ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 117-122

Specification, estimation and evaluation of single hidden-layer feedforward autoregressive artificial neural network models

Gianluigi Rech
Department of Economic Statistics, Stockholm School of Economics,
Box 6501, S-113 83, Stockholm, Sweden. E-mail: stgr@hhs.se

Abstract

This paper considers artificial neural network modelling from a statistical point of view. Specification, estimation and evaluation are carried out using Lagrange multiplier testing. Simulations in samples of moderate size demonstrate the performances of the overall procedure.

Acknowledgments: I am grateful to the Tore Browaldh's Foundation for financial support.

1. Specification and estimation

The capability of single hidden-layer feedforward neural networks (hereafter NN) of approximating any Borel-measurable function to any degree of accuracy has been pointed out in Hornik, Stinchcombe, and White (1989). Nonlinear features in time series can then be successfully modelled applying statistical tools to the data of interest, since the connection between NN and statistics is generally well accepted. I call the following model an autoregressive NN model of order k with q hidden units and a linear component:

$$y_t = \boldsymbol{\alpha}'\mathbf{w}_t + \sum_{j=1}^{q} \beta_j \psi(\boldsymbol{\gamma}_j'\mathbf{w}_t) + u_t, \ t = 1, ..., T. \tag{1.1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_k, \alpha_0)'$, $\boldsymbol{\beta} = \left(\beta_1, ..., \beta_q\right)'$, $\boldsymbol{\gamma}_j = \left(\gamma_{j0}, \gamma_{j1}, ..., \gamma_{j,k-1}, c_j\right)'$, $j = 1, ..., q$; $\mathbf{w}_t = (w_{1t}, w_{2t}, ..., w_{kt}, 1)' = (y_{t-1}, y_{t-2}, ..., y_{t-k}, 1)'$ and u_t is $n.i.d.$. Here, I assume for simplicity that all lags from 1 to k enter model (1.1).

In time series applications, the first issue is the choice of lags to be included in the model. I do this using a semiparametric method as in Rech, Teräsvirta, and Tschernig (1999); hereafter RTT. The authors utilize an h-th order Taylor expansion of the process. Such expansion contains all possible combinations of lags from 1 to k, the maximum lag encompassed, the idea being that the terms involving redundant lags have zero coefficients. This fact is used to omit the

ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 117-122

redundant variables from the model. The variable selection procedure works as follows. First, regress y_t on all variables (products of lags from 1 to k) in the Taylor expansion and compute the value of the SBIC criterion (Rissanen (1978), Schwarz (1978)). Next, leave one lag out from the original model, regress y_t on all products of remaining lags in the Taylor expansion and compute the value of the criterion. Repeat this by omitting each lag in turn. Continue by simultaneously omitting two lags from the original model. Proceed until y_t is just white noise. This amounts to estimating $\sum_{i=1}^{k}\binom{k}{i}+1=2^k$ linear models by ordinary least squares. The combination of lags that yields the lowest value of the model selection criterion is selected. SBIC is consistent in the sense that if there is a finite set of correct lags and if the true model is exactly a h-th order Taylor expansion, the correct lags will asymptotically be selected with probability one. Although the number of regressors grows exponentially with k and h, RTT show that the procedure works well already in small samples and can be applied even in large samples where the nonparametric model selection techniques become computationally infeasible.

The second step is linearity testing. I follow Teräsvirta, Lin, and Granger (1993), where the authors utilize a third-order Taylor expansion to approximate the nonlinear hidden units. This is equivalent to testing for 0 hidden units against 1. Once linearity is rejected, the choice of the number of hidden units determines the goodness of the approximation to the true data generating process (hereafter DGP). It is therefore important to utilize an appropriate method for this choice. I test a NN model with q hidden units against a model with $q+1$ ones, $q=0,1,2,....$, carrying out a sequence of Lagrange Multiplyer (hereafter LM) tests from specific to general until the first acceptance of the null hypothesis of q hidden units. LM testing is based on the likelihood function restricted under H_0. An extensive description of LM testing can be found in Godfrey (1988). The $(q+1)$-th hidden unit is approximated by a Taylor series to the third order around the point $\boldsymbol{\gamma}_{q+1}$=0 as in Teräsvirta, Lin, and Granger (1993). We call the first, second and third order coefficients λ_i, λ_{ij}, λ_{ijl}, respectively. This is done to avoid identification problems since if β_{q+1} equals zero, model (1.1) is not identified:

$$\begin{aligned} y_t &= \boldsymbol{\alpha}^{*\prime}\mathbf{w}_t+\sum_{j=1}^{q}\beta_j\psi(\boldsymbol{\gamma}_j,\mathbf{w}_t)+\sum_{i=1}^{k}\sum_{j\geq i}^{k}\lambda_{ij}^{*}w_{ti}w_{tj} \\ &\quad+\sum_{i=1}^{k}\sum_{j\geq i}^{k}\sum_{l\geq j}^{k}\lambda_{ijl}^{*}w_{ti}w_{tj}w_{tl}+u_t^{*} \end{aligned} \tag{1.2}$$

where $\boldsymbol{\alpha}^{*}=\boldsymbol{\alpha}+\boldsymbol{\lambda}^{*}$, $\lambda_i^{*}=\beta_{q+1}\lambda_i$, $\lambda_{ij}^{*}=\beta_{q+1}\lambda_{ij}$, $\lambda_{ijh}^{*}=\beta_{q+1}\lambda_{ijl}$, $u_t^{*}=\beta_{q+1}\overline{R}_3(\boldsymbol{\gamma}_{q+1},\mathbf{w}_t;0)+u_t$, where $\overline{R}_3(\boldsymbol{\gamma}_{q+1},\mathbf{w}_t;0)$ is the rest of the third-order Taylor expansion of the $(q+1)$-th hidden unit. The null hypotesis corresponds to $\lambda_{ij}^{*}=0, i=1,...,k; j=i,...,k;\ \lambda_{ijl}^{*}=0, i=1,...,k; j=i,...,k; l=j,...,k$ in (1.2). I define the estimated residuals under the null hypothesis, $\widehat{v}_t=y_t-\hat{\boldsymbol{\alpha}}^{*\prime}\mathbf{w}_t-\sum_{j=1}^{q}\hat{\beta}_j\psi(\hat{\boldsymbol{\gamma}}_j,\mathbf{w}_t)$, where "$\widehat{\ }$" denotes a consistent estimator. The test can be performed in three stages:

ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 117-122

(i) compute the estimated residuals $\widehat{v}_t$ and the corresponding sum of the squared residuals $SSR_0 = \sum \widehat{v}_t^2$;

(ii) regress $\widehat{v}_t$ on w_t, $\psi(\widehat{\boldsymbol{\gamma}}_j, \mathbf{w}_t)$, $\left(\partial\psi(\boldsymbol{\gamma}_j, \mathbf{w}_t)/\partial\boldsymbol{\gamma}_j\right)_{\boldsymbol{\gamma}_j=\widehat{\boldsymbol{\gamma}}_j}$, $j = 1, ..., q$, and all the terms at the second and third power, whose number I call m and compute the residuals $\widehat{v}_t'$ and their sum $SSR = \sum \widehat{v}_t'^2$;

(iii) compute

$$F = \frac{(SSR_0 - SSR)/m}{SSR/(T-k-1-m)}$$

Under the null, F is approximately distributed as an $F_{m,T-k-1-m}$.

The estimation procedure for model (1.1) resembles the 2-steps algorithm in White (1989). In order to improve its accuracy, a set of starting values is chosen, linearizing the model and using ordinary least squares. The nonlinear least squares procedure is subsequently carried out by the Broyden-Fletcher-Goldfarb-Shanno algorithm.

Summing up, the overall device works as follows. The set of variables to be included in the model is selected as in RTT. The hypothesis of no nonlinear hidden units (linear model) is tested at a given significance level α. If rejected, a model with a linear part and one hidden unit is estimated and the approximate t-values of the parameters computed, approximating the covariance matrix of the parameters by the outer product of the gradient matrix. The lags with low t-values are removed and the model re-estimated. The whole procedure is redone until the hidden unit contains only significant estimates. Subsequently, the hypothesis of no additional hidden units is tested at the significance level $\alpha/2$. If rejected, a model with two hidden units is estimated, and the dimension of the model reduced by checking the t-values of its estimates as above. The procedure continues halving the significance level again to $\alpha/4$, $\alpha/8$, ..., stopping the procedure at the first acceptance of the null hypothesis of no additional hidden units. Letting the significance level converge to zero as $q \to \infty$ keeps the dimension of the model under control.

2. Evaluation

Evaluating a model requires, as in Eitrheim and Teräsvirta (1996), to develop specific LM tests for the hypotesis of no error autocorrelation and parameter constancy, while additional nonlinearity is already checked when I choose the number of hidden units. The test for error autocorrelation is based on model (1.1), where the residuals u_t follow an autoregressive process of order r, $u_t = \sum_{j=1}^{r} a_j u_{t-j} + \varepsilon_t$, $\varepsilon_t \sim n.i.d.(0, \sigma^2)$. The corresponding LM test for the hypotesis $H_0 : \mathbf{a} = 0$ can be carried out in 3 steps as in testing for q against $q+1$ hidden units. As to parameter constancy, I generalize model (1.1) assuming that the hidden units have constant parameters whereas both βs and αs may change smoothly over time. Therefore $\boldsymbol{\beta}(t) = \beta_0 + \boldsymbol{\lambda}_1 F_j(t, \gamma_1, c_1)$ and $\boldsymbol{\alpha} = \boldsymbol{\alpha}(t) = \alpha_0 + \boldsymbol{\lambda}_2 F_j(t, \gamma_1, c_1)$. The null hypothesis of parameter constancy implies that $F_j(t, \gamma_1, c_1) \equiv$ constant for any t. I consider three possible functional forms for

ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 117-122

the transitional function F_j:

$$F_1(t,\gamma_1,c_1) = \left(1+\exp\left\{-\gamma_1(t-c_1)\right\}\right)^{-1} - 1/2 \tag{2.1}$$

$$F_2(t,\gamma_1,c_1) = 1-\exp\left\{-\gamma_1(t-c_1)^2\right\} \tag{2.2}$$

$$F_3(t,\gamma_1,\underline{c}_1) = \left(1+\exp\left\{-\gamma_1(t-c_1)(t-c_2)(t-c_3)\right\}\right)^{-1} - 1/2 \tag{2.3}$$

I derive a test statistic for the most general case (2.3) . To circumvent the identification problem as in (1.2), I take the first order Taylor expansion of (2.3) about $\gamma_1 = 0$, approximating $\boldsymbol{\beta}(t)$ and $\boldsymbol{\alpha}(t)$ by powers of t. Combining terms in a similar way than in (1.2), I set up three LM tests, LM_i, $i = 1,2,3$. Their null hypothesis require that terms involving powers of t from 1 to i, respectively, have zero coefficients.

3. Simulations

The simulated series, whose lenght is $T = 200$, are generated by the following NN DGP with 2 hidden units and a linear part, time-varying parameters and autocorrelated errors:

$$\begin{aligned} y_t &= H_1(t)\left(0.07+0.2y_{t-1}-0.1y_{t-2}\right) \\ &\quad +0.53H_2(t)\left(1+\exp\{-17(y_{t-1}-1.81y_{t-2}-0.07)\}\right)^{-1} \\ &\quad -0.37H_3(t)\left(1+\exp\{-20(y_{t-1}+2.4y_{t-4}+0.07)\}\right)^{-1}+u_t \end{aligned} \tag{3.1}$$

$$u_t = \rho u_{t-1}+\varepsilon_t, \varepsilon_t \sim N(0,10^{-2}); \rho = 0, 0.1, 0.2 \tag{3.2}$$

$$H_1(t) = \left(1+\exp\left(-2\left(t-100.5\right)\right)\right)-0.5$$

$$H_j(t) = \left(1-\exp\left(-2\left(j-1\right)\left(t-100.5\left(j-1\right)\right)^2\right)\right), j = 2,3$$

I investigate power and size of the evaluation tools in the cases of time-varying parameters and constant ones ($H_j(t) \equiv 1$ for any j and any t in 3.1), autocorrelated and $n.i.d.$ errors ($\rho = 0$ in 3.2). When the DGP has constant parameters (graphs 1-6), the size distortion of the tests for no autocorrelation and parameter constancy are negligible, and its power good even in small samples. This is a general result since parameter constancy tests have some power towards the hypothesis of no autocorrelation. Introducing nonconstant parameters (graphs 7-10) biases the size of the tests for no autocorrelation when the null is true, while their power is still a positive function of ρ. As to parameter constancy tests, the power is good both for $\rho = 0$ and.$\rho = 0.1$.

4. Conclusions

I have discussed a statistically consistent specification technique for NN models which can be applied to a wide range of nonlinear processes. If linearity is not rejected, a single hidden layer NN model is fitted to the data and evaluated in

ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 117-122

a simple but statistically consistent way. Simulations were run to demonstate the performances of the evaluation tests when the true DGP is an NN model, results being satisfactory already at sample size $T = 200$. The whole procedure can be utilized by model builders interested in nonlinear time series modelling.

References

EITRHEIM, Ø., AND T. TERÄSVIRTA (1996): "Testing the Adequacy of Smooth Transition Autoregressive Models," *Journal of Econometrics*, **74**, 59–75.

GODFREY, L. (1988): *Misspecification Tests in Econometrics*. University Press, Cambridge.

HORNIK, K., M. STINCHCOMBE, AND H. WHITE (1989): "Multi-Layer Feedforward Networks and Universal Approximations," *Neural Networks*, 2, 359–66.

RECH, G., T. TERÄSVIRTA, AND R. TSCHERNIG (1999): "A Simple Variable Selection Method for Nonlinear Models," *SSE/EFI Working Paper Series in Economics and Finance, No. 296.*

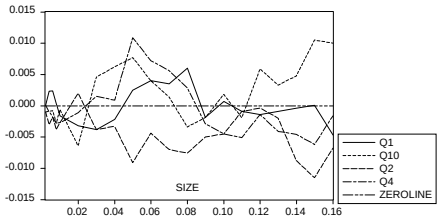
RISSANEN, J. (1978): "Modeling by Shortest Data Description," *Automatica*, 14, 465–471.

SCHWARZ, G. (1978): "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–64.

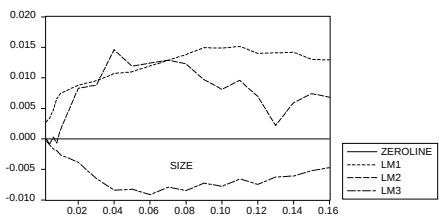
TERÄSVIRTA, T., C. LIN, AND C. GRANGER (1993): "Power of the Neural Network Linearity Test," *Journal of Time Series Analysis*, **14**, 209–220.

WHITE, H. (1989): "Some Asymptotic Results for Learning in Single Hidden-Layer Feedforward Network Models," *Journal of the American Statistical Association*, 84(408), 1003–1013.

Graphs 1-6. Size discrepancy plots and size-power curves of the test of no error autocorrelation up to lags 1 (Q1), 2 (Q2), 4 (Q4) and 10 (Q10), and the three tests for parameter constancy LM_1, LM_2, LM_3 against smooth structural change at the sample size $T = 200$, for 1000 replications of the series generated by process (3.1) with constant parameters:

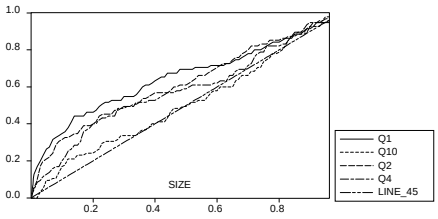


1 p-value discrepancy plots, $\rho = 0$

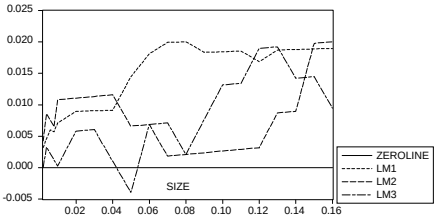


2 p-value discrepancy plots, $\rho = 0$

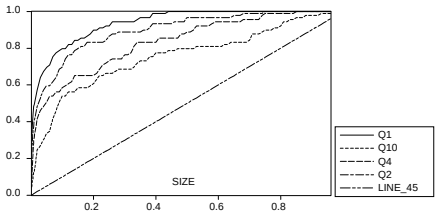
ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 117-122



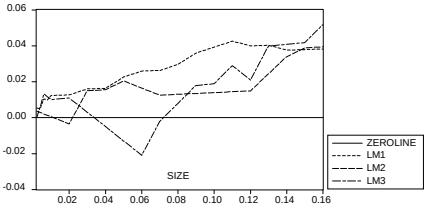
3 size-power curves, $\rho = 0.1$



4 p-value discrepancy plots, $\rho = 0.1$

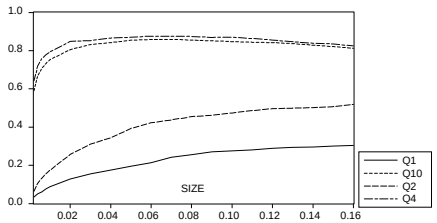


5 size-power curves, $\rho = 0.2$

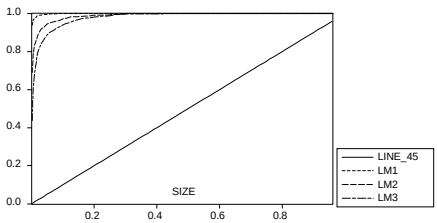


6 p-value discrepancy plots, $\rho = 0.2$

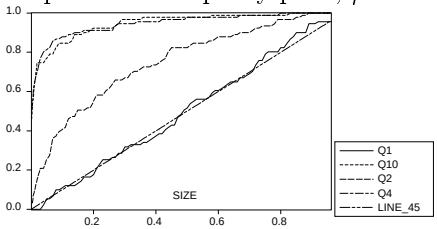
Graphs 7-10. Size discrepancy plots and size-power curves of the test of no error autocorrelation up to lags 1 (Q1), 2 (Q2), 4 (Q4) and 10 (Q10), and the three tests for parameter constancy LM_1, LM_2, LM_3 against smooth structural change at the sample size $T = 200$, for 1000 replications of the series generated by process (3.1) with time-varying parameters:



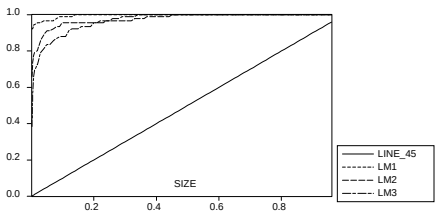
7 p-value discrepancy plots, $\rho = 0$



8 size-power curves, $\rho = 0$



9 size-power curves, $\rho = 0.1$



10 size-power curves, $\rho = 0.1$