Regressor and Structure Selection in NARX Models Using a Structured ANOVA Approach

Ingela Lind, Lennart Ljung

Division of Automatic Control

E-mail: ingela@isy.liu.se, ljung@isy.liu.se

25th June 2007

Report no.: LiTH-ISY-R-2799
Accepted for publication in Automatica, 2007

Address:
Department of Electrical Engineering
Linköpings universitet
SE-581 83 Linköping, Sweden

WWW: http://www.control.isy.liu.se

Abstract

Regressor selection can be viewed as the first step in the system identification process. The benefits of finding good regressors before estimating complex models are especially clear for nonlinear systems, where the class of possible models is huge. In this article, a structured way of using the tool Analysis of Variance (ANOVA) is presented and used for NARX model (nonlinear autoregressive model with exogenous input) identification with many candidate regressors.

Keywords: identification, Structure identification, Analysis of variance

Regressor and Structure Selection in NARX Models Using a Structured ANOVA Approach

Ingela Lind and Lennart Ljung

Division of Automatic Control, Department of Electrical Engineering, Linköpings universitet, SE-583 37 Linköping, Sweden
e-mail: ingela.lind@saab.se, ljung@isy.liu.se

Abstract

Regressor selection can be viewed as the first step in the system identification process. The benefits of finding good regressors before estimating complex models are especially clear for nonlinear systems, where the class of possible models is huge. In this article, a structured way of using the tool Analysis of Variance (ANOVA) is presented and used for NARX model (nonlinear autoregressive model with exogenous input) identification with many candidate regressors.

Key words: Nonlinear system identification, Structure identification, Analysis of variance

1 Introduction

The general setting considered in this article is a nonlinear black-box system identification problem. We assume that the measurements $\mathcal{Z} = \{y(t), u(t)\}_{t=1,\ldots,N}$ can be reasonably well described by a static nonlinear function

$$y(t) = g(\varphi(t)) + e(t), \tag{1}$$

where $e(t)$ is Gaussian noise and $\varphi(t)$ is a regression vector formed from $\mathcal{Z}$. Regressor selection is to select the essential elements of $\varphi(t)$ (the regressors) to be included in the model, preferably without simultaneously estimating g. Structure selection is to get a basic idea of which regressors interact with each other and which do not, in order to keep the complexity of $g(\varphi(t))$ down.

The reason to bother about regressor and structure selection is to significantly reduce the effort needed to choose the function g and estimate its parameters. The cost reduction depends heavily on the number of candidate regressors considered. For a system with three candidate regressors, seven (2^3-1) different regressor combinations should be compared, while for a system with twenty candidate regressors there are $2^{20} - 1 \approx 1.05 \cdot 10^6$ regressor combinations. Any more refined structure identification, such as e.g. ordering of the regressors or considering interactions between regressors, results in an even higher growth rate of the problem size. With high-dimensional system descriptions, we also encounter the problem that the measurement data may not give sufficient basis for parameter estimation in the entire regressor space. This means that there is reason to preselect the regressors and do the time consuming parameter estimation with a fixed regression vector.

For linear systems, regressor selection is often done with criteria like AIC (Akaike, 1974), MDL (Rissanen, 1978, 1986) or several others. In recent years also nonlinear systems have gained more and more interest, e.g., Autin et al. (1992); Gunn and Kandola (2002); Korenberg et al. (1988); Piroddi and Spinelli (2003); Rhodes and Morari (1998); Spinelli et al. (2006). In Haber and Unbehauen (1990); Lind and Ljung (2005) and Roll et al. (2006), surveys of methods for regressor selection in nonlinear systems are given.

In previous work (Lind, 2000; Lind and Ljung, 2005), it was shown that the statistical tool ANOVA (e.g., Miller (1997); Montgomery (1991)) is useful for selecting regressors in quite small system identification problems where the true system is of nonlinear finite impulse response (NFIR) type. It has also been shown in a comparison with four other methods, selected among the ones above to give as varying approaches as possible, that ANOVA gives better and more homogeneous results on small-size problems (Lind, 2006; Mannale, 2006). All the methods suffer when the number of regressors to test grows. The contribution of this article is a systematic divide-and-conquer approach called Test of Interactions using Layout for Intermixed ANOVA (TILIA). This method shows homogeneously good results also for nonlinear autoregressive systems with exogenous input (NARX) and many candidate regressors.

2 ANOVA

The fundamental idea of ANOVA is to use an over-parameterised locally constant model to describe the data, a model that has a constant value for each box in an axis-orthogonal grid of the regressor space. The tool to achieve this end is an ANOVA function expansion (Friedman, 1991) with piecewise constant basis functions;

$$\begin{aligned}
&\mathcal{M}(\mathbf{c}, \theta, \varphi) = \\
&= c_0\theta_0 + \sum_{k=1}^{d} c_k \left(\sum_{i_1=1}^{m_k} \theta_{k;i_1} \mathbf{I}_{b_{(k,i_1)}}(\varphi_k) \right) \\
&+ \sum_{k=1}^{d-1} \sum_{l=k+1}^{d} c_{k,l} \left(\sum_{i_1=1}^{m_k} \sum_{i_2=1}^{m_l} \theta_{k,l;i_1,i_2} \mathbf{I}_{b_{(k,i_1)}}(\varphi_k) \mathbf{I}_{b_{(l,i_2)}}(\varphi_l) \right) \\
&+ \dots \\
&+ c_{1,2,\dots,d} \left(\sum_{i_1=1}^{m_1} \sum_{i_2=1}^{m_2} \dots \sum_{i_d=1}^{m_d} \theta_{1,2,\dots,d;i_1,i_2,\dots,i_d} \prod_{k=1}^{d} \mathbf{I}_{b_{(k,i_k)}}(\varphi_k) \right)
\end{aligned} \tag{2}$$

where $b_{(k,i)}$ is the ith interval relating to the regressor φ_k and $\mathbf{I}_b(x) = 1$ if $x \in b$ and zero otherwise. These intervals form the axis-orthogonal grid. The parameter vector $\mathbf{c}$ acts as decision variable in ANOVA and its elements can only assume the values 0 or 1. The influence of a regressor is described by a number of different terms of increasing complexity. In this expression $c_0\theta_0$ is the total mean, independent of all regressors. The part of the influence of each regressor, which is independent of all other regressors, is described by the terms in the first sum and called the *main effects*. The part of the influence of each regressor that also depend on *one* other regressor is described by the terms in the second sum and called two-factor *interaction effects* and so on. The number of regressors included in an effect is called the *interaction degree* of the effect. Each effect is described by a set of basis functions. Note that each effect has one corresponding parameter in $\mathbf{c}$, but can have many parameters in θ, one for each basis function. Since the model is strongly over-parameterised, many linear constraints are needed to give identifiability. These are denoted

$$\mathbf{A}\theta = 0, \tag{3}$$

where the details of the very structured matrix $\mathbf{A}$ can be found in Roll et al. (2006).

The same principles as for lasso (Tibshirani, 1996) and non-negative garrote (Breiman, 1995) can be used to describe ANOVA (Roll et al., 2006). Start with the model $\mathcal{M}(\mathbf{c}, \theta, \varphi(t))$, which is linear in the parameters θ. Use the objective function

$$V(\mathbf{c}, \theta) = \sum_{t=1}^{N} \Big(y(t) - \mathcal{M}\big(\mathbf{c}, \theta, \varphi(t)\big) \Big)^2. \tag{4}$$

In a first step, let $\hat{\theta}$ be the minimising argument to $V(\mathbf{1}, \theta)$, subject to (3). In the second step, solve

$$\begin{aligned}
\min_{\mathbf{c}} \quad & V(\mathbf{c}, \hat{\theta}) + J\|\mathbf{F}\mathbf{c}\|_1 \\
\text{subject to} \quad & \mathbf{c} \in \{0,1\}^{2^d}.
\end{aligned} \tag{5}$$

The penalty factors J and $\mathbf{F}$ are computed from $V(\mathbf{1}, \hat{\theta})$ and statistical F-tables. Also here, the details are not important in this context but can be found in Roll et al. (2006). The resulting model is usually sparse in the number of used regressors, similar to what is obtained by using the 1-norm constraints of lasso and non-negative garrote. The complexity of the obtained model is naturally restricted by the used function expansion, which forces, e.g., one-dimensional effects to be modelled by the one-dimensional terms only.

The piece-wise constant basis functions form a grid in the regressor space. The area in the regressor space that correspond to a specific combination of active basis functions (one for each regressor) is called a *cell*. If a data set covers the regressor space in such a way that all cells have an equal amount of data points, the data set is called *balanced*.

An alternative introduction to ANOVA is given in Lind and Ljung (2005) and detailed descriptions of both approaches are given in Lind (2006). Therein, also a discussion of the properties of the statistical tests, e.g., that they are robust in a nonlinear setting, and comparisons to several other regressor selection methods are given. In Roll et al. (2006) it was shown that ANOVA is not only similar to non-negative garrote and lasso, but is a special case of group non-negative garrote (Yuan and Lin, 2006).

3 TILIA: A Way to Structure the Use of ANOVA

Recall the problem (1). In Lind and Ljung (2005), the problem of selecting the elements of $\varphi(t)$ among input time lags was successfully solved by using ANOVA. The main hindrance for extension of this result to many candidate regressors is the explosion in the number of necessary data points needed for the analysis in high-dimensional regressor spaces, a problem that is not unique to this method. As will be shown in this contribution, this hindrance can be avoided, although not deleted.

One of the benefits with ANOVA is that the locally constant model (2) used, can be expressed with only a few model terms if the interaction degree is low. A rough model fit is sufficient since only the structure of the system is sought at this stage. This gives the hint that the complexity can be restricted if only interactions of relatively low interaction degree (2–4 regressors) are considered instead of full order interactions. Since data seldom offers support for models of higher dimension than this in the entire regressor space, the restriction on interaction degree will not give large effects on the final system model.

The idea is to treat the full problem with maybe 20–40 candidate regressors as a large number of small problems of more friendly sizes, three to seven candidate regressors each, where the number of regressors depends on correlations between candidate regressors and available data. Each of the smaller problems will be tested with ANOVA, a *basic test*. To test all interactions up to the restricted level in the full problem, each candidate regressor has to be included in many different basic tests. How many depends on which interaction degree is considered and on how many regressors are included in each basic test. The results from this large number of basic tests will be *intermixed* to give a composite value for each tested main and interaction effect for the full problem. By this divide-and-conquer approach, advice on an appropriate model structure can be extracted from the data. To keep the computational complexity down and give a fair comparison of the candidate regressors it is important how the basic tests are selected.

The following steps defines Test of Interactions using Layout for Intermixed ANOVA (TILIA):

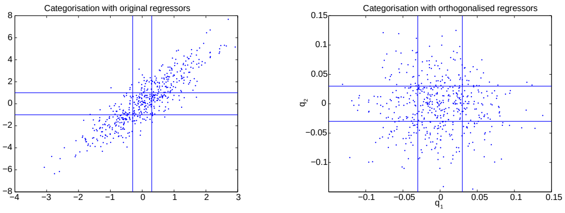
(1) Select regressors to test (Section 4.1).
(2) Select intervals (Section 4.2).
(3) Do test design (Section 4.3).
    For each test:
    (a) Balance data (Section 4.4).
    (b) Run basic test (Section 4.5).
(4) Combine basic test results (Section 4.6).
(5) Interpret results (Section 4.7).

The tasks of selecting candidate regressors, selecting intervals and balancing data are all dependent on each other and closely connected to the spread of the data in the regressor space. They are also important for the success of the method.

4 TILIA: Description of the Method

4.1 Orthogonalisation of Regressors

One issue to consider is whether to orthogonalise the regressors before regressor selection or not. If the regressors are strongly correlated, the piecewise constant basis functions for different regressors will be very close, giving numerical problems and problems caused by the representation of the regressor space. It might then be a better idea to tilt and skew the axis-orthogonal grid (Figure 1) in the regressor space formed by the piecewise constant basis functions. This can easily be done by using a QR-factorisation of the matrix formed by $[\varphi_1, \varphi_2, \dots, \varphi_d]$. An important drawback is that it matters in which order the regressors are put in the above matrix, since we consider nonlinear relations. A sparse representation of the model is only obtained if a good ordering is found.



(a) The sample distribution in the space spanned by the regressors φ_1 and φ_2.

(b) The sample distribution in the space spanned by the orthogonal basis vectors q_1 and q_2, which are obtained by QR-factorisation of φ_1 and φ_2.

Figure 1. Difference between categorisation of original regressors and orthogonalised regressors. The balance in the number of data between the different cells is much better in (b).

The advice is to use orthogonalisation if necessary to get reasonably balanced data, but not otherwise.

4.2 Categorisation of Data

The term *categorisation of data* is here used for the selection of the intervals for the piecewise constant basis functions in (2). These functions can be interpreted as index functions for which *cell* a data point belongs to. Think of a cell as a “voxel” in the space of candidate regressors.

When designing experiments for an intended analysis with ANOVA, balanced designs are in great demand, since they are the easiest to analyse and give the most trustworthy results. The reason is that a balanced design gives independence in the second optimisation problem (5), in the sense that each c_j parameter can be optimised individually and does not influence the other parameters. See also discussion in Lind (2006). In our case, balanced design means the same number of data in all cells. The distribution of the data into the cells depends, of course, strongly on the choice of cells. A bad choice of categorisation gives empty cells and large variations of the number of data in the different cells.

In TILIA, the implemented categorisation is very naive. Each candidate regressor is treated independently. A user defined *proportion vector*, which tells how many intervals to use and how large proportion of the data should be in each interval, is used to determine intervals of the continuous range of each regressor. The proportion vector, e.g., $[\alpha_1, \alpha_2, \alpha_3]$, is used together with the values of the regressor φ_j to find the intervals for the basis functions $\mathbf{I}_{b(j,1)}(\varphi_j)$, $\mathbf{I}_{b(j,2)}(\varphi_j)$ and $\mathbf{I}_{b(j,3)}(\varphi_j)$. The N values of φ_j in the data set are sorted in ascending order depending on magnitude. The interval $b(j,1)$ is selected as the range from the first sorted data point to data point $\alpha_1 N$ and includes the value of data point $\alpha_1 N$. $b(j,2)$ goes from data point $\alpha_1 N$ to data point $(\alpha_1 + \alpha_2)N$ and $b(j,3)$ from data point $(\alpha_1 + \alpha_2)N$ to the last one. Through the proportion vector, it is possible to choose the number of categories and to adjust the interval limits to the distribution of data in the regressor space. The purpose is to get a reasonable balance between the cells, which then is enhanced by the balancing. In our experience, three equal proportions suffices in most cases. There is reason to elaborate with the proportion vector if the distribution of data is very skewed or there is strong correlation between the regressors. A good multi-variable balance is usually obtained by taking a smaller proportion from the data ranges where there is plenty of data. The method is not very sensitive to the choice of proportion vector if data is plentiful, which can be seen in the simulated examples in Section 5. A huge potential for improvements of the interval selection is present, especially multi-variable methods, since the major drawback of the chosen categorisation method is that it treats each regressor independently, which works less well with correlated regressors.

4.3 Test Design

To illustrate the problem to be solved here, we start with a small example:

Example 1 *Assume that we have a dataset where we would like to test what candidate regressors $\{y(t-1), \ldots, y(t-9), u(t), \ldots, u(t-9)\}$ are important to explain $y(t)$, that is, 19 candidate regressors. We are interested in interactions of a degree that is possible to visualise in some way, which means up to 3-factor interactions. The amount of data we have got and the correlation between regressors present in it, limit the analysis to five regressors at a time, since there are empty areas in the regressor space. This means we have $\binom{19}{3} = 969$ 3-factor interactions to test and $\binom{19}{5} = 11628$ basic tests with five regressors each to choose from. Each basic test includes $\binom{5}{3} = 10$ 3-factor interactions. To test all 3-factor interactions means that the basic tests must overlap a bit and that a single 3-factor interaction might be tested more than once. What is the best test sequence?*

The purpose of the test design is to reduce the computational complexity while maintaining a good balance of the number of tests for different candidate regressors. In the example above, if all possible different basic tests would be used, each candidate regressor would be included in 3060 basic tests, and each 3-factor interaction would be tested 120 times. It is important that all effects of the same interaction degree are tested approximately the same number of times to get comparable composite values. In TILIA a randomised procedure to reduce the number of basic tests is used. This test design, which is given in Algorithm 1, reduces the number of tests to do and keeps approximately the same number of tests for all effects of the same interaction degree. A typical run of the randomised test design in the example above consist of 212–218 basic tests with each candidate regressor included in 53–60 basic tests and each 3-factor interaction tested 2–4 times.

Algorithm 1 Test design

```
Let $L_i$ be the matrix defining the interactions $v$ that should be tested, where $v$ is a row vector of the $n_i$ regressor symbols that define an interaction. $L_t$ is the matrix defining the basic tests to perform. The rows in $L_t$ are vectors of $n_r \geq n_i$ regressor symbols. Let $\text{int}_{n_i}(l)$ denote the effects of interaction degree $n_i$ included in the basic test defined by the row vector $l$.
If $n_r = n_i$ set $L_t = L_i$. Otherwise, begin at step 1.
(1) Set $L_t$ to an empty matrix with $n_r$ columns.
(2)a. Set $l$ to an empty row vector.
   b. Set $l = [l \quad v]$, where $v$ is a random row in $L_i$.
   c. Sort $l$ and remove replicates of regressor symbols.
   d. Repeat from 2b until length$(l) \geq n_r$, or $L_i \subseteq \text{int}_{n_i}(l)$.
(3) If $n_r \geq$ length$(l)$, let $\tilde{l} = l$, otherwise select $n_r$ regressor symbols from $l$ at random to obtain $\tilde{l}$.
(4) Delete $\text{int}_{n_i}(\tilde{l})$ from $L_i$.
(5) Add $\tilde{l}$ to $L_t$.
(6) Repeat from step 2 until $L_i$ is empty.
```

This procedure has the drawbacks of not having a perfectly even coverage (the same number of tests on all included regressors) and using more basic tests than what is strictly necessary, but works satisfactorily. A benefit is that since the sequence of tests is randomised, a different test sequence will be obtained if the test design is rerun. This can be used to stabilise the composite results. For example, four different *complete tests* (when all basic tests in a test sequence are run) will still give even coverage of the effects without repeating exactly the same test sequence. When the effects with highest interaction degree are tested more than once, a better balance between tests of effects with low and high interaction degrees is obtained. The tests of the effects with high interaction degree, that has become significant by pure chance, has less influence when the effect is tested more than once. Since also the data points used for each basic test are selected randomly, better confidence in the composite values will be obtained. Even with four complete tests, the computational complexity is much lower than for doing all the possible tests.

Example 2 *We will now make a test design for a simple case. Let $n_r = 3$, $n_i = 2$ and the number of regressors to test is 4. This means that there are three regressors included in each basic test and the maximum interaction degree tested is 2. The effects with highest interaction degree are in this case*

$$L_i = \begin{bmatrix} \varphi_1 & \varphi_2 \\ \varphi_1 & \varphi_3 \\ \varphi_1 & \varphi_4 \\ \varphi_2 & \varphi_3 \\ \varphi_2 & \varphi_4 \\ \varphi_3 & \varphi_4 \end{bmatrix}.$$

To find a sequence of basic tests to run, Algorithm 1 is used.

1 $L_t = [\,]$
2a $l = [\,]$
2b $l = \big[[\,]\ [\varphi_2 \quad \varphi_3]\big]$ *(selected randomly)*
2c $l = [\varphi_2 \quad \varphi_3]$
2d length$(l) < n_r$, *int*$_{n_i}(l) \subset L_i$. *Repeat from 2b*
2b $l = \big[[\varphi_2 \quad \varphi_3] \quad [\varphi_1 \quad \varphi_2]\big]$
2c $l = [\varphi_1 \quad \varphi_2 \quad \varphi_3]$
2d length$(l) = n_r$. *Continue to 3*
3 $\tilde{l} = l$
4 $L_i = \begin{bmatrix} \varphi_1 & \varphi_4 \\ \varphi_2 & \varphi_4 \\ \varphi_3 & \varphi_4 \end{bmatrix}$
5 $L_t = [\varphi_1 \quad \varphi_2 \quad \varphi_3]$
6 There are entries left in L_i. *Repeat from 2*
2a $l = [\,]$
2b $l = \big[[\,]\ [\varphi_2 \quad \varphi_4]\big]$
2c $l = [\varphi_2 \quad \varphi_4]$
2d length$(l) < n_r$, *int*$_{n_i}(l) \subset L_i$. *Repeat from 2b*
2b $l = \big[[\varphi_2 \quad \varphi_4] \quad [\varphi_1 \quad \varphi_4]\big]$
2c $l = [\varphi_1 \quad \varphi_2 \quad \varphi_4]$
2d length$(l) = n_r$. *Continue to 3*
3 $\tilde{l} = l$
4 $L_i = \begin{bmatrix} \varphi_3 & \varphi_4 \end{bmatrix}$
5 $L_t = \begin{bmatrix} \varphi_1 & \varphi_2 & \varphi_3 \\ \varphi_1 & \varphi_2 & \varphi_4 \end{bmatrix}$
6 There are entries left in L_i. *Repeat from 2*
2a $l = [\,]$
2b $l = \big[[\,]\ [\varphi_3 \quad \varphi_4]\big]$
2c $l = [\varphi_3 \quad \varphi_4]$
2d int$_{n_i}(l) = L_i$. *Continue to 3*
3 $\tilde{l} = l$
4 $L_i = [\,]$
5 $L_t = \begin{bmatrix} \varphi_1 & \varphi_2 & \varphi_3 \\ \varphi_1 & \varphi_2 & \varphi_4 \\ \varphi_3 & \varphi_4 & [\,] \end{bmatrix}$
6 L_i *is empty. The test sequence* L_t *is determined.*

Now the sequence of basic tests is determined. In the first basic test, φ_1, φ_2 and φ_3 will be included, in the second test, φ_1, φ_2 and φ_4, and in the third test φ_3 and φ_4. In this case, the interaction between φ_1 and φ_2 will be tested twice. Note also that all regressors are included in two tests each.

4.4 Balance data

When the categorisation intervals are defined for all regressors and the sequence of basic tests decided, the actual balancing can be done. The objective of the balancing is to get an as equal number of data as possible in the cells for the basic test. Excess data from the cells are discarded randomly. The categorisation is determined once and used for all the basic tests, while the balancing is remade for each basic test, only considering the categorisation for the regressors included in the basic test. Some unbalance due to, e.g., an extremely unusual regressor value combination works fine. As was shown in Lind (2006), between four and six data are sufficient in each cell and good results are also obtained with a ratio between the maximum and minimum number of data/cell of 3 to 5. The main considerations when deciding how many data should be the maximum in each cell are:

- The larger difference between the maximum and minimum number of data in the cells — the less reliable estimates both within each basic test and for the composite values.
- The less data in each cell — the less dependence between different basic tests, since the probability that the same data points are reused in different basic tests is lower.
- The more data in each cell — the more reliable estimates within each basic test, due to higher power.

An important point is that the balancing is made for each basic test separately. **This is the reason why a data set can be sufficient for TILIA, but not for a complete analysis with one larger test.** It is much easier to balance data in several lower-dimensional subspaces than in one high-dimensional subspace.

4.5 Basic Tests

The term basic test refer to an ANOVA test of a specific combination of candidate regressors to a specified interaction degree n_i. A basic test is a subproblem of a full structure identification problem with many regressors. First, the data set is balanced with respect to the included regressors, according to Section 4.2, and then a fixed-level ANOVA is run. All effects with interaction degree $\leq n_i$, formed from the included regressors, are tested. The result is the probabilities of the null hypothesis ($c_j = 0$) for each tested effect, which means that the probability values are low for *important* effects. (The statistical interpretation of (5) is F-distributed hypothesis tests for each c_j).

In the implemented version of TILIA, the basic tests are done with the anovan routine in MATLAB. Empty cells and (the restricted) unbalance is automatically treated in this routine. The output is an ANOVA table. In the table, each tested main and interaction effect is given a probability value p. An effect is denoted significant if the value is below a predefined value α, which often is set to 0.05.

When a basic test is analysed, the remaining candidate regressors are viewed as noise contributions. This is of course not a valid assumption, since then they would not have been candidate regressors. By choosing data points from the data set randomly (in each cell defined by the candidate regressors included in the basic test) the effects from the neglected regressors are hopefully minimised. A result of neglecting the rest of the candidate regressors is that the error sum of squares (related to the variance estimate, see the rows marked Error in Table 1) is estimated as too large. This makes it harder to find small effects from the included candidate regressors.

4.6 Combining Test Results/Composite Tests

Assume that, as in Example 2, regressors are included in more than one basic test. In the example we have three different ANOVA tables, where each main effect and also the interaction between φ_1 and φ_2 have got two different probability values, p. If the data set is not large enough to give each test a unique set of test data, the basic tests will be dependent. The problem here is how to combine (or compose) two values of p, obtained in different basic tests, concerning the same main or interaction effect. We would like the following properties for the *composite* value:

- Several significant basic tests should result in a significant composite value.
- Some significant and some insignificant basic tests should result in an insignificant composite value. (These results can occur when several correlated candidate regressors are tested in different basic tests. If only one of them is present in the basic test, the test shows significance since the candidate regressor has explaining power due to its correlation with an important regressor. If any of the important regressors is included in the same basic test, the candidate regressor is tested as insignificant.)
- Several insignificant basic tests should result in a not significant composite value.
- A main or interaction effect included in many basic tests should not be obviously mistreated by being tested many times. The probability of Type I errors in at least one of quite many tests is very large.

Assume that the effect studied has been included in K basic tests and has got the probability level p_k from test $k \in \{1, \dots K\}$. If none of the basic tests including this effect shows significance, the smallest value of p_k is larger than the significance level (1− the probability of Type I error);

$$\min_k p_k > \alpha. \tag{6}$$

This is equivalent to

$$\max_k (1 - p_k) < (1 - \alpha). \tag{7}$$

$\max_k(1 - p_k)$ will be denoted $\lceil 1 - p \rceil$ in the tables. The related quantity $\min_k(1 - p_k)$ will be denoted $\lfloor 1 - p \rfloor$. If either the arithmetic average (AA)

$$\frac{1}{K} \sum_{k=1}^{K} (1 - p_k), \tag{8}$$

or the geometric average (GA)

$$\Big(\prod_{k=1}^{K} (1 - p_k) \Big)^{\frac{1}{K}}, \tag{9}$$

is larger than $1 - \alpha$, there is reason to believe that the effect is significant. The geometric average is closest to what is used in statistics for computing the simultaneous confidence level for multiple comparisons. The arithmetic average is not affected as severely as the geometric average if a single test has a high value of p_k. Also the median of p_k could be interesting to investigate. The effects are ordered in importance by $\prod_{k=1}^{K}(1 - p_k)$, which is denoted $\prod$. A drawback with such an ordering is that effects tested many times have a disadvantage against effects tested few times, due to the probability of Type I errors in each test. Now the procedure is as follows:

(1) Compute the table with the values $\prod$, $\lfloor 1 - p \rfloor$, $\lceil 1 - p \rceil$, arithmetic average and geometric average of $(1 - p)$ for each effect.
(2) Remove effects where $\lceil 1 - p \rceil < 1 - \alpha$, since these effects are never tested as significant.
(3) Remove effects where both averages are less than $1 - \alpha$. This will take care of the cases where an effect is tested as significant in one regressor set but not in combination with other regressor sets.
(4) Sort results according to $\prod_{k=1}^{K}(1 - p_k)$.

From the achieved table, it is most often clear what effects are important for explaining the data (e.g., by a

Table 1
Results from the basic tests in Example 3. These are the tables given by Matlabs `anovan` routine. The first columns give the sums of squares (SS), the corresponding degrees of freedom (df), mean squares (MS) and corresponding test variable value (F). The last column state the probability that there is no significant effect.

Source	SS	df	MS	F	p
φ_1	156.517	2	78.2583	60.99	0
φ_2	65.702	2	32.8509	25.6	0
φ_3	0.164	2	0.0821	0.06	0.9381
(φ_1, φ_2)	5.076	4	1.2689	0.99	0.4203
(φ_1, φ_3)	3.964	4	0.991	0.77	0.5473
(φ_2, φ_3)	1.556	4	0.3891	0.3	0.8747
Error	79.549	62	1.2831		
Total	312.528	80			
φ_1	318.095	2	159.047	115.67	0
φ_2	97.135	2	48.567	35.32	0
φ_4	2.059	2	1.029	0.75	0.4772
(φ_1, φ_2)	13.062	4	3.266	2.37	0.0616
(φ_1, φ_4)	11.49	4	2.873	2.09	0.0929
(φ_2, φ_4)	3.888	4	0.972	0.71	0.5902
Error	85.249	62	1.375		
Total	530.979	80			
φ_3	25.674	2	12.8369	2.31	0.1283
φ_4	34.816	2	17.408	3.13	0.0682
(φ_3, φ_4)	9.464	4	2.3661	0.43	0.7885
Error	100.165	18	5.5647		
Total	170.119	26			

"gap" in the values in some or all of the columns). Ideally, the significance level α is chosen such that all effects above the "gap" are included in the table and all effects below the "gap" excluded. There are cases where it is not obvious though. In these cases one has to manually inspect the obtained table and decide which effects seem reasonable to include in the model.

Example 3 *Consider again Example 2. Assume that the three basic test in L_t give the ANOVA tables in Table 1. Only the last column (p) is considered here. We compute composite p values using all the measures in Section 4.6. This yields Table 2, where no pruning or ordering is done. In this example, the conclusions are clear cut, since only the effects from φ_1 and φ_2 have got values of $\lceil 1-p \rceil > 0.95$. These are the only effects that are significant in any test.*

More examples are given in Section 6.

In the procedure above, the dependence between the different tests is neglected. This means that if the data set is small with respect to how many and how complex basic tests are performed, the neglected dependence is important, and the results *might* be badly influenced by it.

Table 2
Composite p values for the results in Table 1. The number of basic test including the effect is denoted by n_b,and p is the probability value from the ANOVA table. The different columns give the product of these probabilities for each basic test, the minimum probability, the maximum probability and the arithmetic and geometric averages of the probabilities.

Effect	n_b	$\prod$	$\lfloor 1-p \rfloor$	$\lceil 1-p \rceil$	AA	GA
φ_1	2	1	1	1	1	1
φ_2	2	1	1	1	1	1
(φ_1, φ_4)	1	0.907	0.907	0.907	0.907	0.907
(φ_1, φ_2)	2	0.544	0.580	0.938	0.759	0.738
φ_4	2	0.487	0.523	0.932	0.727	0.698
(φ_1, φ_3)	1	0.453	0.453	0.453	0.453	0.453
(φ_2, φ_4)	1	0.410	0.410	0.410	0.410	0.410
(φ_3, φ_4)	1	0.212	0.212	0.212	0.212	0.212
(φ_2, φ_3)	1	0.125	0.125	0.125	0.125	0.125
φ_3	2	0.054	0.062	0.872	0.467	0.232

Another approach to compose values would be to weigh together the sum of squares values directly, keeping track of the dependence between different tests, and as the last step perform hypothesis tests, instead of performing the hypothesis tests first and then weigh them together. The statistics involved seem quite complicated.

4.7 Interpreting Results

The list obtained in Section 4.6 should include all important regressors and their interactions. The results get more stable and reliable if two to four complete test sequences are designed using the randomised test design and combined in the same manner as in Section 4.6. The balance between the main and higher order interactions gets better when the latter are tested more than once, which means that less spurious high order interactions are tested as significant without missing any main effects. The price is the higher amount of tests.

There are a few things to keep in mind:

- Due to correlation between regressors, the list of important effects obtained from the composite values can also include spurious regressors. The spurious regressors should have a lower value of $\prod_{k=1}^{K}(1-p_k)$, since they should have been tested as not significant at least once. The correlation between regressors can be used as a warning sign. In further model building, include the dubious regressors, but scrutinise their contributions extra carefully.

Table 3
User parameters for TILIA used in the test examples. The proportion vector is used for categorisation of data, see Section 4.2. The number of included regressors in each basic test is denoted by n_r and the tested degree of interaction with n_i.

Test example	Proportion vector	n_r	n_i
1: Chen 1	[1/3 1/3 1/3]	4	2
2: Chen 2, test 1	[1/3 1/3 1/3]	4	2
2: Chen 2, test 2	[1/3 1/3 1/3]	4	3
3: Chen 3	[1/3 1/3 1/3]	5	2

- If the number of data grows, the list of important effects usually grows slowly too, since the power (1 - the probability of Type II error) of the tests grows with the number of data. With a fixed number of the maximum amount of data in each cell, this consequence is minimised. Of course, the number of data can be tuned to get a desired power for a certain model (see further Lind (2006)).
- Another issue is that the basic tests are dependent, since they partly will be performed on the same data. An ill-placed outlier could affect the analysis badly, especially if placed in a region with few good data. This means that the influence of the outlier is not averaged out among many data points in the same region and also that it is "reused" for different basic tests more often than a data point from a region with more data. See also Barnett and Lewis (1978).

In Section 5 we will see how TILIA works on simulated data and in Section 6 some real data sets will be treated.

5 Structure Selection on Simulated Test Examples

The following simulated examples are taken from Chen et al. (1995). These are all nonlinear autoregressive (NAR) systems. Several more examples, also including NAR system with exogenous input (NARX), and comparisons with other approaches, are given in Lind (2006). Here, the signal-to-noise ratio varies from example to example. Data series with 3000 input/output data are used for all examples. TILIA was used to identify the structure, using the parameter settings of Table 3.

5.1 Example 1: Chen 1

The first example is a nonlinear additive autoregressive process,

$$y(t) = 2e^{-0.1y^2(t-1)}y(t-1) - e^{-0.1y^2(t-2)}y(t-2) + e(t), \tag{10}$$

where $e(t)$ is Gaussian noise with standard deviation 1. The model is similar to an exponential autoregressive model, but has different time lags in the exponent so that it is additive. The model, in its original context, was used together with Example 2 to test algorithms for spotting additivity in NAR systems. The candidate regressors are $\{y(t-1), \dots, y(t-8)\}$.

The results from the systematic ANOVA tests were that $y(t-1)$ and $y(t-2)$ should be included additively in the model, both if the regressors were used as they are, or if they were orthogonalised in the given order before the ANOVA tests (see Table 4). This test took less than one minute to run and about the same amount of time to interpret.

5.2 Example 2: Chen 2

The second example is almost the same as the first, but this is an exponential autoregressive model,

$$y(t) = e^{-0.1y^2(t-1)}(2y(t-1) - y(t-2)) + e(t), \tag{11}$$

where $e(t)$ is Gaussian noise with standard deviation 1. Also here, the candidate regressors were $\{y(t-1), \dots, y(t-8)\}$.

All the numerical results from TILIA are given in Table 4. These results were obtained by running the method two times; first with candidate regressors $\{y(t-1), \dots, y(t-8)\}$, both with the regressors as they are and with orthogonalised regressors in the given order (see Section 4.1). The first approach gave the results that $y(t-1)$ interacts with $y(t-2)$ and with $y(t-3)$ and that the interaction between $y(t-2)$ and $y(t-4)$ also is important. The orthogonalised approach gives the important interactions $y(t-1)$ with $y(t-2)$, $y(t-2)$ with $y(t-3)$, and $y(t-2)$ with $y(t-5)$. The results gave no strong reason to change the orthogonalisation order, since the important regressors in the results also are first among the orthogonalised ones.

The regressors $\{y(t-6), \dots, y(t-8)\}$ were then excluded and the method rerun on the remaining candidate regressors. At the same time, a higher interaction degree was tested. In the second run, both approaches identified correctly the interaction between $y(t-1)$ and $y(t-2)$, but with the original regressors, also $y(t-3)$ and $y(t-4)$ were found as giving additive contributions. The reason could be the strong correlation with the true lags. (Remember that not all regressors are included in all tests. When $y(t-1)$ is missing, e.g., $y(t-3)$ could explain parts of the contribution from $y(t-1)$.) The number of tests where these effects are tested, n_b, confirm this theory. For all effects with the same interaction degree, the design gives an approximately equal number of test. Depending on the data set, some of the intended tests get empty cells, resulting in no computed probability levels and a smaller value of n_b, which indicates bad balancing. The tests with empty cells tend to be the tests where correlated candidate regressors are tested together, e.g.,

Table 4
Composite p values. The number of basic test including the effect is denoted by n_b, and p is the probability, according to the hypothesis test, that the sum of squares are large not purely by random. The different columns give the product of these probabilities for each basic test, the minimum probability, the maximum probability and the arithmetic and geometric averages of the probabilities.

Test example	Effect	n_b	$\prod$	$\lfloor 1-p \rfloor$	$\lceil 1-p \rceil$	AA	GA
(10)	$y(t-1)$	15	1	1	1	1	1
	$y(t-2)$	15	0.894	0.894	1	0.993	0.993
(10) The regressors were orthogonalised before TILIA.	$\tilde{y}(t-1)$	13	1	1	1	1	1
	$\tilde{y}(t-2)$	14	0.888	0.892	1	0.992	0.992
(11)	$y(t-1)$	14	1	1	1	1	1
	$(y(t-2), y(t-4))$	10	0.890	0.902	1	0.989	0.988
	$(y(t-1), y(t-2))$	8	0.800	0.854	1	0.974	0.973
	$(y(t-1), y(t-3))$	9	0.726	0.786	1	0.968	0.965
(11) The regressors were orthogonalised before TILIA.	$\tilde{y}(t-1)$	13	1	1	1	1	1
	$\tilde{y}(t-2)$	12	1	1	1	1	1
	$(\tilde{y}(t-1), \tilde{y}(t-2))$	4	0.999	0.999	1	1	1
	$(\tilde{y}(t-2), \tilde{y}(t-3))$	5	0.994	0.994	1	0.999	0.999
	$(\tilde{y}(t-2), \tilde{y}(t-5))$	5	0.801	0.896	0.996	0.958	0.957
(11) $\{y(t-6), \ldots, y(t-8)\}$ excluded.	$y(t-1)$	14	1	1	1	1	1
	$y(t-4)$	7	1	1	1	1	1
	$y(t-3)$	9	0.998	0.999	1	1	1
	$(y(t-1), y(t-2))$	13	0.937	0.966	1	0.995	0.995
(11) $\{y(t-6), \ldots, y(t-8)\}$ excluded. The regressors were orthogonalised before TILIA.	$\tilde{y}(t-2)$	19	1	1	1	1	1
	$\tilde{y}(t-1)$	19	1	1	1	1	1
	$(\tilde{y}(t-1), \tilde{y}(t-2))$	11	0.991	0.993	1	0.999	0.999

$y(t-1)$ with $y(t-3)$ which would give a low value of $1-p$ and lower composite values of $y(t-3)$. Based on the n_b-values of the non-orthogonalised test in this case, orthogonalisation should give better results and also did so.

The result was that the interaction between $y(t-1)$ and $y(t-2)$ was correctly identified by using the orthogonalised approach.

5.3 *Example 3: Chen 3*

The third example is an additive threshold autoregressive model (Chen et al., 1995),

$$\begin{aligned} y(t) = &-2y(t-1)I\big(y(t-1) \leq 0\big) \\ &+ 0.4y(t-1)I\big(y(t-1) > 0\big) + e(t), \end{aligned} \qquad (12)$$

where $e(t)$ is Gaussian noise with standard deviation 1 and $I(x)$ is an indicator such that $I(x) = 1$ if x holds. Candidate regressors are $\{y(t-1), \ldots, y(t-8)\}$.

The regressor $y(t-1)$ was found correctly both when the regressors were tested as they are and when they were orthogonalised first. No other candidate regressors were important enough to be included in the table of composite values and the composite values for the 9 basic tests including $y(t-1)$ are all 1.

6 Structure Selection on Measured Data Sets

6.1 *'Silver Box' data*

The 'silver box' data are sampled from an electrical circuit (looking like a silver box) and should in theory be described by

$$m\frac{d^2y(t)}{dt} + d\frac{dy(t)}{dt} + ay(t) + by(t)^3 = u(t). \qquad (13)$$

This dataset is due to Pintelon and Schoukens (2001) and was studied in a special session at the 6th IFAC-Symposium on Nonlinear Control Systems (NOLCOS, 2004). The data set consist of a validation data set, the "arrow head" in Figure 2, with 40000 samples, and an estimation data set of 86916 samples in the trailing part.

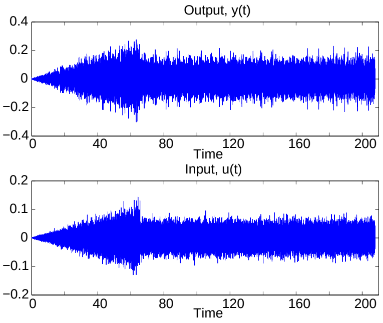


Figure 2. Silver box data. The first 40000 samples are validation data and the last 86916 samples are estimation data.

Table 5
The most important regressors and two-factor interactions for the silver box data when testing candidate regressors without orthogonalisation. The number of basic test including the effect is denoted by n_b,and p is the probability, according to the hypothesis test, that the sum of squares are large not purely by random. The different columns give the product of these probabilities for each basic test, the minimum probability, the maximum probability and the arithmetic and geometric averages of the probabilities. Here y_x and/or u_x is short for $y(t-x)$ and $u(t-x)$ respectively.

Effect	n_b	$\prod$	$\lfloor 1-p \rfloor$	$\lceil 1-p \rceil$	AA	GA
y_4	30	1	1	1	1	1
y_1	32	1	1	1	1	1
u_2	29	1	1	1	1	1
y_5	32	1	1	1	1	1
y_8	29	0.948	0.973	1	0.998	0.998
(y_4, y_9)	6	0.821	0.910	0.998	0.968	0.968
u_7	31	0.803	0.864	1	0.993	0.993
(y_2, y_3)	6	0.743	0.897	1	0.953	0.952
y_2	32	0.721	0.758	1	0.991	0.990
u_1	29	0.719	0.719	1	0.990	0.989
y_9	32	0.681	0.703	1	0.990	0.988
y_6	30	0.616	0.633	1	0.987	0.984
y_7	27	0.576	0.538	1	0.979	0.973
y_3	28	0.451	0.750	1	0.975	0.972
u_3	30	0.197	0.239	1	0.969	0.947
u_6	34	0.026	0.049	1	0.958	0.899

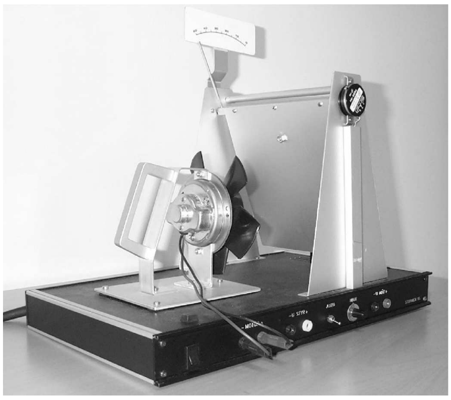


Figure 3. The vane process. The fan to the left blows at the vane to the right. The vane swings about its upper edge.

The sampling period is 0.0016384 seconds. The validation data are chosen to give the ability to test the generalisation capability of the estimated models . TILIA was applied to the estimation data set. Candidate regressors were $\{y(t-1), \ldots, y(t-9), u(t), \ldots, u(t-9)\}$, that is, 19 candidates. The proportion vector was chosen as [1/3 1/3 1/3] for all regressors, the number of regressors included in each basic test was 5 and second order interactions were tested. To stabilise the result, the complete test was repeated four times (see Section 4.3). Each complete test consist of 27 to 29 basic tests. The most important effects are given in Table 5. In this case, no good ordering for orthogonalised regressors was found. The result is that all output regressors, $\{y(t-1), \ldots, y(t-9)\}$, and the input regressors $\{u(t-1), \ldots, u(t-3)\}$, $u(t-6)$ and $u(t-7)$ should be included in the model. Most of the effects are additive. The important interaction effects are between $y(t-2)$ and $y(t-3)$ and between $y(t-4)$ and $y(t-9)$. The suggested model structure is then

$$\begin{aligned} y(t) = & g_1\big(y(t-1)\big) + g_2\big(y(t-2), y(t-3)\big) \\ & + g_3\big(y(t-4), y(t-9)\big) + g_4\big(y(t-5)\big) \\ & + g_5\big(y(t-6)\big) + g_6\big(y(t-7)\big) + g_7\big(y(t-8)\big) \\ & + g_8\big(u(t-1)\big) + g_9\big(u(t-2)\big) + g_{10}\big(u(t-3)\big) \\ & + g_{11}\big(u(t-6)\big) + g_{12}\big(u(t-7)\big). \end{aligned} \tag{14}$$

This model structure uses almost the same regressors as the two best (in RMSE sense) models of Table 1 in Ljung et al. (2004) of the NOLCOS comparison. Note that quite many old outputs are needed in the model.

6.2 Nonlinear Laboratory Process

The vane process consists of an air fan mounted 15 cm in front of a 20×20 cm vane (see Figure 3). The input signal is the voltage over the motor and the output signal is the output voltage from the angle meter. To give the process a more nonlinear behaviour, the motion of the vane is blocked by a slightly damping object. Both the range of the input and the range of the output are limited to −10 V to 10 V, due to limitations in the instrumentation.

6.2.1 Input Selection

The chosen input signal is a pseudo-random multi-level input signal with three levels. To get all $3^7 = 2187$ signal level combinations in the sequence $u(t)$, $u(t-1)$, $u(t-2)$,

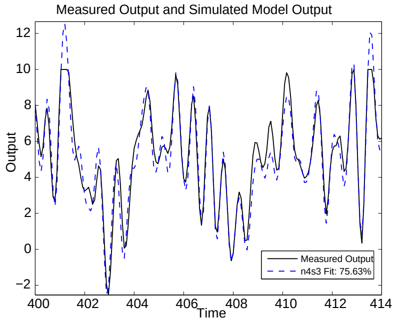


Figure 4. Simulated outputs from the linear model. The solid line is the measured output and the dashed line is the simulated output from the linear model.

$u(t-3)$, $u(t-4)$, $u(t-5)$, $u(t-6)$ the equation

$$u(t) = \operatorname{mod}\big(u(t-5) + 2u(t-7), 3\big) \tag{15}$$

is used to generate the signal. Here $\operatorname{mod}(x,3)$ stands for modulo three. This equation can only generate 3^7-1 different signal level combinations, since if there are seven or more zeros in a row, the output will be constantly zero. Since it is important in the intended analysis to have measurements of all signal level combinations, a zero is appended in each period of the signal. Then the three signal levels $\{0,1,2\}$ are mapped to the desired levels $\{5,1,9\}$ of the signal.

The sampling period is chosen such that 4–8 samples can be taken during the rise time ($= 0.3\,\text{s}$) of a step response. With 4 samples during the rise time this gives the sampling period $0.08\,\text{s}$.

6.2.2 Linear identification

Several linear models of different orders were tried out. To handle the offset from zero mean, a second, constant, input was included in the models. The model that showed the best performance on validation data was a state space model of order three;

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) + Ke(t)\\ y(t) &= Cx(t) + e(t), \end{aligned}$$

where A is a 3×3 matrix, B is a 3×2 matrix, K is a 3×1 matrix and C is a 1×3 matrix. This third order, two input, one output system has 12 identifiable parameters. The model was estimated using the first half (8086 samples) of the data set as estimation data. The fit for the linear model was 75% on the second half of the data set (see Figure 4) for a zoom in on 175 data points. The linear model does not handle the saturation caused by the blocking of the vane very well, giving large overshoots when the measured signal saturates and not enough amplitude in the other oscillations. The residuals from the linear model almost pass a whiteness test and the correlation between the residuals and the input is insignificant, see Figure 5. This means that there is essentially nothing more to gain from a linear model.

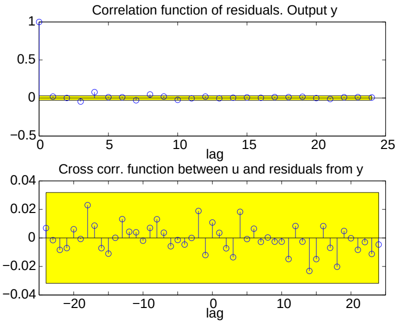


Figure 5. Residual analysis for the linear model.

Table 6
Results from TILIA on data from the vane process. Table headings are explained in Table 2. Here y_x and/or u_x is short for $y(t-x)$ and $u(t-x)$ respectively.

Effect	n_b	$\prod$	$\lfloor 1-p\rfloor$	$\lceil 1-p\rceil$	AA	GA
y_1	428	1	1	1	1	1
u_4	450	0.795	0.962	1	1	1
y_2	402	0.207	0.374	1	0.997	0.996
u_5	420	0.035	0.343	1	0.994	0.992
u_3	435	0.002	0.200	1	0.988	0.985
y_8	397	0.000	0.176	1	0.971	0.961

Table 7
Results from TILIA on data from the vane process. The candidate regressors are orthogonalised. Column headings are explained in Table 2. The orthogonalised regressors are denoted by $\tilde{y}_x$ and $\tilde{u}_x$ respectively.

Effect	n_b	$\prod$	$\lfloor 1-p\rfloor$	$\lceil 1-p\rceil$	AA	GA
$\tilde{y}_1$	451	1	1	1	1	1
$\tilde{y}_2$	421	1	1	1	1	1
$\tilde{y}_3$	433	0.003	0.333	1	0.989	0.987

6.2.3 TILIA

TILIA was applied to the entire data set. Candidate regressors were $\{y(t-1),\dots,y(t-9),u(t),\dots,u(t-9)\}$, that is, 19 candidates. The proportion vector was chosen as $[1/3\ 1/3\ 1/3]$ for all regressors, the number of regressors included in each basic test was 4 and third order interactions were tested. The balance of the tests was enforced by only using three samples from each cell. To stabilise the result, the complete test was repeated four times. Each complete test consisted of about 480 to 530 basic tests. The most important effects are given in Table 6, for the candidate regressors tested directly and in Table 7, for the candidate regressors tested after orthogonalisation by QR-factorisation in the order given

Table 8
Results from TILIA on data from the vane process. The candidate regressors are orthogonalised in the order $y(t-1)$, $y(t-2)$, $y(t-3)$, $u(t-4)$, $u(t-5)$, $u(t-3)$, $y(t-8)$, $y(t-4)$, $y(t-5)$, $u(t)$, $u(t-1)$, $y(t-6)$, $y(t-7)$, $u(t-2)$, $u(t-6)$, $y(t-9)$, $u(t-7)$, $u(t-8)$, $u(t-9)$. The results almost maintain the order among the regressors, so a truncation after the seventh regressor is possible. Column headings are explained in Table 2.

Effect	n_b	$\prod$	$\lfloor 1-p \rfloor$	$\lceil 1-p \rceil$	AA	GA
$\tilde{y}_1$	422	1	1	1	1	1
$\tilde{y}_2$	427	1	1	1	1	1
$(\tilde{y}_1, \tilde{u}_4, \tilde{u}_5)$	7	0.96	0.98	1	0.99	0.99
$\tilde{y}_3$	433	0.00	0.47	1	0.98	0.98
$\tilde{y}_8$	423	0.00	0.03	1	0.98	0.96
$\tilde{u}_5$	460	0.00	0.04	1	0.95	0.93

Table 9
Results from TILIA on data from the vane process. The candidate regressors are orthogonalised in the order $y(t-1)$, $y(t-2)$, $y(t-3)$, $u(t-4)$, $u(t-5)$, $y(t-8)$, $u(t-3)$. The order of the regressors is maintained in the results, so the orthogonalisation is successful and a final regressor selection is reached. Column headings are explained in Table 2.

Effect	n_b	$\prod$	$\lfloor 1-p \rfloor$	$\lceil 1-p \rceil$	AA	GA
$\tilde{y}_1$	32	1	1	1	1	1
$\tilde{y}_2$	30	1	1	1	1	1
$(\tilde{y}_1, \tilde{u}_4, \tilde{u}_5)$	5	0.95	0.98	1	0.99	0.99
$\tilde{y}_3$	31	0.52	0.70	1	0.98	0.98
$\tilde{y}_8$	30	0.15	0.38	1	0.95	0.94

above. To get a more sparse representation, the regressors were reordered according to the order in Table 7 (and the last ones according to Table 6). After rerunning TILIA, the most important effects were among the first seven candidate regressors in Table 8. After truncation of superfluous regressors, the results in Table 9 were obtained. The now maintained ordering of the regressors makes the orthogonalised test useful for regressor selection. The suggested model structure is

$$\begin{aligned} \hat{y}(t) =& g_1(\tilde{y}_1, \tilde{u}_4, \tilde{u}_5) + g_2(\tilde{y}_2) \\ &+ g_3(\tilde{y}_3) + g_4(\tilde{y}_8), \end{aligned} \tag{16}$$

where the tilde-denoted regressors are the original candidate regressors QR-factorised in the order

$$[\tilde{y}_1,\ \tilde{y}_2,\ \tilde{y}_3,\ \tilde{u}_4,\ \tilde{u}_5,\ \tilde{y}_8]R = A \tag{17}$$

with $A = [y(t-1),\ y(t-2),\ y(t-3),\ u(t-4),\ u(t-5),\ y(t-8)]$.

6.2.4 Nonlinear Estimation

A nonlinear model with the structure of (16) using an artificial neural network with totally 40 sigmoids in the hidden layers was estimated. The fit of the model was 85% and it was able to model the saturations in the measured data, see Figure 6. The time needed to run TILIA with nineteen candidate regressors is about the same as the time needed to estimate *one or two* of these nonlinear models.

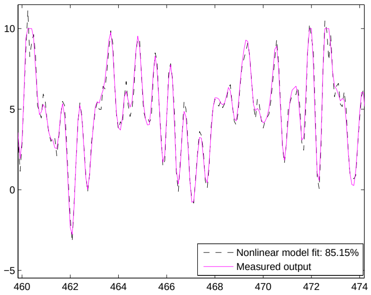


Figure 6. Simulated outputs from the nonlinear model. The dashed line is the measured output and the dashed line is the simulated output from the nonlinear model.

7 Conclusions

A systematic method (TILIA) for regressor selection using ANOVA has been presented and tested both on simulated and measured data sets with many candidate regressors. In earlier work (Lind, 2006; Mannale, 2006), ANOVA has been compared with other methods on small-size identification problems, and given better and more homogeneous results. All methods for regressor selection have more or less severe problems with extensions to many regressors. TILIA is a way to reduce large problems to sets of sub-problems, which can be treated by ANOVA. It has been shown here that TILIA gives homogeneously good results in test cases with up to 19 candidate regressors. As pointed out, for example in Section 4.2, there are nevertheless many possible improvements to TILIA. These include improvements of the categorisation, a more structured way to order regressors for orthogonalisation, the methods to combine probability values, and computational aspects.

References

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

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.

W. Barnett and T. Lewis. *Outliers in Statistical Data.* Wiley, New York, 1978.

L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, Nov 1995.

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

J. H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19(1):1–67, Mar 1991.

S. R. Gunn and J. S. Kandola. Structural modeling with sparse kernels. *Machine Learning*, 48:137–163, 2002.

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

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.

I. Lind. Model order selection of N-FIR models by the analysis of variance method. In *Proceedings of the 12th IFAC Symposium on System Identification*, pages 367–372, Santa Barbara, Jun 2000.

I. Lind. *Regressor and Structure Selection: Uses of ANOVA in System Identification.* PhD thesis, Linköpings universitet, Linköping, Sweden, May 2006.

I. Lind and L. Ljung. Regressor selection with the analysis of variance method. *Automatica*, 41(4):693–700, Apr 2005.

L. Ljung, Q. Zhang, P. Lindskog, and A. Juditsky. Modeling a non-linear electric circuit with black box and grey box models. In *Proceedings of NOLCOS 2004 — IFAC Symposium on Nonlinear Control Systems*, pages 543–548, Stuttgart, Germany, Sep 2004.

R. Mannale. Comparison of regressor selection methods in system identification. Technical Report LiTH-ISY-R-2730, Department of Electrical Engineering, Linköping University, Feb 2006.

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.

NOLCOS. Special session on identification of nonlinear systems: The silver box study. In *Proceedings of the 6th IFAC-Symposium on Nonlinear Control Systems, Stuttgart, Germany. September 01–03*, 2004.

R. Pintelon and J. Schoukens. System identification—a frequency domain approach. IEEE Press., 2001.

L. Piroddi and W. Spinelli. An identification algorithm for polynomial narx models based on simulation error minimization. *International journal of Control*, 76: 1767–1781, 2003.

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

J. Rissanen. Modelling by shortest data description. *Automatica*, 14:465–471, 1978.

J. Rissanen. Prediction minimum description length principles. *Annals of Statistics*, 14:1080–1100, 1986.

J. Roll, I. Lind, and L. Ljung. Connections between optimisation-based regressor selection and analysis of variance. In *The 45th IEEE Conference on Decision and Control*, pages 4907–4914, San Diego, CA, December 2006.

W. Spinelli, L. Piroddi, and M. Lovera. A two-stage algorithm for structure identification of polynomial narx models. In *Proceedings of the American Control Conference*, pages 2387 – 2392, Minneapolis, Minnesota, USA, jun 2006.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

M. Yuan and Y Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B—Statistical Methodology*, 68(1):49–67, 2006.

Avdelning, Institution
Division, Department

Division of Automatic Control
Department of Electrical Engineering

Datum
Date

2007-06-25

Språk
Language

☐ Svenska/Swedish
☒ Engelska/English
☐ ________________

Rapporttyp
Report category

☐ Licentiatavhandling
☐ Examensarbete
☐ C-uppsats
☐ D-uppsats
☒ Övrig rapport
☐ ________________

URL för elektronisk version

http://www.control.isy.liu.se

ISBN
—

ISRN
—

Serietitel och serienummer
Title of series, numbering

ISSN
1400-3902

LiTH-ISY-R-2799

Titel
Title

Regressor and Structure Selection in NARX Models Using a Structured ANOVA Approach

Författare
Author

Ingela Lind, Lennart Ljung

Sammanfattning
Abstract

Regressor selection can be viewed as the first step in the system identification process. The benefits of finding good regressors before estimating complex models are especially clear for nonlinear systems, where the class of possible models is huge. In this article, a structured way of using the tool Analysis of Variance (ANOVA) is presented and used for NARX model (nonlinear autoregressive model with exogenous input) identification with many candidate regressors.

Nyckelord
Keywords

identification, Structure identification, Analysis of variance