Keynote talk presented at the International Symposium on Artificial Neural Networks, Hsinchu, Taiwan, December 1995.

Economic Forecasting: Challenges and Neural Network Solutions

John Moody
Computer Science Dept., Oregon Graduate Institute, PO Box 91000, Portland, OR 97291, USA
Email: moody@cse.ogi.edu FTP: neural.cse.ogi.edu; cd pub/neural/papers

Abstract

Macroeconomic forecasting is a very difficult task due to the lack of an accurate, convincing model of the economy. The most accurate models for economic forecasting, "black box" time series models, assume little about the structure of the economy. Constructing reliable time series models is challenging due to short data series, high noise levels, nonstationarities, and nonlinear effects. This paper describes these challenges and surveys some neural network solutions to them. Important issues include balancing the bias/variance tradeoff and the noise/nonstationarity tradeoff. The methods surveyed include hyperparameter selection (regularization parameter and training window length), input variable selection and pruning, network architecture selection and pruning, new smoothing regularizers, and committee forecasts. Empirical results are presented for forecasting the U.S. Index of Industrial Production. These demonstrate that, relative to conventional linear time series and regression methods, superior performance can be obtained using state-of-the-art neural network models.

1 Challenges of Macroeconomic Forecasting

Of great interest to forecasters of the economy is predicting the "business cycle", or the overall level of economic activity. The business cycle affects society as a whole by its fluctuations in economic quantities such as the unemployment rate (the misery index), corporate profits (which affect stock market prices), the demand for manufactured goods and new housing units, bankruptcy rates, investment in research and development, investment in capital equipment, savings rates, and so on. The business cycle also affects important socio-political factors such as the the general mood of the people and the outcomes of elections.

The standard measures of economic activity used by economists to track the business cycle include the Gross Domestic Product (GDP) and the Index of Industrial Production (IP). GDP is a broader measure of economic activity than is IP. However, GDP is computed by the U.S. Department of Commerce on only a quarterly basis, while Industrial Production is more timely, as it is computed and published monthly. IP exhibits stronger cycles than GDP, and is therefore more interesting and challenging to forecast. (See figure 1.) In this paper, all empirical results presented are for forecasting the U.S. Index of Industrial Production.

Macroeconomic modeling and forecasting is challenging for several reasons:

No *a priori* Models: A convincing and accurate scientific model of business cycle dynamics is not yet available due to the complexities of the economic system, the impossibility of doing controlled experiments on the economy, and non-quantifiable factors such as mass psychology and sociology that influence economic activity. There are two main approaches that economists have used to model the macroeconomy, econometric models and linear time series models:

Econometric Models: These models attempt to model the macroeconomy at a relatively fine scale and typically contain hundreds or thousands of equations and variables. The model structures are chosen by hand, but model parameters are estimated from the data. While econometric models are of some use in understanding the workings of the economy qualitatively, they are notoriously bad at making quantitative predictions.

Linear Time Series Models: Given the poor forecasting performance of econometric models, many economists have resorted to analyzing and forecasting economic activity by using the empirical "black box" techniques of standard linear time series

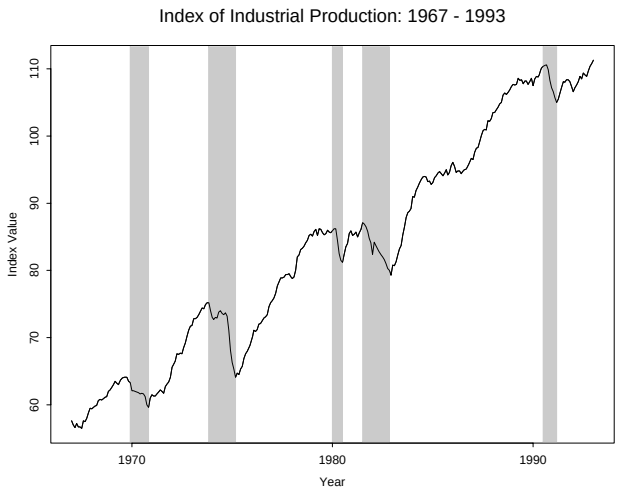

Figure 1: U.S. Index of Industrial Production (IP) for the period 1967 to 1993. Shaded regions denote official recessions, while unshaded regions denote official expansions. The boundaries for recessions and expansions are determined by the National Bureau of Economic Research based on several macroeconomic series. As is evident for IP, business cycles are irregular in magnitude, duration, and structure.

analysis. Such time series models typically have perhaps half a dozen to a dozen input series. The most reliable and popular of these models during the past decade or so have been bayesian vector autoregressive (BVAR) models (Litterman, 1986). As we have found in our own work, however, neural networks can often outperform standard linear time series models. The lack of an *a priori* model of the economy makes input variable selection, the selection of lag structures, and network model selection critical issues.

Noise: Macroeconomic time series are intrinsically very noisy and generally have poor signal to noise ratios. (See figures 2 and 3.) The noise is due both to the many unobserved variables in the economy and to the survey techniques used to collect data for those variables that are measured. The noise distributions are typically heavy tailed and include outliers. The combination of short data series and significant noise levels makes controlling model variance, model complexity, and the *bias / variance tradeoff* important issues (Geman, Bienenstock and Doursat, 1992). One measure of complexity for nonlinear models is P_{eff}, the *effective number of parameters* (Moody, 1992; Moody, 1994*b*). P_{eff} can be controlled to balance bias and variance by using regularization and model selection techniques.

Nonstationarity: Due to the evolution of the world's economies over time, macroeconomic series are intrinsically nonstationary. To confound matters, the definitions of many macroeconomic series are changed periodically as are the techniques employed in measuring them. Moreover, estimates of key series are periodically revised retroactively as better data are collected or definitions are changed. Not only do the underlying dynamics of the economy change with time, but the noise distributions for the measured series vary with time also. In many cases, such nonstationarity shortens the useable length of the data series, since training on older data will induce biases in predictions. The combination of noise and nonstationarity gives rise to a *noise / nonstationarity tradeoff* (Moody, 1994*a*), where using a short training window results in too much model variance or *estimation error* due to noise in limited training data, while using a long training window results in too much model bias or *approximation error* due to nonstationarity.

Nonlinearity: Traditional macroeconomic time series models are linear (Granger and Newbold, 1986; Hamilton, 1994). However, recent work by several investigators have suggested that nonlinearities can improve macroeconomic forecasting models in some cases (Granger and Terasvirta, 1993; Moody *et al.*, 1993; Natter, Haefke, Soni and Otruba, 1994; Swanson and White, 1995). (See table 1 and figures 2 and 3.) Based upon our own experience, the degree of nonlinearity captured by neural network models of macroeconomic series tends to be mild (Moody *et al.*, 1993; Levin, Leen and Moody, 1994; Rehfuss, 1994; Utans, Moody, Rehfuss and Siegelmann, 1995; Moody, Rehfuss and Saffell, 1996; Wu and Moody, 1996). Due to the high noise levels and limited data, simpler models are favored. This makes reliable estimation of nonlinearities more difficult.

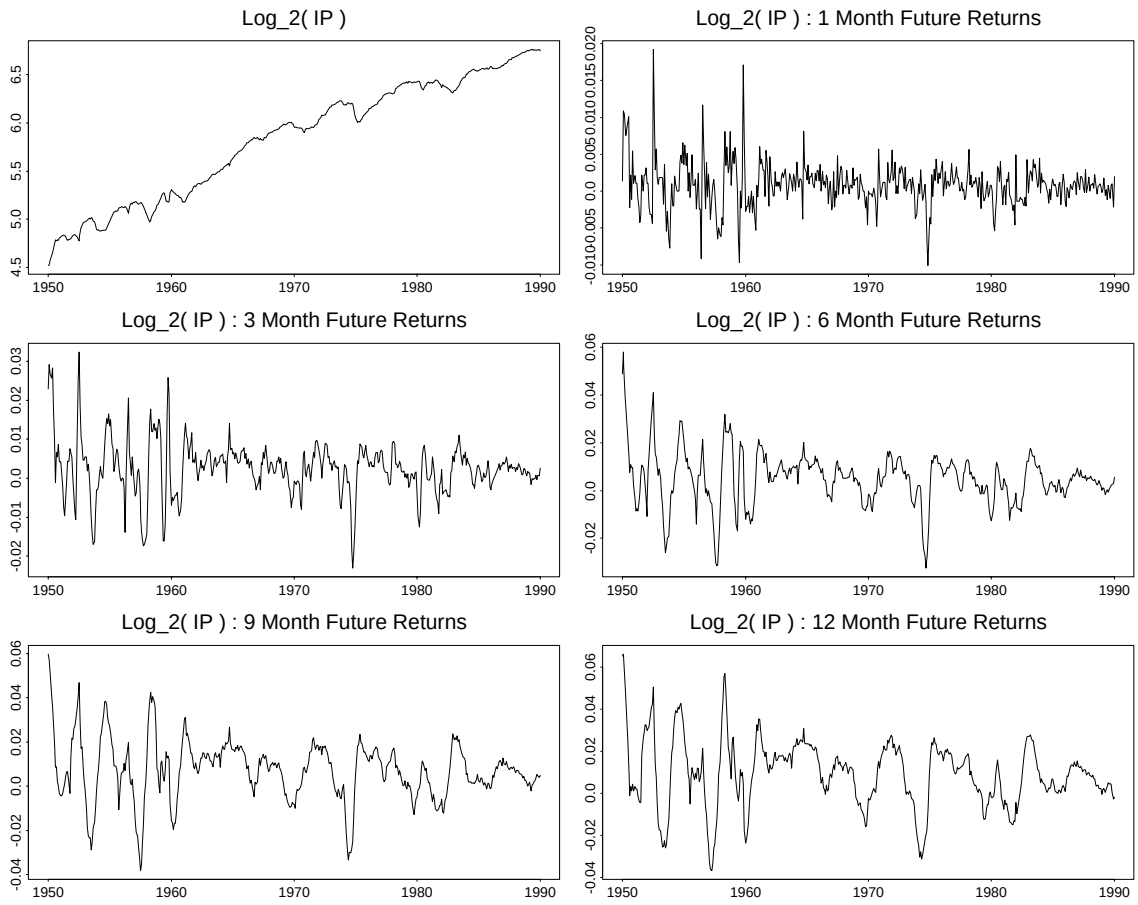

Figure 2: The U.S. Index of Industrial Production and five return series (rates of change measured as log differences) for time scales of 1, 3, 6, 9, and 12 months. These return series served as the prediction targets for the standard Jan 1950 - Dec 1979 / Jan 1980 - Jan 1990 benchmark results reported in Moody *et al.* (1993). The difficulty of the prediction task is evidenced by the poor signal to noise ratios and erratic behavior of the target series. For the one month returns, the performance of our neural network predictor in table 1 suggests that the SNR is around 0.2. For all returns series, significant nonstationarities and deviations from normality of the noise distributions are present.

2 Neural Network Solutions

We have been investigating a variety of algorithms for neural network model selection that go beyond the *vanilla* neural network approach.[1] The goal of this work is to construct models with minimal prediction risk (expected test set error). The techniques that we are developing and testing are described below. Given the brief nature of this survey, I have not attempted to provide an exhaustive list of the many relevant references in the literature.

Hyperparameter Selection: Hyperparameters are parameters that appear in the training objective function, but not in the network itself. Examples include the regularization parameter, the training window length, and robust scale parameters. Examples of varying the regularization parameter and the training window length for a 12 month IP forecasting model are shown in figure 4 a and b. Varying the regularization parameter trades off bias and variance, while varying the training window length trades off noise and nonstationarity.

[1] We define a *vanilla* neural network to be a fully connected, two-layer sigmoidal network with a full set of input variables and a fixed number of hidden units that is trained on a data window of fixed length with backprop and early stopping using a validation set. No variable selection, pruning, regularization, or committee techniques are used.

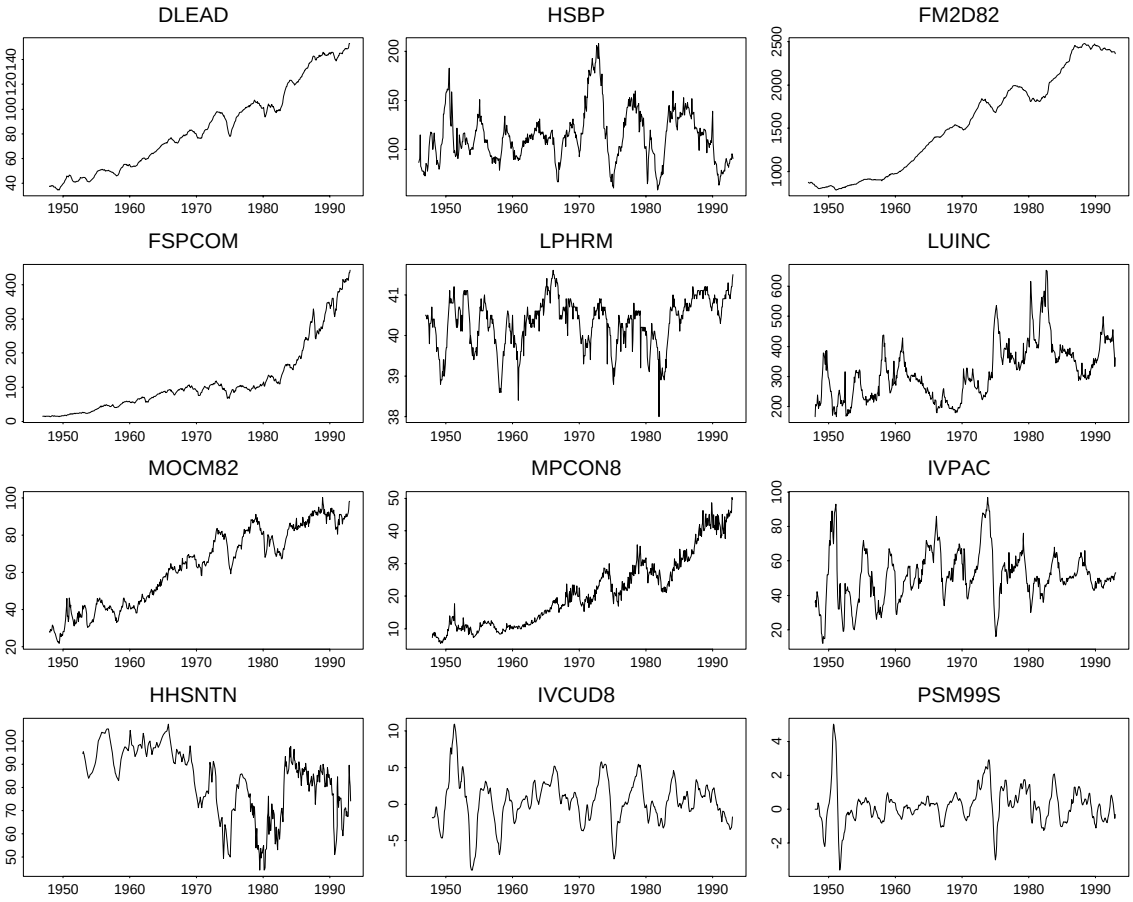

Figure 3: The U.S. Index of Leading Indicators (DLEAD) and its 11 component series as currently defined. The Leading Index is a key tool for forecasting business cycles. The input variables for the IP forecasting models reported in Moody *et al.* (1993) included transformed versions of DLEAD and several of its components. The difficulty of macroeconomic forecasting is again evident, due to the high noise levels and erratic behaviors of DLEAD and its components. (Note that the component series included in DLEAD have been changed several times during the past 47 years. The labels for the various series are those defined in Citibase: HSBP denotes housing starts, FM2D82 is M2 money supply, FSPCOM is the Standard & Poors 500 stock index, and so on.)

Input Variable Selection and Pruning: Selecting an informative set of input variables and an appropriate representation for them is critical to the solution of any forecasting problem. We have been studying the use of both model-independent and model-dependent variable selection procedures. The Delta Test, a model independent procedure, is a nonparametric statistical algorithm that selects meaningful predictor variables by direct examination of the data set (Pi and Peterson, 1994). We are developing some refinements to this approach. Sensitivity-based pruning (SBP) techniques are model-dependent algorithms that prune unnecessary or harmful input variables from a trained network (Mozer and Smolensky, 1990; Moody and Utans, 1994; Moody, 1994*b*; Utans *et al.*, 1995). An example of reducing a set of input variables from 48 to 13 for a 12 month IP forecasting model is shown in figure 5.

Model Selection and Pruning: A key technique for controlling the bias / variance tradeoff for noisy problems is to select the size and topology of the network. For two-layer networks, this includes selecting the number of internal units, choosing a connectivity structure, and pruning unneeded nodes, weights, or weight matrix eigennodes. A constructive algorithm for selecting the number of internal units is sequential network construction (SNC) (Ash, 1989; Moody and Utans, 1994; Moody, 1994*b*). Techniques for pruning weights and internal nodes include sensitivity-based pruning methods like optimal brain damage (OBD) (LeCun, Denker and Solla, 1990) and optimal brain surgeon (OBS) (Hassibi and Stork, 1993). Our recently-proposed supervised principal components pruning (PCP) method (Levin *et al.*, 1994) prunes weight matrix eigennodes, rather than weights. Since PCP does

Prediction Horizon (Months)	Trivial (Average of Training Set)	Univariate AR(14) Model Iterated Pred.	Multivariate Linear Reg. Direct Pred.	Sigmoidal Nets w/ PC Pruning Direct Pred.
1	1.04	0.90	0.87	0.81
2	1.07	0.97	0.85	0.77
3	1.09	1.07	0.96	0.75
6	1.10	1.07	1.38	0.73
9	1.10	0.96	1.38	0.67
12	1.12	1.23	1.20	0.64

Table 1: Comparative summary of normalized prediction errors for rates of return on Industrial Production for the period January 1980 to January 1990 as presented in Moody *et al.* (1993). The four model types were trained on data from January 1950 to December 1979. The neural network models significantly outperform the trivial predictors and linear models. For each forecast horizon, the normalization factor is the variance of the target variable for the training period. Nonstationarity in the IP series makes the test errors for the trivial predictors larger than 1.0. In subsequent work, we have obtained substantially better results for the IP problem (Levin *et al.*, 1994; Rehfuss, 1994; Utans *et al.*, 1995; Moody *et al.*, 1996; Wu and Moody, 1996).

not require training to a local minimum, it can be used with early stopping. It has computational advantages over OBS, and can outperform OBD when input variables or hidden node activities are noisy and correlated. Figure 6 shows reductions in prediction errors obtained by using PCP on a set of IP forecasting models.

Better Regularizers: Introducing biases in a model via regularization or pruning reduces model variance and can thus reduce prediction risk. Prediction risk can be best minimized by choosing appropriate biases. Quadratic weight decay (Plaut, Nowlan and Hinton, 1986; Hoerl and Kennard, 1970*b*; Hoerl and Kennard, 1970*a*), the standard approach to regularization used in the neural nets community, is an *ad hoc* function of the network weights. Weight decay is *ad hoc* in the sense that it imposes direct constraints on the weights independent of the nature of the function being learned or the parametrization of the network model. A more principled approach is to require that the function $f(W,x)$ learned by the network be smooth. This can be accomplished by penalizing the m^{th} order curvature of $f(W,x)$. The regularization or penalty functional is then the smoothing integral

$$S(W,m) = \int d^D x \Omega(x) \left\| \frac{d^m f(W,x)}{dx^m} \right\|^2 \quad , \tag{1}$$

where $\Omega(x)$ is a weighting function and $\| \|$ denotes the Euclidean tensor norm. Since numerical computation of (1) generally requires expensive Monte Carlo integrations and is therefore impractical during training, we have derived algebraically-simple approximations and bounds to $S(W,m)$ for feed forward networks that can be easily evaluated at each training step (Moody and Rögnvaldsson, 1995). Our empirical experience shows that these new smoothing regularizers typically yield better prediction accuracies than standard weight decay. In related work, we have derived an algebraically-simple regularizer for recurrent nets that corresponds to the case $m = 1$ (Wu and Moody, 1996). A comparison of this recurrent regularizer to quadratic weight decay for 1 month forecasts of IP is shown in figure 7.

Committee Forecasts: Due to the extremely noisy nature of economic time series, the control of forecast variance is a critical issue. One approach for reducing forecast variance is to average the forecasts of a committee of models. Researchers in economics have studied and used combined estimators for a long time, and generally find that they outperform their component estimators and that unweighted averages tend to outperform weighted averages, for a variety of weighting methods (Granger and Newbold, 1986; Winkler and Makridakis, 1983; Clemen, 1989). We are exploring several extensions of this approach. Reductions of prediction error variances obtained by unweighted committee averaging for a selection of different IP forecasting models are shown in figure 8.

3 Discussion

In concluding this brief survey of the algorithms that we are developing and testing for improving forecast accuracy with neural networks, it is important to note that many other techniques have been proposed. Also, the empirical results presented herein are preliminary, and further work on both the algorithms and forecasting models is required. As a final comment, I would like to

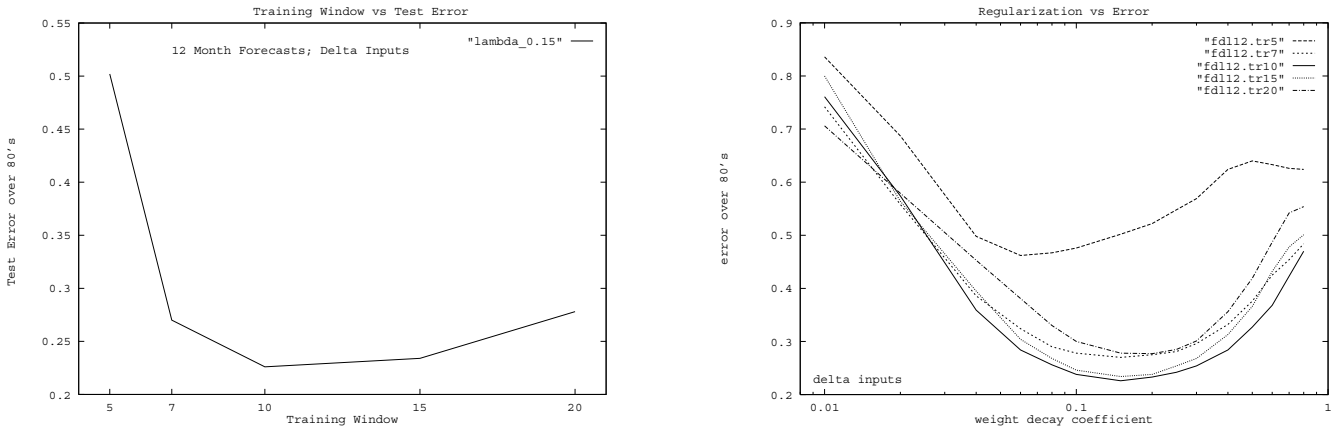

Figure 4: Left: Example of the *Noise / Nonstationary Tradeoff* and selection of the best training window, in this case 10 years (Rehfuss, 1994; Moody *et al.*, 1996). The longer training windows of 15 and 20 years yield higher test set error due to the model bias induced by nonstationarity. The shorter training windows of 5 and 7 years have significantly higher errors due to model variance resulting from noise in the data series and smaller data sets. The test errors correspond to models trained with the best regularization parameter 0.15 indicated in the figure on the right. Right: Example of the effect of regularization (weight decay) parameter on test error (Rehfuss, 1994; Moody *et al.*, 1996). The five curves are for training windows of length 5, 7, 10, 15, and 20 years. The *Bias / Variance Tradeoff* is clearly evident in all the curves; the minimum test set errors occur for weight decay parameters of order 0.1. Larger errors due to bias occur for larger weight decay coefficients, while larger errors due to model variance occur for smaller values of the coefficient.

emphasize that given the difficulty of macroeconomic forecasting, no single technique for reducing prediction risk is sufficient for obtaining optimal performance. Rather, a combination of techniques is required.

Acknowledgements

The author wishes to thank Todd Leen, Asriel Levin, Yuansong Liao, Hong Pi, Steve Rehfuss, Denni Rögnvaldsson, Matthew Saffell, Joachim Utans, and Lizhong Wu for their many contributions to this research. This work was supported at OGI by ONR/ARPA grants N00014-92-J-4062 and N00014-94-1-0071, NSF grants CDA-9309728 and CDA-9503968, and at Nonlinear Prediction Systems by ARPA contract DAAH01-92-CR361.

References

Ash, T. (1989), 'Dynamic node creation in backpropagation neural networks', *Connection Science* **1**(4), 365–375.

Clemen, R. T. (1989), 'Combining forecasts: A review and annotated bibliography', *International Journal of Forecasting* (5), 559–583.

Geman, S., Bienenstock, E. and Doursat, R. (1992), 'Neural networks and the bias/variance dilemma', *Neural Computation* **4**(1), 1–58.

Granger, C. W. J. and Newbold, P. (1986), *Forecasting Economic Time Series*, 2nd edn, Academic Press, San Diego, California.

Granger, C. W. J. and Terasvirta, T. (1993), *Modelling Nonlinear Economic Relationships*, Oxford University Press.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton University Press.

Hassibi, B. and Stork, D. G. (1993), Second order derivatives for network pruning: Optimal brain surgeon, *in* S. J. Hanson, J. D. Cowan and C. L. Giles, eds, 'Advances in Neural Information Processing Systems 5', Morgan Kaufmann Publishers, San Mateo, CA, pp. 164–171.

Hoerl, A. and Kennard, R. (1970*a*), 'Ridge regression: applications to nonorthogonal problems', *Technometrics* **12**, 69–82.

Hoerl, A. and Kennard, R. (1970*b*), 'Ridge regression: biased estimation for nonorthogonal problems', *Technometrics* **12**, 55–67.

LeCun, Y., Denker, J. S. and Solla, S. A. (1990), Optimal brain damage, *in* D. S. Touretzky, ed., 'Advances in Neural Information Processing Systems 2', Morgan Kaufmann Publishers.

Levin, A. U., Leen, T. K. and Moody, J. E. (1994), Fast pruning using principal components, *in* J. Cowan, G. Tesauro and J. Alspector, eds, 'Advances in Neural Information Processing Systems 6', Morgan Kaufmann Publishers, San Francisco, CA.

Litterman, R. B. (1986), 'Forecasting with Bayesian vector autoregressions – five years of experience', *Journal of Business and Economic Statistics* **4**(1), 25–38.

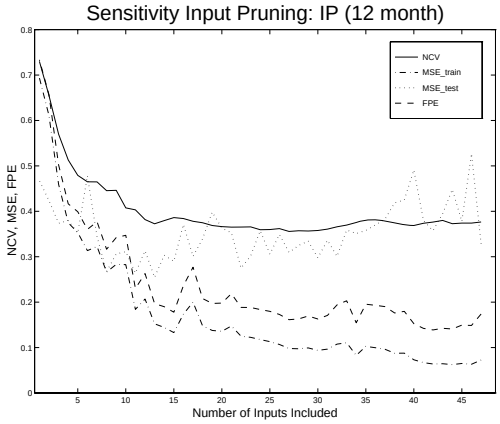

Figure 5: Sensitivity-Based Pruning (SBP) method for selecting a subset of input variables for a neural net forecasting model (Utans *et al.*, 1995). The original network was trained on all 48 input variables to predict the 12 month percentage changes in Industrial Production (IP). The variables have been ranked in order of decreasing importance according to a sensitivity measure. The input variables are pruned one-by-one from the network; at each stage, the network is retrained. The figure shows four curves: the Training Error, Akaike Final Prediction Error (FPE), Nonlinear Cross-Validation Error (NCV) proposed by Moody and Utans (1994) (see also Moody (1994*b*)), and the actual Test Error. NCV is used as a selection criterion and suggests that only 13 of the variables should be included. NCV predicts the actual test error quite well relative to FPE.

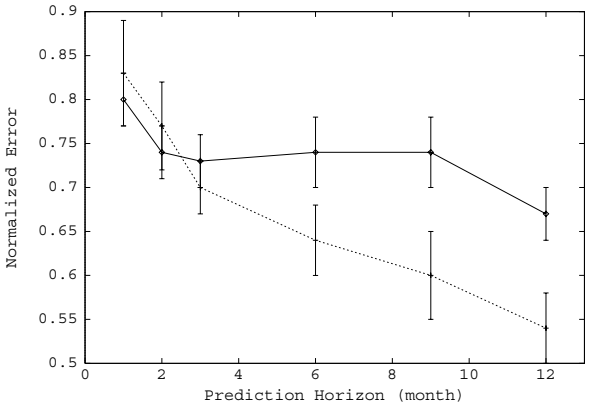

Figure 6: Prediction errors for two sets of neural network models for 12 month returns for IP, with (dotted line) and without (solid line) Supervised Principal Components Pruning (PCP) proposed by Levin *et al.* (1994). Each data point is the mean error for 11 nets, while the error bars represent one standard deviation. Statistically significant improvements in prediction performance are obtained for the 6, 9, and 12 month prediction horizons by using the PCP algorithm to reduce the network complexities. While techniques like optimal brain damage and optimal brain surgeon prune weights from the network, PCP reduces network complexity and hence model variance by pruning eigennodes of the weight matrices. Unlike the unsupervised use of principal components, PCP removes those eigennodes that yield the greatest reduction in estimated prediction error.

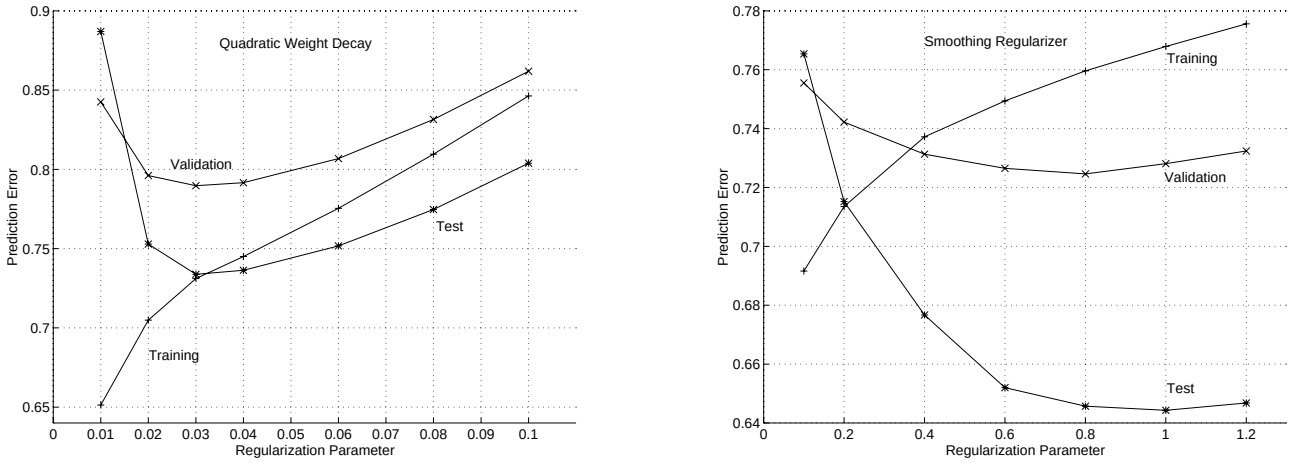

Figure 7: Regularization parameter vs. normalized prediction errors for the task of predicting the one month rates of change of the U.S. Index of Industrial Production (Wu and Moody, 1996). The example given is for a recurrent network trained with standard weight decay (left) or with the new recurrent smoothing regularizer (right). For standard weight decay, the optimal regularization parameter is 0.03 corresponding to a test error of 0.734. For the new smoothing regularizer, the optimal regularization parameter which leads to the least validation error is 0.8 corresponding to a test error of 0.646. The new recurrent regularizer thus yields a 12% reduction in test error relative to that obtained using quadratic weight decay.

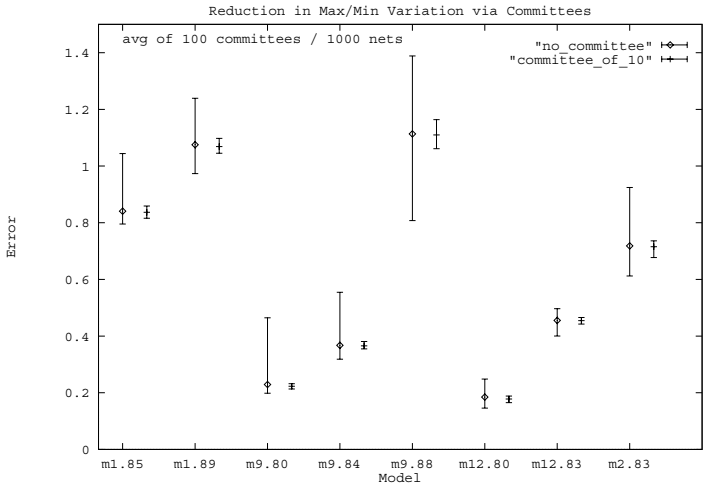

Figure 8: Reduction in error variance for prediction of the U.S.Index of Industrial Production by use of combining forecasts (or committees) (Rehfuss, 1994; Moody *et al.*, 1996). Abscissa points are various combinations of prediction horizon and test period. For example, "m12.80" denotes networks trained to make 12 month forecasts on the ten years prior to 1979 and tested by making true *ex ante* forecasts on the year 1980. Performance metric is normalized mean square error (NMSE) computed over the particular year. All training sets have length 10 years. For each point, bars show range of values for either 1000 individual models, or 100 committees of 10. The individual networks each have three sigmoidal internal units, one linear output, and typically a dozen or so input variables selected by the δ-test from an initial set of 48 candidate variables.

Moody, J. (1992), The effective number of parameters: an analysis of generalization and regularization in nonlinear learning systems, *in* J. E. Moody, S. J. Hanson and R. P. Lippmann, eds, 'Advances in Neural Information Processing Systems 4', Morgan Kaufmann Publishers, San Mateo, CA, pp. 847–854.

Moody, J. (1994*a*), Challenges of Economic Forecasting: Noise, Nonstationarity, and Nonlinearity, Invited talk presented at Machines that Learn, Snowbird Utah, April 1994.

Moody, J. (1994*b*), Prediction risk and neural network architecture selection, *in* V. Cherkassky, J. Friedman and H. Wechsler, eds, 'From Statistics to Neural Networks: Theory and Pattern Recognition Applications', Springer-Verlag.

Moody, J. and Rögnvaldsson, T. (1995), Smoothing regularizers for feed-forward neural networks, Manuscript in preparation.

Moody, J. and Utans, J. (1994), Architecture selection strategies for neural networks: Application to corporate bond rating prediction, *in* A. N. Refenes, ed., 'Neural Networks in the Captial Markets', John Wiley & Sons.

Moody, J., Levin, A. and Rehfuss, S. (1993), 'Predicting the U.S. index of industrial production', *Neural Network World* **3**(6), 791–794. Special Issue: Proceedings of Parallel Applications in Statistics and Economics '93.

Moody, J., Rehfuss, S. and Saffell, M. (1996), Macroeconomic forecasting with neural networks, Manuscript in preparation.

Mozer, M. C. and Smolensky, P. (1990), Skeletonization: A technique for trimming the fat from a network via relevance assessment, *in* D. S. Touretzky, ed., 'Advances in Neural Information Processing Systems 1', Morgan Kaufmann Publishers, San Mateo, CA.

Natter, M., Haefke, C., Soni, T. and Otruba, H. (1994), Macroeconomic forecasting using neural networks, *in* 'Neural Networks in the Capital Markets 1994'.

Pi, H. and Peterson, C. (1994), 'Finding the embedding dimension and variable dependencies in time series', *Neural Computation* pp. 509–520.

Plaut, D., Nowlan, S. and Hinton, G. (1986), Experiments on learning by back propagation, Technical Report CMU-CS-86-126, Dept. of Computer Science, Carnegie-Mellon University, Pittsburgh, Pennsylvania.

Rehfuss, S. (1994), Macroeconomic forecasting with neural networks, Unpublished simulations.

Swanson, N. and White, H. (1995), A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks, Discussion paper, Department of Economics, Pennsylvania State University.

Utans, J., Moody, J., Rehfuss, S. and Siegelmann, H. (1995), Selecting input variables via sensitivity analysis: Application to predicting the U.S. business cycle, *in* 'Proceedings of Computational Intelligence in Financial Engineering'.

Winkler, R. L. and Makridakis, S. (1983), 'The combination of forecasts', *Journal of Royal Statistical Society* (146).

Wu, L. and Moody, J. (1996), 'A smoothing regularizer for feedforward and recurrent networks', *Neural Computation* **8**(2).