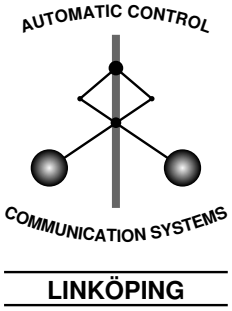
Regressor Selection with the Analysis of Variance Method

Ingela Lind, Lennart Ljung

Division of Automatic Control
Department of Electrical Engineering
Linköpings universitet, SE-581 83 Linköping, Sweden
WWW: http://www.control.isy.liu.se
E-mail: ingela@isy.liu.se, ljung@isy.liu.se

14th September 2004



Report no.: LiTH-ISY-R-2626
Submitted to Automatica

Abstract

Identification of nonlinear dynamical models of a black box nature involves both structure decisions, i.e., which regressors to use, the selection of a regressor function, and the estimation of the parameters involved. The typical approach in system identification seems to be to mix all these steps, which for example means that the selection of regressors is based on the fits that is achieved for different choices. Alternatively one could then interpret the regressor selection as based on hypothesis tests (F-tests) at a certain confidence level that depends on the data. It would in many cases be desirable to decide which regressors to use independently of the other steps.

In this paper we investigate what the well known method of analysis of variance (ANOVA) can offer for this problem. System identification applications violate many of the ideal conditions for which ANOVA was designed and we study how the method performs under such non-ideal conditions.

ANOVA is much faster than a typical parametric estimation method, using e.g. neural networks. It is actually also more reliable, in our tests, in picking the correct structure even under non-ideal conditions. One reason for this may be that ANOVA requires the data set to be balanced, that is regressor value combinations in test/validation data that are very common or unusual are adjusted in order to play more equal roles when deciding their influence on the fit. Just applying tests of fit for the recorded data may give improper weight to very common regressor value combinations.

Keywords: Time delay estimation, Time lag, Structure identification, Non-linear models, Analysis of variance

Regressor Selection with the Analysis of Variance Method

Ingela Lind and Lennart Ljung

Division of Automatic Control, Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden.

Abstract

Identification of nonlinear dynamical models of a black box nature involves both structure decisions, i.e., which regressors to use, the selection of a regressor function, and the estimation of the parameters involved. The typical approach in system identification seems to be to mix all these steps, which for example means that the selection of regressors is based on the fits that is achieved for different choices. Alternatively one could then interpret the regressor selection as based on hypothesis tests (F-tests) at a certain confidence level that depends on the data. It would in many cases be desirable to decide which regressors to use independently of the other steps.

In this paper we investigate what the well known method of analysis of variance (ANOVA) can offer for this problem. System identification applications violate many of the ideal conditions for which ANOVA was designed and we study how the method performs under such non-ideal conditions.

ANOVA is much faster than a typical parametric estimation method, using e.g. neural networks. It is actually also more reliable, in our tests, in picking the correct structure even under non-ideal conditions. One reason for this may be that ANOVA requires the data set to be balanced, that is regressor value combinations in test/validation data that are very common or unusual are adjusted in order to play more equal roles when deciding their influence on the fit. Just applying tests of fit for the recorded data may give improper weight to very common regressor value combinations.

Key words: Time delay estimation, Time lag, Structure identification, Non-linear models, Analysis of variance

1 Introduction

System Identification is the problem of building mathematical models of dynamical systems, that is, systems whose outputs depend not only on the current input, but also past inputs.

The problem. To be concrete, assume that a non-linear model describes the relationship between the measured output y_t and the input u_t from a system with measurement noise e_t, that is,

$$y_t = g(u_t, u_{t-T}, u_{t-2T}, ..., u_{t-nT}) + e_t. \quad (1)$$

Here T is the sampling interval and g is an unknown function of up to $n+1$ variables. Which time lags u_{t-kT} that affect y as well as the order n are unknown. If only past inputs occur in the regression function, as in the case above, the model is known as a Nonlinear Finite Impulse Response (NFIR) model. If also past outputs, y_{t-kT} enter g, the model is known as a Nonlinear ARX (NARX) model, e.g. Sjöberg et al. (1995).

Email addresses: ingela@isy.liu.se (Ingela Lind), ljung@isy.liu.se (Lennart Ljung).

The problem we are faced with in system identification is to find a good model for the system from input/output data, $y_t, u_t, t = rT, r = 1, \ldots, N$. This process contains three steps (e.g., Ljung (1999)):

- First, proper regressors have to be found. In the example above it means to determine the lags, or regressors, $u_{t-kT}, k = k_1, k_2, \ldots, k_d, \quad k_i \in [0, n]$ that should be included in the function arguments.
- Second, the type of function g needs to be chosen. This could for example be a general polynomial of degree r in d variables, or a neural network, or defined by a binary tree, etc. It could also be a function which is locally fit to data, e.g. as local polynomials or piecewise constant approximations. There is obviously an infinite number of choices, and quite a few have been suggested and tested in the literature (e.g., Harris et al. (2002); Sjöberg et al. (1995)) .
- The function selected will contain a number of parameters whose numerical values need to be determined by the data in the third step. For global parameterisations, this could be quite time consuming, since many parameters might be involved ("the curse of dimensionality") and the objective function that is optimised may have several local optima.

In practical use, the three steps are often entangled in each other, in the sense that the search for regressors, type of function and parameter values is done by trial and error. The resulting model's ability to reproduce (validation) data is judged for each choice of regressors, model structure and estimated parameters, until a satisfactory result is reached. We will call this common method *validation based regressor selection*, or *exhaustive search for best regressors*. Obviously, this could be very time-consuming. Comparing the fit for two choices (hypotheses) of regressor selection on validation data can be seen as a hypothesis test at a certain confidence level (Söderström, 1977). It has been noted, e.g., in Piroddi and Spinelli (2003), that this validation based regressor selection could lead to bad results especially for poorly excited systems.

Approaches. It would be desirable to have a more structured route to the model, and we shall in this contribution discuss the first step, disconnected from the other ones. Actually this step has not been that much studied in the system identification literature for nonlinear models, in contrast to the case of linear systems. Haber and Unbehauen (1990) give an early survey. A local method, called the "statistical approach"is described in Poncet and Moschytz (1994). Somewhat related methods are the "false nearest neighbour" method in Kennel et al. (1992); Rhodes and Morari (1998) and the "Lipschitz numbers" method (He and Asada, 1993). Bomberger (1997) compares these two methods. Other methods are the δ-test (Pi and Peterson, 1994), the use of the rank of linearised systems (Autin et al., 1992), the mutual information criterion (Zheng and Billings, 1996), the coherence function of input-output data (Krishnaswami et al., 1995), the orthogonal structure detection routine (Billings et al., 1988; Korenberg et al., 1988), and stepwise regression (Billings and Voon, 1986). Kukreja et al. (1999) have developed a bootstrap-based method.

Not surprisingly, there are many more contributions in the statistical literature. We could point to the nonparametric final prediction error criterion (Auestad and Tjøstheim, 1990; Cheng and Tong, 1992; Tjøstheim and Auestad, 1994a,b; Tschernig and Yang, 2000; Vieu, 1995; Yao and Tong, 1994), the local conditional mean and ANOVA (Chen et al., 1995; Truong, 1993), and the lag dependence function by Nielsen and Madsen (2001).

A simple idea. A simple, intuitive idea has resurfaced a couple of times in the literature: Suppose u is periodic and y depends on u_{t-T} only. Then each time t that u_{t-T} has the same value, y_t should also be the same value, apart from the noise e_t. That is to say, that the variance of y_t taken for these values of t (call it V_1) should be the variance of e_t. The variance of e_t is typically unknown. However, if we check the times t when the pair $[u_{t-T}, u_{t-2T}]$ has the same values, the variance of y_t for these t should also be around V_1 if y_t does not depend on u_{t-2T}. By comparing the variances for different constellations of candidate regressors we could thus draw conclusions about which ones y_t actually depend on.

ANOVA and the current paper. The idea, as we said, is very intuitive and natural. It has been formalised in the statistical literature under the name of ANOVA (ANalysis Of VAriance). In the ideal situation, the regressors are independent and assume only a finite number of values in a well distributed way. This is typically not the situation in applications to dynamical systems, where the regressors are time-lagged inputs and possibly also time-lagged outputs (NARX-models). It is the purpose of this paper to review the basic ideas of the ANOVA method and relate it to structure selection in non-linear system identification. In particular, we shall investigate by simulation how serious the typical violations from the ideal ANOVA situation are for the system identification application. We shall compare the results of ANOVA with the common approach of picking those regressor combinations that give best fit for validation data.

2 The ANOVA idea

The statistical analysis method ANOVA (Miller (1997); Montgomery (1991), among many others) is a widely spread tool for finding out which factors contribute to given measurements. It has been used and discussed since the 1930's and is a common tool in, e.g., medicine and quality control applications.

The method is based on hypothesis tests with F-distributed test variables computed from the residual quadratic sum. There are several slightly different variants (Miller, 1997). Here the fixed effects model with two factors will be described.

Assume that the collected measurement data can be described by a linear statistical model,

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk},$$

where the ϵ_{ijk} are independent Gaussian distributed variables with zero mean and constant variance σ^2. The parameter μ is the overall mean. For each (quantised) level $i = 1, \ldots, a$ of the first regressor $(\varphi_1(t))$ there is a corresponding effect τ_i, and for each level $j = 1, \ldots, b$ of the second regressor$(\varphi_2(t))$ the corresponding effect is β_j. The interaction between the regressors is described by the parameters $(\tau\beta)_{ij}$. The sum of a batch of indexed parameters over any of its index is zero.

For a linear $(y(t) = \theta_1\varphi_1(t) + \theta_2\varphi_2(t) + e(t))$ or a non-linear additive system $(y(t) = g_1(\varphi_1(t)) + g_2(\varphi_2(t)) + e(t))$, the interaction parameters $(\tau\beta)_{ij}$ are zero. These are needed when the non-linearities have a non-additive nature, i.e., $y(t) = g(\varphi_1(t), \varphi_2(t)) + e(t)$.

Since the regressors are quantised, it is a very simple procedure to estimate the model parameters by computing means;

$$\overline{y}_{...} = \frac{1}{abn}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}, \quad \overline{y}_{i..} = \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk},$$
$$\overline{y}_{.j.} = \frac{1}{an}\sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk}, \quad \overline{y}_{ij.} = \frac{1}{n}\sum_{k=1}^{n} y_{ijk}, \tag{2}$$

which are the overall mean, the means over the regressor levels and the cell means. For example, the constant μ would correspond to $\overline{y}_{...}$, while the effects from the first regressor are computed as $\tau_i = \overline{y}_{i..} - \overline{y}_{...}$. The number of free parameters in the model is equal to the number of cells.

ANOVA is used for testing which of the parameters that significantly differ from zero and for estimating the values of the parameters with standard errors, which makes it a tool for exploratory data analysis. The residual quadratic sum, SS_T, is used to design test variables for the different batches (e.g., the τ_i:s) of parameters. Under the assumptions on ϵ_{ijk} stated above and the case when all regressor level combinations are sampled equally, the residual quadratic sum can be divided into four independent parts;

$$\begin{aligned} SS_T &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{...})^2 \\ &= \sum_{i=1}^{a} bn(\overline{y}_{i..} - \overline{y}_{...})^2 + \sum_{j=1}^{b} an(\overline{y}_{.j.} - \overline{y}_{...})^2 \\ &+ \sum_{i=1}^{a}\sum_{j=1}^{b} n(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2 \\ &+ \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{ij.})^2 \\ &= SS_A + SS_B + SS_{AB} + SS_E. \end{aligned} \tag{3}$$

Each part is related to one batch of parameters. If all the parameters in the batch are zero, the corresponding quadratic sum is χ^2-distributed if divided by the true variance σ^2 (see, e.g., Montgomery (1991, page 59)). Since the true variance is not available, the estimate $\hat{\sigma}^2 = \frac{SS_E}{ab(n-1)}$ is used to form F-distributed test variables, e.g., for τ_i;

$$v_A = \frac{SS_A/(a-1)}{SS_E/(ab(n-1))}.$$

If all the τ_i:s are zero, v_A belongs to an F-distribution with $a-1$ and $ab(n-1)$ degrees of freedom. If any τ_i is nonzero it will give a large value of v_A, compared to an F-table. This is, of course, a test of the null hypothesis that all the τ_i:s are zero, which correspond to the case that the regressor φ_1 does not have any main effect on the measurements y.

The most important modeling simplifications made are the assumptions that the variance is equal for all ϵ_{ijk} and that the random error component is Gaussian distributed. The F-tests are quite robust against violations against both assumptions (Krishnaiah, 1980, Chapter 7). When a regressor is quantised to a discrete set of levels both these assumptions are violated. Another complication with using ANOVA in the system identification context is that the regressors often are correlated, e.g., in NARX-models (see Lind (2001)). The correlated regressors violate the assumption that the ϵ_{ijk} are independent and make it hard to obtain a balanced design, even if unequal quantisation intervals are used.

When not all factor level combinations are represented by an equal amount of measurement data, the design is called *unbalanced*. In an unbalanced design the independence between the different sums of squares, Equation (3), is lost. The F-tests should then be done in a different manner (Miller, 1997). An exact treatment of a badly unbalanced design (when the number of measurements in each cell differ a factor 10 or more) is especially hard to interpret. A computationally efficient solution with easy interpretation (suggested in Montgomery (1991)) is to randomly discard excess data in the superfluous cells until a balanced design is obtained. At least two measurements in each cell should be kept. Drawbacks of this solution are that it does not make use of all data and that it might be sensitive to outliers.

Connection to model estimation It is illuminating to see the relationships between ANOVA and direct model estimation. Suppose q regressors are chosen in the predictor model (1). Then g is a function from R^q to R (if y and u are scalars). If u_t only assumes a different distinct values, the domain of the function g is visited only in a^q points in R^q. Lacking structural knowledge of g it is then only meaningful to estimate g's values at these points. The function estimate will then be a table with a^q values. The sample means listed in (2) are the simple and natural estimates of these function values, so in that sense also ANOVA works with estimates of g. Now, not knowing which regressors to include leaves 2^n different combinations, so this function estimation should be performed over all these combinations. On the other hand, it is clear that the function estimates for different regressor combinations are closely related. The ANOVA tests can be seen as an efficient way to evaluate the function fits for all the possible regressor combinations. The fits are constructed from the sums in (3).

3 Simulation results

3.1 The problem considered

We shall consider the problem of identifying models with the NFIR structure

$$y_t = g(u_t, u_{t-T}, u_{t-2T}) + e_t \tag{4}$$

The data will be generated from this model for different functions g, see Table 1, using some or all of the regressors u_t, u_{t-T} and u_{t-2T}. From the generated input-output data, the problem is to decide which regressors to include in the model. This will be done by using ANOVA and comparing with the results when candidate functions g of all 7 combinations of regressors are fitted to the data. In this study we have chosen to work with three candidate regressors. In, e.g., MATLAB's Statistics Toolbox (the Mathworks), there is no software limitation on the number of factors possible to test with ANOVA, but hardware constraints (such as memory limitations) do apply.

3.2 Experiment setup

The simulated measurement error signal e_t is zero mean Gaussian noise with standard deviation 1 in the first two data columns of Table 1 and standard deviation 0.0001 in the other columns.

Fixed-level input. A carefully selected input signal for the identification experiment, with the purpose of finding good regressors with ANOVA, would be a pseudo-random multi-level signal with three or more levels. The range and the particular levels of the signal should correspond to the operating range for the desired model. Here a signal assuming the values -2, 1, 3 and 5 is used. All value combinations among three adjacent time lags can be obtained with the help of a multi-level shift register (Godfrey, 1993), giving a frequency content similar to white noise. Compare with the pseudo-random binary signal often used for linear system identification (Ljung, 1999). There are often reasons to use a symmetric distribution of the values, but it is not necessary for analysis with ANOVA. The signal is repeated four times to give a reasonable amount of data for the analysis, that is, 256 input/output data with 4 in each cell.

Continuous-level input. If the input cannot be chosen by the user, a fixed-level input signal is not realistic. For that reason ANOVA is also evaluated for an input signal, u_t, as an independent, identically distributed random signal from the uniform distribution. It can assume values between -2.5 and 5.5, that is, close to the range used in the earlier experiments. Even though a uniform distribution of the input is not necessarily the most common one in practice, we use it as the most "unfavourable" case for ANOVA.

In the ANOVA treatment, the input range is quantised to four levels, corresponding to intervals of equal length. Three regressors with four levels each gives $4^3 = 64$ cells in the experiment design. Each cell corresponds to a unique combination of regressor levels or a non-overlapping box in the regressor space. The level assignment introduces quantisation error in the ANOVA. The output y_t can now be seen as

$$y_t = E[y_t|(u_t, u_{t-T}, u_{t-2T}) \in C] + e_t + n_t,$$

where $E[y_t|(u_t, u_{t-T}, u_{t-2T}) \in C]$ is the expected value of y_t given that the input is assigned to cell C, and

$$n_t = g(u_t, u_{t-T}, u_{t-2T}) - E[y_t|(u_t, u_{t-T}, u_{t-2T}) \in C]. \tag{5}$$

The distribution of the bias term, n_t, depends on the function g, the distribution of the input u_t and the number of levels used to categorise u_t. This distribution is not necessarily equal in all cells, which violates the ANOVA assumption on equal variances in all cells. It is instructive to think of this construction as a piecewise constant surface (i.e. constant over C) fit to the data. Choosing the size of C is then a classical bias-variance trade-off, and making C small will also result in a "curse-of-dimensionality" problem.

It is possible to elaborate both with the size and placement of C, which gives many possibilities to incorporate process knowledge. The constraint is that the cells should be placed in an axis-orthogonal grid, such that for each level of each regressor, there are cells representing all combinations of levels of the other regressors. One way to deal with the intervals in practice is to start with few, e.g. 3, intervals per regressor, check the ANOVA assumptions with a normal probability plot, and add intervals if needed (Lind, 2001, Chapter 7).

Correlated input.

Whenever the input signal u_t is not a series of independent variables, the factors (regressors) in the ANOVA become correlated and the assumptions on ϵ_{ijk} are violated. To evaluate ANOVA for correlated regressors,

$$u_t = x_t - x_{t-T} + x_{t-2T},$$

where x_t is white noise uniformly distributed between -2.75 and 5.75 is used. This signal is chosen to make a comparison with previous simulations possible. The same quantisation levels as for the continuous-level input signal is used, giving 64 predetermined cells. Due to the rather strong correlation between the adjacent regressors, the number of data in each cell differ considerably. Data series of length 5000 are generated. For each data series, the cell with the lowest number of data is detected. The purpose is to keep 10 data series with at least 2 observations in each cell, 10 data series with at least 3

observations per cell, and so on, up to 6 observations per cell. Data series with less observations per cell are not suitable for evaluation of the ANOVA method. It is also interesting to see if there is any difference in test results for different minimum number of observations per cell.

3.3 The methods to determine the regressors

Validation based regressor selection by exhaustive search with neural networks as analysis method. Within the structure (4) there are 7 possible choices of regressor combinations. The exhaustive search method is to try out how good fit is obtained (on validation data, being the second half of the data record) for each possible combination of regressors. A complication is that the fit is also affected by the choice of function $g(\cdot)$. Another complication is that the best model in the tested class may not be found, due to numerical problems with local minima of the criterion function. We deal with those complications by using several $g(\cdot)$ of varying flexibility as well as several random restarts in the search for the best model. In practice, one of the most used model types for nonlinear black box identification is the one-hidden-layer sigmoid neural network (Sjöberg et al., 1995). We choose to use that model type here, but any model structure with good approximation ability would serve the same purpose.

The analysis is conducted as follows: Divide the data set into estimation data and validation data. Construct neural networks for all $2^3 - 1$ possible combinations of the time lags u_t, u_{t-T} and u_{t-2T}. For each such combination, construct networks with different numbers of parameters. Start with random network parameters and estimate the parameters on the estimation data. Start over 4 or 5 times with new random network parameters to try to avoid getting stuck in a local minimum. Of all these estimated networks, choose the one with the smallest root mean square error(RMSE) on the validation data. This network gives the proposed regressors. The networks used here have $s = 5$, 10 or 15 sigmoids in the hidden layer, giving $(r + 2)s + 1$ parameters to estimate, where r is the number of included regressors. The minimisation algorithm used is Levenberg-Marquardt (MATLAB's Neural Network Toolbox the Mathworks is used).

Note that, in the table, the choice of regressors is considered to be successful if the right set of regressors are selected. One could also seek to determine interaction patterns, i.e. decide if the structure is like $g_1(u_t) + g_2(u_{t-T})$ or $g_3(u_t, u_{t-T})$ etc. If the interaction should be considered in exhaustive search, 18 different model structures have to be tested instead of the seven model structures needed to determine the regressors.

ANOVA as analysis method. A three-factor fixed effects model ANOVA is used to find good regressors, as described above. Model assumptions are checked with a normal probability plot of the residuals. If the plot shows non-Gaussian behaviour, the cells with largest variance are omitted from the analysis (Lind, 2001). This handles, e.g., discontinuities (which give unequal variance in the cells), and leads to much better performance. Notice that the ANOVA decision is regarded as correct in the table only if both the choice of regressors and the interaction pattern is correct. This is a more stringent requirement than for the validation based regressor selection.

3.4 Results from Monte-Carlo simulations

A number of different simulations were performed. For the case of fixed-level inputs, different realizations of the noise were tested using the same input sequence, while for continuous-level inputs, different realizations of both the input and the additive noise were used.

Results from several studies are given in Table 1. The first study, columns 1–4, compares the validation based technique with ANOVA in ideal conditions. The second study investigates the effects of a continuous-level input signal (column 5) and in columns 6 and 7 a correlated input signal is used.

By inspections of columns 1–4 in Table 1 it can be concluded that the ANOVA method is much better at spotting what regressors contribute to the output than the validation based method. The difference in performance between the two methods becomes more pronounced when the functions have a more nonlinear behaviour, e.g., exponential functions. This indicates that the used networks do not handle this kind of functions very well. This can be confirmed by the RMSE values on validation data. The better performance for ANOVA in column 4 compared to column 2 is mostly due to the increased significance level, except for the first function, where the decrease in noise variance is important to explain the better performance (Lind, 2001). It can also be seen that the decrease in noise variance does not have a big impact on the performance for the validation based method either, except for function 1 (compare columns 1 and 3).

For the continuous-level input signal, the quantisation used in the test is very rough. One might expect much worse performance than the results obtained in column 5 (compare with column 2 at the same significance level) which are fairly good and uniform. For function 1 the problem is still low signal to noise ratio for the time lag u_{t-2T}. The results for function 14 are bad. In most of the decisions, the regressor u_t is not included. The reason seems to be that the relatively small contribution from this regressor drowns in the bias term n_t (see (5)), whose size and distribution varies considerably in the different cells.

Analysis method		VB	A	VB	A	A	VB	A
No. of MC runs		100	100	100	100	50	50	50
Significance level		-	0.01	-	0.0001	0.01	-	0.01
Input signal		fix	fix	fix	fix	cont	corr	corr
No. of input/output data		256	256	256	256	800	5000	128-384
Noise standard deviation		1	1	0.0001	0.0001	0.0001	0.0001	0.0001
No.	Function	col. 1	col. 2	col. 3	col. 4	col. 5	col. 6	col. 7
1	$u_t - 0.03u_{t-2T}$	10	5	94	100	52	74	14
2	$\ln \lvert u_t \rvert + u_{t-T} + e^{u_{t-2T}}$	77	98	78	98	92	2	98
3	$u_{t-T} \cdot [u_t + \frac{1}{u_{t-2T}}]$	100	98	100	100	100	60	100
4	$\mathrm{sgn}(u_{t-T})$	84	94	80	100	74	56	94
5	$\mathrm{sgn}(u_{t-T}) \cdot u_{t-2T}$	93	96	92	100	90	84	100
6	$\mathrm{sgn}(u_{t-T}) \cdot u_t \cdot u_{t-2T}$	100	100	100	100	100	100	100
7	$\ln \lvert u_{t-T} + u_{t-2T} \rvert$	95	96	94	100	92	76	86
8	$\ln \lvert u_{t-T} \cdot u_{t-2T} \rvert$	94	90	82	100	90	68	84
9	$u_{t-2T} \cdot \ln \lvert u_{t-T} \rvert$	97	97	95	100	90	58	82
10	$u_{t-2T}^3 \cdot \ln \lvert u_{t-T} \rvert$	50	95	56	100	90	74	90
11	$u_{t-2T} \cdot (\ln \lvert u_{t-T} \rvert)^3$	93	95	91	99	90	54	98
12	$\lvert u_{t-2T} \rvert \cdot e^{u_{t-T}}$	54	96	54	100	90	74	84
13	$u_{t-2T} \cdot e^{u_{t-T}}$	49	94	49	100	88	88	94
14	$u_{t-2T} \cdot e^{u_{t-T} - 0.03u_t}$	58	100	58	100	6	22	6
15	$\lvert u_t \rvert$	83	96	73	100	96	74	98
16	$g(u_{t-T}, u_{t-2T})$	73	88	94	100	-	-	-

Table 1
Results from Monte-Carlo simulations. The first rows state the experiment setup and the last rows give the percentage of tests where the regressors are chosen correctly. Abbreviations used are VB for validation based exhaustive search, A for ANOVA, fix for fixed-level, cont for continuous-level and corr for correlated. Function 16 is a network of the same type as the ones used in the models used for the validation based search.

For the correlated data in column 7, the 50 Monte-Carlo simulations are performed on balanced data, i.e., an equal number of data in all cells. Recall that the chosen inputs guaranteed that all quantised cells contained at least 2 (3, 4, 5, 6, resp.) data points. To obtain balanced data, the minimum available data in each cell (i.e. 2, 3, 4, 5, or 6) were drawn randomly, thus discarding many measurements. Only 128 (192, 256, 320, 384, resp.) data points were actually used. There were no obvious differences in the success rate between the analysis of the data series with at least 2 observations per cell and those with at least 6 observations per cell (Lind, 2001). The results in columns 5 and 7 are comparable, so ANOVA is not seriously corrupted by the correlated regressors in the balanced tests. It shows that throwing away data did not have such a bad effect after all. Validation based search for best regressors for correlated data is not very successful (column 6). 3000 data are used for estimation and 2000 data for validation. The errors are equally distributed between missing contributing regressors and adding spurious regressors. It takes about 1 hour to compute column 7 (including balancing data), while it takes over 320 hours to compute column 6.

4 Conclusions

It can be concluded from experiments that the analysis of variance method is a good alternative to validation based exhaustive search using neural networks for identifying which inputs that contribute to the output in nonlinear finite impulse response (NFIR) models. The ANOVA method manages much better to identify good regressors and reduces the number of erroneous models from 19.4-24.4% to 0.1-9.7% when compared to the exhaustive search method, using a multi-level pseudo-random input signal. The main source of the difficulties in exhaustive search is that the minimisation algorithm gets stuck in local minima, due to the non-convexity of the identification problems. ANOVA is also computationally much more efficient. The ANOVA method (including estimation of a model) is at least 14 times (and for some tested cases 800 times) faster in finding appropriate regressors and a corresponding model. When the complexity increases, due to more possible regressors, the speed differences become more pronounced.

It is also possible to get good results from input/output

data with a continuous-level input signal and a low to intermediate level measurement noise. The extra bias term introduced by the quantisation, can sometimes lead to a more complicated analysis. The two main problems are the reduction of the signal to noise ratio and unequal variances in the cells. Both these problems can be counteracted by changing the quantisation intervals. The simulation results give no reason to be extra cautious when a correlated input signal is used. In our experience (Lind, 2001), also NARX models can be tested with success if a balanced design is attained. The technique was also applied in Lind and Ljung (2003) to find underlying structures for dynamical systems in terms of so called *local linear models*.

A conceptual, but illuminating way to interpret ANOVA from a system identification perspective is to see it as building piece-wise constant models of the response-surface g and comparing the fits for various regressor combinations. This brings out the following features:

- Parameterisations in terms of piece-wise constant predictor functions g can be done as simple linear regressions, which reflects the speed and lack of problems with global optimisations in ANOVA. The same comment would apply for several other types of parametric or non-parametric estimation of g, e.g. spline-based techniques.
- ANOVA provides a systematic way of testing what would be regarded as significant improvement of fit using the F-tests. The validation based approach typically employs best fit to validation data as the criterion (like in Table 1). This could be interpreted as an F-test at a certain level (e.g., Söderström (1977)). Clearly other levels could be applied also in the validation based search case, with a possibly different outcome in Table 1.
- ANOVA offers no "free lunch" regarding typical problems in black-box identification (like the curse of dimensionality) but it brings out the essential statistical features associated with the regressor selection problem. In our tests it showed good performance also when the ideal conditions were violated.
- The search over the fit to validation data gave rather bad results for the correlated data. This is in line with the observations of Piroddi and Spinelli (2003). One explanation is that the data balancing that takes place in ANOVA is healthy for the decision which regressors actually affect the output. A combination of regressor values that is very common in the data set may conceal other effects when just checking the data fit. Differently put: Validation based techniques look for the regressors that may give the best (MSE) fit to data with the same character as the validation set. ANOVA tries to give a direct answer to the question if a certain regressor has a clear influence on the output.

Acknowledgements

This work has been supported by the Swedish Research Council (VR) which is gratefully acknowledged.

References

B.H. Auestad and A. Tjøstheim. Identification of nonlinear time-series - 1st order characterization and order determination. *Biometrika*, 77:669–687, 1990.

M. Autin, M. Biey, and M. Hasler. Order of discrete time nonlinear systems determined from input-output signals. In *IEEE International Symposium on Circuits and Systems, ISCAS '92.*, volume 1, pages 296–299, 1992.

S.A. Billings, M.J. Korenberg, and S. Chen. Identification of non-linear output-affine systems using an orthogonal least squares algorithm. *International Journal of Systems Science*, 19:1559–1568, 1988.

S.A. Billings and W.S.F. Voon. A prediction error and stepwise-regression estimation algorithm for nonlinear systems. *International Journal of Control*, 44: 803–822, 1986.

J.D. Bomberger. *Radial Basis Function Networks for Process Identification.* PhD thesis, University of California, Santa Barbara, Aug 1997.

R. Chen, J.S. Liu, and R.S. Tsay. Additivity tests for nonlinear autoregression. *Biometrika*, 82:369–383, 1995.

B. Cheng and H. Tong. On consistent non-parametric order determination and chaos (with discussion). *Journal of the Royal Statistical Society, Series B*, 54:427–474, 1992.

K. Godfrey. *Perturbation Signals for System Identification.* Prentice Hall, New York, 1993.

R. Haber and H. Unbehauen. Structure identification of nonlinear dynamic systems - a survey on input/output approaches. *Automatica*, 26:651–677, 1990.

C. Harris, X. Hong, and Q. Gan. *Adaptive Modelling, Estimation and Fusion from Data: a neurofuzzy approach.* Springer-Verlag, Berlin Heidelberg, 2002.

X. He and H. Asada. A new method for identifying orders of input-output models for nonlinear dynamic systems. In *Proceedings of the American Control Conference*, pages 2520–2523, 1993.

M.B. Kennel, R. Brown, and H.D.I. Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review A*, 45:3403–3411, 1992.

M. Korenberg, S.A. Billings, Y.P. Liu, and P.J. McIlroy. Orthogonal parameter estimation algorithm for nonlinear stochastic systems. *International Journal of Control*, 48(1):193–210, 1988.

P.R. Krishnaiah, editor. *Handbook of Statistics*, volume 1. North-Holland, Amsterdam, 1980.

V. Krishnaswami, Y.W. Kim, and G. Rizzoni. A new model order identification algorithm with application to automobile oxygen sensor modeling. In *Proceedings*

of the American Control Conference, pages 2113–2117, 1995.

S.L. Kukreja, H.L. Galiana, and R.E. Kearney. Structure detection of NARMAX models using bootstrap methods. In *Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, Arizona USA*, volume 1, pages 1071–1076, 1999.

I. Lind. Regressor selection in system identification using ANOVA, Nov 2001. Licentiate Thesis no. 921, Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden.

Ingela Lind and Lennart Ljung. Structure selection with ANOVA: Local linear models. In P. van der Hof, B. Wahlberg, and S. Weiland, editors, *Proc. 13th IFAC Symposium on System Identification*, pages 51 – 56, Rotterdam, the Netherlands, aug 2003.

L. Ljung. *System Identification, Theory for the User*. Prentice Hall, New Jersey, 2nd edition, 1999.

R.G. Miller, Jr. *Beyond ANOVA*. Chapman and Hall, London, 1997.

D.C. Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, New York, 3rd edition, 1991.

H.Aa. Nielsen and H. Madsen. A generalization of some classical time series tools. *Computational Statistics & Data Analysis*, 37:13–31, 2001.

H. Pi and C. Peterson. Finding the embedding dimension and variable dependencies in time series. *Neural Computation*, 6:509–520, 1994.

L. Piroddi and W. Spinelli. Structure selection for polynomial narx models based on simulation error minimization. In P. van der Hof, B. Wahlberg, and S. Weiland, editors, *Proc. 13th IFAC Symposium on System Identification*, pages 371 – 376, Rotterdam, the Netherlands, aug 2003.

A. Poncet and G.S. Moschytz. Optimal order for signal and system modeling. In *IEEE International Symposium on Circuits and Systems, ISCAS '94.*, volume 5, pages 221–224, 1994.

C. Rhodes and M. Morari. Determining the model order of nonlinear input/output systems. *AIChE Journal*, 44(1):151–163, 1998.

J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Deylon, P.-Y. Glorennec, H. Hjalmarsson, and A. Juditsky. Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, 31:1691–1724, 1995.

T. Söderström. On model structure testing in system identification. *International Journal of Control*, 26: 1–18, 1977.

the Mathworks, Inc. *Matlab*. Natick, MA, USA. `http:\\www.mathworks.com`.

A. Tjøstheim and B.H. Auestad. Nonparametric identification of nonlinear time-series - projections. *Journal of The American Statistical Association*, 89:1398–1409, 1994a.

A. Tjøstheim and B.H. Auestad. Nonparametric identification of nonlinear time-series - selecting significant lags. *Journal of The American Statistical Association*, 89:1410–1419, 1994b.

Y.K. Truong. A nonparametric framework for time series analysis. In D. Billinger, P. Caines, J. Gewekw, E. Parzen, M. Rosenblatt, and M. S. Taqqu, editors, *New Directions in Time Series Analysis*, pages 371–386. Springer-Verlag, New York, 1993.

R. Tschernig and L.J. Yang. Nonparametric lag selection for time series. *Journal of Time Series Analysis*, 21: 457–487, 2000.

P. Vieu. Order choice in nonlinear autoregressive models. *Statistics*, 26:307–328, 1995.

Q. Yao and H. Tong. On subset selection in nonparametric stochastic regression. *Statistica Sinica*, 4: 51–70, 1994.

G.L. Zheng and S.A. Billings. Radial basis function network configuration using mutual information and the orthogonal least squares algorithm. *Neural Networks*, 9:1619–1637, 1996.