

# Subsampling Model Selection in Neural Networks for Nonlinear Time Series Analysis

Michele La Rocca and Cira Perna  
Di.S.E.S., University of Salerno  
*email:* [larocca, perna]@unisa.it

## Abstract

In this paper, the subsampling method is applied to the problem of model selection in neural networks for non linear time series data. A complete strategy, which combines a set of graphical, exploratory and inferential statistical tools, is proposed to select the topology of a neural network model. The procedure allows to choose the number and the type of inputs (by using a formal test procedure based on relevance measures) and to identify the hidden layer size (by looking at the predictive performance of the neural network model). The proposed approach heavily uses the subsampling technique to extend some approaches already available for the *iid* case to dependent data.

## 1 Introduction

The subsampling method was first proposed by Carlstein (1996) and extensively studied by Politis and Romano (1994) as a tool for estimating parameters of the sampling distribution of a statistic computed from a sample from a stationary process. 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 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.

The subsampling technique gives consistent results under quite general and weak assumptions (Politis et al., 1999). Basically the scheme requires the existence of a limiting law for the sampling distribution of the statistic involved in the inference process and it does not require any knowledge of the specific structures of the time series other than its asymptotic stationarity and strong mixing properties.

The aim of the paper is to apply the subsampling method to the problem of model selection in neural networks for non linear time series data. Here we propose a complete strategy, which combines a set of graphical, exploratory and inferential statistical tools, to select the topology of the neural network model.

The procedure allows to select the number and the type of inputs by using a formal test procedure based on relevance measures, and to identify the hidden layer size by looking at the predictive performance of the neural network model. The proposed approach heavily uses the subsampling technique to extend some approaches already available for the *iid* case to dependent data.

The paper is organized as follows. In section 2, we briefly illustrate the subsampling scheme to calibrate a test procedure. In section 3 we describe the data generating process and the neural network model we considered. In section 4 we discuss the proposed procedure for neural network model selection. In section 5, in order to evaluate the performances of the proposed procedure, we report an illustrative example on a simulated data set. Some concluding remarks will close the paper.

## 2 The subsampling technique

In this section the subsampling method for the construction of hypothesis test for general null hypotheses, in the case of dependent data, is briefly reviewed (see Politis et al., 1999, for details).

Let  $\{Y_t, t \in \mathbb{Z}\}$  be a stationary and mixing process governed by a probability law  $P$ , assumed to belong to a certain class of laws  $\mathcal{P}$ . The goal is to construct an asymptotically valid test for the null hypothesis  $H_0 : P \in \mathcal{P}_0$  versus the alternative  $H_1 : P \in \mathcal{P}_1$  with  $\mathcal{P}_0 \cup \mathcal{P}_1 = \mathcal{P}$ . Given the observed time series  $\{Y_1, Y_2, \dots, Y_T\}$ , the test can be based on a test statistic such as

$$W_T = \delta_T w_T = \delta_T w_T(Y_1, Y_2, \dots, Y_T)$$

where  $\delta_T$  is a normalizing sequence. Assume that there exists a constant  $w(P)$  which satisfies  $w(P) = 0$  under the null and  $w(P) > 0$  under the alternative and that  $w_T \rightarrow w(P)$  in probability.

Let  $G_T(x, P) = \Pr_P \{\delta_T w_T \leq x\}$  the cumulative distribution function of the sampling distribution of the test statistic and assume that it converges in distribution to a given limit law at least for  $P \in \mathcal{P}_0$ . Naturally, as long as  $\delta_T \rightarrow \infty$ , this implies that  $w_T \rightarrow 0$  in probability for  $P \in \mathcal{P}_0$ .

The subsampling scheme works as follows.

Fix the subseries length  $b$  and let  $\mathbf{Y}_{b,t} = \{Y_t, Y_{t+1}, \dots, Y_{t+b-1}\}$  be a subseries of  $b$  consecutive observations. The sampling distribution of  $W_T$  can be then approximated by subsampling as

$$\hat{G}_{T,b}(x) = \frac{1}{T-b+1} \sum_{t=1}^{T-b+1} \mathbb{I}\{\delta_b w_{b,t} \leq x\}$$

where  $w_{b,t}$  is the test statistic evaluated at the block of data  $\mathbf{Y}_{b,t}$ . The critical value for the test is obtained as the  $1 - \alpha$  quantile of  $\hat{G}_{T,b}(\cdot)$ , that is

$$g_{T,b}(1 - \alpha) = \inf \left\{ x : \hat{G}_{T,b}(x) \geq 1 - \alpha \right\}.$$

The nominal level  $\alpha$  test rejects the null  $H_0$  if and only if

$$W_T > g_{T,b}(1 - \alpha).$$

Clearly, a subsampling  $p$ -value can be computed as

$$PV_{T,b} = \frac{1}{T - b + 1} \sum_{t=1}^{T-b+1} \mathbb{I}\{\delta_b w_{b,t} \geq \delta_T w_T\}$$

and as a consequence the nominal level  $\alpha$  test rejects the null if and only if  $PV_{T,b} < \alpha$ .

The subsampling method gives consistent results under general minimal assumptions which are valid for both linear and nonlinear processes. Basically the scheme requires that  $\frac{\delta_b}{\delta_T} \rightarrow 0$ ,  $\frac{b}{T} \rightarrow 0$  and  $b \rightarrow \infty$  as  $T \rightarrow \infty$  and the existence of a limiting law for the sampling distribution of the test statistic. The normalizing sequence  $\delta_T$  could also be unknown, and a preliminary round of subsampling can be used to consistently estimate the rate  $\delta_T$  (Bertail et al., 1999). Moreover, the subsampling can be used to approximate the sampling distributions of diverging statistics that are particularly useful in some applications, such as in assessing financial risk (Bertail et al., 2004).

The method 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. Moreover, the subsampling is by no means restricted to stationary series but it gives asymptotically correct inference even for heteroskedastic time series (Politis et al., 1997).

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. They proposed a number of strategies to select  $b$ , but in the following we focus on the minimum volatility method which works under very general conditions.

The procedure runs as follows. For  $b = b_{small}$  to  $b = b_{big}$ , compute a subsampling quantile  $g_{T,b}(1 - \alpha)$ , for the desired significance level  $\alpha$ . For each  $b$  compute a volatility index  $VI_b$  of the quantiles in neighborhood of  $b$ . More specifically, for a small integer  $k$ , let  $VI_b$  be equal to the standard deviation of the values  $\{g_{T,b-k}(1 - \alpha), \dots, g_{T,b+k}(1 - \alpha)\}$ . Pick the value  $b^*$  corresponding to the smallest volatility index and use  $g_{T,b^*}(1 - \alpha)$  as a critical value of the test.

As variability index the standard deviation or others robust alternatives (such as the MAD) can be used.

### 3 The DGP and the neural network model

Let  $\{Y_t, t \in \mathbb{Z}\}$ , a process modeled as:

$$Y_t = g(\mathbf{X}_t) + \varepsilon_t$$

where  $\{Y_t, \mathbf{X}'_t\}$  is a stationary,  $\alpha$ -mixing sequence and  $\mathbf{X}_t = (X_{1t}, \dots, X_{dt})'$  is a vector of  $d$  random variables possibly including explanatory variables, lagged explanatory variables and lagged values of  $Y_t$ . The unknown function  $g(\cdot)$  is assumed to be a continuously differentiable function defined on a compact subset of  $\mathbb{R}^d$ .

The function  $g$  can be approximated by a feed-forward neural network  $NN(d, r)$  defined as

$$f(\mathbf{x}_t, \theta) = \sum_{k=1}^r c_k \phi \left( \sum_{j=1}^d a_{kj} x_{jt} + a_k \right) + c_0$$

where  $\mathbf{x} = (x_1, \dots, x_d)$  is the vector of the  $d$  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;  $c_k(k)$ ,  $k = 1, \dots, r$  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  $\theta = (c_0, c_1, \dots, c_r, \mathbf{a}'_1, \mathbf{a}'_2, \dots, \mathbf{a}'_r)'$  where  $\mathbf{a}'_i = (a_{i0}, a_{i1}, \dots, a_{id})$  with  $\theta \in \Theta \subset \mathbb{R}^{r(d+2)+1}$ .

Artificial neural networks are widely accepted as flexible tools of modeling complex non linear and dynamic systems. They are particularly useful when the underlying physical process relationships are not fully understood or when the nature the phenomenon being modeled may display chaotic properties.

The diffusion of neural network models is due to their great flexibility and their capability of providing a model which fits any data with an arbitrary degree of accuracy. A well known result by Barron (1993) guarantees that, under quite general conditions, there exists a parameter vector  $\theta^*$  such that

$$\|g(\mathbf{x}) - f(\mathbf{x}, \theta^*)\| \leq \frac{(2C_g)^2}{r}$$

where  $C_g > 0$  is a proper chosen constant.

So once fixed the network topology, the parameter vector  $\theta^*$  can be estimated by solving

$$T^{-1} \sum_{t=1}^T \psi(\mathbf{z}_t, \theta) = 0.$$

where the function  $\psi(\cdot)$  can generate different classes of estimators such as least squares, maximum likelihood and generalized method of moments.

The universal approximation property and the wide choice of algorithms available for the estimation procedure have made the neural networks widely used in a variety of statistical applications. Unfortunately, this class of models

is not yet supported by the rich collection of specification and diagnostic tests usually employed in statistical and econometric modeling.

A crucial point, when using a  $NN(d, r)$  model, is the choice of a proper topology which, for feedforward networks, is basically related to the specification of the type and the number of the input variables and to the selection of the hidden layer size. The most used approaches to this problem are based on: (i) pruning and regularization (Reed, 1993); (ii) the use of information criteria, such as the AIC or the BIC; (iii) sequences of tests on sets of weights (Anders and Korn, 1999). 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. Moreover, different topologies can achieve the same approximation accuracy. As a consequence, a proper choice of the network topology cannot be just based on complexity reasons and it should also take into account model plausibility.

Therefore, a model selection strategy should emphasize the role of the explanatory variables (useful for the identification and interpretation of the model) and it should treat the hidden layer size as a smoothing parameter, taking into account the trade-off between estimation bias and variability. The solutions proposed in the statistical literature, following this spirit, for model identification, diagnostic testing and model adequacy, basically deal with *iid* case and so they are not suitable for time series data (Baxt and White, 1995; Refenes and Zapranis, 1999; White and Racine, 2001; Refenes and Holt, 2001).

## 4 The proposed procedure

The strategy we propose in this paper extends some results available for *iid* data to stationary and mixing processes, including the problem of topology selection in the classical model selection approach. The procedure emphasizes the role of the explanatory variables (which is important for the identification and interpretation of the model), and treats the hidden layer size as a smoothing parameter taking into account the trade-off between estimation bias and variability. In this perspective, the proposed procedure identifies the number and the type of inputs by using a formal test procedure and the hidden layer size by looking at the predictive performance of the neural network model, following the suggestions of Swanson et al. (1995) who emphasize the role of predictive criteria over the information based ones.

The procedure uses extensively the subsampling technique which 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 seems to be 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.

## 4.1 The selection of the input: inclusion of irrelevant variables

Here, we focus on the approach based on relevance measures for the input variables. In order to remove irrelevant variables we use a stepwise selection rule which involves

1. Definition of a variable's relevance to the model
2. Estimation of the sampling distribution of the relevance measure
3. Testing the hypothesis that the variable is irrelevant.

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.

In the nonlinear case, the partial derivative is not a constant but it varies through the range of the independent variables. So, the hypothesis that the independent variable  $x_i$  has no effect on  $Y$  can be formulated as

$$\frac{\partial g(x)}{\partial x_i} = 0, \forall x.$$

The hypothesis that a set of independent variables  $\mathcal{X}_0 = \{x_i, i \in I_0\}$  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

$$f_i(x; \theta^*) = \frac{\partial f(x; \theta^*)}{\partial x_i} = 0, \forall x, i \in I_0. \quad (1)$$

since the function  $f(\cdot; \cdot)$  is known and  $\theta^*$  can be closely approximated.

The general form for relevance measures is

$$RM_i(\theta^*) = \mathbb{E}[h[f_i(\mathbf{X}_t, \theta^*)]]$$

where  $h(\cdot)$  is a proper chosen function and  $\mathbb{E}\{\cdot\}$  is the expected value w.r.t. the probability measure of the vector of the explanatory variables. We can get the measures proposed by Refenes and Zapranis (1999) or White and Racine (2001) by choosing for example the average derivative ( $h(x) = x$ ); the absolute average derivative ( $h(x) = |x|$ ); the square average derivative ( $h(x) = x^2$ ) or the maximum and minimum derivative ( $h(x) = \max(x)$  and  $h(x) = \min(x)$ ).

Let  $\mathcal{X}_0 = \{x_i, i \in I_0\}$  be the set of variables to be tested as irrelevant to the model. The hypothesis that the variables in  $\mathcal{X}_0$  are not relevant can be written as

$$H_0 : m^* = \sum_{i \in I_0} \mathbb{E}[m_i(\mathbf{X}_t, \theta^*)] = 0$$

where

$$m_i(x, \theta) = h[f_i(x, \theta)]$$

The null  $H_0$  can be tested by using the statistic,

$$\begin{aligned} \hat{m}_T &= T^{-1} \sum_{i \in I_0} \sum_{t=1}^T m_i(\mathbf{X}_t, \hat{\theta}_T) \\ &= T^{-1} \sum_{t=1}^T m(\mathbf{X}_t, \hat{\theta}_T) \end{aligned}$$

where the parameter vector  $\hat{\theta}_T$  is consistent estimator of the unknown parameter vector  $\theta^*$ .

The distribution of the test statistic, under the null and under quite general assumptions, can be consistently approximated by using the subsampling (LaRocca and Perna, 2004) following the scheme of section 2.

## 4.2 The selection of input: omission of relevant variables

In order to verify if there are any omitted variables, it can be used a procedure which extends the one proposed by White and Racine (2001) for the *iid* case. It is based on the comparison of competing neural network models.

Let  $f_1(x, \theta_1^*)$  and  $f_2(x, \theta_2^*)$  two competing neural network models which are nested and differ only in the inputs. The idea is that if there are no omitted variables the network  $f_1$  is capable of producing an output identical to that of the network  $f_2$ .

The omission of relevant variables can be verified by using a discrepancy measure between the outputs of the two competing neural network models  $f_1(x, \theta_1^*)$  and  $f_2(x, \theta_2^*)$  defined as

$$m^* = \mathbb{E} \left[ (f_1(x, \theta_1^*) - f_2(x, \theta_2^*))^2 \right]$$

Therefore, the hypothesis that the two models are equivalent and so there are no omitted variables, can be written as

$$H_0 : m^* = 0$$

and it can be tested by using the statistic

$$\hat{m}_T = \frac{1}{T} \sum_{t=1}^T \left( f_1(\mathbf{X}_t, \hat{\theta}_{T1}) - f_2(\mathbf{X}_t, \hat{\theta}_{T2}) \right)^2$$

where  $\hat{\theta}_{T1}$  and  $\hat{\theta}_{T2}$  are consistent estimators of, respectively,  $\theta_1^*$  and  $\theta_2^*$ .

Again, the distribution of the test statistic can be consistently estimated by using the subsampling, thus extending the procedure to the case of dependent data.

### 4.3 The selection of the hidden layer size

In order to select the hidden layer size we follow the remarks of Swanson et al. (1995) who showed that information criteria, such as the AIC or the BIC, when applied to neural networks, are not able to select models parsimoniously with low prediction errors. So we employ in the context of neural networks a procedure proposed by Fukuchi (1999) in a general set-up. It is based on a measure of predictive risk estimated by using the subsampling.

Given the model  $NN(d, r)$ , a measure of the predictive risk is

$$\Delta_T(r) = \mathbb{E} \left[ \left( \hat{Y}_{T+1} - Y_{T+1} \right)^2 \right]$$

where  $\hat{Y}_{T+1}$  is the one-step ahead predictor of  $Y_{T+1}$

Estimate  $\Delta_T(r)$ , for  $r = 1, 2, \dots, r^*$ , by subsampling, i.e.,

- Fix  $b$ , the subseries length, and get an estimate of the parameter vector  $\hat{\theta}_{T,b,t}$  by using  $Y_{b,t} = (Y_t, \dots, X_{t+b-1})'$  and  $\mathbf{X}_{b,t} = (\mathbf{X}'_t, \dots, \mathbf{X}'_{t+b-1})'$
- Obtain replicates of the one-step ahead predictor  $\left\{ \hat{Y}_{t+b}^{(t)}, t = 0, \dots, T-b \right\}$  by using the set of models  $\left\{ \hat{\theta}_{T,b,t} \right\}$
- Get the subsampling estimate as  $\hat{\Delta}_{T,b}(r) = \frac{1}{T-b+1} \sum_{t=0}^{T-b} \left( \hat{Y}_{t+b}^{(t)} - Y_{t+b} \right)^2$

Choose  $\hat{r} = \arg \min_r \hat{\Delta}_{T,b}(r)$ .

## 5 An illustrative example on simulated data

To illustrate how the proposed model selection procedure works, the results of an illustrative example on simulated data will be reported. The experimental setup is based on a dataset generated by an Exponential Autoregressive model of order 2, defined as

$$Y_t = (0.5 + 0.9 \exp(-Y_{t-1}^2)) Y_{t-1} - (0.8 - 1.8 \exp(-Y_{t-1}^2)) Y_{t-2} + \epsilon_t$$

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, 1990).

To capture the nonlinear dynamical structure of the data we approximate the data generating process by a proper chosen neural network model  $NN(d, r)$ .

The procedure to get the best neural network approximation can be implemented as follows.

SELECT THE RELEVANT VARIABLES

1. Fix  $r = 1$  and  $d = d^*$  where  $d^*$  is a proper chosen maximum number of lags.
2. Estimate the model  $NN(d^*, r)$ .
3. Plot the derivatives and the relevance measures to identify the set  $I_0$  of the irrelevant variables.
4. Test if the set of variables  $I_0$  is irrelevant (by subsampling)
5. Determine the 'optimal' value  $\hat{d}$ .

SELECT THE HIDDEN LAYER SIZE

1. Fix  $r = r^*$ , the maximum number of hidden units.
2. Estimate the models  $NN(\hat{d}, 1), \dots, NN(\hat{d}, r^*)$ .
3. Compute the predictive risk for each model (by subsampling).
4. Choose  $NN(\hat{d}, \hat{r})$  such that the predictive risk is minimum.

CHECK FOR OMITTED VARIABLES

1. Identify a set of possibly omitted variable.
2. Estimate the new neural model including the new set of variables.
3. Test if the two models give the same output (by subsampling).
4. If the two models are equivalent, choose the most parsimonious one.

To select the set of variables to be tested as irrelevant, we use simple graphical exploratory tools based on plots of the derivatives and plots of the relevance measures for each single lag. Values of the derivatives and of the relevance measures close to zero candidate that lag to be in the set of irrelevant ones. In our application we start with a tentative model  $NN(6, 1)$ . By the plots in figure 1 we identify, as possible relevant variables, lags 1 and 2. This seems to be confirmed by the plot of figure 2 where we report the values of the relevant measure. It is worthwhile to observe that the relevant lags to be chosen are not influenced by the number of units in the hidden layer. The qualitative impression given by the six plots is the same, with a clear cut between the first two lags and the others.

In order to implement a formal test procedure to confirm the exploratory identification of the first two lags as relevant, the subsampling approach is used and the identification of the subseries length is needed. By looking at figure 3,

in which the variance inflation indexes are reported, a value for  $b$  equal to 180 is reasonable. The results of the identification procedure are reported in table 1. Analyzing the the  $p$ -values of the tests, estimated by subsampling, we identify a neural network model with the first two lags as relevant. Again, the decision about the number and type of relevant lags remain stable for different values of the hidden layer size.

Now we can proceed to select a proper hidden layer size by using the predictive risk measured by the mean square error of prediction. In figure 4 we report the distributions of the MSE estimated by subsampling, for different values of the subseries length, corresponding to different hidden layer sizes ranging from 1 to 8. Clearly there is no improvement in the performances by using more than 5 neurons in the hidden layer. So the optimal identified model is  $NN(2, 5)$ . The neural network tests for neglected nonlinearity by Teraesvirta (Teraesvirta et al., 1993) and White (Lee et al., 1993) on the residuals from the identified optimal model, along with the Jarque-Bera test for normality, are reported in table 2. Clearly, the tests do not refuse the null and so the nonlinear structure of the data seems to corretly modeled. Moreover, the residuals can be considered as realization of a gaussian process.

To check if there are omitted variables in the identified model, we also compared neural models with different lag structures. The results are reported in table 3. Clearly, neural network models with more than two neurons in the input layer seem to equivalent to the chosen one. Therefore, there are no omitted lags. Observe that, on the contrary, if we had used a network including just one input neuron (corresponding to the first lag) there would have been better models obtained by including more lags.

## 6 Concluding remarks

The subsampling method for the selection of the neural network topology is presented with an illustrative example on simulated data. The procedure seems to be able to detect correctly the set of input variables and a proper hidden layer size. Clearly, joint usage of neural network models and subsampling is usually quite demanding from a computational point of view. In any case, once fixed the subseries length, the proposed procedure takes just few minutes on a desktop PC.

## References

- Anders, U. and Korn, O. (1999). Model selection in neural networks. *Neural Networks*, 12:309–323.
- Barron, A. R. (1993). Universal approximation bounds for superposition of a sigmoidal function. *IEEE Transactions on Information Theory*, 39:930–945.
- Baxt, W. and White, H. (1995). Bootstrapping confidence intervals for clinical

- input variable effects in a network trained to identify the presence of acute myocardial infarction. *Neural Comput.*, 7:624–638.
- Bertail, P., Haefke, C., Politis, D. N., and White, H. (2004). Subsampling the distribution of diverging statistics with applications to finance. *Journal of Econometrics*, 120:295–326.
- Bertail, P., Politis, D. N., and Romano, J. P. (1999). On subsampling estimators with unknown rate of convergence. *JASA*, 94:569–579.
- Carlstein, E. (1996). The use of subseries values for estimating the variance of a general statistic from a stationary time series. *Annals of Statistics*, 14:1171–1179.
- Fukuchi, J.-I. (1999). Subsampling and model selection in time series analysis. *Biometrika*, 86:591–604.
- LaRocca, M. and Perna, C. (2004). Variable selection in neural network regression models with dependent data: a subsampling approach. *Computational Statistics and Data Analysis*. To appear.
- Lee, T. H., White, H., and Granger, C. W. J. (1993). Testing for neglected nonlinearity in time series models. *Journal of Econometrics*, 56:269–290.
- Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Annals of Statistics*, 22:2031–2050.
- Politis, D. N., Romano, J. P., and Wolf, M. (1997). Subsampling for heteroscedastic time series. *Journal of Econometrics*, 81:281–317.
- Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer NY.
- Reed, R. (1993). Pruning algorithms—a survey. *Neural Networks*, 4:740–747.
- Refenes, A. N. and Holt, W. T. (2001). Forecasting volatility with neural regression: a contribution to model adequacy. *IEEE Transactions on Neural Networks*, 12(4):850–864.
- Refenes, A. P. N. and Zapranis, A. D. (1999). Neural model identification, variable selection and model adequacy. *Journal of Forecasting*, 18:299–332.
- Swanson, N., White, H., and Korn, O. (1995). A model selection approach to assessing the information in the term structure using linear models and artificial neural networks. *Journal of Business and Economic Statistics*, 13:265–275.
- Teraesvirta, T., Lin, C. F., and Granger, C. W. J. (1993). Power of the neural network linearity test. *Journal of Time Series Analysis*, 14:209–220.
- White, H. (1994). *Estimation, inference and specification analysis*. Cambridge University Press.

White, H. and Racine, J. (2001). Statistical inference, the bootstrap, and neural-network modeling with application to foreign exchange rates. *IEEE Transactions on Neural Networks*, 12:657–673.

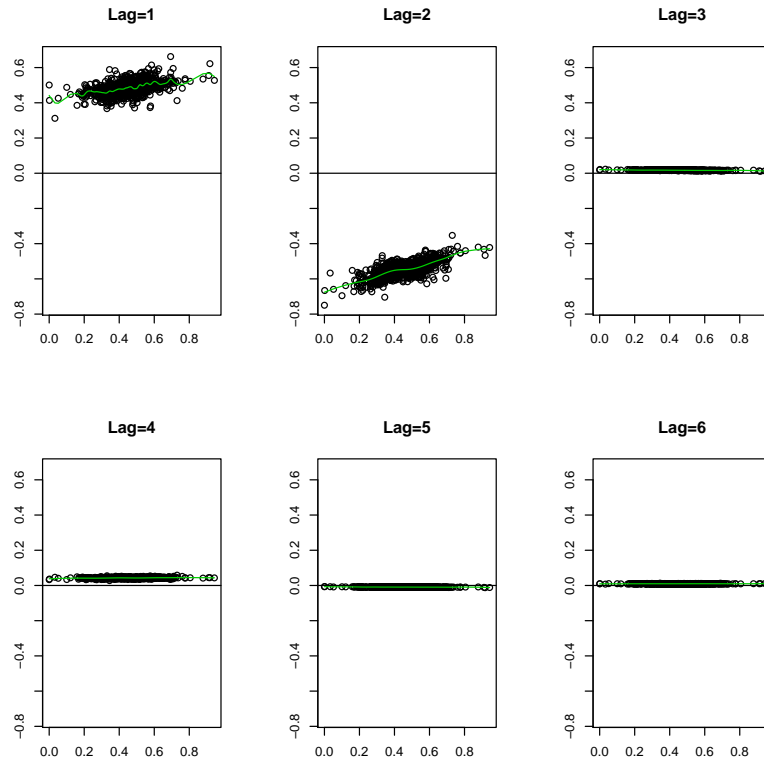


Figure 1: Plots of the derivatives for different lags. Neural network model  $NN(6, 1)$ .

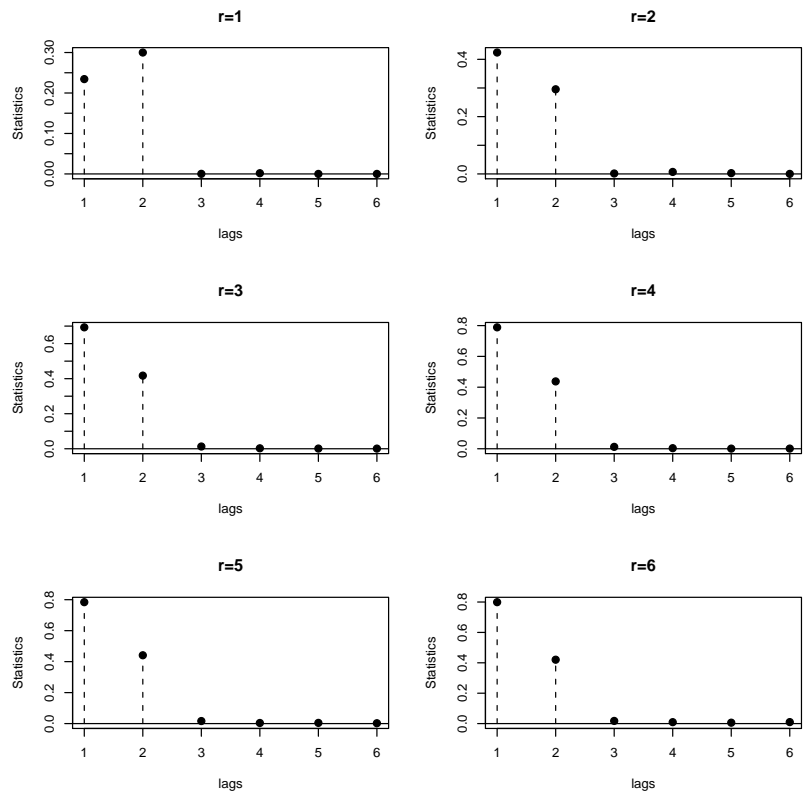


Figure 2: Plots of the values of the relevance measure for different lags and different hidden layer sizes.

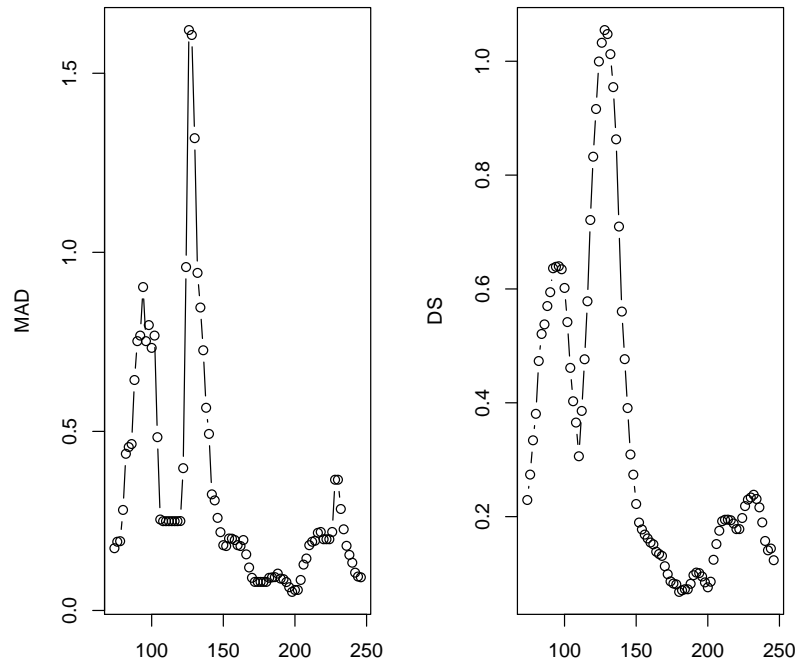


Figure 3: Values of the VIF indicator for different values of the block length.

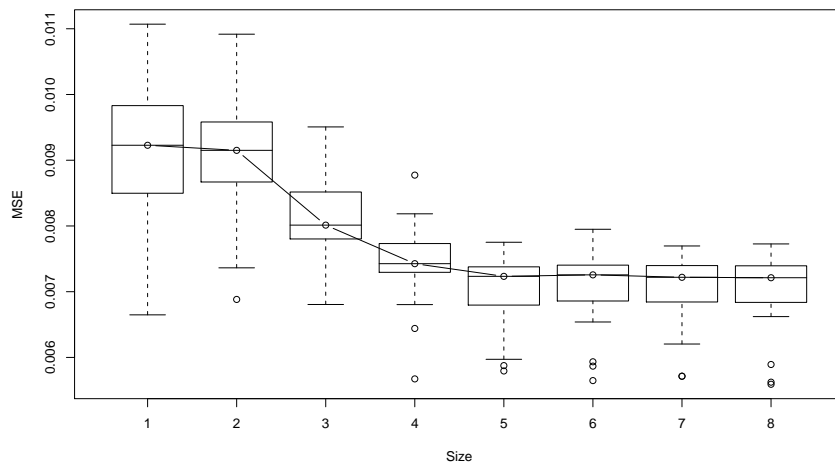


Figure 4: Boxplots of the distributions of the predictive accuracy measure for different values of the subseries length. Each distribution refers to a given hidden layer size.

$r$	$I_0$	{1}	{2}	{3}	{4}	{5}	{6}
1	Test Statistic	243.02	299.98	0.2859	1.8345	0.0842	0.0931
	$p$ -value	<i>0.0000</i>	<i>0.0000</i>	<i>0.7698</i>	<i>0.2692</i>	<i>0.8453</i>	<i>0.6760</i>
2	Test Statistic	421.28	316.86	1.6937	6.1157	2.6152	0.1861
	$p$ -value	<i>0.0000</i>	<i>0.0000</i>	<i>0.7393</i>	<i>0.2521</i>	<i>0.4604</i>	<i>0.9367</i>
3	Test Statistic	698.24	413.57	12.1104	2.6203	1.2834	1.0792
	$p$ -value	<i>0.0000</i>	<i>0.0000</i>	<i>0.0804</i>	<i>0.6236</i>	<i>0.7540</i>	<i>0.7333</i>
4	Test Statistic	781.28	446.56	12.9795	3.9536	1.0705	1.3363
	$p$ -value	<i>0.0000</i>	<i>0.0000</i>	<i>0.4994</i>	<i>0.7284</i>	<i>0.9622</i>	<i>0.9050</i>
5	Test Statistic	728.02	444.57	16.3791	3.7655	3.9889	2.8014
	$p$ -value	<i>0.0000</i>	<i>0.0000</i>	<i>0.4848</i>	<i>0.8940</i>	<i>0.8758</i>	<i>0.8149</i>
6	Test Statistic	790.27	429.05	16.7634	10.2584	7.3156	10.9096
	$p$ -value	<i>0.0000</i>	<i>0.0000</i>	<i>0.6821</i>	<i>0.6833</i>	<i>0.8161</i>	<i>0.7150</i>

Table 1: Values of the test statistic ( $p$ -values) for different variables set  $\mathcal{X}_0$ , (Subseries length  $b = 180$ ).

Test	ORIGINAL DATA		RESIDUALS	
	Statistic	<i>p</i> -value	Statistic	<i>p</i> -value
<i>Teraesvirta</i>	68.0038	0.0000	11.5667	0.1157
<i>White</i>	73.1060	0.0000	1.9058	0.3856
<i>Jarque-Bera</i>	84.4625	0.0000	1.1218	0.5707

Table 2: Teraesvirta and White neural network tests for neglected nonlinearity on the original data and on the residuals from the "optimal" estimated neural network model  $NN(2, 3)$  and alternative network models.

Model	$NN(1, 5)$	$NN(2, 5)$
$NN(2, 5)$	5.1924 (0.0000)	–
$NN(3, 5)$	5.2106 (0.0000)	0.0617 (0.3801)
$NN(4, 5)$	5.0483 (0.0000)	0.0724 (0.8478)
$NN(5, 5)$	5.2799 (0.0000)	0.1349 (0.6529)
$NN(6, 5)$	5.2909 (0.0000)	0.1599 (0.8977)

Table 3: Values of the test statistics and p-values (estimated by subsampling) in parenthesis for the hypotheses of equivalence of neural network models.