Variable selection in neural network regression models with dependent data: a subsampling approach

Michele La Rocca*, Cira Perna

Department of Economics and Statistics, University of Salerno, via Ponte Don Melillo, Fisciano, Salerno 84084, Italy

Received 1 March 2003; received in revised form 1 January 2004; accepted 15 January 2004

Abstract

The problem of variable selection in neural network regression models with dependent data is considered. In this framework, a test procedure based on the introduction of a measure for the variable relevance to the model is discussed. The main difficulty in using this procedure is related to the asymptotic distribution of the test statistic which is not one of the familiar tabulated distributions. Moreover, it depends on matrices which are very difficult to estimate because of their complex structure. To overcome these analytical issues and to get a consistent approximation for the sampling distribution of the statistic involved, a subsampling scheme is proposed. The procedure, which takes explicitly into account the dependence structure of the data, will be justified from an asymptotic point of view and evaluated in finite samples by a small Monte Carlo study.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Artificial neural networks; Dependent data; Subsampling

1. Introduction

Over the past decade artificial neural networks have become increasingly widely used in a variety of statistical problems but they are not yet supported by the rich collection of specification and diagnostic tests usually employed in statistical and econometric
modelling. In a regression framework, a key issue is variable selection, to avoid omission of relevant variables or inclusion of irrelevant ones. In addition, when dealing with dependent data a lag selection procedure is also needed. Due to the black-box nature of the network, this problem cannot be faced focusing on single weights since it is difficult, if not impossible, to interpret them. Testing whether the weights are significantly different from zero could be misleading since a given approximation accuracy can be obtained with different network topologies. Moreover, this approach can be inadequate in testing the overall significance of an explanatory variable. Hence, we focus on statistical procedures based on the introduction of a measure for the variable relevance to the model, which can be used as a basis for a formal statistical test.

The aim of the paper is twofold. Firstly, we suggest a test procedure to select a proper set of input variables and we derive its asymptotic distribution by extending some results available for the iid framework to the case of dependent data. Secondly, we propose to use a subsampling scheme, which takes explicitly into account the dependence structure of the data, to get an alternative approximation for the sampling distribution of the test statistic. The use of a resampling scheme is necessary since the asymptotic distribution of the test statistic is not one of the familiar tabulated distributions and it depends on matrices which are difficult to estimate because of their complex structure. The proposed subsampling test procedure will be justified from an asymptotic point of view, by proving its consistency, and evaluated in finite samples by a small Monte Carlo study.

The paper is organized as follows. In Section 2 we describe the structure of the data generating process and the neural network model employed. In Section 3 we discuss the relevance measure approach to variable selection in neural network regression models with dependent data. In Section 4 we derive the asymptotic distribution of a proper class of test statistics. In Section 5 we introduce a subsampling scheme to get an approximation for its sampling distribution and we prove a consistency result. In Section 6 in order to evaluate the performance of the proposed procedure for finite samples, we discuss the results of a small Monte Carlo study. Some concluding remarks will close the paper.

2. The data generating process and the neural network model

Let \( \{Y_t\}, \ t \in \{1, \ldots, T\} \), a time series modeled as \( Y_t = g(X_t) + \epsilon_t \), where \( g(\cdot) \) is a continuously differentiable function defined on a compact subset \( \mathcal{X} \) of \( \mathbb{R}^d \) and \( X_t = (X_{1t}, \ldots, X_{dt})' \) is a vector of \( d \) random variables possibly including explanatory variables, lagged explanatory variables and lagged values of \( Y_t \).

We will assume that the following framework holds.

**Assumption A (Data generating process).** (i) \( \mathbb{E}(\epsilon_t | X_t) = 0, \ \forall t \).

(ii) Let \( Z_t = \{(Y_t, X_t)\}' \) with \( X_t \) bounded and \( \|Y_t\|_p < \infty \) where \( \| \cdot \|_p = (\mathbb{E} | \cdot |^p)^{1/p} \). \( Z_t \) is a stationary, \( \alpha \)-mixing sequence on a complete probability space \( (\Omega, \mathcal{F}, \mathbb{P}) \) with \( \alpha(m) \) of size \( -p/(p - 2) \) \( p \geq 4 \).
Condition A(i) clearly implies that
\[ E[Y_t | X_t] = g(X_t), \forall t. \]
This data generating process is quite general; it includes as particular cases regression models with dependent errors and with lagged variables and nonlinear autoregressive models for time series.

The function \( g \) can be approximated with a single hidden layer feed-forward neural network (Hornik et al., 1989; White, 1989) of the form
\[
f(x, \theta) = \sum_{k=1}^{r} c_k \phi \left( \sum_{j=1}^{d} a_{kj} x_j + a_{k0} \right) + c_0,
\]
where \( x = (x_1, \ldots, x_d) \) is the vector of the input variables, \( a_{kj} \) is the weight of the connection between the \( j \)th input neuron and the \( k \)th neuron in the hidden level; \( c_k, k = 1, \ldots, r \) is the weight of the link between the \( k \)th neuron in the hidden layer and the output; \( a_{k0} \) and \( c_0 \) are, respectively, the bias term of the hidden neurons and of the output; \( \phi(\cdot) \) is the activation function of the hidden layer. We define \( \theta = (c_0, c_1, \ldots, c_r, a'_1, a'_2, \ldots, a'_r)' \) where \( a'_i = (a_{0i}, a_{1i}, \ldots, a_{di}) \) and we suppose that \( \theta \in \Theta \) with \( \Theta \) compact subset of \( \mathbb{R}^{(d+2)+1} \).

As usual in neural network applications, we will assume a sigmoidal activation function such as the logistic or the hyperbolic tangent function. In this case, being \( g(\cdot) \) continuously differentiable on a compact set, the universal approximation theorem of Barron (1993) guarantees that, for each \( C_g > 0 \), there exists a single hidden feed forward neural network with weights \( \theta^* \), so that
\[
\|g(x) - f(x, \theta^*)\|_2 \leq \frac{(2sC_g)^2}{r} \tag{2}
\]
with \( s > 0 \). The coefficients of the neural network can be restricted to satisfy \( \sum_{i=1}^{r} |c_i| \leq 2sC_g \) and \( c_0 = g(0) \).

So we can approximately model the time series as \( Y_t \equiv f(X_t; \theta^*) + \varepsilon_t \) but, since the difference \( f(x; \theta^*) - g(x) \) is negligible if compared to the noise, we neglect this theoretical imprecision and suppose that \( g(x) \) can be exactly represented by \( f(x; \theta^*) \).

An estimator \( \hat{\theta}_T \) of the parameter vector can be obtained by solving
\[
\sum_{t=1}^{T} \psi(Z_t, \theta) = 0, \tag{3}
\]
where \( \psi: \mathcal{Z} \times \Theta \to \mathbb{R} \) with \( \mathcal{Z} \subset \mathbb{R}^{d+1} \), is a suitably chosen function leading to different classes of estimators including the least squares, maximum likelihood and generalized method of moments. A solution to the problem can be given by using a recursive \( M \)-estimator defined (Kuan and White, 1994) as
\[
\hat{\theta}_{t+1} = \hat{\theta}_t + a_t \psi(Z_t, \hat{\theta}_t), \quad t = 1, 2, \ldots,
\]
where \( \{a_t\} \) is a sequence of real positive integers.
3. Variable selection by relevance measures

In a neural network regression framework a key issue is the choice of a proper topology for a given problem. This requires selecting both an appropriate number of hidden units and a suitable set of the explicative variables and, as a consequence, the connections thereof. The most popular approaches used so far are pruning, stopped training and regularization (Reed, 1993). Although these techniques may lead to satisfactory results, they focus on single weights and this can be misleading due to the black-box nature of the neural network model. Indeed, they do not give any information on the most “significant” variables, which is useful in any model building strategy and, moreover, different topologies can achieve the same approximation accuracy. Therefore, a proper choice of the network topology cannot be just based on complexity reasons and should also take into account model plausibility.

All the techniques based on weights selection are much more on the side of computational standpoint than on the side of a statistical perspective. Instead, it would be of some interest to look at the choice of the network topology by including this problem in the classical statistical model selection approach. In this view, Anders and Korn (1999) propose a complete strategy based on a sequence of LM tests on sets of weights. Although the statistical perspective is very strong, again the role of the variables is not stressed. To emphasize this role, the choice of the neural network topology should follow a different strategy. Generally, a desirable neural network contains as few hidden units as necessary for a good approximation of the true function, taking into account the trade-off between estimation bias and variability. As a consequence, once fixed the hidden layer size by using some asymptotic results (Perna and Giordano, 2001, inter alia) or some alternative ad hoc procedures (Faraway and Chatfield, 1998), attention can be focused on model identification in a classical sense: the selection of the explanatory variables, in order to remove the “irrelevant” ones.

The approach we consider here is based on the introduction of some measures for the variable relevance which also, in an exploratory view, allows a ranking of the available explanatory variables. This involves (i) quantifying a “relevance” measure, (ii) estimating the sampling distribution (or at least variability) of this measure and (iii) testing the hypothesis of insignificance. This approach strictly follows the usual one generally employed when selecting a model in the classical regression framework.

In a linear regression model the relevance of a variable is measured by its coefficient which is also the magnitude of the partial derivative of the dependent variable with respect to the variable itself. So, in this set up, testing whether this coefficient is zero is equivalent to testing the hypothesis that the variable is not relevant for the model. When dealing with nonlinear functions, the partial derivative is not a constant but it varies through the range of the independent variables. So, the hypothesis that a set of independent variables \{x_i, i \in I_0\} has no effect on Y can be formulated as

$$\frac{\partial g(x)}{\partial x_i} = 0, \ \forall x, \ i \in I_0.$$
Of course, the function $g(\cdot)$ is unknown but we can investigate the hypothesis

$$f_i(x; \theta^*) = \frac{\partial f(x; \theta^*)}{\partial x_i} = 0, \quad \forall x, i \in I_0,$$

since the function $f(\cdot; \cdot)$ is known and $\theta^*$ can be closely approximated. The two sets of hypotheses are not strictly equivalent since hypotheses on $g$ are related to the structure of the data generating process whereas those on $f$ concern the model used to approximate the underlying process. Nevertheless, as pointed out previously, $f$ can approximate $g$ to any degree of accuracy (Barron, 1993) and so the difference between the two functions could be considered negligible and the two sets of hypotheses can be confused.

As a consequence a relevance measure of a set of explanatory variables to the model can be based on some functions of the partial derivatives of $f$, namely $m(x; \theta^*) = \sum_{i \in I_0} h(f_i(x, \theta^*))$. Some proposals for choosing the function $h$ (and so the function $m$) are reported in Refenes and Zapranis (1999). For example, if $h$ is the identity function we get the average derivative as relevance measure. However, cancellations between negative and positive values can make this measure not representative of the effect of the set $\{x_i, i \in I_0\}$ on the dependent variable. Alternatively, to avoid the problem, the absolute value or the square function could be used to get a suitable measure. Evidently, there are potentially many possible quantifications of the sensitivity of the dependent variable to the independent ones. Any proposed measure reflects different aspects of this relationship which are to be related to the applications at hand. However, to use these measures as inferential tools, their sampling distribution is needed and so, in the next section, we focus on a class of measures based on smooth functions of the partial derivatives.

4. A test procedure to select input variables

The hypothesis in Eq. (4) can be equivalently written as

$$H_0 : \mathbb{E}(m(X_t, \theta^*)) = 0$$

and tested by using the statistic

$$\hat{m}_T = T^{-1} \sum_{t=1}^T m(X_t, \hat{\theta}_T).$$

To derive the asymptotic distribution of this statistic, we now pose some general conditions on the function $m$.

**Assumption B (Test statistic).** (i) Let $m$ be a function such that $m(x; \cdot)$ is continuously differentiable of order 2, $\forall x \in \mathcal{X}$ and $m(\cdot; \theta)$ is measurable $\forall \theta \in \Theta$.

(ii) $m(x; \theta^*) = 0, \quad \forall x$.

(iii) $\nabla m(x; \theta^*) = 0, \forall x$ where $\nabla m(x; \cdot)$ is the gradient of $m(x; \cdot)$ with respect to $\theta$.

(iv) $\mathbb{E}[\sup_{\theta} |\nabla^2 m(X_t, \theta)|] < \infty$. 

Moreover, we need some mild regularity conditions on the structure of the function $\psi$ used in the recursive estimation algorithm.

**Assumption C (Estimating function).** (a) Let $\psi$ be a function such that $\psi(z; \theta)$ is continuously differentiable $\forall z \in \mathcal{Z}$ and $\psi(\cdot; \theta)$ is measurable $\forall \theta \in \Theta$. There exists a function $h_1: \mathbb{R}^+ \rightarrow \mathbb{R}^+$ with $h_1(u) \rightarrow 0$ as $u \rightarrow 0$ and a measurable function $h_2: \mathcal{Z} \rightarrow \mathbb{R}^+$ such that

$$|\nabla \psi(z, \theta) - \nabla \psi(z, \theta^*)| \leq h_1(|\theta - \theta^*|)h_2(z),$$

for each $(z, \theta) \in \mathcal{Z} \times \Theta^o$, $\Theta^o$ is an open neighbourhood of $\theta^*$.

(b) There exist a unique vector $\theta^* \in \Theta$ such that

(i) $E[\psi(Z_t, \theta^*)] = 0$;

(ii) $||\psi(Z_t, \theta^*)||_6 < \infty$;

(iii) $||\nabla \psi(Z_t, \theta^*)||_2 < \infty$;

(iv) the eigenvalues of $H = H^* + I/2$ where $H^* = E[\nabla^2 m(X_t, \theta^*)]$ and $I$ is the identity matrix, have negative real parts.

Now we are in a position to prove a result which straightforward extends Theorem 2.1 in White and Racine (2001) to the case of stationary and mixing data.

**Theorem 1.** If the assumptions A, B and C hold, then, for $T \rightarrow \infty$,

$$\sum_{i=1}^{T} m(X_t, \hat{\theta}_T) \xrightarrow{d} \sum_{i=1}^{k} \omega_i \lambda_{m}^2,$$

where $\omega_1, \ldots, \omega_k$ are the roots (with multiplicity $\lambda_1, \ldots, \lambda_k$) of the equation $|\nu I - M^* C^*| = 0$ with $M^* = E[\nabla^2 m(X_t, \theta^*)]/2$ and $C^*$ variance–covariance matrix of the asymptotic distribution of the estimator $\hat{\theta}_T$.

**Proof.** See the appendix.

As in White and Racine (2001) we consider the function

$$m(x, \theta) = \sum_{i \in I_0} f_i^2(x, \theta),$$

which satisfies the requirements of Theorem 1. So, letting

$$m^* = \sum_{i \in I_0} E(f_i^2(X_t, \theta^*)),$$

we have that $m^*$ is zero if and only if Eq. (4) is true and $m^*$ is positive if the set of variables $\{x_i, i \in I_0\}$ contains at least one relevant variable. So, the hypothesis $H_0: m^* = 0$ can be tested by using the statistic

$$\hat{m}_T = T^{-1} \sum_{i \in I_0} \sum_{t=1}^{T} f_i^2(X_t, \hat{\theta}_T).$$
The estimator \( \hat{m}_T \), under the null, converges to a nondegenerate limit law as shown by the following corollary of the previous theorem.

**Corollary 2.** Under the null, if \( \mathbb{E}[X_t'X_t] < \infty \) and if the assumptions A and C hold, then, for \( T \to \infty \),

\[
T \hat{m}_T = \sum_{i \in I_0} \sum_{t=1}^{T} f_i^2(X_t, \hat{\theta}_T) - \sum_{i=1}^{k} \omega_i \lambda_i^2,
\]

where \( \omega_1, \ldots, \omega_k \) are the roots (with multiplicity \( \lambda_1, \ldots, \lambda_k \)) of the equation \(|\omega I - M^* C^*| = 0\) with

\[
M^* = \mathbb{E} \left( \sum_{i \in I_0} [\nabla f_i(X_t, \theta^*) \nabla f_i(X_t, \theta^*)' + f_i(X_t, \theta^*) \nabla^2 f_i(X_t, \theta^*)] \right).
\]

**Proof.** See the appendix.

Theorem 1 and Corollary 2 show that the procedure proposed by White and Racine (2001) can be extended to dependent data and the additional requirements basically involve only stationarity and strong mixing of the underlying process. Weaker conditions could be imposed; nevertheless the class of processes considered here contains a wide variety of linear and nonlinear structures (Doukhan, 1994) and in any case, restrictions to this class should be imposed when considering the resampling technique we focus on in this paper.

From an operational point of view, a key point is the selection of the set of variables \( I_0 \) to be tested as irrelevant. A simple graphic exploratory device based on a plot of the relevance measure of every single variable i.e. \( (i, R_i) \) where \( R_i = \sum_{t=1}^{T} f_i^2(X_t, \hat{\theta}_T) \), can be used as a guideline. Clearly, values of this statistic close to zero candidate the variable to be in the set of irrelevant ones.

5. **The subsampling approach**

The asymptotic distribution of the test statistic is very difficult to deal with since it involves matrices with a complex structure and difficult to estimate. Therefore, the unknown sampling distribution can be better approximated by some types of resampling techniques. In the context of neural network regression models with iid data, the bootstrap method has been pursued in Tibshirani (1995), Refenes and Zapranis (1999) and White and Racine (2001) inter alia. As a matter of fact, as pointed out by Singh (1981), the standard bootstrap, when applied to dependent data, gives results that are not consistent. A way to preserve the original dependence structure in the resampled data is to use the subsampling approach which gives valid statistical inference under very general and minimal conditions. It only requires some regularity and dependency conditions and that the sampling distribution of the properly normalized statistic of interest has a nondegenerate limiting law (Politis et al., 1999).
The subsampling method is probably the simplest way to take into account the dependent structure of the data. In this approach, blocks of consecutive observations are obtained from the original observed time series, looking upon each individual subseries of observations as a valid “sub-time series” in its own right. Each block, being a part of the original time series, has been generated by the true underlying data generating process and so, information on the sampling distribution of a given statistic can be gained by evaluating the statistic on all subseries.

Let \( Y_{b,t} = (Y_t, \ldots, Y_{t+b-1})' \) and \( X_{b,t} = (X_t', \ldots, X_{t+b-1}')' \) subseries of consecutive observations, obtained by splitting the observed data series \( Z_1 = (Y_1, X_1), \ldots, Z_T = (Y_T, X_T) \) into overlapping blocks of size \( b \). Let \( \{\hat{\theta}_{T,b,t}\}, t = 1, \ldots, T - b + 1 \) be the estimates of the parameter vector based on these subseries and let \( \{\hat{m}_{T,b,t}\}, t = 1, \ldots, T - b + 1 \) be the statistics computed on these subsample estimates.

The sampling distribution of the test statistic \( T \hat{m}_T \) is approximated by

\[
\hat{G}_{T,b}(x) = \frac{1}{T - b + 1} \sum_{t=1}^{T-b+1} I(b \cdot \hat{m}_{T,b,t} \leq x),
\]  

where \( I(\cdot) \) is the indicator function.

Therefore, the critical value for the test is obtained as the \( 1 - \alpha \) quantile of \( \hat{G}_{T,b}(\cdot) \) defined as \( g_{T,b}(1 - \alpha) = \inf \{x : \hat{G}_{T,b}(x) \geq 1 - \alpha\} \). So, the test rejects the null hypothesis that the set of variables \( \{x_i, i \in I_0\} \) is not relevant if and only if \( T \hat{m}_T > g_{T,b}(1 - \alpha) \).

The consistency of the proposed subsampling procedure is ensured by the following result.

**Theorem 3.** Let \( b/T \to 0, b \to \infty, \) as \( T \to \infty \) and assume that A, B and C hold.

(a) Under the null, we have, as \( T \to \infty \),

\[
\Pr\{T \hat{m}_T > g_{T,b}(1 - \alpha)\} \to \alpha.
\]

(b) Under the alternative, if

\[
\mathbb{E} \left[ \sup_{\theta} |m(X_t, \theta)| \right] < \infty,
\]

we have, as \( T \to \infty \),

\[
\Pr\{T \hat{m}_T > g_{T,b}(1 - \alpha)\} \to 1.
\]

**Proof.** See the appendix.

The subsampling method gives consistent results under general minimal assumptions which are valid for both linear and nonlinear processes. It does not require any knowledge of the specific structures of the time series other than its stationarity and strong mixing property, so it is robust against misspecified models. This property is particularly useful in neural network models that are basically “atheoretical”, employed for the lack of knowledge about the functional form of the data generating process, and so intrinsically misspecified in the sense of White (1994), being an approximation of the underlying model.
The main problem when applying the subsampling procedure lies in choosing the length of the block. Even if the conditions on $b$ are quite weak, they do not give any guidelines for its choice and this parameter, which is related to the amount of dependence assumed in the series, has to be chosen on the data at hand. Nevertheless, Theorem 2.7.1 in Politis et al. (1999) ensures that the asymptotic results are still valid for a broad range of choices for the subsample size.

The solutions proposed in the literature for the estimation of $b$ are really computer-intensive even for simple cases and become soon impractical for neural network models. Therefore, the choice of a feasible estimation procedure for block length in this framework still remains an open issue.

6. The results of a Monte Carlo study

To evaluate the performance of the proposed test procedure in finite samples a small Monte Carlo study was performed. The experimental setup is based on datasets generated by an Exponential Autoregressive model of order 2 (EXPAR(2)), defined as

$$Y_t = [\beta_{01} + \beta_{11} \exp(-Y_{t-1}^2)]Y_{t-1} + [\beta_{02} + \beta_{12} \exp(-Y_{t-1}^2)]Y_{t-2} + \epsilon_t,$$

where the innovations $\epsilon_t$ are distributed as standard normal. This nonlinear model is very flexible and allows generation of quite different time series structures. Moreover, the skeleton of the model is defined by a function which belongs to the class of continuously differentiable functions and so the universal approximation theorem of Barron (1993) applies. In addition, the EXPAR process is geometrically ergodic which implies that it is stochastically stable and it is also strongly mixing with geometrically decreasing mixing coefficients (Gyorfi et al., 1990).

To study the level of the test, we fixed $\beta_{01} = 0.5, \beta_{11} = 0.9, \beta_{02} = 0.0$ and $\beta_{12} = 0.0$. We modeled the observed series with a neural network with two input variables, lag 1 and lag 2, and we tested the hypothesis that the latter is not relevant. We used two different network topologies estimated by using a least square loss function, with hidden layer size $r = 1$ and 2. These values guarantee a good approximation of the network to the time series generated from an EXPAR model while higher values can lead to overfitting for this simple data generating process. The distribution of the test statistic was approximated by the subsampling technique with blocks of length $b$ which varies in the range $[30; 190]$ in steps of 10. We considered two different time series lengths ($T = 500$ and 1000) and 1000 Monte Carlo runs.

In Table 1 we reported the observed level of the test corresponding to two different choices of nominal level $\alpha = 0.05$ and 0.10. The values reported in bold are inside the asymptotic acceptance interval of 0.95. As expected, a wrong choice of $b$ can lead to under or over estimation of the size. Nevertheless, for a quite large range (roughly, from 60 to 160 for series with $T = 500$ and from 90 to 190 for series with $T = 1000$ observations) the observed size cannot be considered statistically different from the nominal one. Therefore, the choice of the subseries length $b$, although it is considered critical in general, in this framework it can be made in a quite large interval without affecting the observed size of the test in a severe way.
### Table 1
Oberved level for nominal size $\alpha = 0.05$ and 0.10. Rejection proportions, under the null, computed over 1000 Monte Carlo runs

<table>
<thead>
<tr>
<th>$T$</th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>0.05</td>
<td>0.10</td>
</tr>
<tr>
<td>$r$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.026</td>
<td>0.008</td>
</tr>
<tr>
<td>2</td>
<td>0.029</td>
<td>0.020</td>
</tr>
<tr>
<td>3</td>
<td>0.023</td>
<td>0.021</td>
</tr>
<tr>
<td>4</td>
<td><strong>0.038</strong></td>
<td><strong>0.025</strong></td>
</tr>
<tr>
<td>5</td>
<td>0.036</td>
<td>0.036</td>
</tr>
<tr>
<td>6</td>
<td>0.044</td>
<td>0.036</td>
</tr>
<tr>
<td>7</td>
<td>0.036</td>
<td>0.042</td>
</tr>
<tr>
<td>8</td>
<td>0.036</td>
<td>0.053</td>
</tr>
<tr>
<td>9</td>
<td>0.042</td>
<td>0.036</td>
</tr>
<tr>
<td>10</td>
<td>0.062</td>
<td>0.061</td>
</tr>
<tr>
<td>11</td>
<td>0.044</td>
<td>0.064</td>
</tr>
<tr>
<td>12</td>
<td>0.063</td>
<td>0.049</td>
</tr>
<tr>
<td>13</td>
<td>0.061</td>
<td>0.059</td>
</tr>
<tr>
<td>14</td>
<td>0.064</td>
<td>0.059</td>
</tr>
<tr>
<td>15</td>
<td>0.067</td>
<td>0.070</td>
</tr>
<tr>
<td>16</td>
<td>0.073</td>
<td>0.069</td>
</tr>
<tr>
<td>17</td>
<td>0.076</td>
<td>0.067</td>
</tr>
</tbody>
</table>

Values reported in bold are inside the 0.95 asymptotic acceptance interval which is $(0.036, 0.064)$ for $\alpha = 0.05$ and $(0.081, 0.119)$ for $\alpha = 0.10$.

This behaviour can be explained by looking at Fig. 1 where we reported the box-plots of the distribution of the 0.95 quantile estimated by subsampling for the different values of the subseries length $b$, with time series of length $T=1000$. Clearly, the distributions become roughly stable in the range going from 90 to 190.

In Fig. 2 we reported the power functions of the proposed test procedure for selected block lengths. In the data generating process we fixed $\beta_{01} = 0.5$, $\beta_{11} = 0.9$, $\beta_{02} = 0.0$ and $\beta_{12}$ ranging in the interval $[-0.4,0.4]$. For each value of $\beta_{12}$ we computed the empirical rejection frequency for the test $H_0 : \beta_{12} = 0$ with a nominal level of $\alpha = 0.05$ (panel (a)) and $\alpha = 0.10$ (panel (b)). The functions present the classical bell-shaped form where the power increases as the distance between the true value of $\beta_{12}$ and its hypothesized value under the null grows. Again, when considering different values of the subseries length (namely $b = 60, 90, 120$ and 180) the differences are not very large. Obviously, the problem of the optimal choice of $b$ is still an open question and should be further investigated.

### 7. Some concluding remarks

In this paper we propose a test procedure, calibrated by subsampling, to select the set of relevant variables for a neural network regression model with dependent data. The overall results are encouraging. The test procedure, which is justified from an
Fig. 1. Distributions, over 1000 Monte Carlo runs, of the 0.95 quantile estimated by subsampling for different values of $b$. Neural networks estimated with two neurons in the hidden layer. Time series with $T = 1000$ observations.

Fig. 2. Observed power functions for different values of $b$. Panel on the left refers to $z = 0.05$ and panel on the right to $z = 0.10$. The curves correspond to $b = 60$ (the higher line), $b = 90, 120$ and 180 (the lower line). The lower dotted lines refer to the nominal level of the test.
asymptotic point of view, has empirical level close to nominal one and power functions with a correct shape, rising steadily to one.

In any case, several different aspects should be further explored to get a better insight into the joint usage of neural networks and resampling methods.

From a computational point of view, the resulting combined procedure is really computer intensive, calling for implementing very efficient algorithms for neural network learning and for the subsampling scheme. Of course, the increasing power computing available even on PC desktops and the increasing ease of access to parallel computing make the problem easier to face with, in the next future.

From a statistical point of view, interesting points arise when considering the relationship between the block length and the hidden layer size of the network. A comparison with alternative resampling techniques, such as the moving block bootstrap, would also be of interest.

Acknowledgements

We would like to thank two anonymous referees for their helpful comments and suggestions.

Appendix A. Proofs

Proof of Theorem 1. By using a Taylor expansion around $\theta^*$ we get

$$
\sum_{t=1}^T m(X_t, \hat{\theta}_T) = \sum_{t=1}^T m(X_t, \theta^*) + \sum_{t=1}^T \nabla m(X_t, \theta^*)(\hat{\theta}_T - \theta^*)
$$

$$
+ \frac{T}{2}(\hat{\theta}_T - \theta^*)' \left\{ \frac{1}{T} \sum_{t=1}^T \nabla^2 m(X_t, \tilde{\theta}_T) \right\} (\hat{\theta}_T - \theta^*),
$$

where $\tilde{\theta}_T$ is a proper chosen point between $\hat{\theta}_T$ and $\theta^*$.

By using conditions B(ii) and B(iii), it follows

$$
\sum_{t=1}^T m(X_t, \hat{\theta}_T) = \frac{T}{2}(\hat{\theta}_T - \theta^*)' \left\{ \frac{1}{T} \sum_{t=1}^T \nabla^2 m(X_t, \tilde{\theta}_T) \right\} (\hat{\theta}_T - \theta^*)
$$

which, by adding and subtracting the quantity $E[\nabla^2 m(X_t, \theta^*)]$, becomes

$$
\sum_{t=1}^T m(X_t, \hat{\theta}_T) = \frac{T}{2}(\hat{\theta}_T - \theta^*)' \Xi^*(\hat{\theta}_T - \theta^*),
$$

where

$$
\Xi^* = \left\{ \frac{1}{T} \sum_{t=1}^T \nabla^2 m(X_t, \tilde{\theta}_T) - E[\nabla^2 m(X_t, \theta^*)] + E[\nabla^2 m(X_t, \theta^*)] \right\}.
$$
Being \( \tilde{\theta}_T \) on the segment joining points \( \hat{\theta}_T \) and \( \theta^* \), it can be written as \( \tilde{\theta}_T = (1 - \lambda) \hat{\theta}_T + \lambda \theta^* \) with \( \lambda \in [0, 1] \). By Assumptions A and C and being \( \psi(\cdot) \) a bounded, continuously differentiable function, we have (Kuan and White, 1994),

\[
\sqrt{T}(\hat{\theta}_T - \theta^*) \overset{d}{\rightarrow} N(0, \mathbf{C}^*)
\]

where \( \mathbf{C}^* = \int_{0}^{+\infty} \exp(\mathbf{H}c) \Sigma \exp(\mathbf{H}'c) dc \) is the unique solution to the matrix equation

\[
\mathbf{H} \mathbf{C}^* + \mathbf{C}^* \mathbf{H}' = -\Sigma.
\]

So we have that \( \sqrt{T}(\hat{\theta}_T - \theta^*) = O_p(1) \), \( \hat{\theta}_T = \theta^* + o_p(1) \) and \( \tilde{\theta}_T = \theta^* + o_p(1) \).

Moreover, by conditions B(i) and B(iv), \( \nabla^2 m(X_t, \theta) \) obeys the uniform weak law of large numbers (Bierens, 1994, Theorem 6.3.2) that is

\[
\frac{1}{T} \sum_{t=1}^{T} \nabla^2 m(X_t, \hat{\theta}_T) - \mathbb{E}[\nabla^2 m(X_t, \theta^*)] = o_p(1)
\]

and, as a consequence,

\[
\sum_{t=1}^{T} m(X_t, \hat{\theta}_T) = T(\hat{\theta}_T - \theta^*)' \mathbb{E}[\nabla^2 m(X_t, \theta^*)](\hat{\theta}_T - \theta^*) + o_p(1).
\]

So, by using the Slutzky theorem (Serfling, 1980), the quantity \( \sum_{t=1}^{T} m(X_t, \hat{\theta}_T) \) and \( T(\hat{\theta}_T - \theta^*)' \mathbb{E}[\nabla^2 m(X_t, \theta^*)](\hat{\theta}_T - \theta^*) \) have the same asymptotic distribution. By a theorem in White (1994) the quadratic form is distributed as a linear combination of independent chi-square distributions.

**Proof of Corollary 2.** The conditions assumed here imply those of the previous theorem. Condition B(i) is true since \( \psi(\cdot) \) is bounded and continuously differentiable. In addition, the assumptions on \( \psi(\cdot) \), if \( \mathbb{E}[X_t'X_t] < \infty \), guarantee that

\[
\mathbb{E} \left[ \sup_{\theta} |\nabla f_i(X_t, \theta)\nabla f_i(X_t, \theta)' + f_i(X_t, \theta)\nabla^2 f_i(X_t, \theta)| \right] < \infty
\]

and so B(iv) holds. Moreover B(iii) is valid since

\[
\nabla m(x, \theta^*) = 2 \sum_{i \in I_0} \nabla f_i(x, \theta^*)f_i(x, \theta^*) = 0
\]

and B(ii) is implied by Eq. (4).

**Proof of Theorem 3.** (a) By Theorem 1, under Assumptions A–C, the sampling distribution of the test statistic, \( G_T(x) = \Pr\{T \cdot \tilde{m}_T \leq x\} \), converges to an absolutely continuous limit law \( G(x) \), namely a mixture of independent chi-square distributions. From this result and from the assumptions on the subseries length, it follows that \( \hat{G}_{T,h}(x) \rightarrow G(x) \) in probability (Politis et al., 1999, Theorem 2.2.1). It follows that the quantile \( g_{T,h}(1 - \alpha) \) converges, in probability as \( T \rightarrow \infty \), to the \( 1 - \alpha \) quantile of the distribution \( G(x) \). Thus, by Slutsky’s theorem, the asymptotic rejection probability of the event \( T \cdot \tilde{m}_T > g_{T,h}(1 - \alpha) \) is exactly \( \alpha \).
(b) Assumptions A and C imply those of Corollary II.4.1 in Kuan and White (1994) ensuring that \( \hat{\theta}_T \rightarrow \theta^* \). Assuming that Assumptions B(i), B(iii), B(iv) and Assumption (6) hold, we get \( \hat{m}_T \rightarrow m^* \) by the uniform strong law of large numbers for mixing processes (Bierens, 1994, Theorem 6.3.4). Moreover, \( m^* = 0 \) under the null and \( m^* > 0 \) under the alternative. Now, define

\[
\hat{G}^0_{T,b}(x) = \frac{1}{T - b + 1} \sum_{t=1}^{T-b+1} I(\hat{m}_{T,b,t} \leq x) = \hat{G}_{T,b}(hx).
\]

Following Politis et al. (1999), observe that \( \hat{G}^0_{T,b}(x) \) is a U-statistic of degree \( b \) and kernel \( h(y) = I(y \leq x) \) (Serfling, 1980) with expectation

\[
E(\hat{G}^0_{T,b}(x)) = \Pr\{\hat{m}_b \leq x\}.
\]

Therefore, since \( \hat{m}_T \rightarrow m^* \), \( \hat{G}^0_{T,b}(x) \) converges in distribution to a point mass at \( m^* \). As a consequence, the \( 1 - \alpha \) quantile \( g^0_{T,b}(1 - \alpha) \) of \( \hat{G}^0_{T,b}(x) \) converges in probability to \( m^* \). But, the proposed test rejects when

\[
\frac{T}{b} \cdot \hat{m}_T > g^0_{T,b}(1 - \alpha).
\]

Since \( T/b > 1 \) and \( \hat{m}_T \rightarrow m^* \) in probability (with \( m^* > 0 \)) it follows by Slutsky’s theorem that the asymptotic rejection probability is one. □

References

White, H., Racine, J., 2001. Statistical inference, the bootstrap, and neural-network modeling with application