ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 133-138

New criterion of identification in the multilayered perceptron modelling

Morgan Mangeas, Marie Cottrell*, Jian-Feng Yao

SAMOS, Université Paris I.
90, rue de Tolbiac , 75013 Paris, FRANCE

Abstract. This paper deals with multilayered perceptrons modelling for time series analysis. Based on recent results about the least-squares estimation for non-linear time series, we propose a complete and feasible statistical methodology for both parameter estimation (learning process) and model selection (architecture selection). In particular, we solve the parameter pruning problem for multilayer perceptron models with a stepwise search method by using a least squares BIC* criterion which is proved to be consistent.

1. Introduction

One of the most important applications of the multilayered perceptrons is time series modeling and forecasting. See for example (Weigend and Gershenfeld, 1994, [8] or Cottrell et al., 1995, [1]) for references. However, estimation and identification of such models often require sophisticated techniques and it is not easy to determine the suited architecture. Many papers deal with the techniques of *pruning* irrelevant parameters, mainly in the regression setting. See for example (Le Cun et al., 1990, [4], Moody, 1992, [6], Murata et al., 1994, [7], etc.). The present paper proposes theoretical results in the time series setting (valid in the regression setting too).

We consider a family of models called *Neural AutoRegressive model* (NAR), defined by :

$$\begin{aligned} Y_t &= f_W(Y_{t-1},\ldots,Y_{t-p}) + \varepsilon_t \\ &= \alpha_0 + \sum_{j=1}^{K} \alpha_j \phi\left(\sum_{i=1}^{p} \beta_{ij} Y_{t-i} + \beta_{0j}\right) + \varepsilon_t \end{aligned} \quad (1)$$

where $Y_{t-i}, i = 1,\ldots,p$ are lags of the series (Y_t), f_W denotes a function implemented by a multilayered perceptron with one output unit, p input linear units, K hidden units supplied with a sigmoid function ϕ. The variable ε_t is an i.i.d. noise, with mean 0 and constant variance σ^2, independent from the past of the series.

*email: Cottrell@univ-paris1.fr

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 133-138

In this work, we will take into account only real variables. But all the properties we present here can be straightforwardly expanded to the multi-dimensional case.

Let $W = \{(\alpha_j)_{0\leq j\leq K}, (\beta_{ij})_{0\leq i\leq p, 1\leq j\leq K}\}$ be the parameter vector. Given $T+p$ observations $(Y_{-p+1}, \ldots, Y_0, Y_1, \ldots, Y_T)$ of the series, we estimate W by minimizing the sum of squared residuals (the Error Function):

$$S_T(W) = \sum_{t=1}^{t=T} \left(Y_t - f_W(Y_{t-1}, \ldots, Y_{t-p})\right)^2 . \tag{2}$$

Let $\hat{W}_T = \text{Argmin} S_T(W)$ denote the least squares estimate of W. Its computation can be carried out with any minimization method. In this paper, we will not be concerned with the delicate minimization problem and $\hat{W}_T$ is assumed to be the true minimum of the Error Function $S_T(W)$. As the maximum likelihood estimator, the least squares estimator is a particular form of *minimum contrast estimator* (see Guyon, 1995, [2]).

2. Asymptotic results on the least squares estimation

Let us consider the NAR model defined by equation (1). We will denote the components of the parameter vector W by $W = (w_i)_{1\leq i\leq m}$ where $m = (p+2)K + 1$ and by W_0 its (unknown) true value. Recently, (Mangeas and Yao, 1996, [5]) have considered the asymptotic properties of the least squares estimator for general NAR processes. Let $Y^{(p)} = (Y_t^{(p)})_{t>0}$ be the *vector process*, defined by $Y_t^{(p)} = (Y_t, \ldots, Y_{t-p+1})$ for $t > 0$. The sequence $(Y_t^{(p)})$ is homogeneous Markov chain with state space $\mathcal{R}^p$. The vector $(y_1, \ldots, y_p) \in \mathcal{R}^p$ is denoted by $\tilde{y}$. The assumptions that will be made below ensure the stability of the chain $Y^{(p)}$. In particular, this chain will have unique invariant distribution μ_{W_0}. In (Mangeas and Yao, 1996, [5]), the authors prove :

Theorem 1 Strong consistency and asymptotic normality *For the model (1), with $\phi(x) = \tanh(x)$, assume the following:*

1. *$(\varepsilon_t)_{t>0}$ is a centered, i.i.d. sequence, independent of the initial states $(Y_{-p+1}, \ldots, Y_0)$, such that $E\,\varepsilon_t^6 < \infty$,*
2. *W belongs to a compact subset $\mathcal{W}$ of the m-dimensional Euclidean space $\mathcal{R}^m$, such that $W_0 \in \overset{\circ}{\mathcal{W}}$.*
3. *(Identifiability condition) For any W different from W_0, $f_W \neq f_{W_0}$ in the sense that there is a $\tilde{y} \in \mathcal{R}^p$ such that $f_W(\tilde{y}) \neq f_{W_0}(\tilde{y})$.*
4. *The following $m \times m$ matrix*

$$\Sigma_0 = \int_{\mathcal{R}^p} \left[\frac{\partial}{\partial w_i} f_W(\tilde{y}) \frac{\partial}{\partial w_j} f_W(\tilde{y})\right]_{1\leq i,j\leq m} \mu_{W_0}(d\tilde{y}), \tag{3}$$

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 133-138

is positive definite. Then,

(a) The least squares estimator $\hat{W}_T$ is strongly consistent, that is it converges almost surely to W_0 when T tends to $+\infty$.

(b) Independently of any initial distribution of the Markov chain $Y^{(p)}$, the term $\sqrt{T}\left[\hat{W}_T - W_0\right]$ converges in distribution to the multivariate Gaussian distribution $\mathcal{N}(0, \sigma^2\Sigma_0^{-1})$.

The residual variance σ^2 is in practice estimated by $\hat{\sigma}^2 = \frac{1}{T}S_T(\hat{W}_T)$, and the matrix Σ_0 by $\hat{\Sigma}_0 = \frac{1}{2T}\nabla^2 S_T(\hat{W}_T)$ which can also be approximated by

$$\frac{1}{T}\sum_t [\nabla f_{\hat{W}_T}(Y_t^{(p)})][\nabla f_{\hat{W}_T}(Y_t^{(p)})]'.$$

It is worthy to note that these conditions are very weak with respect to the usual normality assumptions.

3. Almost sure identification of a true model

Thus, by applying the existing results on model selection by penalized contrasts (Guyon, 1995, [2]), we establish an almost sure identification of a true model, when there are a finite number of possible models having a common dominant model.

More precisely, assume that we have a fixed bound M for all possible model dimensions. Thus let $\mathcal{W} \subset \mathcal{R}^{\mathcal{M}}$ and F_{max} be a dominant model, whose parameter vector is denoted $W_{\max} = (w_1, w_2, \ldots, w_M)$. Consider the finite family $\mathcal{F} = \{(w_1, w_2, \ldots, w_M)/\text{some components are set to } 0\}$, *respecting a set of constraints* linked to the interpretation of the components of W in the neural network. These restrictions are equivalent to giving a priority to the different ways for pruning weights (to first prune the weights between the input units and the hidden units).

For a $F \in \mathcal{F}$, sub-model of $F_{\max}$, we denote by $m(F)$ the number of its non null parameters, i.e. the dimension of the parameter vector W, and $\mathcal{W}_F$ the set of possibles values of W. The true model, which is a sub-model of $F_{\max}$, is denoted by F_0 and the true value of the parameter vector is W_0 with dimension $m(F_0)$.

Let $\hat{W}_{T,F}$ be the least squares estimator of W restricted to F,

$$\hat{W}_{T,F} = \text{Arg} \min_{W \in \mathcal{W}_F} S_T(W) \quad ,$$

and $S_T(F)$ (instead of $S_T(\hat{W}_{T,F})$) the associated minimum of the Error Function. Also let $(c(t))$ be a positive sequence of real numbers. The *penalized least-squares contrast* with penalization rate $(c(t))$ takes the form

$$\text{CWP}(T, F) = \frac{S_T(F)}{T} + \frac{c(T)}{T}m(F). \qquad (4)$$

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 133-138

Let $\hat{F}_T = \text{Arg}\min_{F \in \mathcal{F}} \text{CWP}(T, F)$ be the estimated model, which is the result of two successive minimizations for a fixed T: a minimization on a continuous space, to compute $\hat{W}_{T,F}$ and $S_T(F)$, and a minimization on a finite space, to compute $\hat{F}_T$.

With these definitions, the following result and its complete proof can be found in (Mangeas and Yao, 1996, [5]).

Theorem 2 *Assume that the conditions of theorem (1) hold. Suppose also the penalization rate $c(T)$ is such that*

$$\lim_T \frac{c(T)}{T} = 0, \quad \textit{and} \quad \liminf_T \frac{c(T)}{2 \ln \ln T} > \sigma^2 \frac{\Lambda}{\lambda} \tag{5}$$

where Λ (resp. λ) is the largest (resp. smallest) eigenvalue of the matrix Σ_0. Then, the pair $(\hat{F}_T, \hat{W}_{T,\hat{F}_T})$ converges a.s. to the true value (F_0, W_0).

From Theorem 2, we can now propose a an almost sure identification methodology to determine the *true model* within the set of the $F_{\max}$ sub-models:

Let the term γ be some positive constant[1]. A logarithmic penalization rate $c(t) = \gamma \ln t$ clearly meets the above conditions (5). Taking such a penalization rate yields the following *least squares* BIC* *criterion* for model selection:

$$\text{BIC}^* = \text{BIC}^*(T, F) = \frac{S_T(F)}{T} + \gamma \frac{\ln T}{T} m(F) \tag{6}$$

We will use this criterion in the sequel. One should note the difference between BIC* and the usual BIC criterion: $\text{BIC} = \ln \frac{S_T(F)}{T} + \frac{\ln T}{T} m(F)$. They have both a logarithmic penalization rate but their first terms differ since the BIC criterion is an approximation based on maximum likelihood estimation.

Let $\sigma^2_{F_{\max}}$ be the residual variance associated to $F_{\max}$. We remind also that $W_{\max} = (w_1, w_2, \ldots, w_M)$ denotes the associated dominant parameter vector. Theoretically, in order to estimate the true model, we would have to exhaustively explore a finite family and compute BIC* for all sub-models $F \in \mathcal{F}$. But the number of these sub-models is exponentially large (as 2^M) and it is impossible to do it in practice. So, as in linear regression analysis, we propose a Statistical Stepwise Method (SSM) to guide the search in $\mathcal{F}$. Such a descending strategy is based on the asymptotic normality of the estimator $\hat{W}_T$. See [1] for previous presentations of the SSM algorithm, with several examples. The SSM pruning method is related to the OBD algorithm defined by (Le Cun et al., 1990, [4]), because they choose in the same way the next parameter to be candidate for pruning. But their algorithm does not provide any stopping criterion, thus it needs a performance estimate on a set of external data.

Using the results about the almost sure identification model, we have a theoritical stopping criterion: the BIC* criterion. The principle is to stop the deletion as soon as the criterion BIC* increases. This the main difference between this new strategy and the one described in [1].

[1]The constant γ have in practice the same order of magnitude as the variance σ^2. So the criterion BIC* does not depend on the scale of the series terms.

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 133-138

4. Computer-generated example

In order to test the accuracy of the BIC* criterion and of the identification strategy described section 3, we investigate a simulation experiment. The true architecture as well as the true parameter vector W_0 and the associated noise (ε_t) are known. The choosen model (see also figure 1) is: $X_t = \tan(-0.5X_{t-1} - 1.5X_{t-3} + 0.5) + \tan(X_{t-3} - 0.5) + 0.5 + \varepsilon_t$. The goal is to determine if we can retrieve the true architecture within a set of over and sub connected architectures. Since this simulation is not that complex, we can compare the exhaustive search with the SSM methodology described previously. The search is performed over the set of the sub-architectures of the dominant model described figure 2.

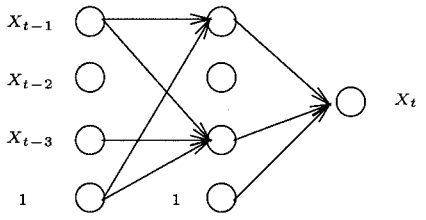

Figure 1: *True architecture*

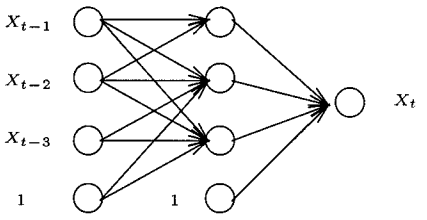

Figure 2: *Dominant Architecture.*

We generate independantly 50 sequences of 1000 points. The true architecture, with 8 connexions, is shown fig 1. The associated noise (ε_t) is an i.i.d. Gaussian sequence with $\sigma_0^2 = 0.1$. The used dominant model, with 16 connexions, is shown fig 2. Concerning the BIC* criterion, we set $\gamma = \sigma_0^2 = 0.1$. In order to avoid local minima, the parameter estimation is computed by taking the best one among 10 independant runs.

Final architecture	Percentage over 50 runs
T	0.73
A	0.12
B	0.10

Table 1: *Exhaustive search performances.*

Final architecture	Percentage over 50 runs
T	0.62
A	0.22
B	0.16

Table 2: *Statistical Stepwise Method performances.*

Table 1 shows the three "best" architectures: T,A,B (minimizing the BIC* criterion) selected by the exhaustive search over 50 runs. As expected, the true architecture T appears 73% of the time. The other winners A et B are equivalent to the architecture T but with respectively one connexion more and

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 133-138

one connexion off. They appear respectively 12% and 10% of the time. Table 2 shows the three best architectures selected by SSM over 50 runs. One can note that this strategy yield exactly the same three architectures. T is found in 62% of the runs.

5. Conclusion

Currently we are working to extend the main results of the paper in three directions. First, we want to add exogeneous variables as inputs in the network, and consider a NARX model. Second, we introduce lags of the error in order to use a non linear neural "ARMA" model. Third, we deal with multi-dimensional time series leading to perceptrons with multiple output units. In this multi-dimensional case, we minimize the generalized sum of squares, weighted by the inverse of the variance-covariance matrix of the data vector.

References

[1] M. Cottrell, B. Girard, Y. Girard, M. Mangeas and C.Muller, "Neural modeling for time series : a statistical stepwise method for weight elimination", *IEEE Tr. on Neural Networks*, Vol. No. 6, 1355-1364, 1995.

[2] X.Guyon, *Random Fields on a Network - Modeling, Statistics and Applications*, Springer-Verlag, 1995.

[3] C. Jutten and R. Chentouf, "A new scheme for incremental learning", *Neural Processing Letters*, Vol. 2, No 1, 1-4, 1995.

[4] Y. Le Cun, J.S. Denker and S.A. Solla, "Optimal Brain Damage", in: Advances in Neural Information Processing Sust. II, ed. D.S. Touretzsky, Morgan Kaufman, 598-605, 1990.

[5] M.Mangeas and J.Yao, "Sur l'estimateur des moindres carrés des modèles auto-regressifs fonctionnels", Tech. Rep. No. 53, Université Paris 1, 1996.

[6] J.Moody, "The effective number of parameters: An Analysis of Generalization and Regularization in Nonlinear Learning Sysytems", in *Advances in neural Information Processing Systems 4*, Morgan Kaufman Publishers, 1992.

[7] N.Murata, S. Yoshizawa and S.Amari, "Network Information Criterion-Determining the Number of Hidden Units for an Artificial Neural Network Model", *IEEE Trans. on Neural Networks*, Vol. 5, No. 6, 865-872, 1994.

[8] A.S.Weigend and N.A.Gershenfeld, *Time Series Prediction*, Proceedings Volume XV, Santa Fe Institute, Addison-Wesley, 1994.