Regressor Selection with the Analysis of Variance Method

Ingela Lind

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

19th December 2001

Report No.: LiTH-ISY-R-2407
Submitted to IFAC 2002, Barcelona

Technical reports from the Automatic Control group in Linköping are available at http://www.control.isy.liu.se/publications.
Abstract

Identification of non-linear finite impulse response (N-FIR) models is studied. In particular the selection of model structure, i.e., to find the best regressors, is examined.

In this report it is shown that a statistical method, the analysis of variance, is a better alternative than exhaustive search among all possible regressors, in the identification of the structure of non-linear FIR-models. The method is evaluated for different conditions on the input signal to the system. The results will serve as a foundation for the extension of the ideas to non-linear autoregressive processes.

Keywords: Time delay estimation, Time lag, Identification, Non-linear models, Analysis of variance
REGRESSOR SELECTION WITH THE ANALYSIS OF VARIANCE METHOD

Ingela Lind

Division of Automatic Control, Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden. E-mail: ingela@isy.liu.se

Abstract: Identification of non-linear finite impulse response (N-FIR) models is studied. In particular the selection of model structure, i.e., to find the best regressors, is examined.

In this report it is shown that a statistical method, the analysis of variance, is a better alternative than exhaustive search among all possible regressors, in the identification of the structure of non-linear FIR-models. The method is evaluated for different conditions on the input signal to the system. The results will serve as a foundation for the extension of the ideas to non-linear autoregressive processes.

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

1. INTRODUCTION

Assume that a non-linear FIR model describes the measurements \( y_t \) from a system with input \( u_t \), that is,

\[
y_t = g(u_t, u_{t-T}, u_{t-2T}, ..., u_{t-kT}) + e_t.
\]

The value of \( k \) is unknown in addition to which time lags of \( u_t \) that contributes to the value of \( y_t \) and \( g \) is an unknown static non-linear function of up to \( k + 1 \) variables.

The problem faced in system identification is to find a good model for the system from input/output data. This process is threefold: First, proper regressors have to be found. Then, the function \( g \) need to be estimated, and at last the parameters should be estimated (Ljung, 1999). The third problem is the easiest one to handle. The second often needs many data and much computer power, due to the curse of dimensionality. The first problem is fundamental and not much studied in the system identification literature, in contrast to the case for linear systems. If proper regressors could be found without too much effort, the problem of estimating the function \( g \) would be much easier. In this report a common method from statistics, Analysis of Variance (ANOVA), will be applied to this problem in an approach, novel to the system identification area.

2. THE ANOVA IDEA

The statistical analysis method ANOVA is a widely spread tool for finding out which factors contribute to given measurements. Though common in medicine and quality control applications, it does not seem to have been tried in system identification applications.

The method is based on hypothesis tests with F-distributed test variables computed from the residual quadratic sum. Below, the fixed effects variant of the method is stated in a statistical framework for two factors. The complexity grows rapidly with the number of factors.
2.1 Two-way analysis of variance

Assume having a batch of \( a \cdot b \cdot n \) observations, corresponding to the \( ab \) treatment/level combinations of the factors \( A \) and \( B \). \( A \) has \( a \) different levels, \( B \) has \( b \) different levels and the experiment is repeated \( n \) times. The measurements are made in random order to be sure to avoid effects of time dependency etc.. The observations may be described by a linear statistical model,

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

where \( i = 1, \ldots, a \), \( j = 1, \ldots, b \), \( k = 1, \ldots, n \), \( \mu \) is the overall mean effect, \( \tau_i \) is the effect of the \( i \)th level of the factor \( A \), \( \beta_j \) is the effect of the \( j \)th level of the factor \( B \), \((\tau \beta)_{ij}\) is the effect of the interaction between the \( i \)th level of factor \( A \) and the \( j \)th level of the factor \( B \) and \( \epsilon_{ijk} \) is a random error component from a Gaussian distribution with constant variance. This means that deterministic effects from the level of the factors are assumed, and that the stochastic effects are totally due to measurement noise. The treatment effects are defined to be fixed deviations from the overall mean, so \( \sum_{i=1}^{a} \tau_i = 0 \), \( \sum_{j=1}^{b} \beta_j = 0 \), \( \sum_{i=1}^{a} (\tau \beta)_{ij} = 0 , \forall \ j \) and \( \sum_{j=1}^{b} (\tau \beta)_{ij} = 0 , \forall \ i \).

For example, to use ANOVA in a system identification application, consider the non-linear FIR-model

\[
y_t = g(u_t, u_{t-1}) + \epsilon_t.
\]

\(y_t\) is seen as the measurements, \(u_t\) as factor \( A\) and \(u_{t-1}\) as factor \( B\). This structure gives that \( a = b = m \), the number of levels \( u_t \) is allowed to assume.

The focus of interest is to test the following hypotheses regarding the treatment effects:

\[
H_{0AB} : (\tau \beta)_{ij} = 0 , \forall \ i,j,
\]

that is, the measurements do not depend on the level combination of factors \( A \) and \( B \) (interaction effect), against

\[
H_{1AB} : \text{at least one } (\tau \beta)_{ij} \neq 0,
\]

that there is at least one significant interaction effect. (The null hypothesis corresponds to the model \( y_t = g_1(u_{t-1}) + g_2(u_{t-2}) + \epsilon_t \). If the null hypothesis \( H_{0AB} \) is accepted, also the following tests are of interest:

\[
H_{0A} : \tau_1 = \tau_2 = \ldots = \tau_a = 0,
H_{0B} : \beta_1 = \beta_2 = \ldots = \beta_b = 0,
\]

against, respectively,

\[
H_{1A} : \text{at least one } \tau_i \neq 0,
H_{1B} : \text{at least one } \beta_j \neq 0.
\]

If any of the null hypotheses is rejected, it is assumed that the factor involved does have some effect on the measurements. (If the null hypotheses cannot be rejected, the appropriate model corresponds to: for \( H_{0A} \) that \( y_t = g_2(u_{t-2}) + \epsilon_t \), for \( H_{0B} \) that \( y_t = g_1(u_{t-1}) + \epsilon_t \), and for both \( H_{0A} \) and \( H_{0B} \) that \( y_t \) can not be explained by \( u_{t-1} \) and \( u_{t-2} \). )

To do the hypothesis tests a two-factor analysis of variance is used. The overall mean, the cell means and the means over the factor levels are computed. Let

\[
\bar{y}_... = \frac{1}{abn} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk},
\]

\[
\bar{y}_{i..} = \frac{1}{bn} \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk},
\]

\[
\bar{y}_{.j.} = \frac{1}{an} \sum_{i=1}^{a} \sum_{k=1}^{n} y_{ijk},
\]

\[
\bar{y}_{ij.} = \frac{1}{n} \sum_{k=1}^{n} y_{ijk}.
\]

The total residual quadratic sum can be written

\[
SS_T = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (y_{ijk} - \bar{y}_{...})^2
= SS_A + SS_B + SS_{AB} + SS_E,
\]

where

\[
SS_A = \sum_{i=1}^{a} bm(\bar{y}_{i..} - \bar{y}_{...})^2,
\]

\[
SS_B = \sum_{j=1}^{b} an(\bar{y}_{.j.} - \bar{y}_{...})^2,
\]

\[
SS_{AB} = \sum_{i=1}^{a} \sum_{j=1}^{b} n(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2
\]

and

\[
SS_E = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (y_{ijk} - \bar{y}_{ij.})^2.
\]

From a theorem by Cochran (Montgomery, 1991, page 59) it is possible to show that

- the stochastic variables \( SS_A, SS_B \) and \( SS_{AB} \) are independent,
- the stochastic variable \( \frac{1}{ab} \cdot SS_E \sim \chi^2(ab(n-1)) \),
- if \( \tau_1 = \ldots = \tau_a = 0 \), then \( \frac{1}{a} \cdot SS_A \sim \chi^2(a-1) \),
• if $\beta_1 = \ldots = \beta_0 = 0$, then $\frac{1}{\sigma^2} \cdot SS_B \sim \chi^2(b - 1)$,

• if $(\tau \beta)_{ij} = 0, \forall i, j$, then $\frac{1}{\sigma^2} \cdot SS_{AB} \sim \chi^2((a - 1)(b - 1))$.

These observations are used to design test variables to test the proposed hypotheses. The test variable associated with factor $A$ is chosen as

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

If $H_{0A}$ is true, then $v_A \sim F(a - 1, ab(n - 1))$, i.e., $v_A$ is $F$-distributed with $a - 1$ and $ab(n - 1)$ degrees of freedom. The null-hypothesis is rejected if a large value of $v_A$ is obtained, that is, reject $H_{0A}$ if $v_A > c$, where $c$ is taken from an $F(a - 1, ab(n - 1))$-table and $\alpha$ denotes the level of significance (the probability to reject $H_0$ though $H_0$ is true). $H_{0B}$ and $H_{0AB}$ are tested analogously. It is also possible to estimate the standard deviation, $\sigma$, associated with the random error component, as

$$\hat{\sigma}^2 = \frac{SS_E}{ab(n - 1)}.$$ 

The degrees of freedom are $ab(n - 1)$. Note that $n$ needs to be larger than 1 to do the analysis with all interaction effects.

The results from the hypotheses testing can be used to determine which factors have effect on the measurements and if there are interaction effects between different factors.

The most important modelling simplifications made are the assumptions that the variance is constant through the batch and that the random error component is Gaussian distributed. The F-tests are quite robust against violations against both assumptions (Krishnaiah, 1980, Chapter 7).

### 2.2 Unbalanced design

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 1 to 4, is lost. The F-tests should then be done in a different manner, see Miller (1997). Badly unbalanced designs, when the number of measurements in each cell differ a factor 10 or more are especially hard to handle.

### 2.3 Other statistical models

There are other types of linear statistical models that can be analysed with ANOVA, the random effects model and the mixed model. The random effects model assumes that each effect, $\tau_i$, $\beta_j$ and $(\tau \beta)_{ij}$, is a normally distributed random variable instead of a fixed value as in the fixed effects model. In the mixed model there are some variables that are considered as giving random effects and others that are considered as giving random effects.

The random effects model gives analysis results that are better suited to generalize, for example, to the entire range of a continuous variable (only a sample of the possible values can be treated in the analysis). There are some drawbacks with the model though. The most important drawback is that the tests for the interaction effects are very sensitive to non-normality of both the specific effects (the $\tau_i$s etc.) and the error component, $\epsilon_{ijk}$, when the null hypothesis is false (Miller, 1997). This could be problematic when identifying nonlinear systems, which would give non-normal effects. This problem is one of the reasons why the fixed effects model is used in this report.

### 3. SIMULATION RESULTS

#### 3.1 Experiment setup

**3.1.1. Fixed-level input** A carefully selected input signal for the identification experiment, with the purpose to find good regressors with ANOVA, would be a multi-level pseudo-random signal with three or more levels. The range of the signal should correspond to the working range for the wanted 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, see Godfrey (1993). The signal is repeated four times to give plenty of data for the analysis, that is, 256 input/output data with 4 in each cell.

**3.1.2. Random input** A fixed-level input signal is not practical in all situations. It is not very likely that full control over the input signal is achievable in all identification experiments. For that reason ANOVA will be evaluated for an uniformly distributed random input signal to be able to cope in those situations.

The input signal, $u_t$, is an independent, identically distributed random signal from the uniform distribution. It can assume values between -2.5 and 5.5, that is, close to range used in the earlier experiments.

Now, these simulated time series cannot be used directly. The input data range is divided into equal intervals, each assigned to one factor level for use in the ANOVA. Four factor levels for each time lag are used, which gives 64 cells in the experiment design, each cell corresponding to a unique combination of factor levels. The signal assignment introduces a new type of error in the ANOVA. The output $y_t$ can now be seen as
where $E[\text{cell}(u_t, u_{t-1}, u_{t-2})]$ is the expected value of $y_t$ in the cell which the input is assigned to, and

$$n_t = g(u_t, u_{t-1}, u_{t-2}) - E[\text{cell}(u_t, u_{t-1}, u_{t-2})].$$

The distribution of the new error term, $n_t$, depends on the function $g$, the distribution of the input $u_t$ and the number of levels used to categorize $u_t$. It is not necessarily equal in all cells, which violates the ANOVA assumption on equal variances in all cells.

In the Monte-Carlo simulations 800 input/output data are used for each run, trying to get an approximately balanced design. The number of data needed grows rapidly if more time lags are tested or if a finer grid for the input signal is used.

3.1.3. Correlated input Whenever the input signal $u_t$ is not a series of independent variables, the factors in the ANOVA become correlated. In many applications, ANOVA is used after a proper experiment design. Then care is taken that only the examined factors change during the experiment and in such a fashion that all cells in the design are covered by the observations in an equal manner. In those cases, the experiment design guarantees orthogonality of the factors. In identification applications, it would be nice to be able to use a wide range of different input data. So far, multi-level pseudo-random signals and uniformly distributed random signals are covered. These corresponds to the completely planned experiment and to the simplest choice of a persistently exciting input signal respectively.

When neither of these situations are applicable, for instance when there is no possibility to choose the input signal, it is necessary to know if the analysis get tarnished by correlated input signals. This analysis is meant to give an indication if extra care is needed when the input is correlated.

It is still necessary to get observations in all cells, so the input signal $u_t$ must be persistently exciting in the considered dimension. To get results comparable to the other input signals, $u_t$ would be preferred to assume values roughly between $-2.5$ and $5.5$, jump around in the state space given by the first three time lags and have a rather strong correlation between adjacent time lags. One choice that fulfills these criteria is

$$u_t = 0.95u_{t-1} - 0.9u_{t-2} + x_t - x_{t-1} + x_{t-2},$$

where $x_t$ is white noise uniformly distributed between $-4$ and $8$. If $1160$ samples from $u_t$ are taken, $300-350$ observations with values between $-2.5$ and $5.5$ for three adjacent samples, covering all cells, are obtained. The remaining observations, with values outside this range, are simply not considered, following the procedure stated for the additivity tests in Chen et al. (1995). The few observation series with empty cells are not analyzed. As before, the range between $-2.5$ and $5.5$ is divided into four intervals to give four factor levels, i.e., $64$ cells.

The number of useful observations is quite low, but the extra effort in each simulation, needed to increase it, did not seem worthwhile. This will probably give an unbalanced design.

3.1.4. Exhaustive search as analysis method The exhaustive search method is used to provide some comparison numbers, to get some perspective on how good or bad ANOVA is.

The idea behind the exhaustive search method to find the model structure is to enumerate all possible combinations of the input time lags and estimate a model for each such combination. When all the models are estimated, their prediction performance are compared with some good measure, e.g., the root mean square error between measured and predicted output. The one model that has the best performance is chosen as the model for the system. With some luck, this model has the same structure as the examined system. This method does not distinguish between the task of finding good regressors and the task of finding a model, thereby a lot of tuning is done to improve models before it is known if they are going to be used or not.

The models used to parameterize the system depend on the assumptions on how the system is working, i.e., if it is linear, non-linear, dynamic or static etc. There are no really efficient and completely reliable methods to estimate a non-linear model of a non-linear system. It is common to use neural networks to do it, but these suffer from the lack of search algorithms that guarantee that the global minimum is found. They also take lots of computations to find a minima at all. This means that it is not possible to be sure that the global minimum is found for each compared model, and thereby not that the best regressors are found.

In the simulation study, one input/output data set with length 256 with fixed-level input is used for each Monte-Carlo simulation. Then the analysis is made as follows: Divide the data set into estimation data and verification data. Construct neural networks for all possible combinations of time lags under the assumption that $k$ is at most $3$. For each such combination, construct networks with different numbers of parameters. The networks used have two layers with 5, 10 or 15 tansig neurons in
the first layer and one purelin neuron in the second layer. 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 minima. The minimization algorithm used is Levenberg-Marquardt (Matlab’s Neural Network Tool-box is used). Of all these estimated networks, chose the one with the smallest RMS on the verification data. This network gives the proposed regressors.

3.1.5. ANOVA as analysis method A three-factor fixed effects model ANOVA is used to find good regressors. In data columns two and four of Table 1 the implementation of ANOVA is made by the author in Matlab and in data columns five and six, Statistica (StatSoft, 1995) is used.

The demands on the method is somewhat higher than on the exhaustive search method. The main interaction pattern between the regressors is considered when deciding whether the method is successful in a simulation or not, which is not the case for the exhaustive search method.

3.2 Results from Monte-Carlo simulations

Results from several studies are lumped together in Table 1. The first study, columns one to four, compares exhaustive search with ANOVA, with different noise levels and different level of significance for ANOVA. The second study investigates what happens when a random input signal is used (column five). The third study is made to see if extra caution was necessary when a correlated input signal is used (column six).

The output signal $y_t$ was computed according to the equation

$$y_t = g(u_t, u_{t-1}, u_{t-2}) + e_t,$$

where the function $g(\cdot)$ is given in Table 1.

The simulated measurement error signal $e_t$ was 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.

From columns one to four the conclusion can be drawn, that the ANOVA method is much better at spotting what input time lags contribute to the output than the exhaustive search method. The results for function 1 shows that it is important to have high enough signal to noise ratio (SNR). If the SNR is increased by a factor 4 the theoretical probability of finding the correct model structure by ANOVA increases from 4% to 89%.

The difference in performance between the two methods becomes more profound when the functions have a more non-linear behaviour, e.g., exponential functions. This indicates that the used networks does not handle this kind of functions very well, which can be confirmed by looking at the RMS values on validation data.

The better performance for ANOVA in column four compared to column two 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. It can also be seen that the decrease in noise variance does not affect the performance for the exhaustive search method either, except for the function 1 (compare columns one and three).

The loss in performance when using a random input signal (column five) compared with the case with a fixed-level input signal (column two) is not as great as anticipated. The division into intervals is, after all, very rough. For function 1 the problem is still low signal to noise ratio for the time lag $u_{t-2}$. The results for function 14 are bad. In most of the simulations, the regressor $u_t$ is not included. The reason seems to be that the relatively small contribution from this regressor drowns in the noise $e_t$, which size and distribution varies a lot in the different cells.

Comparing the result in column six with the result in column five it can be seen that the frequency of correct answers is somewhat lower, but not considerably. This can be partially explained by the smaller amount of data used in the analysis. The analysis can be enhanced using a normal probability plot of the model residuals to check for normality and excluding factor levels which give extremely large within-cell standard deviations. These methods are used for functions 2, 3 and 14, but could prove useful also for function 4 and 11.

4. Conclusion

It can be concluded from experiments that the analysis of variance method is a good alternative to exhaustive search for identifying what input contribute to the output in nonlinear finite impulse response (N-FIR) 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 differences is that the minimization 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 faster in finding appropriate input time lags and a corresponding model. When the complexity increases, due to more possible input time lags, it is an even greater advantage to use the ANOVA method.

It is also possible to get good results from input/output data with a random input signal and a low level measurement noise. The extra noise term introduced by the division of the input into intervals, 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. The first problem can be counteracted by a finer interval grid in combination with more data and/or more control over the input signal with less variation around fixed input levels. The second problem is most pronounced when the functional relationship between input and output features discontinuities, e.g., function 4, or large changes of the derivate, e.g., function 2. This problem can be counteracted by excluding the cells with the largest within-cell standard deviation, e.g., the cells including the discontinuity. Functions with high interaction order and large differences between the size of the contributions from different time lags, can be analyzed erroneously. This should not pose a large problem, since the small contributions will not enhance the fit of a model by very much.

The conclusions from the study is that ANOVA together with a good choice of input signal is an efficient and reliable tool for finding proper regressors in system identification applications.

The ideas in this report are at the moment extended to non-linear autoregressive processes with and without exogenous variables.

References


