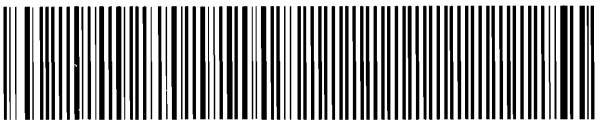
Discussion Paper

Discussion Paper No. 96-21

Model Selection in Neural Networks

Ulrich Anders
Olaf Korn



ZEW
Zentrum für Europäische Wirtschaftsforschung GmbH

International Finance Series

Model Selection in Neural Networks

Ulrich Anders, Olaf Korn

Centre for European Economic Research (ZEW), Mannheim

Tel: 0621/1235-141 or -147
Fax: 0621/1235-223
Email: anders@zew.de or korn@zew.de

PO Box 10 34 43
68034 Mannheim

December 1996

Abstract

In this article we examine how model selection in neural networks can be guided by statistical procedures such as hypotheses tests, information criteria and cross validation. The application of these methods in neural network models is discussed, paying attention especially to the identification problems encountered. We then propose five specification strategies based on different statistical procedures and compare them in a simulation study. As the results of the study are promising, it is suggested that a statistical analysis should become an integral part of neural network modelling.

Key words: *Neural Networks, Statistical Inference, Model Selection, Identification, Information Criteria, Cross Validation.*

Acknowledgements

We are indebted to Daniel Schwamm for his capable research assistance.

1 Introduction

Recently, there has been a growing interest in the modelling of nonlinear relationships and a variety of test procedures for detecting nonlinearities has been developed.[1] If the aim of analysis is prediction, however, it is not sufficient to uncover nonlinearities. Moreover, we need to describe them through an adequate nonlinear model. Unfortunately, for many applications theory does not guide the model building process by suggesting the relevant input variables or the correct functional form.

This particular difficulty makes it attractive to consider an 'atheoretical' but flexible class of statistical models. Artificial neural networks are well suited for this purpose as they can approximate virtually any (measurable) function up to an arbitrary degree of accuracy (Hornik/Stinchcombe/White, 1989). This desired flexibility, however, makes the specifcation of an adequate neural network model even harder. Despite the huge amount of network theory and the importance of neural networks in applied work, there is still little experience with a statistical approach to model selection.

The aim of this article is to develop model selection strategies for neural networks which are based on statistical concepts. Taking a statistical perspective is especially important for 'atheoretical' models like neural networks, because the reason for applying them is the lack of knowledge about an adequate functional form.[2] Furthermore, when based on a clearly defined decision rule, model selection becomes more comprehensible and reconstructable. The concepts considered in this article are *hypothesis testing*, *information criteria* and *cross validation* methods. These concepts are the building blocks which constitute the basis of the different model selection strategies which we develop and evaluate in a simulation study. To our kowledge this article provides the first systematic comparison of statistical selection strategies for neural network models.

The overall results of the simulation study are promising as they lead to neural networks which closely approximate the simulated models. Our results demonstrate that a sequence of hypothesis tests produces neural network architectures with the best overall performance. Strategies based on cross validation and information criteria are very accurate for some models although they tend to overfit or underfit others. When information criteria are to be employed, we cannot recommend the use of an estimated penalty term.

The remaining part of the article is organized as follows: Section 2 defines neural network models and briefly describes some network specification methods frequently used in applied work. Section 3 reviews the theory of hypothesis testing and inference in neural networks. Such an analysis is severly complicated by possibly non-identified parameters. Sections 3.1 and 3.2 shortly describe two workarounds for the identification problem due to White (1989a) and Teräsvirta/Lin/Granger (1993). Section 4 briefly introduces the network information criterion NIC, and investigates the applicability of the AIC in neural network models. It is pointed out that the use of the AIC may theoretically not be justified. After that, cross validation techniques, frequently referred to in the neural networks literature, are introduced. Section 5 defines our model selection strategies. In section 6

[1] See Granger/Teräsvirta (1993) for an overview.
[2] See Ripley (1993) and Kuan/White (1994).

the strategies are compared in a simulation study. Section 7 summarizes the main results and concludes the article.

2 Neural Network Models

Neural networks build a class of very flexible models which can be used for a variety of different applications, e.g. nonlinear regression or discriminant analysis.[3] Unfortunately, the term 'neural network' is not uniquely defined. Instead it comprises many different network types and models. In this article, we will deal exclusively with so called 'multilayer perceptron networks', an example of which is shown in Figure 1 below.

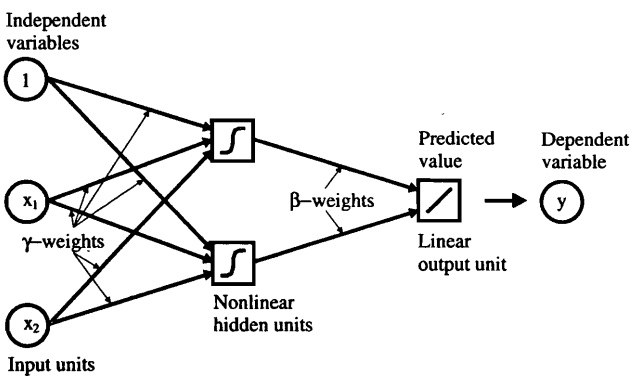


Figure 1: A multilayer perceptron neural network.

The network outputs shall serve as predicted values of the dependent variable y and can be expressed by a function $f(X, w)$ of the input data $X = [x_0, x_1, \ldots, x_I]$ and the network parameters w commonly called weights. For an architecture of the type shown in Figure 1 the network function takes the following form

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

with network weights $w = (\beta_1, \ldots, \beta_H, \gamma_{10}, \ldots, \gamma_{HI})'$. The variable x_0 is defined to be constant and set to $x_0 \equiv 1$. The scalars I and H denote the number of input and hidden

[3]For a more detailed discussion of how neural networks compare to classical statistical procedures see Sarle (1994).

units in the net and $g(\cdot)$ is a nonlinear transfer function attached to each hidden unit. Usually, $g(\cdot)$ is either the logistic function or the hyperbolic tangent function. Because of its symmetry around the origin and its easily computable derivatives we prefer to use the tanh-function.

By adding an error term to equation (1) the network can be interpreted as a parametric nonlinear regression of y on X. Thus, when applied to a data set, a certain functional form, the network architecture, is assumed and the parameters of the network are estimated thereafter.[4]

One of the most unresolved questions in the literature on neural networks is what architecture should be used for a given problem. Architecture selection requires choosing both the appropriate number of hidden units and the connections thereof (Sarle, 1995). A desirable network architecture contains as little hidden units and connections as necessary for a good approximation of the true function, taking into account the trade-off between estimation bias and variability due to estimation errors. Unfortunately, the form of the true function is rarely known (otherwise, one would not use a neural net to approximate it). It is therefore necessary to develop a methodology to select an appropriate network model for a given problem.[5]

The usual approaches pursued in the network literature are *regularization*, *pruning*, and *stopped training*. Reed (1993) provides a survey. Although some of these methods may lead to satisfactory results, their derivation has been rather heuristic. For example, in regularization methods the network weights are chosen such that they minimize the network error function (e.g. sum of squared errors) plus a penalty term for the networks complexity. Usually the penalty terms do not result from theoretical reasoning but are set in a rather ad hoc fashion. In order to justify the regularization term the method has been formalized and interpreted in a bayesian framework. This was mentioned by Weigend/Huberman/Rumelhart (1991) and reviewed in Bishop (1995) or Ripley (1996). In our view the approach suffers from the difficulty to choose prior distributions for network parameters which have no intuitive interpretation. Though the bayesian approach solves some of the identification problems which complicate classical inference, it is difficult to apply in practical situations. For this reason, we will not consider a bayesian approach here and confine ourselves to applications of classical statistics.

The aim of pruning methods is to identify those parameters that do not 'significantly' contribute to the overall network performance. However, this 'significance' is not judged on the basis of test statistics. Instead, pruning methods use the so called 'saliency' as a measure of a weight's importance. The saliency of a weight is defined as the increase in network model error (e.g. sum of squared errors) incurred by setting this weight to zero. The idea is to remove the weights with low saliency; however the method does not provide any guidelines whether or not a saliency should be judged as low.

[4]The estimation procedure is usually called *network training*.
[5]See Anders (1997).

In the application of stopped training the dataset is split into a training and a validation set. If the model errors in the validation set begin to grow during the training process the training algorithm is stopped. In statistical terms, the method tries to make up for the model being overparameterized by stopping the estimation algorithm before the minimum of the network error function is reached. In our view, this does not lead to sensible estimates of the network parameters. Instead, growing errors in the validation set should be seen as an indication to reduce the network's complexity.

The main disadvantage of regularization, pruning and stopped training is that these methods comprise of a strong judgemental component, which makes the model building process difficult to reconstruct. The next two sections describe statistical concepts which can serve as building blocks for modelling strategies which overcome this deficiency.

3 Hypothesis Testing in Neural Networks

Since multilayer perceptron neural networks are nonlinear regression models the standard procedures for testing parameter significance like Wald–tests or LM–tests in principle apply. However, to perform these tests the (asymptotic) distribution of the network parameters is needed. This distribution was derived by White (1989b). If there exists a unique set of optimum parameters w^* that leads to the best approximation of the true function by a certain network model, White shows that w^* can be consistently estimated through a set of parameters $\hat{w}$ obtained by quasi maximum likelihood methods under quite general conditions. Furthermore, $\hat{w}$ is asymptotically normal with mean w^* and covariance matrix $\frac{1}{T}C$, or

$$\sqrt{T}\,(\hat{w} - w^*) \sim \mathcal{N}(0, C), \tag{2}$$

where T is the number of observations. Since neural networks are 'atheoretical' in spirit and therefore only approximations to the true underlying functions the derivation of the covariance matrices relies on the theory of misspecified models developed by White (1981, 1982, 1994). If a neural network is estimated via the maximum likelihood method with log likelihood contribution $\mathcal{L}_t$, the covariance matrix of the parameters is given by $\frac{1}{T}C = \frac{1}{T}A^{-1}BA^{-1}$. The matrices A and B are defined as $A \equiv -E[\nabla^2 \mathcal{L}_t]$ and $B \equiv E[\nabla \mathcal{L}_t \nabla \mathcal{L}_t']$ where ∇ denotes the gradient. Note that the covariance matrix C accounts for misspecification in mean as well as in variance, i.e. a consistent estimate of C converges to the covariance matrix of the limiting distribution of $\hat{w}$. This holds for example if the network does not encompass the true underlying function and a possibly heteroscedastic error term has not been taken account of in the specification of the likelihood function. As can be seen the covariance matrix C is a generalization of the covariance formula in standard maximum likelihood theory. If the model is correctly specified, the asymptotic covariance matrix of the limiting distribution can be estimated by both the inverse of the matrix A and the inverse of the matrix B. In this case A asymptotically equals B and BA^{-1} converges to I.

However, one severe problem remains. The optimum w^* is not unique for the network model given in (1), i.e. the model is not globally identified. This problem is a rather hard one to deal with, and the question of identification will complicate the application of hypothesis tests and information criteria considerably.

There are two characteristics of neural networks which cause the non-identifiability: the first one is due to symmetries in the architecture of a neural network, which lead to multiple optima. For example, if the hidden units in Figure 1 swap places, the architecture would remain unchanged, while the optimum weight vector would become permutated. Fortunately, the possible presence of multiple optima has no essential effect on the result in (2). It is simply required that the estimated weight vector $\hat{w}$ is a consistent estimator for a single w_p^* of all optimum weight vectors building a set W^*. In other words, the above result (2) remains unchanged if an optimum weight vector is (locally) unique in a small neigbourhood.

The second reason why network parameters are not identified is the mutual dependence of the β- and γ-weights shown in Figure 1. A β-weight between a hidden unit and the output unit equals zero, the corresponding γ-weights leading into that hidden unit can take any value and are thus not unique. If the γ-weights which lead into a hidden unit are all zero the corresponding β-weight is not identified. In this case, the set W^* of optimum solutions contains values corresponding to flat regions of the quasi likelihood function. If convergence of $\hat{w}$ to one of these flat regions occurs, the limiting distribution of the estimated parameters $\hat{w}$ is no longer normal. Instead, the distribution of the parameters belongs to the family of 'mixed Gaussian' as was shown by Phillips (1989) for 'partially identified models'. In other words, if we want to carry out parameter inference in a neural network on the basis of an asymptotic normal distribution, we must guarantee that the parameters are at least locally unique. In order to guarantee this, it is necessary to ensure that a given network model contains no irrelevant hidden units. The question, whether or not there are any irrelevant hidden units in a neural net, can in principle be investigated by tests on parameter significance. However, since β- and γ-weights are mutually dependent on each other, there are two ways to formulate the null hypothesis. Either we test whether a β-weight is significantly different from zero, or we test whether at least one of the corresponding γ-weights is significantly different from zero. In the first case the null is $H_0 : R\beta = 0$ with the alternative $H_1 : R\beta \neq 0$, in the second case the null is $\tilde{H}_0 : \tilde{R}\gamma = 0$ with the alternative $\tilde{H}_1 : \tilde{R}\gamma \neq 0$, where R and $\tilde{R}$ are restriction matrices that pick out the weights in question. In both cases, if the null is true the parameters are not locally unique and thus the estimator does not follow an asymptotic normal distribution. This is exactly the problem of 'hypothesis testing when a nuisance parameter is present only under the alternative' studied by Davies (1977, 1987). In this case, the resulting test statistics of the Wald- or the LM–test no longer follow a χ^2 distribution and further analysis is complicated. However, two techniques have been proposed in the literature which yield a χ^2-statistic for the testing problem and avoid the difficulties mentioned. One technique was developed by White (1989a) and its properties investigated by Lee/White/Granger (1993). The resulting test has good power against a variety of nonlinear alternatives. The other technique was devised by Teräsvirta/Lin/Granger (1993) and compared to the former.

Both groups of authors start from the model $y = F(X) + \varepsilon$, where $F(\cdot)$ is the true function, ε an iid random noise with $E[\varepsilon\varepsilon'|X] = \sigma I$, $E[\varepsilon|X] = 0$ and $E[X'\varepsilon] = 0$. It is further assumed, that $F(X)$ has already been approximated by a parametric function $f(X, \hat{w})$.[6] The question is now: can the approximation of $F(X)$ through $f(X, \hat{w})$ be improved by adding Q (one or more) hidden units in order to capture some neglected nonlinearities? If the answer is yes, the data can be explained more accurately by the following equation:

$$y = f(X, \hat{w}) + \sum_{q=H+1}^{H+Q} \hat{\beta}_q g(\sum_{i=0}^{I} \hat{\gamma}_{qi} x_i) + \hat{\eta} \tag{3}$$

The new residual is denoted by $\hat{\eta}$. The appropriate tests on the additional parameters β_q or γ_q are called 'tests against neglected nonlinearities'. If $f(X, \hat{w})$ in (3) is the linear function $f(X, \hat{w}) = X\hat{w}$, which implies $H = 0$, the test against neglected nonlinearities becomes a test of model linearity against model nonlinearity.

3.1 LM–Test Procedure using Random Sampling

The neural network test proposed by White (1989a), relies on the hypothesis $H_0 : \beta_q = 0$ for all $q = H+1, \ldots, H+Q$ against the alternative $H_1 : \beta_q \neq 0$ for all $q = H+1, \ldots, H+Q$. The idea is based on the following consequence of H_0. If H_0 is true, then $F(X) = f(X, w)$ and the residuals ε from the regression of y on $f(X, w)$ is independent from X by definition. Consequently, the residuals ε are independent from any function of the X, say $s(X)$, which implies that $E[s(X)'\varepsilon] = 0$. Thus, if the signals $s_q(X, \gamma_q) = g(\sum_i \gamma_{qi} x_i)$ from the additional hidden units are correlated with the residuals $\hat{\varepsilon}$, that is $E[s_q(X, \gamma_q)'\varepsilon] \neq 0$, the hypothesis H_0 cannot be true.

In order to avoid the identification problem, the weigths $\gamma_q = (\gamma_{q0}, \ldots, \gamma_{qI})'$ are drawn from a uniform distribution, which is denoted with $\dot{\gamma}_q$. The weights are chosen such that the signals $s_q(X, \dot{\gamma}_q)$ are not collinear to the gradient $\nabla f(X, \hat{w})$. Otherwise the signals of the additional hidden units would only provide information that is already present in the model $f(X, \hat{w})$. Drawing weights $\dot{\gamma}_q$ from a random distribution amounts to a random choice in the parameter space of the γ_q. The test is carried out conditional on the sampled values of γ_q. Due to the random choice of the γ_q it often occurs that the signals $s_q(X, \dot{\gamma}_q)$ of the additional hidden units are heavily correlated. The problem can be remedied by sampling a large number of hidden unit signals, say $\tilde{Q} \gg Q$, and by subsequently choosing the Q most important principal components which are not collinear to $\nabla f(X, \hat{w})$ to be signals s_q of the additional hidden units.

[6] The function $f(X, \hat{w})$ may be an arbitrary function (e.g. a linear function). It may but need not be a neural network.

The test procedure is pursued in the fashion of a standard Lagrange multiplier test.

1. Regress y on $f(X,w)$ and compute the residuals $\hat{\varepsilon}$.
2. Regress the residuals $\hat{\varepsilon}$ on the gradient $\nabla f(X,\hat{w})$ and the Q signals $s_q(X,\gamma_q)$ from the additional hidden units. This regression is commonly called *Gauss–Newton–regression* (GNR).[7] Compute the uncentered squared multiple correlation R_u^2 from the GNR.
3. According to White (1989a) the test statistic is TR_u^2 which is asymptotically χ_Q^2 distributed. However, as Davidson/MacKinnon (1993) point out, it is probably safer to use $(T-K+Q)R_u^2$ as test statistic in finite samples, where K is the number of parameters in the unrestricted model and Q is the number of restrictions. The hypothesis H_0 is rejected if the value of the test statistic exceeds the appropriate value of the χ_Q^2 distribution. Furthermore, Davidson/MacKinnon (1993) suggest that the ordinary F–statistic from the GNR for $\beta_q = 0$ $(q = H+1,\ldots,H+Q)$ may have even better finite sample properties.

3.2 LM–Test Procedure using Taylor Expansions

An alternative test procedure has been proposed by Teräsvirta/Lin/Granger (1993). As opposed to the method suggested by White, this test relies on the null hypothesis $\tilde{H}_0$: $\gamma_q = 0$ for all $q = H+1,\ldots,H+Q$. The problem hereby is that the β_q's are not identified under the null. The identification problem can be solved in the spirit of Davies (1977) by using a Taylor series approximation of the additional hidden unit transfer functions. A third order Taylor expansion of a hidden unit transfer function $g(X\gamma) = \tanh(X\gamma)$ around $X\gamma = 0$ results in:

$$\tanh(X\gamma) \approx \sum_{i=0}^{I} \gamma_i x_i - \frac{1}{3}\sum_{i=0}^{I}\sum_{j=i}^{I}\sum_{k=j}^{I} \delta_{ijk}\, x_i x_j x_k, \tag{4}$$

where I is the number of inputs used in the model and δ_{ijk} are the corresponding coefficients of the cubical terms. By replacing the function $g(\cdot)$ in the second term of equation (3) the unrestricted model becomes:

$$y = f(X,w) + \sum_{q=H+1}^{H+Q} \beta_q [\sum_{i=0}^{I} \gamma_{qi} x_i - \frac{1}{3}\sum_{i=0}^{I}\sum_{j=i}^{I}\sum_{k=j}^{I} \delta_{q,ijk} x_i x_j x_k] + \eta, \tag{5}$$

where the error term is denoted with η. Collecting terms leads to

$$y = f(X,w) + \sum_{i=0}^{I} (\sum_{q=H+1}^{H+Q} \beta_q \gamma_{qi})\, x_i - \frac{1}{3}\sum_{i=0}^{I}\sum_{j=i}^{I}\sum_{k=j}^{I} (\sum_{q=H+1}^{H+Q} \beta_q \delta_{q,ijk})\, x_i x_j x_k + \eta. \tag{6}$$

[7] See Davidson/McKinnon (1993).

By defining the parameters to be estimated in this model as $\theta_i \equiv \sum_{q=H+1}^{H+Q} \beta_q \gamma_{qi}$ and $\theta_{ijk} \equiv \sum_{q=H+1}^{H+Q} \beta_q \delta_{q,ijk}$ equation (6) simplifies to

$$y = f(X, w) + \sum_{i=0}^{I} \theta_i \, x_i - \frac{1}{3} \sum_{i=0}^{I} \sum_{j=i}^{I} \sum_{k=j}^{I} \theta_{ijk} \, x_i x_j x_k + \eta. \tag{7}$$

First, note that some summands may merge with summands of the model $f(X, w)$. Second, the values of the thetas are not dependent on the number of additional hidden unit transfer functions being approximated. Therefore, if the test accepts the Taylor expansion to be significant we can only add one further hidden unit to the original network architecture. Third, as can be reckoned from equation (4), the third order Taylor expansion may add quite a large number of summands to the nested model $f(X, w)$. The cubical sum already enlarges the nested model by $\binom{(I+1)+3-1}{3}$ linear terms, which for instance amounts to 56 if the number of inputs including the constant is $(I + 1) = 6$. As all these terms result from combinations of the same input variables, multicollinearities amongst them are very likely. To improve the power of the test we propose to replace the additional terms by some of their most important principal components. The number of principal components to use in (7) can be choosen such that a high proportion, say 99 percent, of the total variance is explained. This procedure is valid as the information contained in the Taylor expansion remains in the principal components.

A test for an additional hidden unit is now performed in model (7). The null hypothesis $\tilde{H}_0 : \theta = 0$ is tested against the alternative $\tilde{H}_1 : \theta \neq 0$, where θ is the vector of restricted parameters whose (principal component) terms have not merged with the nested model $f(X, \hat{w})$. The test procedure runs analogous to the one given in the last section.

3.3 Wald–Test

When the network does not contain irrelevant hidden units, one can test for arbitrary parameter restrictions on the γ–weights by help of a Wald–test. The test statistic is given by $(R\hat{w})'(R\hat{C}R')^{-1}(R\hat{w}) \sim \chi^2_q$, where R determines the form of restrictions and q denotes their number. The matrix $\hat{C}$ is the estimate of the covariance matrix defined earlier. In this paper we only apply exclusion restrictions which test for the relevance of γ–weights. One very important variant is a test for an irrelevant input variable. In this case, R selects those γ–weights which are linked to the input unit in question.

4 Information Criteria for Neural Networks

The underlying idea of information criteria is to find an optimal trade-off between an unbiased approximation of the underlying model and the loss of accuracy caused by estimating an increasing number of parameters. To achieve this, information criteria combine some measure of fit with a penalty term to account for model complexity. A variety of different criteria can be found in the literature.[8] The most prominent and still widely used criterion is probably the AIC (Akaike, 1973 and 1974), which in principle applies to any model estimated by maximum likelihood. The AIC is defined as

$$\text{AIC} = -\frac{2}{T} \ln L(\hat{w}) + \frac{2K}{T}, \tag{8}$$

where $\ln L(\hat{w})$ is the estimated maximum log likelihood. Unfortunately an application of the AIC is complicated as soon as we turn to neural networks. As mentioned in section 3, it is reasonable to think of a neural network as an approximation to an underlying model and analyse it as being misspecified in the sense of White (1981, 1982, 1994). In this context the AIC does not apply, since it assumes the model structure to be the true one. Fortunately, a generalization to the AIC for misspecified models has been proposed. The criterion is called NIC (Network Information Criterion) and was developed by Murata/Yoshizawa/Amari (1994).[9] The NIC chooses a specification for which the following expression (9) takes a minimum:

$$\text{NIC} = -\frac{1}{T} \ln L(\hat{w}) + \frac{\text{tr}[BA^{-1}]}{T}, \tag{9}$$

The matrices A and B are defined to be $A \equiv -E[\nabla^2 \mathcal{L}_t]$ and $B \equiv E[\nabla \mathcal{L}_t \nabla \mathcal{L}_t']$ like in section 3. If the class of models investigated includes the true model, A equals B asymptotically, thus, $\text{tr}[BA^{-1}] = \text{tr}[I]$ is the number of model parameters K. By multiplying with 2 the NIC reduces to the AIC.

Even if a neural network model encompasses the true structure, we are not in general allowed to apply either NIC or AIC. Both criteria were derived under the assumption of asymptotic normality of the maximum likelihood estimators. Hence the criteria are not valid for overparameterized networks, e.g. networks with irrelevant hidden units. As mentioned in section 3 such networks contain unidentified parameters, whose limiting distribution is 'mixed gaussian' instead.[10] Since the purpose of a model selection strategy is just to recognize overparameterized models the use of information criteria for neural networks is questionable, at least if they are naively applied.

[8] See Judge et. al. (1985), p. 870 ff.

[9] A similar criterion which is known as the 'effective number of parameters' in the network literature was developed by Moody (1992).

[10] The same reasoning applies to the SIC (see Schwartz, 1978), which was used by Swanson/White (1995) to select a network architecture. Their strategy does not take into account the identification problem, however.

In order to avoid the difficulties mentioned one possible strategy would be the use of information criteria only for those models which were decided to be identified on other grounds. Identification can for example be judged by tests of significance as they were introduced in section 3. A further strategy we put forward here is to proceed as Teräsvirta/Lin/Granger (1993) in the context of hypothesis testing. When the identification problem is circumvented by help of a Taylor series expansion of the additional hidden unit transfer function, the hidden unit reduces to linear terms, which allows the use of AIC, NIC or other information criteria.

An alternative model selection method, often referred to in the neural networks literature, is the so called cross validation (CV), or more specific v-leave-out cross validation.[11] The motivation for this model selection is similar to the line of arguments leading to information criteria. Adding model complexity need not result in a better description of an underlying function due to increasing estimation errors. In order to find an appropriate degree of complexity, it is appealing to compare the mean squared prediction errors (MSPE) of different model specifications. Such prediction errors are obtained by dividing the sample into M subsets which contain v observations each. The model is repeatedly reestimated, leaving out one different subset each time. The average mean squared prediction error on the M subsets that have been left out defines the cross validation error CV.

$$\mathrm{CV} = \frac{1}{M} \sum_{m=1}^{M} \mathrm{MSPE}_m \,. \tag{10}$$

The model with the lowest cross validation error is finally chosen. An advantage of cross validation lies in its independence of probabilistic assumptions, especially the properties of maximum likelihood estimators. On the other hand, splitting the data results in a loss of efficiency. Furthermore, the calculation of cross validation errors can be cumbersome due to the frequent reestimation of the models considered.

[11] See e.g. Stone (1974) or Efron/Tibshirani (1993).

5 Model Selection Strategies

In order to specify a network architecture we have to choose the relevant input variables and the appropriate number of hidden units, i.e. the complexity of functional form. Both problems can be dealt with statistically, because irrelevant units result in zero restrictions of the network parameters w^*.

Whenever test statistics or information criteria are applied, we have to ensure the (local) identification of our model. Therefore, we cannot adopt a pure top down approach which starts with a large (and probably overparameterized) neural net. To obtain statistically valid results we always begin with an empty model and successively add hidden units. We assume that a number I of input variables which possibly enter the model is given. When the appropriate number of hidden units is determined single input connections will sucessively be removed as to reach the optimal architecture. This general structure is common to all suggested specification strategies in order to make the different approaches comparable and to fulfill some practical restrictions that are imposed by the demand on computer time.

5.1 Strategies based on Sequential Tests

The first two selection strategies rely on sequential hypotheses tests. As a starting point all I input variables are combined with one hidden unit and the relevance of this unit is tested by the LM–test procedures of White (1989a) or Teräsvirta/Lin/Granger (1993). If the test fails to show significance, the whole procedure would stop; if the unit is relevant, it is included in the model. In this case the network is estimated and a further fully connected hidden unit tested for significance. The procedure continues until no further additional hidden unit shows relevance. Once the number of hidden units is determined Wald–tests are applied in a top down approach to decide on the significance of single input connections. If there are insignificant connections, the one with the highest p-value is removed from the model and the reduced network reestimated thereafter. This procedure is carried on until only significant connections remain in the model. The two strategies using a sequence of hypotheses tests are summarized in Figures 2 and 3.

The proposed strategies ensure that the percentage of selected models which are overparameterized with respect to the number of hidden units is bounded by the sizes of the LM–tests.[12] In how far the procedure favours too small models depends on the power of the tests and will be investigated in the simulation.

It is clear that many different specification strategies can be devised which combine the LM– and Wald–tests. One restriction, however, is that no inference on single input connections should be drawn until the relevance of the associated hidden units is examined. Although the identification cannot be guaranteed it can at least be tested for.

[12]The test sizes may be different from the chosen significance levels in finite samples. Simulation results for the size of White's neural network test are given in Lee/White/Granger (1993, p. 280)

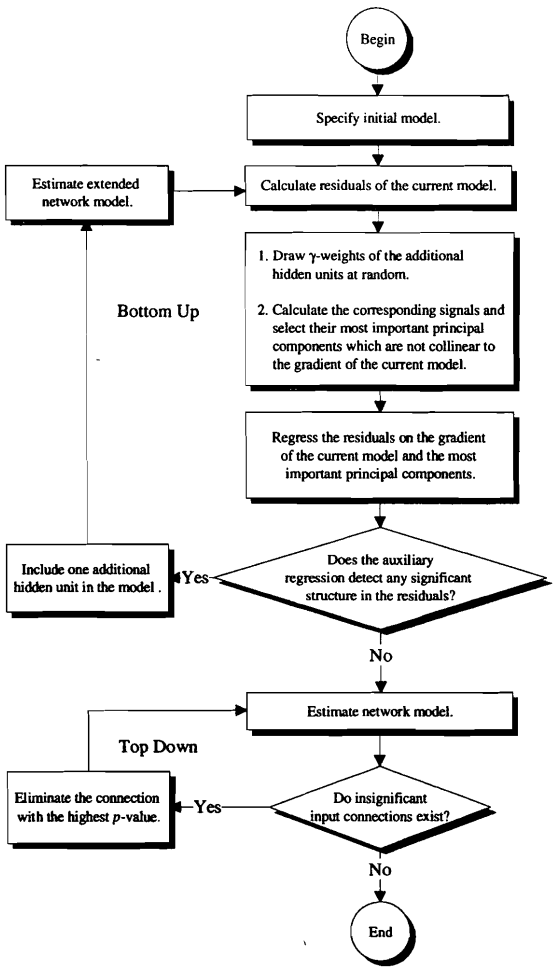


Figure 2: Model selection by help of sequential tests based on the LM–Test of White (1989a).

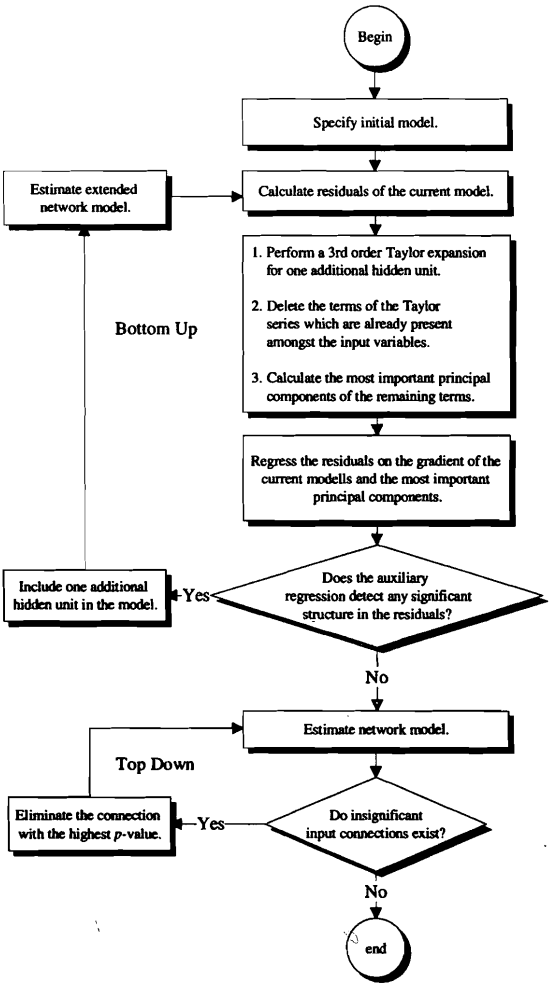


Figure 3: Model selection by help of a sequential tests based on the LM–Test of Teräsvirta/Lin/Granger (1993).

5.2 Strategies based on Information Criteria

Our model selection strategies based on information criteria are similar to the test procedures described above. We estimate a restricted model and decide on the grounds of an information criterion whether the model can be improved by lifting the restrictions. Again a bottom up procedure is used to determine the number of hidden units, followed by a top down approach to specify the appropriate input connections.

In the first step we compute the residuals from the initial model and check whether some structure in the residuals can be explained by a single, fully connected hidden unit extension to the initial model. Identified model extensions are obtained by a third order Taylor series approximation of the transfer function. Subsequently, the value of either AIC or NIC is calculated and the hidden unit is accepted when criterion values show an improvement over an empty model. When this is the case an enlarged network is estimated, new residuals are computed and the relevance of an additional hidden unit is examined via the information criteria. The procedure stops when an additional hidden unit does not lead to further improvements.

The top down strategy starts from the fully connected network obtained in the first step and tries to detect irrelevant input connections. All submodels with one of the input connections removed are estimated and compared with the full network by means of the information criteria. If the full network turns out to show the lowest criterion value, the specification strategy stops. Otherwise the best submodel is chosen, which serves as the starting network for the next round of the specification process. Again the starting network is compared with all submodels containing one input connection less. Thus in each round of the top down strategy either the procedure stops or one input connection is removed. The IC–strategies are summarized in Figure 4.

As soon as the number of hidden units is determined one could in principle compare the criterion values for all combinations of input variables and hidden units. But even in small networks this results in an enormous number of specifications which have to be estimated. Therefore, we decided to proceed in the fashion of sequential testing and run a top down strategy which successively eliminates one input connection in each step.

A particularly interesting aspect of this study is the comparison between the alternative criteria AIC and NIC. The NIC is theoretically valid even for misspecified models but requires, in contrast to the AIC, the estimation of the penalty term. It is therefore an empirical question which criterion turns out to be superior in which situation.

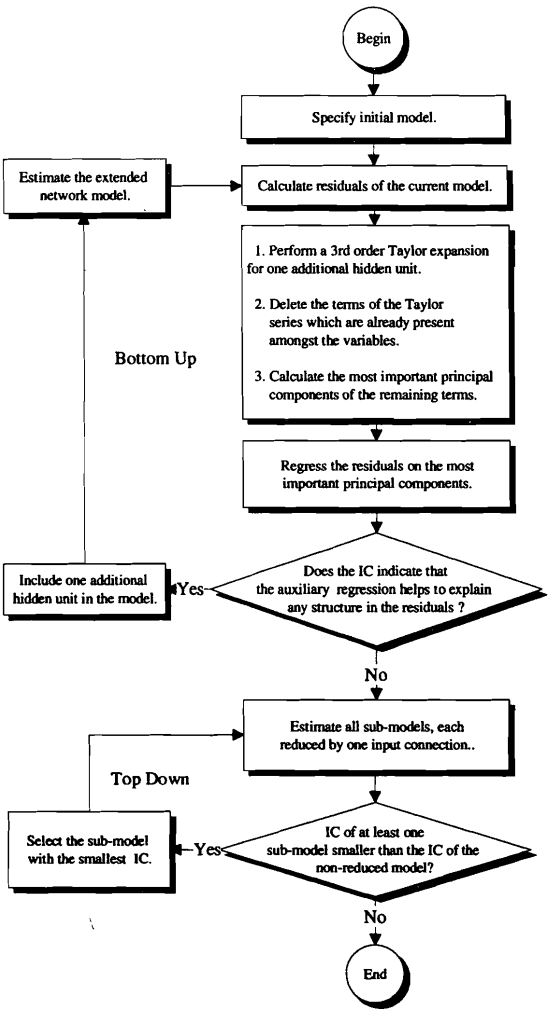


Figure 4: Model selection by help of information criteria.

5.3 Strategy based on Cross Validation

Cross Validation is the most generally applicable strategy for model selection in neural networks since it does not rely on any probabilistic assumptions and is not affected by identification problems. In principle, all combinations of input variables and hidden units can be compared. The resulting models are repeatedly estimated for subsamples with v observations left out at a time, and the model with the smallest averaged mean squared prediction error is selected. However, for all but the smallest networks this is hardly feasible.

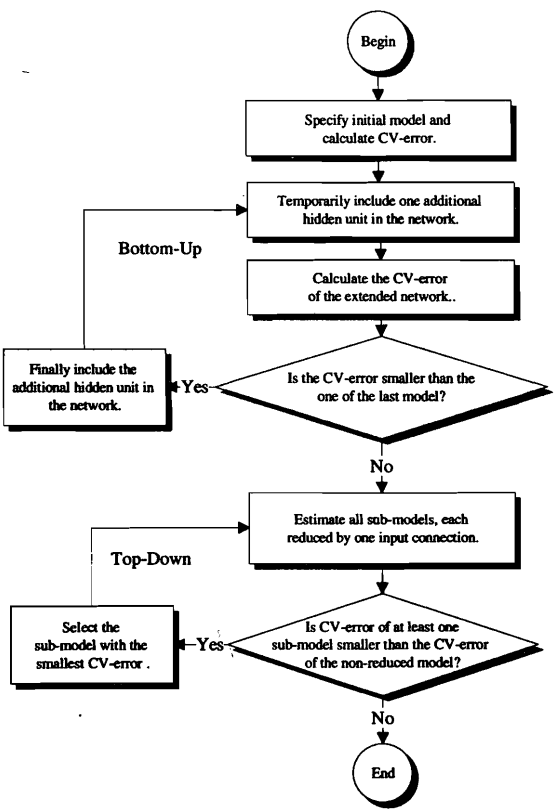


Figure 5: Model selection by help of cross validation.

Therefore, we again start by moving bottom up in order to determine the appropriate number of hidden units. In each step the cross validation errors of two models are compared, one of which contains an additional hidden unit. If the more complex network turns out to show a smaller cross validation error, the additional hidden unit is accepted and the network enlarged. The procedure stops when no further hidden unit is able to reduce the cross validation error of the previous model. Subsequently the top down part of the strategy follows. The selection of input connections runs analogous to the procedure based on information criteria, except that the cross validation errors are employed instead of the criterion values. In each round an initial network is compared with all submodels containing one input connection less. In the finally chosen network model no input connection can be removed without increasing the cross validation error. The CV–strategy is summarized in Figure 5.

6 Monte Carlo Comparison of Strategies

6.1 The Simulation Design

The simulation study is designed to highlight some aspects of the different selection strategies. As a first aspect we want to gain some experience on whether a given strategy tends to underfit or overfit the data. Therefore we need to simulate from a true model which itself is a neural network. A second important aspect is to see how the selection strategies work when the true model is not nested in the class of neural networks. In this case one would expect that a correction for misspecification, as it is employed e.g. by the NIC, leads to a superior performance.

In the simulation study we consider three different models. The first one (M1) is a neural network which consists of three input units in addition to a constant, two hidden units and a linear output unit. The network is not fully connected, as the last input unit is not linked to the second hidden unit. Thus the network model contains 7 parameters and can be described through:

$$y = \beta_1 g\left(\sum_{i=0}^{3} \gamma_{1i} x_i\right) + \beta_2 g\left(\sum_{i=0}^{2} \gamma_{2i} x_i\right) + \varepsilon, \qquad \text{(M1)}$$

where ε is a zero mean error term. For this and the subsequent models the X–variables are drawn from a standard normal distribution. We generate several weight vectors randomly and choose the one that leads to the lowest correlation between the signals of the hidden units. This is done in order to give the hidden units a high justification. The errors are drawn from a normal distribution whose standard deviation equals twenty percent of the unconditional standard deviation of y, i.e. $\sigma_\varepsilon = 0.2\, sigma_y$.

As the true number of hidden units in the first model is known, we can conclude which strategy rather leads to overparameterized and which to underparameterized network architectures. The model further allows us to learn about the size and power of the testing strategies.

Apart from this network model, we consider two further models which a neural network can only approximate, i.e. the resulting networks are misspecified. The second model has been chosen as

$$y = \ln(x_1 + 4) + \sqrt{x_2 + 4} + \varepsilon, \tag{M2}$$

the third model as

$$y = -0.5 + 0.2x_1^2 - 0.1\exp(x_2) + \varepsilon. \tag{M3}$$

Both of these models are motivated by transformations which are quite common in econometrics. The functions underlying models M2 and M3 are depicted in the following Figure. As model M3 shows a more complex nonlinear structure than model M2 we expect the approximating networks to consist of more hidden units and input connections.

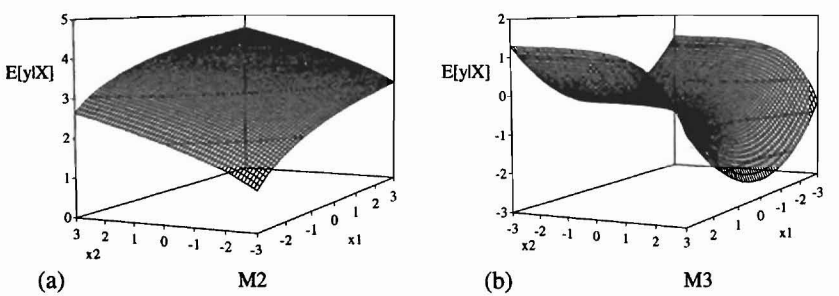


Figure 6: The structure of the models M2 and M3.

For all models the independent variables $X = [x_1, x_2]$ are drawn from a standard normal distribution. The standard deviation σ_ε of the error terms ε_t is chosen to be twenty percent of the unconditional standard deviation of each model's output. The whole set of simulated data consists of a 1000 observations, which we split into an in-sample and an out-of-sample set containing 500 data points each.

In the simulation study we compare the out-of-sample mean squared prediction errors (MSPE_{MSS}) of the networks resulting from our model selection strategies (MSS) with the known out-of-sample mean squared error (MSE_{TM}) of the true model (TM). We repeated the experiment a thousand times, each time redrawing the in-sample random errors, while the out-of-sample errors where kept constant. The best out-of-sample performance which the model selection strategy can — apart from chance — achieve, is the out-of-sample MSE_{TM} of the true model. Therefore, this value is taken as a benchmark.

We use nonlinear least squares as the estimation method. All tests are carried out at the 5 percent level. In the cross validation strategy the in-sample data set is split into ten sets each containing 50 data points. In order to reduce the problem of convergence to local minima we performed several runs from different starting values whenever we had to estimate a model. The pure computing time needed to perform the whole simulation study was about three month on a Pentium-120 computer.

6.2 Simulation Results

The results of the simulation study are given in the following tables. Column 1 contains the abbreviations for the different strategies. TESTS (W) and TESTS (T) denote sequences of hypotheses tests applying the techniques due to Teräsvirta/Lin/Granger (1993) and White (1989a). AIC and NIC stand for the strategies based on information criteria and CV for cross validation.

Column 2 reports how far the model chosen by the model selection strategy (MSS) deteriorates from the true model (TM) in terms of out-of-sample performance (MSPE). This is measured by the relative differences in the MSPE between the true model and the selected model calculated as:

$$\frac{\text{MSPE}_{\text{MSS}} - \text{MSE}_{\text{TM}}}{\text{MSE}_{\text{TM}}} \tag{11}$$

The numbers given in the tables are the averages over the thousand replications of the simulation study.

Column 3 shows the ranking of the strategies with regard to the out-of-sample MSPE. Columns 4 to 6 reveal the complexity of the model. Column 4 shows how often a certain number of hidden units has been allocated, columns 5 and 6 summarize the size of the networks by giving the average number of hidden units ($\#\beta$) and γ–weights ($\#\gamma$) that remained in the selected models.

The overall results of the simulation study are encouraging. They provide for a ranking that is similar in the different models. All in all the best and most stable results are obtained by the sequence of hypotheses tests employing the LM–test procedure of Teräsvirta/Lin/Granger (1993). This strategy leads to the best out-of-sample performances in the models M1 and M2 and is second best in model M3 where the three best model selection strategies have very similar performances. In model M1 strategy TESTS (T) achieves an out-of-sample error which only is 3.3% higher than the true out-of-sample error. This is a small error given that a model with the true network structure and estimated parameters already leads to an error that is 2.0% above the true error. In model M2 the TEST (T)–strategy produces an out-of-sample error that is only 2.8% worse than the true MSE. Compared to the out-of-sample performance of the other strategies the MSPE in model M3 is again relatively small, even if it is 30.9% worse than the true MSE. This relatively high deviation probably stems from the apparent complexity of the underlying function, especially in the tail areas of the independent variables' distribution, as shown in

Figure 6. Hence it is quite complicated for the network to arrive at a good approximation given the limited number of 500 observations and the noise level of 20 percent.

Strategies	Eq. (11)	Rank	1	2	3	4	5	$\#\beta$	$\#\gamma$
TESTS (T)	3.30%	1	0.0	87.5	12.5	0.0	0.0	2.1	7.8
TESTS (W)	6.09%	4	15.3	76.4	6.9	1.4	0.0	1.9	6.9
CV	3.32%	2	0.0	67.7	32.0	0.3	0.0	2.3	8.0
AIC	4.48%	3	17.2	81.5	1.3	0.0	0.0	1.8	6.2
NIC	15.76%	5	47.3	45.6	7.1	0.0	0.0	1.6	5.6

Table 1: Results of the model selection strategies for model M1.

Strategies	Eq. (11)	Rank	1	2	3	4	5	$\#\beta$	$\#\gamma$
TESTS (T)	2.8%	1	0.2	8.6	25.2	34.0	32.0	3.9	11.3
TESTS (W)	7.9%	2	19.8	23.2	23.3	19.2	14.5	2.9	8.1
CV	13.3%	3	50.2	35.6	10.9	2.3	1.0	1.7	4.4
AIC	14.1%	4	57.1	30.7	10.0	2.3	0.0	1.6	4.1
NIC	18.4%	5	91.3	8.3	0.4	0.0	0.0	1.1	2.6

Table 2: Results of the model selection strategies for model M2.

Strategies	Eq. (11)	Rank	1	2	3	4	5	$\#\beta$	$\#\gamma$
TESTS (T)	30.9%	2	0.0	0.0	71.1	27.9	1.0	3.3	7.3
TESTS (W)	31.3%	3	0.0	0.0	87.7	12.3	0.0	3.2	6.7
CV	53.6%	4	0.0	0.0	29.8	69.2	0.0	3.7	9.7
AIC	28.1%	1	0.0	0.0	82.9	17.0	0.1	3.2	7.5
NIC	1894.6%	5	58.2	41.0	0.7	0.1	0.0	1.4	3.5

Table 3: Results of the model selection strategies for model M3.

Of particular interest are the results from model M1 as they provide some evidence about size and power of the underlying test procedures. As all tests have been carried out on a significance level of 5 percent, an insignificant hidden unit should have been accepted in only 5 per cent of all cases. However, the TEST (T)–strategy allocates three hidden units for 12.5 per cent of all replications and thus does not keep to the chosen size of the LM–test. In order to further investigate this problem we performed a thousand LM–tests for the same underlying model M1 with an increased sample size of 1500 observations. We tested a model already owning two hidden units for an additional hidden unit. It turned out that the percentage of models where the third hidden unit was accepted reduced to 6.1%. This results suggest that the problem with the test size arose from the limited number of observations, as the test is an asymptotic one. On the other hand the power

of the LM-tests of Teräsvirta/Lin/Granger (1993) seems to be very high, since in none of the cases the true number of hidden units was underestimated.

For all three models the TESTS (T)–strategy delivers better results than the TESTS (W)–strategy. The latter strategy produces an out-of-sample error that is 6.09% worse than the true error in model M1, 7.9% worse in model M2 and 31.3% worse in model M3. In comparison with the TESTS (T)–strategy the TESTS (W)–strategy is handicapped by the random selection of the parameters needed to perform the LM–test. If the random selection of the parameters of the additional hidden unit is unfortunate, the LM–test can not recognize a correlation between the unexplained structure in the residuals and the signal of the additional hidden unit. In this case a neccessary additional hidden unit would not have been selected. For this reason the TESTS (W)–strategy tends to produce smaller models than the TESTS (T)–strategy. In our simulation the strategy allocates only one hidden unit in 15.3 percent of the cases which shows a considerably lower power of the LM-test due White (1989a) compared to the one of Teräsvirta/Lin/Granger (1993).

The out-of-sample performances of the AIC– and the CV–strategy differ considerably for different models. Both strategies perform very well in model M1, tend to underfit the function of model M2 and behave differently in model M3. It is interesting to note that the AIC strategy selects network architectures that are less complex than the architectures chosen by the CV–strategy. Which architecture is better seems to depend upon the true model structure to be approximated. Some models apparently bear to be approximated by too large a network without showing a deterioration in out-of-sample performance, whilst others do not. The relatively smooth surface of model M2 seems to allow for larger network architectures without showing a decreasing out-of-sample error. In model M3 the optimum number of hidden units given the data is probably three. It appears that a fourth hidden unit significantly contributes to a deterioration of the network approximation, so that the out-of-sample error increases.

For all three models the NIC–strategy leads to the worst results. This strategy is very reluctant to accept hidden units and choses the least complex network architectures for all three models. Thus the selected networks tend to considerably underfit the true model structures. This characteristic of the NIC-strategy becomes particularly apparent in model M3, where the out-of-sample MSPE is almost 2000% larger than the true model's MSE. The NIC–strategy only allocates 1 or 2 hidden units, which is far too small compared to the number of hidden units accepted by the other strategies. From a theoretical point of view the AIC-strategy should perform better than the NIC–strategy in model M1 and worse in models M2 and M3 as the NIC takes account of the misspecification of the networks. However, it seems that the estimates of the penalty term BA^{-1} given in section 3 penalizes extra parameters too strongly.

We found that a slight overparameterization of network models leads to lower out-of-sample errors than an underparametrization. A too small a model produces a relatively high bias, whereas the variance of too large a model does not increase so much. This may be caused by the hidden units transfer function. In opposition to e.g. polynomials the tanh-function apparently behaves more stable even in regions where only a few observations are available.

Comparing all strategies the simulation results show that the sequence of hypotheses tests employing a Taylor expansion leads to the most reliable results. If statistical model selection strategies are to be applied, we recommend this strategy. A further advantage of the sequence of hypotheses tests is that it needs the least amount of computing time, as decisions about the significance of parameters are drawn within the model and not by comparison between models.

7 Conclusion

In this paper we suggest different model selection strategies for neural networks which are based on statistical concepts. As in general neural networks are applied to nonlinear problems where little is known about the correct functional form, a statistical approach to model selection seems particularly important.

The building blocks of the model selection strategies are hypotheses tests, information criteria and cross validation. We discuss the applicability of these concepts for neural network specification and emphasize that care is needed due to the identification problem inherent in network models.

The proposed selection strategies account for this identification problem by combining a top down and a bottom up approach. In the bottom up part the number of hidden units, i.e. the general model complexity, is determined. By means of the subsequent top down step, irrelevant input connections are removed. Since the decisions taken in each step of the model building process are based on a clearly defined rule, model selection in neural networks becomes more comprehensible. This allows to arrive at the same network model when a study is repeated with the same data set.

In a Monte Carlo simulation the selection strategies based on different concepts are compared for three different models, including a network and two non-network models. The overall results are encouraging, as in most cases the strategies lead only to a small increase in the out-of-sample MSPEs compared to the MSE of the true model. It is shown that a sequence of hypotheses tests based on an LM-procedure due to Teräsvirta/Lin/Granger (1993) produces the most stable out-of-sample performance of the resulting networks while the test-procedure due to White (1989a) leads to a worse network specification. Strategies based on cross validation and information criteria are very accurate for some models though tend to overfit or underfit others. When information criteria are to be employed we recommend the use of the AIC instead of the NIC as the estimation of the penalty term results in too small network architectures with an unsatisfactory out-of-sample performance.

This study shows how statistical methods can be employed for the specification of neural networks. Although the simulation study presented is encouraging, it can just be a first step. Much experience has to be gained through further simulations with different underlying models, sample sizes and level to noise ratios. Moreover, applications with real world data sets will show in how far statistical methods can improve the model building process for neural networks. We hope that such methods will become a standard tool of the network practitioner.

References

Akaike H. (1973): *Information Theory and an Extension of the Maximum Likelihood Principle.* In Petrov B.N., Csaki F. (eds.): *Second International Symposium on Information Theory.* (Budapest: Akademiai Kiado), 267–281.

Akaike H. (1974): *A New Look at the Statistical Model Identification.* IEEE Transactions on Automatic Control, AC-19, 716–723.

Anders U. (1997): *Statistische neuronale Netze.* Vahlen Verlag.

Bishop C.M. (1995): *Neural Networks for Pattern Recognition.* Clarendon Press.

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

Davies R.B. (1977): *Hypothesis Testing when a Nuisance Parameter is present only under the Alternative.* Biometrika, 64, 247–254.

Davies R.B. (1987): *Hypothesis Testing when a Nuisance Parameter is present only under the Alternative.* Biometrika, 74, 33–34.

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

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

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

Judge G.G., Griffiths W.E., Hill R.C., Lütkepohl H., Lee T.-S., (1985): *The Theory and Practice of Econometrics, 2nd edition,* Wiley.

Kuan C.-M., White H. (1994): *Artificial Neural Networks: An Econometric Perspective.* Econometric Reviews, Vol 13, 1–91.

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

Moody J.E. (1992): *The Effective Number of Parameters: An Analysis of Generalization and Regularization in Nonlinear Learning Systems.* Advances in Neural Information Processing Systems, 4, 847–854.

Murata N., Yoshizawa S., Amari S. (1994): *A Criterion for Determining the Number of Parameters in an Artificial Neural Network Model.* IEEE Trans. Neural Networks, Vol 5, No 6, 865–872.

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

Reed R. (1993): *Pruning Algorithms — A Survey.* IEEE Transactions on Neural Networks, 4, 740–747.

Ripley B.D. (1993): *Statistical Aspects of Neural Networks.* In: Barndorff-Nielsen O.E., Lensen J.L., Kendall W.S. (eds): *Networks and Chaos - Statistical and Probabalistic Aspects.* Chapman and Hall, 40–123.

Ripley B.D. (1996): *Pattern Recognition and neural networks.* Cambridge University Press.

Sarle, W.S. (1994): *Neural Networks and Statistical Models.* Proceedings of the N19th Annual SAS Users Group International Conference, Cary, NC. SAS Institute Inc., 1538-1550.

Sarle, W.S. (1995): *Stopped Training and Other Remedies for Overfitting.* To appear in Proceedings of the 27th Symposium on the Interface.

Schwarz G. (1978): *Estimating the Dimension of a Model.* The Annals of Statistics, 6, 461–464.

Sin C.-Y., White H. (1996): *Information Criteria for Selecting Possibly Misspecified Parametric Models.* Journal of Econometrics, 71, 207–225.

Stinchcombe M.B., White H. (1995): *Consistent Specification Testing with Nuisance Parameters Present only under the Alternative.* University of Texas.

Stone M., (1974): *Cross Validation Choice and Assessment of Statistical Predictions.* Journal of the Royal Statistical Society, B, 36, 111–147.

Swanson N.R., White H. (1995): *A Model-Selection Approach to Assessing the Information in the Term Structure Using Linear Models and Artificial Neural Networks.* Journal of Business & Economic Statistics, Vol 13, No 3, 265–275.

Teräsvirta T., Lin C.-F., Granger C.W. (1993): *Power of the Neural Network Linearity Test.* Journal of Time Series Analysis, Vol 14, No 2, 209–220.

Weigend A.S., Huberman B.A., Rumelhart D.E. (1991): *Predicting Sunspots and Exchange Rates with Connectionist Networks.* In: Casdagli M., Eubank S.: *Nonlinear Modeling and Forecasting.* Addison–Wesley, 395–432.

White H. (1981): *Consequences and Detection of Misspecified Nonlinear Regression Models.* Journal of the American Statistical Association, Vol 76, No 374, 419–433.

White H. (1982): *Maximum Likelihood Estimation of Misspecified Models.* Econometrica, Vol 50, No 1, 1–25.

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

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

White H. (1994): *Estimation, inference and specification analysis.* Cambridge University Press.