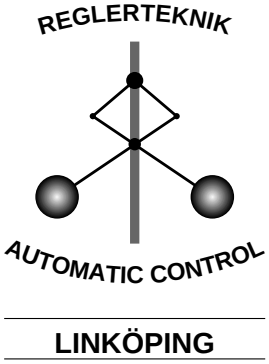
Linköping Studies in Science and Technology
Thesis No. 921

Regressor Selection in System Identification using ANOVA

Ingela Lind



Department of Electrical Engineering
Linköpings universitet, SE–581 83 Linköping, Sweden

Linköping 2001

Regressor Selection in System Identification using ANOVA

ingela@isy.liu.se
http://www.control.isy.liu.se
Department of Electrical Engineering
Linköpings universitet
SE–581 83 Linköping
Sweden

ISBN 91-7373-231-1 ISSN 0280-7971

Printed by UniTryck, Linköping, Sweden 2001

To Mattias and Elsa

Abstract

The aim of this work is to find a good method to select regressors for nonlinear system identification. A literature survey over possible methods to select the model structure for nonlinear systems, mainly autoregressive processes, is first given. The main ideas are:

1. Compare estimated models, using different regressor vectors, with each other.

2. Compare the variability of the output data, given one regressor vector, with the variability of the output data given other regressor vectors.

The second idea is investigated further by applying a common statistical tool, analysis of variance, to system identification applications. This method differs from most of the suggested methods by treating the variability in a stochastic framework, instead of treating the problem from a geometrical point of view. An investigation of the properties of analysis of variance (ANOVA), practical considerations with its use and Monte-Carlo simulations covering several aspects of the use of ANOVA in system identification applications is performed. It is shown that ANOVA is reliable, useful for different types of input signals and not critically sensitive to the amount of measurement noise. Moreover, the computations are fast, without iterations or minimisations. The result of this work is a suggested procedure for selecting a model structure from input/output data, using analysis of variance.

Acknowledgements

I would like to thank Prof. Lennart Ljung for all his advise and encouragement during the past three years. This thesis would not exist without him. Dr. Mikael Norrlöf, Jonas Elbornsson and David Lindgren have proofread this thesis, which has been a great help. Thank you.

All the people at the Division of Automatic Control are also important to me: for sharing thoughts, answering questions, stupid or not, and just being friendly. Ulla Salaneck, thank you for always being considerate and for your patience with all practical details.

This work has been supported by the Swedish Research Council(VR), the Swedish Research Council for Engineering Sciences(TFR) and the ECSEL graduate school in Linköping, which are gratefully acknowledged.

Finally, I would like to thank my family for their love and support. Especially Mattias, who share my life, and Elsa, who has brought more laughter into our home and shows the absolute joy of discovery and learning. I love you.

Linköping, 7 November, 2001

Ingela Lind

Contents

1

Introduction

The problem of system identification is to find a good model for a system from measured input/output data, without necessarily knowing anything about the physical laws controlling the system. A system can be any object we are interested in, physical or imaginary. For example, a water-tap, an industrial robot, the growing rate of a child or the happiness of a country's inhabitants. Output data are typically things that are important to us, such as the flow and temperature of the water, the movements of the robot arm, the length and weight of the child and, e.g., the suicide rate in a country. Things that affect the system are divided into two groups: inputs and disturbances. Inputs can be controlled, e.g., the flow of cold water and the flow of warm water, the voltages to the robot motors, the amount and contents of the food fed to the child and the quality of arts and sports activities available to people. Disturbances cannot be controlled, such as the hot water temperature, the load of the robot arm, how much the child moves and the interest in arts and sports activities among people. The questions of what affects the system and whether it is an input or not, are not always easy to answer. The example of the happiness of a country's inhabitants is one such system.

A model is a mathematical description of the system. It will never be complete, but good approximations are possible. The model can have several purposes: give greater insight of how the system works, give a foundation for how to control the system or give a possibility to tell if something new has happened in the system.

The process of finding a good model can be divided into five tasks:

Experiment design Decide what input signal(s) should be used in the identification experiment (Godfrey, 1993; Ljung, 1999) and what sampling interval to use. Signal range and frequency content should be considered as well as in what working points the model will be used. Good experiment design is necessary to get informative measurement data that can be used for estimation of useful models. In some systems the possible experiments are strictly limited, due to, e.g., safety constraints or cost.

Regressor selection Decide what regressors, i.e., which present and old inputs and old outputs, to use for explaining the output of the model. The regressor selection can be done completely guided by measurement data or in combination with knowledge gained from other sources, e.g., physical laws. If proper regressors are found, the tasks of choosing an appropriate model type and estimate the model parameters are much easier. For nonlinear systems, the regressor selection is not extensively studied in the system identification literature.

Model type selection Determine what function is suitable to describe the relation between the regressors and the output. There are several versatile model types available for both linear (see, e.g., Ljung (1999)) and nonlinear relations (Sjöberg et al., 1995). The flexibility of the model type has to be weighted against the amount of introduced parameters. Nonlinear model types tend to have a large number of parameters, even for few regressors, due to the curse of dimensionality. A large number of parameters makes it necessary to have a large amount of estimation data.

Parameter estimation The parameters associated with the chosen model type have to be estimated. This is done by minimising some criterion based on the difference between measurements and predictions from the model (e.g., Ljung (1999)). This is often the easiest task to handle, but time consuming.

Model validation The estimated model has to be validated to make certain that it is good enough for its proposed use. Prediction and simulation performance, model errors and stability are important to check. The input/output data used for estimation should not be reused for validation, but instead a new data set should be used (Ljung, 1999). The importance of the model validation cannot be overrated.

In this thesis, the focus is on regressor selection when the input/output data originates from a nonlinear system. The motivation to put some effort on regressor selection is to dramatically reduce the effort necessary to select a model type and estimate the associated parameters. If the regressors are not fixed beforehand, several models have to be tried to determine which regressor set that works best. For nonlinear model types, the parameters often takes considerable time to estimate, and it is not only regressors that have to be chosen. Also the degree of nonlinearity (or model order) and other structural issues need to be considered. All in all, the amount of tried models can grow very large.

Several existing regressor selection methods are presented in the thesis and a common method from statistics, Analysis of Variance (ANOVA), is applied to the regressor selection problem in an approach novel to the system identification area.

1.1 Thesis Outline

The thesis begins with a short introduction to system identification and some common models for nonlinear systems in Chapter 2. In Chapter 3, available methods for regressor selection are presented. In Chapter 4, the Analysis of Variance method is presented in some detail, and in Chapter 5, the practical aspects of the method are discussed. The theoretical chapters are then followed by several simulation experiments with different types of input signals, with the objective to understand the possibilities and limitations of the analysis of variance method for this problem. In Chapter 6 a multi-level pseudo-random signal is used, in Chapter 7 a random input signal and in Chapter 8 an input signal with autocorrelation is used to identify the structure of NFIR-models. In Chapter 9, several NARX-systems are investigated with ANOVA. Finally, the conclusions are given in Chapter 10.

2

System identification

In general, system identification is the method to obtain a mathematical description for an unknown dynamical system, see Figure 2.1, given a measured output signal y_t and the corresponding input signal u_t. To keep things nice and easy, the usual thing to do is to assume that a linear model can describe the relations between the input and output signals good enough. That is, y_t can be written

$$y_t = G(q)u_t + H(q)e_t, \tag{2.1}$$

with $G(q)$ and $H(q)$ the transfer functions from the input signal u_t and the disturbance e_t, see, e.g., Ljung (1999). One important subgroup of the

Figure 2.1: System with input signal u_t and output signal y_t.

linear model structures is the autoregressive (ARX) model structure,

$$\begin{aligned} y_t &= -a_1 y_{t-1} - a_2 y_{t-2} - \ldots - a_n y_{t-n} + \\ &\quad b_0 u_t + b_1 u_{t-1} + b_2 u_{t-2} + \ldots + b_m u_{t-m} + e_t. \end{aligned} \tag{2.2}$$

This model structure uses both old values of the output signal and the present and old values of the input signal to explain the present output. The noise model is $H(q) = 1/A(q)$, with $A(q) = 1 + a_1 q + \ldots + a_n q^n$. Since the noise, e_t, is assumed to be white, an estimator $\hat{y}_t(\theta)$ for y_t is obtained by omitting e_t in the formula above:

$$\begin{aligned} \hat{y}_t(\theta) &= -a_1 y_{t-1} - a_2 y_{t-2} - \ldots - a_n y_{t-n} + \\ &\quad b_0 u_t + b_1 u_{t-1} + b_2 u_{t-2} + \ldots + b_m u_{t-m}. \end{aligned} \tag{2.3}$$

The parameter vector θ contains the parameters $a_1, \ldots, a_n$ and $b_0, \ldots, b_m$, and is estimated by minimising a loss function, often formed by the sum of squares of the residuals between measured output data and estimated output from the model:

$$V(\theta) = \min_{\theta} \sum_{t=1}^{N} l(y_t - \hat{y}_t(\theta)). \tag{2.4}$$

This gives the maximum likelihood estimate $\hat{\theta}$ of θ, in case $l(x)$ is chosen as the negative logarithm of the probability density function of the disturbance e_t. For Gaussian disturbances this gives $l(x) = 1/2x^2$.

When a linear model structure is not sufficient to describe the relation between the input and output signals, nonlinear model structures are introduced. The class of nonlinear models is huge. The nonlinear autoregressive (NARX) model structure (Billings, 1980; Haber and Unbehauen, 1990; Mehra, 1979) is similar to the ARX-model above:

$$y_t = g(y_{t-1}, \ldots, y_{t-n}, u_t, \ldots, u_{t-m}, \theta) + e_t. \tag{2.5}$$

Also the nonlinear finite impulse response (NFIR) model structure,

$$y_t = g(u_t, \ldots, u_{t-m}, \theta) + e_t, \tag{2.6}$$

and the NAR model structure

$$y_t = g(y_{t-1}, \ldots, y_{t-n}, \theta) + e_t, \tag{2.7}$$

which are special cases of the NARX model structure, will be discussed. The estimation of the parameters is done by minimising the same loss function as for the linear case, with $\hat{y}_t(\theta)$ given by

$$\hat{y}_t(\theta) = g(y_{t-1}, \dots, y_{t-n}, u_t, \dots, u_{t-m}, \theta). \tag{2.8}$$

This gives a nonlinear least squares problem, which can be solved with the Gauss-Newton algorithm (Rao (1973): 'the method of scoring', Dennis and Schnabel (1983): 'damped Gauss-Newton'):

$$\hat{\theta}^{(i+1)} = \hat{\theta}^{(i)} - \mu^{(i)}[R^{(i)}]^{-1}V'(\hat{\theta}^{(i)}), \tag{2.9}$$

where $\hat{\theta}^{(i)}$ is the ith iterate and the norm is $l(x) = \frac{1}{2}x^2$. The step size $\mu^{(i)}$ is chosen such that

$$V(\hat{\theta}^{(i+1)}) < V(\hat{\theta}^{(i)}), \tag{2.10}$$

and the search direction is chosen as $R^{(i)} = H(\hat{\theta}^{(i)})$. $H(\theta)$ is a positive semidefinite approximation of the Hessian:

$$\begin{aligned} V''(\theta) &= \frac{1}{N}\sum_{t=1}^{N}\frac{\partial}{\partial\theta}\hat{y}_t(\theta)(\frac{\partial}{\partial\theta}\hat{y}_t(\theta))^T - \frac{1}{N}\sum_{t=1}^{N}\frac{\partial^2}{\partial\theta^2}\hat{y}_t(\theta)(y_t - \hat{y}_t(\theta)) \\ &\approx \frac{1}{N}\sum_{t=1}^{N}\frac{\partial}{\partial\theta}\hat{y}_t(\theta)(\frac{\partial}{\partial\theta}\hat{y}_t(\theta))^T = H(\theta). \end{aligned} \tag{2.11}$$

The minimisation procedure is more difficult than in the linear case since there may be many local minima, and many iterations are needed to find one of these, regardless of the parameterisation. The global minimum cannot be guaranteed to be found. If the global minimum is found and the noise e_t is assumed to be white Gaussian, $\hat{\theta}$ is the maximum likelihood estimator of the parameter vector. If the noise is not Gaussian, a change of norm l will give the maximum likelihood estimator.

The additional problem for nonlinear model structures is to select the function $g(\cdot)$. This problem can be divided into two parts by introducing the notation

$$g(y_{t-1}, \dots, y_{t-n}, u_t, \dots, u_{t-m}, \theta) = g(\varphi_t, \theta). \tag{2.12}$$

The first problem is now to determine the regressor vector φ_t and the second problem to choose a function g. The selection of the regressor vector φ_t is what this thesis is all about.

The options for the function g are quite many: artificial neural networks (Haykin, 1994; Kung, 1993), fuzzy models (Brown and Harris, 1994; Wang, 1994), hinging hyper planes (Breiman, 1993; Chua and Kang, 1977; Pucar and Sjöberg, 1998), local polynomial models (De Boor, 1978; Schumaker, 1981), kernel estimators (Nadaraya, 1964; Watson, 1969), nearest neighbour (Ljung, 1999) etc. Most of these methods can be described by the function expansion

$$g(\varphi_t, \theta) = \sum_k \alpha_k \kappa(\beta_k(\varphi_t - \gamma_k)), \tag{2.13}$$

where α_k, β_k and γ_k are parameters with suitable dimensions, and κ is a 'mother basis function'. For example, the Fourier series expansion has $\kappa(x) = \cos(x)$ with β_k corresponding to the frequencies and γ_k corresponding to the phases. For an overview of the possible choices, see Sjöberg et al. (1995). Two special model structures that will be used later, are the linear in the parameters NARMAX model structure and the radial basis neural network. The NARMAX model structure (Leontaritis and Billings, 1985) is function expansion of the form

$$y_t = F^l(y_{t-1}, \ldots, y_{t-n}, u_t, \ldots, u_{t-m}, e_t, \ldots, e_{t-p}) + e_t. \tag{2.14}$$

The nonlinear mapping F^l may include a variety of nonlinear terms, for example, terms raised to an integer power (e.g.,u_{t-1}^4), products of past inputs (e.g.,$u_t u_{t-4}$), past outputs (e.g.,$y_{t-1}u_{t-4}$) or cross terms (e.g.,$u_t y_{t-4}$). The radial basis neural network is a function expansion of the form

$$g(\varphi_t, \theta) = \sum_k \alpha_k r(\beta_k(\varphi_t - \gamma_k)), \tag{2.15}$$

where r is a radial function. The γ_k's are the centers of the radial function and the β_k's decide how large the support for the function is. Also the sigmoidal neural network, which is defined in the example below, is used later in this thesis.

The amount of parameters for the nonlinear case grows very rapidly with the number of regressors — the curse of dimensionality. Any scheme to reduce the number of parameters would be useful, since each parameter is estimated with an error. Additivity is one inner structure of the regressors which reduces the amount of parameters, see the example below.

Definition 2.1 (Additivity)
By additivity means that the nonlinear function g can be divided into ad-

ditive functions dependent on one regressor each:

$$\begin{aligned} y_t &= g(y_{t-1}, \ldots, y_{t-n}, u_t, \ldots, u_{t-m}, \theta) + e_t \\ &= g_1(y_{t-1}, \theta) + \ldots + g_n(y_{t-n}, \theta) + \\ &\quad g_{n+1}(u_t, \theta) + \ldots + g_{n+m+1}(u_{t-m}, \theta) + e_t. \end{aligned} \tag{2.16}$$

Definition 2.2 (Interaction)
In the cases where the division in (2.16) *is not possible, the regressors interact. The interaction can be of different order, e.g., two-factor interaction or four-factor interaction. Full interaction is when all regressors interact. In*

$$y_t = g_1(u_{t-1}) + g_2(u_{t-2}, u_{t-3}, u_{t-4}) + e_t, \tag{2.17}$$

there is a three-factor interaction between u_{t-2}, u_{t-3} *and* u_{t-4} *and in*

$$y_t = g_1(u_{t-1}, u_{t-4}) + g_2(u_{t-2}, u_{t-3}) + e_t, \tag{2.18}$$

there are two two-factor interactions between u_{t-1} *and* u_{t-4} *and between* u_{t-2} *and* u_{t-3}*. A simple function which gives (two-factor) interaction is*

$$y_t = u_t \cdot u_{t-1} + e_t. \tag{2.19}$$

Example 2.1 (Parameter reduction by use of inner structure.)
In this example, sigmoidal neural networks will be used. These are function expansions of the form

$$g(\varphi_t, \theta) = \beta + \sum_k LW_k \sigma(\varphi_t IW_k + b_k), \tag{2.20}$$

with the sigmoid function given by

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{2.21}$$

and IW_k *(Input Weight) and* LW_k *(Layer Weight) are weight matrices of appropriate dimensions. The regressor vector* φ_t *consist of three lagged inputs:*

$$\varphi_t = [u_t, u_{t-1}, u_{t-2}]^T. \tag{2.22}$$

We assume that these affect the output additively, that is, the function g *can be divided in the following manner:*

$$g(u_t, u_{t-1}, u_{t-2}) = g_1(u_t) + g_2(u_{t-1}) + g_3(u_{t-2}). \tag{2.23}$$

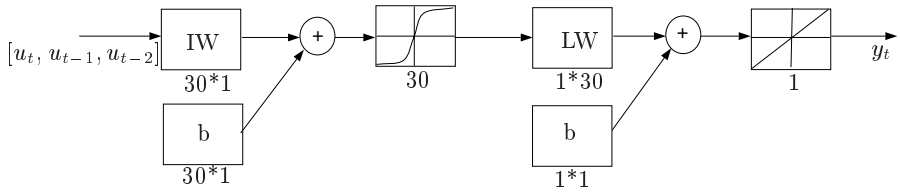


Figure 2.2: Neural network with 30 sigmoidal neurons and a full structure. The picture shows the signal flow for one neuron and the remaining 29 should be thought of as parallel, all summed in the last summation. The input to the net is $[u_t, u_{t-1}, u_{t-2}]^T$ and the output is y_t. IW stands for Input Weight matrix, b for bias matrix and LW for Layer Weight matrix. The dimensions of the matrices are given below the boxes. The total number of parameters is 151.

A full network with 30 sigmoidal neurons, as depicted in Figure 2.2, for three regressors use 151 parameters. If the inner structure is taken advantage of, as in Figure 2.3, only 91 parameters are used. The reduction is 40%. The number of parameters can probably be reduced further, since more neurons (for each dimension) should be needed to describe a multi-dimensional surface than a scalar function.

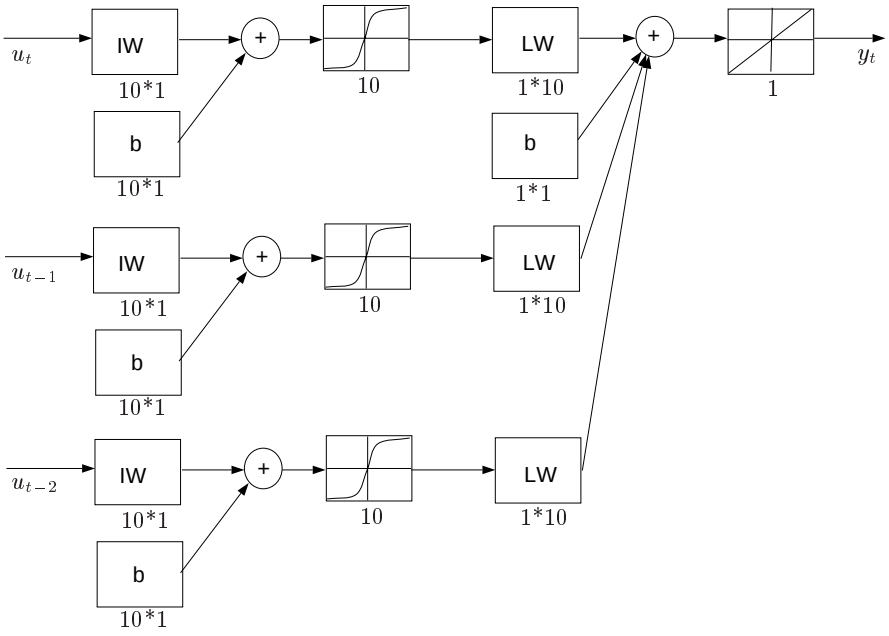


Figure 2.3: Neural network with 30 sigmoidal neurons and an additive structure, which gives 10 neurons for each input. The inputs to the net are u_t, u_{t-1} and u_{t-2} and the output is y_t. IW stands for Input Weight matrix, b for bias matrix and LW for Layer Weight matrix. The dimensions of the matrices are given below the boxes. The total number of parameters is 91.

3

Methods for finding significant regressors in nonlinear regression

Some existing methods to select the regressor vector φ in Equation (2.12) will be described briefly in this chapter. Many of the methods have their origin in the literature of statistics or chaos theory.

3.1 Non-parametric methods

3.1.1 Non-parametric FPE (final prediction error) and related methods

This method aims at minimising the expected squared prediction error for NAR processes, Equation (2.7), in a non-parametric setting,that is

$$FPE(\hat{g}) = E[(\tilde{y}_t - \hat{g}(\tilde{\varphi}_t))^2 w(\tilde{\varphi}_{M,t})]. \tag{3.1}$$

In Tschernig and Yang (2000) the tilde-denoted time series are independent, but identically distributed to y_t, φ_t. w is a continuous nonnegative weight function with compact support, mapping the vector $\tilde{\varphi}_{M,t}$ of length M to a real number. The noise e_t is allowed to have time-varying variance and may be non-white. $\hat{g}$ is a non-parametric estimate of the true function g.

A locally constant or locally linear estimate of g is obtained with w chosen as a kernel weight function. The FPE-criterion (introduced by Akaike) is computed for different choices of the regressor, φ_t, and the one with smallest FPE is chosen as indicator of the correct regressors to use in the model. Auestad and Tjøstheim (1990) give a heuristic justification, which is followed by a theoretical investigation in the companion articles Tjøstheim and Auestad (1994a) and Tjøstheim and Auestad (1994b). Tschernig and Yang (2000) prove consistency and make some improvements, for example, modifications to the local linear estimator to achieve faster computation. A Monte-Carlo study is made which confirms the theoretical reasoning.

Cross-validation

Cheng and Tong (1992), Yao and Tong (1994) and Vieu (1995) proposed an order selection method for smooth stationary autoregressive functions. The objective is to minimise the prediction error without including too many explanatory variables. The proposed method has much in common with the non-parametric FPE (above). The main difference is that the residual variance is computed by the 'leave-one-out' method in the cross-validation approach.

3.1.2 Local conditional mean and ANOVA

Truong (1993) investigated the convergence properties of local conditional mean and median estimators, which inspired Chen et al. (1995) to use Analysis of Variance (ANOVA, see Chapter 4) together with Truong's local conditional mean estimator to do additivity tests on autoregressive processes. The method is similar to the approach used in this thesis, but the application was limited to check if the function $g(X)$ can be divided into additive functions of one regressor each, which reduces the dimensionality of the function estimation process, see Example 2.1.

3.1.3 'Statistical approach'

Poncet and Moschytz (1994) suggest an optimal model order selection method, which in spirit is close to ANOVA. Their idea is to estimate the minimum mean squared prediction error realizable from data. An estimation theoretic argument claims that the prediction error lower-bound of order m (the length of φ), σ_m^2, is equal to the conditional variance,

$$\sigma_m^2 = Var(y|\varphi) = E[Var(y|\varphi = \mathbf{x})]. \tag{3.2}$$

To estimate the quantity in the right hand side, the local variance of the output signal y given regressor length m, several data points with exactly the same regressor $\mathbf{x}$ are needed. Since this is very rare in practice, some approximation is done. One could discretize the space by making a grid with size 2ϵ and centers $\mathbf{x}'_\mathbf{j}$, compute the estimates

$$Var(y|\|\varphi - \mathbf{x}'_\mathbf{j}\| \leq \epsilon) \tag{3.3}$$

and averaging them to get the estimate $\delta_m^2(\epsilon)$ of σ_m^2. Another possibility is to use pairs of data points, y and y', where $\|\varphi - \varphi'\| \leq \epsilon$. Since, under weak conditions,

$$\sigma_m^2 = E[\frac{1}{2}(y - y')^2|\varphi = \varphi'], \tag{3.4}$$

provided that y and y' are uncorrelated, the practical estimate of the conditional mean-square difference of order m,

$$\delta_m^2(\epsilon) = E[\frac{1}{2}(y - y')^2|\|\varphi - \varphi'\| \leq \epsilon] \tag{3.5}$$

is made. This estimate is more efficient with respect to the given data than the previous one, Equation (3.3), and has some monotonic properties which are useful for order selection. The minimum order m_0 is chosen such that the confidence intervals for the numeric estimate of (3.5) are approximately equal for all $m \geq m_0$, where ϵ should be chosen much smaller than the output signal variance.

3.1.4 False nearest neighbours

The false nearest neighbours method is based more on geometrical than stochastic reasoning. Kennel et al. (1992) introduced one version of the false nearest neighbours concept to find the correct embedding dimension, that is, the number of regressors, needed to give a reasonable description of a nonlinear time series. Their purpose was to find the smallest embedding dimension needed to recreate the dynamics of autonomous chaotic systems.

The idea is to compare the distance between two neighbouring explanatory regressors with the distance between their respective output observations. If the length of the regressor is sufficient, the distance between the observations should be short when the distance between the regressors is short, assuming that the nonlinear function is smooth. If the regressors are too short, the distance between the observations could be long even though

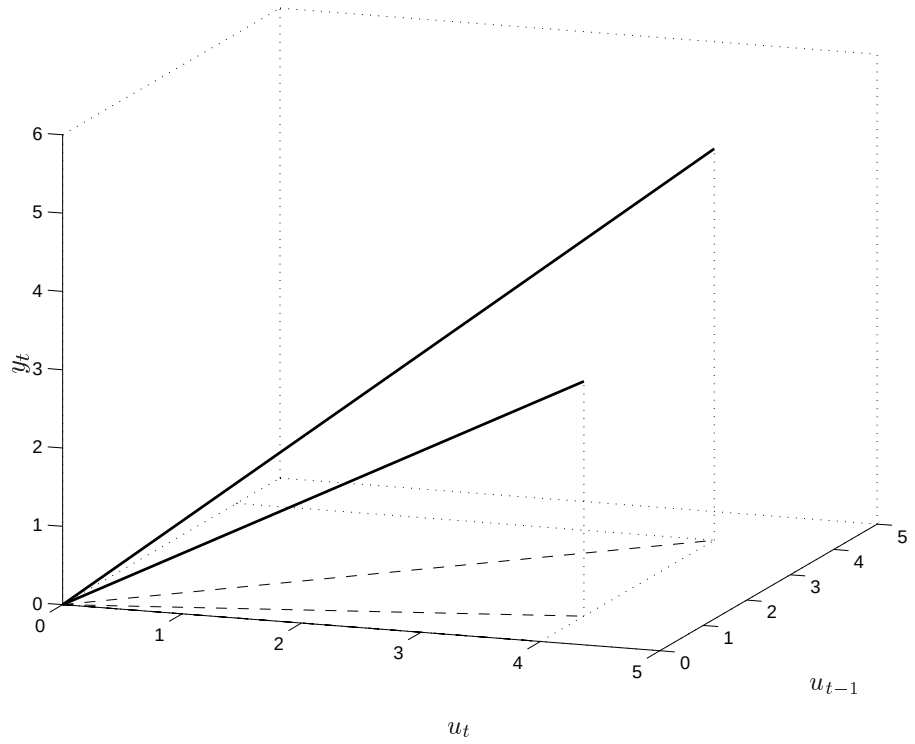


Figure 3.1: Two input/output data are plotted in this figure. Both u_t and u_{t-1} are needed to explain y_t. When the data are projected into the plane given by u_t and y_t, there is no explanation for the difference in output value, while the explanation is obvious in the higher dimension — false nearest neighbours.

the distance between the regressors is short, due to the projection of the regressor space, see figure 3.1. This is what is called a false nearest neighbour.

Kennel et al. (1992) propose a measure, which tells if two points are false nearest neighbours, and uses it to calculate the percentage of false nearest neighbours for each length of the regressor. When this percentage drops to nearly zero (or low enough), for growing dimensions, the correct embedding dimension is found.

Rhodes and Morari (1998) develops the idea to include systems with exogenous input and consider noise corrupted data in more detail.

3.1.5 Lipschitz numbers

He and Asada (1993) compare the distances between measurements with the distances between regressors. They compute the Lipschitz quotient,

$$q_{ij}^k = \frac{|y_i - y_j|}{\|\mathbf{x_i} - \mathbf{x_j}\|_2}, \ (i \neq j), \tag{3.6}$$

for each regressor length k and all data pairs (i, j). Then the geometrical average of the p largest quotients is formed. p is chosen to be approximately one percent of the number of data. For k less than optimal, this average decreases rapidly with growing k, but for k larger than optimal, the average is more or less constant. This is used to find the correct regressor length. The method is claimed to work also when a low level noise is present, that is, not only noise-free conditions, and is applicable to autoregressive models with exogenous input.

Bomberger (1997) compares the Lipschitz numbers method and the False Nearest Neighbours method on a few chemical engineering processes.

3.1.6 δ-test

Pi and Peterson (1994) worked with nonlinear autoregressive processes. They used the continuity of the nonlinear function to obtain a statistical measure of the dependence on each regressor, $y_{t-1}, \ldots, y_{t-n}$. The measure is the probability that two measurements are close when the regressor vector φ_t of length k are closer to each other than a specified value, δ. This probability depends on the amount of noise in the process and the presence of a functional dependence on the regressors. The achieved information can be used to determine the amount of noise in the process, the embedding dimension and which regressors contribute to the output.

3.1.7 Rank of linearised system

Autin et al. (1992) worked with input-output systems with measurement errors. They suggest to linearise the system around several operating points, by using a constant input signal to fix the operating point and add white noise to form the input signal. From an input-output data matrix, the order of the linearised system can be estimated by examination of the singular value decomposition, i.e., rank estimation. The size of the added input noise is important. It should be small enough to ensure a good approximation by a linear system and large enough to give a good signal to noise ratio in the presence of measurement errors, see (Autin et al., 1992).

3.1.8 The BRUTO algorithm

Chen and Tsay (1993) compare two methods to form non-parametric estimates of nonlinear additive autoregressive processes with exogenous signals, see Equation (2.16). One of the methods, the BRUTO algorithm of Hastie and Tibshirani (1990), can also be used to select the regressors that should be included in the model. The other method studied, alternating conditional expectation, is used in a fashion similar to the exhaustive search method described in this thesis, see Section 3.2.4. The BRUTO algorithm is an adaptive back-fitting procedure using ideas of cross-validation in selecting the smoothing parameter. The back-fitting idea is that, if the assumption of additivity is valid, then an estimate of $f_l(x_l)$ is given by a smoothed

$$y - \sum_{k \neq l} \hat{f}_k(x_k), \tag{3.7}$$

where x_k is the regressor that goes with function f_k. All $\hat{f}$'s are estimated in the same manner sequentially. The procedure is repeated until convergence of each function is achieved. Various smoothers with a smoothing parameter, λ, can be used, e.g., the Gaussian kernel smoother. For each $\hat{f}$, the λ-value minimising a global criterion is chosen as the candidate for the next iteration model. In each iteration of the BRUTO algorithm only one change of $\hat{f}$ is incorporated, namely the one which decreases the global criterion most. Also linear fits or no fit at all can be chosen by the λ values -1 and 0 respectively. Thereby regressors can be excluded from the model. For more details, see Chen and Tsay (1993).

3.1.9 Genetic algorithms

The paper by Gray et al. (1998) is not closely connected to the subject of finding the correct regressors. Their work focus on finding a proper model structure, including specific nonlinear functions and time delays. Their idea is to combine block libraries with a wide range of possible functions to chose from with a genetic programming algorithm. This is an interesting approach but beyond the scope of this thesis.

3.1.10 Mutual information

Zheng and Billings (1996) proposed the use of mutual information to select input nodes to radial basis networks, see Equation (2.15). This is applicable to determine what regressors should be used in system identification

applications. Mutual information is a fundamental information measure in information theory. It is a measure of the general dependence, including nonlinear dependence, between two variables, or, alternatively, a measure of the degree of predictability of the output knowing the input. The suggested algorithm aims at finding the subset of explanatory variables that maximises the mutual information. For details see Zheng and Billings (1996).

3.1.11 Coherence function of input-output data

Krishnaswami et al. (1995) suggest the use of the coherence function to find the significant regressors. Their version of the coherence function is defined as

$$\gamma_{yu} = \frac{|S_{yu}|^2}{S_{yy}S_{uu}}, \tag{3.8}$$

where S_{yu} is the cross-correlation between the signals y and u, and S_{yy} and S_{uu} are the autocorrelations of the signals. If γ_{yu} is small the dependence between the signals is weak and if it is large the dependence is strong. Also nonlinear functional dependencies are claimed to have large values of γ_{yu}. The procedure to find the significant regressors is then as follows:

1. Calculate the coherence function $\gamma_{y_t y_{t-i}}$ between y_t and all y_{t-i} up to $i = k$, for some large k. The computational effort is not high. Choose the $i = \alpha_1$ with the largest value of γ.

2. Use Gram-Schmidt orthogonalisation to orthogonalise y_t with respect to $y_{t-\alpha_1}$, giving $\tilde{y}_t$.

3. Then the coherence function $\gamma_{y_t y_{t-i}}$ should be computed for $\tilde{y}_t$ as in step 1, giving α_y. In the same manner the coherence function $\gamma_{y_t u_{t-i}}$ between $\tilde{y}_t$ and all u_{t-i} should be computed, giving α_u. Let $\alpha_2 = \max(\alpha_y, \alpha_u)$.

4. Orthogonalise $\tilde{y}_t$ with respect to both $y_{t-\alpha_2}$ and $u_{t-\alpha_2}$.

5. Repeat coherence function computation and orthogonalisation until $\tilde{y}_t$ approaches zero.

The order of the input is then the maximum α_u value and the order of the output is the maximum α_y value.

The suggested algorithm is quite sketchy with few explanations on the choices made. Some comparison with the Lipschitz quotient method is done on simulated data.

3.2 Parametric methods

3.2.1 Orthogonal structure detection routine

Korenberg et al. (1988) work with linear-in-the parameters NARMAX models, see Equation (2.14), where the model order is assumed to be known. They suggest a method to calculate the least squares problem for the parameter estimation, which gives the side effect that the contribution to the mean squared error for each parameter can be calculated. The suggested error reduction ratio can provide an indication of which terms to include in the model. Thresholds are needed to give useful result, but are not much discussed in their article. The noise is assumed to be zero-mean and white. In Billings et al. (1988) the method is extended to output-affine models without noise.

3.2.2 Stepwise regression

Billings and Voon (1986) have done successful modelling of nonlinear systems using NARMAX models in combination with stepwise regression to select significant terms. The idea is to include one term at a time, the one which contributes most, and deleting terms one at a time if they are found insignificant when more terms are added. This common method for selecting regressors in linear system identification is here adapted to NARMAX models. The method is known to have problems with convergence.

3.2.3 Bootstrap-based

A bootstrap-based method of reducing the number of parameters in the NARMAX models is suggested by Kukreja et al. (1999). It should be seen as an alternative to the method suggested in Section 3.2.1. They start with computing a parameter estimate with the extended least squares method. To get estimated confidence intervals on the parameters the following procedure is done.

1. The parameter estimate is used to compute the residuals from the linear regression.
2. The residuals are sampled with replacement to form new sets of data (the 'residuals' for the bootstrap data series).
3. The predicted output and the re-sampled residuals are used to form new 'measurements'. Each such data series gives a new parameter estimate, here called the bootstrap parameter estimate.

4. A confidence interval of the parameter estimate can then be formed using all the bootstrap parameter estimates.

If zero is contained in the confidence interval for a parameter, the parameter is considered as spurious.

The method is claimed to work well for moderately over-parameterised models. An important drawback in this context though, is that the maximum model order is considered known, that is, the maximum number of lagged inputs, the maximum number of lagged outputs, the maximum number of lagged errors and the maximum order on the polynomial expansion are considered as known.

3.2.4 Exhaustive search

The idea behind the exhaustive search method to find the model structure is to enumerate all possible combinations of the regressors and estimate a model for each such combination. There are mainly three ways to do the model selection:

Cross-validation The one model that has the best prediction performance on an untouched set of input/output data ('validation data') is chosen as the model for the system (Ljung, 1999). For example, the root mean square error (RMS) between measured and predicted output,

$$\mathrm{RMS} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2}, \tag{3.9}$$

can be used as comparison measure. This approach is used in this thesis.

AIC If the minimisation criterion, Equation (2.4), in the parameter estimation is replaced by

$$V_{AIC} = (1 + \frac{2\dim\,\theta}{N})V, \tag{3.10}$$

a proper choice among the estimated models can be done without having a validation data set (Akaike, 1981). The model with the lowest V_{AIC} should be chosen. Akaike's information criterion introduces in this way an extra penalty for the amount of parameters used in the model, in an attempt to avoid over-fit to the data.

Hypothesis tests If two models $\hat{g}_1(\varphi_t, \theta)$ and $\hat{g}_2(\psi_t, \eta)$, where ψ_t is a subset of φ_t, are compared, a hypothesis test can be used to determine if the difference in performance is significant (Ljung, 1999). The null hypothesis,

$$H_0: \text{ the data have been generated by } \hat{g}_2(\psi_t, \eta), \tag{3.11}$$

is tested against

$$H_1: \text{ the data have been generated by } \hat{g}_1(\varphi_t, \theta). \tag{3.12}$$

That means that we are prejudiced against the larger model. The test variable used is

$$N \cdot \frac{V(\hat{g}_2(\psi_t, \eta)) - V(\hat{g}_1(\varphi_t, \theta))}{V(\hat{g}_1(\varphi_t, \theta))}, \tag{3.13}$$

computed for estimation data, which is asymptotically χ^2-distributed with $(\dim \varphi - \dim \psi)$ degrees of freedom (at least for linear regressions and ARMAX models). If the value of the test variable is large enough, compared to a $\chi^2_\alpha(\dim \varphi - \dim \psi)$-table, the null hypothesis is rejected at the confidence level α.

If we are lucky, the chosen model has the same structure as the system we want to identify the structure of. This method does not distinguish between the task of finding the model structure and the task of finding a model, thereby a lot of tuning is done to improve models before we know if they are going to be used or not.

The models used to parameterise the system depend on our assumptions of the system structure, i.e., if it is linear, nonlinear, 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 noise effects in the signals. Here, the issue is to identify the structure of nonlinear systems, which gives more points to consider. There are no really efficient and completely reliable methods to estimate a nonlinear model of a nonlinear 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 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.

3.3 Comparison of methods

Most of the methods above belong to one of two main categories of methods. The first category can be called neighbour methods. These methods use the idea to compare distances between output values with distances between the corresponding regressor vectors of different length. Methods that belong to this category are: Local conditional mean and ANOVA, Section 3.1.2, 'Statistical approach', Section 3.1.3, False nearest neighbours, Section 3.1.4 and Lipschitz numbers, Section 3.1.5. A probably better alternative to many of these methods is to use ANOVA, which is a statistical tool for this kind of problem, see next chapter.

The second category can be called estimate and compare. Several different models are estimated and their performance compared. Methods that belong to this category are: The non-parametric FPE, Section 3.1.1, the BRUTO algorithm, Section 3.1.8, genetic algorithms, Section 3.1.9, orthogonal structure detection routine, Section 3.2.1, and stepwise regression, Section 3.2.2. A complete investigation of all possible models should be done if one wants to make certain that the correct regressors are found, as in Section 3.2.4.

4

The ANOVA idea

The statistical analysis method Analysis of Variance (ANOVA) (Miller, 1997; Montgomery, 1991) 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 has been discussed in the statistical literature since the 1940's.

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. But first a few words about sampling distributions.

One of the most important sampling distributions is the normal distribution. If z is a normal random variable, the probability distribution of z is

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma}}, \tag{4.1}$$

where μ is the finite mean of the distribution and $\sigma^2 > 0$ the variance. If $z_1, \ldots, z_u$ are normally and independently distributed random variables

with zero mean and variance 1, then the random variable

$$\chi^2(u) = \sum_{i=1}^{u} z_i^2 \tag{4.2}$$

follows the chi-square distribution with u degrees of freedom. If $\chi_1^2(u)$ and $\chi_2^2(v)$ are two independent chi-square random variables with u and v degrees of freedom, then the ratio

$$F(u,v) = \frac{\chi_1^2(u)/u}{\chi_2^2(v)/v} \tag{4.3}$$

follows the F-distribution with u nominator and v denominator degrees of freedom. The probability distribution of F is

$$h(F) = \frac{\Gamma(\frac{u+v}{2})(\frac{u}{v})^{u/2}F^{(u/2)-1}}{\Gamma(\frac{u}{2})\Gamma(\frac{v}{2})[(\frac{u}{v})F+1]^{(u+v)/2}} \quad 0 < F < \inf. \tag{4.4}$$

Also a non-central F-distribution is defined, as $F(u,v,\delta)$, where δ is a non-centrality parameter. If $\delta = 0$ the non-central F-distribution becomes the usual F-distribution.

4.1 Two-way analysis of variance

Assume that we have 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. To use ANOVA in a system identification application, we can consider, e.g., the nonlinear FIR-model

$$y_t = g(u_{t-i_1}, u_{t-i_2}) + e_t. \tag{4.5}$$

y_t is seen as the measurements, u_{t-i_1} as factor A and u_{t-i_2} as factor B. This structure gives that $a = b = m$, which is the limited number of levels we allow u_t to assume.

The observations may be described by a linear statistical model,

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}, \tag{4.6}$$

where $i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, n$, μ is the overall mean effect, τ_i is the effect of the ith level of the factor A, β_j is the effect of the jth

level of the factor B, $(\tau\beta)_{ij}$ is the effect of the interaction between the ith level of factor A and the jth level of the factor B and ϵ_{ijk} is a random error component from a Gaussian distribution with constant variance σ^2. 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 effects from A and B 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:

$$H_{0AB}: \ (\tau\beta)_{ij} = 0, \forall\ i, j, \tag{4.7}$$

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

$$H_{1AB}: \text{ at least one } (\tau\beta)_{ij} \neq 0, \tag{4.8}$$

that there is at least one significant interaction effect. (The null hypothesis corresponds to the model $y_t = g_1(u_{t-i_1}) + g_2(u_{t-i_2}) + e_t$.) If the null hypothesis H_{0AB} is accepted, we are also interested in testing:

$$H_{0A}: \ \tau_1 = \tau_2 = \ldots = \tau_a = 0, \tag{4.9}$$

$$H_{0B}: \ \beta_1 = \beta_2 = \ldots = \beta_b = 0, \tag{4.10}$$

against, respectively,

$$H_{1A}: \text{ at least one } \tau_i \neq 0, \tag{4.11}$$

$$H_{1B}: \text{ at least one } \beta_j \neq 0. \tag{4.12}$$

If any of the null hypotheses is rejected, we assume that the factor involved does have some effect on the measurements. (If the null hypotheses can not be rejected, the appropriate model corresponds to: for H_{0A} that $y_t = g_2(u_{t-i_2}) + e_t$, for H_{0B} that $y_t = g_1(u_{t-i_1}) + e_t$, and for both H_{0A} and H_{0B} that y_t can not be explained by u_{t-i_1} and u_{t-i_2}.)

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

$$\overline{y}_{...} = \frac{1}{abn} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk}, \tag{4.13}$$

$$\overline{y}_{i..} = \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}, \tag{4.14}$$

$$\overline{y}_{.j.} = \frac{1}{an}\sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk}, \tag{4.15}$$

$$\overline{y}_{ij.} = \frac{1}{n}\sum_{k=1}^{n} y_{ijk}. \tag{4.16}$$

The total residual quadratic sum can be written

$$\begin{aligned} SS_T &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{...})^2 \\ &= \sum_{i=1}^{a} bn(\overline{y}_{i..} - \overline{y}_{...})^2 + \sum_{j=1}^{b} an(\overline{y}_{.j.} - \overline{y}_{...})^2 \\ &\quad + \sum_{i=1}^{a}\sum_{j=1}^{b} n(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{ij.})^2 \\ &= SS_A + SS_B + SS_{AB} + SS_E, \end{aligned} \tag{4.17}$$

where

$$SS_A = \sum_{i=1}^{a} bn(\overline{y}_{i..} - \overline{y}_{...})^2, \tag{4.18}$$

$$SS_B = \sum_{j=1}^{b} an(\overline{y}_{.j.} - \overline{y}_{...})^2, \tag{4.19}$$

$$SS_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b} n(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2 \tag{4.20}$$

and

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{ij.})^2. \tag{4.21}$$

From a theorem by Cochran (Montgomery, 1991, 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 $\tau_1 = \ldots = \tau_a = 0$, then $\frac{1}{\sigma^2} \cdot SS_A \sim \chi^2(a-1)$,
- if $\beta_1 = \ldots = \beta_b = 0$, then $\frac{1}{\sigma^2} \cdot SS_B \sim \chi^2(b-1)$,
- if $(\tau\beta)_{ij} = 0, \forall i, j$, then $\frac{1}{\sigma^2} \cdot SS_{AB} \sim \chi^2((a-1)(b-1))$,

under the assumptions that the measurement noise is Gaussian with constant variance σ^2 and the design is balanced. The $\sim$ sign means 'distributed as'. A design is *balanced* if the number of data is equal in all cells, and *unbalanced* otherwise. These observations are used to design test variables to test the proposed hypotheses. The test variable associated with factor A is chosen as

$$v_A = \frac{SS_A/(a-1)}{SS_E/(ab(n-1))}. \tag{4.22}$$

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

4.1.1 Assumptions

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). To test if the assumption of normal distribution is valid, a normal probability plot of the residuals from the linear statistical model can be used: Order the N residuals ϵ_i in ascending order. Plot the ϵ_i's versus $\Phi^{-1}(i/(N+1))$ for $i = 1 \ldots N$, where $\Phi(\cdot)$ is the cumulative distribution function for the

		Factor A			
		Low		High	
Factor B	Low	$y_{111} = 2$	$y_{112} = 1$	$y_{211} = -2$	-1
		$y_{113} = 0$	$y_{114} = 1$	-1	0
	High	$y_{121} = -5$	-4	5	6
		-6	-3	4	4

Table 4.1: Measurement data divided into cells. There are 4 measurements for each combination of factor levels.

normal distribution with zero mean and variance 1. If the residuals belong to a normal distribution, the result is a straight line.

Example 4.1 (This example is meant to enlighten the formulas above.) *Consider the data in Table 4.1. These are collected from an experiment where both factor A and factor B have two levels, low and high. To compute SS_A, we need the column means $\bar{y}_{1..} = -14/8$ and $\bar{y}_{2..} = 15/8$ and the overall mean $\bar{y}_{...} = 1/16$. We get*

$$\begin{aligned} SS_A &= na\sum_{i=1}^{2}(\bar{y}_{i..} - \bar{y}_{...})^2 \\ &= 4 \cdot 2((\frac{-14}{8} - \frac{1}{16})^2 + (\frac{15}{8} - \frac{1}{16})^2) = 52.5625 \end{aligned} \tag{4.23}$$

SS_B is computed using the row means $\bar{y}_{.1.} = 0$ and $\bar{y}_{.2.} = 1/8$:

$$\begin{aligned} SS_B &= nb\sum_{j=1}^{2}(\bar{y}_{.j.} - \bar{y}_{...})^2 \\ &= 4 \cdot 2((0 - \frac{1}{16})^2 + (\frac{1}{8} - \frac{1}{16})^2) = 0.0625 \end{aligned} \tag{4.24}$$

For SS_{AB} also the cell means $\bar{y}_{11.} = 1$, $\bar{y}_{21.} = -1$, $\bar{y}_{12.} = -18/4$ and $\bar{y}_{22.} = 19/4$ are needed:

$$\begin{aligned} SS_{AB} &= n\sum_{i=1}^{2}\sum_{j=1}^{2}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 = 4((1 + \frac{14}{8} - 0 + \frac{1}{16})^2 \\ &+ (-1 - \frac{15}{8} - 0 + \frac{1}{16})^2 + (-\frac{18}{4} + \frac{14}{8} - \frac{1}{8} + \frac{1}{16})^2 \\ &+ (\frac{19}{4} - \frac{15}{8} - \frac{1}{8} + \frac{1}{16})^2) = 126.5625, \end{aligned} \tag{4.25}$$

Effect	Degrees of Freedom	Mean Square	F	p-level
A	*1	*52.56	*53.68	0.0000*
B	1	0.06	0.06	0.8048
AB	*1	*126.56	*129.26	0.0000*
Error	12	0.97		

Table 4.2: Analysis of Variance Table for the data in Table 4.1. The columns are from the left; the degrees of freedom associated with each sum of squares, the sum of squares divided by its degrees of freedom, the value of the F-distributed test variable associated with the corresponding null hypothesis and, finally, the probability level of the null hypothesis. The rows marked with stars denotes rejected null hypotheses at the significance level $\alpha = 0.01$.

and, finally, SS_E is computed as:

$$\begin{aligned} SS_E &= \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{4}(y_{ijk} - \bar{y}_{ij\cdot})^2 \\ &= (2-1)^2 + (1-1)^2 + \ldots + (4 - 19/4)^2 = 11.75. \end{aligned} \quad (4.26)$$

The sums of squares are customarily collected in an ANOVA table, see Table 4.2, in the form of mean squares, that is, each sum of squares is divided by its degrees of freedom. In the table are also the degrees of freedom, df, associated with each sum of squares, the value of the F-distributed test variable associated with the corresponding null hypothesis (for row 1 that is (4.22)*) and its probability level stated. The last column can be interpreted as the significance level α, which must be used in the hypothesis tests in order to accept the null hypothesis. The rows marked with stars correspond to rejected null hypotheses at the significance level $\alpha = 0.01$.*
The ANOVA table is read from the bottom. In this example it is clear that there are interaction effects between factor A and factor B since the p-level is close to zero. Normally the significance level α is 0.05 or 0.01 and any p-level lower than the chosen α will correspond to rejecting the null hypothesis. Since there are interaction effects present, the linear statistical model will not be simplified by further testing. Both factor A and factor B are considered as good regressors for these data. If the null hypothesis H_{0AB} had been accepted, it had been interesting to check the rows above to see if there had been significant main effects. The results from all the

hypothesis test is that we have selected a model structure for the system: the regressors A and B, and their interaction pattern, which is that they enter the model additively.

4.2 Random effects and mixed models

For sure, it is not always possible to view the factors A and B as giving fixed effects τ_i, β_j and $(\tau\beta)_{ij}$. One example is when the factors represent continuous variables instead of discrete variables. Another thing to consider is if the conclusions will be generalised to more than the analysed factor levels. In these cases the random effects model is the proper model to use.

The random effects model is

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}, \tag{4.27}$$

where $i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, n$, μ is the overall mean effect, and

$$\tau_i \text{ independent } N(0, \sigma_a^2), \tag{4.28}$$

$$\beta_j \text{ independent } N(0, \sigma_b^2), \tag{4.29}$$

$$(\tau\beta)_{ij} \text{ independent } N(0, \sigma_{ab}^2), \tag{4.30}$$

$$\epsilon_{ijk} \text{ independent } N(0, \sigma_e^2), \tag{4.31}$$

with independence between the different lettered variables. Here the interesting thing to do is to estimate the variance components σ_a^2, σ_b^2, σ_{ab}^2 and σ_e^2. The estimators for the variance components (balanced design) are based on the expected mean squares for the sums of squares, (4.18) - (4.21). The expectations are:

$$E[MS_A] = E[\frac{SS_A}{a-1}] = \sigma_e^2 + n\sigma_{ab}^2 + nb\sigma_a^2, \tag{4.32}$$

$$E[MS_B] = E[\frac{SS_B}{b-1}] = \sigma_e^2 + n\sigma_{ab}^2 + na\sigma_b^2, \tag{4.33}$$

$$E[MS_{AB}] = E[\frac{SS_{AB}}{(a-1)(b-1)}] = \sigma_e^2 + n\sigma_{ab}^2, \tag{4.34}$$

$$E[MS_E] = E[\frac{SS_E}{ab(n-1)}] = \sigma_e^2, \tag{4.35}$$

which gives the estimators:

$$\hat{\sigma}_a^2 = \frac{MS_A - MS_{AB}}{nb}, \tag{4.36}$$

$$\hat{\sigma}_b^2 = \frac{MS_B - MS_{AB}}{na}, \tag{4.37}$$

$$\hat{\sigma}_{ab}^2 = \frac{MS_{AB} - MS_E}{n}, \tag{4.38}$$

$$\hat{\sigma}_e^2 = MS_E. \tag{4.39}$$

For unbalanced designs, i.e., not all n_{ij} equal, the estimators are complicated, see (Searle, 1971, Chapters 10 and 11).

Here, we will not use the estimators directly. Instead we will test the null hypotheses $H_{0AB}: \sigma_{ab}^2 = 0$, $H_{0A}: \sigma_a^2 = 0$ and $H_{0B}: \sigma_b^2 = 0$ against the non-zero alternatives. We use the test variable $v_{AB} = MS_{AB}/MS_E$ which has a $F((a-1)(b-1), ab(n-1))$-distribution if H_{0AB} is true. For $v_{AB} > c_{AB}$, where c_{AB} is a critical limit taken from an $F_\alpha((a-1)(b-1), ab(n-1))$-table, the null hypothesis is rejected. $v_A = MS_A/MS_{AB}$ follows the $F(a-1, (a-1)(b-1))$-distribution if H_{0A} is true and for $v_B = MS_B/MS_{AB}$ the distribution is $F(b-1, (a-1)(b-1))$ for H_{0B} true. Note that the test statistics are not the same as for the fixed effects case. If the null hypotheses are false, the test statistics are still central F distributions.

If the design is poorly balanced these tests should not be used, since the independence between the different mean squares of sums of squares is lost. Any better tests for the unbalanced case does not seem to have been derived.

Mixed models

When some factors are treated as fixed and others are treated as random the model is called mixed. For a discussion on this type of model, see Miller (1997).

4.3 Significance and power of ANOVA

In order to decide the appropriate amount of measurements necessary to gain an acceptable performance of the hypothesis tests, we need to know how to calculate the power of the tests. There are two measures of performance often used:

$$\text{significance level} = \alpha = P(H_0 \text{ rejected}|H_0 \text{ true}) \tag{4.40}$$

$$\text{power} = 1 - \beta = P(H_0 \text{ rejected}|H_0 \text{ false}) \tag{4.41}$$

We want both α and β to be small. Do not confuse β with the effects β_j. We use the wanted α to calculate the critical limit for the test variable v, that is, we regard α as a design parameter. It is harder to get a value of β, since we need an assumption of in what way the null hypothesis is false and the distribution of the test variable according to this assumption. For the hypothesis test associated with factor A in the two-way fixed model analysis of variance we have that (Scheffé, 1959):

$$v_A = \frac{SS_A/(a-1)}{SS_E/ab(n-1)} \sim F(a-1, ab(n-1)) \text{ if } H_{0A} \text{ true} \tag{4.42}$$

and

$$v_A = \frac{SS_A/(a-1)}{SS_E/ab(n-1)} \sim \text{non-central } F(a-1, ab(n-1), \delta) \text{ if } H_{0A} \text{ false}, \tag{4.43}$$

where the two first parameters in the F-distribution are the degrees of freedom and the third, δ, is a non-centrality parameter with

$$\delta = na \sum_{i=1}^{a} \frac{\tau_i^2}{\sigma^2}, \tag{4.44}$$

which is closely related to the signal to noise ratio through τ_i. The formula for δ depends on how many factors are included in the test and which interaction effect is tested, see Krishnaiah (1980, p 201). The power of the test depends on the number of repetitions of the measurements, n, the number of levels of factor A, a, and the deviation from the null hypothesis we want to test. The power is different for the tests of main effects and for the tests of interaction effects of different orders.

Example 4.2 (Compute power of hypothesis tests) *As an example of how to compute the power of the hypothesis tests, we will describe how to compute the probability to find the correct model structure of a test function,*

$$y_t = u_t - 0.03u_{t-2} + e_t. \tag{4.45}$$

We are using a three-way analysis of variance and want to find what inputs have a significant effect on the output and if they interact. Input/output data from the function $y_t = u_t - 0.03u_{t-2} + e_t$ *are examined. We have 4 different levels that* u_t *can assume, and each measurement is repeated 4 times, that is,* $a = b = c = m = 4$ *and* $n = 4$*. The level of significance,* $1-\alpha$*, equals 0.99 in the test and the noise is Gaussian with standard deviation 1. The factor* A *is associated with* u_{t-2}*,* B *with* u_{t-1} *and* C *with* u_t*.*
To find the correct model structure we need to:

- *accept the null hypotheses for the interaction effects* ABC*,* AB*,* AC*,* BC *and the main effect* B*, and*
- *reject the null hypotheses for the effects* A *and* C*.*

The probability to accept the null hypothesis when it is true is given by $1-\alpha$ *and the probability to reject the null hypothesis when it is false is given by* $1-\beta$*. We neglect the fact that the different tests for the null hypotheses are not truly independent, due to the division by the estimated variance instead of the true variance in the test variables. We get an upper level (due to the neglected dependence) for the probability to find the correct model,*

$$P(\textit{find the correct model structure}) \leq (1-\alpha)^5(1-\beta_A)(1-\beta_C). \tag{4.46}$$

β_A *is given by* $\beta_A = P(v_A < c_A | H_{0A}\ false)$*, where* c_A *is the critical limit with confidence level* α *for the test variable* v_A*, which belongs to the distribution*

$$v_A \sim \textit{non-central } F(m-1, m^3(n-1), \delta_A) \tag{4.47}$$

with

$$\delta_A = nm^2 \sum_{i=1}^{m} \frac{\tau_i^2}{\sigma^2} \tag{4.48}$$

when H_{0A} *is false. To find the critical limit* c_A *we also need the distribution for* v_A *when* H_{0A} *is true,*

$$v_A \sim F(m-1, m^3(n-1)). \tag{4.49}$$

		A level					
		-2	1	3	5	means	γ_k
C level	-2	-1.94	-2.03	-2.09	-2.15	-2.0525	-3.75
	1	1.06	0.97	0.91	0.85	0.9475	-0.75
	3	3.06	2.97	2.91	2.85	2.9475	1.25
	5	5.06	4.97	4.91	4.85	4.9475	3.25
	means	1.81	1.72	1.66	1.6	$\mu = 1.6975$	
	τ_i	0.1125	0.0225	-0.0375	-0.975		

Table 4.3: Function values, means and effects, $u_t - 0.03u_{t-2}$.

See also Figures 4.1(a) and 4.1(b).
The value of β_C is computed analogously. The deterministic values for all factor combinations, mean values and factor effects are given in Table 4.3. It is easy to compute the effects for factor A, τ_i, $i = 1, \ldots, 4$ and for factor B, γ_k, $k = 1, \ldots, 4$. We get $\delta_A = 1.54$ and $\delta_C = 1713$, and use tables to find the corresponding values $\beta_A = 0.95$ and $\beta_C = 0$. The result is that the probability to find the correct model structure is 4.2%. We can also verify that

$$P(\textit{find only factor } C) \leq (1-\alpha)^6(1-\beta_C) = 0.94, \tag{4.50}$$

which means that we are very likely to assume that only factor C, u_t, explains the output from the function.

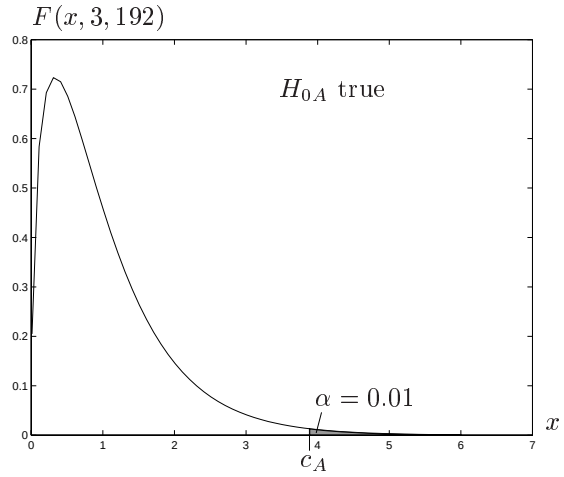
4.4 Unbalanced design

There are several reasons why unbalanced data (unequal cell counts) might occur. For example, a balanced design might have been planned, but for some reason observations have been lost. The orthogonality property of the main effects and interactions are lost when the data are unbalanced (Montgomery, 1991). This means that the usual ANOVA techniques do not apply.

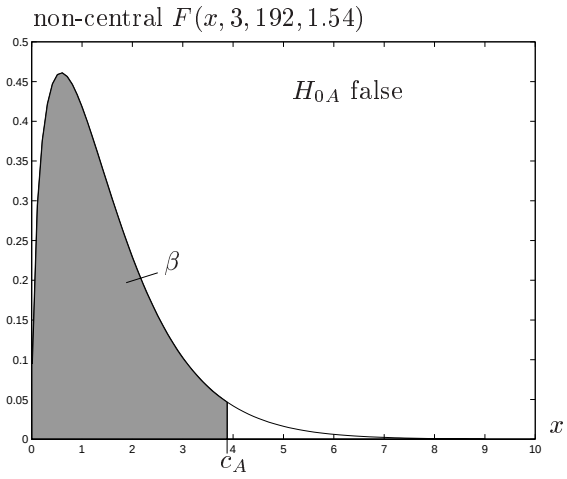
4.4.1 Proportional data

Only minor modifications are needed for proportional data. That is

$$n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}, \tag{4.51}$$



(a) Distribution for v_A when H_{0A} true.



(b) Distribution for v_A when H_{0A} false.

Figure 4.1: Distributions for v_A in Example 4.2. Observe that the x-axes have different scaling in the two plots.

where n_{ij} is the number of data in cell ij, $n_{i\cdot}$ is the number of data in the ith row, $n_{\cdot j}$ the number of data in the jth column and $n_{\cdot\cdot}$ the total number of data. Proportional data is not very likely to occur in our application, so the modifications of the ANOVA will not be described here.

4.4.2 Approximate methods

When the data are nearly balanced, some of the following methods could be used to force the data into balance. The analysis will then be only approximate. The analysis of balanced data is so easy that these methods are often used in practice (Miller, 1997; Montgomery, 1991). The analyst has to take care that the degree of approximation is not too great. If there are empty cells, the exact method has to be used.

Estimation of missing observations

If only a few data are missing, it is reasonable to estimate the missing values. The estimate that minimises SS_E is the cell mean. Treat the estimates as real data in the following analysis, but reduce the degrees of freedom for the error with the number of estimated data.

Discard data

If a few cells have more data than the others, estimating missing data would not be appropriate, since then many estimates would be used in the analysis. It would be better to set the excess data aside. The data that are set aside should be chosen at random. One alternative to completely discard excess data, could be to repeat the analysis with different data set aside (chosen at random).

Unweighted means

In this approach, which can be used without misleading result if the ratios of sample sizes do not exceed 3 (Rankin, 1974), the error sum of squares is computed as:

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij\cdot})^2. \tag{4.52}$$

In the rest of the analysis the cell means $\bar{y}_{ij\cdot}$ are treated as if they were all the averages of n^* data, where n^* is the harmonic mean of the sample sizes,

$$n^* = (\frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{1}{n_{ij}})^{-1}. \tag{4.53}$$

The cell means are used instead of the data in the computation of the other sums of squares. The degrees of freedom for SS_E are adjusted to $n_{\cdot\cdot} - ab$ instead of $ab(n-1)$, which is the degree of freedom in the balanced case. Here $n_{\cdot\cdot}$ is the total amount of data and n is the number of data in each cell in the balanced case.

The advantage of the method is its computational simplicity.

4.4.3 Exact method

If there are empty cells or if the ratios between cell sample sizes are large, an exact method has to be used. The prudent analysis is to resort to multiple regression,

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}, \tag{4.54}$$

where the regression matrix $\mathbf{X}$ should be constructed of 1's, 0's and -1's to insert or leave out the appropriate parameters and satisfy the constraints by expressing some parameters as negative sums of others. The parameter sets are not orthogonal, so the analysis is more complicated than for the balanced case. It makes a difference in what order the hypothesis tests are made and the interpretation is not simple. For details regarding the exact method, see Hocking (1984), since the algebra involved is extended.

5

Practical considerations with the use of ANOVA

5.1 Which variant of ANOVA should be used?

If a variable should be seen as fixed or random depends on what conclusions are to be drawn from the analysis. If the conclusions only concern the factor levels studied in the analysis, the variable associated with the factor can be seen as fixed. If the conclusions are to be generalised to more levels of the factor (e.g., a continuous variable), the variable should be interpreted as random.

In system identification, input signals should sometimes be viewed as continuous. This should call for a random effects model. When using a random number generator, though, it is hard to get a signal that can be divided into intervals such that all cells in an experiment design are covered by equally many data. The unbalanced design gives some drawbacks when using the random effects model, drawbacks that are poorly investigated (Miller, 1997). To give a better motivation on why we instead choose to work with the fixed effects model, some pros and cons for the different models are listed below.

Fixed effects model

Pros

- The F-tests are relatively simple to derive and use.
- Non-normality of the error component does not have a considerable effect on the significance level of the F-tests.
- Unbalanced designs can be treated quite well, but they are sensitive to outliers.

Cons

- It is not possible to use any information on how the null hypotheses might be false to enhance the tests.
- Non-normality can reduce the power of the tests.

Random effects model

Pros

- The results from the analysis are suited for generalisation over the entire range of the continuous variable.
- Power calculations are easier to perform for the random effects model, since the test variables always follow a central F-distribution, even when the null hypothesis is false.

Cons

- It is hard to handle unbalanced designs. Miller (1997) calls it a 'horror story'.
- No work has been done on what happens if the measured data are serially correlated, which is precisely the kind of data we would like to apply ANOVA to. This means that we have absolutely no idea if we can or cannot trust the results from an analysis of real measurement (not simulated) data, since there probably is going to be serial correlations in the data we can not build into our proposed model.
- If the null hypothesis is false, the tests for the interaction effects are very sensitive to non-normality of both the specific effects (the τ_i's etc.) and the error component ϵ_{ijk}. If we have lots of data the tests for

the main effects might still be robust. When applying the analysis to nonlinear systems, we can surely count on getting non-normal effects. This will probably be a problem.

How easy or hard it is to implement the tests might be beside the point here, because there are lots of commercial software that can perform all kinds of ANOVA imaginable. It is more important to know if and why the results are to be trusted.

For the purpose to find out what regressors of a signal contributes to the level of another signal it should be enough to have results applicable to some points in the regressor space. See, e.g., Autin et al. (1992), who uses this as an argument to work with linearised systems to find the proper time lags to use in the model of nonlinear systems. We need to be extremely unlucky to pick out only such points, that the effect from one regressor is not visible in any of them. If we choose an input signal that cover the range of the input we would like to get a good model for, it is unlikely that we would miss time lags that gives significant effect with this scheme.

These are the reasons behind the choice to work with the fixed effects model.

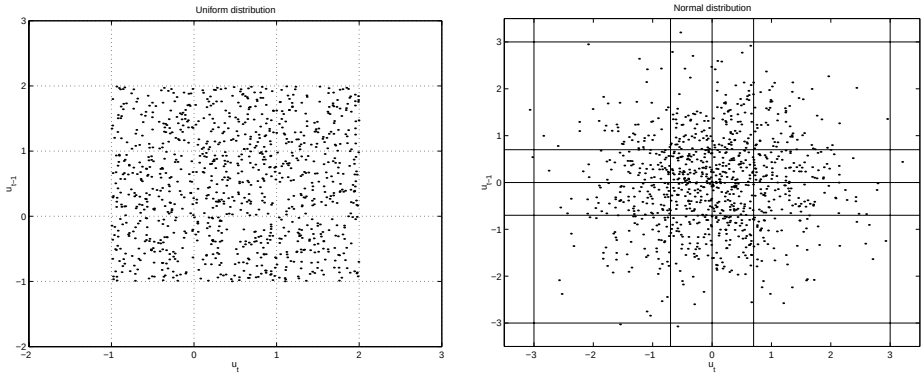
5.2 Division into levels

One of the largest practical problems one encounters with the use of ANOVA is that the explanatory variables have to be grouped into levels. How should that be done? We will examine the grouping of some different input signals, u_t, and also discuss the grouping of the measurements, y_t, when working with NARX-models.

5.2.1 Fixed levels input signal

A pseudo-random multi-level signal is a typical choice of a fixed levels input signal. u_t can assume a number of different values (e.g. -2, 1, 3 and 5 as used in Chapter 6) and does so in a more or less random order. Multi-level shift registers can be used to construct such a signal, see Godfrey (1993).

The natural grouping into factor levels for the ANOVA is to associate each level of the input signal with one group. For NFIR-models this gives ideal conditions for using the fixed effects model ANOVA.



(a) Uniform distribution. u_t is plotted against u_{t-1} for 1000 data.

(b) Normal distribution. u_t plotted against u_{t-1} for 1000 data.

Figure 5.1: Examples on groupings of random input signals. The lines give limits for the different groups.

5.2.2 Random input signal

Here, a random input signal denotes a series of independent identically distributed random variables.

By dividing the range of a uniformly distributed random signal into three or four equal intervals, each associated with one group, it is possible to get an approximately even cell count for the ANOVA design, see Figure 5.1(a).

A normally distributed random signal needs to be divided into unequal intervals, shorter ones close to the mean value, to give approximately equal cell counts, see Figure 5.1(b). Note that the cells in the corners get fewer data than the cells in the middle.

5.2.3 Correlated input signal

With autocorrelated input signals some problems are encountered. Most divisions of the original input signal range lead to unequal cell counts, often varying a factor 10 or more, see Figure 5.2. How large this effect is depends on how large the correlation between the different time lags is. The unbalanced design affects the analysis badly.

One solution to the problem is to work with a shrunken range (Chen and Tsay, 1993), that is, only consider data inside a closed cube of the same dimension as the number of factors analysed, see Figure 5.3. The numbers

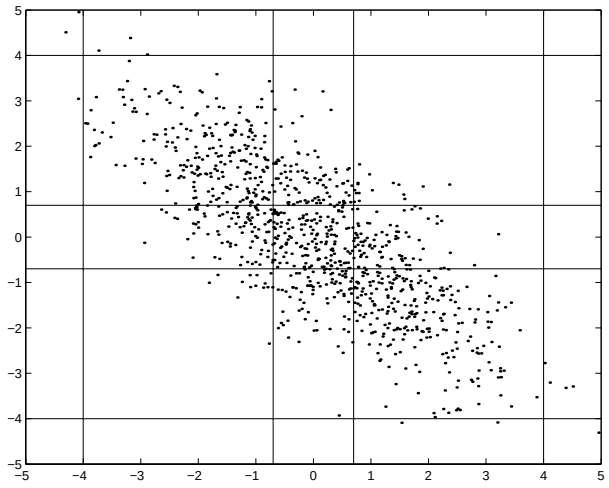


Figure 5.2: The entire range for a correlated signal divided into intervals with almost equal numbers of data. The grouping still leads to unequal cell counts. u_t is plotted against u_{t-1} for 1000 data.

of data in each cell are still not equal, but the unbalance is not as severe as before. The intervals, (a_i, a_{i+1}) for $i = 0, \ldots, (m-1)$, are constructed as follows:

$$a_i = u_{min} + (1-\delta)(y_{max} - y_{min}) + i\delta(y_{max} - y_{min})/m, \qquad (5.1)$$

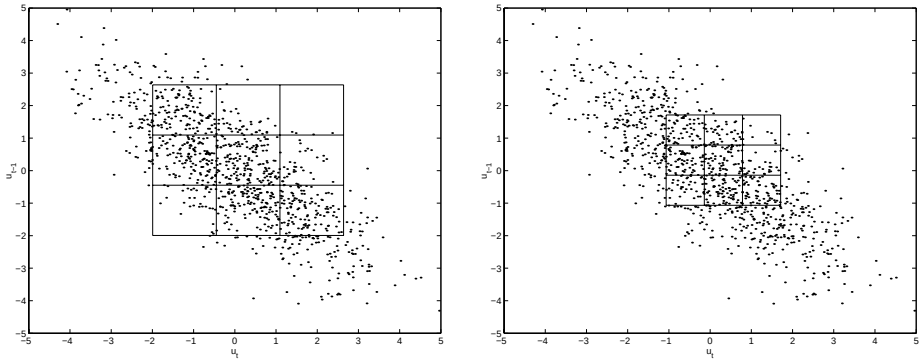
where $\delta \in (0, 1)$ is the shrinking factor. This partitions the shrunken range $\delta(y_{min}, y_{max})$ into m equal intervals.

The method with the shrunken range might consume large numbers of data depending on how large the autocorrelation of the signal is. If a choice of input signal is possible, heavily correlated ones are not to be preferred.

The problem is the same with correlations between time lags (autocorrelation) in one signal or with correlation between different input signals if there are more than one.

5.2.4 Autoregressive processes

The uneven cell counts are even more pronounced when autoregressive processes, see Equation (2.5), are analysed. In many cases it is hard to get data that is informative enough, especially if the signal-to-noise ratio is high.



(a) Shrinking factor $\delta = 0.5$. u_t plotted against u_{t-1} for 1000 data.

(b) Shrinking factor $\delta = 0.3$. u_t plotted against u_{t-1} for 1000 data.

Figure 5.3: In these plots the data from Figure 5.2 is divided according to the shrunken range method.

Then the signal does not jump around enough to cover all cells, depending on what time lags are chosen for analysis, see Figure 5.4.

5.2.5 Discard data

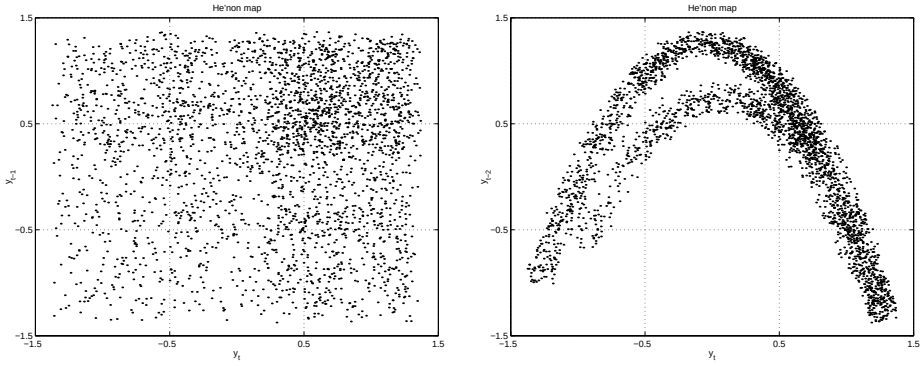
One way to get around the problems with uneven cell counts could be to pick out just a few data in each cell. Then it would be possible to choose the number of data in the cells such that the design gets balanced. If the data included in the analysis are picked at random from all the data in the same cell, this procedure would get us close to the experiment design usually associated with ANOVA. The procedure would be as follows:

1. Choose what regressors should be included in the analysis, i.e., considered for inclusion in the process model.

2. Divide the data into cells, that is, choose what grouping is appropriate for the regressors.

3. Let the smallest cell count $n_{min} = n$.

4. Pick n data at random from each cell and perform the ANOVA on these data.



(a) Hénon map (definition in Section 9.1.8). y_t plotted against y_{t-1} for 3000 data.

(b) Hénon map. y_t plotted against y_{t-2} for 3000 data.

Figure 5.4: This is an example of an autoregressive process with strong correlation between the time lags y_t and y_{t-2}, but small correlation between y_t and y_{t-1}. It is hard to make a grid with more than two intervals for each regressor without getting any empty cells.

5.3 How many regressors can be tested?

The answer to this question depends partly on the available data. For regressors without correlation, more regressors can be tested at the same time than for regressors with correlation. The reason is that the data reflects the correlation and makes it impossible to obtain a grouping without empty cells for regressors with strong correlation. From a statistical point of view, this limits the possible tests, since all possible regressors should be included in the test at the same time. It seems that this restriction is not strictly necessary to get useful insights of the structure of the data. A feasible way to extract as much information as possible from available data is as follows:

Group the data. Try to get a grouping which enables as many regressors as possible to enter the test at the same time. If nonlinear systems are considered, at least three groups for each regressor are needed to cover most types of nonlinearities. Some knowledge about the system is needed to select the best groups. Perform a test with as many possible regressors as can be done. Start with the ones assumed most likely to explain the output signal or use a systematic scheme. The result will be that some regressors show significant effect and others do not. Discard the regressors that did

not show any effect on the output and keep the ones that did. Now the discarded regressors can be replaced by yet untested ones and a new test performed. The same procedure can be repeated until all possible regressors are tested or until all the regressors in the same test show significant effect.

If the possible regressors are uncorrelated, it should be feasible to restart the testing with only untested regressors in the test and keep on until all possible regressors are covered by tests. The regressors with significant effects from earlier tests are of course kept in mind. Interaction effect between regressors tested in different tests can unfortunately not be considered in this scheme. Another drawback is that if the regressors are correlated, spurious ones can show significant effect if they are correlated with a contributing regressor tested in a separate test.

Conclusion

The conclusion is that at least main effects from a large number of possible regressors can be tested, and interaction effects to a limited extent. The results can include spurious regressors and significant interaction effects of higher order can be missed.

Example 5.1 (Testing procedure) *Consider data from a system with the following structure:*

$$y_t = g_1(u_t, u_{t-6}) + g_2(u_{t-1}) + g_3(u_{t-5}) + e_t. \tag{5.2}$$

Suppose that a grouping with three groups, allowing test with three regressors at a time has been done. Assume that the regressors u_t, u_{t-1} *to* u_{t-10} *are considered as possible regressors, that is, the model*

$$y_t = g(u_t, u_{t-1}, \dots, u_{t-10}) + e_t \tag{5.3}$$

is considered. In test 1, u_t, u_{t-1} *and* u_{t-2} *are included. No interaction effects are found and* u_t *and* u_{t-1} *show significant main effects. For test 2,* u_t *and* u_{t-1} *are kept and* u_{t-2} *discarded. In test 2,* u_t, u_{t-1} *and* u_{t-3} *are included. As before, only main effects from* u_t *and* u_{t-1} *are significant, so* u_{t-3} *is discarded. In test 3,* u_t, u_{t-1} *and* u_{t-4} *are included. The result is that* u_{t-4} *is discarded. In test 4,* u_t, u_{t-1} *and* u_{t-5} *are included. All show significant main effects, but no interaction effects are found. Since it is no longer possible to discard regressors, next test includes only yet untested regressors.* u_{t-6}, u_{t-7} *and* u_{t-8} *are included in test 5. Only* u_{t-6} *show a significant main effect, so* u_{t-7} *and* u_{t-8} *are discarded. In test 6,* u_{t-6}, u_{t-9}

and u_{t-10} are included, with the result that only u_{t-6} is kept. This means that the tested model is

$$\begin{aligned} y_t &= g_1(u_t, u_{t-1}, u_{t-2}) + g_2(u_t, u_{t-1}, u_{t-3}) \\ &\quad + g_3(u_t, u_{t-1}, u_{t-4}) + g_4(u_t, u_{t-1}, u_{t-5}) \\ &\quad + g_5(u_{t-6}, u_{t-7}, u_{t-8}) + g_6(u_{t-6}, u_{t-9}, u_{t-10}) + e_t. \end{aligned} \tag{5.4}$$

Note that most of the possible interaction effects can not be tested due to the low number of regressors in each test and that this search scheme does not consider all three-factor interactions, which of course could be done. The resulting model is

$$y_t = g_1(u_t) + g_2(u_{t-1}) + g_3(u_{t-5}) + g_4(u_{t-6}) + e_t, \tag{5.5}$$

since no interaction effects between u_{t-6} and the other significant regressors have been tested. This can be done with two tests including u_t, u_{t-1} and u_{t-6} in the first and at least u_{t-5} and u_{t-6} in the second. If for some reason these complementary tests are not done, the model structure

$$y_t = g_1(u_t, u_{t-6}) + g_2(u_{t-1}, u_{t-6}) + g_3(u_{t-5}, u_{t-6}) + e_t \tag{5.6}$$

can be considered for further model building.

5.3.1 Linear systems and time delays

If only linear systems are considered, some simplifications can be done. A linear system is a subgroup of the additive systems, see 2.16. This means that no interaction effects need to be considered. A complete testing can be done with only one data in each cell and it is not necessary to have more than two groups for each regressor, which means that the minimum amount of data needed for the analysis is 2^k, where k is the number of regressors. More data give better power for the tests. For example, a binary input signal can be used. If the signal covers most frequencies it should also be informative enough to enable tests with many regressors at the same time. The testing 'window' can be moved to make it possible to spot time delays by testing new batches of regressors.

6

Determine the structure of NFIR-models with a fixed levels input signal

We want to design an experiment to determine which regressors have effect on the output of a system, assumed to be described by a nonlinear FIR model, $y_t = g(u_t, u_{t-1}, \ldots, u_{t-k}) + e_t$, 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 first experiment is designed to give a setup as ideal for analysis by ANOVA as possible. This involves to select an input signal to the system that can only assume a small number of different fixed values.

To be able to evaluate the results a second regressor selection scheme is necessary. The one chosen for comparison, exhaustive search (see Section 3.2.4), is based on a simple idea, but is computationally very demanding.

The networks used here to estimate the function g, are one hidden layer neural networks with sigmoidal neurons in the hidden layer and linear neurons in the output layer. The minimisation algorithm used is Levenberg-Marquardt in MATLAB™'s Neural Network Toolbox. See Haykin (1994) for a treatment of neural networks.

6.1 Input signal design

To do identification experiments with the described variant of ANOVA, see Chapter 4, we need to construct an input signal u_t that contains all possible combinations of levels of the time lags we want to be able to examine. Assume that we have p regressors we want to examine. These could be different input signals or, as in this case, time-shifted variants of the same signal u_t. We have m levels of u_t and want n repetitions of each combination. The most straightforward way to design this signal is to enumerate all possible sequences of length p, with different level combinations of the regressors, 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 pth measurement of the output from the system. The jumps are done to get an equal number of data for each regressor level combination, but of course, all measurements could be used.

It is also possible to create a periodic signal u_t of length $n(m^p - 1)$, with the help of a p-stage, m-level shift register (Godfrey, 1993, page 41). In this signal all except one of the possible combinations occur n times and we use all the output from the system. Measurements of the missing combination (all time lags assuming the lowest level) might be easy to add afterwards, since only a constant input signal is needed.

6.1.1 More details for this experiment

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

$$y_t = g(u_t, u_{t-1}, u_{t-2}) + e_t, \tag{6.1}$$

see Table 6.1. That is, we will perform a three-way analysis of variance ($p = 3$). The level of the input signal u_t can assume the values -2, 1, 3 and 5 ($m = 4$). All possible combinations of these levels in the sequence u_{t-2}, u_{t-1}, u_t are constructed. The signal is repeated $n = 4$ times. Then the output y_t is computed according to Equation 6.1, with e_t Gaussian distributed noise with mean 0 and variance 1. The number of input/output data, N is set to $N = n * m^p = 4 * 4^3 = 256$. Two such data sets are constructed in each Monte-Carlo run to give an estimation data set and a validation data set.

6.2 Experiment setup and results

The regressor selection methods are used in the following way to estimate comparable system models. The performance criterion used in this experiment is the root mean square error (RMS) between measured output and simulated output from the model.

- **Exhaustive search** Divide the first data set into estimation data and verification data. Construct neural networks for all possible combinations of regressors under the assumption that the largest k is equal to 3, which gives seven possible networks: $g(u_t)$, $g(u_{t-1})$, $g(u_{t-2})$, $g(u_t, u_{t-1})$, $g(u_t, u_{t-2})$, $g(u_{t-1}, u_{t-2})$ and $g(u_t, u_{t-1}, u_{t-2})$. For each such model structure, construct networks with different numbers of parameters (e.g. 5, 10 and 15 neurons in the hidden 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.

- **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). This defines the regressor structure. To obtain a model, construct neural networks with different numbers of neurons (5, 10 ,15) for the chosen model. The information of the interaction pattern (model structure) is not used in this test. Estimate as in the exhaustive search method and choose the network with the smallest RMS value on the verification data.
  Note that one benefit of the ANOVA test, when a fixed levels input signal is used, is that we get a good estimate of the noise variance more or less automatically. This information can be used to determine when the minimisation algorithm has failed to find a minimum close to the global minimum. If the RMS value 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 a smaller RMS value, try to get more estimation data or change to another type of network. In this test, this useful feature is not taken advantage of.

Finally, 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 of our test functions given in Table 6.1. We also include 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. The results are collected in Table 6.1 and Table 6.2. The theoretical probability of finding the correct model structure was computed as in Example 4.2.

6.3 Conclusion

From Table 6.1 and Table 6.2 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, $u_t - 0.03u_{t-2}$, show that it is important to have large 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 nonlinear 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 can be confirmed by looking at RMS values on validation data.

In Table 6.2 the better performance for ANOVA as compared to Table 6.1 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.

One important issue for further modelling is how the regressors should enter the model, the interaction pattern. That information can be used to reduce the number of parameters needed in the model, which gives better variance properties for the model. If ANOVA is used, this information is gained at the same time as selection of the regressors. If exhaustive search is used, the search must be done among all possible interaction patterns to gain the same information. For the case when only three regressors are considered for inclusion in the model, the number of possible model

No.	Function	Exhaustive search	ANOVA regressors	ANOVA model structure	Theoret. average
1	$u_t - 0.03u_{t-2}$	10	6	5	4
2	$\ln\|u_t\| + u_{t-1} + e^{u_{t-2}}$	77	100	98	96
3	$u_{t-1} \cdot [u_t + \frac{1}{u_{t-2}}]$	100	100	98	98
4	$\mathrm{sgn}(u_{t-1})$	84	94	94	94
5	$\mathrm{sgn}(u_{t-1}) \cdot u_{t-2}$	93	96	96	96
6	$\mathrm{sgn}(u_{t-1}) \cdot u_t \cdot u_{t-2}$	100	100	100	100
7	$\ln\|u_{t-1} + u_{t-2}\|$	95	96	96	96
8	$\ln\|u_{t-1} \cdot u_{t-2}\|$	94	92	90	95
9	$u_{t-2} \cdot \ln\|u_{t-1}\|$	97	97	97	96
10	$u_{t-2}^3 \cdot \ln\|u_{t-1}\|$	50	95	95	96
11	$u_{t-2} \cdot (\ln\|u_{t-1}\|)^3$	93	95	95	96
12	$\|u_{t-2}\| \cdot e^{u_{t-1}}$	54	96	96	96
13	$u_{t-2} \cdot e^{u_{t-1}}$	49	94	94	96
14	$u_{t-2} \cdot e^{u_{t-1} - 0.03u_t}$	58	100	100	100
15	$\|u_t\|$	83	96	96	94
16	network $g(u_{t-1}, u_{t-2})$	73	88	88	-
	TOTAL	75.6	90.3	89.9	90.2

Table 6.1: Results from Monte Carlo simulations, 100 runs. Two regressor selection methods, exhaustive search and ANOVA are tested. Stated are percentage of correctly chosen models. The third column states how often ANOVA also picks out the correct interaction pattern, see Definition 2.2, of the regressors. The fourth column states the theoretical average of finding the correct model structure with ANOVA, which is computed as in Example 4.2. $N = 256$, $\sigma = 1$ and $\alpha = 0.01$.

structures would increase from seven to eighteen (if the search is done in one step).

In this experiment the computation time for ANOVA is roughly one to two seconds for each test. The computation time for exhaustive search is about six to seven minutes for each test. This sums up to roughly two weeks CPU time for the two result tables.

No.	Function	Exhaustive search	ANOVA regressors	ANOVA model structure	Theoret. average
1	$u_t - 0.03u_{t-2}$	94	100	100	99.95
2	$\ln\|u_t\| + u_{t-1} + e^{u_{t-2}}$	78	100	98	99.96
3	$u_{t-1} \cdot [u_t + \frac{1}{u_{t-2}}]$	100	100	100	99.98
4	$\mathrm{sgn}(u_{t-1})$	80	100	100	99.94
5	$\mathrm{sgn}(u_{t-1}) \cdot u_{t-2}$	92	100	100	99.96
6	$\mathrm{sgn}(u_{t-1}) \cdot u_t \cdot u_{t-2}$	100	100	100	100
7	$\ln\|u_{t-1} + u_{t-2}\|$	94	100	100	99.96
8	$\ln\|u_{t-1} \cdot u_{t-2}\|$	82	100	100	99.95
9	$u_{t-2} \cdot \ln\|u_{t-1}\|$	95	100	100	99.96
10	$u_{t-2}^3 \cdot \ln\|u_{t-1}\|$	56	100	100	99.96
11	$u_{t-2} \cdot (\ln\|u_{t-1}\|)^3$	91	99	99	99.96
12	$\|u_{t-2}\| \cdot e^{u_{t-1}}$	54	100	100	99.96
13	$u_{t-2} \cdot e^{u_{t-1}}$	49	100	100	99.96
14	$u_{t-2} \cdot e^{u_{t-1} - 0.03u_t}$	58	100	100	100
15	$\|u_t\|$	73	100	100	99.94
16	network $g(u_{t-1}, u_{t-2})$	94	100	100	-
	TOTAL	80.6	99.9	99.9	99.96

Table 6.2: Results from Monte Carlo simulations, 100 runs. Two regressor selection methods, exhaustive search and ANOVA are tested. Stated are percentage of correctly chosen models. The third column states how often ANOVA also picks out the correct interaction pattern, see Definition 2.2, of the regressors. The fourth column states the theoretical average of finding the correct model structure with ANOVA, which is computed as in Example 4.2. $N = 256$, $\sigma = 0.0001$ and $\alpha = 0.0001$.

7

Determine the structure of NFIR-models with a random input signal

A fixed levels input signal is not practical in all situations. We would also like to be able to use ANOVA when a random signal is used as input to the system. In this chapter we evaluate ANOVA with a uniformly distributed random input signal to test what happens for that case. This is also a first step towards analysing autoregressive processes, since old outputs from the system cannot be viewed as fixed levels signals.

7.1 Experiment setup

The output signal y_t is computed according to the equation

$$y_t = g(u_t, u_{t-1}, u_{t-2}) + e_t, \tag{7.1}$$

where the function $g(\cdot)$ is given in Table 7.1.

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.

The simulated measurement error signal e_t is zero-mean Gaussian noise with standard deviation 0.0001.

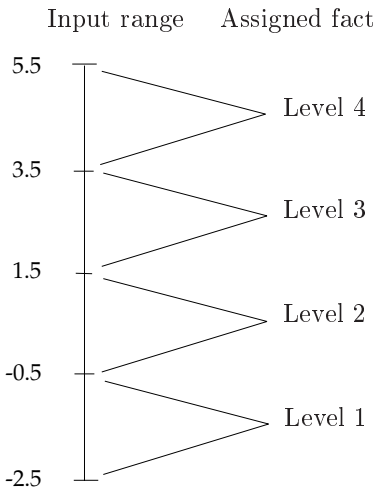


Figure 7.1: The range of the input signal is divided into intervals, each of length two. Each interval is assigned a factor level, used in the ANOVA.

Now, these simulated time series can not be used directly. Levels need to be assigned to the input values for use in the ANOVA. This is done according to Figure 7.1. We get four factor levels for each time lag, which gives 64 cells in the experiment design, each cell corresponding to a unique combination of factor levels.

The level assignment introduces a new type of error in the ANOVA. The output y_t can now be seen as

$$y_t = E[y_t|(u_t, u_{t-1}, u_{t-2}) \in C] + e_t + n_t, \tag{7.2}$$

where $E[y_t|(u_t, u_{t-1}, u_{t-2}) \in C]$ is the expected value of y_t, given that the input is assigned to cell C, and

$$n_t = g(u_t, u_{t-1}, u_{t-2}) - E[y_t|(u_t, u_{t-1}, u_{t-2}) \in C]. \tag{7.3}$$

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 categorise u_t. The distribution is not necessarily equal in all 64 cells, which violates the ANOVA assumption on equal variances in all cells.

To get a good experiment design for the ANOVA, all cells need to have at least one observation. Preferably all cells should have the same number of observations. With a random input signal this is impossible to guarantee,

but if we let the number of input data be moderately high, we can be quite sure to get observations in each cell. In the Monte-Carlo simulations, 800 input/output data are used for each run. The number of data needed grows rapidly if more time lags are tested or if a finer grid for the input signal is used.

Exhaustive search was not used for comparison here. There is no reason for that method to work either better nor worse with this type of input signal. As long as the same type of input signal is used for both estimation and validation of the models, the results should be comparable to the results in the previous chapter.

7.2 Results from Monte-Carlo simulations

In Table 7.1 we see that for most of the selected functions the performance of ANOVA is about 90%. There are four exceptions to this; three of them will be discussed in detail later in this chapter. For function 1 the problem is too low signal to noise ratio for the time lag u_{t-2}, since the standard deviation of n_t is 0.6 in all cells. See Section 4.3 for a discussion on the power of the tests.

The loss in performance compared with the case with fixed input levels is not as great as anticipated. The division into intervals is, after all, very rough.

Conclusions from the study will be given after the discussion of functions 2, 3 and 14.

7.3 What is the problem with function 2?

As we can see from Table 7.1, the time lags of function 2,

$$y_t = \ln|u_t| + u_{t-1} + e^{u_{t-2}} + e_t, \tag{7.4}$$

are almost never identified correctly. What is the problem? In almost all cases only the time lag u_{t-2} is found. If we check the function we see that the error term, n_t in cell C, is given by

$$n_t = \ln|u_t| + u_{t-1} + e^{u_{t-2}} - E[y_t|(u_t, u_{t-1}, u_{t-2}) \in C]. \tag{7.5}$$

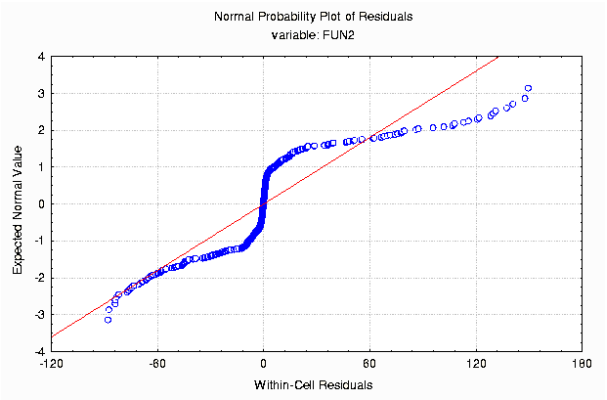
The variance of n_t is large in the cells where u_{t-2} is large and does not depend as strongly on u_t and u_{t-1}. The large within-cell variation leads

No.	Function	% correct results	Comment
1	$u_t - 0.03u_{t-2}$	52	In 46% of the cases only u_t was found.
2	$\ln\|u_t\| + u_{t-1} + e^{u_{t-2}}$	0	Interrupted after 40 runs.
3	$u_{t-1} \cdot [u_t + \frac{1}{u_{t-2}}]$	8	Interrupted after 40 runs.
4	$\mathrm{sgn}(u_{t-1})$	76	
5	$\mathrm{sgn}(u_{t-1}) \cdot u_{t-2}$	90	
6	$\mathrm{sgn}(u_{t-1}) \cdot u_t \cdot u_{t-2}$	100	
7	$\ln\|u_{t-1} + u_{t-2}\|$	92	
8	$\ln\|u_{t-1} \cdot u_{t-2}\|$	90	
9	$u_{t-2} \cdot \ln\|u_{t-1}\|$	90	
10	$u_{t-2}^3 \cdot \ln\|u_{t-1}\|$	90	
11	$u_{t-2} \cdot (\ln\|u_{t-1}\|)^3$	90	
12	$\|u_{t-2}\| \cdot e^{u_{t-1}}$	90	
13	$u_{t-2} \cdot e^{u_{t-1}}$	88	
14	$u_{t-2} \cdot e^{u_{t-1} - 0.03u_t}$	6	Interrupted after 40 runs.
15	$\|u_t\|$	96	

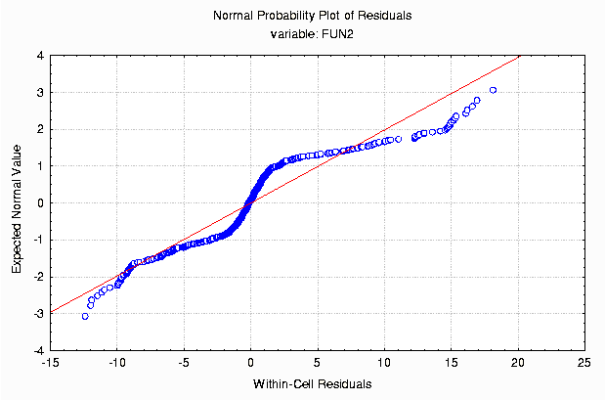
Table 7.1: Results from Monte-Carlo simulations, 50 runs. ANOVA was performed on 800 input/output data, where the uniformly distributed input data was divided into four equal intervals. The normally distributed 'measurement' noise has standard deviation 0.0001, and the level of significance for ANOVA is 0.01. After 40 runs, the analysis of functions 2, 3 and 14 was changed, due to the obvious failure, and the simulation rerun, see Table 7.14.

Interval	-2.5 - (-0.5)	-0.5 - 1.5	1.5 - 3.5	3.5 - 5.5
u_t	1.6	Inf	0.8	0.5
u_{t-1}	2	2	2	2
u_{t-2}	0.5	3.9	29	210

Table 7.2: Contributions to the range of n_t (Equation (7.5)) from different sources. The entries state the difference between the maximum and minimum value in the stated interval for the contribution from each source. The contributions should be summed to give the range of n_t in each cell. Note: The possible infinite variation of u_t is in practice mostly about 1.



(a) All cells included in the analysis. The plot shows that the assumption of normal distribution of the random error component is not valid, since it is not a straight line.



(b) One level of u_{t-2} excluded from the analysis. The residuals are still belonging to a non-normal distribution.

Figure 7.2: Normal probability plots of within-cell residuals for function 2.

Effect	Degrees of Freedom	Mean Square	F	p-level
u_t	3	1638	2.0	0.11
u_{t-1}	3	934	1.1	0.33
u_{t-2}	*3	*459191	*561.5	0.0000*
u_t, u_{t-1}	9	556	0.7	0.73
u_t, u_{t-2}	*9	*2562	*3.1	0.001*
u_{t-1}, u_{t-2}	9	155	0.2	0.99
u_t, u_{t-1}, u_{t-2}	27	441	0.5	0.97
Error	734	818		

Table 7.3: Analysis of Variance Table for Function 2, all cells included. For an explanation of the table, see Table 4.2.

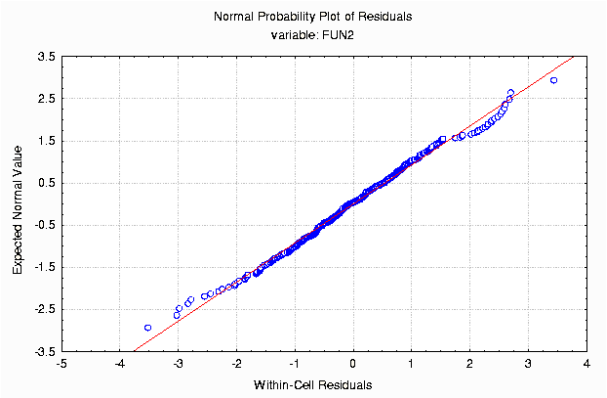


Figure 7.3: Normal probability plot of within-cell residuals for function 2. Two levels of u_{t-2} excluded from the analysis. Here the assumption of normal distribution is valid.

to a large residual quadratic sum. The between-cell variations of the other parts of the function drown in the noise from u_{t-2}, see Table 7.2.

This was not a problem when the input signal u_t took only fixed values, since then all the within-cell variation came from the measurement noise e_t, which has equal variance in all cells.

This problem is not as bad as it looks though. If we perform the ANOVA in a wiser fashion and check the assumptions before accepting the result we will get a fair warning that everything is not as nice as usual. There are two

Effect	Degrees of Freedom	Mean Square	F	p-level
u_t	*3	*156	*6.6	0.0002*
u_{t-1}	*3	*1286	*54.1	0.0000*
u_{t-2}	*2	*11381	*478.4	0.0000*
u_t, u_{t-1}	9	28	1.2	0.31
u_t, u_{t-2}	6	38	1.6	0.14
u_{t-1}, u_{t-2}	6	25	1.1	0.39
u_t, u_{t-1}, u_{t-2}	18	21	0.9	0.58
Error	567	24		

Table 7.4: Analysis of Variance Table for Function 2, one level of u_{t-2} excluded.

Effect	Degrees of Freedom	Mean Square	F	p-level
u_t	*3	*87	*70	0.0000*
u_{t-1}	*3	*640	*515	0.0000*
u_{t-2}	*1	*312	*251	0.0000*
u_t, u_{t-1}	9	1.3	1.0	0.42
u_t, u_{t-2}	3	1.7	1.4	0.25
u_{t-1}, u_{t-2}	3	0.1	0.1	0.95
u_t, u_{t-1}, u_{t-2}	9	2.2	1.7	0.08
Error	370	1.2		

Table 7.5: Analysis of Variance Table for Function 2, two levels of u_{t-2} excluded.

checks that will help us along. The first is the normal probability plot of the within-cell residuals (they should be normal), see Section 4.1.1, and the second is the within-cell standard deviations (which should be approximately equal). For function 2, a typical normal probability plot of a complete (all cells) analysis looks like Figure 7.2(a). This is clearly not normal and the test results in the corresponding ANOVA table, Table 7.3, should not be trusted. The power of the tests, which means the ability to spot contributing regressors, is probably affected, see Section 5.1. The within-cell standard deviations are very large whenever u_{t-2} takes a value between 3.5 and 5.5. In the ANOVA table we see that u_{t-2} clearly contributes to the output, so

we could decrease the number of levels to test for this factor and exclude the 16 cells with large standard deviations. This leads to the ANOVA table, Table 7.4, and the corresponding normal probability plot, Figure 7.2(b). We can see that the estimated variance has decreased (MS Error) significantly, that now the contributions from all the time lags are found significant, but the normal probability plot still tells that the analysis should not be trusted. Another check of the within-cell standard deviations gives that the analysis might be improved if we exclude also the cells where u_{t-2} takes values between 1.5 and 3.5. This gives another dramatic reduction of estimated variance and the same test result as previous analysis, see Table 7.5. This time the normal probability plot looks all right, see Figure 7.3.

A proper analysis can be made without knowing the function in advance, even if this knowledge has been used here to explain the failure of the head-on approach of the Monte-Carlo simulations.

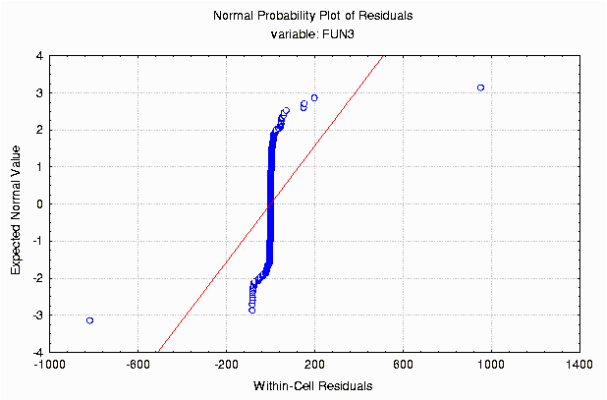
7.4 A closer look at function 3

There are some problems with the analysis of function 3,

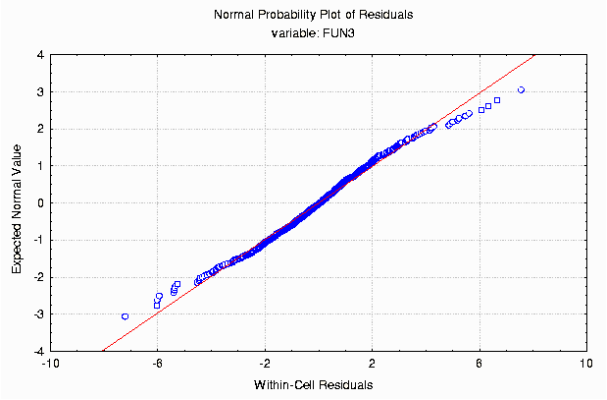
$$y_t = u_{t-1} \cdot [u_t + \frac{1}{u_{t-2}}] + e_t, \tag{7.6}$$

as well. As for function 2, see Section 7.3, the problems are not as severe as they look at first sight. A typical ANOVA table from the Monte-Carlo study looks like Table 7.7. The corresponding normal probability plot can be seen in Figure 7.4(a). We can see that the within-cell residuals are clearly non-normally distributed, so the test results should not be trusted. When a study of the within-cell standard deviations for the 64 cells is made, see Table 7.6, it is found that when the time lag u_{t-2} takes values between -0.5 and 1.5, we get large contributions to the residual quadratic sum.

This is not surprising, as it might give division by very small numbers, see Equation (7.6). When the problematic cells, with large variance of n_t, are excluded from the analysis, we get the ANOVA table in Table 7.8 and the normal probability plot in Figure 7.4(b). Now the residuals are close to normally distributed and we get test results that are much closer to the truth, even if we on this data set find an erroneous structure.



(a) All cells included in the analysis. The deviation from normal distribution — a straight line — is clear.



(b) One level of u_{t-2} excluded from the analysis. Here the residuals have a normal distribution.

Figure 7.4: Normal probability plots of within-cell residuals for function 3.

u_t	u_{t-2}	u_{t-1}			
		-2.5 – (-0.5)	-0.5 – 1.5	1.5 – 3.5	3.5 – 5.5
-2.5 – (-0.5)	-2.5 – (-0.5)	1.9	1.2	1.6	3.4
	-0.5 – 1.5	5.0	3.9	81.3	20.5
	1.5 – 3.5	1.2	0.8	1.5	3.2
	3.5 – 5.5	1.3	1.0	2.0	2.5
-0.5 – 1.5	-2.5 – (-0.5)	1.2	0.2	2.0	2.8
	-0.5 – 1.5	48.5	3.0	14.7	10.1
	1.5 – 3.5	0.8	0.7	1.5	2.5
	3.5 – 5.5	0.5	0.4	1.7	2.8
1.5 – 3.5	-2.5 – (-0.5)	1.7	0.7	2.5	1.9
	-0.5 – 1.5	22.7	3.0	59.4	14.6
	1.5 – 3.5	1.6	1.6	1.9	3.1
	3.5 – 5.5	1.8	1.3	2.3	3.3
3.5 – 5.5	-2.5 – (-0.5)	1.9	1.9	2.3	3.3
	-0.5 – 1.5	4.9	4.0	34.3	102.6
	1.5 – 3.5	2.7	2.2	2.9	3.6
	3.5 – 5.5	1.5	3.1	2.6	3.5

Table 7.6: Within-cell standard deviations for function 3. The cells are indicated by the given intervals.

Effect	Degrees of Freedom	Mean Square	F	p-level
u_t	*3	*8560	*3.5	0.016*
u_{t-1}	3	2929	1.2	0.31
u_{t-2}	3	340	0.2	0.92
u_t, u_{t-1}	9	4207	1.7	0.08
u_t, u_{t-2}	9	1572	0.6	0.77
u_{t-1}, u_{t-2}	9	1915	0.8	0.64
u_t, u_{t-1}, u_{t-2}	27	1839	0.7	0.82
Error	734	2467		

Table 7.7: Analysis of Variance Table for Function 3, all cells included.

Effect	Degrees of Freedom	Mean Square	F	p-level
u_t	*3	*2210	*507	0.0000*
u_{t-1}	*3	*1911	*438	0.0000*
u_{t-2}	*2	*205	*47	0.0000*
u_t, u_{t-1}	*9	*1422	*326	0.0000*
u_t, u_{t-2}	6	7.3	*1.7	0.13
u_{t-1}, u_{t-2}	*6	*169	*39	0.0000*
u_t, u_{t-1}, u_{t-2}	*18	*9.2	*2.1	0.005*
Error	552	4.4		

Table 7.8: Analysis of Variance Table for Function 3, one level of u_{t-2} excluded.

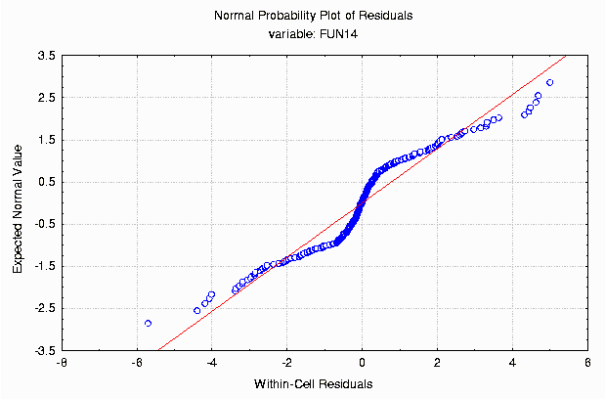


Figure 7.5: Normal probability plot of within-cell residuals for function 14 with two levels of u_{t-1} excluded from the analysis. The plot shows that the residuals does not have a normal distribution.

7.5 Function 14

It is not as easy to obtain a good result for function 14,

$$y_t = u_{t-2} \cdot e^{u_{t-1} - 0.03 u_t} + e_t, \tag{7.7}$$

see Table 7.1. When discarding levels as for function 2 and 3, Sections 7.3 and 7.4, we get the normal probability plot in Figure 7.5. This means we have discarded, in this case, more than half of the available data and still

Effect	Degrees of Freedom	Mean Square	F	p-level
u_t	3	2.1	0.4	0.72
u_{t-1}	*1	*624	*136	0.0000*
u_{t-2}	*3	*782	*171	0.0000*
u_t, u_{t-1}	3	1.2	0.3	0.86
u_t, u_{t-2}	9	3.9	0.8	0.58
u_{t-1}, u_{t-2}	*3	*496	*108	0.0000*
u_t, u_{t-1}, u_{t-2}	9	*3.2	0.7	0.70
Error	369	4.6		

Table 7.9: Analysis of Variance Table for Function 14, two levels of u_{t-1} excluded.

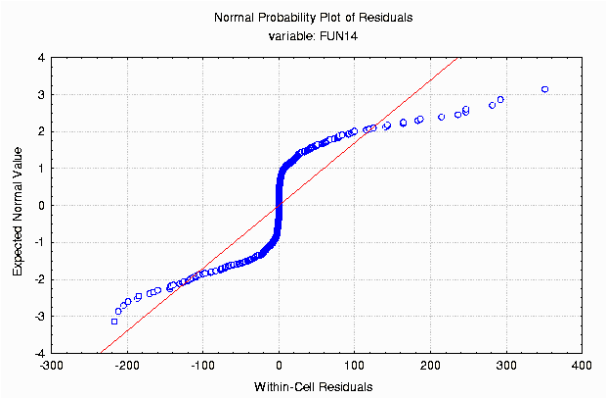
have no trustworthy result, Table 7.9. One thing can still be done. The division of the input into intervals is very rough. A finer grid with more intervals will reduce the variation in each cell. This could give us a more trustworthy result. And, of course, if possible, more data could be collected to help the analysis along. The number of intervals one can use to categorise the data depends heavily on the available data. It is important that no cells are empty and the same number of data in each cell makes a better analysis.

7.5.1 Finer grid

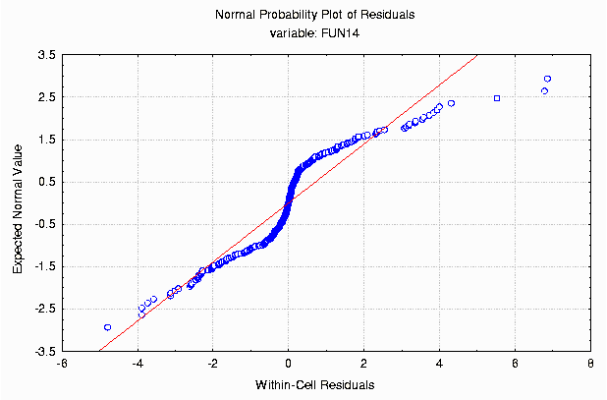
In Figure 7.6(a) we see the normal probability plot of the within-cell residuals when we have divided u_{t-1} into 8 intervals while u_t and u_{t-2} are still divided into 4 intervals each to get enough data in each cell. The data in this plot is clearly not from a normal distribution.

When the four intervals with largest variation of u_{t-1} are excluded from the analysis, we get the normal probability plot in Figure 7.6(b). This contains half of the original data, but is still not satisfactory. If we exclude two more intervals of u_{t-1}, Figure 7.7, the plot looks better, but now only 1/4 of the data are left.

The corresponding ANOVA tables are given in Table 7.10, Table 7.11 and Table 7.12. As we can see in the latter one, the contribution from u_t is not found.



(a) All cells included in the analysis. There are strong non-normal effects.



(b) Four levels of u_{t-1} excluded from the analysis.

Figure 7.6: Normal probability plots of within-cell residuals for function 14 with u_{t-1} divided into eight intervals.

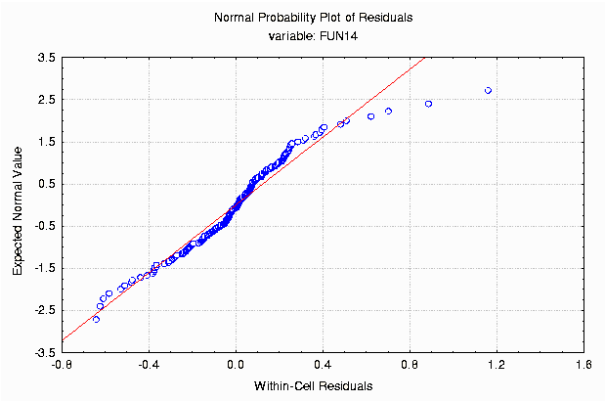


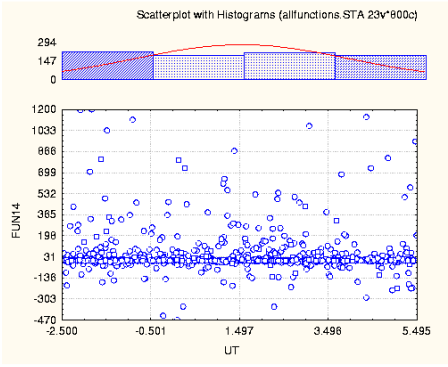
Figure 7.7: Normal probability plot of within-cell residuals for function 14 with u_{t-1} divided into eight intervals. Six levels of u_{t-1} excluded from the analysis. Here the effects are almost normal.

Effect	Degrees of Freedom	Mean Square	F	p-level
u_t	3	3758	1.5	0.21
u_{t-1}	*7	*519198	*209	0.0000*
u_{t-2}	*3	*818947	*330	0.0000*
u_t, u_{t-1}	21	1750	0.7	0.83
u_t, u_{t-2}	9	5541	2.2	0.02
u_{t-1}, u_{t-2}	*21	*330683	*133	0.0000*
u_t, u_{t-1}, u_{t-2}	*63	*4731	*1.9	0.0001*
Error	670	2480		

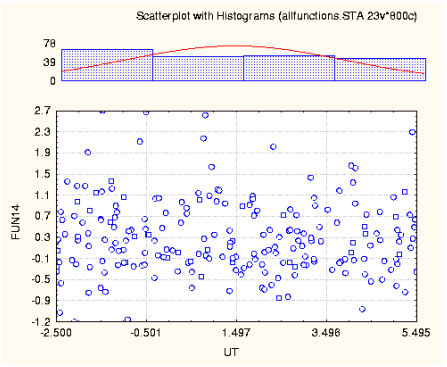
Table 7.10: Analysis of Variance Table for Function 14, all cells included with u_{t-1} in fine grid.

7.5.2 A look at the data

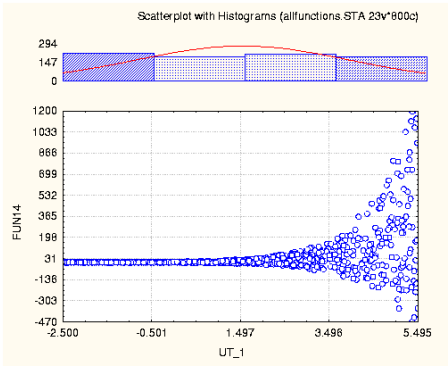
If we take a closer look at typical input/output data for function 14, see Figures 7.8 and 7.9, we can see that the contributions from time lags u_{t-1} and u_{t-2} are clear, while the contribution from u_t is not obvious at all. The scatter plots should be interpreted carefully since the data are projected into one dimension only.



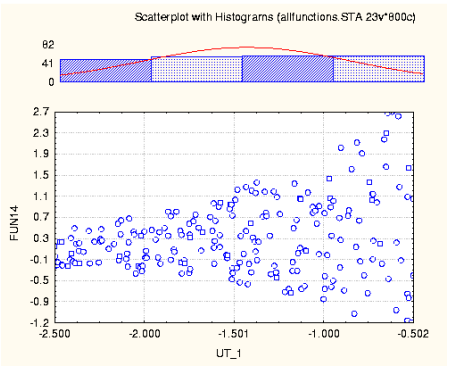
(a) Output values against time lag u_t. Notice that the variation of the output data seems to be the same over the whole range of u_t.



(b) Output values against time lag u_t. Notice that the variation of the output data still seems to be the same over the whole range of u_t.



(c) Output values against time lag u_{t-1}. Notice that the variation of the output data grows approximately exponentially with u_{t-1}.



(d) Output values against time lag u_{t-1}.

Figure 7.8: Scatter plot with histograms for function 14. To the left all data are plotted. To the right only 1/4 of the data, from the group with smallest within-cell variation of u_{t-1}, are included.

Effect	Degrees of Freedom	Mean Square	F	p-level
u_t	3	1.4	0.7	0.56
u_{t-1}	*3	*295	*144	0.0000*
u_{t-2}	*3	*619	*302	0.0000*
u_t, u_{t-1}	9	0.8	0.4	0.94
u_t, u_{t-2}	9	2.6	1.3	0.26
u_{t-1}, u_{t-2}	*9	*231	*113	0.0000*
u_t, u_{t-1}, u_{t-2}	27	2.0	1.0	0.53
Error	337	2.1		

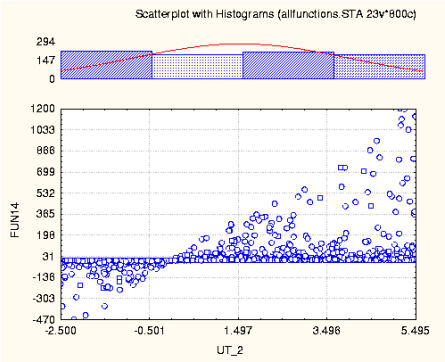
Table 7.11: Analysis of Variance Table for Function 14, four out of eight levels of u_{t-1} excluded.

Effect	Degrees of Freedom	Mean Square	F	p-level
u_t	3	0.1	1.4	0.23
u_{t-1}	*1	*4.5	*67	0.0000*
u_{t-2}	*3	*17	*255	0.0000*
u_t, u_{t-1}	3	0.03	0.4	0.75
u_t, u_{t-2}	9	0.13	1.9	0.05
u_{t-1}, u_{t-2}	*3	*3.5	*51	0.0000*
u_t, u_{t-1}, u_{t-2}	9	0.07	1.1	0.37
Error	172	0.07		

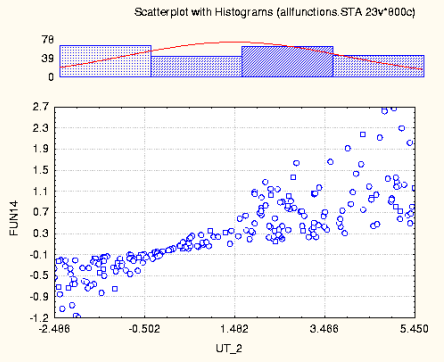
Table 7.12: Analysis of Variance Table for Function 14, six out of eight levels of u_{t-1} excluded.

7.5.3 Conclusions for function 14

It is probably not possible to find the contribution from the time lag u_t because of its size compared to the other two contributions, at least not when using a random input signal like in this study. The division into intervals makes n_t so large that the signal to noise ratio with respect to the time lag u_t gets too small. See Table 7.13 to get a feeling for how large n_t is in the different cells.



(a) Output values against time lag u_{t-2}. Here the variation of the output data varies more linearly.



(b) Output values against time lag u_{t-2}.

Figure 7.9: Scatter plot with histograms for function 14. To the left all data are plotted. To the right only 1/4 of the data, from the group with smallest within-cell variation of u_{t-1}, are included.

7.6 New Monte-Carlo study for the problematic functions

A new Monte-Carlo study with the same conditions as in Section 7.1 was made for functions 2, 3 and 14 with the modifications described in Sections 7.3, 7.4 and 7.5.1, respectively. The results are given in Table 7.14. A dramatic improvement of the performance on functions 2 and 3 can be noted. As concluded in Section 7.5.3, the failure to find the contribution from u_t for function 14, is obvious.

7.7 Higher noise level

To complete this investigation, a Monte-Carlo simulation was run on simulated data with a higher noise level. The experiment setup was exactly as in previous sections, with the exception that the standard deviation of e_t in Equation (7.1) is 1 instead of 0.0001. Functions 2, 3 and 14 were analysed with the modifications used in Section 7.6. Also the functions 4, 5, 12 and 13 needed a different analysis, which can be explained by a more care-

u_t	u_{t-2}	u_{t-1}			
		-2.5 – (-0.5)	-0.5 – 1.5	1.5 – 3.5	3.5 – 5.5
-2.5 – (-0.5)	-2.5 – (-0.5)	1.6	12.4	86.7	642.6
	-0.5 – 1.5	1.0	7.5	55.8	412.4
	1.5 – 3.5	2.2	16.0	118.1	872.7
	3.5 – 5.5	3.3	24.4	180.4	1333.0
-0.5 – 1.5	-2.5 – (-0.5)	1.5	11.1	81.9	605.1
	-0.5 – 1.5	1.0	7.1	52.6	388.4
	1.5 – 3.5	2.0	15.1	111.2	821.9
	3.5 – 5.5	3.1	23.0	169.9	1255.3
1.5 – 3.5	-2.5 – (-0.5)	1.5	11.4	84.6	625.0
	-0.5 – 1.5	1.0	7.3	54.0	398.8
	1.5 – 3.5	2.1	15.6	115.2	851.1
	3.5 – 5.5	3.2	23.9	176.4	1303.4
3.5 – 5.5	-2.5 – (-0.5)	1.6	12.2	90.1	665.4
	-0.5 – 1.5	1.0	7.7	57.1	421.7
	1.5 – 3.5	2.3	16.7	123.0	909.1
	3.5 – 5.5	3.5	25.6	189.0	1396.5

Table 7.13: Contributions to the range of n_t from different sources. The entries are the differences between largest and smallest functional values in the cells indicated by the given intervals.

No.	Function	% correct results	Comment
2	$\ln\|u_t\| + u_{t-1} + e^{u_{t-2}}$	92	See section 7.3.
3	$u_{t-1} \cdot [u_t + \frac{1}{u_{t-2}}]$	100	See section 7.4.
14	$u_{t-2} \cdot e^{u_{t-1} - 0.03u_t}$	6	See Section 7.5.1. In 80% of the cases the regressor u_t was not selected.

Table 7.14: Results from Monte-Carlo simulations, 50 runs. ANOVA was performed on 800 input/output data, where the uniformly distributed input data was divided into intervals according to the referenced sections. The normally distributed 'measurement' noise has standard deviation 0.0001, and the level of significance for ANOVA is 0.01.

No.	Function	% correct results		Comment
1	$u_t - 0.03u_{t-2}$	4	52	In 92% of the cases only u_t was found.
2	$\ln\|u_t\| + u_{t-1} + e^{u_{t-2}}$	96	92	
3	$u_{t-1} \cdot [u_t + \frac{1}{u_{t-2}}]$	100	100	
4	$\mathrm{sgn}(u_{t-1})$	92	76	
5	$\mathrm{sgn}(u_{t-1}) \cdot u_{t-2}$	90	90	
6	$\mathrm{sgn}(u_{t-1}) \cdot u_t \cdot u_{t-2}$	100	100	
7	$\ln\|u_{t-1} + u_{t-2}\|$	98	92	
8	$\ln\|u_{t-1} \cdot u_{t-2}\|$	88	90	
9	$u_{t-2} \cdot \ln\|u_{t-1}\|$	90	90	
10	$u_{t-2}^3 \cdot \ln\|u_{t-1}\|$	82	90	
11	$u_{t-2} \cdot (\ln\|u_{t-1}\|)^3$	82	90	
12	$\|u_{t-2}\| \cdot e^{u_{t-1}}$	98	90	
13	$u_{t-2} \cdot e^{u_{t-1}}$	98	88	
14	$u_{t-2} \cdot e^{u_{t-1} - 0.03u_t}$	2	6	In 96% of the cases only (u_{t-1}, u_{t-2}) was found.
15	$\|u_t\|$	90	96	

Table 7.15: Results from Monte-Carlo simulations, 50 runs. ANOVA was performed on 800 input/output data, where the uniformly distributed input data was divided into four equal intervals. The normally distributed 'measurement' noise has standard deviation 1 in the first data column and 0.0001 in the second data column (the values are taken from Tables 7.1 and 7.14), and the level of significance for ANOVA is 0.01.

ful inspection of assumptions before the Monte-Carlo simulation. That is, these modifications should have been done also in the former Monte-Carlo simulation (Section 7.1). For functions 4 and 5, the data with u_{t-1} between -0.5 and 1.5 were discarded, and for functions 12 and 13 the data with u_{t-1} between 1.5 and 5.5 were discarded. The results from the Monte-Carlo simulation are collected in Table 7.15. The introduction of more noise does not give any surprising result. The functions 1 and 14, which have bad signal-to-noise ratio, are still not giving any good results. The rest of the functions get a test performance comparable to the results in Table 7.1.

7.8 Conclusion

It seems like it is possible to get good results from input/output data with a random input signal. The ANOVA test seem to be more sensitive to the noise term with non-normal characteristics, introduced by the random input signal, than to the variance of the normally distributed 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 analysed erroneously. This might not pose a large problem, since the small contributions might not enhance the fit of a model by very much anyway.

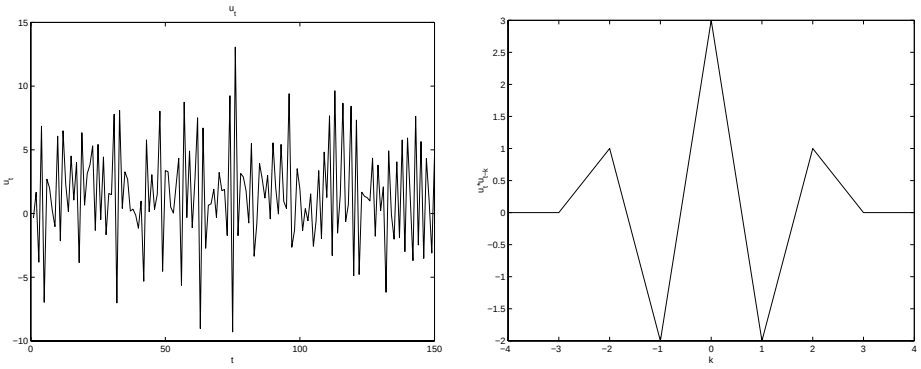
8

Determine the structure of NFIR-models with a correlated input signal

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.

What happens then if normal operations data are used? In identification applications, it would be nice to be able to use a wide range of different input data. So far, we have covered multi-level pseudo-random signals and uniformly distributed random signals. These correspond 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 we have no possibility to choose the input signal, we need to know if the analysis get tarnished by correlated input signals. The following analysis is meant to give an indication if extra care is needed when the input is correlated.



(a) 150 samples of u_t plotted against time.

(b) The covariance function for u_t.

Figure 8.1: One data sequence from Equation (8.2)

8.1 Experiment setup

The output signal y_t is computed according to the equation

$$y_t = g(u_t, u_{t-1}, u_{t-2}) + e_t, \tag{8.1}$$

where the function $g(\cdot)$ is given in Table 8.1, as in previous chapters.

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 results in previous chapters we would like u_t to assume values roughly between -2.5 and 5.5 and jump around in the regressor space given by the first three time lags. In this test, we will also have a correlation between adjacent time lags. One choice that fulfills these criteria is the moving average

$$u_t = x_t - x_{t-1} + x_{t-2}, \tag{8.2}$$

where x_t is white noise uniformly distributed between -2.75 and 8.25, see also Figure 8.1. There are two possible ways to proceed with this kind of input signal. One is to use all the measurements in the data series, obtaining a severely unbalanced design. The other one is to use the scheme described in Section 5.2.5, obtaining a balanced design. The second method to use the data is preferred and studied more carefully than the first one. 10

Monte-Carlo simulations using all the data in the data set are run and 50 Monte-Carlo simulations with data selected from the data sets are run.

The same grouping as in Chapter 7 is used. That is, the range between -2.5 and 5.5 is divided into four intervals to give four factor levels, i.e., 64 cells. To select a proper length, N, for the data series used in the Monte-Carlo simulations, the lowest cell counts are investigated for some different lengths of the data series (2000 data series used). For $N = 1000$, 1815 data series were giving empty cells and the rest had lowest cell count one, which makes all the data series useless for analysis with ANOVA, at least with this grouping. For $N = 2000$, the amount of useful data series was 16%, for $N = 3000$ 49%, for $N = 4000$ 74% and for $N = 5000$ 87%. See Figure 8.2 for the distribution of the lowest cell counts in the 2000 data series. Data series with length $N = 5000$ will be used in the Monte-Carlo simulations. This gives about 1600 useful observations with the level of u_t between -2.5 and 5.5 for three adjacent time lags.

For the Monte-Carlo simulations with unbalanced design, data series with empty cells will not be considered, to give a practical setup. For a real case with empty cells, other groupings can be tried. For the Monte-Carlo simulations with balanced design we will analyse 10 data series with lowest cell count 2, 10 data series with lowest cell count 3, and so on up to lowest cell count 6, which gives the total 50 runs. Especially the lowest cell counts 5 and 6 will be over-represented compared to the distribution in Figure 8.2. Data series with other lowest cell counts are not considered.

8.2 Results from Monte-Carlo simulations

The results are collected in two tables. Table 8.1 which gives the total percentage of correct results for the 10 unbalanced Monte-Carlo simulations and the 50 balanced Monte-Carlo simulations, and Table 8.2 which gives the number of correct results depending on the lowest cell counts for the balanced cases.

The analysis in this chapter is enhanced using the methods described in Chapter 7. The reason this is done here and not in the previous experiments is that a more careful inspection of the data and a check to see if the assumptions were fulfilled was done. For functions 2, 3, 4, 5, 11, 12, 13 and 14, groups with especially high within-cell variation had to be removed from the analysis to get good normal probability plots and trustworthy results. In the balanced case, it seemed like the analysis results were more sensitive to the non-normal noise introduced by the grouping of the input

		% correct results	
No.	Function	unbalanced	balanced
1	$u_t - 0.03u_{t-2}$	60	14
2	$\ln\|u_t\| + u_{t-1} + e^{u_{t-2}}$	100	98
3	$u_{t-1} \cdot [u_t + \frac{1}{u_{t-2}}]$	100	100
4	$\mathrm{sgn}(u_{t-1})$	100	94
5	$\mathrm{sgn}(u_{t-1}) \cdot u_{t-2}$	80	100
6	$\mathrm{sgn}(u_{t-1}) \cdot u_t \cdot u_{t-2}$	100	100
7	$\ln\|u_{t-1} + u_{t-2}\|$	70	86
8	$\ln\|u_{t-1} \cdot u_{t-2}\|$	90	84
9	$u_{t-2} \cdot \ln\|u_{t-1}\|$	100	82
10	$u_{t-2}^3 \cdot \ln\|u_{t-1}\|$	90	90
11	$u_{t-2} \cdot (\ln\|u_{t-1}\|)^3$	90	98
12	$\|u_{t-2}\| \cdot e^{u_{t-1}}$	100	84
13	$u_{t-2} \cdot e^{u_{t-1}}$	100	94
14	$u_{t-2} \cdot e^{u_{t-1} - 0.03u_t}$	10	6
15	$\|u_t\|$	70	98

Table 8.1: Results from Monte-Carlo simulations. ANOVA was performed on input/output data, where the input data was generated by (8.2). The normally distributed 'measurement' noise has standard deviation 0.0001, and the level of significance for ANOVA is 0.01. $N = 5000$, but the number of useful input/output data varies in each simulation. The labels balanced and unbalanced refer to the way the input/output data were treated before analysis. 10 runs were made for the unbalanced case and 50 for the balanced.

data, compared to the unbalanced case. One reason could be that the non-normal noise is not averaged over that many data points, which would give a nearly normal distribution.

Comparing the result in Table 8.1 with previous results, see Table 7.1 on page 60 and Table 7.14 on page 74 (for functions 2, 3 and 14), we can see that the frequency of correct answers is comparable for the case with balanced design. The division of the results depending on the lowest cell count, see Table 8.2, shows that it does not really matter how much data in each cell that are used, with one exception. For function 12, it seems like the drop in performance for cell count 2 might depend on a power reduction due to the low number of data used.

The results for the unbalanced case should be interpreted with more care,

		No. of correct results				
No.	Function	6	5	4	3	2
1	$u_t - 0.03u_{t-2}$	1	1	4	0	1
2	$\ln\|u_t\| + u_{t-1} + e^{u_{t-2}}$	10	10	10	10	9
3	$u_{t-1} \cdot [u_t + \frac{1}{u_{t-2}}]$	10	10	10	10	10
4	$\mathrm{sgn}(u_{t-1})$	8	10	10	9	10
5	$\mathrm{sgn}(u_{t-1}) \cdot u_{t-2}$	10	10	10	10	10
6	$\mathrm{sgn}(u_{t-1}) \cdot u_t \cdot u_{t-2}$	10	10	10	10	10
7	$\ln\|u_{t-1} + u_{t-2}\|$	9	9	9	9	8
8	$\ln\|u_{t-1} \cdot u_{t-2}\|$	9	9	7	8	9
9	$u_{t-2} \cdot \ln\|u_{t-1}\|$	10	7	9	6	9
10	$u_{t-2}^3 \cdot \ln\|u_{t-1}\|$	10	10	9	8	9
11	$u_{t-2} \cdot (\ln\|u_{t-1}\|)^3$	10	10	10	10	9
12	$\|u_{t-2}\| \cdot e^{u_{t-1}}$	10	9	9	10	4
13	$u_{t-2} \cdot e^{u_{t-1}}$	10	10	9	10	8
14	$u_{t-2} \cdot e^{u_{t-1} - 0.03u_t}$	0	0	0	1	2
15	$\|u_t\|$	9	10	10	10	10
	No. of used data	384	320	256	192	128

Table 8.2: Results from Monte-Carlo simulations, 50 runs with balanced design, 10 in each column. ANOVA was performed on input/output data, where the input data was generated by (8.2). The number of correct results are given in different columns depending on the lowest cell count. In the bottom line the total amount of data for each test is given. The normally distributed 'measurement' noise has standard deviation 0.0001, and the level of significance for ANOVA is 0.01.

since there are fewer analysed data series. The good results for function 1 was a bit surprising, and the results for function 15 a bit worse than in previous experiments, but otherwise there is nothing new to say about these results.

8.3 Exhaustive search

Exhaustive search among neural networks with one to three regressors, as in Chapter 6, was performed with data sets of length $N = 1160$. The networks used here have 30 neurons in the hidden layer. Of 100 Monte-Carlo runs on function 15 *only 2% gave correct results.* Another Monte-Carlo simulation

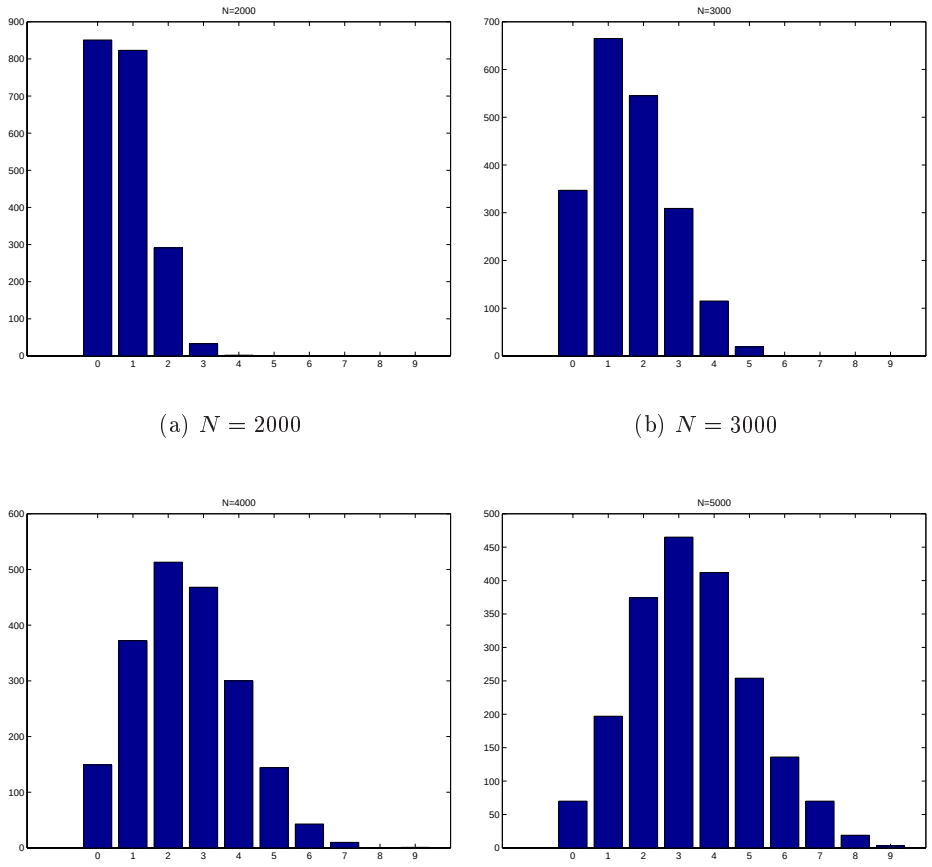
with $N = 5000$ was run with *7% correct results.* In 59% of the cases, all regressors were included. The simulations took more than two weeks computation time to finish. The sigmoidal network does not model this function very well. One reason could be that there are more neurons than necessary to build a good model for this simple function, $y_t = |u_t| + e_t$. Then it does not matter if some neurons are used to model the non-existing contributions from the other regressors.

8.4 Conclusion

The ANOVA results give us no reason to be extra cautious when a correlated input signal is used in the identification experiment, provided that all cells in the design are covered by observations. In severely unbalanced cases it can be a good idea to pick out an equal amount of data from each cell to get a balanced design. When the input signal is strongly correlated it can be hard to group the data such that no empty cells occur.

The ANOVA results are not very sensitive to the amount of measurement noise. An increase in the noise variance a factor 10^8 does not change the result, but an increase a factor 10^{12} give a too low signal-to-noise ratio to get correct results for any function. These results are obtained by adding different noise sequences with varying variance to the function values obtained by using one input signal.

The simulations show that results from exhaustive search should not be trusted.



(a) $N = 2000$ (b) $N = 3000$

(c) $N = 4000$ (d) $N = 5000$

Figure 8.2: Distribution of the lowest cell counts in 2000 data series of varying length. The bars labelled with a nine contains all data series with a lowest cell count ≥ 9.

9

Determine the structure of NARX-models

The previous chapters have all aimed at giving a sound foundation for the extension to NAR- and NARX-models. A NAR-model is given by

$$y_t = g(y_{t-1}, \dots, y_{t-k}) + e_t, \tag{9.1}$$

and a NARX-model by

$$y_t = g(y_{t-1}, \dots, y_{t-k_y}, u_t, \dots, u_{t-k_u}) + e_t. \tag{9.2}$$

As before, e_t is assumed to be additive Gaussian noise.

The following investigation was made to give some indication on how ANOVA works on nonlinear auto-regressive processes and what the difficulties are. It is not complete in any way. For the NAR-models, also some of the methods from Chapter 3 could be applied to find appropriate regressors.

9.1 Test examples

The following examples were taken from different articles treating aspects on the identification of NAR-models. The signal-to-noise ratio varies from

example to example. Most of the examples are pure autoregressive without exogenous input variables (see Equations (2.5) and (2.7)). The example setup is taken from the different papers, while the grouping and analysis is new for this thesis. Data series with 3000 input/output data are used for all examples.

9.1.1 Example 1: Chen 1

The first example NAR system is taken from Chen et al. (1995). It is a nonlinear additive autoregressive process,

$$y_t = 2e^{-0.1y_{t-1}^2}y_{t-1} - e^{-0.1y_{t-2}^2}y_{t-2} + e_t, \tag{9.3}$$

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, since it was used together with Example 2 to test algorithms for spotting additivity in NAR systems.

Grouping

The signal y_t has heavy correlation between the time lags, which makes it hard to get data in all cells, especially if many time lags are to be tested at the same time. The data was divided into six intervals:

Group	Interval
0	$[2, \infty]$
1	$[1, 2]$
2	$[0, 1]$
3	$[-1, 0]$
4	$[-2, -1]$
5	$[-\infty, -2]$

Of these, only groups 1 up to 4 were used in the analysis, following the idea of the shrunken range (see Section 5.2.3), with slightly different intervals. With this division, it is only possible to test for two time lags at a time, due to the amount of empty cells in higher dimensions.

Analysis

First, all time lags up to y_{t-8} were tested pairwise with the data in groups 1 up to 4 for both time lags. y_{t-5} to y_{t-8} could be excluded as they had no significant effect on the signal y_t, see Table 9.1. When y_{t-1} was tested together with y_{t-3} and y_{t-4} respectively, also y_{t-4} could be excluded.

Effect	p-level	groups
y_{t-1}	0	1–4
y_{t-2}	0	1–4
y_{t-3}	0.002	1–4
y_{t-4}	0.003	1–4
y_{t-5}	0.13	1–4
y_{t-6}	0.98	1–4
y_{t-7}	0.51	1–4
y_{t-8}	0.26	1–4
y_{t-1}	0	1–4
y_{t-3}	0.003	1–4
y_{t-1}	0	1–4
y_{t-4}	0.5	1–4

Table 9.1: Results from ANOVA test for Example 1. If the p-level is low enough, below 0.01, the null hypothesis for the effect is rejected. The time lags were tested pairwise. Since no two-factor interactions were found significant, they were not included in the table.

Effect	p-level	groups
y_{t-1}	0	1–4
y_{t-2}	0.003	2–3
y_{t-3}	0.93	2–3

Table 9.2: Results from ANOVA test 2 for Example 1, only main effects collected in the table.

Then another test was made for y_{t-1}, y_{t-2} and y_{t-3}. It was not possible to perform the analysis with all factor levels included for all time lags, since then some cells got empty. Instead groups 1–4 were used for y_{t-1}, while only groups 2 and 3 were used for y_{t-2} and y_{t-3}, see Table 9.2. The result is that y_{t-3} can be excluded as regressor.

Conclusion

y_{t-1} and y_{t-2} should certainly be included in further model building. Since we suspect nonlinear functions, two levels of the factor y_{t-3} could possibly be too little to draw any certain conclusions from this analysis. A careful analyst would probably also build one model with y_{t-3} included, and

postpone further regressor exclusion to the model validation phase. This analysis took two to three hours to complete. The major part of the time was spent on the grouping of the data and interpretation of the results. Each ANOVA test is completed in a few seconds.

Comparison with exhaustive search

For this function, also regressor selection with the exhaustive search method, see Section 3.2.4, was tried. Eight regressors, y_{t-1} to y_{t-8}, were tried, giving 256 network models to compare. Each network has a single hidden layer with 30 sigmoidal neurons and a linear output layer. The networks were trained with the Levenberg-Marquardt minimisation algorithm with random starting values of the parameters. Ten restarts with new random values were used for each network to give a larger probability to find a good minimum for the loss function. The data sequence was split in half to give 1500 samples for training data and 1500 samples for validation data.

It took about 5 hours to prepare a MATLAB™ script for the exhaustive search and 300 CPU hours to run it, but the results were lost, and the script had to be run a second time. The second try took 75 CPU hours to complete. One important thing to consider when the computation time is so long is what happens if there is a sudden computer shutdown. The result was that the network with y_{t-1} and y_{t-2} as inputs had best RMS values on validation data, which is the correct structure.

Networks with the suggested structure from the ANOVA tests were also trained. The first network used y_{t-1} and y_{t-2} as regressors, entering additively:

$$y_t = g_1(y_{t-1}) + g_2(y_{t-2}). \tag{9.4}$$

The second network used y_{t-1}, y_{t-2} and y_{t-3} as regressors, entering additively, which means that no interaction effects are considered in the model:

$$y_t = g_1(y_{t-1}) + g_2(y_{t-2}) + g_3(y_{t-3}). \tag{9.5}$$

The MATLAB™ script took 30 minutes to prepare and 20 minutes to run.

The networks from the different approaches were compared on a new set of data of length 3000. Their simulation performance is almost equal. The net from the exhaustive search has a slightly worse fit, RMS value 1.175, than the others. The network with two additive regressors has the RMS value 1.143, and the network with three additive regressors has the RMS value 1.145. In Figure 9.1, the simulated output from the networks with

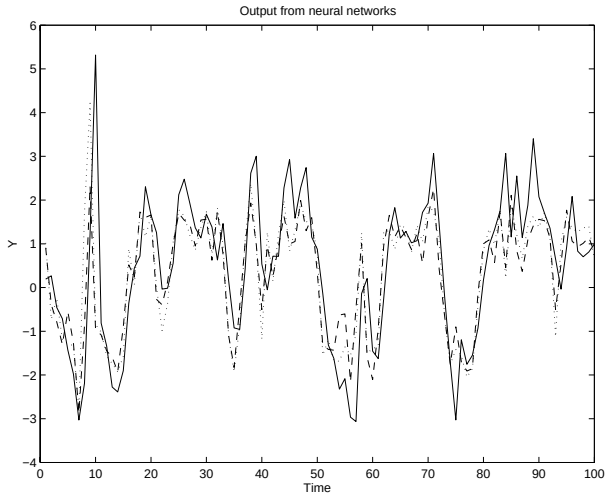


Figure 9.1: Simulated output from networks with two regressors. The dotted line corresponds to the net from the exhaustive search, the dashed line to the output from the network structure suggested by ANOVA and the solid line to the measured output.

two regressors and the real output are plotted for the first 100 data points. The results are quite good, considering that the noise added to the signal has variance 1.

9.1.2 Example 2: Chen 2

The second example, also from Chen et al. (1995), is almost the same as the first, but this is an exponential autoregressive model,

$$y_t = 2e^{-0.1y_{t-1}^2}(y_{t-1} - y_{t-2}) + e_t, \tag{9.6}$$

where e_t is Gaussian noise with standard deviation 1.

Grouping

The same grouping as for Example 1 is used. Also in this example the strong correlation makes it impossible to test more than two time lags at a time if all groups 1 to 4 should be included, since there is the problem with empty cells in higher dimensions.

Effect	p-level	groups
y_{t-1}	0	1–4
y_{t-2}	0	1–4
y_{t-3}	0.0003	1–4
y_{t-4}	0	1–4
y_{t-5}	0	1–4
y_{t-6}	0	1–4
y_{t-7}	0.55	1–4
y_{t-8}	0.35	1–4

Table 9.3: Results from ANOVA test for Example 2. If the p-level is low enough, below 0.01, the null hypothesis for the effect is rejected. The time lags were tested pairwise. Since no two-factor interactions were found significant, they were not included in the table.

Analysis

When a pairwise testing is done, only the time lags y_{t-7} and y_{t-8} can be excluded, see Table 9.3.

Since the pairwise testing could not give much information, a new analysis run with tests where three time lags were included were run. To avoid empty cells only data from groups 2 and 3 were included, giving the results in Table 9.4. The first test was made on y_{t-1}, y_{t-2} and y_{t-3}. y_{t-3} could be excluded and there was room to test another factor, y_{t-4}, which also was insignificant. Since y_{t-1} and y_{t-2} were clearly significant, they were included in all the remaining tests. No other regressors were found significant. After the second analysis the remaining time lags were y_{t-1} and y_{t-2}.

Conclusion

The second analysis round should give the suspicion that the data only depends on y_{t-1} and y_{t-2}, possibly with interaction between them, but two levels for each time lag are too few to give any confidence, since nonlinear behaviour is suspected. Possible regressors are y_{t-1} to y_{t-6}, as given by the first analysis round. Better grouping is needed.

Effect	p-level	groups
y_{t-1}	0	2–3
y_{t-2}	0.02	2–3
y_{t-3}	0.19	2–3
y_{t-1}	0	2–3
y_{t-2}	0.001	2–3
y_{t-3}	0.06	2–3
y_{t-1}	0	2–3
y_{t-2}	0	2–3
y_{t-5}	0.23	2–3
$y_{t-1} * y_{t-2}$	0.02	2–3
y_{t-1}	0	2–3
y_{t-2}	0	2–3
y_{t-6}	0.79	2–3
y_{t-1}	0	2–3
y_{t-2}	0	2–3
y_{t-7}	0.11	2–3
y_{t-1}	0	2–3
y_{t-2}	0	2–3
y_{t-8}	0.81	2–3

Table 9.4: Results from ANOVA test for Example 2. If the p-level is low enough, below 0.01, the null hypothesis for the effect is rejected. The time lags were tested three and three.

9.1.3 Example 3: Chen 3

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

$$y_t = -2y_{t-1}I(y_{t-1} \leq 0) + 0.4y_{t-1}I(y_{t-1} > 0) + e_t, \tag{9.7}$$

where e_t is Gaussian noise with standard deviation 1 and $I(x)$ is an indicator such that $I(x) = 1$ if x holds.

Grouping

The output data was divided into six intervals:

Effect	p-level	groups
y_{t-1}	0	1–4
y_{t-2}	0.57	1–4
y_{t-3}	0.93	1–4
y_{t-1}	0	1–4
y_{t-4}	0.53	1–4
y_{t-5}	0.71	1–4
y_{t-1}	0	1–4
y_{t-6}	0.60	1–4
y_{t-7}	0.48	1–4
y_{t-1}	0	1–4
y_{t-8}	0.22	1–4

Table 9.5: Results from ANOVA test for Example 3, only main effects collected in the table. If the p-level is low enough, below 0.01, the null hypothesis for the effect is rejected. The time lags were tested three and three, except in the last test.

Group	Interval
0	$[2, \infty]$
1	$[1, 2]$
2	$[0.5, 1]$
3	$[0, 0.5]$
4	$[-1, 0]$
5	$[-\infty, -1]$

which gave an approximately equal number of data in groups 1 to 4. Three time lags at a time could be analysed without empty cells when groups 1 to 4 were used.

Analysis

The first test was made with the time lags y_{t-1}, y_{t-2} and y_{t-3}. Only y_{t-1} was found significant and included in the next test. This was the case for the second and third tests too. In the fourth test only y_{t-1} and y_{t-8} were tested.

Conclusion

Only y_{t-1} was found to be a proper regressor for the analysed data series, which coincides with the true model.

9.1.4 Example 4: Chen 4

This example is similar to the previous, a threshold autoregressive model,

$$\begin{aligned} y_t &= (0.5y_{t-1} - 0.4y_{t-2})I(y_{t-1} < 0) \\ &+ (0.5y_{t-1} + 0.3y_{t-2})I(y_{t-1} \geq 0) + e_t, \end{aligned} \tag{9.8}$$

where e_t is Gaussian noise with standard deviation 1 and $I(x)$ is an indicator such that $I(x) = 1$ if x holds. This model is not additive, which means that it cannot be separated in the following manner:

$$y_t = g(y_{t-1}, y_{t-2}) = g_1(y_{t-1}) + g_2(y_{t-2}). \tag{9.9}$$

Grouping

Compared to the example in Section 9.1.3, a different grouping was used here. The data range was divided into three intervals:

Group	Interval
0	$[-\infty, 0]$
1	$[0, 1]$
2	$[1, \infty]$

The data is not as strongly correlated as in previous examples and fewer intervals are used to group the data. In this case four time lags can be tested at the same time.

Analysis

The time lags were included in the ANOVA four at a time. Beginning with y_{t-1} to y_{t-4}, four tests were needed to cover the first eight time lags. The results are given in Table 9.6. The time lags y_{t-1} and y_{t-2} were included in all tests, since they proved to give significant main and two-factor interaction effects. When y_{t-5} was included in the test, the interaction effect (y_{t-1}, y_{t-5}) proved significant, so also y_{t-5} was included in the remaining tests.

Effect	p-level	groups
y_{t-1}	0	0–2
y_{t-2}	0	0–2
y_{t-3}	0.95	0–2
y_{t-4}	0.75	0–2
(y_{t-1}, y_{t-2})	0	
y_{t-1}	0	0–2
y_{t-2}	0	0–2
y_{t-5}	0.48	0–2
y_{t-6}	0.80	0–2
(y_{t-1}, y_{t-2})	0	
(y_{t-1}, y_{t-5})	0.009	
y_{t-1}	0	0–2
y_{t-2}	0	0–2
y_{t-5}	0.72	0–2
y_{t-7}	0.43	0–2
(y_{t-1}, y_{t-2})	0	
(y_{t-1}, y_{t-5})	0.005	
y_{t-1}	0	0–2
y_{t-2}	0	0–2
y_{t-5}	0.45	0–2
y_{t-8}	0.38	0–2
(y_{t-1}, y_{t-2})	0	
(y_{t-1}, y_{t-5})	0.004	

Table 9.6: Results from ANOVA test for Example 4. All significant effects, at $\alpha = 0.01$, and all main effects are collected in the table. If the p-level is low enough, below 0.01, the null hypothesis for the effect is rejected. The time lags were tested four and four.

Conclusion

The results from the ANOVA indicates that the data comes from a model on the form:

$$y_t = g_1(y_{t-1}, y_{t-2}) + g_2(y_{t-1}, y_{t-5}) + e_t, \tag{9.10}$$

where $g_1(y_{t-1}, y_{t-2})$ probably explains most of y_t.

9.1.5 Example 5: Chen 5

Next out is a functional-coefficient AR(1) model with a sine function of lag two,

$$y_t = y_{t-1} \sin(y_{t-2}) + e_t, \tag{9.11}$$

where e_t is Gaussian noise with standard deviation 1 (Chen et al., 1995).

Grouping

The range of y_t was divided into three intervals:

Group	Interval
0	$[1, \infty]$
1	$[-1, 1]$
2	$[-\infty, -1]$

This grouping makes it possible to test three time lags at a time.

Analysis

First, the time lags y_{t-1} to y_{t-3} were tested. The interaction effect $(y_{t-1}, y_{t-2}, y_{t-3})$ was significant. It was not possible to include any more factors in the test due to empty cells, so y_{t-4} to y_{t-6} were included in the next test. Here, the interaction effect (y_{t-4}, y_{t-5}) was significant, so y_{t-4} and y_{t-5} were included also in the following tests, see Table 9.7.

Conclusion

The model resulting from the ANOVA should have the following structure:

$$y_t = g_1(y_{t-1}, y_{t-2}, y_{t-3}) + g_2(y_{t-4}, y_{t-5}) + e_t. \tag{9.12}$$

As this is a rather big model, it could be worth the effort to collect more data and see if it was just an unlucky data sequence that led to the large number of regressors.

Comparison with exhaustive search

Exhaustive search with eight regressors, y_{t-1} to y_{t-8}, was run on two different data sets with 3000 data from this function. The results, after 75 CPU hours each, was that for the first data set the model with the correct

Effect	p-level	groups
y_{t-1}	0.06	0–2
y_{t-2}	0.08	0–2
y_{t-3}	0	0–2
(y_{t-1}, y_{t-2})	0	0–2
$(y_{t-1}, y_{t-2}, y_{t-3})$	0	
y_{t-4}	0.25	0–2
y_{t-5}	0.56	0–2
y_{t-6}	0.09	0–2
(y_{t-4}, y_{t-5})	0	
y_{t-4}	0.64	0–2
y_{t-5}	0.68	0–2
y_{t-7}	0.12	0–2
(y_{t-4}, y_{t-5})	0	
y_{t-4}	0.24	0–2
y_{t-5}	0.96	0–2
y_{t-8}	0.33	0–2
(y_{t-4}, y_{t-5})	0	

Table 9.7: Results from ANOVA test for Example 5. All significant effects, at $\alpha = 0.01$, and all main effects are collected in the table. If the p-level is low enough, below 0.01, the null hypothesis for the effect is rejected. The time lags were tested three and three.

regressors y_{t-1} and y_{t-2}, had best performance on the part of the data used for validation. For the second data set a model with the regressors y_{t-1}, y_{t-2}, y_{t-4} and y_{t-6} was best. It seems like there is not enough information in the input/output data to do regressor selection with useful results.

9.1.6 Example 6: Chen and Lewis

This example is an adaptive spline threshold auto-regression, which exhibits limiting cycle behaviour (Chen et al., 1995; Lewis and Stevens, 1991). The model is:

$$\begin{aligned} y_t &= 14.27 + 0.46y_{t-1} - 0.02y_{t-1}(y_{t-2} - 30)_+ \\ &+ 0.047y_{t-1}(30 - y_{t-2})_+ + e_t, \end{aligned} \tag{9.13}$$

where $(x)_+ = x$ if $x > 0$ and $(x)_+ = 0$ otherwise and e_t is Gaussian noise with standard deviation 1.

Effect	p-level	groups
y_{t-1}	0	0–2
y_{t-2}	0	0–2
y_{t-3}	0.0006	0–2
y_{t-4}	0.22	0–2
y_{t-5}	0	0–2
y_{t-6}	0.63	0–2
y_{t-5}	0	0–2
y_{t-7}	0.28	0–2
y_{t-8}	0.38	0–2
y_{t-1}	0	0–2
y_{t-2}	0	0–2
y_{t-3}	0.003	0–2
y_{t-5}	0.17	0–2

Table 9.8: Results from ANOVA test for Example 6. All significant effects, at $\alpha = 0.01$, and all main effects are collected in the table. If the p-level is low enough, below 0.01, the null hypothesis for the effect is rejected. The time lags were tested three and three.

Grouping

The range of the data was divided into three intervals:

Group	Interval
0	$[-\infty, 27.6]$
1	$[27.7, 28.8]$
2	$[28.8, \infty]$

In this data sequence, the correlation between every second sample is strong, as can be seen from the distribution in the different cells. For example, the cells corresponding to data belonging to groups 0, 2, 0, 2 and to groups 2, 0, 2, 0 are empty.

Analysis

First, the time lags were tested three at a time. y_{t-1} to y_{t-3} all had significant main effects, so the next test was performed on y_{t-4} to y_{t-6}. y_{t-5} had a significant main effect in this test, so the third test was made on y_{t-5}, y_{t-7} and y_{t-8}, see Table 9.8. Then, y_{t-1} to y_{t-3} and y_{t-5}, were tested

together. As the time lags are not all adjacent, this was possible, despite the strong correlation. This last test showed that y_{t-5} is spurious. The difference compared to the former tests depend on the fact that the real regressors are included in the same test.

Conclusion

The resulting model is

$$y_t = g_1(y_{t-1}) + g_2(y_{t-2}) + g_3(y_{t-3}) + e_t. \tag{9.14}$$

The interaction effect (y_{t-1}, y_{t-2}) has not been picked up by the test.

9.1.7 Example 7: Yao

This example is NARX model structure (see Equation (2.5)). The model was found in Yao and Tong (1994).

$$y_t = 0.3y_{t-1}e^{u_{t-1}} + \sin(u_{t-1}) + e_t, \tag{9.15}$$

where u_t is an AR(2) model:

$$u_t = 0.1u_{t-1} - 0.56u_{t-2} + n_t. \tag{9.16}$$

The noise term e_t has the same distribution as the noise term $0.6n_t$. n_t is the sum of 48 independent uniformly distributed random variables, in the range $[-0.25, 0.25]$. According to the central limit theorem the noise terms can then be treated as coming from a Gaussian distribution, but with the support bounded to $[-12, 12]$.

Grouping

The range of y_t was divided into the intervals:

Group	Interval
0	$[-\infty, -1.4]$
1	$[-1.4, 1]$
2	$[1, \infty]$

and the range of u_t was divided into the intervals:

Group	Interval
0	$[-\infty, -1]$
1	$[-1, 1]$
2	$[1, \infty]$

With this grouping, four factors at a time could be tested.

Effect	p-level	groups
y_{t-1}	0.016	0–2
y_{t-2}	0.12	0–2
y_{t-3}	0.83	0–2
y_{t-4}	0.87	0–2
y_{t-1}	0	0–2
y_{t-5}	0.93	0–2
u_t	0.21	0–2
u_{t-1}	0.017	0–2
(y_{t-1}, u_t)	0.008	-
(y_{t-1}, u_{t-1})	0	-
(y_{t-1}, u_t, u_{t-1})	0.008	-
y_{t-1}	0	0–2
u_t	0.68	0–2
u_{t-1}	0.39	0–2
u_{t-2}	0.23	0–2
(y_{t-1}, u_{t-1})	0	-
(y_{t-1}, u_{t-2})	0	-
$(y_{t-1}, u_{t-1}, u_{t-2})$	0	-

Table 9.9: Results from ANOVA test for Example 7, tests 1 to 3. All significant effects, at $\alpha = 0.01$, and all main effects are collected in the table. If the p-level is low enough, below 0.01, the null hypothesis for the effect is rejected. The time lags were tested four and four.

Analysis

Six tests were necessary to cover y_{t-1} to y_{t-5} and u_t to u_{t-5}, see Tables 9.9 and 9.10. In the first test, which was made on y_{t-1} to y_{t-4}, only y_{t-1} was found to have a significant effect. In the second test, also u_t and u_{t-1} were significant, in interactions. When u_{t-2} was included in the third test, u_t lost its importance. No other time lags had significant effects. The normal probability plots of the residuals show some non-normal behaviour, which indicates that the the analysis should not be completely trusted. With more data, some cells with large within-cell variation could be excluded to do something about the non-normal residuals.

Effect	p-level	groups
y_{t-1}	0	0–2
u_{t-1}	0.27	0–2
u_{t-2}	0.19	0–2
u_{t-3}	0.83	0–2
(y_{t-1}, u_{t-1})	0	-
(y_{t-1}, u_{t-2})	0	-
$(y_{t-1}, u_{t-1}, u_{t-2})$	0	-
y_{t-1}	0	0–2
u_{t-1}	0.01	0–2
u_{t-2}	0.01	0–2
u_{t-4}	0.44	0–2
(y_{t-1}, u_{t-1})	0	-
(y_{t-1}, u_{t-2})	0	-
(u_{t-1}, u_{t-2})	0.005	-
$(y_{t-1}, u_{t-1}, u_{t-2})$	0	-
y_{t-1}	0	0–2
u_{t-1}	0.02	0–2
u_{t-2}	0.009	0–2
u_{t-5}	0.71	0–2
(y_{t-1}, u_{t-1})	0	-
(y_{t-1}, u_{t-2})	0	-
$(y_{t-1}, u_{t-1}, u_{t-2})$	0	-

Table 9.10: Results from ANOVA test for Example 7, tests 4 to 6. All significant effects, at $\alpha = 0.01$, and all main effects are collected in the table. If the p-level is low enough, below 0.01, the null hypothesis for the effect is rejected. The time lags were tested four and four.

Conclusion

The resulting model is:

$$y_t = g(y_{t-1}, u_{t-1}, u_{t-2}) + e_t. \tag{9.17}$$

The time lag u_{t-2} is spurious, and was possibly tested significant due to its importance in explaining u_t.

9.1.8 Example 8: Pi

This example is a Hénon map (Pi and Peterson, 1994),

$$y_t = 1 - 1.4(y_{t-2} - e_{t-2})^2 + 0.3(y_{t-4} - e_{t-4}) + e_t, \tag{9.18}$$

where e_t is i.i.d. noise uniformly distributed on $[-0.10122, 0.10122]$. In simulations this model can be viewed as an autoregressive model with or without exogenous variables, depending on whether e_t is treated as an input signal or noise. For all examples, 3000 input/output data were used.

Grouping

The range of the data was divided into three intervals:

Group	Interval
0	$[-\infty, -0.16]$
1	$[-0.16, 0.64]$
2	$[0.64, \infty]$

The data was strongly correlated, see Figure 5.4, which led to problems with empty cells. At most three time lags could be tested at the same time.

Analysis

Different numbers of groups had to be used for the different factors in each test, to avoid empty cells. The tests were done according to Table 9.11. Every second time lag seem to have a strong influence on y_t. They are also strongly dependent on each other, which can be concluded from the number of groups included in each test. When not all groups are included, it is to avoid empty cells.

Conclusion

The result from the tests is that y_{t-2}, y_{t-4}, y_{t-6}, y_{t-8} and y_{t-10} influence y_t. This result should not be trusted, due to the low number of groups included in the tests and the strong correlation between these time lags. The only useful result is that all odd time lags can be excluded from further model building.

Effect	p-level	groups
y_{t-1}	0.35	1–2
y_{t-2}	0	0–2
y_{t-3}	0.08	0–2
y_{t-2}	0	1–2
y_{t-4}	0	0–2
y_{t-5}	0.97	0–2
(y_{t-2}, y_{t-4})	0	-
y_{t-2}	0	1–2
y_{t-4}	0	1–2
y_{t-6}	0	1–2
(y_{t-2}, y_{t-4})	0	-
y_{t-2}	0	1–2
y_{t-4}	0	0–2
y_{t-7}	0.84	0–2
(y_{t-2}, y_{t-4})	0	-
y_{t-8}	0	1–2
y_{t-9}	0.11	0–2
y_{t-10}	0	0–2
(y_{t-8}, y_{t-10})	0	-
y_{t-2}	0	1–2
y_{t-8}	0	1–2
y_{t-10}	0	0–2
$(y_{t-2}, y_{t-8}, y_{t-10})$	0	-

Table 9.11: Results from ANOVA test for Example 8. All significant effects, at $\alpha = 0.01$, and all main effects are collected in the table. If the p-level is low enough, below 0.01, the null hypothesis for the effect is rejected. The time lags were tested three and three.

9.2 Discussion

For the NARX models, ANOVA manages to pick out at least the true regressors. Spurious regressors are included quite often. In one case, Example 6, interaction effects were missed, but all true regressors were included. In the examples where exhaustive search was tried, there was no indication that it would perform any better than ANOVA.

The main problem with the analysis is the grouping that has to be done. It seems like it is better to divide the data into few groups than into many.

With the strong correlation between time lags, natural to autoregressive processes, there are still problems to get observations into all cells. The strong correlation also leads to the inclusion of spurious regressors, since it is not possible to test all possible regressors at the same time with limited data.

One idea that came up during the experiments is that if it is possible to introduce more noise in the NAR processes during data collection or excite NARX processes with the input signal, more information could be extracted from the measurement data when using ANOVA for the analysis. The more the signal jumps around in the regressor space, the easier it gets to get data covering all cells.

So, what information has been gained? The grouping is crucial to good analysis of NAR and NARX models. It is not likely that ANOVA will miss contributing regressors, but spurious ones could be included. In most cases, that will be an improvement and lead to more sparse models than without any tests.

9.3 Open questions

There are some open questions that need further study.

Is it the strong correlation between different regressors itself that leads to spurious regressors in the results or does it depend on the unbalanced design of the tests?

Is there any good way to group the data? How much does the signal-to-noise ratio influence the possibility to get a good grouping?

More NARX models should be investigated to be able to draw any conclusions on how ANOVA works for that type of systems. It would also be interesting to compare ANOVA results with other methods to find proper regressors for NAR-models.

10

Conclusions

The aim of this work was to find a good method to select regressors for nonlinear system identification. To begin with, a literature survey over possible methods to select the model structure for nonlinear systems, mainly autoregressive processes, has been done. The main ideas were:

1. Compare estimated models, using different regressor vectors, with each other.

2. Compare the variability of the output data, given one regressor vector, with the variability of the output data given other regressor vectors.

The second idea has been investigated further by applying a common statistical tool, analysis of variance, to system identification applications. This method differs from most of the suggested methods by treating the variability in a stochastic framework, instead of treating the problem from a geometrical point of view. An investigation of the properties of analysis of variance (ANOVA), practical considerations with its use and Monte-Carlo simulations covering several aspects of the use of ANOVA in system identification applications has been done. The result of this work is the suggested

procedure for selecting a model structure from input/output data, given below.

Suggested procedure

1. Start with designing the input signal. Decide what range it should have, depending on what input signal range the model should be valid for, and the practical limits of the system. Select the sampling rate (Ljung, 1999). The best choice of input signal when ANOVA will be used is a pseudo-random multi-level signal, which can be constructed with a multi-level shift register (Godfrey, 1993). For nonlinear system identification, some physical insight is needed to select the best levels, but if at least three levels are used, many nonlinearities will be covered. The required length of the signal depends on what regressors will be included in the tests, see step 3. The second best choice of input signal is a random signal, see Chapter 7. If neither of these choices can be made, normal operations data can be used if there is enough variability of the input signal. This can be checked by trying a few different groupings of the data, see Section 5.2. Watch out for input signals with strong correlation — they seldom have enough variability.

2. The step which is most open to further research is the grouping of the data. The grouping has to be done to get something to associate the factor levels in ANOVA with. For pseudo-random multi-level signals this is easy: one input signal level = one group. For other kinds of signals, it is necessary to experiment with the grouping intervals, trying to get an equal amount of data in each cell. For examples, see Section 5.2 and Chapter 9. For nonlinear systems, at least three groups for each regressor are needed, but if only linear systems are considered, it is enough with two groups. For many signals, the grouping leads to an unbalanced design for ANOVA, which can affect the sensitivity to outliers badly (Section 5.1). This can be avoided by discarding excess data at random in the cells with most data (Section 5.2.5).

3. Decide which regressors should be tested for effects on the output. The possible regressors could be different input signals with several different time lags and/or different time lags of old output signals. Semi-physical modelling can also be used, which means that any partial physical insight of the model structure should be taken advantage of. If, e.g., it is likely that the power (current times voltage) should

be a regressor for the model, this product can be constructed and tested like the other possible regressors. Since such products affect the distribution of the measurement noise, an extra careful check of the assumptions is necessary.

It is best if all possible regressors can be tested at the same time. This puts quite strong demands on the variability of the input signal(s). Also, the chosen grouping of the data affects the amount of possible regressors that can be tested at the same time. The limitation is that all cells, given by that the values of each regressor belong to specific intervals, have to include at least one observation to perform a test. At least two observations are needed if all interaction effects (Definition 2.2) should be tested. If there are empty cells, concentrated to one level of one regressor, the entire block of cells, corresponding to this level, can be excluded from the analysis, see, e.g., Section 9.1.1. This can sometimes make it possible to include more regressors in the same test. If it is not possible to include all regressors in the same test, the procedure in Section 5.3 can be used. The drawbacks of that procedure are that not all interaction effects can be tested (those of higher order) and that spurious regressors can be found significant if there is correlation between the regressors in different tests.

Compute the sums of squares and the F-distributed test variables as in Section 4.1. A suitable computation tool is needed for data sets including more than a few data. For example, MATLAB^{TM} or any tool for statistical computations can be used. The result from the computations is an analysis of variance table (Table 4.2 on page 31). In the table, the probability levels of the null hypotheses corresponding to each interaction effect are stated. If the probability levels are below the chosen significance level (often 0.01–0.05), the null hypothesis is rejected. Start reading the table from the bottom with the interaction effect of highest order. For each accepted null hypothesis, the full model including all regressors, can be reduced, either by excluding regressors or by finding less complex inner structure. It is not meaningful to check for significant effects of a lower interaction order if an interaction of higher order (with the same regressors) is included in the model. Check that the assumptions are valid. These are that the measurement noise is Gaussian with constant variance. The most important checks are the normal probability plot of the residuals (Section 4.1.1) and the within-cell standard deviations, which should be approximately equal. If the assumptions are not valid, the situation

can sometimes be saved by the methods used in Section 7.3.

4. The test result is a model structure. Left to consider is what model type should be used, estimate the parameters and validate the model (Ljung, 1999). The given model structure both gives which regressors should be used and what interaction pattern they have. This information can be used to build a sparse model, with the parameters in places where they can do most good. Not that many parameters will be used to estimate non-existing relations between input signals and output signal, as if a full model structure had been used. The input signal used for the analysis of variance tests can be reused for parameter estimation, since only a small fraction of its information content has been used.

Benefits by using ANOVA for model structure selection

In contrast to methods which compare estimated models in a structured (or non-structured) manner, ANOVA has the following benefits:

- The computations are fast and straightforward, without iterations or minimisations.
- ANOVA is easy to interpret.
- The method is reliable: ANOVA seldom fails to give correct results (some exceptions among the NARX-models in Chapter 9), and when it does, warning signals are given as failing assumption checks.
- ANOVA is useful for varying kinds of input signals.
- ANOVA puts demands on the information contents in the input/output data, which will make further identification easier.
- ANOVA is not critically sensitive to the amount of measurement noise.

Finally, let us reiterate some results from the Monte-Carlo simulations for the example function $y_t = |u_t| + e_t$. With a pseudo-random multi-level signal with 256 input/output data, ANOVA finds the correct regressor in 100% of the cases, while the exhaustive search method (Section 3.2.4) manages 73% of the cases (Table 6.2). The difference is more pronounced with a correlated input signal of length 5000, where ANOVA finds the correct regressor for 98% of the cases (Table 8.1) in approximately ten minutes and exhaustive search only manages 7% (Section 8.3) in approximately 200 hours.

Bibliography

Akaike, H. (1981). Modern development of statistical methods. In Eykhoff, P., editor, *Trends and Progress in System Identification*. Pergamon Press, Elmsford,N.Y.

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

Autin, M., Biey, M., and Hasler, M. (1992). 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.

Billings, S. (1980). Identification of nonlinear systems - a survey. In *IEE Proc. D.*, volume 130, pages 193–199.

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

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

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

Breiman, L. (1993). Hinging hyperplanes for regression, classification, and function approximation. *IEEE Transactions on Information Theory*, 39(3):999–1013.

Brown, M. and Harris, C. (1994). *Neurofuzzy Adaptive Modeling and Control*. Prentice Hall, New York.

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

Chen, R. and Tsay, R. S. (1993). Nonlinear additive ARX models. *Journal of the American Statistical Association*, 88:955–967.

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

Chua, L. and Kang, S. (1977). Section-wise piecewise-linear functions: Canonical representation, properties and applications. In *Proceedings of the IEEE*, volume 65, pages 915–929.

De Boor, C. (1978). *Practical Guide to Splines*. Springer Verlag, New York.

Dennis, Jr., J. and Schnabel, R. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Upper Saddle River, N.J.

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

Gray, G. J., Murray-Smith, D. J., Li, Y., Sharman, K., and Weinbrenner, T. (1998). Nonlinear model structure identification using genetic programming. *Control Engineering Practice*, 6:1341–1352.

Haber, R. and Unbehauen, H. (1990). Structure identification of nonlinear dynamic systems - a survey on input/output approaches. *Automatica*, 26:177–202.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Linear Models.* Chapman and Hall, London.

Haykin, S. (1994). *Neural Networks - A Comprehensive Foundation.* Macmillan College Publishing Company, New York.

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

Hocking, R. R. (1984). *The Analysis of Linear Models.* Brooks/Cole, Monterey.

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

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

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

Krishnaswami, V., Kim, Y. W., and Rizzoni, G. (1995). A new model order identification algorithm with application to automobile oxygen sensor modeling. In *Proceedings of the American Control Conference*, pages 2113–2117.

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

Kung, S. (1993). *Digital Neural Networks.* Prentice Hall, New Jersey.

Leontaritis, I. and Billings, S. (1985). Input-output parametric models for nonlinear systems. *International Journal of Control*, 41:303–344.

Lewis, P. A. W. and Stevens, J. G. (1991). Nonlinear modeling of time-series using multivariate adaptive regression splines (MARS). *Journal of The American Statistical Association*, 86:864–877.

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

Mehra, R. (1979). Nonlinear system identification. In *Proc. 5th IFAC Symp. Identification and System Parameter Estimation*, volume paper S-4, pages 77–85, Darmstadt, FRG. Pergamon Press, New York.

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

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

Nadaraya, E. (1964). On estimating regression. *Thory of Prob. and Applic.*, 9:141–142.

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

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

Pucar, P. and Sjöberg, J. (1998). On the hinge finding algorithm for hinging hyperplanes. *IEEE Trans. on Information Theory*, 44(3):1310–1319.

Rankin, N. O. (1974). The harmonic mean method for one-way and two-way analysis of varance. *Biometrika*, 61:117–122.

Rao, C. (1973). *Linear Statistical Inference and its Applications*. Wiley, New York.

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

Scheffé, H. (1959). *The Analysis of Variance*. John Wiley & Sons, New York.

Schumaker, L. (1981). *Spline Functions: Basic Theory*. Wiley, Chichester.

Searle, S. R. (1971). *Linear Models*. Wiley, New York.

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

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

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

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

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

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

Wang, L. (1994). *Adaptive fuzzy Systems and Control: Design and Stability Analysis*. Prentice Hall, New Jersey.

Watson, G. (1969). Smooth regression analysis. *Sankhya, Ser. A*, 26:359–372.

Yao, Q. and Tong, H. (1994). On subset selection in non-parametric stochastic regression. *Statistica Sinica*, 4:51–70.

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