Model Order Selection of N-FIR Models by the Analysis of Variance Method

Ingela Lind

Department of Electrical Engineering
Linköping University, SE-581 83 Linköping, Sweden
WWW: <http://www.control.isy.liu.se>
Email: ingela@isy.liu.se

September 22, 1999

Report no.: LiTH-ISY-R-2210
Submitted to SYSID 2000

Technical reports from the Automatic Control group in Linköping are available by anonymous ftp at the address [ftp.control.isy.liu.se](ftp://ftp.control.isy.liu.se). This report is contained in the pdf file 2210.pdf.

Model Order Selection of N-FIR Models by the Analysis of Variance Method

Ingela Lind

September 22, 1999

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 contributing input time lags, is examined. A common method, exhaustive search among models with all possible combinations of the input time lags, has some undesired drawbacks, as a tendency that the minimization algorithm gets stuck in local minima and heavy computations. To avoid these drawbacks we need to know the model structure prior to identifying a model.

In this report we show that a statistical method, the multivariate analysis of variance, is a good alternative to exhaustive search in the identification of the structure of non-linear FIR-models.

We can reduce the risks of getting an erroneous model structure due to the non-convexity of the minimization problems, reduce the computation time needed and also get a good estimate of how far we can enhance the fit of the desired model.

Keywords: Model order selection, Identification, Non-linear FIR models, Analysis of variance

1 Problem Description

Assume that a non-linear finite impulse response model [7] describes the measurements y_t from a system with input u_t , that is,

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

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 is to identify from measurements the model order k , the contributing time lags u_{t+iT} and the function g .

A common method used to estimate the function g is to parameterize it as a non-linear neural network. If we want to get a small network, with few parameters to estimate, and information about what inputs contribute to the measurements, this method has some drawbacks:

Since the model order is not known in advance, it is necessary to assume a largest k and estimate a network of each combination of the time lags of u_t under that assumption. Of these models, the one with the best performance on

validation data is chosen. Each model takes time to estimate and depending on the assumption of k , there can be many models.

During the estimation of the neural network, the optimization algorithm can stop in a local minimum, which can lead to a model with wrong assumptions on relevant input time lags coming out best.

Both these drawbacks can be avoided if the structure of the model is known in advance, i.e., the contributing time lags and k are known. In order to achieve this knowledge, it is possible to use the Analysis of Variance method (ANOVA) from statistics [5, 3]. This way, we only need to estimate one model, thereby reducing the computation time considerably, and need not worry about getting an erroneous model structure because the optimization algorithm has been stuck in a local minimum.

2 Exhaustive Search for Finding Model Structure

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 goodness measure, e.g., the root mean square error between measured and predicted output. The model that has the best performance is chosen as the model for the system. If we are lucky, this model has the same structure as the system, whose structure we want to identify.

This method does not distinguish between the task of finding the model structure and the task of finding a model, thereby tuning is done to improve models before we know if they are going to be used or not.

The models used to parameterize the system depend on our assumptions on how the system is working, i.e., if it is linear, non-linear, dynamic or static etc.. If we had been identifying a linear system there would not have been any big problems with using the exhaustive search method to find the correct structure. The identification methods for finding good linear models are efficient and reliable. Of course, we could still do the wrong choices sometimes, due to the non-deterministic nature of the system. In this report we have been trying to identify the structure of non-linear systems though, and that is where the problems start. 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 guarantees that the global minimum is found. They also take lots of computations to find a minimum at all. So, we cannot be sure that the global minimum is found for each of the models we compare and can thereby not be sure that we have found the model structure that describes the system best.

3 The Analysis of Variance

The statistical analysis method ANOVA [5, 3] 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, one variant of the method is stated in a statistical framework for two factors. The complexity grows rapidly with the number of factors.

Assume that we have a batch of abn 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,

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}, \quad (2)$$

where $i = 1, \dots, a$, $j = 1, \dots, b$, $k = 1, \dots, n$, μ is the overall mean effect, τ_i is the effect of the i th level of the factor A , β_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 ϵ_{ijk} is a random error component. This means that we assume that there are deterministic effects from the level of the factors, 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 $\sum_{j=1}^b (\tau\beta)_{ij} = 0, \forall i$.

We are interested in testing the following hypotheses regarding the treatment effects:

$$H_{0AB} : (\tau\beta)_{ij} = 0, \forall i, j, \quad (3)$$

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

$$H_{1AB} : \text{at least one } (\tau\beta)_{ij} \neq 0, \quad (4)$$

that there is at least one significant interaction effect. If the null hypothesis is accepted, we are also interested in testing:

$$H_{0A} : \tau_1 = \tau_2 = \dots = \tau_a = 0, \quad (5)$$

$$H_{0B} : \beta_1 = \beta_2 = \dots = \beta_b = 0, \quad (6)$$

against, respectively,

$$H_{1A} : \text{at least one } \tau_i \neq 0, \quad (7)$$

$$H_{1B} : \text{at least one } \beta_j \neq 0. \quad (8)$$

If any of the null hypotheses is rejected, we assume that the factor involved does have some effect on the measurements. To do the hypothesis tests we use a two-factor analysis of variance. We compute the overall mean, the cell means and the means over the factor levels. Let

$$\bar{y}_{...} = \frac{1}{abn} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n y_{ijk}, \quad (9)$$

$$\bar{y}_{i..} = \frac{1}{bn} \sum_{j=1}^b \sum_{k=1}^n y_{ijk}, \quad (10)$$

$$\bar{y}_{.j} = \frac{1}{an} \sum_{i=1}^a \sum_{k=1}^n y_{ijk}, \quad (11)$$

$$\bar{y}_{ij.} = \frac{1}{n} \sum_{k=1}^n y_{ijk}. \quad (12)$$

The total residual quadratic sum can be written

$$\begin{aligned} SS_T &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (y_{ijk} - \bar{y}_{...})^2 \\ &= \sum_{i=1}^a bn(\bar{y}_{i..} - \bar{y}_{...})^2 + \sum_{j=1}^b an(\bar{y}_{.j.} - \bar{y}_{...})^2 \\ &\quad + \sum_{i=1}^a \sum_{j=1}^b n(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 + \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (y_{ijk} - \bar{y}_{ij.})^2 \\ &= SS_A + SS_B + SS_{AB} + SS_E, \end{aligned} \quad (13)$$

where

$$SS_A = \sum_{i=1}^a bn(\bar{y}_{i..} - \bar{y}_{...})^2, \quad (14)$$

$$SS_B = \sum_{j=1}^b an(\bar{y}_{.j.} - \bar{y}_{...})^2, \quad (15)$$

$$SS_{AB} = \sum_{i=1}^a \sum_{j=1}^b n(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 \quad (16)$$

and

$$SS_E = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (y_{ijk} - \bar{y}_{ij.})^2. \quad (17)$$

From a theorem by Cochran [5, page 59] it is possible to show that

- the stochastic variables SS_A , SS_B , SS_{AB} and SS_E are independent,
- the stochastic variable $\frac{1}{\sigma^2} \cdot SS_E \sim \chi^2(ab(n-1))$,
- if $\tau_1 = \dots = \tau_a = 0$, then $\frac{1}{\sigma^2} \cdot SS_A \sim \chi^2(a-1)$,
- if $\beta_1 = \dots = \beta_b = 0$, 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 interesting hypotheses. The test variable associated with factor A is chosen as

$$v_A = \frac{SS_A/(a-1)}{SS_E/(ab(n-1))}. \quad (18)$$

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 we get a

large value of v_A , that is, we reject H_{0A} if $v_A > c$, where c is taken from an $F_\alpha(a-1, ab(n-1))$ -table and α denotes the level of significance (the probability to reject H_0 though H_0 is true). H_{0B} and H_{0AB} are tested analogously. We can also estimate the standard deviation, σ , associated with the random error component, as $\hat{\sigma}^2 = \frac{SSE}{ab(n-1)}$. 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 [3, Chapter 7]. If the levels of the input cannot be regarded as fully deterministic or if the effects of the input are partly stochastic it is better to use another variant of the analysis of variance method [5, 6]. There are also ways to treat missing data, see [2].

Now, to use ANOVA in a system identification application, we can consider the non-linear FIR-model

$$y_t = g(u_t, u_{t-1}) + e_t. \quad (19)$$

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 we allow u_t to assume.

4 Experimental Design

We want to design an experiment to determine which time lags have effect on the output of a system, assumed to be described by a non-linear FIR model, $y_t = g(u_t, u_{t-1}, \dots, u_{t-k}) + e_t$, where y is the output, g is an unknown function of the input and e is additive Gaussian white noise with variance σ^2 . The result will both give us the important time lags and the main structure of the function g , e.g., with k assumed equal to 3,

$$g(u_t, u_{t-1}, u_{t-2}, u_{t-3}) = g_1(u_t, u_{t-2}) + g_2(u_t, u_{t-3}) + g_3(u_{t-1}). \quad (20)$$

To do identification experiments with the described variant of ANOVA we need to construct a signal u_t that contains all possible combinations of levels of the time lags we want to examine. Assume that we have p time lags we want to examine, m levels of u_t and need n repetitions of each combination. The most straightforward way is to enumerate all possible sequences of length p , with different level combinations of the time lags, repeat them n times, and sort them in random order. This gives us a sequence u_t of length pnm^p . In this longer sequence, the short sequences will occur more than n times each, and we use only every p th measurement of the output from the system.

5 Test Specification and Results

To compare the different methods, we test their performance on a list of different functions that fits into

$$y_t = g(u_t, u_{t-1}, u_{t-2}) + e_t, \quad (21)$$

see Table 1. That is, we have assumed that $k = 2$ and will perform a three-way analysis of variance ($p = 3$). Assume that we have two sets of input/output data for the system we want to identify. Also assume that the datasets are suited for analysis with ANOVA, that is, all possible level combinations of the input are included. We put the second data set aside for validation use. From these data sets we want to find what input time lags are relevant, and if possible, a good model for the system.

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. We vary the number of input time lags between one and three and get seven different combinations of possible time lags. This means there are 16 to 76 parameters to estimate, depending on the number of input time lags included. The minimization algorithm used is Levenberg-Marquardt [1].

The only performance criterion used in this test is the root mean square error (RMS) between measured output and simulated output from the model.

Our own implementation of the ANOVA test used in this experiment does not employ all useful features of the theory. It is possible to get the test more efficient and to gain more information by using it.

We will compare the two structure identification methods in the following way.

1. **Exhaustive search** Divide the first data set into estimation data and verification data. Construct neural networks for all possible combinations of time lags under the assumption of largest k . For each such combination, construct networks with different numbers of parameters (e.g. 5, 10 and 15 neurons in the first 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 minimum. Of all these estimated networks, choose the one with the smallest RMS on the verification data.
2. **ANOVA** Use all of the first data set and perform ANOVA. This gives a model (which input time lags contribute to measurements) and a model structure (the interaction pattern of the inputs). In this report, we do not use the information about the model structure to improve the estimation of the model, that is, only the information about the contributing input time lags is used. Construct neural networks with different numbers of neurons (5, 10, 15) for the chosen model. Estimate as in the exhaustive search and choose the network with the smallest RMS on the verification data.

Note that one benefit of the ANOVA test is that we get a good estimate of the noise variance more or less automatically. This information can be used to determine when the minimization algorithm has failed to find a minimum close to the global minimum. If the RMS is much larger than the ANOVA estimated standard deviation we know that we have got a bad model. Try to restart the estimation process more times with new random network parameters to find a better local minimum. If this does not lead to smaller RMS try to get more estimation data or change to another type of network. In the following test, this feature is not used.

Function	$\sigma = 1, \alpha = 0.01$		$\sigma = 0.0001, \alpha = 0.0001$	
	Exhaustive search	ANOVA	Exhaustive search	ANOVA
$u_t - 0.03u_{t-2}$	10	6	94	100
$\ln u_t + u_{t-1} + e^{u_{t-2}}$	77	100	78	100
$u_{t-1} \cdot [u_t + \frac{1}{u_{t-2}}]$	100	100	100	100
$\text{sgn}(u_{t-1})$	84	94	80	100
$\text{sgn}(u_{t-1}) \cdot u_{t-2}$	93	96	92	100
$\text{sgn}(u_{t-1}) \cdot u_t \cdot u_{t-2}$	100	100	100	100
$\ln u_{t-1} + u_{t-2} $	95	96	94	100
$\ln u_{t-1} \cdot u_{t-2} $	94	92	82	100
$u_{t-2} \cdot \ln u_{t-1} $	97	97	95	100
$u_{t-2}^3 \cdot \ln u_{t-1} $	50	95	56	100
$u_{t-2} \cdot (\ln u_{t-1})^3$	93	95	91	99
$ u_{t-2} \cdot e^{u_{t-1}}$	54	96	54	100
$u_{t-2} \cdot e^{u_{t-1}}$	49	94	49	100
$u_{t-2} \cdot e^{u_{t-1} - 0.03u_t}$	58	100	58	100
$ u_t $	83	96	73	100
network				
$g(u_{t-1}, u_{t-2})$	73	88	94	100
TOTAL	75.6	90.3	80.6	99.9

Table 1: Results from Monte Carlo simulations to test structure identification methods. Stated are the number of correctly chosen models out of 100. $n = 4$.

Last, we compare the network chosen by exhaustive search and the network chosen by ANOVA on the validation data.

We are interested in how often the two methods can pinpoint the correct input time lags, and for ANOVA, also the correct model structure, for a specific function. We test this by running 100 Monte Carlo simulations on a list of test functions, see Table 1. We also include an additional test function, a random network function, which is the same type of network as the ones we try to estimate the function g with. We vary the signal to noise ratio and the level of significance for the ANOVA method and let the number of repeated measurements, n , equal 4. The possible levels of u_t are -2, 1, 3 and 5, that is, $m = 4$, so we get datasets of length $n * m^p = 192$. The results are collected in Table 1. It is also possible to compute the theoretical probability to find the correct model structure by the ANOVA method [4].

From Table 1 we can draw the conclusion 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 the first function, $u_t - 0.03u_{t-2}$, shows that it is important to have not too small a 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 network we have been working with does not handle this kind of functions very well, which we can confirm by looking at the

RMS values for the validation data.

In Table 1, Column 5, the better performance for ANOVA as compared to Column 3 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. We can also see that the decrease in noise variance does not affect the performance for the exhaustive search method either, except for the first function.

6 Conclusions

We can conclude from experiments that the multivariate analysis of variance method is a good alternative to exhaustive search for identifying what input contribute to the output in non-linear finite impulse response (N-FIR) models.

The ANOVA method manages much better to identify the correct input signals (contributing time lags) and reduces the number of erroneous models from 19.4-24.4% to 0.1-9.7% when compared to the exhaustive search method. 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 effective. 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.

References

- [1] Howard Demuth and Mark Beale. *Neural Network Toolbox: Manual*. The MathWorks, Inc., Natick, Mass., 1998.
- [2] Ronald R. Hocking. *The Analysis of Linear Models*. Brooks/Cole, Monterey, 1984.
- [3] Paruchuri R. Krishnaiah, editor. *Handbook of Statistics*, volume 1. North-Holland, Amsterdam, 1980.
- [4] Ingela Lind. Use of the analysis of variance method for model order selection of N-FIR models. Technical Report LiTH-ISY-R-2177, Department of Electrical Engineering, Linköping University, Aug 1999.
- [5] Douglas C. Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, New York, 3rd edition, 1991.
- [6] Henry Scheffé. *The Analysis of Variance*. John Wiley & Sons, New York, 1959.
- [7] Jonas Sjöberg, Q Zhang, Lennart Ljung, A Benveniste, B Deylon, P-Y Glorennec, Håkan Hjalmarsson, and A Juditsky. Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, 31:1691–1724, Dec 1995.