New Procedures for Model Selection in Feedforward Neural Networks for Time Series Forecasting

Suhartono
Statistics Department Institut Teknologi Sepuluh Nopember

ABSTRACT

The aim of this paper is to propose two new procedures for model selection in Neural Networks (NN) for time series forecasting. Firstly, we focused on the derivation of the asymptotic properties and asymptotic normality of NN parameters estimator. Then, we developed the model building strategies based on statistical concepts particularly statistics test based on the Wald test and the inference of $R^2_{incremental}$. In this paper, we employ these new procedures in two main approaches for model building in NN, i.e. fully bottom-up or forward scheme by using the inference of $R^2_{incremental}$, and the combination between forward (by using the inference of $R^2_{incremental}$) and top-down or backward (by implementing Wald test). Bottom-up approach starts with an empty model, whereas top-down approach begins with a large NN model. We used simulation data as a case study. The results showed that a combination between statistical inference of $R^2_{incremental}$ and Wald test was an effective procedure for model selection in NN for time series forecasting.

Keywords: Time series, neural networks, asymptotic normality, Wald test, $R^2_{incremental}$

INTRODUCTION

In recent years, an impressive array of publications has appeared claiming considerable successes of neural networks (NN) in data analysis and engineering applications. NN model is a prominent example of such a flexible functional form. The use of the NN model in applied work is generally motivated by a mathematical result stating that under mild regularity conditions, a relatively simple NN model is capable for approximating any Borel-measureable function to any given degree of accuracy (see e.g. Hornik *et al.* 1989, 1990).

In the application of NN, it contains limited number of parameters (weights). How to find the best NN model, i.e. how to find an accurate combination between number of input variables and nodes in hidden layer, is a central topic on the some NN literatures that discussed on many articles and books (see e.g. Bishop 1995, Haykin 1999, Ripley 1996). In general, there are two procedures usually used to find the best NN model (the optimal architecture), those are "general-to-specific" or "top-down" and "specific-to-general" or "bottom-up" procedures. "Top-down" procedure is started from complex model and then applies an algorithm to reduce number of parameters by using some stopping criteria, whereas "bottom-up" procedure works from a simple model. The first procedure in some literatures is also known as "pruning" (see Reed 1993), or "backward" method in statistical modeling. The second procedure is also known as "constructive learning" and one of the most popular is "cascade correlation" (see e.g. Fahlman & Lebiere 1990, Prechelt 1997), and it can be seen as "forward" method in statistical modeling.

The aim of this paper is to discuss and propose two new procedures for model selection in FFNN for time series forecasting. These procedures are developed based on the inference of R2 incremental and Wald test. The inference of R2 incremental is implemented on forward scheme, whereas Wald test is employed on backward scheme. We emphasize on the used of NN for time series forecasting.

Feedforward neural networks

Feed forward Neural Networks (FFNN) is the most popular NN models for time series forecasting applications. Figure 1 shows a typical three-layer FFNN used for forecasting purposes. The input nodes are the previous lagged observations, while the output provides the forecast for the future values. Hidden nodes with appropriate nonlinear transfer functions are used to process the information received by the input nodes. The model of FFNN in Figure 1 can be written as:

$$y_t = \beta_0 + \sum_{j=1}^{q} \beta_j \psi\left(\sum_{i=1}^{p} \gamma_{ij} y_{t-i} + \gamma_{oj}\right) + \varepsilon_t \quad \ldots\ldots (1)$$

where p is the number of input nodes, q is the number of hidden nodes, ψ is a sigmoid transfer function such as the logistic, $\{\beta_j, j = 0,1,\cdots,q\}$ is a vector of weights from the hidden to output nodes and $\{\gamma_{ij}, i = 0,1,\cdots,p; \quad j = 1,2,\cdots,q\}$ are weights from the input to hidden nodes. Note that equation (1) indicates a linear transfer function is employed in the output node.

Functionally, the FFNN expressed in equation (1) is equivalent to a nonlinear AR model. This simple structure of the network model has been shown to be capable of approximating arbitrary function (see e.g. Hornik *et al*. 1989,1990). However, few practical guidelines exist for building a FFNN for a time series, particularly the specification of FFNN architecture in terms of the number of input and hidden nodes is not an easy task.

Kaashoek & Van Dijk (2002) introduced a "pruning" procedure by implementing three kinds of methods to find the best FFNN model; those are incremental contribution(R2 incremental), principal component analysis, and graphical analysis. Whereas, Swanson and White (1995,1997) applied a criterion of model selection, SIC, on "bottom-up" procedure to increase number of nodes in hidden layer and input variables until finding the best FFNN model.

Recently, Suhartono *et al.* (2006) proposed a new forward procedure based on the statistical inference of R2 incremental contribution.

Backpropagation algorithm

Backpropagation algorithm is an algorithm that usually used to estimate the FFNN weights (parameters). Ripley (1996) stated that the existence of the function approximation was not useful if there was not known the way to find this function. This condition affected many researches about NN for many years.

The main idea of the approximation by using NN is started by Rumelhart-McClelland learning for fitting parameters by employing least squares method. The training of the NN involves adjusting the weights of the network such that the output generated by the network for the given input $x_{(k)}$ is as "close" to $\hat{y}_{(k)} = f(x; w)$ as possible. Formally, this can be formulated as the optimization problem by finding weights, $w = (\gamma_{ij}, \beta_j)$, to minimize

$$E(w) = \sum_{k=1}^{n} \left\| y_{(k)} - f(x_{(k)}; w) \right\|^2 \quad \ldots\ldots\ldots\ldots\ldots (2)$$

as done in nonlinear regression (see e.g. Bates & Watts 1988, Seber & Wild 1989).

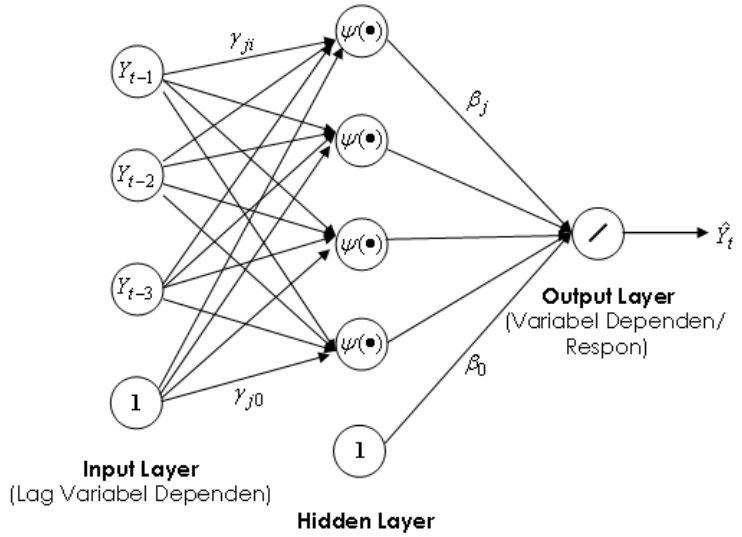

Figure 1. Architecture of neural network model with single hidden layer, i.e three input units, four nodes in the hidden layer, and one output unit.

.

Gradient descent is known as one of the oldest optimization methods. This method is based on a linear approximation of the error function given by

$$E(w+\Delta w) \approx E(w) + \Delta w^T E'(w) \quad \ldots\ldots\ldots\ldots\ldots(3)$$

The weights update is

$$\Delta w = -\eta E'(w), \quad \eta > 0, \quad \ldots\ldots\ldots\ldots\ldots\ldots(4)$$

where η is learning rate. Suhartono *et al.* (2005) derived a corollary about back-propagation algorithm to find the optimal weights of FFNN for time series forecasting as illustrated in Figure 1.

Asymptotic properties of FFNN estimator

The large-sample properties of learning backpropagation in single hidden layer feedforward networks have been studied further by White (1989a, 1989b). The aim of learning networks by using backpropagation is to find the solution w^* on the optimization problem $\arg\min_{w \in W} Q(w)$, i.e.

$$w^* = \arg\min_{w \in W} \left(Q(w) = E[(Y - f(X,w))^2 / 2 \right) \ldots\ldots(5)$$

where w^* is index of an optimal networks.

With squared error penalty, learning must arrive at w^*, which solve

$$\min_{w \in W} \Big(E[(Y - f(X,w))^2 / 2] = E([Y - E(Y \mid X)]^2 / 2)$$

$$+ E([E(Y \mid X) - f(X,w)]^2 / 2) \Big) \ldots\ldots\ldots\ldots\ldots(6)$$

Finding w^* is precisely the problem of finding the parameters of an optimal least squares approximation to $E(Y \mid X)$, the conditional expectation of Y given X. Specifically, given target/input pairs (Y_t, X_t) with $t = 1, 2, \ldots, n$, randomly drawn from the operating environment, then $\hat{w}_n$ is the nonlinear least squares estimator, i.e.

$$\hat{w}_n = \arg\min_{w \in W} Q_n(w) = n^{-1} \sum_{t=1}^{n} (Y_t - f(X_t, w))^2 / 2 \ldots\ldots(7)$$

Nonlinear regression is an established method that has been completely analyzed in statistics and econometrics literatures.

White (1989b) provided a formal statement of condition sufficient to guarantee convergence of $\hat{w}_n$, as stated in the following theorem.

Theorem 2.1. (White 1989b). *Let* (Ω, F, P) *be a complete probability space on which is defined the sequence of independent identically distributed random variables* $\{Z_t\} = (Z_t : \Omega \to \Re^v,\ t = 1, 2, \ldots)$, $v \in \mathbb{N} \equiv \{1, 2, \ldots\}$. *Let* $l : \Re^v \times W \to \Re$ *be a function such that for each* w *in* W, *a compact subset of* $\Re^s$, $s \in \mathbb{N}$, $l(\cdot, w)$ *is measurable-*B^v (*where* B^v *is the Borel* σ*-field generated by the open sets of* $\Re^v$), *and for each* z *in* $\Re^v$, $l(z, \cdot)$ *is continuous on* W. *Suppose further that there exists* $d : \Re^v \to \Re^+$ *such that for all* w *in* W, $|l(z,w)| \le d(z)$ *and* $E(d(Z_t)) < \infty$ (*i.e.,* l *is dominated on* W *by an integrable function*). *Then for each* $n = 1, 2, \ldots$ *there exists a solution* $\hat{w}_n$ *to the problem* $\min_{w \in W} \hat{Q}_n(w) \equiv n^{-1} \sum_{t=1}^{n} l(Z_t, w)$ *and* $\hat{w}_n \to W^*$ *a.s.–P, where* $W^* \equiv \{w^* \in W : Q(w^*) \le Q(w)$ *for all* $w \in W\}$, $Q(w) = E(l(Z_t, w))$.

Asymptotic normality of FFNN estimator

The appropriate formal concept for studying the limiting distribution of $\hat{w}_n$ is that of convergence in distribution. Asymptotic distribution of $\hat{w}_n$ depends on the nature of W^*. In general, W^* may consist of isolated points and/or isolated "flat". If convergence to a flat occurs, then the estimated weights $\hat{w}_n$ have a limiting distribution that can be analyzed using the theory of Phillips (1989) for "partially identified" models. These distributions belong to the "limiting mixed Gaussian" (LMG) family introduced by Phillips. When w^* is locally unique, the model is said to be "locally identified" and estimated weights $\hat{w}_n$ converging to w^* have a limiting multivariate normal distribution.

The condition ensuring that $\hat{w}_n$ is the multivariate normal distribution have been studied further by White (1989b). The following theorem is one of the results of White's works.

Theorem 2.2. (White 1989b). *Let* (Ω, F, P), $\{Z_t\}$, W *and* l *be as in Theorem* 2.1, *and suppose that* $\hat{w}_n \to w^*$ *a.s.* $-P$ *where* w^* *is an isolated element of* W^* *interior to* W.

Suppose in addition that for each z *in* $\Re^v$, $l(z, \cdot)$ *is continuously differentiable of order 2 on* $\int W$; *that* $E(\nabla l(Z_t, w^*)' \nabla l(Z_t, w^*)) < \infty$;

that each element of $\nabla^2 l$ *is dominated on* W *by an integrable function; and that* $A^* \equiv E(\nabla^2 l(Z_t, w^*))$ *and* $B^* \equiv E(\nabla l(Z_t, w^*)\nabla l(Z_t, w^*)')$ *are nonsingular* $(s \times s)$ *matrices, where* ∇ *and* ∇^2 *denote the* $(s \times 1)$ *gradient and* $(s \times s)$ *Hessian operators with respect to* w.

Then $\sqrt{n}(\hat{w}_n - w^*) \xrightarrow{d} N(0, C^*)$, *where* $C^* = A^{*-1}B^*A^{*-1}$. *If in addition each element of* $\nabla l \nabla l'$ *is dominated on* W *by an integrable function, then* $\hat{C}_n \to C^*$ *a.s.* $-P$, *where* $\hat{C}_n = \hat{A}_n^{-1}\hat{B}_n\hat{A}_n^{-1}$, *and*

$$\hat{A}_n = \frac{\sum_{t=1}^{n} \nabla^2 l(Z_t, \hat{w}_n)}{n},$$

$$\hat{B}_n = \frac{\sum_{t=1}^{n} \nabla l(Z_t, \hat{w}_n)\nabla l(Z_t, \hat{w}_n)'}{n}.$$

White (1989a) stated that taking one Nonlinear Least Squares (NLS) Newton-Raphson step from the backpropagation estimator asymptotically equivalent to NLS. Thus, tests of hypotheses bases on $\hat{w}_n$ can be conducted for selecting the optimal architecture of FFNN.

Hypothesis testing by using wald test

The Wald statistic allows the simplest analysis, although it may or may not the easiest statistic to compute in a given situation. The motivation for the Wald statistic is that when the null hypothesis is correct $\mathbf{S}\hat{\mathbf{w}}_n$ should be close to $\mathbf{S}\mathbf{w}^* = \mathbf{s}$, so a value of $\mathbf{S}\hat{\mathbf{w}}_n - \mathbf{s}$ far from zero is evidence against the null hypothesis.

The theorem about Wald statistic that be used for hypothesis testing of parameters in NN model is constructed as the following results.

Theorem 2.3. (Suhartono 2007) *Let the conditions of Theorem 2.2 hold, i.e.*

(i) $\mathbf{C}^{*-1/2}\sqrt{n}(\hat{\mathbf{w}}_n - \mathbf{w}^*) \xrightarrow{d} N(\mathbf{0}, \mathbf{I})$,

where $\mathbf{C}^* \equiv \mathbf{A}^{*-1}\mathbf{B}^*\mathbf{A}^{*-1'}$, *and* $\mathbf{C}^{*-1}$ *is* $O(1)$,

(ii) *there exists a matrix* $\hat{\mathbf{B}}_n$ *positive semidefinite and symmetric such that* $\hat{\mathbf{B}}_n - \mathbf{B}^* \xrightarrow{p} \mathbf{0}$. *Then* $\hat{\mathbf{C}}_n - \mathbf{C}^* \xrightarrow{p} \mathbf{0}$, *where* $\hat{\mathbf{C}}_n = \hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n\hat{\mathbf{A}}_n^{-1}$, *and*

$$\hat{\mathbf{A}}_n = \frac{\sum_{t=1}^{n} \nabla^2 l(Z_t, \hat{\mathbf{w}}_n)}{n}, \quad \hat{\mathbf{B}}_n = \frac{\sum_{t=1}^{n} \nabla l(Z_t, \hat{\mathbf{w}}_n)\nabla l(Z_t, \hat{\mathbf{w}}_n)'}{n}.$$

and, let $\text{rank}(S) = q \le k$. *Then under* $H_0 : \mathbf{S}w^* = s$,

(i) $\mathbf{\Gamma}_n^{-1/2}\sqrt{n}(\mathbf{S}\hat{w}_n - s) \xrightarrow{d} N(\mathbf{0}, \mathbf{I})$,

where $\mathbf{\Gamma}_n \equiv \mathbf{S}\mathbf{C}^*\mathbf{S}' = \mathbf{S}\mathbf{A}^{*-1}\mathbf{B}^*\mathbf{A}^{*-1'}\mathbf{S}'$.

(ii) *The Wald statistic*

$$W_n \equiv n(\mathbf{S}\hat{w}_n - s)'\hat{\mathbf{\Gamma}}_n^{-1}(\mathbf{S}\hat{w}_n - s) \xrightarrow{d} \chi_q^2 \quad \ldots\ldots(8)$$

where $\hat{\mathbf{\Gamma}}_n \equiv \mathbf{S}\hat{\mathbf{C}}_n\mathbf{S}'$.

Proof: We use Corollary 4.24, Proposition 2.30 and Theorem 4.30 in White (1999) to prove Theorem 2.3 and the results are as follow:

(i) Under H_0, $\mathbf{S}\hat{\mathbf{w}}_n - \mathbf{s} = \mathbf{S}(\hat{\mathbf{w}}_n - \mathbf{w}^*)$, so

$$\mathbf{\Gamma}_n^{-1/2}\sqrt{n}(\mathbf{S}\hat{\mathbf{w}}_n - \mathbf{s}) = \mathbf{\Gamma}_n^{-1/2}\mathbf{S}\mathbf{C}^{*1/2}\mathbf{C}^{*-1/2}\sqrt{n}(\hat{\mathbf{w}}_n - \mathbf{w}^*).$$

It follows from Corollary 4.24 in White (1999) that $\mathbf{A}_n = \mathbf{S}$ and $\mathbf{b}_n = \sqrt{n}(\hat{\mathbf{w}}_n - \mathbf{w}^*)$, so that $\mathbf{\Gamma}_n^{-1/2}\sqrt{n}(\mathbf{S}\hat{\mathbf{w}}_n - \mathbf{s}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I})$.

(ii) Based on Theorem 2.2 we have that $\hat{\mathbf{C}}_n - \mathbf{C}^* \xrightarrow{a.s} 0$, so it imply that $\hat{\mathbf{C}}_n - \mathbf{C}^* \xrightarrow{p} 0$. By using Proposition 2.30 in White (1999), where $\hat{\mathbf{\Gamma}}_n = g(\hat{\mathbf{C}}_n)$ and $\mathbf{\Gamma}_n = g(\mathbf{C}^*)$, so that $\hat{\mathbf{\Gamma}}_n - \mathbf{\Gamma}_n \xrightarrow{p} 0$. Given the result in (i), i.e.

$\mathbf{\Gamma}_n^{-1/2}\sqrt{n}(\mathbf{S}\hat{\mathbf{w}}_n - \mathbf{s}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I})$,

so by implementing Theorem 4.30 at [19], this yields

$W_n \equiv n(\mathbf{S}\hat{w}_n - s)'\hat{\mathbf{\Gamma}}_n^{-1}(\mathbf{S}\hat{w}_n - s) \xrightarrow{d} \chi_q^2$.

Thus, a test about the relevance (significance) of input where the hypothesis can be stated as $H_0 : \mathbf{S}w^* = 0$ against $H_1 : \mathbf{S}w^* \neq 0$, can be evaluated by applying Theorem 2.3. As an example, Wald statistic to evaluate this hypothesis testing is

$\hat{W}_n = n\hat{w}_n'\mathbf{S}'(\mathbf{S}\mathbf{C}^*\mathbf{S}')^{-1}\mathbf{S}\hat{w}_n$, where $\mathbf{C}^*$ as defined in earlier section.

Statistically inference of R^2 incremental

Suhartono *et al*. (2006) used statistical inference of R^2 incremental contribution on the forward procedure to determine the best architecture of FFNN. This approach involves three basic steps, which can be described in the following theorem.

Theorem 2.4. (Suhartono 2007) *Let the Reduced Model is defined as*

$$Y_t = f(X_t, \hat{\mathbf{w}}_n^{(R)}) + \varepsilon_t^{(R)} \quad \text{……(9)}$$

where l_R *is the number of parameters to be estimated. And, let the Full Model that is more complex than Reduced Model is defined as*

$$Y_t = f(X_t, \hat{\mathbf{w}}_n^{(F)}) + \varepsilon_t^{(F)} \quad \text{.....(10)}$$

where l_F *is the number of parameters in the Full Model,* $l_F > l_R$. *Then, under* $H_0 : \mathbf{w}^{*+} = \mathbf{0}$ *or testing for and additional parameters in the Full Model equal to zero, the* F *statistic can be constructed, i.e.*

$$F = \frac{(SSE_{(R)} - SSE_{(F)})/(df_{(R)} - df_{(F)})}{SSE_{(F)} / df_{(F)}} \quad \text{……..(11)}$$

$$\text{or } F = \frac{R^2_{\text{incremental}} \big/ (df_{(R)} - df_{(F)})}{(1 - R^2_{(F)}) \big/ df_{(F)}} \quad \text{……(12)}$$

where $R^2_{\text{incremental}} = R^2_{(F)} - R^2_{(R)}$, $df_{(R)} = n - l_R$ *is degree of freedom at Reduced Model, and* $df_{(F)} = n - l_F$ *is degree of freedom at Full Model.*

New procedures for ffnn model building.

Based on the Wald test and statistically inference of $R^2_{\text{incremental}}$, we proposed two new procedures for FFNN model building that applied for time series forecasting. In the first step, nonlinearity test is employed to validate whether a nonlinear time series model must be used for analyzing the time series data.

These two algorithms are started with the same approach, i.e. forward scheme by using inference of $R^2_{\text{incremental}}$ for determining the optimal number of hidden nodes. Then, the first procedure continues with the same forward scheme for selecting the optimal input units, and illustrated as Figure 2. Whereas, the second procedure uses backward scheme by implementing Wald test for selecting the optimal input units. This combination between inference of $R^2_{\text{incremental}}$ and Wald test is illustrated in Figure 3.

METHODS

In this paper, these two new procedures are applied on the simulated data. The simulation experiment is carried out to show how the proposed FFNN modeling procedures work. Finally, the result is compared to the procedures proposed by Kaashoek & Van Dijk (2002) and Suhartono *et al*. (2006). Simulated data are generated as ESTAR (Exponential Smoothing Transition Auto-regressive) model, i.e.

$$y_t = 6.5 y_{t-1}.\exp(-0.25 y_{t-1}^2) + u_t \quad \text{…..…..(13)}$$

where $u_t \sim \text{nid}(0, 0.5^2)$.

Time series and the lags plots of this simulated data can be seen in Figure 4. We can observe that data follow nonlinear autoregressive pattern at lag 1.

Empirical results

In this section, the empirical results for the two proposed procedures as illustrated in Figure 2 and 3 are presented and discussed. It contains three sub sections, i.e. the results of the first procedure by using inference of $R^2_{\text{incremental}}$, the results of the second procedure by implementing combination between inference of $R^2_{\text{incremental}}$ and Wald test, and the comparison result of these two new procedures.

The results of the first procedure

In this procedure, firstly we apply the proposed forward procedure starting with a FFNN with six variable inputs $(y_{t-1}, y_{t-2}, \ldots, y_{t-6})$ and one constant input to find the optimal nodes in the hidden layer. It's done by implementing inference of $R^2_{\text{incremental}}$. The result of an optimization steps are reported in Table 1. Based on the results in Table 1, we can see that two hidden nodes are the optimal result and further optimization runs are not needed.

Then, we continue an optimization to find the optimal input units. The results are presented in Table 2. It shows that input unit 1, i.e. y_{t-1}, is the optimal input unit of the network. Hence, the first procedure based on the forward scheme by implementing inference of $R^2_{\text{incremental}}$ yields the optimal network of FFNN with one input unit and two hidden nodes or FFNN(1,2).

COMPARISON RESULTS

There are two main evaluations for the comparison results between these two proposed procedures and other procedures proposed Kaashoek & Van Dijk (2002), i.e. the final result of FFNN architecture and the number of running steps. In general, the results of this simulation study show that the optimal FFNN architecture yielded by these procedures is the same, i.e. FFNN(1,2).

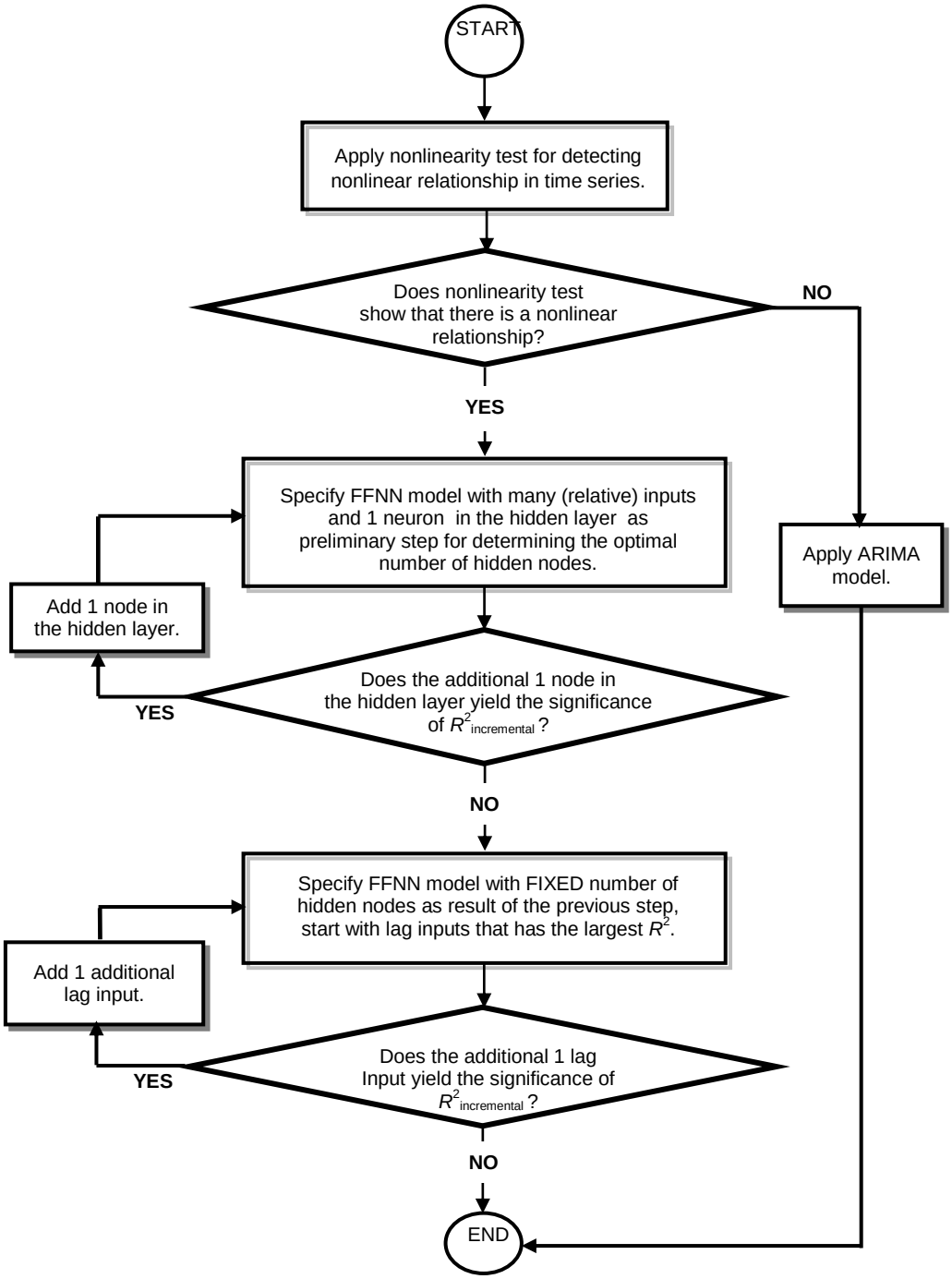

Figure 2. The first proposed procedure of FFNN model building for time series forecasting.

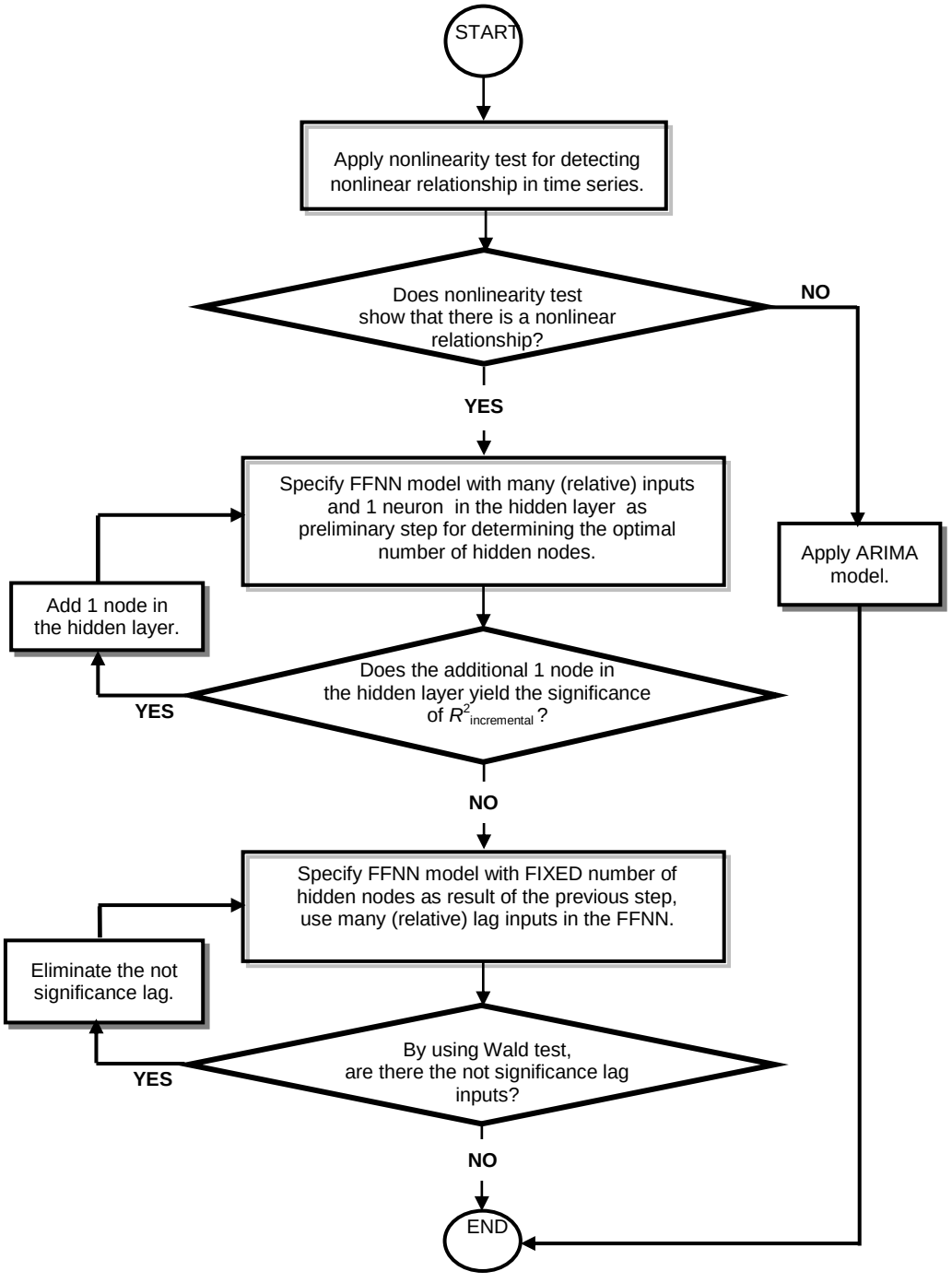

Figure 3. The second proposed procedure of FFNN model building for time series forecasting.

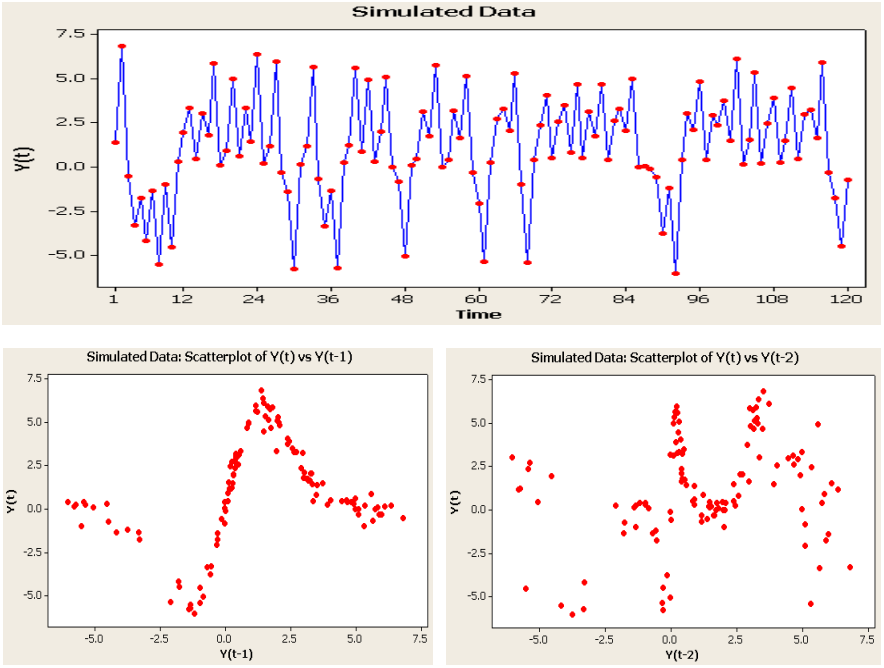

Figure 4. Time series and lags (y_{t-1} and y_{t-2}) plots of simulated data.

Table 1. The results of the optimal hidden nodes determination in the first procedure based on the forward scheme by implementing inference of $R^2_{\text{incremental}}$.

Number of hidden nodes	SBC	R^2	$R^2_{\text{INCREMENTAL}}$	F test	p-value
0	234.4843	0.161569	-	-	-
1	182.0737	0.547258	0.385689	28.5667	0.00000
2	-72.8918	0.975535	0.428277	7.7719	0.00000
3	-61.4821	0.981029	0.005494	0.0518	0.99993
4	-45.5007	0.984601	0.003572	0.0300	0.99999
5	-33.6011	0.987999	0.003398	0.0251	1.00000
6	2.70047	0.988065	0.000066	0.0004	1.00000

Table 2. The results of the optimal input units determination by in the first procedure based on the forward scheme by implementing inference of $R^2_{\text{incremental}}$.

Input lags	SBC	R^2	$R^2_{\text{incremental}}$	F test	p-value
1	-137.764	0.972463	-	-	-
2	235.233	0.383648	-	-	-
3	272.478	0.159330	-	-	-
4	284.580	0.070129	-	-	-
5	285.902	0.059832	-	-	-
6	278.594	0.115375	-	-	-
1 – 2	130.9003	0.973078	0.000615	1.23954	0.29349
1 – 3	129.0907	0.972669	0.000206	0.41539	0.66109
1 – 4	129.1086	0.972673	0.000210	0.42346	0.65583
1 – 5	128.5544	0.972547	0.000083	0.16829	0.84531
1 – 6	130.3262	0.972949	0.000485	0.97934	0.37877

Table 3. The results of the optimal input units determination in the second procedure based on backward scheme by implementing Wald test.

Weights	COEFFICIENT	S.E.	WALD TEST	P-VALUE
b ->h1	-0.0122	0.0352	0.1203	0.728733
1->h1	0.9630	0.0556	300.0898	0.000000
2->h1	-0.0165	0.0108	2.3532	0.125021
3->h1	-0.0016	0.0068	0.0555	0.813763
4->h1	-0.0060	0.0068	0.7712	0.379829
5->h1	-0.0009	0.0071	0.0162	0.898732
6->h1	0.0020	0.0069	0.0846	0.771153
b->h2	-0.0005	0.0369	0.0002	0.989196
1->h2	1.3477	0.0746	326.0336	0.000000
2->h2	-0.0175	0.0116	2.2753	0.131440
3->h2	-0.0038	0.0081	0.2198	0.639206
4->h2	-0.0048	0.0080	0.3584	0.549406
5->h2	-0.0006	0.0080	0.0057	0.939963
6->h2	-0.0008	0.0078	0.0104	0.918691
b->o	0.3878	0.1474	6.9216	0.008515
h1->o	-77.4291	23.8600	10.5307	0.001174
h2->o	76.5030	23.9097	10.2381	0.001376

The comparison result on the number of running steps shows that the second procedure based on the combination between inference of $R^2_{\text{incremental}}$ in forward scheme and Wald test in backward scheme yields the least running steps. The results in Table 1 and 3 show that the second proposed procedure need 4 running steps, i.e. 3 running for determining the optimal hidden nodes and 1 running for input layer cells.

The results of the second procedure

As stated in the previous section, these two new procedure start with the same approach, i.e. forward scheme by implementing inference of R2incremental for determining the optimal number of hidden nodes. Hence, the optimal

number of hidden nodes in this second procedure is exactly the same with the result of the first procedure, i.e. two hidden nodes as presented at Table 1.

Then, an optimization at the second procedure continue to find the optimal input units. It is done in backward scheme by using Wald test. The results of the significance Wald test for FFNN estimator are presented in Table 3. It shows that only input unit 1, i.e. y_{t-1}, is the input cell of the network which has significance estimator, both to hidden node 1 and 2 (h1, h2). Hence, this backward procedure yields the optimal network is FFNN with one input unit (i.e. y_{t-1}) and two nodes in the hidden layer or FFNN(1,2).

CONCLUSION

Based on the results at the previous sections, we can make two main conclusions, i.e.

i. Two new proposed procedures for FFNN model selection based on the inference of R2incremental and Wald test work properly for determining the best FFNN architecture.

ii. The second proposed procedure based on the combination between inference of R2incremental in forward scheme and Wald test in backward scheme yields the least running steps.

In general, the results also show that the proposed procedures give an advantage for FFNN modeling, i.e. the building process of FFNN model is not a black box. Additionally, we can do further research particularly on the application of this proposed procedure in the real time series data.

REFERENCES

Bates DM. & Watts DG. 1988. Nonlinear Regression Analysis and Its Applications. Wiley: New York.

Bishop CM. 1995. Neural Network for Pattern Recognition, Oxford: Clarendon Press.

Fahlman SE. & Lebiere C. 1990. The Cascade-Correlation Learning Architecture. In Touretzky, D. S. (ed.). Advances in Neural Information Processing Systems **2**. Los Altos, CA: Morgan Kaufmann Publishers, pp. 524-532.

Haykin H. 1999. Neural Networks: A Comprehensive Foundation. Second edition, Prentice-Hall: Oxford.

Hornik K., Stinchombe M. & White H. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks*. **2**: 359-366.

Hornik K., Stinchombe M. & White H. 1990. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*. **3**: 551-560.

Kaashoek JF. & Van Dijk HK. 2002. Neural Network Pruning Applied to Real Exchange Rate Analysis. *Journal of Forecasting*. **21**: 559-577.

Phillips PCB. 1989. Partially identified econometric models. *Econometric Theory*. **5**: 181-240.

Prechelt L. 1997. Investigation of the CasCor Family of Learning Algorithms, *Neural Networks*. **10**: 885-896.

Reed R. 1993. Pruning algorithms – A survey. *IEEE Transactions on Neural Networks*. **4**: 740-747.

Ripley BD. 1996. Pattern Recognition and Neural Networks. Cambridge University Press: Cambridge.

Seber GAF. & Wild CJ. 1989. Nonlinear Regression. Wiley: New York.

Suhartono. 2007. *Feedforward Neural Networks for Time Series Forecasting*. [Unpublished Dissertation, Department of Mathematics, Gadjah Mada University, Yogyakarta].

Suhartono, Rezeki S., Subanar & Guritno S. 2005. Optimisation of Backpropagation Algorithm of Feedforward Neural Networks for Regression and Time Series Modeling. *Proceeding International Regional Conference on Mathematics, Statistics and It's Application (IRCMSA)*. Danau Toba, Medan.

Suhartono, Subanar & Guritno S. 2006. Model Selection in Neural Networks by Using Inference of R^2 incremental, PCA, and SIC Criteria for Time Series Forecasting. JOURNAL OF QUANTITATIVE METHODS: *Journal Devoted to The Mathematical and Statistical Application in Various Fields*. **2**(1): 41-57.

Swanson NR. & White H. 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.

Swanson NR. & White H. 1997. A model-selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks. *Review of Economic and Statistics*. **79**: 540-550.

White H. 1989a. Some asymptotic results for learning in single hidden layer feedforward networks. *Journal of the American Statistical Association*. **84**(408): 1003-1013.

White H. 1989b. Learning in neural networks: a statistical perspective. *Neural Computation*. **1**: 425-464.

White H. 1999. Asymptotic Theory for Econometricians. Academic Press Inc.: New York.