Statistical Model Building for Neural Networks

Ulrich Anders

Abstract

Neural networks are a new, very flexible class of statistical and - if applied to economic data - econometric models. Basically, neural networks are a generalization of nonlinear regression models and can therefore be applied to all kinds of regression problems. Since neural networks do not require the specification of a certain structural form, they are particularly suited for modelling very complex functions as observed on the capital markets.

Since neural networks are statistical procedures, they offer the opportunity to perform a thorough statistical analysis in order to build an adequate model. However, this is not the status-quo. Most network practitioners pursue a rather heuristic approach for determining a network architecture which is prone to finding only suboptimal network models. It is the aim of this article to propose a structurized process for modelling neural network architectures, which relies on statistical methods.

Résumé

Les réseaux neuronaux forment une classe nouvelle et flexible de modèles statistiques et économétriques. Fondamentalement, les réseaux neuronaux constituent une généralisation des modèles de régression non linéaire et peuvent s'appliquer à tout type de problème de régression. Comme les réseaux neuronaux ne requièrent la spécification d'une certaine forme structurelle, ils conviennent particulièrement à la modélisation de relations complexes telles que celles observées sur les marchés financiers.

Les réseaux neuronaux étant procédure statistique, il convient d'utiliser cette opportunité lors de leur modélisation. Toutefois aucun consensus n'émerge dans ce domaine. La plupart des praticiens des réseaux neuronaux adoptent une démarche plutôt heuristique pour déterminer l'architecture d'un réseau, ce qui conduit au choix d'un modèle non optimal. Le but de cet article est de proposer une procédure structurée pour modéliser l'architecture des réseaux neuronaux tout en se basant sur des méthodes statistiques.

Keywords

Neural networks, model building, regression problems.

Mots clefs

Réseaux neuronaux, modélisation, problème de régression.

Centre for European Economic Research (ZEW), Kaiserring 14-16, D-68161 Mannheim (Germany); Tel: + 49-621-1235-141, Fax: + 49-621-1235-223, E-mail: anders@zew.de

1 Introduction

Few statistical procedures have received as much attention as neural networks have of late. Neural networks show their potential in many different areas, but they are particularly well suited to very complex applications. A good example for these is the explanation or the forecasting of financial data, like stock prices, exchange rates or option prices. Even in such difficult tasks neural networks achieved reasonably good results.[1]

However, in many practical applications the possibilities of neural networks are not fully employed. Neural networks and especially the multilayer perceptron (MLP) networks are nothing other than a very flexible class of statistical regression models. As such they can be thoroughly analyzed by the usual statistical procedures. These procedures can help to build adequate network models and to recognize insignificant variables as well as data problems.

In this article, the most important statistical procedures applicable to neural networks are summarized. We will only consider the MLP-type of neural networks as shown in figure 1. Although other types of networks may be equally well suited to regression analysis,[2] only MLP-networks allow for the application of statistical inference. The application of statistical methods for neural networks will be termed 'neurometrics'.

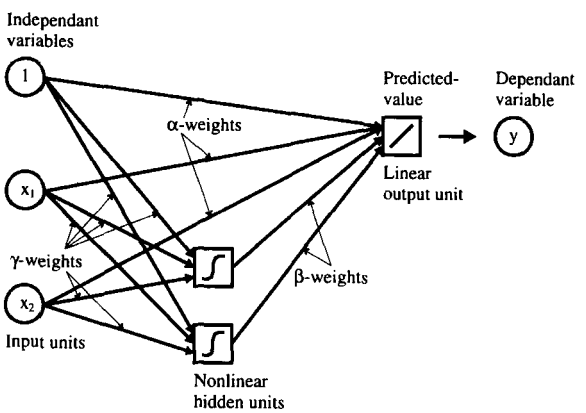

Figure 1: MLP-network with three layers.

Each neural network can be expressed as a function of the explaining variables $X = [x_0, x_1, \ldots, x_I]$, and the network weights $w = (\alpha', \beta', \gamma')'$. x_0 is constant and defined

as $x_0 \equiv 1$ for all observations. I denotes the number of non-constant explaining variables and H the number of hidden units employed. The network shown in figure 1 has the following functional form $f(X, w)$:

$$f(X, w) = X\alpha + \sum_{h=1}^{H} \beta_h g\left(\sum_{i=0}^{I} \gamma_{hi} x_i\right) \tag{1}$$

The function $g(\cdot)$ is the so-called transfer function of a neuron and is usually chosen to be the logistic or tangens hyperbolicus function.

2 Statistical neural networks

Neural networks are able to solve the nonlinear regression problem

$$y = F(X, w) + \varepsilon, \tag{2}$$

where F is an unknown relationship and ε an iid error term with $E[\varepsilon\varepsilon'] = \sigma I$, $E[\varepsilon] = 0$ and $E[\varepsilon|X] = 0$. For neural networks are able to approximate any function — and especially the unknown function F — up to an arbitrary degree of accuracy given the architecture of the network is sufficiently complex.[3] Therefore, neural networks are nothing else than nonlinear regression models. The advantage of neural networks is that they do not require an explicit assumption about the structural form of the unknown function, since the network extracts all information about the unknown function purely from the data. Nevertheless, even with neural networks one assumption is made: it is presupposed, that the chosen network architecture is complex enough to actually approximate the unknown function up to a satisfactory degree.

For this reason, the question is which network architecture is optimal for a given task. The selection of an adequate architecture is one of the most difficult problems when dealing with neural networks. If the function to be approximated is very irregular, the network must be much more complex than if the function is rather smooth. If the dimension of a network is too low, the unknown function can not be approximated. Forecasts of such a network will always have a systematic error (bias). If the dimension of a network is too high, the network will overfit the function and

forecasts will show high variance. In addition to this, each parameter in a network has a random element. The more parameters which are estimated in a network the more insecurity is introduced into the system. Therefore, one should only accept parameters that are not redundant, since each redundant parameter unneccessarily increases the variabilitiy of the forecast.

Thus, the aim of neural network model building should be to select an architecture that is just big enough to capture the unknown function. The forecasts of such a network are unbiased and have the least possible variance. The search for an adequate network model is an iterativ process which is described in section 3.

2.1 Nonlinear least squares

In most applications of neural nets the weights are determined such that they minimize the sum of squared errors (SSE) of the regression:[4]

$$\text{SSE}(w) = [y - f(X, w)]'[y - f(X, w)] \rightarrow \text{Min!} \tag{3}$$

Together with the assumptions of the nonlinear regression model (2) and under some regularity conditions for f — in particular, the parameters of the model must be uniquely determinable — it can be proven (White, 1989a) that the parameter estimator $\hat{w}$ is consistent with an asymptotic normal distribution. In principle, this allows for the application of the usual hypotheses tests for nonlinear models such as the Wald-test or the LM-test.

Since a neural network does not normally map the unknown function exactly but only approximates it, neural networks belong to the class of misspecified models. The corresponding theory has been developed by White (1994). He proved that the application of asymptotic standard tests is still valid if the misspecification is taken account of when calculating the covariance matrix of the parameters. The estimated parameters $\hat{w}$ are normally distributed with mean w^* and covariance matrix $\frac{1}{T}C$. The parameter vector w^* corresponds to the best projection of the misspecified model onto the true model. This leads to:

$$\sqrt{T}(\hat{w} - w^*) \sim N(0, C), \tag{4}$$

where T is the number of observations. According to the theory of misspecified models the covariance matrix is calculated by $\frac{1}{T}C = \frac{1}{T}A^{-1}BA^{-1}$. The matrices A and B are defined as $A \equiv E[\nabla^2 \mathrm{SE}_t]$ and $B \equiv E[\nabla \mathrm{SE}_t \nabla \mathrm{SE}_t']$. SE_t denotes the squared error contribution of the t-th observation and ∇ the gradient with respect to the weights.

In general, the parameters of a nonlinear regression problem cannot be determined analytically and the application of numerical procedures is neccessary. The search for optima in nonlinear functions is a standard problem in numerical mathematics. For this reason, a wide variety of efficient algorithms exist (for instance BFGS, DFP, Levenberg-Marquardt, etc.).[5] The backpropagation algorithm which is often used for estimating neural networks is due to its simplicity inferior to those algorithms.[6]

3 Neurometric model building

The search for a good network architecture is similar to the approach usually employed in econometrics or statistics. The search is an iterativ process as shown in figure 2. It is continued as long as the resulting network solves the given task in a satisfactory way.

In the literature of neural networks relatively little attention has been given to the specification of neurals networks. The result of this oversight is, that a large number of investigations rely on a hopelessly overparameterized model, or too many input variables.[7] In principle the number of observations should be larger than the number of free parameters in the model. Reliable insight into a model can only be achieved if the number of observations exceeds the number of model parameters by about a factor of ten (White, 1989a).

The selection of the input variables for a neural network runs analogous to the selection of variables when building statistical or econometric models. The first step for the modeller should be to investigate the quality of the data and to correct data problems. In addition to this, he should be aware of the degree of integration as well as potential cointegrating relationships between the variables.

A good starting point for the construction of network architecture is to select a linear regression model. The analysis of the linear model may reveal difficulties such as heteroscedasticity of autocorrelation. Furthermore, one may recognize model instabilities or structural breaks. If the linear model neglects nonlinearities the application

of a neural network is justified.[8] The linear regression model may then be nested into the neural network as it is done in figure 1.

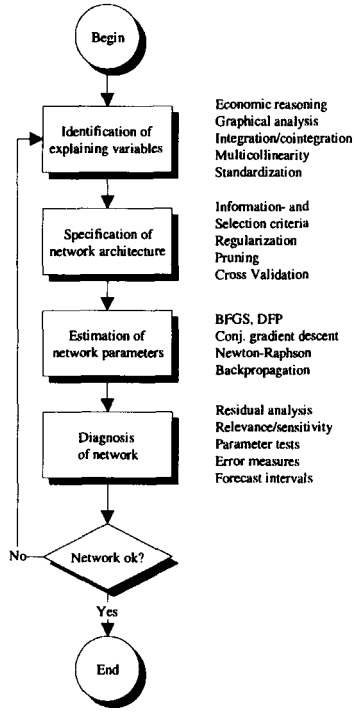

Figure 2: Neurometric model building.

For selecting an adequate network architecture two types of approaches exist: the heuristic procedures of the network literature (e.g. regularization or pruning) and statistical procedures (e.g. hypotheses tests). The following sections portray the most important statistical procedures applicable to neural nets.

3.1 Tests for parameter significance

An important instrument for the diagnosis of statistical and econometric models are hypotheses tests for the significance of parameters. The computation of the so-called t-values has become standard in linear models, whereas it is overlooked in neural networks. The computation of the significance of the model parameters is of great

importance, for non-significant parameters can be removed from the model. This is valid for linear models as well as for neural networks.[9] The t-values are defined as

$$\frac{\hat{w}_k - r_k}{\hat{\sigma}_{\hat{w}_k}}, \tag{5}$$

where r_k denotes the restriction to be tested for, usually zero. The estimated standard deviation $\hat{\sigma}_{\hat{w}_k}$ of the weight $\hat{w}_k$ stands on the main diagonal of the covariance matrix $\hat{C}$. The latter can consistently be estimated by help of the matrices $\hat{A}$ and $\hat{B}$:

$$\hat{C} = \frac{1}{T}\hat{A}^{-1}\hat{B}\hat{A}^{-1}. \tag{6}$$

In case of the least squares method the matrices $\hat{A}$ und $\hat{B}$ can be computed as follows:

$$\hat{A} = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 \mathrm{SE}_t}{\partial\hat{w}\partial w'} \text{ und } \hat{B} = \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_t^2\left(\frac{\partial f(X_t,\hat{w})}{\partial\hat{w}}\right)\left(\frac{\partial f(X_t,\hat{w})}{\partial\hat{w}}\right)' \tag{7}$$

Since it is known from equation (4) that the parameters of the network are asymptotically normally distributed, it is possible to test for the significance of each parameter using the estimated covariance matrix $\hat{C}$. According to the theory of misspecified models both a Wald- and an LM-Test are applicable.

Before the application of the parameter tests it must be ensured that the parameters — apart from symmetric solutions — are uniquely identified. This is not the case when a network contains irrelevant hidden units. Since the β-weight of an irrelevant hidden unit is zero the γ-weights which lead into this hidden unit can take any value and are thus not identified. The problem arises from the fact that the distribution of non identified parameters is no longer normal but belongs to a more general class of mixed Gaussian distributions (Phillips, 1989). Therefore, parameter inference which relies on normal theory is not admissible any more.

In order to make standard inference applicable, all irrelevant hidden units have to be removed from the network in question. However, the identification of irrelevant hidden units is very difficult. For this reason, Anders/Korn (1996) suggest to include an additional hidden unit only when it significantly contributes to the explanation of

the endogenous variable. The significance of the additional hidden unit can be tested for by LM–test procedures as they were proposed by White (1989b) or Teräsvirta (1993).

Under the assumption that the network does not contain any irrelevant hidden units, one can test for arbitrary parameter restrictions on the γ–weights. The hypotheses $H_0 : R\hat{w} = 0$ of an irrelevant input unit deserves particular attention. R selects those γ–weights which are linked to the input unit in question.

For testing such hypotheses it is easiest to use a Wald–test. This tests allows for analysing the unrestricted model. A neural network must therefore be trained only once. The test statistic is given by:

$$(R\hat{w})'(R\hat{C}R')^{-1}(R\hat{w}) \sim \chi^2_q, \tag{8}$$

where q denotes the number of restricted parameters.

3.2 Information- and selection criteria

Alternative instruments for model selection are the so-called information criteria. They usually relate the squared residuals to the number of free model parameters. The intention of this is to weigh the error of the model against the number of its parameters. An additional parameters should only be introduced into the model if the value of the applied information criterion decreases. Amongst several competing model the model which is selected has the smallest value of the information criterion.

The best know information criterion is the Akaike (1973) information criterion (AIC)

$$\text{AIC} = \ln\left(\frac{\varepsilon'\varepsilon}{T}\right) + \frac{2K}{T}. \tag{9}$$

The error term of the regression is denoted with ε, K is the number of free model parameters and T corresponds to the number of observations. However, this criterion is not valid for misspecified models. Therefore, Murata/Yoshizawa/Amari (1994) developed a criterion — the network information criterion (NIC) — suited for neural networks.

The NIC is a generalization of the AIC should the chosen model not encompass the true function F.[10] When the least squares method (3) is applied the NIC is

computed from the mean squared error of the regression, and a penalty term for the number of the parameters:

$$\text{NIC} = \text{MSE} + \frac{1}{T} \cdot \text{tr}[BA^{-1}], \tag{10}$$

The matrices A and B have been defined in section 2.1. Should the network be able to map the true function F exactly, which means the network is not misspecified, the asymptotic relationsship $B = 2\sigma^2 A$ can be derived. This leads to $\text{tr}[BA^{-1}] = 2\sigma^2 \,\text{tr}[I] = 2\sigma^2 K$ and the NIC reduces to the AIC formula given by Amemiya (1980):

$$\text{AIC} = \text{MSE} + \sigma^2 \frac{2K}{T} \tag{11}$$

Unfortunately, the application of information criteria is theoretically valid only if the neural network does not contain irrelevant hidden units. The NIC can thus only help to decide which γ-weights should remain in a network model.

Nevertheless, the NIC gives an indication as to whether or not the neural network contains irrelevant hidden units: if it does, the NIC takes ernormously high values. This is due to the fact, that overparameterized models own badly behaved covariance matrices. The matrices $\hat{A}$ and $\hat{B}$ strongly deviate from each other and thus lead to the high values of the NIC.

Apart from the information criteria there exist many alternative criteria which are employed for model selection (for instance, FPE, PC, $\bar{R}^2$). Best known is the coefficient of determination R^2. The coefficient of determination is defined as:

$$R^2 = 1 - \frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{T}(y_t - \bar{y}_t)^2} \tag{12}$$

It measures the goodness of fit of the regression: the closer the value of the coefficient of determination to one, the better the fit; the closer the value to zero, the worse the fit.[11]

3.3 Cross Validation

A procedure that is often used for neural networks and nonparametric models is cross validation (CV) or more specific v-leave-out cross validation.[12] By help of this

procedure one tries to forecast the expected error for unknown data. Between several competing models, the model with the smallest prediction error is selected. For estimating this error, the observation set is divided into M seperate subsets with v observations each (v is often chosen as $v = 1$). Now each of the subsets is taken successively as a validation set, and the parameters of the model are estimated without using the observations of this set. The mean squared prediction errors MSPE_m on the different validation sets are averaged and the result serves as an estimator for the expected prediction error. Thus, the cross validation error is computed as

$$\mathrm{CV} = \frac{1}{M} \sum_{m=1}^{M} \mathrm{MSPE}_m \,. \tag{13}$$

In opposition to the information criteria the cross validation procedure is extremly cumbersome, since the estimation of the parameters has to be repeated M times. The procedure of cross validation must not be confused with stopped training. In stopped training the observation set is split into two distinct subsets, a training and a validation set. In the subsequent iterative training process the squared errors are simultaneously determined with respect to both the training and the validation set. The iteration is stopped when the sum of squared errors with respect to the validation set achieves a minimum.

The intention which lies behind stopped training is to break off the training when the network begins to overfit. Although this seems to be a good idea on first sight, stopped training has three serious drawbacks. First, a network that is able to overfit is obviously overparameterized. Consequently the network would need fewer units than actually used. Instead of stopping the training at a suboptimal point it is therefore more reasonable to reduce the number of free parameters. Second, the parameter estimates are very sensitive to the choice of the training and validation set, especially if these sets have been selected by help of a random mechanism. The paramater estimates are thus quite unreliable. Third, when the network training is stopped at a suboptimal point, that means the gradient is not zero, statistical inference is no longer possible. In particular, the hessian will not be positive definite in many cases and thus the covariance matrix (6) cannot be determined.

3.4 Quality of forecasts

The quality of an econometric model depends on its ability to forecast economic developments. The use of neural networks can only be justified if its prediction error is smaller than that of less complex models. In order to evaluate the different models, the out-of-sample–analysis is of major importance. However, in the neural network literature it is seldom undertaken in an appropriate way. In financial market applications, one can often find that numbers of correct hits are counted or profits are calculated, which would result from a neural network trading strategy. However, such measures do not reveal the true performance of the network, since the results may be achieved only by chance, i.e. a favourable development on the market. For this reason, it is alway neccessary to compare the result of neural networks to alternative models or strategies, in order to evaluate its relative performance.

Two possibilities of prediction exist: the forecast of the expected value and the forecast of a prediction interval. The prediction error is comprised — even in a correctly specified and estimated model — of a large insecurity due to the error terms of the regression and the variance of the estimated parameters. This insecurity is the basis for the construction of the prediction interval. The bigger the error term and the larger the variance of the parameter estimates, the wider the prediction interval.

3.4.1 Prediction error

In order to quantify the mean prediction error, one can use several different statistics, which are based on the estimated residual errors $\hat{\varepsilon}$. These measures should be computed for the training set as well as for the validation set and then be compared to each other:

- *root mean squard error* $\mathrm{RMSE} = \sqrt{\frac{\hat{\varepsilon}'\hat{\varepsilon}}{T}}$
- Theil's $U = \sqrt{(\frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'y})}$
- *mean error* $\mathrm{ME} = \frac{1}{T}\sum_t \hat{\varepsilon}_t$
- *mean percentage error* $\mathrm{MPE} = \frac{100}{T}\sum_t \frac{\hat{\varepsilon}_t}{y_t}$
- *mean absolute error* $\mathrm{MAE} = \frac{1}{T}\sum_t |\hat{\varepsilon}_t|$
- *mean absolute percentage error* $\mathrm{MAPE} = \frac{100}{T}\sum_t |\frac{\hat{\varepsilon}_t}{y_t}|$

These measures help to evaluate the quality of the models. Amongst them the RMSE is the most popular, since it is directly comparable to the standard error of regression. However, as it is an absolute measure it can only be interpreted if the scale of y is known. The advantage of the relative measures, therefore, is their independence from the size of y. Unfortunately, their values have little meaning if some of the y_t are close to zero and they cannot be computed if at least one y_t is zero.

3.4.2 Prediction intervals

The construction of a prediction interval is based on the variance of the prediction error. The computation of the prediction intervals in linear models $\hat{y} = X\hat{\beta}$ usually takes account of the regression residuals as well as the variance of the parameter estimates. The prediction error $\hat{\varepsilon}_{T+1} = y_{T+1} - \hat{y}_{T+1}$ is thus comprised of

$$\hat{\varepsilon}_{T+1} = y_{T+1} - \hat{y}_{T+1} = X_{T+1}(\beta - \hat{\beta}) + \varepsilon_{T+1} \tag{14}$$

and its variance is hence given through

$$\begin{aligned} \sigma^2_{\hat{\varepsilon}_{T+1}} &= \mathrm{Var}[X_{T+1}(\beta - \hat{\beta})] + \sigma^2_\varepsilon \\ &= \mathrm{Var}[\hat{y}_{T+1}] + \sigma^2_\varepsilon \\ &= X'_{T+1} C_{\hat{\beta}} X_{T+1} + \sigma^2_\varepsilon, \end{aligned} \tag{15}$$

where $C_{\hat{\beta}} = \sigma^2(X'X)^{-1}$ is the covariance matrix of the parameters in the linear model.

The $(1-\alpha)$–prediction interval of the exogenous variable is now calculated by

$$y_{T+1} \in [\hat{y}_{T+1} - \varphi_{\alpha/2}\sigma_{\hat{\varepsilon}_{T+1}}; \hat{y}_{T+1} + \varphi_{\alpha/2}\sigma_{\hat{\varepsilon}_{T+1}}], \tag{16}$$

where $\varphi_{\alpha/2}$ denotes the value of the standard normal distribution at the $\alpha/2$–significance level.[13]

In the nonlinear case the insecurity due to the parameter variance is ignored in most cases. The then given interval thus underestimates the true width of the prediction interval. However, the true prediction interval of $\hat{y}_{T+1}$ can be computed by help of a linear taylor expansion around w:

$$f(X_{T+1}, \hat{w}) \approx f(X_{T+1}, w) - \nabla f(X_{T+1}, w) \cdot (\hat{w} - w), \tag{17}$$

where ∇f denotes the gradient of the nonlinear regression f.[14] The variance of the prediction error results analoguous to equation (15)

$$\sigma^2_{\hat{\epsilon}_{T+1}} = \nabla f(X_{T+1}, w)' C_{\hat{w}} \nabla f(X_{T+1}, w) + \sigma^2_{\epsilon} \tag{18}$$

and the forecast interval corresponding to equation (16). The covariance matrix $C_{\hat{w}}$ of the weights can be estimated according to equation (6).

At this point it is easy to recognize the importance of parsimonious modelling. Redundant parameter are responsible for a badly behaved variance matrix $\hat{C}$ with which the prediction interval becomes larger than neccessary.

3.4.3 Bootstrapping

As an alternative to the estimation of the prediction interval, the interval can be simulated. The appropriate procedure for this is bootstrapping.[15] In this procedure, one generates (say 200) bootstrap sets by randomly taking observations from the original set. The observations are drawn with replacements until the bootstrap sets have the size of the original set. Afterwards the model is estimated with respect to each bootstrap set and any interesting statistic is calculated. In particular one can compute the different forecast values, whose histogram gives the corresponding forecast interval.[16]

Moreover, the mean and variance of such a histogram are good estimates of the expected forecast value as well as the forecast variance. It is obvious that the method of bootstrapping is very powerful, but unfortunately it is extremly computer intensive, since as many models must be trained as bootstrap sets have been agreed upon.

4 Summary

Neural networks are a new very flexible class of statistical regression procedures. Their main advantage is that they do not require an explicit assumption about the

structural form of the unknown function. Since neural networks are nothing else than nonlinear statistical procedures they offer the opportunity for a thorough statistical analysis. In this paper, the application of statistical methods for neural networks is termed 'neurometrics'.

A neurometric model building of neural networks is particular helpful, if an optimal network architecture has to be chosen for a given task. The aim is a network that is just complex enough to approximate the unknown function up to a satisfactory degree. Such a network does not show a bias and has the least possible variance when forecasting future values.

The statistical procedures usable for model building in neural networks are significance tests for parameters, information and selection criteria and cross validation. It was shown that the method of stopped training is not useful for determining an adequate network model.

It is of major importance to perform a thorough diagnosis of neural networks, in order to find out whether the chosen network architecture is actually appropriate for approximating an unknown function. For this purpose the diagnostics usually applied in econometrics (e.g. residual analysis) are extremly helpful.

The forecast quality of neural networks can be measured by the help of many different statistics which reveal the deviation in mean. In addition to this, the forecast intervals can either be computed analytically or simulated by help of the bootstrapping procedure.

Notes

[1]Compare Rehkugler/Zimmermann (1994) and Anders/Korn/Schmitt (1996).

[2]For example, the so-called radial basis function (RBF) networks, compare Poggio/Girosi (1990).

[3]Compare Hornik/Stinchcombe/White (1989).

[4]The least squares method is a special case of the maximum likelihood method if the joint distribution of the residuals is assumed to be multivariat normal with homoscedastic variance. If these assumptions are not fulfilled the application of the least squares method is statistically inefficient.

[5]Compare Press/Flannery/Teukolsky/Vetterling (1992).

[6]Compare Anders (1995).

[7]See for instance Weigend/Huberman/Rumelhart (1990).

[8]Neglected nonlinearities can for instance be tested for by help of the neural network test (compare Lee/White/Granger, 1993).

[9]Since the t-values in neural networks are usually not t-distributed, Davidson/McKinnon (1993) denote them with pseudo–t-values.

[10]Neither AIC nor NIC are consistent criteria.

[11]In nonlinear models R^2 can actually take values outside the interval [-1;1].

[12]Compare Craven/Wahba (1979) or Härdle (1990).

[13]At the usual significance level of 95% the value of φ is 1.96.

[14]Compare Seber/Wild (1989).

[15]Compare Efron/Tibishirani (1993).

[16]Compare Efron/Tibishirani (1986).

References

Akaike H. (1973): *Information Theory and an Extension of the Maximum Likelihood Principle.* In Petrov B., Csake F. (eds): *2nd International Symposium on Information Theory.* Budapest.

Amemiya T. (1980): *Selection of Regressors.* International Economic Review, 21, 331–354.

Anders U. (1996): *Was neuronale Netze wirklich leisten.* Die Bank, 3, 162–165.

Anders U. (1995): *Neuronale Netzwerke in der Ökonometrie.* ZEW Discussion Paper 95-26.

Anders U., Korn O. (1996): *Model Selection in Neural Networks.* ZEW Discussion Paper 96-21.

Anders U., Korn O., Schmitt C. (1996): *Improving the Pricing of Options — A Neural Network Approach.* ZEW Discussion Paper, 96-04.

Arminger G. (1993): *Ökonometrische Schätzverfahren für neuronale Netze.* In Bol G., Nakhaeizadeh G., Vollmer K.-H. (Eds): *Finanzmarktanwendungen Neuronaler Netze und ökonometrischer Verfahren.* Physica–Verlag, 25–39.

Craven P., Wahba G. (1979): *Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation.* Numerische Mathematik, 31, 377–403.

Davidson R., MacKinnon J.G. (1993): *Estimation and Inference in Econometrics.* Oxford University Press.

Efron B., Tibshirani R. (1986): *Bootstrap Methods for Standard Errors Confidence Intervals, and Other Measures of Statistical Accuracy.* Statistical Science, 1(1), 54–77.

Efron B., Tibshirani R.J. (1993): *An Introduction to the Bootstrap.* Chapman & Hall.

Härdle W. (1990): *Applied Nonparametric Regression.* Cambridge University Press.

Hornik K., Stinchcombe M., White H. (1989): *Multilayer Feedforward Networks are Universal Approximators.* Neural Networks, 2, 359–366.

Lee T.-H., White H., Granger C.W.J. (1993): *Testing for Neglected Nonlinearity in Time Series Models.* Journal of Econometrics, 56, 269–290.

Murata N., Yoshizawa S., Amari S. (1994): *Network Information Criterion Determining the Number of Hidden Units for Artificial Neural Network Models.* IEEE Trans. Neural Networks, 5, 865–872.

Phillips P.C.B. (1989): *Partially Identified Econometric Models.* Econometric Theory, 5, 181–240.

Poggio T., Girosi F. (1990): *Networks for Approximation and Learning.* Proceedings of the IEEE, 78(9), 91–106.

Press W.H., Flannery B.P., Teukolsky S.A., Vetterling W.T. (1992): *Numerical Recipes in C.* Cambridge University Press.

Rehkugler H., Zimmermann H.G. (1994): *Neuronale Netze in der Ökonomie.* Vahlen.

Seber G.A.F., Wild C.J. (1989): *Nonlinear Regression.* John Wiley & Sons.

Teräsvirta T., Lin C.-F., Granger C.W. (1993): *Power of the Neural Network Linearity Test.* Journal of Time Series Analysis, 14(2), 209–220.

Weigend A.S., Hubermann B.A., Rummelhart D.E. (1990): *Predicting the Future: a Connectionist Approach.* International Journal of Neural Systems, 1, 193–209.

White H. (1989a): *Learning in Neural Networks: A Statistical Perspective.* Neural Computation, 1, 425–464.

White H. (1989b): *An Additional Hidden Unit Test for Neglected Non-linearity in Multilayer Feedforward Networks.* Proceedings of the International Joint Conference on Neural Networks, Washington, DC. San Diego: SOS Printing, II, 451–455.

White H. (1994): *Estimation, Inference and Specification Analysis.* Cambridge University Press.