Neural Networks for Signal Processing IV, J. Vlontzos, J. Hwang, E. Wilson, eds, pp. 229-237 (IEEE , 1994).

THE SELECTION OF NEURAL MODELS OF NON-LINEAR DYNAMICAL SYSTEMS BY STATISTICAL TESTS

D. URBANI, P. ROUSSEL-RAGOT,
L. PERSONNAZ, G. DREYFUS
Ecole Supérieure de Physique et de Chimie Industrielles de la Ville de Paris
Laboratoire d'Electronique
10, rue Vauquelin
F - 75005 PARIS - FRANCE
Phone: 33 1 40 79 45 41 ; Fax: 33 1 40 79 44 25
e-mail: dreyfus@neurones.espci.fr

Abstract - A procedure for the selection of neural models of dynamical processes is presented. It uses statistical tests at various levels of model reduction, in order to provide optimal tradeoffs between accuracy and parsimony. The efficiency of the method is illustrated by the modeling of a highly non-linear NARX process.

INTRODUCTION

The representation of the behaviour of dynamical processes is a conceptually straightforward application of neural networks, whether feedforward or recurrent, as non-linear regressors. In practice, however, the modeling of a process requires solving several problems:
(i) the choice of the nature of the model (static model *vs* dynamic model, input-output representation *vs* state representation, ...) requires an analysis of the future use of the model (for instance, whether it will be used for predicting the future evolution of the process, or whether it will be used within a control system), and an analysis of the *a priori* knowledge on the phenomena involved in the process;
(ii) the choice of the structure of the model, defined by the number of its inputs, by the number of its outputs, by the type of input-output relationship (linear, polynomial, radial-basis function, multi-layer neural network, etc.), and by its structural parameters (degree of the polynomial approximation, number of radial basis functions, number of neurons, etc.);
(iii) the estimation of the optimal set of adjustable coefficients (synaptic weights in the case of neural net models) of the chosen structure ("identification" in automatic control, "training" in neural network parlance);

The first problem is fully application-dependent: no general statement can be made. The third problem has been investigated in great depth in the case of

linear models [1]; in the case of neural network models, a variety of training algorithms is available [2], and it has been shown that the choice of a training algorithm, in the context of dynamical process modeling, is based on the nature of the noise present in the process to be modeled [3].

In the present paper, we investigate the second problem, namely, that of model selection, which is a key factor for a model to be successful [4]. We suggest a pragmatic model selection procedure for dynamical input-output non-linear models, which features three steps in succession: first, the inputs (external inputs and feedback inputs) of *linear* models of the process around operating points are selected; in a second step, the relevant inputs of the *non-linear* model are selected, thereby determining the order of the model; finally, the structural parameter of the model is determined. An optimized model of a dynamical process is thus derived.

We describe the selection procedure in the case of stable (within the range of operation for which a model is needed), single-input-single-output processes. We assume that the process is NARX:
$y_p(t) = \Phi[y_p(t-1), ..., y_p(t-\nu), u(t-1), ..., u(t-\mu)] + w(t)$
where $\{w(t)\}$ is a gaussian sequence of zero mean independent random variables, ν is the order of the assumed model, and μ is the memory span of the control sequence $\{u(t)\}$.
The following predictor is used:
$y(t) = \Psi[y_p(t-1), ..., y_p(t-n), u(t-1), ..., u(t-m)]$;
We know from [3] that such a predictor (trained with a directed, or teacher-forcing, algorithm) is optimal as a predictor for a NARX process.
If $n = \nu$, if $m = \mu$, and if $\Psi(.)$ is an accurate approximation of $\Phi(.)$, then the predictor is optimal for the process.
In the following, we describe the three steps of the procedure, in the case of a neural network model.

THE PROCEDURE

First step

In the stability domain of the process, operating points (u_i, y_i) are chosen. The process is subjected to time-dependent control sequences of length N in the ranges $[u_i + \Delta u_i, u_i - \Delta u_i]$, such that a *linear* model of the process can be considered valid in each of these ranges. For each operating point, we select, as described below, a linear model which is a satisfactory tradeoff between accuracy and parsimony. At the end of the first step, the set of all inputs which were selected is available for use in the second step of model selection.

For each operating point, we make the assumption that the process can be described as an ARX model :

$$y_p(t) = \sum_{i=1}^{\nu} \alpha_i \, y_p(t-i) + \sum_{i=1}^{\mu} \alpha_{\nu+1} \, u(t-i) + w(t) \ .$$

where ν et μ are unknow parameters.

We consider a training set of size N, and a family of predictors of the form:

$$y(t) = \sum_{i=1}^{n} \theta_i \, y_p(t-i) + \sum_{i=1}^{m} \theta_{n+i} \, u(t-i) \ .$$

The aim of the procedure is to find a predictor such that $n = \nu$, $m = \mu$.

We denote by $\mathbf{y}_p$, $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$, $\mathbf{x}_{n+1}$, …, $\mathbf{x}_{n+m}$, $\mathbf{w}$, $\mathbf{y}$ the N-vectors, corresponding to the values $y_p(t)$, $y_p(t-1)$, ..., $y_p(t-m)$, u(t-1), ..., u(t-n), w(t), y(t), for t=1 to N; thus:

$\mathbf{y} = [\mathbf{x}_1, \ ... \ \mathbf{x}_M] \, \theta$, where M = m + n.

We have to find M regressors, corresponding to M independent vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ such that the subspace spanned by these vectors is the subspace of smallest dimension containing $E[\mathbf{y}_p]$. In order to find this subspace, we start with a complete model, whose parameters n' and m' are chosen to be larger than can be expected from the *a priori* knowledge available on the process. We thus make the assumption that the subspace H spanned by the M'=n'+m' vectors contains $E[\mathbf{y}_p]$, and we expect to extract the satisfactory subset of significant regressors from the initial set. This could be achieved by computing and comparing all possible regressions; however, this method becomes too expensive for large M'.

In order to decrease the amount of computation, we build from the initial set $\{\mathbf{x}_1, \ldots, \mathbf{x}_{M'}\}$ an ordered set of orthonormal vectors $\{\mathbf{p}_1, \ldots, \mathbf{p}_{M'}\}$ such that the model defined by $\{\mathbf{p}_1, \ldots, \mathbf{p}_k\}$, for all $1 \le k \le M'$, gives a sum of squares of errors (SSE) which is smaller than the SSE given by all other models with k regressors [5].

We first choose, among the M' vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_{M'}\}$, the vector $\mathbf{x}_j$ giving the largest square regression $|\mathbf{p}_1^T \mathbf{y}_p|^2$, with $\mathbf{p}_1 = \mathbf{x}_j / \|\mathbf{x}_j\|$. The (M'-1) remaining $\{\mathbf{x}_i\}$ vectors are orthonormalized with respect to $\mathbf{p}_1$.

Consider the k^{th} step of the ordering procedure, where $\mathbf{p}_1, \ldots, \mathbf{p}_{k-1}$ have been selected. We denote by SSE(k) the SSE obtained with the selected model having k regressors, thus :

$SSE(k-1) - SSE(k) = |\mathbf{p}_k^T \mathbf{y}_p|^2$,

with :

$SSE(0) = \|\mathbf{y}_p\|^2$.

This contribution decreases as k increases. This procedure is iterated M'-1 times for $\mathbf{p}_2$, $\mathbf{p}_3$, ... until completion of the list. Thus :

$$\|y_p\|^2 = \sum_{k=1}^{M'} |\mathbf{p}_k^T \mathbf{y}_p|^2 + SSE(M')$$

where SSE(M') is the sum of squares of errors for the complete model.

Subsequently, the above list is scanned in the inverse order of its construction, and each model is compared with the complete model, using the Log Determinant Ratio Test (LDRT). The number of models we have to take into account is at most equal to M'. Note that the comparison between these models by LDRT is easy (see Appendix for further details about this test), since the variable used to compare the k-regressor model and the complete model is :

$$X_{LDRT} = N \frac{\log\left[SSE(k)\right]}{\log\left[SSE(M')\right]}.$$

We select the smallest predictor model accepted by the test.

In order to further decrease the number of tests, we introduce a simple stopping criterion during the formation of the subset $\{\mathbf{p}_1, \ldots, \mathbf{p}_{M'}\}$: at the k^{th} step, the procedure is terminated if $|\mathbf{p}_k^T \mathbf{y}_p|^2 < \rho \|\mathbf{y}_p\|^2$. The choice of ρ is not critical provided it is small (typically $\rho<10^{-8}$).

In the present work, we use LDRT, but Fisher-Snedecor test, Akaike's Information Criterion (AIC) test are also available (for a review see [4]) and lead to similar results.

Thus, for each chosen operating point, a linear model is available, which achieves a satisfactory tradeoff between accuracy and parsimony. Note that the techniques which are used in the linear context of this step are not computationally expensive, so that a large number of external inputs n and feedback inputs m can be used as a starting model for selection.

At the end of the first step, each regressor which was selected for at least one operating point is available for consideration in the second step of model selection.

Second step

In this step, the process is subjected to large-amplitude control signals corresponding to the conditions of operation which the model is expected to account for. A non-linear model is defined (e.g. a neural network), whose inputs are the set of inputs which were determined during the previous step, and whose structural parameters are deemed to be appropriate for the non-linear input-output function to be accurately approximated (e.g. a neural network with an appropriate, possibly too large, number of neurons, trained by an algorithm which allows an efficient minimization of the SSE). Such methods tend to be computationally expensive, so that the chosen number of neurons should not be excessively large. The best subset of inputs is selected by statistical tests (LDRT or AIC criterion (see appendix)) : we compare the complete non-linear model with all these sub-models with one input less. If all the models are rejected, this step of the procedure is terminated. Otherwise, the best submodel is chosen, and compared with all these sub-models having one input less, and so on.

At the end of this step, a non-linear model M_1 is available, whose inputs have been selected.

Third step

The final step aims at determining the structural parameter of the model: in the case of a neural network model, this parameter is the number of hidden neurons. Here, the accuracy/parsimony tradeoff is expressed by the fact that too large a number of hidden neurons leads to overtraining (small SSE on the training set, large SSE on the test set), whereas too small a number of neurons leads to poor approximation (large SSE on the training set itself). The model M_1 resulting from the previous two steps is considered as the complete model, and models with a smaller number of hidden neurons than M_1 are considered for selection. As in the previous steps, statistical tests are used in order to find a satisfactory tradeoff. Note that most model reduction algorithms used for neural networks aim at eliminating connections [6], whereas this final step aims at eliminating neurons.

EXAMPLE

The efficiency of the above procedure is illustrated by the modeling of a second-order, highly non-linear NARX process, which is simulated by the following equation:

$$y_p(t) = 50 \tanh\left(2.\ 10^{-3}\left[\frac{24 + y_p(t-1)}{3}\, y_p(t-1) - 8\,\frac{u(t-1)^2}{1 + u(t-1)^2}\, y_p(t-2)\right]\right) + 0.5\, u(t-1) + w(t)\ ,$$

where w(t) is white noise with variance $(\sigma_w)^2$. The behaviour of this process is essentially that (i) of a linear first-order low-pass filter for amplitudes smaller than or on the order of 0.1, and (ii) of a second-order, oscillatory, linear ($0.1 < |u| < 0.5$), or non-linear ($0.5 < |u| < 5$) system for larger amplitudes; it becomes almost static for positive signals of very large amplitude; in addition, it is not symmetrical with respect to zero. Figure 1 shows the response of the process to steps of random amplitude in the region of interest, with $(\sigma_w)^2 = 10^{-2}$.

First step

The operating points were $u_i = \{-10, -8, -5, -2, -1, -0.5, 0.1, 1, 2, 5, 8, 10\}$. At each of these points, a uniformly distributed random sequence was added to the control input, with maximum amplitude $\Delta u_i = 0.1$ ($\sigma_u^2 = 3.10^{-3}$). The initial model was chosen to have n' = m' = 100. The training sequence was of length N = 1000. The orthonormalization procedure retained 15 inputs, and the subsequent LDRT tests (with 1% risk) led to the selection of n+m = 2 to 5 inputs, depending on the operating points.

Second step

The training set was a sequence of large-amplitude steps, such as shown on Figure 1. M_1 was a fully connected neural network, with the 5 inputs (n = 3, m = 2) selected in the first step, and with 10 hidden neurons. After training, the variance of the prediction error (as estimated by SSE/N) was on the same order of magnitude as σ_W , which shows that the network was sufficiently large, and had been trained efficiently. Subsequently, the networks obtained by suppressing 1 input, then 2 inputs, etc., were trained and submitted to the LDRT procedure, as illustrated on Table 1: the full model M_1 is compared to $M_2, M_3, ..., M_6$. The test selected only M_2 and M_4 (the deletion of one input leads to the deletion of 11 connections; the corresponding value of the χ^2 variable for a 1% risk is 24.7). Since the SSE of M_4 was smallest, it was selected for comparison with all models smaller than M_4[1]; M_7 is the only three-input model which was selected. All models smaller than M_7 were rejected. Therefore, M_7 was finally accepted. The success of the procedure is shown by the fact that M_7 is indeed the only model which has the same inputs as the simulated process. A similar result is obtained if the AIC test is used.

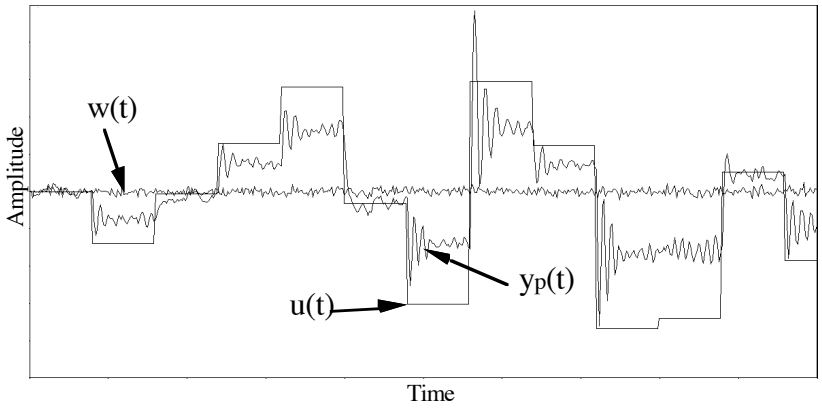

FIGURE 1
Sequence of control input and process output.

Third step

Model selection is performed on neural nets having the inputs of M_7, and 0 to 10 hidden neurons, with the same training set for all nets. The result of the selection depends on σ_W. With $\sigma_W = 10^{-2}$, a model with 9 neurons is selected. With $\sigma_W = 10^{-1}$, the same inputs are selected by the first two steps and the third step leads to a neural network with 4 neurons. As should be

[1] Actually, the SSE's of M_2 and M_4 are very close; if M_2 is selected instead of M_4, the same result is obtained, since M_7 is a sub-model of both M_2 and M_4.

Model	$y_p(t-1)$	$y_p(t-2)$	$y_p(t-3)$	$u(t-1)$	$u(t-2)$	$SSEx10^2$	X_{LDRT}
1	X	X	X	X	X	19.1	
2	X	X	X	X	–	19.6	11
3	X	X	X	–	X	13.0	832
4	X	X	–	X	X	19.5	10
5	X	–	X	X	X	31.6	218
6	–	X	X	X	X	31.8	221
7	X	X		X	–	19.6	1.2
8	X	X		–	X	97.7	697
9	X	–		X	X	11.8	980
10	–	X		X	X	39.4	1303
11	X	X		–		25.4	1114
12	X	–		X		18.7	978
13	–	X		X		18.2	968

TABLE1
Models labelled by boldface figures are those
whose inputs include the inputs of the process.

expected, the procedure selects a smaller number of neurons if the noise level is high than if it is low.

CONCLUSION

A pragmatic three-step procedure for non-linear dynamical model selection has been proposed, which uses statistical tests at various levels of model reduction. It relies on the fact that efficient training procedures are available. It allows the selection of the delayed external inputs, of the feedback inputs (hence the determination of the order of the model) and of the structural parameters such as the number of hidden neurons. Its main shortcoming seems to be the fact that its application is subject to the availability of two types of data from the process, namely, small-signal responses around chosen operating points, and large-signal responses in "normal" operation. Its efficiency is shown on an illustrative example: the neural modeling of a highly non-linear NARX process.

APPENDIX

The Logarithm Determinant Ratio Test (LDRT) [4]

The problem of the selection of one model out of two can be formulated as a statistical testing problem. We suppose that an accurate model M_1, described by the vector of paramters θ, is available to explain a set of N experimental data. The null hypothesis states that a part θ_2 of the vector parameter θ is

equal to zero; if this assumption is true, $\theta=[\theta_1, \theta_2]$ can be reduced to θ_1. If the alternative hypothesis is true, then θ_2 cannot be taken equal to a zero vector. A very efficient test to solve such a problem is the Likelihood Ratio Test (LRT), but this test requires the expression of the likelihood function. In our case, with very large N, it reduces to the Log Determinant Ratio Test (LDRT) : under the null hypothesis $\theta_2=0$, with a scalar output, the distribution of the statistics :

$$X_{LDRT} = N \log \frac{SSE(\theta_1)}{SSE(\theta)}$$

converges to a chi-square distribution with $\dim(\theta_2)$ degrees of freedom.

The Akaike's Information Criterion Tests (AIC)

The AIC is an alternative way of selecting a model from a set of models, using statistical tests. For each model of the set, we compute the AIC value :

$AIC = 2 N \log(SSE/N) + 2M$

where N is the number of data and M is the number of parameters of the model.

The model corresponding to the smallest AIC value is thus selected as the best model of the set, with respect to this criterion. This procedure requires no assumptions on the models. There exist more efficient variants of the classical AIC [4], such as the AIC*, used in this work :

$AIC^*= 2 N \log(SSE/N) + 4 M$

.

REFERENCES

[1] See for instance:
L. Ljung, System Identification: Theory for the User: Prentice Hall, 1987.
G.C. Goodwin, R.L. Payne, Dynamic System Identification: Experiment Design and Data Analysis: Academic Press,1977.

[2] O. Nerrand, P. Roussel-Ragot, L. Personnaz, G. Dreyfus, "Neural Networks and Non-linear Adaptive Filtering: Unifying Concepts and New Algorithms", Neural Computation, vol. 5, pp.165-197, 1993..

[3] O. Nerrand, P. Roussel-Ragot, D. Urbani, L. Personnaz, G. Dreyfus, "Training Recurrent Neural Networks: Why and How ? An Illustration in Dynamical Process Modeling", IEEE Transactions on Neural Networks, vol. 5, pp. 178-184, 1994.

[4] I.J. Leontaritis, S.A. Billings, "Model Selection and Validation for Non-Linear Systems", International Journal of Control, vol. 1, pp. 311-341, 1987.

[5] S. Chen, S.A. Billings, W. Luo, "Orthogonal Least Squares Methods and their Application to Non-Linear System Identification" International Journal of Control, vol. 50, pp. 1873-1896, 1989.

[6] R. Reed, "Pruning Algorithms - A Survey", IEEE Transactions on Neural Networks, vol. 4, pp. 740-747, 1993.