On the Consistency of the Blocked Neural Network Estimator in Time Series Analysis

A. Sarishvili
alex.sarishvili@itwm.fraunhofer.de
Institute of Industrial Mathematics, Kaiserslautern, Germany

Ch. Andersson
christina.andersson@esi.oru.se
Department of Economy, Statistics, and Computer Science, University of Örebro, Sweden

J. Franke
franke@mathematik.uni-kl.de
Department of Mathematics, University of Kaiserslautern, Germany

G. Kroisandt
gerald.kroisandt@itwm.fraunhofer.de
Institute of Industrial Mathematics, Kaiserslautern, Germany

We describe a nonlinear regression problem, where the regression functions have an additive structure and the dependent variable is a one-dimensional time series. Multivariate time series with unknown time delay operators are used as independent variables. By fitting a feedforward neural network with block structure to the data, we estimated the additive regression function and, parallel to this, the time lags. We present the consistency proof of neural network weights estimator and the time lag estimator independently from each other. In the practical part of the article, we present the useful feature of blocked neural networks to estimate the relevance measures of each input variable in a simple way. Furthermore, we propose an approach to solve the well-known variable selection problem for the class of nonlinear multivariate beta-mixing time series models considered here. Finally, we apply the methodology to an artificial example.

1 Introduction

Let $Y_t \in \mathbb{R}^d$, $t = 1, \ldots, N$ be a $d$-dimensional time series and let the $i$th component $y^i_t \in \mathbb{R}$ with $i = 1, \ldots, d$ and $t = 1, \ldots, N$ have the representation

$$y^i_t = m_{i1}(y^1_{t-\tau_1}) + m_{i2}(y^2_{t-\tau_2}) + \ldots + m_{id}(y^d_{t-\tau_d}) + \epsilon_t, \quad (1.1)$$

Neural Computation 18, 2568–2581 (2006) © 2006 Massachusetts Institute of Technology
where $\epsilon_t$ independently and identically (i.i.d). $N(0, \sigma^2)$-distributed with finite variance and $l_{ij} \in \{1, 2, \ldots, A_i\}$ is the time lag with $A_i \in \mathbb{N}$. The function $m_{ij}$ describes the relationship between the two components $y^j_t$ and $y^i_t$.

In this article, we estimate $m_{ij}$ nonparametrically by fitting feedforward neural networks with block structure to the data for each possible lag vector $L^i = (l_{1i}, \ldots, l_{id})'$. Afterward, we minimize the average one-step forecasting error of the optimal network with respect to the lag vector $L^i$. This gives us an estimate of the network parameters, together with an estimate for the lag vector $L^i$. Furthermore, we compute a measure of relevancy of component $j$ for component $i$, which helps us to identify the significant independent variables.

2 Neural Network Lag-Dependent Model Estimation

Our objective is to introduce a neural network–based method to determine an optimal subset of the stochastic regressors to fit the underlying regression model. For simplicity, we introduce the function $g$ for the true underlying regression function:

$$g(Y_{t-L^i}) = m_{i1}(y^1_{t-l_{1i}}) + m_{i2}(y^2_{t-l_{2i}}) + \ldots + m_{id}(y^d_{t-l_{id}}).$$

If we take into account that the composite function $g(Y_{t-L^i})$ is in particular a result of the addition of maximal $d$ functions each of one real variable, then the neural network without “nonparallel” connections seems to have the right architecture to estimate parameters from lag dependent models. Figure 1 shows a neural network with $d$ inputs, one output, and no non-parallel connections (a so-called blocked neural network). Each neuron in the hidden layer accepts only one variable as input apart from the constant $y^0 = 1$, also known as the bias of the neural network.

The output of the blocked neural network shows the following equation:

$$y^j_t = f_{nn}(y^1_t, \ldots, y^d_t, \Theta)$$

$$= b_0 + \sum_{j=1}^{H(1)} \beta_j \phi(y^1_t \gamma_j + b_j) + \ldots + \sum_{j=H(d-1)+1}^{H(d)} \beta_j \phi(y^d_t \gamma_j + b_j), \quad (2.1)$$

where the difference $H(j) - H(j - 1)$ for $j = 2, \ldots, d$ is the number of neurons in block $j$, and $H(d)$ is the total number of neurons in the hidden layer. $\beta_j$ refers to the weights from the hidden layer to the output layer, $\gamma_j$ to the weights from the input to the hidden layer, the $b_j$ is the constant term, and $\Theta_{H}$ is the set of all these $3 \cdot H(d) \cdot d + 1$ neural network weights. $\phi(y) = \frac{e^{\gamma y} - 1}{e^{\gamma y} + 1}$ is the sigmoid neuron activation function.
The small modification of the theorem presented in Funahashi (1989) guarantees that such a neural network is able to estimate the composite function $g(Y)$.

**Theorem 1.** Let $\phi(y)$ be a nonconstant, bounded, and monotonic increasing continuous function. Let $K$ be a compact subset of $\mathbb{R}^d$ and fix an integer $k \geq 1$. Then any continuous mapping $F : K \rightarrow \mathbb{R}$ with $F(y^1, \ldots, y^d) = f_1(y^1) + \ldots + f_d(y^d)$, which is a composition of functions $f_1, \ldots, f_d$, where $f_i : D(f_i) \rightarrow \mathbb{R}$, $i = 1, \ldots, d$ continuous and $K \subset D(f_i)$, can be approximated in the sense of the uniform topology on $K$ by input-output mappings of $k$ hidden-layer networks without nonparallel connections whose output functions for the hidden layers are $\phi(y)$ and for the input and output layers are linear.

### 3 Estimators for the Lag-Dependent Model

In this section we determine an estimate of the neural network parameters $\theta_{N,L}^i$ and lag vector $\hat{L}^i = (\hat{l}_{i1}, \hat{l}_{i2}, \ldots, \hat{l}_{id})'$, $i = 1, \ldots, d$ and show their consistency. In the following, we fix the coordinate $i$, that is, we omit the dependence on $i$ of the neural network weights and the estimated lags.

First, we choose a so-called lag horizon $\hat{A}^i \in \mathbb{N}$, that is, the maximal time delay, which is reasonable to take into consideration. This choice is based on the experience of the observer. The lag horizon $\hat{A}^i$ and the dimension of the time series, $d$, are both discrete variables. This fact allows us to find the
The Blocked Neural Network 2571

global minimum of the error function by training the neural network \( (A^i)^d \) times as follows:

\[ \hat{\theta}_{N,L^i} = \arg\min_{\theta \in \Theta_H} (D_N(L^i, \theta)), \]

where \( D_N(L^i, \theta) \) denotes the criterion function

\[ D_N(L^i, \theta) = \frac{1}{N} \sum_{t=1}^{N} [y_t^i - f_m(Y_{t-L^i}, \theta)]^2, \]

where \( Y_{t-L^i} = (y_{t-l_1}^1, \ldots, y_{t-l_d}^d) \). Afterward, we estimate the time lags \( \hat{L}^i \) by

\[ \hat{L}^i = \arg\min_{L^i \in \Pi(A^i, d)} (D_N(L^i, \hat{\theta}_{N,L^i})), \]

where \( \Pi(A^i, d) = \{1, \ldots, A^i\} \times \ldots \times \{1, \ldots, A^i\} = \{1, \ldots, A^i\}^d \) is the set of all possible time lags.

We furthermore introduce the notation \( D_0(L^i, \theta) \) as

\[ D_0(L^i, \theta) = E \left[ y_t^i - f_m(Y_{t-L^i}, \theta) \right]^2. \]

### 3.1 Consistency of the Network Weights Estimator.

The consistency is well known for independent observations \( Y_t \) (see White, 1989) if we fix \( L^i \). Since we have several possibilities for choosing \( L^i \), we must formulate the restriction in White (1989) as follows:

**Assumption 1.** There exists a separable subset \( \Theta_H(c) \) of \( \Theta_H \) such that for any \( L^i \in \Pi(A^i, d) \), the criterion function \( D_0(L^i, \theta) \) has a unique global minimum \( \theta_{0,L^i} \) in \( \Theta_H(c) = \{ \theta \in \Theta_H; ||\theta - \theta_{0,L^i}|| \leq c \} \) for any \( c > 0 \):

\[ \theta_{0,L^i} = \arg\min_{\theta \in \Theta_H} (D_0(L^i, \theta)) = \arg\min_{\theta \in \Theta_H(c)} (D_0(L^i, \theta)). \]

**Remark 1.** An alternative to the strategy of reducing the uniform convergence problem to a compact subset of the parameter space is to extend the neural network training procedure to a compactification of the parameter space. This way to show the classical consistency proof of time lag and weight parameters in a blocked neural network–based, lag-dependent model will not be considered here.
Remark 2. As the neural network function \( f_{nn}(x, \theta) \) depends on \( \theta \) in a very smooth and simple manner, assumption 1 is not a severe restriction, but it excludes only a few degenerate constellations of the random process and the network function. \( \Theta_H(c) \) can be chosen independent of \( L^i \) as we consider only finitely many values of \( L^i \). The only problem is if the neurons are not interchangeable, that is, in the definition of \( \Theta \), we must introduce some restriction like

\[
\gamma_j \leq \gamma_{j+1} \quad \text{where} \quad H(i) \leq j < H(i+1) \quad \text{for some} \quad i.
\]

To come from the independently and identically distributed (i.i.d.) case to a time series, we must impose some restrictions on the time series. We must mainly assume that there is ergodicity and some mixing condition:

Assumption 2.

1. \( Y_{t-L^i} \) and \( \epsilon_t \) are independent of each other and the \((d+1)\)-dimensional process \((Y_{t-L^i}, \epsilon_t)'\) is strictly stationary and ergodic.

2. \( g \) is bounded and \( Y_t \) is \( \beta \)-mixing with exponentially decaying mixing coefficients (see Douc&-an, 1994, and Bosq, 1996, for the definition of \( \beta \)-mixing and its equivalence to geometric ergodicity).

Remark 3. The mixing assumption is not unusual for nonlinear autoregressive processes. Franke, kreiss, Mammen, & Neumann, (2000) gives simple sufficient conditions on the autoregression function \( g \) such that this assumption is satisfied.

If we want to use other innovations \( \epsilon \), which are not normally distributed, then we must be able to use the Bernstein’s inequality for a mixing time series:

Assumption 3. The innovations \( \epsilon_t \) are i.i.d. zero mean random variables, for which all moments \( E|\epsilon_t|^n, \ n = 1, 2, \ldots \), are finite and

\[
E|\epsilon_t|^n \leq c^{n-1}n! E\epsilon_t^2 < +\infty,
\]

for \( n = 3, 4, \ldots \) and some \( c > 0 \).

In order to be able to use the consistency proof as in Pötscher and Prucha (1997), we must show the following:

Theorem 2. Under assumptions 1, 2, and 3,

\[
\sup_{\theta \in \Theta_H(c)} |D_N(L^i, \theta) - D_0(L^i, \theta)| \to 0
\]
in probability as $N \to \infty$ for all $L^i \in \Pi(A^i, d)$.

**Proof.** Using Bernstein’s inequality for a mixing time series (see Doukhan, 1994) in conjunction with a simple truncation argument (which is possible by assumption 3), we have

$$
\sup_{\theta \in \Theta_H(c)} \{ pr \left( |\bar{D}_N(L^i, \theta) - D_0(L^i, \theta)| > \delta \right) \} \to 0, \text{ as } N \to \infty \tag{3.1}
$$

for all $\delta > 0$ and $L^i \in \Pi(A^i, d)$ (compare Franke & Neumann, 2000, for a similar argument). So the only problem is to interchange the probability and the supremum.

Let $\alpha > 0$ be a small real number to be chosen later. Let $\theta_k, k = 1, \ldots, h(\alpha)$, be a set of vectors in $\Theta_H(c)$ such that for all $\theta \in \Theta_H(c)$, there is $\theta_k$ such that $||\theta - \theta_k|| \leq \alpha$. Let

$$
\Delta_N(\theta) = \bar{D}_N(L^i, \theta) - D_0(L^i, \theta).
$$

Then for arbitrary $\delta > 0$,

$$
pr \left( \sup_{\theta \in \Theta_H(c)} |\Delta_N(\theta)| > \delta \right) \leq pr \left( \sup_{\theta \in \Theta_H(c)} |\Delta_N(\theta)| > \delta, \sup_{||\theta - \theta_k|| \leq \alpha} |\Delta_N(\theta) - \Delta_N(\theta_k)| \right)
$$

$$
\leq \frac{\delta}{2} \forall k = 1, \ldots, h(\alpha)
$$

$$
+ pr \left( \sup_{||\theta - \theta_k|| \leq \alpha} |\Delta_N(\theta) - \Delta_N(\theta_k)| > \frac{\delta}{2} \right), \text{ for at least one } k \in \{1, \ldots, h(\alpha)\} \tag{3.2}
$$

The first term on the right-hand side of equation 3.2, is bounded from above by

$$
pr \left( \sup_{k \leq h(\alpha)} |\Delta_N(\theta_k)| > \frac{\delta}{2} \right) \leq \sum_{k=1}^{h(\alpha)} pr \left( |\Delta_N(\theta_k)| > \frac{\delta}{2} \right)
$$

$$
\leq h(\alpha) \cdot \sup_{\theta \in \Theta_H(c)} \left\{ pr \left( |\Delta_N(\theta)| > \frac{\delta}{2} \right) \right\} \to 0
$$

for $N \to \infty$ by equation 3.1.
For the second term on the right-hand side of equation 3.2, we have

\[
|\Delta_N(\theta) - \Delta_N(\theta_k)| \leq \frac{1}{N} \sum_{t=1}^{N} \left| \left( y_t - f_{mn}(Y_{t-L^i}, \theta) \right)^2 - \left( y_t - f_{mn}(Y_{t-L^i}, \theta_k) \right)^2 \right| + \left| E \left[ y_t - f_{mn}(Y_{t-L^i}, \theta) \right]^2 - E \left[ y_t - f_{mn}(Y_{t-L^i}, \theta_k) \right]^2 \right|.
\]

Using the particular form of the network function \( f_{mn} \), we have that the derivative of \( f_{mn} \) with regard to one of the parameters \( \beta_i \) is bounded as the activation function \( \phi(\cdot) \) is bounded. The derivative with regard to one of the parameters \( \gamma_i \) or \( b_j \) is bounded because the derivative of our activation function is also bounded and the parameters \( \beta_i \) are out of a compact set. Using the mean value theorem, we obtain

\[
\left| E \left[ y_t - f_{mn}(Y_{t-L^i}, \theta) \right]^2 - E \left[ y_t - f_{mn}(Y_{t-L^i}, \theta_k) \right]^2 \right| \leq b \cdot ||\theta - \theta_k||,
\]

for a suitable constant \( b \) that is independent of \( \delta \) and \( \alpha \). Analogously,

\[
\frac{1}{N} \sum_{t=1}^{N} \left| \left( y_t - f_{mn}(Y_{t-L^i}, \theta) \right)^2 - \left( y_t - f_{mn}(Y_{t-L^i}, \theta_k) \right)^2 \right| \leq B_N \cdot ||\theta - \theta_k||,
\]

where \( B_N \) is a nonnegative random variable with \( B_N \rightarrow b \) in probability for \( N \rightarrow \infty \) by the law of large numbers for mixing time series.

Therefore, the second term on the right-hand side of equation 3.2 is bounded by

\[
pr \left\{ \sup_{||\theta - \theta_k|| \leq \alpha} (B_N + b) \cdot ||\theta - \theta_k|| > \frac{\delta}{2} \quad \text{for at least one} \quad k = 1, \ldots, h(\alpha) \right\} \leq pr \left( \alpha \cdot (B_N + b) > \frac{\delta}{2} \right) = pr \left( B_N - b > \frac{\delta}{2 \cdot \alpha} - 2 \cdot b \right) \rightarrow 0,
\]

if we choose \( \alpha \) small enough, such that \( \frac{\delta}{2 \alpha} - 2 \cdot b > 0 \).

**Corollary 1.** From theorem 2 and assumption 1, it follows immediately that \( \hat{\theta}_{N,L^i} \) is a uniquely identifiable sequence of minimizers of \( \bar{D}_N^i(L^i, \theta) \) for given \( L^i \).

Consistency of the least squared estimator \( \hat{\theta}_{N,L^i} \) for \( \theta_{0,L^i} \) can now be inferred from the following lemma, applied to the compact subset \( \Theta_{H(c)} \):
Lemma 1. Let $D_N, \bar{D}_N : R^d \times \Theta_1(c) \rightarrow R$ be two arbitrary sequences of functions such that in probability,

$$\sup_{\theta \in \Theta_1(c)} |\bar{D}_N(L^i, \theta) - D_0(L^i, \theta)| \rightarrow 0$$

as $N \rightarrow \infty$.

Let $\hat{\theta}_{N,L^i}$ be an unique identifiable sequence of minimizers of $\bar{D}_N(L^i, \theta)$. Then for any sequence $\bar{\theta}_{N,L^i}$ such that

$$D_0(L^i, \bar{\theta}_{N,L^i}) = \inf_{\theta \in \Theta_1(c)} D_0(L^i, \theta),$$

holds, we have $||\hat{\theta}_{N,L^i} - \bar{\theta}_{N,L^i}|| \rightarrow 0$ in probability as $N \rightarrow \infty$.

The proof of this lemma is given in Pötscher and Prucha (1997).

3.2 Consistency of the Lag Estimator $\hat{L}^i$. We need a similar assumption for the lag vector as assumption 1, but the space $\Pi(A^t, d)$ is already compact:

Assumption 4. There is a unique $L^i_0$ such that with $\theta_0 \equiv \theta_{0,L^i_0}$ the pair $(L^i_0, \theta_0)$ satisfies

$$(L^i_0, \theta_0) = \arg\min_{\theta \in \Theta_1, L^i \in \Pi(A^t, d)} (D_0(L^i, \theta))$$

and for every $L^i \neq L^i_0$: $D_0(L^i, \theta_{0,L^i}) > D_0(L^i_0, \theta_0)$.

Remark 4. Assumption 4 follows from

$$\text{pr} \left( g(y^1_{t-l_{i1}}, \ldots, y^d_{t-l_{id}}) \neq g(y^1_{t-\lambda_{i1}}, \ldots, y^d_{t-\lambda_{id}}) \right) > 0,$$

for all $(l_{i1}, \ldots, l_{id}) \neq (\lambda_{i1}, \ldots, \lambda_{id})$. This is automatically satisfied if the innovations $\epsilon_t$ have a density, which is positive everywhere.

Now we can formulate the theorem about the consistency of the lag estimator:

Theorem 3.

$$\hat{L}^i \overset{N \rightarrow \infty}{\longrightarrow} L^i_0, \text{ in probability}$$

where

$$\hat{L}^i = \arg\min_{L^i \in \Pi(A^t, d)} (D_N(L^i, \hat{\theta}_{N,L^i})).$$
\[ L_0^i = \arg\min_{L^i \in \Pi(A^i, d)} \left( D_0(L^i, \theta_{0,L^i}) \right). \]

**Proof.** For a fixed \( L^i \), we know that the neural network weights are consistently estimated (see Lemma 1), that is,

\[
\hat{\theta}_{N,L^i} \xrightarrow{N \to \infty} \theta_{0,L^i} = \arg\min_{\theta \in \Theta(c)} \left\{ E \left[ \left( Y_{t-L^i} - f_{nn}(Y_{t-L^i}, \theta) \right)^2 \right] \right\},
\]

and therefore we conclude that for the continuous function \( D_N \),

\[
D_N(L^i, \hat{\theta}_{N,L^i}) \xrightarrow{N \to \infty} D_0(L^i, \theta_{0,L^i}) \text{ in probability.}
\]

Since the minimum \( L_0^i \) of \( D_0 \) is unique and \( \Pi(A^i, d) \) has only finitely many elements,

\[
\hat{L}_i = \arg\min_{L^i \in \Pi(A^i, d)} \left[ D_N(L^i, \hat{\theta}_{N,L^i}) \right] \xrightarrow{N \to \infty} \arg\min_{L^i \in \Pi(A^i, d)} \left[ D_0(L^i, \theta_{0,L^i}) \right] = L_0^i \text{ in probability.}
\]

### 4 Example

Consider the following nonlinear additive dynamical system, which generates a two-dimensional time series,

\[
y_1^t = m_{11}(y_{1,t-2}^1) + m_{12}(y_{1,t-4}^1) + \epsilon_t, \quad t \in [1, \ldots, 240]
\]

\[
y_2^t = m_{22}(y_{1,t-1}^2) + m_{21}(y_{1,t-2}^1) + \epsilon_t, \quad t \in [1, \ldots, 240],
\]

where \( \epsilon_t \) is i.i. \( \text{N}(0,0.3) \)-distributed. We choose the following regression functions:

\[
m_{11}(x) = \frac{1}{\exp(x)}, \quad m_{12}(x) = 1.7 \cos(x),
\]

\[
m_{22}(x) = \frac{1}{0.2 \exp(x)}, \quad m_{21}(x) = 2 \cos(x) + 0.3 \sin(x).
\]

The simulation of this time series with constant starting values is shown in Figure 2.
We train a blocked neural network with two neurons in each block that is presented in Figure 3. We choose the first variable as output and estimate simultaneously the prediction mean squared error (PMSE) on the validation data set (15% of the whole data set) shown in Figure 4.

The network function according to the definition of blocked neural network function in equation 2.1 then has the form

$$f_{nn}(Y, \theta) = b_0 + \sum_{j=1}^{2} \beta_j \phi(b_j + \gamma_j y^1) + \sum_{j=3}^{4} \beta_j \phi(b_j + \gamma_j y^2),$$

where $y^j$ is the $j$th coordinate of $Y$. If $\phi$ is differentiable, as we always
assume, the partial derivatives (PD) with regard to $y_1$ are calculated as follows:

$$\frac{\partial f_{nn}}{\partial y_1}(Y, \theta) = \sum_{j=1}^{2} \beta_j \phi'(b_j + y_1 y_j) y_j,$$
and analogously for the second component, \( y^2 \). Because of nonlinearity and the additive topology of the neural network, this measure of relevancy is a function of the corresponding component only and therefore could be plotted for further analysis. The plots of PD versus the input variable (PD plots) can be used to analyze the impact of corresponding input variables. For similar considerations of relevancy measures in the context of the fully
connected feedforward neural networks, see Refenes, Zapranis, and Utans (1996) and Sarishvili (2002). A large PD value indicates that the influence of the related input variable is strong for the considered output value. Small changes of the input value will already cause large changes on the output value. The PD plots of each of the input variables to the output variable are showed in Figures 5 and 6. The time delay values of the model were correctly estimated.
The same plots were generated with the second variable as output variable. The performance plot of the blocked neural network is shown in Figure 7. The PD plots are shown in Figures 8 and 9.

The PD and time delay value of \( y^2 \) to \( y^2 \) were poorly predicted because of the small true PD values of \( y^2 \) to \( y^2 \), as shown in Figure 9. The training algorithm of the blocked neural network was the Levenberg-Marquardt algorithm.

5 Conclusion

We have used the framework of nonparametric regression assuming an additive, block structure of the regression functions to model a multivariate time series. Using a blocked feedforward neural network, we showed how both the regression functions and the vector of time lags can be estimated. Finally, we calculated a measure of relevancy.

The major advantage of blocked neural networks compared to other nonlinear time-series approximators is the interpretability of the related partial derivatives and relevance measures with respect to the impact of the single input variables. Plots of the partial derivatives can be used to estimate the functional influence of each input variable with respect to the output.

References


Received March 29, 2005; accepted April 5, 2006.