5(5)$	$X_6(t)$	$Y(t)$
0.003348	0.064516	−0.004617	−0.017452	0.002174	12.900000	−0.027766
0.007804	0.064516	0.004712	−0.015734	−0.031470	12.300000	0.021049
0.007752	0.056338	0.005140	−0.008803	−0.031980	12.100000	0.063288
0.004405	0.064021	0.001852	0.001770	−0.029272	12.100000	0.001052
0.001100	0.078680	−0.001994	0.001770	0.009122	12.400000	−0.018091
−0.018559	0.014652	−0.000570	0.000000	−0.013956	11.900000	0.033024
−0.036559	0.014652	−0.004128	0.007092	−0.006136	11.800000	0.046509
−0.022581	0.007273	0.000196	0.003527	−0.046519	11.500000	0.121959

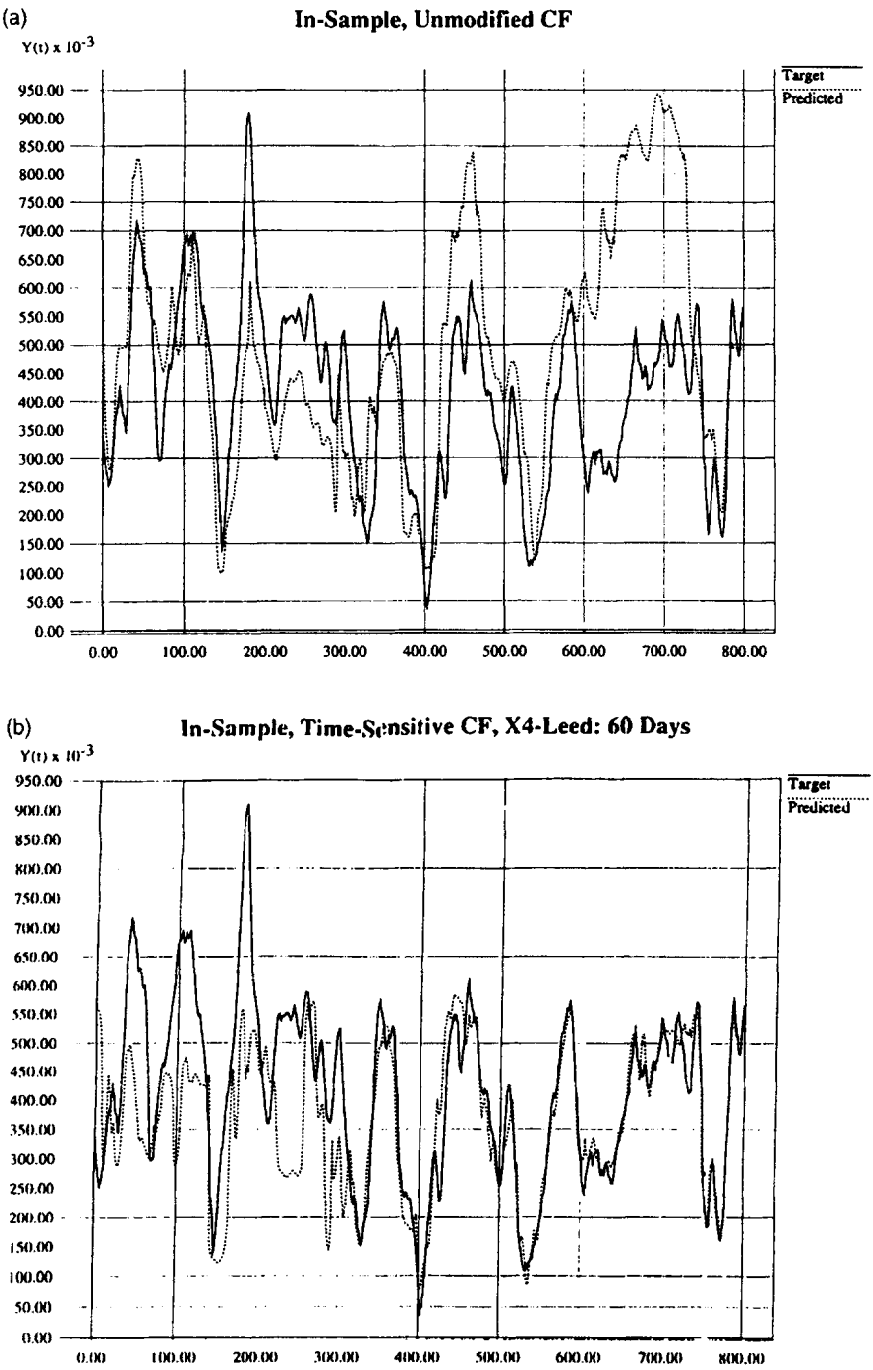


Fig. 9. Target and Predicted $Y(t)$, *in-sample*, (a) LS – fitting well the first part of the data but overshooting the most recent observations, (b) DLS for the exchange rate – near perfect fit for recent observations, less so for early data but still detecting major features of least recent data.

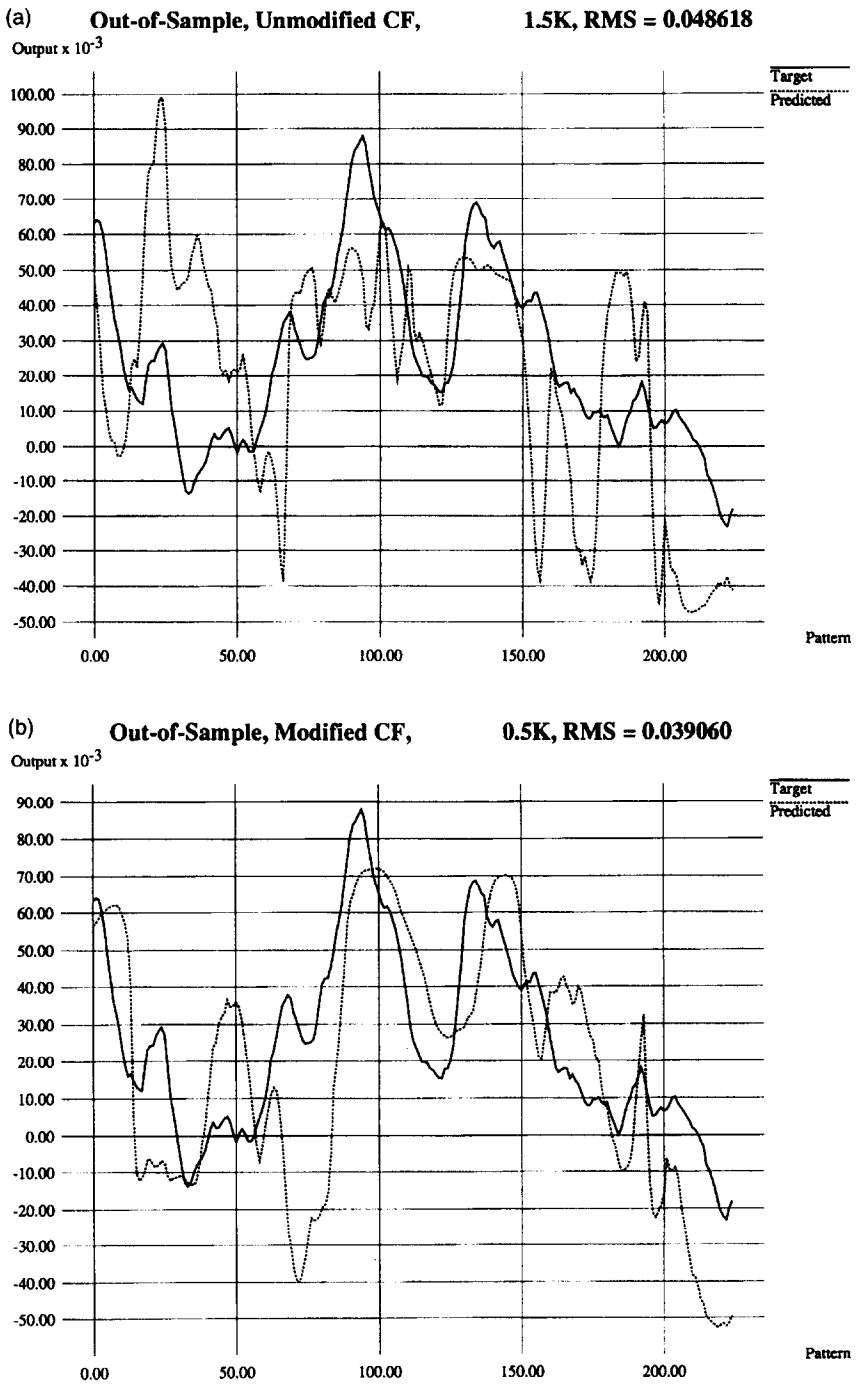


Fig. 10. Target and Predicted $Y(t)$, *out-of-sample*, for 220 days; (a) LS; (b) DLS.

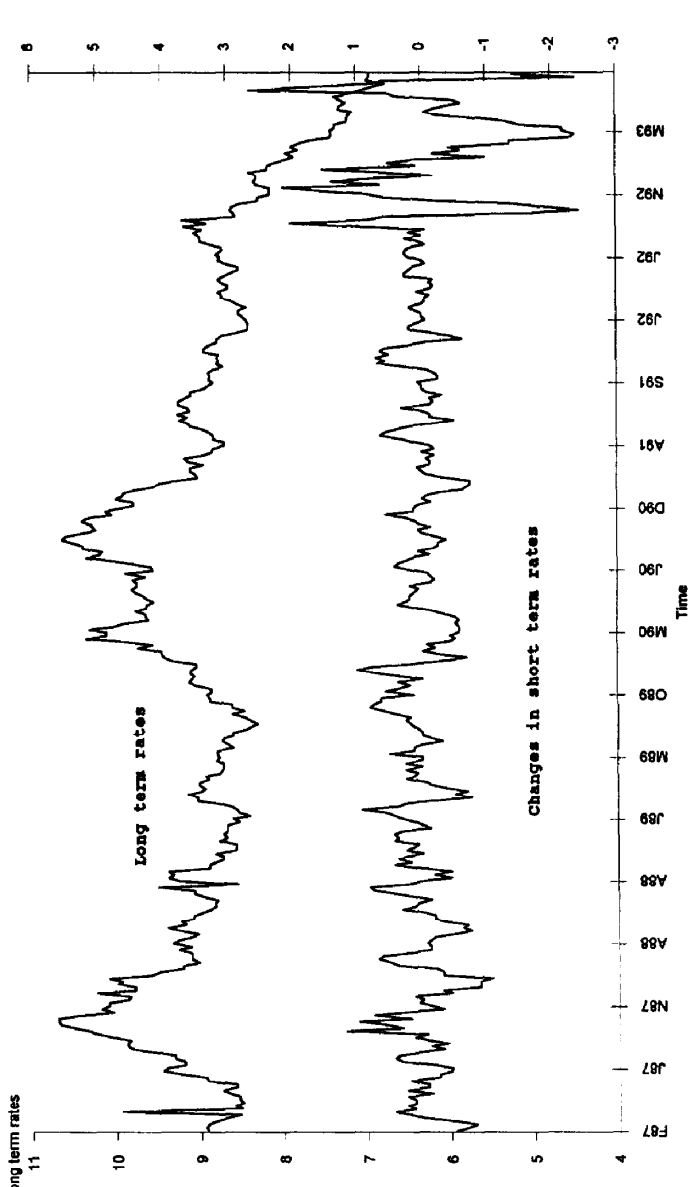


Fig. 11. Non-stationarity in mean (long rates) and variance (short rates).

Fig. 10 shows the goodness of fit of the two methods *out-of-sample*. Again the solid line is the actual return, the dotted line is the predicted return for 200 days immediately after training. The DLS procedure shows a substantial improvement in generalisation performance which is visible with the naked eye (a 30.5% improvement in terms of RMS).

This is clearly a desirable result in terms of the practical financial implications. It is also interesting, however, to investigate why the use of DLS has this effect in this particular case. For example, the relevance and behaviour of short term rates (one of the input variables) can be seen to differ substantially through time. During the first part of the data set short term rates show little volatility but this increases during the later period, whilst long rates also exhibit weak non-stationarity but in terms of the mean level of the series rather than its variance. This is shown clearly in Fig. 11. From 1987 through to September 1992, the volatility in short rates is fairly constant. Since the ERM crisis, short rates have started to play a more central role in economic policy and their volatility nearly doubled. On the other hand up to this period long-rates, although not strictly stationary, follow a more or less consistent pattern, i.e. there are the expected trends as they are used as an instrument to control inflationary pressures. However there is a striking change in mean from September 1992 onwards, which is substantially different than the mean over the previous period.

The importance of techniques to handle non-stationarity of the type described here is more striking if one considers that the non-stationarities occur in the input rather than the output variables. They are not always visible to the naked eye and very often involve non-stationarities in co-variance as policy makers change the ways in which they utilise these variables in order to influence economic development.

5. Concluding remarks

We described a Discounted Least Squares procedure for backpropagation networks designed to deal with the problems of weak non-stationarity in financial data series. We evaluated the procedure in a controlled simulation experiment and in a real application. The simulation results confirmed our expectations that DLS is a more efficient procedure for "weakly" non-stationary data series.

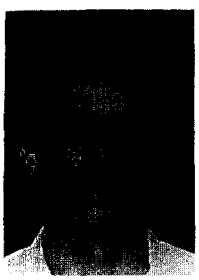
The results are particularly important for financial economics where it is generally believed that structural changes tend to occur slowly over time as the economic environment evolves and although short term trends may have diversionary effects many of the underlying economic laws still apply. Therefore models must be re-estimated regularly to make use of the most recent data but this must be done in a way that retains the integrity of the sample size. The choice of the parameters which control the decay rate was at first felt to be important. However our empirical evidence is that as long as these parameters are estimated conservatively, the danger of overweighing/underweighing recent observations is not critical.

Acknowledgements

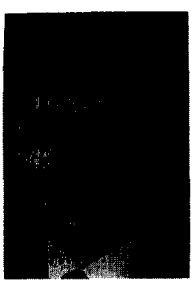
We would like to thank the anonymous referees for their constructive comments in finalising this paper. Also the members of the Neuroforecasting Club for their material and technical support.

References

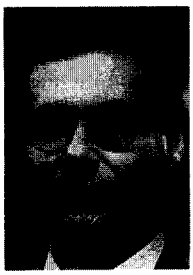
[1] A. Banerjee, J. Dolado, J.W. Galbraith and D.F. Hendry, *Co-integration, Error-Correction, and the Econometric Analysis of Non-stationary Data* (Oxford University Press, Oxford, 1993).
[2] W.A. Barnett, E.R. Berndt and H. White, *Dynamic Econometric Modeling* (Cambridge University Press, Cambridge, 1988).
[3] P.L. Bartlett, Learning with a slowly changing distribution, in: *Proc. Fifth ACM Confer. on Computational Learning Theory* (ACM Press, New York, 1992).
[4] C.R. Bonini and J.R. Freeland, Forecasting by smoothed regression, in: S. Makridakis and S.C. Wheelwright, eds., *Forecasting* (North-Holland, Amsterdam, 1979).
[5] P. Hackl and A. Weslund, *Economic Structural Change: Analysis and Forecasting* (Springer, Berlin, 1991).
[6] A. Harvey, *Time Series Models* (Harvester Wheatsheaf, Hemel Hempstead, 1993).
[7] K. Holden and D. Peel, Unbiasedness, efficiency and the combination of forecasts, *J. Forecasting* 8 (1989) 175–188.
[8] A.-P.N. Refenes, A.D. Zapranis and Y. Bentz, Modelling stock returns with neural networks, in: A.N. Refenes ed., *Neural Networks in the Capital Markets*, Proc. NnCM'93 London Business School (1993).
[9] A. Zellner, ed., *Bayesian Analysis in Econometrics and Statistics* (North-Holland, Amsterdam, 1980).



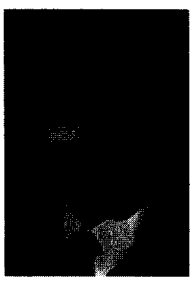
Apostolos-Paul Refenes (Bsc, PhD) is currently Associate Professor of Decision Science and Director of the NeuroForecasting Programme at the London Business School. He has held previous appointments at University College London, University of Athens and the DTI. Consultant to both government and private organisation is the Europe, US, and Japan. Author of over 70 papers and editor of two books on the subjects of neural computing and financial engineering applications. His current research interests include: neural networks, model selection/specification, hypothesis testing, and applications in financial engineering, portfolio management/asset allocation and derivative and term structure models.



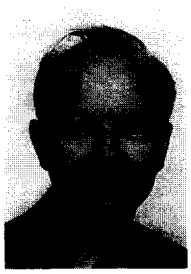
Neil Burgess is Research Fellow at the Neuroforecasting Unit, London Business School working on advanced decision technologies to problems in finance and marketing. Previously he worked for Thorn-EMI Central Research Laboratories where he played a key role in the European ESPRIT project, HANSA, integrating new technologies with existing software products such as spreadsheets and databases. He has been active in the field of neural networks for the past 7 years and has fielded live applications in the areas of both database marketing and financial trading, as well as lecturing and consulting widely in the US, South Africa, Europe, and the UK.



Achileas Zapranis (MSC) is Doctoral candidate at London Business School, Decision Technology Centre. His area of research is the development of diagnostics for misspecified neural network models, parameter estimation and prediction risk with applications to tactical asset allocation. He has published on neural networks, stock selection, and factor models for tactical asset allocation.



Yves Bentz (Msc) works for Société Générale, Asset Management, and is currently seconded to the Decision Technology Centre at London Business School. His current research investigates the use of non-linear techniques such as neural networks in the area of investment strategy, stock selection, portfolio management, asset allocation and risk management. This research is part of his Ph.D. Thesis.



Derek W. Bunn (MA Cambridge, 1971, MSc London, 1972, PhD London, 1975) is currently Professor and Chairman of the Decision Sciences subject area and Director of the Decision Technology Centre at the London Business School. He has held previous appointments at the universities of Oxford, Stanford and at IIASA. Author of over 100 papers and 6 books in the areas of forecasting, decision analysis and energy economics. Recipient of research awards from several public and private funds. Blackett Memorial Lecturer (1993) and Goodeve Medal Winner (1994) of the UK Operational Research Society. Editor of both the *Journal of Forecasting* and *Energy Economics*. Associate Editor of *Management Science*, *European Journal of Operational Research*, and 5 other journals. Elected international council member of The Institute of Management Sciences (1992–1994). Consultant to both government and private organisations. Keynote speaker at many international symposia.