Modelling Non-linear and Non-stationary Time Series

Henrik Madsen and Jan Holst

Summer school 2018 DTU - CITIES and NTNU - ZEN:

Time series analysis - with a focus on modelling and forecasting in energy systems

August 26, 2018

Contents

Chapter 1

Models for non-linear time series

The modeling of linear and non-linear time series starts with a choice of structure. In the linear case this leads to determination of system order, delays for the possible external signals and structure indices when the system is multivariate, cf. e.g. [**?**] for a thorough discussion. In the non-linear case, the question of structure for the system is much more open, giving a very large variety of possibilities. In this chapter parametric as well as non-parametric system representations are presented, for the moment leaving out the semi-parametric variants.

The time parameter may be discrete or continuous. In this chapter the presentation is confined to discrete time, postponing the discussion on continuous-time processes to Chapter **??**.

1.1 Non-linear vs. linear model building

The aim of the modeling effort may be generally expressed as:

General model: For a given (stationary) time series $\{X_t\}$ find a function h such that $\{\epsilon_t\}$ defined by

$$h(\{X_t\}) = \epsilon_t \tag{1.1}$$

is a sequence of independent random variables.

If the 'residuals', $\{\epsilon_t\}$, are not independent, there is a structure in the variability of $\{X_t\}$ that not is described by the model, and which should be persued to improve the modelling.

A model which is globally invertible, i.e. by which it is possible to determine the original time series $\{X_t\}$ from a sequence of residuals $\{\epsilon_t\}$ and a given initial value, is the best possible model for the time series $\{X_t\}$.

In practice there are various different approaches in finding the function h:

- To use a black box approach using either a parametric approach, where the parameters in a number of possible non-linear models are fitted to the time series (e.g. in a SETAR, FAR, BL or NEAR model, all acronyms to be defined below) or using a non parametric approach, like Volterra series or Neural Nets.
- To use a grey box approach where information from other sources about the system to be modeled, e.g. basic relations from physics, is combined with information obtained by statistical methods from measurements on the system.
- Use only externally obtained information from e.g. physics to describe the generating system (if well-established relations exists). In this case statistical methods are obsolete, hence this type of white box modeling will no more be considered.

In the special case with a *linear model*, $\{\epsilon_t\}$ is defined by

$$\sum_{-\infty}^{\infty} h_{t,k} X_{t-k} = \epsilon_t, \tag{1.2}$$

where $\{h_{t,k}\}$ is a weighting function which determines at time t the influence of the process value at time $t-k$ in the formation of the actual residual. In practice, one would usually assume that the system is *causal*, i.e. $h_k = 0$ for $k < 0$.

The parametric black box approach when working with a linear structure on a stationary system then leads to fitting an ARMA model. If the generating system is non stationary, the model might be of ARIMA, or seasonal type. The literature on modeling of linear time series is abundant, see e.g. [Box & Jenkins 1976], [**?**], [Madsen 1995*b*] or [Olbjer, Holst & Holst 1994].

1.2 Non-parametric descriptions

In the linear system case, the modeling of a stationary system may be performed in the time or in the frequency domain. In the non-linear case this twofold possibility do in general not exist and the concept of *transfer function* is therefore in general not defined.

However, when the system modeling starts from a non-parametric representation of the system with a Volterra series, an infinite sequence of generalized transfer functions may be defined. The resulting inference problem will not be treated.

1.2.1 Volterra Series Expansions

The work on Volterra based modelling is early attempt to systematically model non-linear system on an input-output form, starting by [Volterra 1930] and with important contributions given also in the middle of the last century, cf. [Wiener 1958]. Suppose the model is *causal*, i.e. we seek a function h such that

$$h(X_t, X_{t-1}, \dots) = \epsilon_t. \tag{1.3}$$

Suppose also that the model is *causally invertible*, i.e. (1.3) may be 'solved' such that we may write

$$X_t = h'(\epsilon_t, \epsilon_{t-1}, \dots). \tag{1.4}$$

Furthermore, suppose that h' is sufficiently well-behaved so that it can be expanded in a Taylor series. Then we can formally expand the right hand side of (1.4):

$$\begin{aligned} X_t &= \mu + \sum_{k=0}^{\infty} g_k \epsilon_{t-k} + \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} g_{kl}\epsilon_{t-k}\epsilon_{t-l} \\ &\quad + \sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty} g_{klm}\epsilon_{t-k}\epsilon_{t-l}\epsilon_{t-m} + \dots \end{aligned} \tag{1.5}$$

where

$$\mu = h'(0),\ g_k = (\frac{\partial h'}{\partial \epsilon_{t-k}}),\ g_{kl} = (\frac{\partial^2 h'}{\partial \epsilon_{t-k}\partial \epsilon_{t-l}}),\ \text{etc.} \tag{1.6}$$

This is called the *Volterra series* for the process $\{X\}$. The sequences $g_k, g_{kl}, \dots$ are called the *kernels* of the Volterra series. It can be shown that any continuous functional over a finite time interval can be arbitrarily well approximated in a Volterra expansion, but also that it does not allow a representation of discontinuities, jump phenomena or limit cycles, cf. e.g. [Brockett 1977] and the references therein. The first two terms in (1.5) correspond to a linear causally invertible model.

1.2.2 Generalized Transfer Functions

The kernel based description in (1.5) is the basis for the derivation of a transfer function concept. Let U_t and X_t denote the input and the output of a non-linear system respectively. By using the Volterra series representation of the dependence of $\{X_t\}$ of $\{U_t\}$ and omitting any disturbance (or regarding them as a possible input signal), we get

$$\begin{aligned} X_t &= \mu + \sum_{k=0}^{\infty} g_k U_{t-k} + \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} g_{kl} U_{t-k}U_{t-l} \\ &\quad + \sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty} g_{klm} U_{t-k}U_{t-l}U_{t-m} + \dots \end{aligned} \tag{1.7}$$

$$\mu = h'(0),\ g_k = (\frac{\partial h'}{\partial \epsilon_{t-k}}),\ g_{kl} = (\frac{\partial^2 h'}{\partial \epsilon_{t-k}\partial \epsilon_{t-l}}),\ \text{etc.} \tag{1.8}$$

Recall that for a stationary *linear system* the *transfer function* is defined as

$$H(\omega) = \sum_{k=0}^{\infty} g_k e^{-i\omega k} \tag{1.9}$$

and it is completely characterizing the system. For linear systems it is furthermore well known that

L1 if the input is a single harmonic $U_t = A_0 e^{i\omega_0 t}$ then the output is a single harmonic of *the same frequency* but with the amplitude scaled by $|H(\omega_0)|$ and the phase shifted by $\arg H(\omega_0)$.

L2 due to the linearity, the *principle of superposition* is valid, and the total output is the sum of the outputs corresponding to the individual frequency components of the input. Hence the system is complete described by knowing the response to all frequencies – that is what the transfer function supplies.

For *non-linear systems*, however, neither of the properties (L1) or (L2) holds. More specifically we have that

NL1 for an input with frequency ω_0, the output will, in general, contain also components at the frequencies $2\omega_0, 3\omega_0, \dots$ (*frequency multiplication*).

NL2 for two inputs with frequencies ω_0 and ω_1, the output will contain components at frequencies $\omega_0, \omega_1, (\omega_0 + \omega_1)$ and all harmonics of the frequencies (*intermodulation distortion*).

Hence, in general there is *no such thing as a transfer function* for non-linear systems. However, an *infinite sequence of generalized transfer functions* may be defined as:

$$\begin{aligned} H_1(\omega_1) &= \sum_{k=0}^{\infty} g_k e^{-i\omega_1 k} && (1.10)\\ H_2(\omega_1,\omega_2) &= \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} g_{kl} e^{-i(\omega_1 k+\omega_2 l)} && (1.11)\\ H_3(\omega_1,\omega_2,\omega_3) &= \sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty} g_{klm} e^{-i(\omega_1 k+\omega_2 l+\omega_3 m)} && (1.12)\\ &\vdots \end{aligned}$$

In order to get a frequency interpretation of this sequence of functions, consider the input U_t to be a stationary process with *spectral representation* :

$$U_t = \int_{-\pi}^{\pi} e^{it\omega} dZ_U(\omega). \tag{1.13}$$

Using the Volterra series in (1.7) we may write the output as

$$\begin{aligned} X_t &= \int_{-\pi}^{\pi} e^{it\omega_1} H_1(\omega_1) dZ_U(\omega_1) \\ &\quad + \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} e^{it(\omega_1+\omega_2)} H_2(\omega_1,\omega_2) dZ_U(\omega_1) dZ_U(\omega_2) \\ &\quad + ... \end{aligned} \tag{1.14}$$

When U_t is single harmonic, say $U_t = A_0 e^{i\theta_0 t}$, then $dZ_U(\omega) = A_0$, $\omega = \theta_0$ and zero otherwise. Hence Eq. (1.14) becomes

$$X_t = A_0 H_1(\theta_0) e^{i\theta_0 t} + A_0^2 H_2(\theta_0,\theta_0) e^{2i\theta_0 t} + \ldots \tag{1.15}$$

The output thus consists of components with frequencies $\theta_0, 2\theta_0, 3\theta_0, \ldots$, which demonstrates the assertion (NL1). The same signal representation, equation (1.13), may be used to illustrate (NL2). Hence, when working with non-linear systems, there is no simple and useful connection between the time- and frequency representations of the system as is the case for linear systems.

1.3 Non-linear parametric models

In the previous section non-parametric methods for non-linear time series were discussed. In this section an overview of some of the most important parametric non-linear models is given.

1.3.1 The (linear) ARMA Model

Many of the non-linear models take the basic linear ARMA(p, q) model as their point of departure for different types of extensions. The linear model is :

$$X_t = \theta_0 + \sum_{i=1}^{p} \theta_i X_{t-i} + \sum_{i=1}^{q} \phi_i \epsilon_{t-i} + \epsilon_t, \tag{1.16}$$

where ϵ_t is a strict white noise process, i.e. a sequence of independent zero mean random variables with variance σ_ϵ^2.

1.3.2 Threshold Models

This is a very rich class of models, which have been discussed e.g. by Tong in a monograph from 1983, [Tong 1983], and the book from 1990, [Tong 1990*b*]. It has shown to be useful if, for instance, the dynamical behavior of the system depends on the actual state or process value, e.g. by approximating the dynamics in some *regimes* by 'simple' models (usually linear). *Threshold*-values determine the actual regime (or mix of regimes). Below some versions of models with thresholds are listed and named. The analysis of the probabilistic structure of these types of models including e.g. discussions of stationarity and stability is in general very complicated, early results on stochastic stability may be found in [Kushner 1971] and more recently in [Tong 1990*b*].

Self-Exciting Threshold AR (SETAR)

Define intervals $R_1, \ldots, R_l$ such that $R_1 \cup \cdots \cup R_l = \mathbb{R}$ and $R_i \cap R_j = \emptyset \; \forall i, j$. Each interval R_i is given by $R_i =]r_{i-1}; r_i]$, where $r_0 = -\infty$ and $r_1, \ldots, r_{l-1} \in R$ and $r_l = \infty$. The values $r_0, \ldots, r_l$ are called *thresholds*.

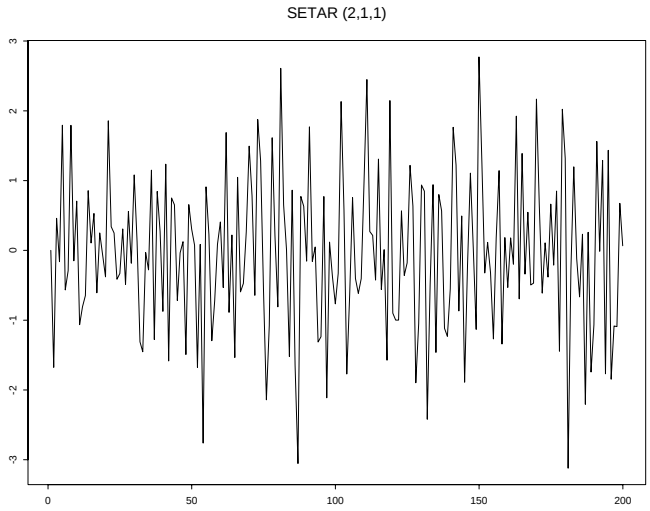

Figure 1.1: A simulation of a SETAR model with two regimes, SETAR(2;1;1)

The SETAR$(l; d; k_1, k_2, \ldots, k_l)$ model is given by :

$$X_t = a_0^{(J_t)} + \sum_{i=1}^{k_{J_t}} a_i^{(J_t)} X_{t-i} + \epsilon_t^{(J_t)}, \tag{1.17}$$

where

$$J_t = \begin{cases} 1 & X_{t-d} \in R_1 \\ 2 & X_{t-d} \in R_2 \\ \vdots & \vdots \\ l & X_{t-d} \in R_l \end{cases}. \tag{1.18}$$

The parameter d is the *delay parameter*. Hence the model has l regimes, a delay-parameter d, and in the j'th regime the process is simply an AR-process of order k_j.

If the AR-processes all have the same order k we often write SETAR$(l; d; k)$.

A simulation of the SETAR(2;1;1)-model

$$X_t = \begin{cases} 1.0 + 0.6X_{t-1} + \epsilon_t & \text{for } X_{t-1} \leq 0 \\ -1.0 + 0.4X_{t-1} + \epsilon_t & \text{for } X_{t-1} > 0 \end{cases},$$

where $\epsilon_t \in$ N(0,1) is shown in figure 1.1. Note that it is not too easy (?) to see the separation of the process performance into different regimes in the figure. The estimation of parameters and structure in this type of models is treated in Chapter **??**.

Self-Exciting Threshold ARMA (SETARMA)

The SETARMA model is an obvious generalization to different ARMA models in the different regimes : SETARMA$(l; d; k_1, \ldots, k_l; k_1^{'}, \ldots, k_l^{'})$:

$$X_t = a_0^{(J_t)} + \sum_{i=1}^{k_{J_t}} a_i^{(J_t)} X_{t-i} + \sum_{i=1}^{k'_{J_t}} b_i^{(J_t)} \epsilon_{t-i} + \epsilon_t, \tag{1.19}$$

where J_t is determined as above in equation (1.18).

Open Loop Threshold AR (TARSO)

A second possible generalization of the basic SETAR structure above, is to allow for an external signal, e.g. an input signal, U_t, and let the external signal determine the regime, as e.g. in the TARSO$(l; d; (k_1, k_1^{'}), \ldots, (k_l, k_l^{'}))$ model :

$$X_t = a_0^{(J_t)} + \sum_{i=1}^{k_{J_t}} a_i^{(J_t)} X_{t-i} + \sum_{i=0}^{k'_{J_t}} b_i^{(J_t)} U_{t-i} + \epsilon_t. \tag{1.20}$$

Now the regime shifts are governed by

$$J_t = \begin{cases} 1 & U_{t-d} \in R_1 \\ 2 & U_{t-d} \in R_2 \\ \vdots & \vdots \\ l & U_{t-d} \in R_l \end{cases}, \tag{1.21}$$

i.e. for each value of the regime variable the system is described by an ordinary ARX-model. The extension of this structure to e.g. ARMAX-performance in each regime is immediate.

Independently Governed AR model (IGAR)

In this thresholding variant the changes in the selection scheme between different regimes is determined by a stochastic variable $\{J_t\}$, which is independent of the noise sources in the various regimes. Then the *stochastic* selection of the regimes in the IGAR(l;k) model is given by :

$$X_t = a_0^{(J_t)} + \sum_{i=1}^{k_{J_t}} a_i^{(J_t)} X_{t-i} + \epsilon_t^{(J_t)}, \tag{1.22}$$

where

$$J_t = \begin{cases} 1 & \text{with prob } p_1 \\ 2 & \text{with prob. } p_2 \\ \vdots & \vdots \\ l & \text{with prob. } 1 - \sum_{i=1}^{l-1} p_i \end{cases} . \tag{1.23}$$

This formulation of the thresholding illustrates that also the properties of the noise process forcing the autoregressions may depend on the regime.

A particular case appears when the regime variable J_t is given by a stationary Markov chain, i.e. when there exists a matrix of stationary transition probabilities P describing the switches between the basic auto regressions (MMAR(l;k)). Analysis and inference for this type of models is given in e.g. [Holst, Lindgren, Holst & Thuvesholmen 1994].

This kind of Markov modulations are also used in description of stochastic processes in telecommunication theory by giving a mechanism for switches between different Poisson processes. cf. [Rydén 1993].

Example 1.1. Let the AR processes be defined by

$$\begin{aligned} (a_1^{(1)}, a_2^{(1)}) &= (1.1, -0.5) \\ (a_1^{(2)}, a_2^{(2)}) &= (-1.2, -0.5) \\ \epsilon_t^{(1)} &\in \mathcal{N}(0,1) \\ \epsilon_t^{(2)} &\in \mathcal{N}(0,1), \end{aligned}$$

i.e. an MMAR(2;2) model, and let the transitions between the different regimes be governed by the matrix

$$P = \begin{pmatrix} 0.95 & 0.05 \\ 0.05 & 0.95 \end{pmatrix} \tag{1.24}$$

The performance of this system is shown in figure 1.2. The example and the figure are from [Thuvesholmen 1994]. ♦

Smooth Threshold AR (STAR)

The threshold models considered so far, have an abrupt change of regime when the decision variable transfers the boundary between regimes. This section will show an alternative to construct models with smooth transition between regimes. Start by considering the SETAR(2;d;k) model

$$X_t = a_0 + \sum_{j=1}^{k} a_j X_{t-j} + \left(b_0 + \sum_{j=1}^{k} b_j X_{t-j} \right) I(X_{t-d}) + \epsilon_t, \tag{1.25}$$

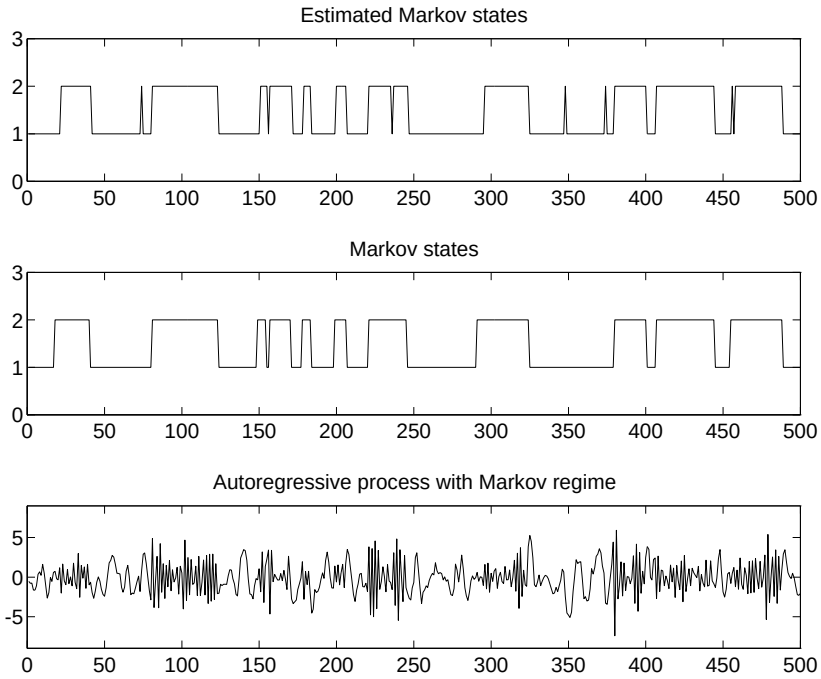

Figure 1.2: A Markov modulated regime model with two AR processes and a rather inert Markov chain – MMAR(2;2)-process

where $I(X_{t-d})$ is a step function to separate between the models with the AR-process characterized by the parameters $\{a_i\}$ from the process characterized by $\{a_i + b_i\}$. A model with smooth transition between the regimes is then obtained simply by letting the function $I(\cdot)$ be a smooth function – for instance a sigmoidal function where the parameters may be tuned to the desired smoothness or a standard Gaussian distribution.

The resulting model is for general values of the structural parameters be denoted a STAR(l;d;k)-model. The smoothing procedure may analogously be used when the regime variable is external as in the TARSO model or independent stochastic as in the IGAR and MMAR models above.

Hence summarize the most important difference in the transition between the regimes to get:

- In SETAR-models the abrupt transition depends on X_{t-d}.
- In TARSO-models the abrupt transition depends on U_{t-d}.
- In IGAR and MMAR-models the abrupt transition is stochastic and is determined of a regime variable that is independent of the forcing noise in the system.
- In STAR models transition between regimes is smooth and is determined by X_{t-d}, U_{t-d} or by an independent regime variable as above.

- Models with periodically varying coefficients are also possible to include under the threshold heading as are models where a stochastic regime may be used to select between different deterministic systems as shown in cf. [Tong 1990*b*] and [**?**].

1.3.3 Amplitude-dependent Exponential Autoregressive (EXPAR)

A particular form for threshold models with smooth transition is given by the Amplitude-dependent Exponential Autoregressive, EXPAR(d;k) model

$$X_t = \sum_{j=1}^{k} [\alpha_j + \beta_j e^{-\delta X_{t-d}^2}] X_{t-j} + \epsilon_t, \qquad \delta > 0. \tag{1.26}$$

Note that if $|X_{t-d}|$ is very large, then an ordinary AR(k) model is obtained, and if $|X_{t-d}|$ is very small another ordinary AR(k) appears. Hence, the EXPAR model contains a smooth transition between regimes.

1.3.4 Piecewise Polynomial Models

The skeleton is

$$y = f(X_{t-1}, X_{t-2}, \ldots, X_{t-k}) + \epsilon_t, \tag{1.27}$$

where f is a polynomial. In general

$$X_t = y. \tag{1.28}$$

Examples which also include the threshold principle :

Hard Censoring

$$X_t = \begin{cases} A & & y > A \\ y & -A \leq & y \leq A \\ -A & & y < -A \end{cases} \tag{1.29}$$

where y is defined by (1.27).

Soft Censoring

An extension of the Hard Censoring principle is given by the Soft Censoring :

$$X_t = \begin{cases} y & |y| \leq A \\ g(X_{t-1}, X_{t-2}, \ldots, X_{t-k}) + \epsilon_t & |y| > A \end{cases} \tag{1.30}$$

where y again is defined by (1.27) and g is a q'th order polynomial. Note that values outside $[-A, A]$ now are allowed.

If only first order terms enter into the polynomials the model is also a SETAR(2;k) model.

1.3.5 Fractional Autoregressive Models (FAR)

Rational functions in $\{X_{t-1}\}$.

The FAR(k;p;q) model :

$$X_t = \frac{a_0 + \sum_{i=1}^{k} \sum_{j=1}^{p} a_{ij} X_{t-i}^j}{b_0 + \sum_{i=1}^{k} \sum_{j=1}^{q} b_{ij} X_{t-i}^j} + \epsilon_t, \ 0 \leq p \leq q+1 \tag{1.31}$$

A simulation of the FAR(2,2,2) model

$$X_t = \frac{-0.35X_{t-1} - 0.1X_{t-1}^2 + 0.4X_{t-2} + 0.3X_{t-2}^2}{0.3X_{t-1} - 0.4X_{t-1}^2 + 0.2X_{t-2} - 0.3X_{t-2}^2} + \epsilon_t$$

where $\epsilon_t \in$ N(0,0.1) is shown in figure 1.3.

1.3.6 Product Autoregressive Model (PAR)

So far all the non-linear models involve *additive* noise.

An example of a non-linear model with non-additive noise is the PAR(k) model :

$$X_t = \epsilon_t \left(\sum_{i=1}^{k} a_i X_{t-i}^{\alpha_i} \right) \tag{1.32}$$

A simulation of the PAR(2) model

$$X_t = (0.3X_{t-1}^{-0.9} + 0.1X_{t-2}^{0.4})\epsilon_t$$

where $\epsilon_t \in$ N(1,0.1).is shown in figure 1.4

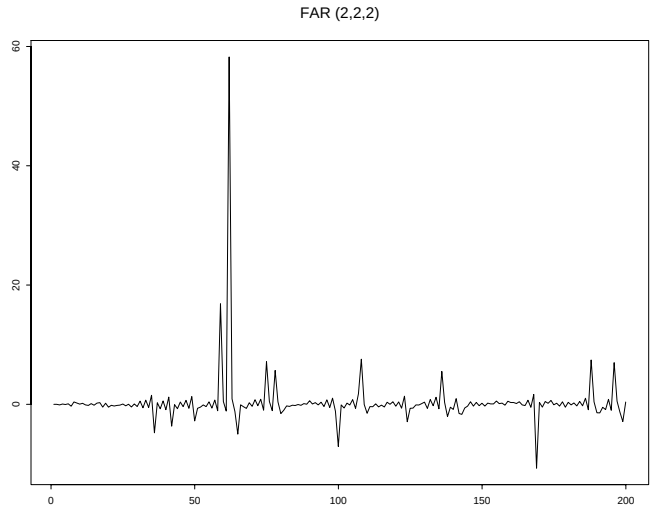

Figure 1.3: Simulation of a Fractional Autoregressive Model

1.3.7 Random Coefficient AR (RCAR)

Now the coefficients are random variables.

The RCAR(k) model :

$$X_t = \sum_{i=1}^{k} \{\beta_i + B_i(t)\} X_{t-i} + \epsilon_t, \tag{1.33}$$

where the sequences $B_i(s), i = 1, \ldots, k$ is white noise independent of ϵ_t for all values of t and s and the $\beta_i, i = 1, \ldots .k$ are scalars.

A simulation using the RCAR(2) model

$$X_t = B_1(t)X_{t-1} + B_2(t)X_{t-2} + \epsilon_t$$

where $\mathrm{E}[B_i(t)] = 0$ and $\mathrm{V}[B_i(t)] = 1.0$ is found in figure 1.5

1.3.8 The Bilinear Model (BL)

Simple - but important - extension of the ARMA model with product terms of $\mathrm{D}\{X_{t-1}\}$ and $\{\epsilon_t\}$.

Important reference : [Subba Rao & Gabr 1984*a*].

The BL(p,q,m,k) model :

$$X_t + \sum_{j=1}^{p} a_j X_{t-j} = \sum_{j=0}^{q} c_j e_{t-j} + \sum_{i=1}^{m}\sum_{j=1}^{k} b_{ij} X_{t-i} e_{t-j} \tag{1.34}$$

where $c_0 = 1$ and e is white noise.

Vector form of bilinear model

Important special case – BL$(p, 0, p, 1)$:

$$X_t + \sum_{j=1}^{p} a_j X_{t-j} = e_t + \sum_{l=1}^{p} b_l X_{t-l} e_{t-1} \tag{1.35}$$

Define

$$\boldsymbol{A} = \begin{bmatrix} -a_1 & -a_2 & -a_3 & \cdots & -a_{p-1} & -a_p \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \boldsymbol{B} = \begin{bmatrix} b_{11} & b_{21} & \cdots & b_{p1} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \tag{1.36}$$

and $\boldsymbol{C}^T = (1, 0, \ldots, 0)$, $\boldsymbol{H}^T = (1, 0, \ldots, 0)$ and $\boldsymbol{x}_t^T = (X_t, X_{t-1}, \ldots, X_{t-p+1})$ all of dimension p.

Now we have the *Vector Form* of the bilinear model

$$\begin{aligned} \boldsymbol{x}_t &= \boldsymbol{A}x_{t-1} + \boldsymbol{B}x_{t-1}e_{t-1} + \boldsymbol{C}\boldsymbol{e}_t \\ X_t &= \boldsymbol{H}^{'}\boldsymbol{x}_t \end{aligned} \tag{1.37}$$

which is denoted VBL(p).

Very like the Markov form of ordinary ARMA models; but it is *not* a Markovian representation.

Define $\boldsymbol{Z}_t = (\boldsymbol{A} + \boldsymbol{B}\boldsymbol{e}_t)\boldsymbol{x}_t$. Now

$$\boldsymbol{Z}_t = (\boldsymbol{A} + \boldsymbol{B}e_t)\boldsymbol{Z}_{t-1} + (\boldsymbol{A} + \boldsymbol{B}e_t)e_t \tag{1.38}$$

is Markovian.

Consider now the BL$(p, 0, p, q)$ model.

Define

$$\boldsymbol{B}_j = \begin{bmatrix} b_{1j} & b_{2j} & \cdots & b_{pj} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \qquad j = 1, 2, \ldots, q \tag{1.39}$$

now the BL$(p, 0, p, q)$ is written in the vector form

$$\begin{aligned} \boldsymbol{x}_t &= \boldsymbol{A}x_{t-1} + \sum_{j=1}^{q} \boldsymbol{B}_j \boldsymbol{x}_{t-1} e_{t-j} + \boldsymbol{C}e_t \\ X_t &= \boldsymbol{H}^T \boldsymbol{x}_t \end{aligned} \tag{1.40}$$

Finally, for the general BL model :

$$\begin{aligned} \boldsymbol{x}_t &= \boldsymbol{A}\boldsymbol{x}_{t-1} + \sum_{j=1}^{q} \boldsymbol{B}_j \boldsymbol{x}_{t-1} e_{t-j} + \boldsymbol{C}\boldsymbol{e}_t \\ X_t &= \boldsymbol{H}^T \boldsymbol{x}_t \end{aligned} \tag{1.41}$$

where, since $p' = \max(p, m)$, x_t has the dimension p' and

$$\boldsymbol{A} = \begin{bmatrix} -a_1 & -a_2 & -a_3 & \cdots & -a_{p-1} & -a_{p'} \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \tag{1.42}$$

$$\boldsymbol{B}_j = \begin{bmatrix} b_{1j} & b_{2j} & \cdots & b_{p'j} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \tag{1.43}$$

and $\boldsymbol{C}$ is given by

$$\boldsymbol{C} = \begin{bmatrix} c_0 & c_1 & \cdots & c_q \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \tag{1.44}$$

$\boldsymbol{e}_t$ is given by $\boldsymbol{e}_t = (e_t, e_{t-1}, \ldots, e_{t-r})$ and $a_i = 0$ for $i > p$, and $b_{ij} = 0$ for $i > m$.

Markovian representation - see [Pham Dinh Tuan 1985].

Simulation :

$$
\begin{aligned}
BL1 \quad & X_t = 0.4X_{t-1} + 0.4X_{t-1}\epsilon_{t-1} + \epsilon_t \\
BL2 \quad & X_t = 0.4X_{t-1} + 0.8X_{t-1}\epsilon_{t-1} + \epsilon_t \\
BL3 \quad & X_t = 0.8X_{t-1} - 0.4X_{t-2} + 0.6X_{t-1}\epsilon_{t-1} + 0.7X_{t-2}\epsilon_{t-1} + \epsilon_t
\end{aligned}
$$

In all cases $\epsilon_t \in$ N(0,1).

1.3.9 Auto Regressive models with Conditional Heteroscedasticity (ARCH)

Most models assume a constant one-step prediction variance. The ARCH process is introduced by [Engle 1982] to allow the conditional variance to change over time as a function of past errors leaving the unconditional variance constant. Hence, the recent past gives information about the one-step prediction variance.

This type of models behavior has proven to be useful in 0 several economic phenomena.

One of the most quoted examples are :

$$X_t = \epsilon_t \sqrt{V_t} \tag{1.45}$$

where

$$V_t = \gamma + \phi_1 X_{t-1}^2 + \cdots + \phi_p X_{t-p}^2 \tag{1.46}$$

where $\gamma > 0$ and $\phi_i > 0$.

In the more general setup in [Engle 1982] the following models is considered :

$$X_t | \psi_{t-1} \sim N(z_t\beta, h_t^2) \tag{1.47}$$

where ψ is the information set, β a vector of unknown parameters, z a vector of lagged variables included in the information set ψ, and

$$
\begin{aligned}
h_t &= h(\epsilon_t, \epsilon_{t-1}, \ldots, \epsilon_{t-p}, \boldsymbol{\theta}) && (1.48) \\
\epsilon_t &= X_t - z_t\beta && (1.49)
\end{aligned}
$$

where $\boldsymbol{\theta}$ is a vector of unknown parameters. The likelihood function can be found in [Engle 1982].

Several extensions of the basic ARCH model is possible. In [Engle, Lilien & Robins 1987] the ARCH model is extended to allow the conditional variance to affect the

mean. In this way changing conditional variances directly affect for instance the expected return. The models are called ARCH in means or ARCH-M models. The ARCH-M is obtained by a slight modification of equation (1.47)

$$X_t|\psi_{t-1} \sim N(z_t\beta + \delta h_t, h_t^2) \tag{1.50}$$

1.3.10 The Doubly Stochastic Model

A generalization of most of the models considered so far.

Tjøstheim [Tjøstheim 1986] suggests the doubly stochastic model :

$$X_t = A_0(t-1;X) + \sum_{i=1}^{P} A_i(t-1;X)\theta X_{t-i} + \sum_{j=0}^{Q} B_j(t-1;X)\epsilon_{t-j} \tag{1.51}$$

where the parameters θ_i, A_i, B_i all are random variables.

It is possible to select the parameters in the doubly stochastic model such that most of the non-linear models considered appears – see Table 1.1.

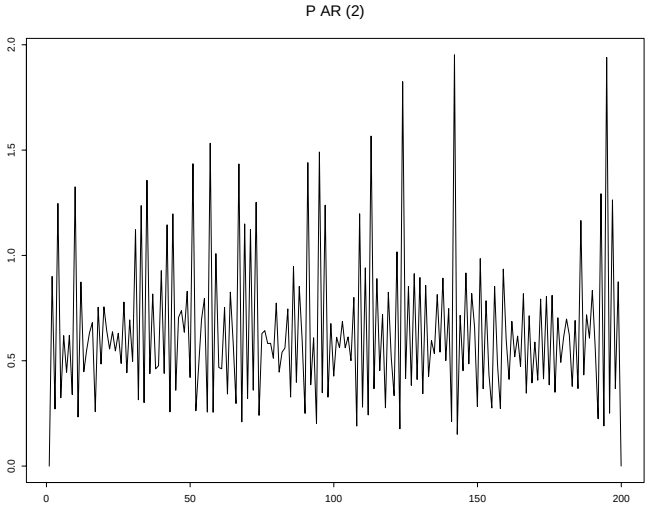

Figure 1.4: Product Autoregressive model

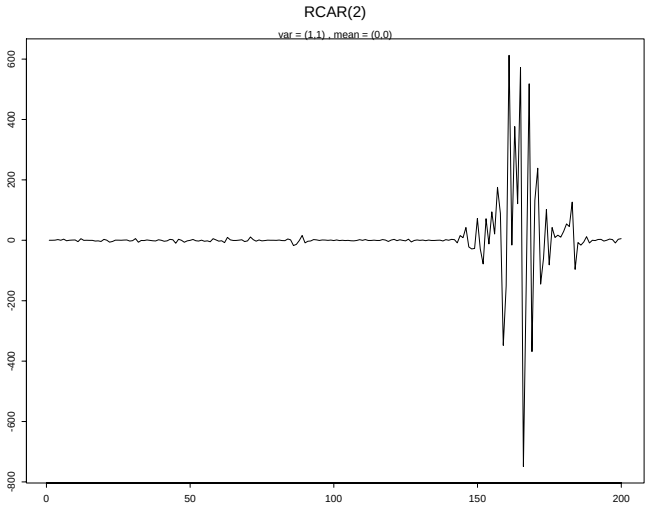

Figure 1.5: Random Coefficient AR

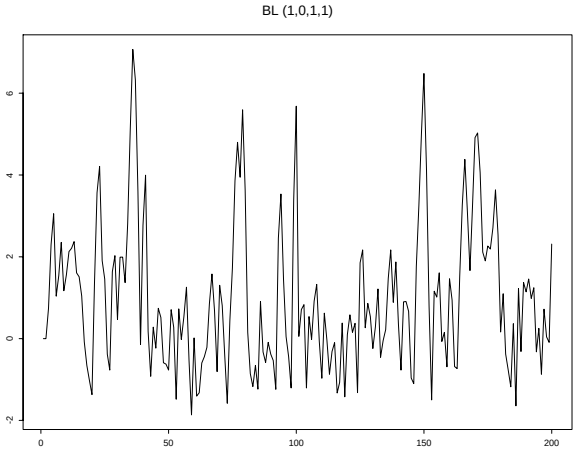

Figure 1.6: BL1

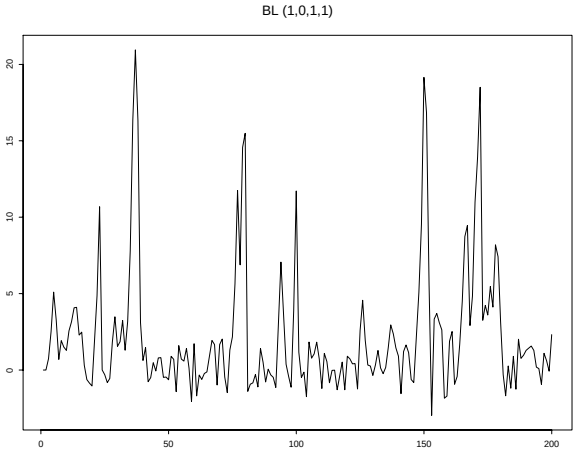

Figure 1.7: BL2

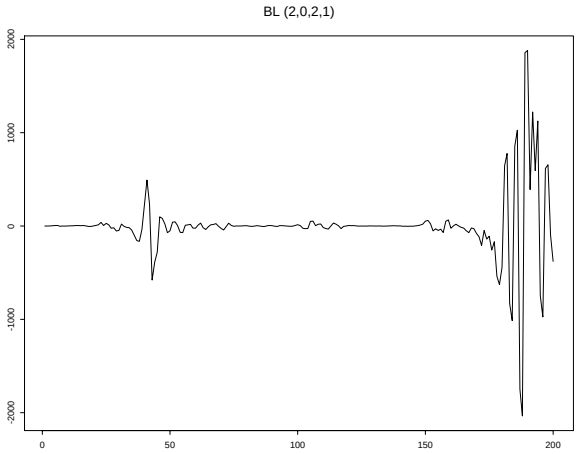

Figure 1.8: BL3

Model	Parameters in a Doubly Stoc. Model					Remarks/ Ext.param.
	P	Q	A_i	θ_i	B_i	
ARMA(p,q)	p	q	a_i	const.	$B_0 = 1$ $B_i = b_i$	
SETARMA (l, k_1, k_2)	k_1	k_2	$a_i^{(k)}$	const.	$B_0 = 1$ $B_i = B_i^{(k)}$	$X_{t-d} \in \mathcal{R}_k$ d=delay
IGAR(k)	k	0	$a_i^{(k)}$	const.	$B_0 = 1$	$D \in \mathcal{R}_k$ $D \in U[0;1]$
IGAR(k)	k	0	$\delta_{ij} a_i^{(1)}$	const.	$B_0 = 1$	$D \in \mathcal{R}_j$ $D \in U[0;1]$
Fractal(k)	l	0	$a_i^{(k)}$	const.	$B_0 = 0$	$D \in \mathcal{R}_k$ $D \in U[0;1]$
STAR(k)	k	0	$a_i + I(X_{t-d}) b_i$	const.	$B_0 = 1$	$I(\cdot)$ known pdf.
Per.coeff. ARMA(p,q)	p	q	$a_i(t)$	const.	$b_1(t)$	
Hard Censoring (k,p,A)	k	0	$a_i^{(l)}, a_i^{(1)} = f(X_{t-i})$ $a_i^{(2)} = sign(a_i^{(1)})A$	const.	$b^{(l)}, b^{(1)} = 1$ $b^{(2)} = 0$	l=1 For $\mid X_t \mid > A$: l=2
Soft Censoring (k,p,A)	k	0	$a_i^{(l)}, a_i^{(1)} = f(X_{t-i})$ $B_0 = 1$ $a_i^{(2)} = g(X_{t-i})$	const.	$b^{(l)}, b^{(1)} = 1$	l=1 For $\mid X_t \mid > A$ l=2
FAR(p,k)	0	0	$\frac{a_0 + \Sigma_{i=l}^{p} \Sigma_{j=l}^{k} a_{ij} x_{t-i}^{j}}{1 + \Sigma_{i=l}^{p} \Sigma_{j=l}^{k} b_{ij} x_{t-i}^{j}}$	const.	$B_0 = 1$	Only for A_0
PAR(k)	0	0			$a_0 +$ $\Sigma_{i=1}^{k} a_i X_{t-i}^{\alpha_i}$	
EXPAR(k)	k	0	$b_i + \beta_i e^{-\delta X_{t-d}^2}, i > 0$	const.	$B_0 = 1$	δ d=delay
RCAR(k)	k	0	const.	r.v.	$B_0 = 1$	
BL(p,q,r,s)	p	q'	a_i	const.	$b_i +$ $\Sigma_{j=l}^{r} c_{ij} X_{t-j}$	q'=max[q,s]

Table 1.1: A family of doubly stochastic models.

1.4 Stationarity and moments for the VBL(p) model

Brief summary of results from [Subba Rao & Gabr 1984*a*], who shows the conditions for stationarity and expressions for moments for the VBL(p) model.

Assume for simplicity that $e_t \in N(0,1)$.

1.4.1 First order stationarity

Take the expectation on both sides of (1.37) (note that $\mathrm{E}[\boldsymbol{x}_t e_t] = \boldsymbol{C}$)

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{A}\boldsymbol{\mu}_t + \boldsymbol{B}\boldsymbol{C} = \boldsymbol{A}^t\boldsymbol{\mu}_1 + (\sum_{j=0}^{t-1} \boldsymbol{A}^j)\boldsymbol{B}\boldsymbol{C} \tag{1.52}$$

It is seen that if $\boldsymbol{\mu}_1 = \boldsymbol{0}$ and either $\boldsymbol{B} = \boldsymbol{0}$ or $\boldsymbol{C} = \boldsymbol{0}$ then $\boldsymbol{\mu}_t = \boldsymbol{0}$ for $t \geq 1$.

Define (the spectral radius)

$$\rho(\boldsymbol{A}) = \max_i |\lambda_i(\boldsymbol{A})| \tag{1.53}$$

where $\lambda_i(\boldsymbol{A})$ is the i'th eigenvalue.

A sufficient condition for $\boldsymbol{\mu} = \lim_{t\to\infty}[\boldsymbol{A}^t\boldsymbol{\mu}_1 + (\sum_{j=0}^{t-1} \boldsymbol{A}^j)\boldsymbol{B}\boldsymbol{C}]$ to be finite is that

$$\rho(\boldsymbol{A}) < 1 \tag{1.54}$$

Hence, under this condition the VBL(p) is asymptotically first order stationary with mean

$$\boldsymbol{\mu} = (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{B}\boldsymbol{C} \tag{1.55}$$

1.4.2 Second order stationarity

Assuming that (1.54) is satisfied, then a necessary condition for second order stationarity is given by [Subba Rao & Gabr 1984*a*] as

$$\rho[\boldsymbol{A} \otimes \boldsymbol{A} + \boldsymbol{B} \otimes \boldsymbol{B}] < 1 \tag{1.56}$$

For a discussion about stationarity for the general bilinear model - see [Pham Dinh Tuan 1986]

1.4.3 Linear versus bilinear model

Assuming second order stationarity, and let us introduce the covariance

$$\boldsymbol{K}(s) = \mathrm{E}[(\boldsymbol{x}_{t+s} - \boldsymbol{\mu})(\boldsymbol{x}_t - \boldsymbol{\mu})^T] \tag{1.57}$$

It is shown by [Subba Rao & Gabr 1984*a*] that

$$\boldsymbol{K}(0) = \boldsymbol{A}\boldsymbol{K}(0)\boldsymbol{A}^T + \boldsymbol{B}\boldsymbol{C}(0)\boldsymbol{B}^T + \boldsymbol{\Delta}_1 \tag{1.58}$$
$$\boldsymbol{K}(1) = \boldsymbol{A}\boldsymbol{C}(0) + \boldsymbol{\Delta}_2 \tag{1.59}$$
$$\boldsymbol{K}(s) = \boldsymbol{A}\boldsymbol{K}(s-1) = \boldsymbol{A}^{s-1}\boldsymbol{K}(1), \qquad s = 2, 3, \ldots \tag{1.60}$$

where the constant matrices $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$ are given as

$$\begin{aligned}\boldsymbol{\Delta}_1 &= \boldsymbol{B}\boldsymbol{\mu}\boldsymbol{\mu}^T\boldsymbol{B}^T + \boldsymbol{A}\boldsymbol{\mu}\boldsymbol{\mu}^T\boldsymbol{A}^T + \boldsymbol{A}\boldsymbol{S}\boldsymbol{B}^T + \boldsymbol{B}\boldsymbol{S}^T\boldsymbol{A}^T + 2\boldsymbol{B}\boldsymbol{C}\boldsymbol{C}^T\boldsymbol{B}^T + \boldsymbol{C}\boldsymbol{C}^T - \boldsymbol{\mu}\boldsymbol{\mu}^T \\ \boldsymbol{\Delta}_2 &= \boldsymbol{A}\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{B}\boldsymbol{S} - \boldsymbol{\mu}\boldsymbol{\mu}^T\end{aligned} \tag{1.61}$$

where $\boldsymbol{S} = \boldsymbol{S}_t = \mathrm{E}[\boldsymbol{x}_t\boldsymbol{x}_t^T e_t]$ ($\boldsymbol{S}$ is constant due to the assumption about stationarity).

By using $\boldsymbol{A}$ from (1.36) in (1.60) we obtain the simple matrix equation :

$$\begin{bmatrix} \gamma(s) & \gamma(s+1) & \cdots & \gamma(s+p-1) \\ \gamma(s-1) & \gamma(s) & \cdots & \gamma(s+p-2) \\ \vdots & \vdots & & \vdots \\ \gamma(s-p+1) & \gamma(s-p+2) & \cdots & \gamma(s) \end{bmatrix} =$$
$$\begin{bmatrix} -a_1 & -a_2 & -a_3 & \cdots & -a_{p-1} & -a_{p'} \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \gamma(s-1) & \gamma(s) & \cdots & \gamma(s+p-2) \\ \gamma(s-2) & \gamma(s-1) & \cdots & \gamma(s+p-3) \\ \vdots & \vdots & & \vdots \\ \gamma(s-p) & \gamma(s-p+1) & \cdots & \gamma(s-1) \end{bmatrix} \tag{1.62}$$

where γ is the autocovariance function for X_t.

It is easily seen that the above matrix equation reduces to p identical equations (and some trivial identities), which mean that we are able to write (1.60) on the form

$$\gamma(s) + a_1\gamma_(s-1) + \cdots + a_p\gamma(s-p) = 0 \qquad s \geq 2 \tag{1.63}$$

Remark. The equation above is obviously the same as the Yule-Walker equations for an ARMA$(p, 1)$-process. Hence, for any BL$(p, 0, p, 1)$ a linear model with exactly the same covariance structure is found. ▼

It is not possible to use ordinary Box-Jenkins methods to identify non-linear systems. No matter the number of parameters in the BL$(p, 0, p, 1)$ we are able to find a linear model with only $p + 1$ parameters which describes the autocorrelation function (and the spectrum) as well as the true model.

Conclusion : In order to identify non-linearities we must use other methods. For instance methods based on the third moment, or methods based on non-parametric estimates of the conditional mean. We shall look on that in the following.

Chapter 2

Non–parametric methods and time series analysis

2.1 Introduction

Non-parametric methods are widely used in non-linear model building. Such methods are particularly useful if no prior information about the structure is available, since the estimation procedure is free of parameters and model structure (apart from a smoothing constant).

A non-parametric 'regression curve' is *a direct function of the data*. It is shown later on in this chapter that the basic idea of local averaging in kernel estimation is equivalent to finding a locally weighted least squares estimate, and the local weight is given by a sequence $\{w_s(x); s = 1, \cdots, N\}$, where N is the number of observations.

In this chapter we will concentrate on *kernel estimation* and *smoothing*. Some other non-parametric methods which are only briefly mentioned are based on *splines*, k *nearest neighbor (k-NN)* or *orthogonal series smoothing*. Each method has a specific weighting sequence $\{w_s(x); s = 1, \cdots, N\}$. These weighting sequences are related to each other and it can be argued [Härdle 1990] that one of the simplest ways of computing a weighting sequence is kernel smoothing.

The method is in general introduced using a regression setting. Later on many very important applications within non-linear and non-stationary time series analysis are described. Very useful methods for *identifying* both *the needed lags* in a non-linear time series models, and methods for identifying for instance *the non-linear functional relationship on lagged values* are based on non-parametric methods. For more information

about non-parametric methods see [Robinson 1983], [Härdle 1990], [Silverman 1986] and [Hastie & Tibshirani 1990].

2.2 Kernel estimators for time series

2.2.1 Introduction

This chapter considers kernel estimation in general and its use in time series analysis. The *non-parametric estimators* are of particular relevance for non-Gaussian or non-linear time series. From kernel estimates of *probability density functions* Gaussianity can be verified or the nature of *non-Gaussianity* can be emerged. Thus non-parametric methods provide information that complements that given for instance by higher-order spectral analysis – see Chapter 3.

In time series analysis and system modelling the conditional mean and variance plays a central rule. For *linear models* it is well-known that the conditional mean $\mathrm{E}[Y_t|Y_{t-1},\ldots]$ is linear (in $Y_{t-1},\ldots$) and the conditional variance $\mathrm{V}[Y_t|Y_{t-1},\ldots]$ is constant. This is not the case for non-linear models.

Non-parametric estimates of *the conditional mean* can be used to detect non-linear behavior and to *identify* a family of relevant time series models. Consider for instance the time series generated by the non-linear model:

$$Y_t = g(Y_{t-1},\ldots,Y_{t-p}) + h(Y_{t-1},\ldots,Y_{t-p})\epsilon_t. \tag{2.1}$$

For this non-linear model the conditional mean and variance is

$$\begin{aligned} \mathrm{E}[Y_t|Y_{t-1},\ldots] &= g(Y_{t-1},\ldots,Y_{t-p}) && (2.2)\\ \mathrm{V}[Y_t|Y_{t-1},\ldots] &= h^2(Y_{t-1},\ldots,Y_{t-p})\sigma_\epsilon^2 && (2.3) \end{aligned}$$

where σ_ϵ^2 is the variance of the white noise process ϵ_t.

Thus the non-parametric methods can be used to estimate for instance $g(y) = \mathrm{E}[Y_{t+1}|Y_t = y]$ in the simple case $p = 1$. Using this non-parametric estimate a relevant parameterization of $g(\cdot)$ can be suggested. This very important aspect will be considered further in Section 2.4.2 and in Chapter 3.

In classical (i.e. linear) time series analysis the order of, for instance, an ARMA(p,q) model, or equivalently the number of needed lagged values of the process itself and the noise, is identified using the autocorrelation functions (ACF, PACF and IACF). For

non-linear time series analysis equivalent functions will be suggested. These functions, called *Lag Dependent Functions* (LDF), are introduced using non-parametric methods.

Furthermore non-parametric estimates can be used to provide *predictors*, i.e. estimators of future values of the time series. Use of non-parametric estimators for getting insight in the *time-varying* behaviour and a complex *dependency on exogenous variables* will also be considered.

In this section the most frequently considered non-parametric estimators for time series, the kernel estimators, are firstly introduced. Next the kernel estimator are described in more detail. Finally, some examples on how to use kernel estimation in time series analysis are given.

2.2.2 The kernel estimator

Let $\{Y_t; t = 0, \pm 1, \cdots\}$ be *a strictly stationary process*. A realization is given as the time series $\{y_t; t = 1, \cdots, N\}$, and that single realization of the process is the basis for inference about the process.

Introduce $Z_t = (Y_{t+j_1}, \cdots, Y_{t+j_n})$ and $Y_t^* = (Y_{t+h_1}, \cdots, Y_{t+h_m})$.

In order to simplify the notation we will consider the case $Y_t^* = (Y_t)$ in the rest of this section. It is obvious how to generalize the equations.

The *basic quantity estimated* is

$$E[G(Z_t)|Y_t = y]f_Y(y) \tag{2.4}$$

where f_Y is the pdf of Y_t and G is a known function.

The *estimator* of (2.4) is (using a special notation)

$$[G(Z_t); y] = (N'h)^{-1} \sum_{t=1}^{N'} G(Z_t)k(\frac{y - Y_t}{h}) \tag{2.5}$$

where $k(u)$ is a real, bounded, function such that $\int k(u)du = 1$, and h is a real number. N' is some number which corrects for the fact that not all N observations can be used in the sum – typically $N' = N - j_n$. It is important to notice that $k(u)$ is called the *kernel function* and h the *bandwidth*.

Let us consider two examples of using equation (2.5). In the first example it is illustrated that equation (2.5) is a reasonable estimator.

Example 2.1 (Non-parametric estimation of a pdf). The pdf of Y_t at y is estimated by

$$\hat{f}(y) = [1; y] = \frac{1}{Nh} \sum_{t=1}^{N} k(\frac{y - Y_t}{h}) \tag{2.6}$$

where the meaning of the notation $[1; y]$ follows from (2.5).

Assume that Y has pdf $f(y)$, then

$$f(y) = \lim_{h \to 0} \frac{1}{2h} P(y - h < Y \leq y + h) \tag{2.7}$$

An estimator of $f(y)$ is

$$\hat{f}(y) = \frac{1}{2hN} [\text{Number of observations in } (y - h; y + h)] \tag{2.8}$$

Define the *rectangular kernel*

$$w(u) = \begin{cases} \frac{1}{2} & \text{if } |u| < 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

then the estimator (2.8) can be written as

$$\hat{f}(y) = \frac{1}{N} \sum_{t=1}^{N} \frac{1}{h} w(\frac{y - Y_t}{h})$$

Now consider the general kernel ($\int k(u)du = 1$). Then the estimator is

$$\hat{f}(y) = \frac{1}{Nh} \sum_{t=1}^{N} k\left(\frac{y - Y_t}{h}\right) \tag{2.10}$$

which clearly is equal to equation (2.6). ♦

Example 2.2 (Non-parametric estimation of the conditional mean). The estimator of the conditional expectation of $G(Z_t)$, given $Y_t = y$ is

$$\begin{aligned} \hat{\mathrm{E}}[G(Z_t)|Y_t = y] &= \frac{[G(Z_t); y]}{[1; y]} \\ &= \frac{\frac{1}{N'} \sum G(Z_t) k(\frac{y - Y_t}{h})}{\frac{1}{N} \sum k(\frac{y - Y_t}{h})} \end{aligned} \tag{2.11}$$

where $[1; y]$ was found in the previous example.

Of special interest is the case $G(Z_t) = Y_{t+1}$, where (2.11) estimates $E[Y_{t+1}|Y_t = y]$. ♦

2.2.3 Central limit theorems

Central limit theorems can be established under various weak dependence condition on the process $\{Y_t\}$. Let $\mathcal{F}_u^v$ be the σ-field of events generated by $Y_t; u \leq t \leq v$. Then introduce the coefficient

$$\alpha_j = \sup_{A \in \mathcal{F}_{-\infty}^t, B \in \mathcal{F}_{t+j}^{\infty}} |P(A \cap B) - P(A)P(B)|;\ j > 0 \tag{2.12}$$

Definition 2.1 (Strong mixing condition). If $\alpha_j \to 0$ as $j \to \infty$ the process X_t is said to be *strongly mixing*. ▲

The strong mixing condition is used frequently in the asymptotic theory of estimators for time series; but the condition is often very difficult to check.

Central limit theorems for non-parametric estimators of pdfs of a process $\{Y_t\}$, as well as of conditional pdfs and conditional expectations at continuity points, are given by [Robinson 1983] under the strong mixing condition.

2.3 Kernel estimation for regression

This section describes in more detail the kernel estimation technique. Since the method is useful for non-parametric regression in general a notation which is related to non-parametric regression will be used; but the problem is highly related to the estimator in (2.11), and thus also useful in time series analysis.

2.3.1 An estimator for regression

Kernel based non-parametric regression is first proposed by Nadaraya (1964). Let Y denote the dependent variable (also called the response, target, output), and let $X^{(1)}, \ldots, X^{(q)}$ denote the q independent variables (also called regressor, covariate, feature, input).

Assume that the theoretical relation is

$$Y = g(X^{(1)}, \ldots, X^{(q)}) + \epsilon \tag{2.13}$$

where ϵ is white noise, and g is an unknown continuous function.

Given N observations

$$O = \{(X_1^{(1)}, \ldots, X_1^{(q)}, Y_1), \ldots, (X_N^{(1)}, \ldots, X_N^{(q)}, Y_N)\}$$

the goal is now to estimate the function g.

Using the result in Example 2.2 the *kernel estimator* for g given the observations in the case $q = 1$ is

$$\hat{g}(x) = \frac{\frac{1}{N}\sum_{s=1}^{N} Y_s k\{h^{-1}(x - X_s)\}}{\frac{1}{N}\sum_{s=1}^{N} k\{h^{-1}(x - X_s)\}} \tag{2.14}$$

Note that if we compare with equation (2.5) then $G(Z_t) = Z_t = Y_t$, and the independent variable is X.

Various candidate functions for the *kernel* k have been proposed. [Härdle 1990] has shown that the parabolic kernel with bounded support, called the Epanechnikov kernel minimizes (asymptotically) the mean squared error among all kernels in which the bandwidth h is optimally chosen.

The *Epanechnikov kernel* with bandwidth h is given by :

$$K_h^{Epa}(u) = h^{-1}k^{Epa}(uh^{-1}) = \frac{3}{4h}(1 - \frac{u^2}{h^2})I_{\{|u| \leq h\}}. \tag{2.15}$$

Other possibilities are the rectangular, triangular or the *Gaussian kernel* given by:

$$K_h^{Gau}(u) = h^{-1}k^{Gau}(uh^{-1}) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{u^2}{2h^2}\right). \tag{2.16}$$

These kernels are depicted in Figure 2.1. The Epanechnikov kernel is shown for $h = 1$, and the Gaussian kernel for $h = 0.45$.

If we allow $q > 1$ in (2.14) a kernel of q'th order have to be used, i.e. a real bounded function $k_q(u_1, \ldots, u_q)$, where $\int k_q(\boldsymbol{u})d\boldsymbol{u} = 1$.

By a generalization of the bandwidth h to the quadratic matrix $\boldsymbol{h}$ with the dimension $q \times q$ the more general kernel estimator for $g(X^{(1)}, \ldots, X^{(q)})$ is

$$\hat{g}(\boldsymbol{x}) = \frac{\frac{1}{N}\sum_{s=1}^{N} Y_s k_q[(\boldsymbol{x} - \boldsymbol{X_s})\boldsymbol{h}^{-1}]}{\frac{1}{N}\sum_{s=1}^{N} k_q[(\boldsymbol{x} - \boldsymbol{X_s})\boldsymbol{h}^{-1}]} \tag{2.17}$$

where $\boldsymbol{x} = (x_1, \ldots, x_q)$ and $\boldsymbol{X_s} = (X_s^{(1)}, \ldots, X_s^{(q)})$. This is called the *Nadaraya-Watson estimator*.

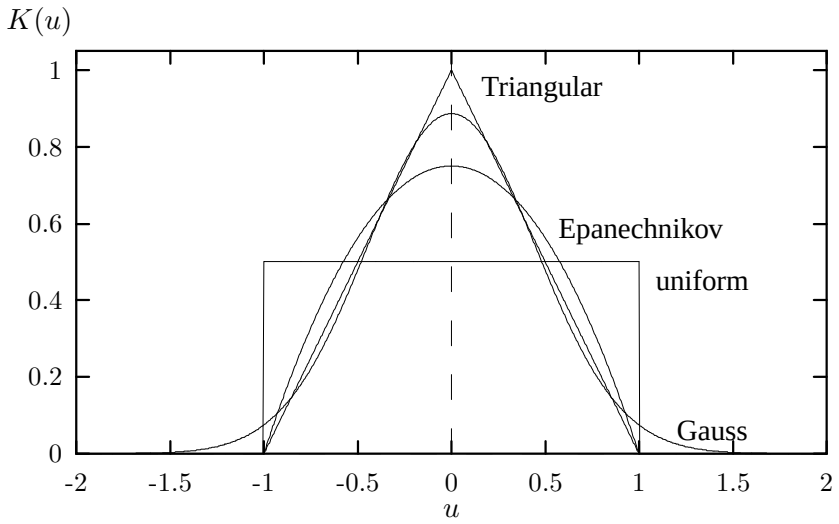

Figure 2.1: Some univariate kernel functions.

2.3.2 Product kernel

In practice product kernels are used, i.e.

$$k_q(\boldsymbol{x}\boldsymbol{h}^{-1}) = \prod_{i=1}^{q} k(x_i/h_i) \tag{2.18}$$

where k is a kernel of order 1.

Assuming that $h_i = h, \quad i = 1, \ldots, q$ the following generalization of the one dimensional case in (2.14) is obtained

$$\hat{g}(\boldsymbol{x}) = \frac{\frac{1}{N}\sum_{s=1}^{N} Y_s \prod_{i=1}^{q} k\{h^{-1}(x_i - X_s^{(i)})\}}{\frac{1}{N}\sum_{s=1}^{N} \prod_{i=1}^{q} k\{h^{-1}(x_i - X_s^{(i)})\}}. \tag{2.19}$$

2.3.3 Non-parametric LS

Let us now describe how the kernel estimation compares to locally-weighted least squares (LWLS). LWLS techniques for non-parametric estimation will be further considered in Section 2.5.2.

If we define the weight

$$w_s(\boldsymbol{x}) = \frac{\prod_{i=1}^{q} k\{h^{-1}(x_i - X_s^{(i)})\}}{\frac{1}{N}\sum_{s=1}^{N}\prod_{i=1}^{q} k\{h^{-1}(x_i - X_s^{(i)})\}} \tag{2.20}$$

then it is seen that equation (2.19) corresponds to the local average:

$$\hat{g}(\boldsymbol{x}) = \frac{1}{N}\sum_{s=1}^{N} w_s(\boldsymbol{x}) Y_s. \tag{2.21}$$

The estimate is actually a *non-parametric least squares estimate* at the point $\boldsymbol{x}$. This is recognized from the fact that the solution to the least squares problem

$$\arg\min_{\theta} \frac{1}{N}\sum_{s=1}^{N} w_s(\boldsymbol{x})(Y_s - \theta)^2 \tag{2.22}$$

is given by

$$\hat{\theta}_{LS}(\boldsymbol{x}) = \frac{\sum_{s=1}^{N} w_s(\boldsymbol{x}) Y_s}{\sum_{s=1}^{N} w_s(\boldsymbol{x})}. \tag{2.23}$$

This shows that at each $\boldsymbol{x}$, $\hat{g}$ is essentially a weighted LS location estimate, i.e.

$$\hat{g}(\boldsymbol{x}) = \hat{\theta}_{LS}(\boldsymbol{x})\frac{1}{N}\sum_{s=1}^{N} w_s(\boldsymbol{x}) = \hat{\theta}_{LS}(\boldsymbol{x}). \tag{2.24}$$

since $\frac{1}{N}\sum_{s=1}^{N} w_s(\boldsymbol{x}) = 1$ cf. Eq.(2.20).

2.3.4 Bias problems

Kernel estimators tends to give bias problems at the boundary of the predictor space (i.e. the space spanned by the independent variables). This is due to the fact that the kernel neighborhood at the boundary is asymmetric. This problem can somehow be accomplished for by modifying the kernel and for instance using a non-symmetric kernel – see [Härdle 1990].

Bias can also be a problem in the interior of the predictor space if the predictors are non-uniform or if the true regression curve has a substantial curvature. These problems are most dominant when the predictors are multidimensional. A variety of kernel modifications have been proposed to provide approximate and asymptotic adjustments

for these biases; but in practice it is difficult to implement the necessary modifications, in particular in the multivariate case.

A family of procedures which does not have the above problem is the regression smoothers, which will be described in Section 2.5.2.

2.3.5 The bandwidth

In analogy with smoothing in spectrum analysis the bandwidth h determines the smoothness of $\hat{g}$:

- If h is large the variance is small but the bias is large.
- If h is small the variance is large but the bias is small.

In the limits:

As $h \to \infty$ it is seen that $\hat{g}(\boldsymbol{x}) = \bar{Y}$.

As $h \to 0$ it is seen that $\hat{g}(\boldsymbol{x}) = Y_i$ for $\boldsymbol{x} = (X_i^{(1)}, \ldots, X_i^{(q)})$ and otherwise 0

2.3.6 Selection of bandwidth - Cross validation

The ultimative goal for the selection of h is to minimize (see [Härdle 1990])

$$\mathrm{MSE1}(h) = \frac{1}{N}\sum_{i=1}^{N}[\hat{g}(\boldsymbol{X}_i) - g(\boldsymbol{X}_i)]^2\pi(\boldsymbol{X}_i) \tag{2.25}$$

Where $\pi(\ldots)$ is a weighting function which screens off some of the extreme observations, and g is the unknown function.

An easy solution is the 'plug-in' method, where Y_i is used as an estimate of $g(\boldsymbol{X}_i)$ in (2.25).

Hence, the criterion is

$$\mathrm{MSE2}(h) = \frac{1}{N}\sum_{i=1}^{N}[\hat{g}(\boldsymbol{X}_i) - Y_i]^2\pi(\boldsymbol{X}_i) \tag{2.26}$$

but it is clear that $\widehat{\mathrm{MSE2}}(h) \to 0$ for $h \to 0$, since $y_i - \hat{g}(\boldsymbol{x}_i) \to 0$ when $h \to 0$. Hence a modification is needed.

In the "leave one out" estimator the idea is to avoid that $(\boldsymbol{X}_i)$ is used in the estimate for Y_i. For every data $(\boldsymbol{X}_i, Y_i)$ we define an estimator $\hat{g}_{(i)}$ for Y_i based on all data except $(\boldsymbol{X}_i)$.

The N estimators $\hat{g}_{(1)}, \ldots, \hat{g}_{(N)}$ (called the "leave one out" estimators) are written

$$\hat{g}_{(j)}(\boldsymbol{x}) = \frac{\frac{1}{N-1}\sum_{s \neq j} Y_s \prod_{i=1}^{q} k\{h^{-1}(x_i - X_s^{(i)})\}}{\frac{1}{N-1}\sum_{s \neq j} \prod_{i=1}^{q} k\{h^{-1}(x_i - X_s^{(i)})\}} \tag{2.27}$$

Now the *Cross-Validation criterion* using the "leave one out" estimates $\hat{g}_{(1)}, \ldots, \hat{g}_{(N)}$ is

$$CV(h) = \frac{1}{N}\sum_{i=1}^{N} [Y_i - \hat{g}_{(i)}(\boldsymbol{X}_i)]^2 \pi(\boldsymbol{X}_i). \tag{2.28}$$

and the optimal value of h is then found by minimizing $CV(h)$.

It can be shown that under weak assumptions the estimate of the bandwidth $\hat{h}$ that is obtained when minimizing the Cross-Validation criterion is asymptotic optimal, i.e. it minimizes (2.25) – see [Härdle 1990].

An example of using the CV procedure is shown in Figure 2.7.

2.3.7 Variance of the non-parametric estimates

To assess the variance of the curve estimate at point $\boldsymbol{x}$ Härdle (1990) proposes the point-wise estimator given by

$$\hat{\sigma}^2(\boldsymbol{x}) = \frac{1}{n}\sum_{s=1}^{n} w_s(\boldsymbol{x})(Y_s - \hat{g}(\boldsymbol{X}_s))^2. \tag{2.29}$$

where the weights are given as shown previously by

$$w_s(\boldsymbol{x}) = \frac{\prod_{i=1}^{q} k\{h^{-1}(x_i - X_s^{(i)})\}}{\frac{1}{n}\sum_{s=1}^{n} \prod_{i=1}^{q} k\{h^{-1}(x_i - X_s^{(i)})\}} \tag{2.30}$$

Remembering the WLS interpretation this estimate seems reasonable.

2.4 Applications of kernel estimators

2.4.1 Identification of the model order

In linear time series analysis the order of, for instance, an ARMA model is determined by considering the autocorrelation function (ACF), the partial autocorrelation function (PACF) and/or the inverse autocorrelation function (IACF).

In non-linear time series some equivalent functions, called *Lag Dependent Functions* (LDF), have resently been introduced in [Nielsen & Madsen 2001]. These methods are based on non-parametric methods and described in Section 3.3.

2.4.2 Non-parametric estimation of the conditional mean and variance

Assume that a realization $X_1, \ldots, X_N$ of a stochastic process $\{X_t\}$ is available. The goal is now to use the realization to estimate the functions $g(\cdot)$ and $h(\cdot)$ in the model

$$X_t = g(X_{t-1}, \ldots, X_{t-p}) + h(X_{t-1}, \ldots, X_{t-q})\epsilon_t. \tag{2.31}$$

As in Tjöstheim (1994) we shall use the notation

$$M(x_{i_1}, \ldots, x_{i_d}) = \mathrm{E}[X_t | X_{t-i_1} = x_{i_1}, \ldots, X_{t-i_d} = x_{i_d}] \tag{2.32}$$

$$V(x_{i_1}, \ldots, x_{i_d}) = \mathrm{V}[X_t | X_{t-i_1} = x_{i_1}, \ldots, X_{t-i_d} = x_{i_d}] \tag{2.33}$$

where i_j are positive integers arranged in increasing order.

One solution is to use the kernel estimator (other possibilities are splines, nearest neighbor or neural network estimates).

Using a product kernel we obtain

$$\hat{M}(x_{i_1}, \ldots, x_{i_d}) = \frac{\frac{1}{N-i_d} \sum_{s=i_d+1}^{N} X_s \prod_{j=1}^{d} k\{h^{-1}(x_{i_j} - X_{s-i_j})\}}{\frac{1}{N-i_d+i_1} \sum_{s=i_d+1}^{N+i_1} \prod_{j=1}^{d} k\{h^{-1}(x_{i_j} - X_{s-i_j})\}} \tag{2.34}$$

Assuming that $E(X_t) = 0$ the estimator for $V(\cdot)$ is

$$\hat{V}(x_{i_1}, \ldots, x_{i_d}) = \frac{\frac{1}{N-i_d} \sum_{s=i_d+1}^{N} X_s^2 \prod_{j=1}^{d} k\{h^{-1}(x_{i_j} - X_{s-i_j})\}}{\frac{1}{N-i_d+i_1} \sum_{s=i_d+1}^{N+i_1} \prod_{j=1}^{d} k\{h^{-1}(x_{i_j} - X_{s-i_j})\}} \tag{2.35}$$

If $\mathrm{E}[X_t] \neq 0$ it is clear that the above estimator is changed to

$$\hat{V}(x_{i_1},\ldots,x_{i_d}) = \frac{\frac{1}{N-i_d}\sum_{s=i_d+1}^{N} X_s^2 \prod_{j=1}^{d} k\{h^{-1}(x_{i_j}-X_{s-i_j})\}}{\frac{1}{N-i_d+i_1}\sum_{s=i_d+1}^{N+i_1}\prod_{j=1}^{d} k\{h^{-1}(x_{i_j}-X_{s-i_j})\}} - \hat{M}^2(x_{i_1},\ldots,x_{i_d}) \tag{2.36}$$

Example 2.3 (Estimation of conditional mean and variance). We have simulated 10 stochastic independent time series for both of the following non-linear models:

$$\begin{aligned} A: \quad X_t &= \begin{cases} -0.8X_{t-1}+\epsilon_t, & X_{t-1}\geq 0 \\ 0.8X_{t-1}+1.5+\epsilon_t, & X_{t-1}<0 \end{cases} \\ B: \quad X_t &= \sqrt{1+X_{t-1}^2}\epsilon_t \end{aligned}$$

Using the simulated time series 10 estimates of the conditional mean and the conditional variance are found using a Gaussian kernel with $h = 0.6$. The results for $M(x)$ and $V(x)$ are shown in Figure 2.2 and 2.3, respectively.

The theoretical conditional mean and variance are shown with '+'. ♦

2.4.3 Non-parametric estimation of the pdf

An estimate of the probability density function (pdf) is [Robinson 1983] :

$$\hat{f}(x_{i_1},\ldots,x_{i_d}) = \frac{1}{N-i_d+i_1}\sum_{s=i_d+1}^{N+i_1}\prod_{j=1}^{d} h^{-1}k\{h^{-1}(x_{i_j}-X_{s-i_j})\}. \tag{2.37}$$

Compare with the univariate pdf in (2.6).

2.4.4 Non-parametric estimation of non-stationarity - An example

Non-parametric methods can be applied to an identification of the structure of the existing relationships leading to proposals for a parametric model. An example of identifying the diurnal variation in a non-stationary time series is given in the following.

Measurements of heat supply from 16 terrace houses in Kulladal, a suburb of Malmö in Sweden, are used to estimate the heat load as a function of the time of the day. The

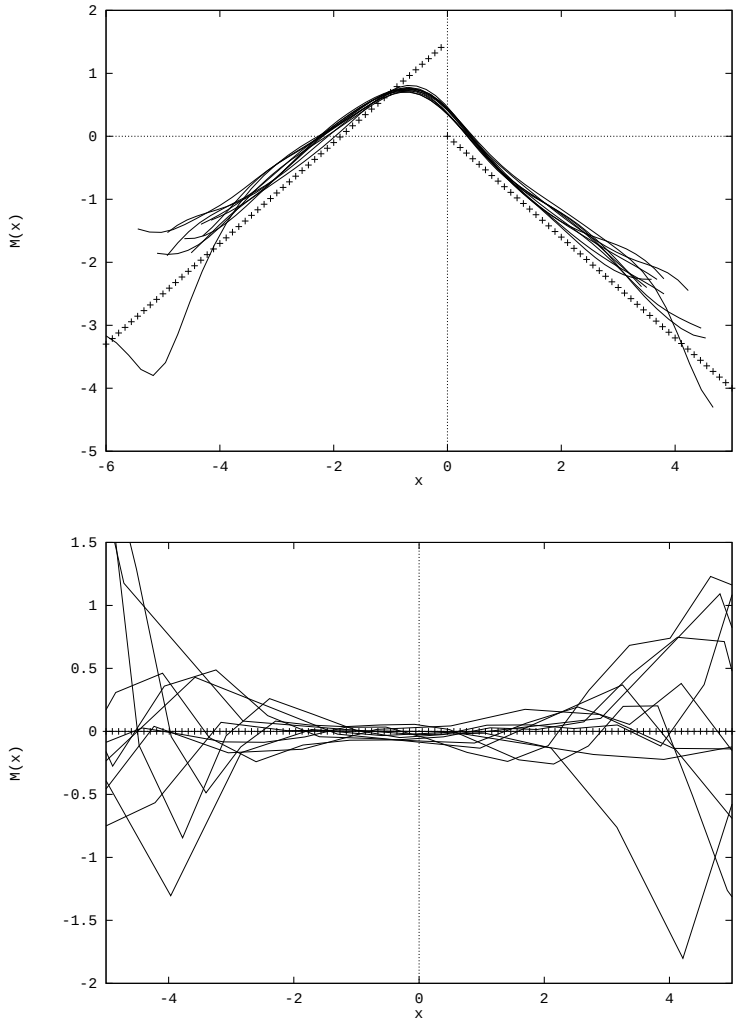

Figure 2.2: Estimated $M(x)$ for model A (above) and model B

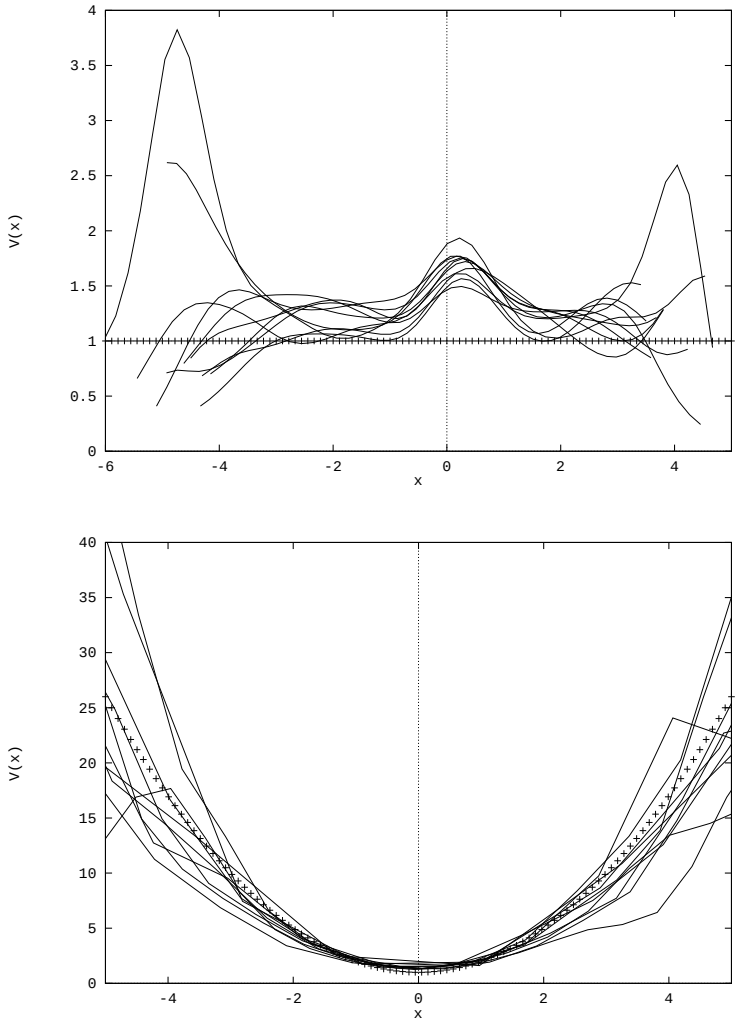

Figure 2.3: Estimated $V(x)$ for model A (above) and model B

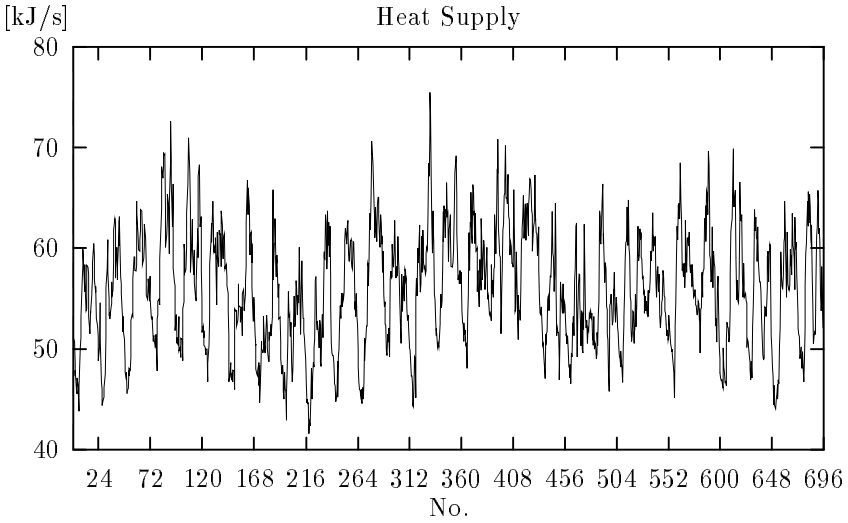

Figure 2.4: Supplied heat to 16 houses in Kulladal/Malmö during February 1989.

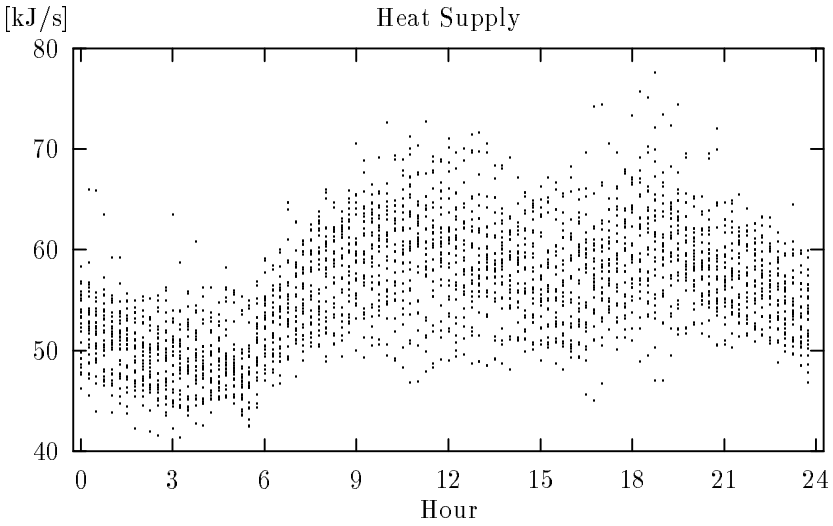

Figure 2.5: Supplied heat versus time of day and night.

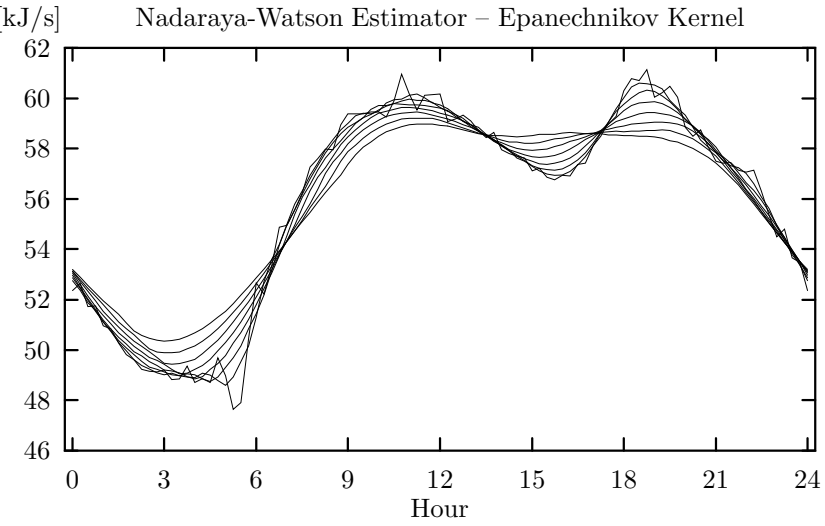

Figure 2.6: Smoothing of the diurnal power load curve using an Epanechnikov kernel and bandwidths 0.125, 0.625,. . . , 3.625.

heat supply was measured every 15 minutes for 27 days. The transport delay between the consumers and the measuring instruments is not more than a couple of minutes.

In the present investigation it is assumed that the heat supply can be related to the time of day, and that there is no difference between the days for the considered period. Then the regression curve is a function only of the time of day, and it is this functional relationship, which will be considered.

An approach for describing the dynamics, using filtering of explanatory variables such as e.g. weather variables, is shown later on.

Thus, the assumption of independence between measurements is not fulfilled. The consequence of this is that the information about the regression function contained in data is minor compared to a set of data with the same number of independent observations.

Figure 2.6 shows the Epanechnikov curve estimate calculated for a spectrum of bandwidths. It is clear, that the characteristics of the non-smoothed averages gradually disappear, when the bandwidth is increased.

Bringing the Cross-Validation function into action for bandwidth selection on the original data gives the result shown in Figure 2.7. Obviously the minimum of the CV-function is obtained for a bandwidth close to 1.3. This value most likely somewhat

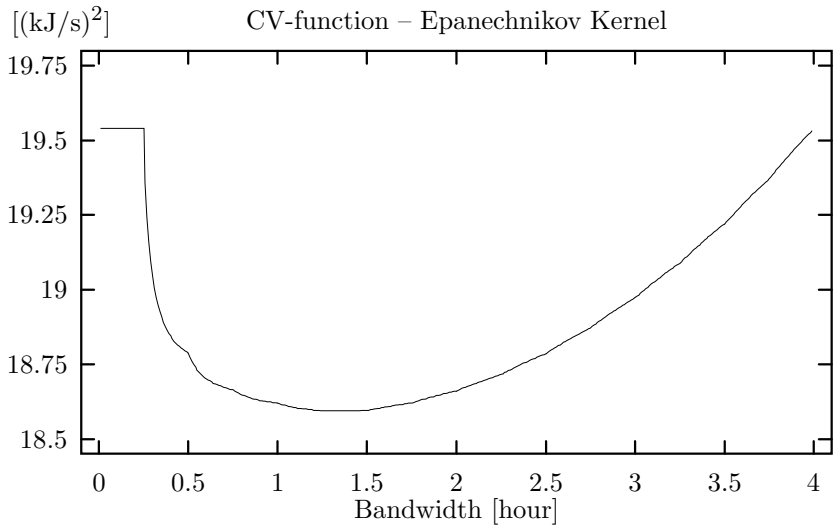

Figure 2.7: CV-function applied on the original data using the Epanechnikov kernel.

bigger than the visually chosen, and the drop at 5.15 in the morning will clearly disappear in a curve estimate with this bandwidth.

These results indicate the need of non-parametric curve estimates in which the bandwidth is not only a function of the density of the predictor variables, but for instance also of the gradient of the regression curve.

If a Gaussian kernel is used in the smoothing procedure the effect of choosing varying bandwidth is shown in Figure 2.8. Clearly it is difficult visually to observe any difference between the curve estimates obtained with the Epanechnikov and Gaussian kernel.

A general conclusion : The choice of the kernel (window) is not important as long as it is reasonably smooth. *It is the choice of the bandwidth that matters*.

2.4.5 Non-parametric estimation of dependence on external variables – An example

This examples illustrates the use of kernel estimators for exploring the (static) dependence on external variables. Data are collected in the district heating system in Esbjerg, Denmark, during the period August 14 to December 10, 1989.

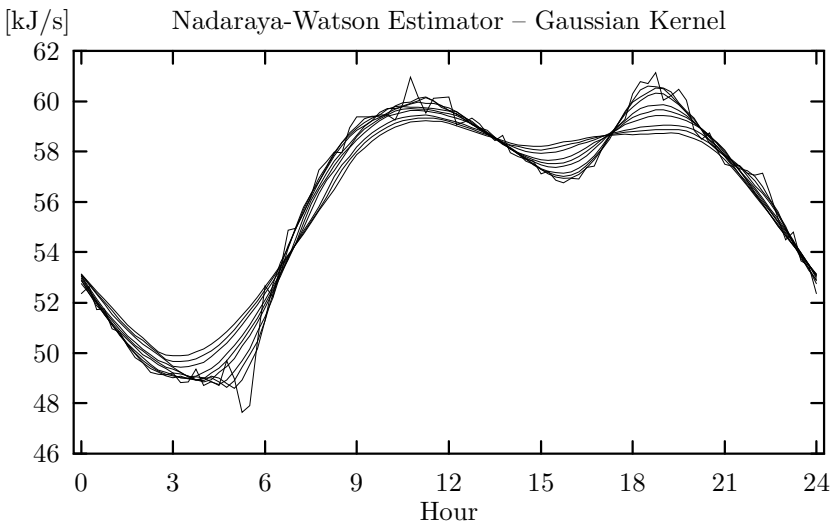

Figure 2.8: Smoothing of the diurnal power load curve using a Gaussian kernel and bandwidths 0.025, 0.125,. . . , 0.925.

The dependent variable is the heat load in a district heating system, and the independent variables are the two most influential explanatory variables in district heating systems, i.e. ambient air temperature and supply temperature. The data will also be used later for estimating some parametric non-linear time series models.

For the estimation of the regression surface the kernel method is applied using the Epanechnikov kernel (2.15). A product kernel is used, i.e. the power load estimate, p, at time t, at ambient air temperature a and for supply temperature s is estimated as

$$\hat{p}_{h_t,h_a,h_s}(t,a,s) = \frac{\sum_{i=1009}^{3843} p_i K_{h_t}(t-t_i) K_{h_a}(a-a_i) K_{h_s}(s-s_i)}{\sum_{i=1009}^{3843} K_{h_t}(t-t_i) K_{h_a}(a-a_i) K_{h_s}(s-s_i)}. \tag{2.38}$$

where $K_h(u) = h^{-1}k(u/h)$. The bandwidths were chosen using a combination of Cross-Validation and visual inspection.

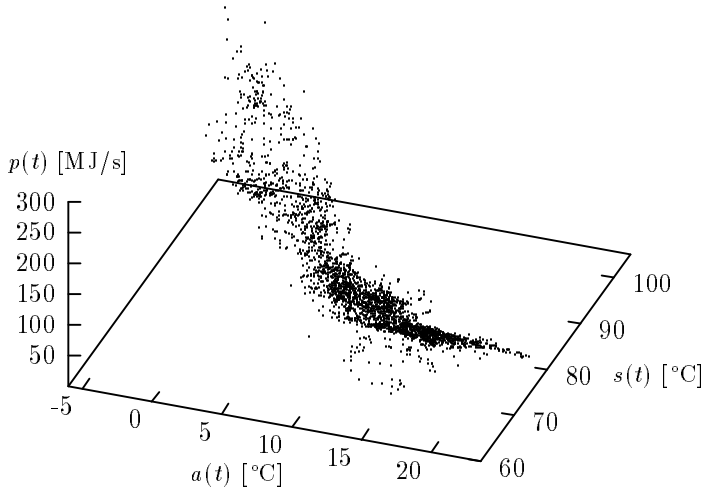

Figure 2.9: Heat load $p(t)$ versus ambient air temperature $a(t)$ and supply temperature $s(t)$ from August 14th to December 10th in Esbjerg, Denmark.

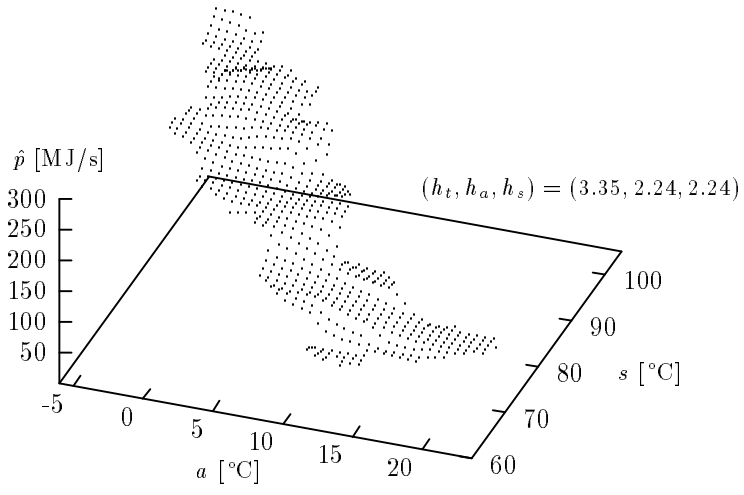

Figure 2.10: Kernel estimate of the dependence on time of day, t_d, ambient air temperature, a and supply temperature, s, shown at $t_d = 7$.

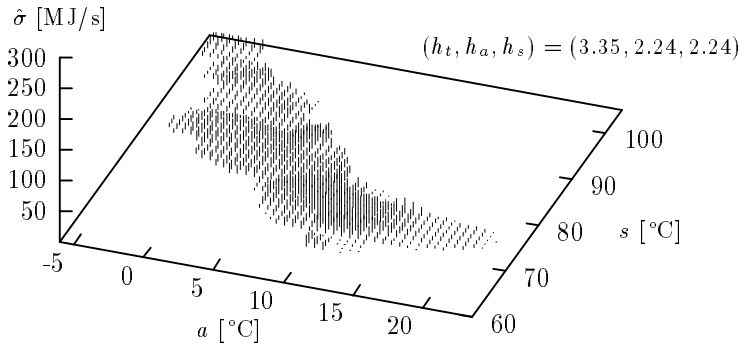

Figure 2.11: Point-wise estimate of standard deviation of the surface estimates.

2.5 Smoothing

The kernel estimates considered in the previous section provide a local *constant* estimate of the response or dependent variable Y as a function of the predictor or independent variable X. Hence the kernel estimate produces a smooth estimate of the variation which are less variable than Y itself. This tool is therefore also called a *kernel smoother*.

In general a *smoother* is a tool for estimating the response variable as a function of the predictor variables. An obvious generalization of the principles of using a local constant approximation is to use a local polynomial approximation.

In this section some other smoothing principles will be outlined and some applications related to time series analysis will be given. First a basic smoothing principle, namely the bin smoother, will be introduced. Secondly, the principles of smoothing by *local weighted least squares* regression is outlined. This method is generalized using the *conditional parametric models*, and extended to time series models by introducing the *conditional parametric ARX-models*. Finally, some examples will be provided.

2.5.1 Regressogram (Bin smoother)

The *bin smoother*, also called *regressogram* [Tukey 1961], is simply a rectangular kernel estimator based on a partitioning of the predictor values into a number of regions, and then averaging the response in each region. In the following the regressogram estimates for the conditional mean and variance are defined.

Definition 2.2 (Regressogram – bin smoother). Let $I_1, \ldots, I_{d_N}$ be a partition of an interval $[a, b]$ made up of intervals of length h_j, and denote $I_x = I_j$ for $x \in I_j$, and $h(x) = h_j$ for $x \in I_j$. The regressogram estimators $\hat{M}$ and $\hat{V}$ for the conditional mean and variance, respectively, are defined by

$$\hat{M}(x) = \frac{(Nh(x))^{-1} \sum_{t=1}^{N} Y_t I\{X_t \in I_x\}}{(Nh(x))^{-1} \sum_{t=1}^{N} I\{X_t \in I_x\}} \tag{2.39}$$

$$\hat{V}(x) = \frac{(Nh(x))^{-1} \sum_{t=1}^{N} (Y_t - \hat{M}(x))^2 I\{X_t \in I_x\}}{(Nh(x))^{-1} \sum_{t=1}^{N} I\{X_t \in I_x\}} \tag{2.40}$$

where $I\{\cdot\}$ is an indicator function. ▲

Often the intervals are of equal length. Typically five regions are chosen, and the cutoff points are selected so that there is approximately an equal number of points in each region [Hastie & Tibshirani 1990].

The estimate provided by the bin smoother is not very smooth because it jumps between the intervals.

2.5.2 Locally-weighted polynomial regression

Let us return to the non-parametric least squares estimate as described in Section 2.3.3 where the following weighted least squares problem is considered

$$\arg\min_{\theta} \frac{1}{N} \sum_{s=1}^{N} w_s(\boldsymbol{x})(Y_s - \theta)^2 \tag{2.41}$$

The underlying model is obviously that θ is (locally) constant, i.e. $Y_t = \theta + \epsilon_t$. The estimate becomes, however, a function of $\boldsymbol{x}$. Note the similarity with adaptive LS and adaptive RLS – see for instance Chapter 2 and 9 in [Madsen 1995*a*].

If the curvature of the true regression curve is substantial then only a small fraction of the observations can be used to estimate a local constant if that estimate shall provide a reasonable (local) estimation of the regression curve. Therefore it is obvious to consider a generalization to local regression models where, for instance, a local polynomial approximation can be used. This is the main idea behind locally-weighted regression, which is also called *Locally-Weighted Least Squares (LWLS).*

Model and estimation

The underlying model for local regression is

$$Y_t = \mu(\boldsymbol{X}_t) + \epsilon_t \tag{2.42}$$

where the function μ is assumed to be smooth and estimated by fitting a polynomial $P(\boldsymbol{X}_t, \boldsymbol{x})$ model within a sliding window, and parameterized such that

$$\mu(\boldsymbol{x}) = P(\boldsymbol{x}, \boldsymbol{x}) \tag{2.43}$$

For each fitting point, and for a given parameterization $\boldsymbol{\theta}$ of the polynomium, the following locally least squares problem is considered

$$\arg\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{s=1}^{N} w_s(\boldsymbol{x})(Y_s - P(\boldsymbol{X}_s, \boldsymbol{x}))^2 \tag{2.44}$$

The local least squares estimate of $\mu(\boldsymbol{x})$ is now $\hat{\mu}(\boldsymbol{x}) = \hat{P}(\boldsymbol{x}, \boldsymbol{x})$.

Example 2.4 (Locally-weighted linear model). Consider the local linear approximation

$$P(X_t, x) = \theta_0 + \theta_1(X_t - x) \tag{2.45}$$

By taking $\boldsymbol{\theta} = (\theta_0, \theta_1)$ the parameters are estimated locally at each fitting point x by solving

$$\arg\min_{\theta} \frac{1}{N} \sum_{s=1}^{N} w_s(x)(Y_s - (\theta_0 + \theta_1(X_s - x))^2 \tag{2.46}$$

and the local linear estimate of $\mu(x)$ is

$$\hat{\mu}(x) = \hat{\theta}_0 \tag{2.47}$$

♦

Note that each least squares problem produces $\hat{\mu}(\boldsymbol{x})$ for that single point – to estimate at additional points a new locally weighted problem has to be solved.

Weight functions

In Section 2.3.3 the weights $w_s(\boldsymbol{x})$ were defined using the kernel; but in the non-parametric least squares setting the weight normally arises using a nowhere increasing function W. Some weight functions are given in Table 2.1.

Name	Weight function
Box	$W(u) = \begin{cases} 1, & u \in [0;1) \\ 0, & u \in [1;\infty) \end{cases}$
Triangle	$W(u) = \begin{cases} 1-u, & u \in [0;1) \\ 0, & u \in [1;\infty) \end{cases}$
Tri-cube	$W(u) = \begin{cases} (1-u^3)^3, & u \in [0;1) \\ 0, & u \in [1;\infty) \end{cases}$
Gauss	$W(u) = \exp(-u^2/2)$

Table 2.1: Some weight functions.

In the case of a spherical weight function the weight is given by

$$w_s(\boldsymbol{x}) = W(||\boldsymbol{x} - \boldsymbol{X}_s||/h(\boldsymbol{x})), \tag{2.48}$$

where $||\boldsymbol{x} - \boldsymbol{X}_s||$ is the Euclidean distance between $\boldsymbol{x}$ and $\boldsymbol{X}_s$. A product kernel is characterized by distances being calculated one dimension at a time, i.e.

$$w_s(\boldsymbol{x}) = \prod_{i=1}^{q} W(|x_i - X_s^{(i)}|/h(\boldsymbol{x})), \tag{2.49}$$

where q is the dimension of $\boldsymbol{x}$. The scalar $h(\boldsymbol{x}) > 0$ is called the bandwidth. Compare with (2.20).

For the choice of the *bandwidth* either a *metric* or a *rank* principle can be used. If $h(\boldsymbol{x})$ is chosen so that a certain fraction (α) of the observations ($\boldsymbol{X}_s$) is within the bandwidth it is denoted a *nearest neighbor (NN) bandwidth*. If $h(\boldsymbol{x})$ is constant for all values of $\boldsymbol{x}$ it is denoted a *fixed bandwidth*. If $\boldsymbol{x}$ has dimension of two or larger, a *scaling* of the individual elements of $\boldsymbol{x}$ before applying the method should be considered, see e.g. [Cleveland & Devlin 1988]. Other coordinate systems than that dictated by $\boldsymbol{x}$ might also be relevant as e.g. Mahalanobis type measures ($\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}$) – compare also with (2.17).

Example 2.5 (Locally weighted LS). Figure 2.12 shows the locally LWLS estimate using a local constant model and a NN method, whereas Figure 2.13 shows the results using a local quadratic model.

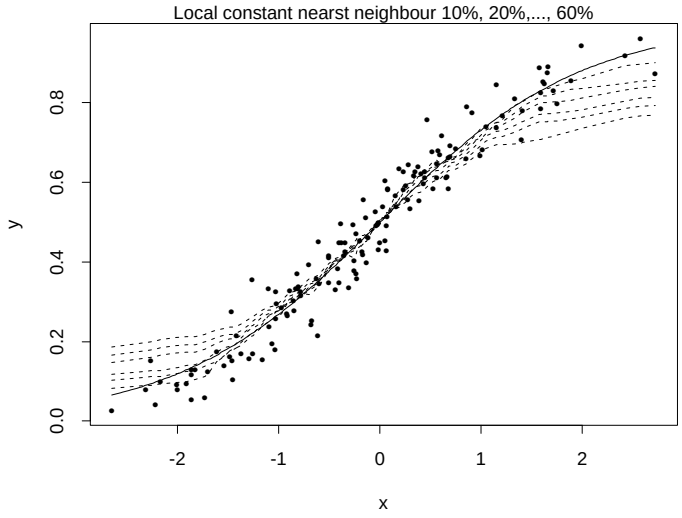

Figure 2.12: LWLS – Local constant approximation.

It is clearly seen that the bias is much larger for the local constant model than for the local quadratic model.

♦

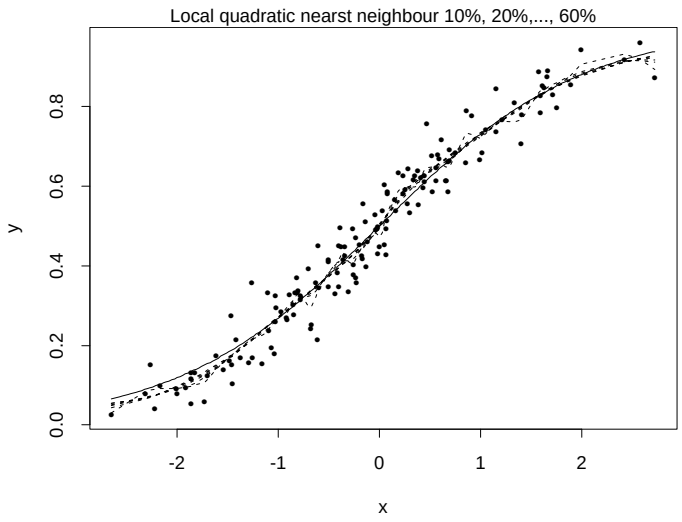

Figure 2.13: LWLS – Local quadratic approximation.

2.6 Conditional parametric models

This class of models is a further generalization to models which are (globally) linear in the some of the regressors; but the coefficients of the model are assumed to change smoothly as an unknown function of other variables. The models are also called *varying-coefficient models* [Hastie & Tibshirane 1993]. When all coefficients depend on the same variable the model is, however, denoted a *conditional parametric model*.

The model

Assume that observations of the response Y_t and the predictor variables $\boldsymbol{X}_t$ and $\boldsymbol{Z}_t$ exist for observation numbers $t = 1, \ldots, N$. The *conditional parametric model* is defined as

$$Y_t = \boldsymbol{Z}_t^T \boldsymbol{\theta}(\boldsymbol{X}_t) + \epsilon_t; \;\; t = 1, \ldots, N, \tag{2.50}$$

where $\{\epsilon_t\}$ is zero mean iid random variables.

The parameters of the above model is again estimated locally by using the straightforward generalization of (2.44), namely by solving

$$\arg\min_{\theta} \frac{1}{N} \sum_{s=1}^{N} w_s(\boldsymbol{x})(Y_s - \boldsymbol{Z}_s^T \boldsymbol{\theta}(\boldsymbol{x}))^2 \tag{2.51}$$

where $\boldsymbol{\theta}(\boldsymbol{x})$ most frequently is parameterized using polynomials, as in (2.44). Regarding the weight function we refer to the previous section. Let us illustrate the method by an example.

Example 2.6 (Locally-weighted linear model). Assume that the global variable is $\boldsymbol{Z}_t^T = (1, Z_{1t})$, and $\dim(\boldsymbol{x}) = 1$.

In a *constant locally-weighted linear* setting $\boldsymbol{\theta}^T = (\theta_0, \theta_1)$.

In a *quadratic locally-weighted linear* setting were the local constant parameter assumption is replaced by a local quadratic assumption of $\boldsymbol{\theta}(x)$ the above value of $\boldsymbol{Z}_t$ is replaced by

$$\boldsymbol{Z}_t^T = (1, X_t, X_t^2, Z_{1t}, Z_{1t}X_t, Z_{1t}X_t^2) \tag{2.52}$$

and the parameter vector is now

$$\boldsymbol{\theta}^T = (\theta_{00}, \theta_{01}, \theta_{02}, \theta_{10}, \theta_{11}, \theta_{12}) \tag{2.53}$$

After (locally) estimating the parameters in (2.53) the local estimates of θ_0 and θ_1 are

$$\begin{aligned} \hat{\theta}_0(x) &= \hat{\theta}_{00}(x) + \hat{\theta}_{01}(x)x + \hat{\theta}_{02}(x)x^2 \\ \hat{\theta}_1(x) &= \hat{\theta}_{10}(x) + \hat{\theta}_{11}(x)x + \hat{\theta}_{12}(x)x^2 \end{aligned}$$

♦

2.6.1 Conditional parametric ARX-models

In this section we shall use the previously defined varying-coefficients models to define a very useful class of non-linear ARX models, where the coefficients depend on some other external variables. Most of the text and the discussion is taken from [Nielsen, Nielsen & Madsen 1997]. In this reference also the effect of correlation in the input is studied in some detail.

It is well known that a linear ARX-model can be written in the form (2.50), where t is the time index, $\boldsymbol{\theta}(\cdot)$ is a constant parameter vector and $\boldsymbol{Z_t}$ contains input and lagged values of Y_t. Hence, if the parameters depend on an external variable, as for instance X_{t-m}, it is obvious to define *conditional parametric ARX-models* as models written in the form

$$Y_t = \sum_{i \in L_y} a_i(X_{t-m}) Y_{t-i} + \sum_{i \in L_u} b_i(X_{t-m}) U_{t-i} + \epsilon_t, \tag{2.54}$$

where t is the time index, Y_t is the output variable, X_t and U_t are input variables, $\{\epsilon_t\}$ is zero mean white noise, L_y and L_u are sets of positive integers defining the

autoregressive and input lags in the model, and m is a positive integer. $a_i(\cdot)$ and $b_i(\cdot)$ are unknown but smooth functions which are to be estimated. Extensions to multivariate X_t and U_t are strait forward.

A simulation study in [Nielsen et al. 1997] has shown that it is reasonable to use the techniques from the conditional parametric models to estimate parameters in the above defined class of non-linear dynamic models.

2.6.2 Functional-coefficient AR-models

The functional-coefficient AR-models ([Chen & Tsay 1993] is a model where the coefficients are function of previous values of the process itself. More specifically the *functional-coefficient AR-model* (FAR model) is a model on the form

$$Y_t = \sum_{i=1}^{p} a_i(\boldsymbol{Y}_{t-1}^*)Y_{t-i} + \epsilon_t, \tag{2.55}$$

where $a_i(\cdot)$ are unknown but smooth functions of previous values of the process as collected in the vector $\boldsymbol{Y}_{t-1}^*$. This model is a special case of the *state-dependent models*, which will be described in more detail in Chapter 7. It is also seen that for instance the EXPAR model is a special case of the above FAR model. Note also that the functional-coefficient AR-model is different from the fractional autoregressive model defined in Chapter 1.

2.6.3 Case study

Let us illustrate the use of the conditional parametric ARX model by consider real data and a real problem which focus on modelling the dynamic relation of temperatures in a district heating system. In such a system the energy needed for heating and hot tapwater in the individual households is supplied from a central heating utility. The energy is distributed as hot water through a system of pipelines covering the area supplied.

Models of the relationship between supply temperature and inlet temperature are of high interest from a control point of view. Previous studies have lead to a library of ARX-models with different time delays and with a diurnal variation in the model parameters. Methods for on-line estimating of the varying time delay as well as a controller which takes full advantage of this model structure have previously been published [Søgaard & Madsen 1991, Palsson, Madsen & Søgaard 1994, Madsen, Nielsen & Søgaard 1996].

A more direct approach is, as suggested in [Nielsen et al. 1997], to use one ARX-model but with parameters replaced by smooth functions of the flow rate. This approach is addressed in the following. A big advantage for this approach is that the need for on-line estimation of the time delay is eliminated.

Data

The data covers the periods from 1 April until 31 May and 1 September until 17 December, 1996. Data consists of five minute samples of supply temperature and flow at the plant together with the network temperature at a major consumer, consisting of 84 households. In 1996 that consumer used 1.2% of the produced energy. Based on the observations half-hour averages were calculated, and these values are used in the following.

Non-linear FIR and ARX models

The network temperature is modelled by models of the structure (2.54), where one time step corresponds to 30 minutes, Y_t represents the network temperature, U_t represents the supply temperature, and X_t represents a filtered value of the flow.

Below results corresponding to two different model structures are presented, (i) the *non-linear Finite Impulse Response (FIR) model*

$$Y_t = \sum_{i=0}^{30} b_i(X_t)U_{t-i} + \epsilon_t \tag{2.56}$$

and the *non-linear ARX-model*

$$Y_t = a(X_t)Y_{t-1} + \sum_{i=3}^{15} b_i(X_t)U_{t-i} + \epsilon_t \tag{2.57}$$

It will be illustrated how some of the terms used for ordinary linear models, like the stationary gain and poles, can be used also for this type of non-linear models.

Results

For both the FIR-model and the ARX-model described in the previous section local quadratic estimates and nearest neighbor bandwidths are used. The coefficient func-

tions are estimated at 50 equally spaced points ranging from the 2% to the 98% quantile of the filtered flow.

The coefficient functions of the FIR-model are estimated using a nearest neighbor (NN) bandwith for α equal to $0.1, 0.2, \ldots, 0.9$. In Figure 2.14 the impulse response as a function of the flow is displayed for $\alpha = 0.4$. Equivalent plots for the remaining bandwidths revealed that for $\alpha \leq 0.2$ the fits are too noisy, whereas in all cases sufficiently smoothness is obtained at $\alpha = 0.5$. Only minor differences in the fits are observed for $\alpha \in \{0.3, 0.4, 0.5\}$.

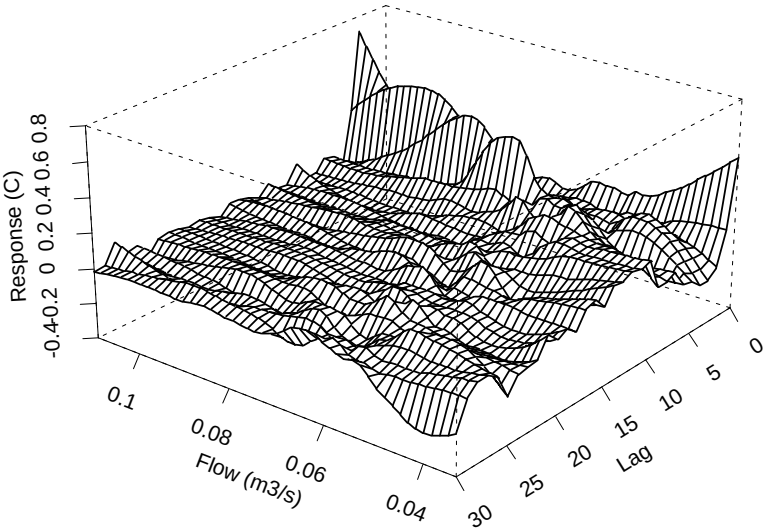

Figure 2.14: Impulse response of the FIR-model.

In Figure 2.15 a contour plot corresponding to Figure 2.14 is shown. From the plot the varying time delay of the system is revealed, it seems to vary from three lags when the flow is large to approximately ten lags when the flow is near its minimum. The peak at lag zero for the lower flows is clearly an artifact.

The residuals of the FIR-model show diurnal variation. The sample inverse autocorrelation function of the residuals indicates that an AR(1)-model can account for most of the correlation.

Based on the results obtained for the FIR-model the ARX-model described in Section 2.6.3 is purposed. The coefficient functions of the model are estimated for α equal to 0.3, 0.4, and 0.5. Very similar results are obtained for the three different bandwidths. For $\alpha = 0.4$ the impulse response as a function of the flow is displayed in Figure 2.16. The varying time delay is clearly revealed.

In Figure 2.17 the stationary gain of the two models and the pole of the ARX-model

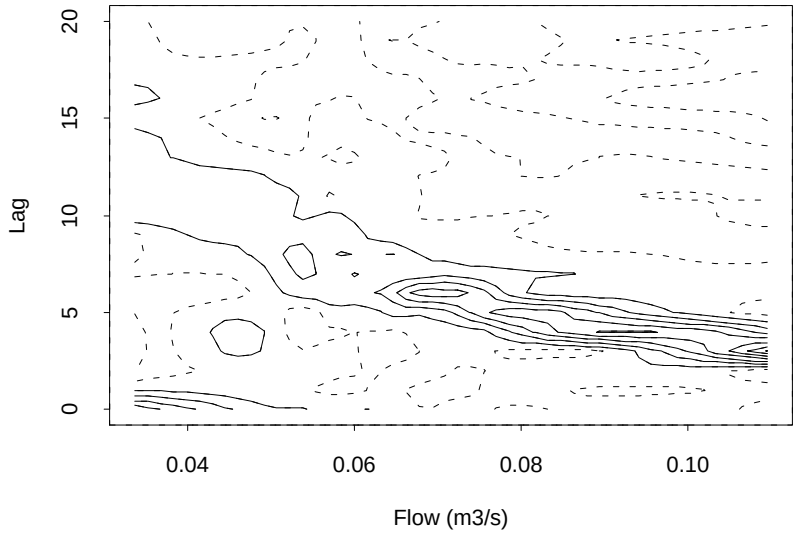

Figure 2.15: Contour plot of the impulse response of the FIR-model ($\alpha = 0.4$). The contour lines is plotted from -0.1 to 0.7 ^oC in steps of 0.1 ^oC, lines corresponding to non-positive values are dotted.

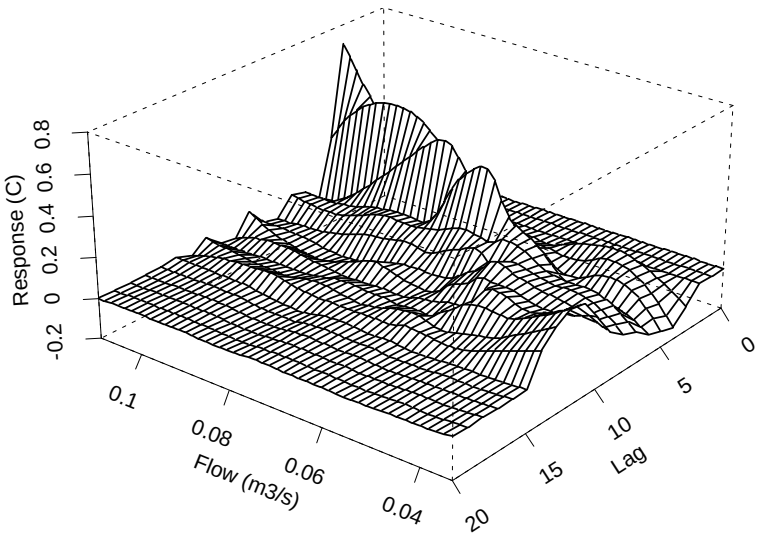

Figure 2.16: Impulse response of the ARX-model.

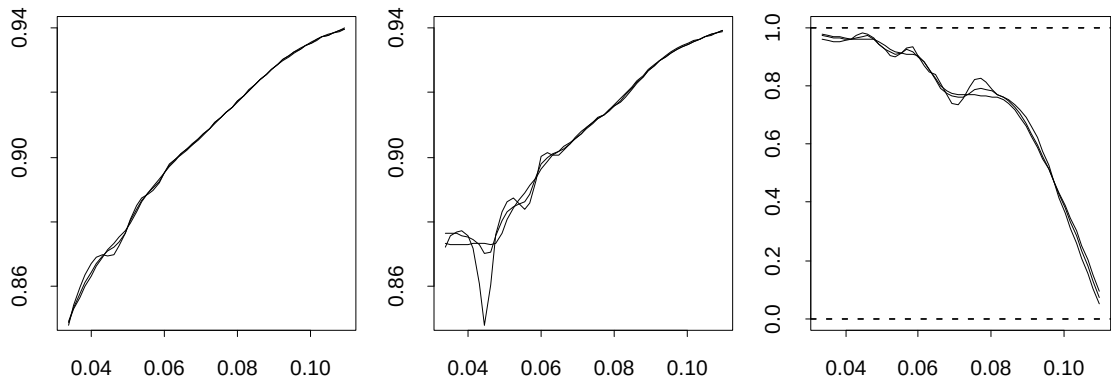

Figure 2.17: Stationary gain of the FIR-model (left) and ARX-model (middle) and the pole of the ARX-model (right) all plotted against the flow and for NN bandwith α equal to 0.3, 0.4, and 0.5.

are shown. The model describes the relation between the temperature at the plant and the network temperature. Hence, in the case of no temperature loss the stationary gain (the stationary gain for constant flow) must be 1.

From the estimated values of the stationary gain it is seen that the temperature loss changes from 6% when the flow is large to 12-15% when the flow is small. This clearly illustrates that no single values exists for the stationary gain as for the linear models; in the considered case the stationary gain and the location of the pole depend on the flow.

Chapter 3
Identification

3.1 Introduction

This chapter focus on various aspects related to the *identification* of a non-linear (or non-Gaussian) time series model. The identification step involves determining the structure of the functional description of a (possibly) non-linear relation. A method, which is related to the kernel estimation of the conditional mean described in the previous chapter, is outlined in Section 3.2.

The identification of the functional relation is based on the assumption that the number of needed lags in the function is identified. In linear time series analysis the number of needed lags is traditionally determined by considering the sample ACF and sample PACF for the time series. However, these tools are not suitable for identifying some non-linear lag dependency. Therefore some equivalent tools, called the *Lag Dependent Functions* are introduced in Section 3.3, for identifying the lag dependency in the more general non-linear case.

3.2 Identification of a functional relation

In this section a method for detecting nonlinearities and determining the structure of the functional relationship is considered. The method uses estimated cumulative conditional mean and variance functions.

Assume that the conditional mean and variance of the stationary process $\{X_t\}$ given

$X_{t-1}, X_{t-2}, \ldots$ only depend on X_{t-1}. This holds for the very important nonlinear model

$$X_t = g(X_{t-1}) + h(X_{t-1})\epsilon_t, \tag{3.1}$$

where $\{\epsilon_t\}$ is white noise with unit variance.

In section 2.4.2, non-parametric estimates of the conditional mean and variance were used to obtain non-parametric estimates of the functions in the nonlinear model (3.1). For convenience we will use the following notation :

$$\lambda(x) = E[x_t \mid x_{t-1} = x], \tag{3.2}$$
$$\gamma(x) = V[x_t \mid x_{t-1} = x]. \tag{3.3}$$

Given a time series $\{X_1, \ldots, X_n\}$ the above estimates can therefore be used to *identify* the functions. It is important to note that asymptotic results for the estimators of λ and γ are only useful *pointwise*.

Introduce *cumulative* versions of the conditional mean and variances

$$\Lambda(\cdot) = \int_a^{\cdot} \lambda(x)dx \tag{3.4}$$
$$\Gamma(\cdot) = \int_a^{\cdot} \gamma(x)dx \tag{3.5}$$

where attention is restricted to some fixed interval $[a, b]$. Obviously estimates of Λ and Γ can be used to *identify* the nonlinear functions as well.

The advantages of considering the cumulative versions are that

- It is possible to derive functional limit theorems (whereas asymptotic results for kernel estimators of λ and γ are only useful pointwise).
- As an implication of the item above, tests can be constructed for testing a particular model.
- The estimators for Λ and Γ are relatively insensitive to variations in the bandwidth compared to the estimator for λ and γ.

In the following we restrict the attention to the *regressogram* [Tukey 1961] as defined in Section 2.5.1.

Let $I_1, \ldots, I_{d_n}$ be a partition of $[a, b]$ made up of intervals of equal length h_n, and denote $I_x = I_j$ for $x \in I_j$. The regressogram estimators $\hat{\lambda}$ and $\hat{\gamma}$ follows using (2.39)

and (2.39) with $Y_t \sim X_t$ and $X_t \sim X_{t-1}$, i.e.

$$\hat{\lambda}(x) = \frac{(nh_n)^{-1}\sum_{t=1}^{n} X_t I\{X_{t-1} \in I_x\}}{(nh_n)^{-1}\sum_{t=1}^{N} I\{X_{t-1} \in I_x\}} \tag{3.6}$$

$$\hat{\gamma}(x) = \frac{(nh_n)^{-1}\sum_{t=1}^{n} (X_t - \hat{\lambda}(x))^2 I\{X_{t-1} \in I_x\}}{(nh_n)^{-1}\sum_{t=1}^{n} I\{X_{t-1} \in I_x\}} \tag{3.7}$$

Now introduce the estimators

$$\hat{\Lambda}(\cdot) = \int_a^{\cdot} \hat{\lambda}(x)dx, \tag{3.8}$$

$$\hat{\Gamma}(\cdot) = \int_a^{\cdot} \hat{\gamma}(x)dx. \tag{3.9}$$

Simulation studies have shown that there should be at least five observations in each interval [McKeague & Zhang 1994]. It is clear that more sophisticated estimators (e.g. the kernel estimators considered in the previous chapter) would outperform the regressograms in estimating λ and γ. However, the integration has a strong smoothing effect, so the gain obtained by using kernel estimators applied to the cumulative characteristics might be small. For higher order models the kernel estimators probably would be useful.

Under some weak conditions (e.g. that the marginal density of X_t does not vanish on $[a, b]$ – see [McKeague & Zhang 1994] for a further description of the conditions) an asymptotic $100(1-\alpha)\%$ *confidence band* for Λ is given by

$$\hat{\Lambda}(x) \pm c_\alpha n^{-1/2} \hat{H}(b)^{1/2} \left(1 + \frac{\hat{H}(x)}{\hat{H}(b)}\right); \qquad x \in [a, b], \tag{3.10}$$

where $\hat{H}(\cdot) = \int_a^{\cdot} \hat{\gamma}/\hat{f}dx$, and $\hat{f}$ is the histogram estimator for f, i.e.

$$\hat{f}(x) = (nh_n)^{-1}\sum_{t=1}^{n} I\{x_{t-1} \in I_x\} \tag{3.11}$$

c_α is the upper α quantile of the distribution $\sup_{t\in[0,1/2]} |B^0(t)|$, where B^0 is a Brownian bridge. Values of c_α are given in Table 3.2.

A confidence band for Γ can be found in a similar way.

Test of simple hypotheses $\lambda = \lambda_0$ and $\gamma = \gamma_0$, where λ_0 and γ_0 are specified, can be made by checking whether the confidence bands contain Λ_0 and Γ_0.

α	0.99	0.95	0.90	0.75	0.50
c_α	1.552	1.273	1.133	0.920	0.720

Table 3.1: Some percentiles of distribution $\sup_{t\in[0,1/2]} |B^0(t)|$.

3.3 Identification of lag dependencies

In classical time series analysis the sample autocorrelation function ($SACF$) and the sample partial autocorrelation function ($SPACF$) has gained wide application for identification of the model order, or lag dependency, for linear time series models. For non-linear time series these tools are not sufficient since they only address linear dependencies.

This section describes generalizations applicable for identification of the model order, in terms of the needed lags, for non-linear time series models. The presentation follows [Nielsen & Madsen 2001]. The generalized functions are the *Lag Dependence Functions* (LDF), and the *Partial Lag Dependence Functions* (PLDF). Furthermore a measure of the distance from linearity, called Non-linear Lag Dependence Function (NLDF), and a Cross Dependence Function is briefly mentioned.

The tools are founded on the non-parametric methods as described in Chapter cha:kernel. The generalizations do not prescribe a particular smoothing technique. In fact, when the smoother is replaced by a linear regression the sample versions of LDF and PLDF reduce to close approximations of the sample versions of ACF and PACF. The generalizations allow for graphical presentations, very similar to those used for SACF and SPACF.

Under a hypothesis of independence bootstrap confidence intervals [Efron & Tibshirani 1993] of the lag dependence function are readily calculated, and it is proposed that these can also be applied for the partial lag dependence function. Furthermore, under a specific linear hypothesis, bootstrapping can be used to construct confidence intervals for the non-linear lag dependence function. Since the size of the confidence intervals depend on the flexibility of the smoother used it is informative to apply the tools using a range of smoothing parameters.

The lag dependence function and the non-linear lag dependence function are readily calculated in that only univariate smoothing is needed, whereas multivariate smoothing or backfitting is required for the application of the partial lag dependence function.

In Section 3.3.1 the study is motivated by considering a simple deterministic non-linear process for which the sample autocorrelation function is non-significant. Section 3.3.2 describes some preliminaries leading to the generalization. The proposed tools are described in Sections 3.3.3, 3.3.4, and 3.3.5 and bootstrapping is considered in Section 3.3.6. Examples of application by considering simulated linear and non-linear processes and the Canadian lynx data [Moran 1953] are found in Section 3.3.7.

3.3.1 Motivation

The sample autocorrelation function, commonly used for structural identification in classical time series analysis, measures only the degree of linear dependency. In fact deterministic series exists for which the sample autocorrelation function is almost zero, see also [Granger 1983]. Consider, for instance, the deterministic process:

$$X_t = 4X_{t-1}(1 - X_{t-1}) \tag{3.12}$$

for which Figure 3.1 shows 1000 values using $X_1 = 0.8$ and the corresponding sample autocorrelation function $SACF$ together with an approximative 95% confidence interval of the estimates under the hypothesis that the underlying process is i.i.d. Furthermore lagged scatter plots for lag one and two are shown. From the plot of the series and the $SACF$ the deterministic structure is not revealed. However, the lagged scatter plots clearly reveal that the series contains a non-linear dynamic dependency.

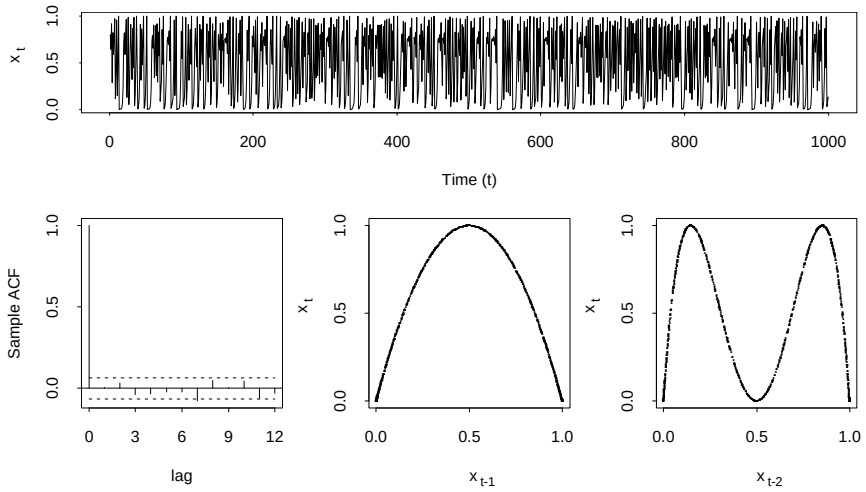

Figure 3.1: The time series (top), $SACF$ (bottom, left), X_t versus X_{t-1} (bottom, middle), and X_t versus X_{t-2} (bottom, right) for 1000 values the deterministic proces $X_t = 4X_{t-1}(1 - X_{t-1})$ with $X_1 = 0.8$.

In practice the series will often be contaminated with noise and it is then difficult to judge from the lagged scatter plots whether any dependence is present. Smoothing the lagged scatter plots will aid the interpretation but different smoothing parameters may result in quite different estimates. Therefore it is important to separate the variability of the smooth from the underlying dependence.

From Figure 3.1 it is revealed that, in principle, x_t can be regarded as a function of x_{t-k} for any $k > 0$, but $k = 1$ is sufficient, since x_t can be predicted exactly from x_{t-1} alone. This indicates that there may exist a non-linear equivalent to the partial autocorrelation function.

3.3.2 Multiple linear regression, correlation and partial correlation

In this section the relations between multiple linear regression, correlation, and partial correlation are presented with a focus on aspects leading to the generalized tools.

Estimates of correlation and partial correlation are closely related to values of the *coefficient of determination* (R-squared) obtained using linear regression models. The generalizations of the sample autocorrelation function $SACF$ and the sample partial autocorrelation function $SPACF$ are based on similar R-squared values obtained using non-linear models.

Consider the multivariate stochastic variable $(Y, X_1, \ldots, X_k)$. The squared multiple correlation coefficient $\rho^2_{0(1\ldots k)}$ between Y and $(X_1, \ldots, X_k)$ can be written [Kendall & Stuart 1961, p. 334, Eq. (27.56)]

$$\rho^2_{0(1\ldots k)} = \frac{V[Y] - V[Y \mid X_1, \ldots, X_k]}{V[Y]}. \tag{3.13}$$

Given observations $y_i, x_{1i}, \ldots, x_{ki};\ i = 1, \ldots, N$ of the stochastic variables $(Y, X_1, \ldots, X_k)$ and assuming normality the maximum likelihood estimate of $\rho^2_{0(1\ldots k)}$ is

$$R^2_{0(1\ldots k)} = \frac{SS_0 - SS_{0(1\ldots k)}}{SS_0}, \tag{3.14}$$

where $SS_0 = \sum (y_i - \sum y_i/N)^2$ and $SS_{0(1\ldots k)}$ is the sum of squares of the least squares residuals when regressing y_i linearly on $x_{1i}, \ldots, x_{ki}$ $(i = 1, \ldots, N)$. $R^2_{0(1\ldots k)}$ is also called the coefficient of determination of the regression. $R^2_{0(1\ldots k)}$ can be interpreted as the relative reduction in variance due to the regressors.

Hence it follows that when regressing y_i linearly on x_{ki} the coefficient of determination $R^2_{0(k)}$ equals the squared estimate of correlation between Y and X_k, and furthermore it follows that $R^2_{0(k)} = R^2_{k(0)}$.

The partial correlation coefficient $\rho_{(0k)|(1\ldots k-1)}$ between Y and X_k given $X_1, \ldots, X_{k-1}$ measures the extent to which, by using linear models, the variation in Y, which cannot be explained by $X_1, \ldots, X_{k-1}$, can be explained by X_k. Consequently, the partial correlation coefficient is the correlation between $(Y \mid X_1, \ldots, X_{k-1})$ and $(X_k \mid X_1, \ldots, X_{k-1})$, see also [Rao 1973, p. 270]. Using [Whittaker 1990, p. 140] we obtain

$$\rho^2_{(0k)|(1\ldots k-1)} = \frac{V[Y \mid X_1, \ldots, X_{k-1}] - V[Y \mid X_1, \ldots, X_k]}{V[Y \mid X_1, \ldots, X_{k-1}]}. \tag{3.15}$$

For $k = 1$ it is readily seen that $\rho^2_{(0k)|(1\ldots k-1)} = \rho^2_{0(1)}$. If the variances are estimated using the maximum likelihood estimator, assuming normality, it follows that an estimate

of $\rho^2_{(0k)|(1\ldots k-1)}$ is

$$R^2_{(0k)|(1\ldots k-1)} = \frac{SS_{0(1\ldots k-1)} - SS_{0(1\ldots k)}}{SS_{0(1\ldots k-1)}}. \tag{3.16}$$

Besides an estimate of $\rho^2_{(0k)|(1\ldots k-1)}$ this value can also be interpreted as the relative decrease in the variance when including x_{ki} as an additional predictor in the linear regression of y_i on $x_{1i}, \ldots, x_{k-1,i}$. Note that (3.16) may also be derived from [Ezekiel & Fox 1959, p. 193].

Interpreting $R^2_{0(1\ldots k)}$, $R^2_{0(k)}$, and $R^2_{(0k)|(1\ldots k-1)}$ as measures of variance reduction when comparing models [Anderson-Sprecher 1994], these can be calculated and interpreted for a wider class of models such as non-parametric and additive models.

In the following "$\tilde{}$" will be used above values of SS and R^2 obtained from models other than linear models.

3.3.3 Lag Dependence Function

Assume that observations $\{x_1, \ldots, x_N\}$ from a stationary stochastic process $\{X_t\}$ exists. It is readily shown that the squared $SACF(k)$ can be closely approximated by the coefficient of determination when regressing x_t linearly on x_{t-k}, i.e. $R^2_{0(k)}$.

This observation leads to a generalization of $SACF(k)$, based on $\tilde{R}^2_{0(k)}$ obtained from a smooth of the k-lagged scatter plot, i.e. a plot of x_t against x_{t-k}. The smooth is an estimate of the conditional mean $f_k(x) = E[X_t \,|\, X_{t-k} = x]$. Thus, the Lag Dependence Function in lag k, $LDF(k)$, is calculated as

$$LDF(k) = \text{sign}\left(\hat{f}_k(b) - \hat{f}_k(a)\right)\sqrt{(\tilde{R}^2_{0(k)})_+} \tag{3.17}$$

where a and b are the minimum and maximum over the observations and the subscript "$+$" indicates truncation of negative values. The sign is included to provide information about the direction of the average slope. The truncation is necessary to ensure that (3.17) is defined. However, the truncation will only become active in extreme cases. By definition $LDF(0)$ is set equal to one.

Due to the reasons mentioned above, when $\hat{f}_k(\cdot)$ is restricted to be linear, $LDF(k)$ is a close approximation of $SACF(k)$ and, hence, it can be interpreted as a correlation. In the general case $LDF(k)$ can be interpreted as (the signed square-root of) the part of the overall variation in x_t which can be explained by x_{t-k}.

3.3.4 Partial Lag Dependence Function

For the time series $\{x_1, \ldots, x_N\}$ the sample partial autocorrelation function in lag k, denoted $SPACF(k)$ or $\hat{\phi}_{kk}$, is obtainable as the Yule–Walker estimate of ϕ_{kk} in the $AR(k)$ model

$$X_t = \phi_{k0} + \phi_{k1} X_{t-1} + \ldots + \phi_{kk} X_{t-k} + e_t, \tag{3.18}$$

where $\{e_t\}$ is i.i.d. with zero mean and constant variance. An additive, but non-linear, alternative to (3.18) is

$$X_t = \varphi_{k0} + f_{k1}(X_{t-1}) + \ldots + f_{kk}(X_{t-k}) + e_t. \tag{3.19}$$

This model may be fitted using the backfitting algorithm [Hastie & Tibshirani 1990]. The function $f_{kk}(\cdot)$ can be interpreted as a partial dependence function in lag k when the effect of lags $1, \ldots, k-1$ is accounted for. If the functions $f_{kj}(\cdot),\ (j = 1, \ldots, k)$ are restricted to be linear then $\hat{f}_{kk}(x) = \hat{\phi}_{kk} x$ and the function can be uniquely identified by its slope $\hat{\phi}_{kk}$.

However, since the partial autocorrelation function in lag k is the correlation between $(X_t \,|\, X_{t-1}, \ldots, X_{t-(k-1)})$ and $(X_{t-k} \,|\, X_{t-1}, \ldots, X_{t-(k-1)})$, the squared $SPACF(k)$ may also be calculated as $R^2_{(0k)|(1\ldots k-1)}$, based on linear autoregressive models of order $k-1$ and k. Using models of the type (3.19) $SPACF(k)$ may then be generalized using an R-squared value obtained from a comparison of models (3.19) of order $k-1$ and k. This value is denoted $\tilde{R}^2_{(0k)|(1\ldots k-1)}$ and we calculate the Partial Lag Dependence Function in lag k, $PLDF(k)$, as

$$PLDF(k) = \text{sign}\left(\hat{f}_{kk}(b) - \hat{f}_{kk}(a)\right) \sqrt{(\tilde{R}^2_{(0k)|(1\ldots k-1)})_+}. \tag{3.20}$$

When (3.19) is replaced by (3.18) $PLDF(k)$ is a close approximation of $SPACF(k)$. As for $LDF(k)$, generally, $PLDF(k)$ cannot be interpreted as a correlation. However, $PLDF(k)$ can be interpreted as (the signed square-root of) the relative decrease in one-step prediction variance when lag k is included as an additional predictor. For $k = 1$ the model (3.19) corresponding to $k-1$ reduces to an overall mean and the R-squared value in (3.20) is thus $\tilde{R}^2_{0(1)}$, whereby $PLDF(1) = LDF(1)$ if the same smoother is used for both functions. For $k = 0$ the partial lag dependence function is set equal to one.

The non-linear additive autoregressive model (3.19) is can be fitted by the the backfitting algorithm [Hastie & Tibshirani 1990], see also [Nielsen & Madsen 2001] for details.

3.3.5 Strictly Non-Linear Lag Dependence

The lag dependence function described in Section 3.3 measures both linear and non-linear dependence. If, in the definition of $\tilde{R}^2_{0(k)}$, the sum of squares from a overall mean SS_0 is replaced by the sum of squares $SS^L_{0(k)}$ of the residuals from fitting a straight line to the k-lagged scatter plot, a measure of non-linearity is obtained. In this paper this will be called the strictly Non-linear Lag Dependence Function in lag k, or $NLDF(k)$. Hence

$$NLDF(k) = \text{sign}\left(\hat{f}_k(b) - \hat{f}_k(a)\right)\sqrt{\left(\frac{SS^L_{0(k)} - \widetilde{SS}_{0(k)}}{SS^L_{0(k)}}\right)_+}, \qquad (3.21)$$

where $\hat{f}_k$ is obtained as for (3.17), i.e. it is a smooth of the k-lagged scatter plot and $\widetilde{SS}_{0(k)}$ is the sum of squares of the residuals from this smooth.

3.3.6 Confidence Intervals

In principle, the smoothing parameters associated with the smooth can be selected to obtain R-squared values arbitrarily close to one. For this reason it is important to obtain confidence intervals for, e.g., the lag dependence function under the hypothesis that the underlying process is i.i.d. and for a given set of smoothing parameters.

As indicated above it is clear that the range of the confidence intervals will depend on the flexibility of the smoother. To detect a general non-linearity a flexible smoother must be used whereby the range of the confidence interval will be increased compared to the case where we are only interested in detecting minor departures from linearity or departures in the direction of near-global higher order polynomials. Thus, the bandwidth of the smoother can be used to adjust the properties of the test. It is recommended to apply the methods using a range of bandwidths and smoothers. These aspects are exemplified in Section 3.3.7.

Under the hypothesis that the time series $\{x_1, \ldots, x_N\}$ are observations from an i.i.d. process the distribution of any of the quantities discussed in the previous sections can be approximated by generating a large number of i.i.d. time series of length N from an estimate of the distribution function of the process and recalculating the quantities for each of the generated time series. Such methods are often denoted bootstrap methods and in this context various approaches to the calculation of approximate confidence intervals have been addressed extensively in the literature, see e.g. [Efron & Tibshirani 1993]. An extensive description of the bootstrap methods for calculating the confidence limits for LDF and PLDF estimates are found in [Nielsen & Madsen 2001].

3.3.7 Examples

Non-linear processes

Three non-linear processes are addressed, namely (i) the non-linear autoregressive process ($NLAR(1)$)

$$X_t = \frac{1}{1 + \exp(-5X_{t-1} + 2.5)} + e_t, \tag{3.22}$$

where $\{e_t\}$ i.i.d. $N(0, 0.1^2)$, and (ii) the non-linear moving average process ($NLMA(1)$)

$$X_t = e_t + 2\cos(e_{t-1}), \tag{3.23}$$

where $\{e_t\}$ i.i.d. $N(0, 1)$ and (iii) the non-linear and deterministic process described in Section 3.3.1, called $DNLAR(1)$ in the following. For all three cases 1000 observations are generated. The starting value for $NLAR(1)$ is set to 0.5 and for $DNLAR(1)$ it is set to 0.8. Plots of the series $NLAR(1)$ and $NLMA(1)$ are shown in Figure 3.2. The plot of $DNLAR(1)$ is shown in Figure 3.1.

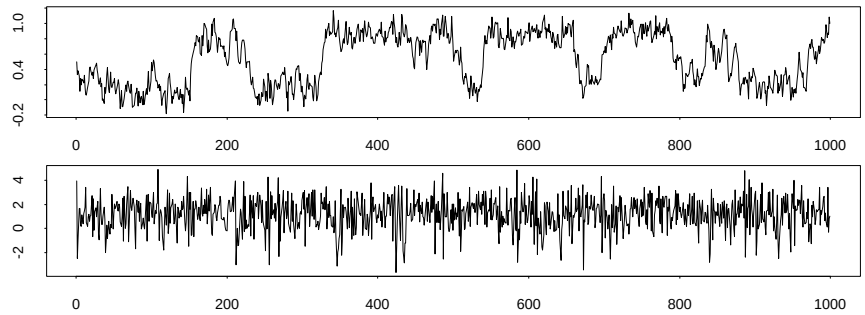

Figure 3.2: Plots of the series $NLAR(1)$ (top) and $NLMA(1)$ bottom.

For the calculation of LDF, $PLDF$, and $NLDF$ a local linear smoother with a nearest neighbour bandwidth of 0.5 is used. Actually, lagged scatter plots indicate that a local quadratic smoother should be applied, at least for $NLMA(1)$ and $DNLAR(1)$, but to avoid a perfect fit for the deterministic series a local linear smoother is used. Confidence intervals are constructed using standard normal intervals, since normal QQ-plots of the absolute values of the 200 bootstrap replicates showed this to be appropriate. The confidence interval obtained for LDF is included on the plots of $PLDF$.

Figure 3.3 shows $SACF$, $SPACF$, LDF, and $PLDF$ for the three series. For $NLMA(1)$ and $DNLAR(1)$ the linear tools, $SACF$ and $SPACF$, indicate independence and LDF shows that lag dependence is present. From these observations it can be concluded that $NLMA(1)$ and $DNLAR(1)$ are nonlinear processes. From the plots of LDF and $PLDF$ it cannot be inferred whether $NLMA(1)$ is of the autoregressive or of the moving average type. For $DNLAR(1)$ the autoregressive property is more

clear since $PLDF$ drops to exactly zero after lag two. In the case of $DNLAR(1)$ a more flexible smoother will result in values of LDF being significantly different from zero for lags larger than two, while, for $NLMA(1)$, LDF will be close to zero for lags larger than one independent of the flexibility of the smoother used. This is an indication of $DNLAR(1)$ being of the autoregressive type and $NLMA(1)$ being of the moving average type.

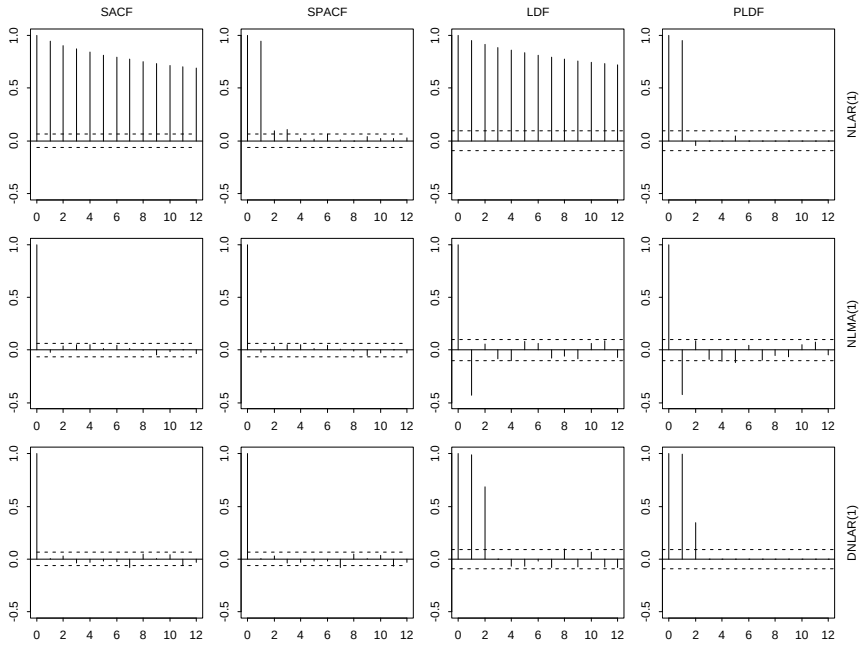

Figure 3.3: $SACF$, $SPACF$, LDF, and $PLDF$ (columns, left to right) for series $NLAR(1)$, $NLMA(1)$, and $DNLAR(1)$ (rows, top to bottom).

For $NLAR(1)$ the linear tools indicate that the observations come from an $AR(1)$ process. This is not seriously contradicted by LDF or $PLDF$, although LDF declines somewhat slower to zero than $SACF$. To investigate if the underlying process is linear, a Gaussian $AR(1)$ model is fitted to the data and this model is used as the hypothesis under which 200 (parametric) bootstrap replicates of $NLDF$ are generated. Normal QQ-plots show that the absolute values of the bootstrap replicates are approximately Gaussian. Figure 3.4 shows $NLDF$ and 95% standard normal intervals, constructed under the hypothesis mentioned above. From Figure 3.4 it is concluded that the underlying process is not the estimated $AR(1)$-model, and based on $PLDF$ it is thus concluded that the observations originate from a non-linear process of the $AR(1)$ type.

For $NLAR(1)$ Figure 3.5 shows the estimated relation between x_t and x_{t-1} using the same smoother as used above. This should only be regarded as a preliminary estimate. Note that although the investigation above indicates that $NLMA(1)$ is of the non-linear $MA(1)$ type the models fitted with the purpose of calculating LDF, $PLDF$, and $NLDF$ are not of this type. A different method is required for estimation.

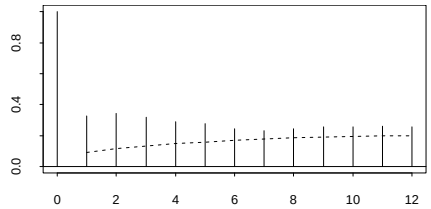

Figure 3.4: $NLDF$ for $NLAR(1)$, including a 95% confidence interval under the assumption of an $AR(1)$ process (dotted).

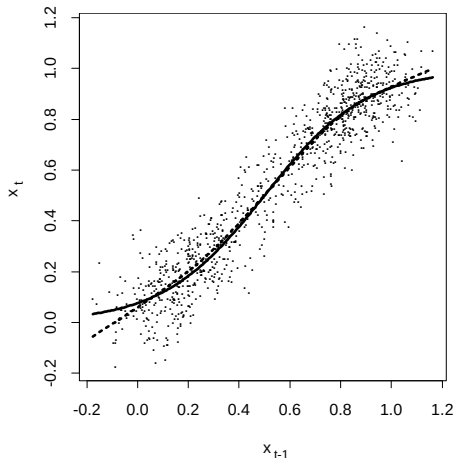

Figure 3.5: Scatter plot of x_t against x_{t-1} for $NLAR(1)$, together with the true (solid) and estimated (dotted) relation using a local linear smoother and a nearest neighbour bandwidth of 0.5.

Canadian lynx data

Lin & Pourahmadi (1998) analyzed the Canadian lynx data [Moran 1953] using non-parametric methods similar to the methods presented in this paper. The data is also described in [Tong 1990*a*, Section 7.2]. In this paper a thorough analysis of the data will not be presented, but the data will be used to illustrate how the methods suggested can be applied. As in [Lin & Pourahmadi 1998] the data is $\log_{10}$-transformed prior to the analysis.

For the transformed data LDF, $PLDF$, and $NLDF$ are computed using a local quadratic smoother and nearest neighbour bandwidths of 0.5 and 1. A local quadratic smoother is used since lagged scatter plots indicate that for some lags the underlying dependence may contain peaks. For LDF 200 bootstrap replicates are generated under the i.i.d. hypothesis and QQ-plots indicate that standard normal intervals are appropriate. The same apply for $NLDF$ with the exception that the bootstrap replicates are generated under the hypothesis that the $AR(2)$ model of Moran (1953), also described by Lin & Pourahmadi (1998), is true. Confidence intervals are computed also for $PLDF$ for the nearest neighbour bandwidth of 1.0. The intervals are based one hundred bootstrap replicates of $PLDF$ generated under the i.i.d. hypothesis.

In Figure 3.6 plots of LDF, $NLDF$, and $PLDF$ are shown. Dotted lines indicate 95% confidence intervals under the i.i.d. hypothesis (LDF) and under the $AR(2)$ model of Moran (1953) ($NLDF$). The intervals obtained for LDF are also shown on the plots of $PLDF$. Furthermore, for the nearest neighbour bandwidth of 1.0, a 95% confidence interval for white noise is included on the plot of $PLDF$ (solid lines).

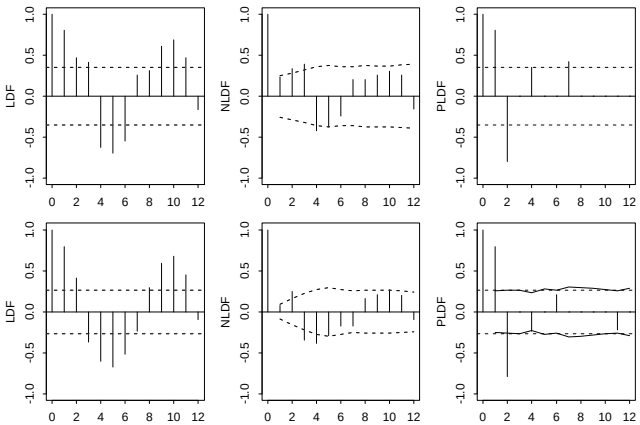

Figure 3.6: Canadian lynx data ($\log_{10}$-transformed). Plots of LDF, $NLDF$, and $PLDF$ using local quadratic smoothers and nearest neighbour bandwidths 0.5 (top row) and 1.0 (bottom row).

From the plots of LDF it is clearly revealed that the process is not i.i.d. The plots of $NLDF$ for a nearest neighbour bandwidth of 0.5 show hardly any significant values,

but when a nearest neighbour bandwidth of 1.0 is used lags two to four show weak significance. This indicates that a departure from linearity in the direction of an almost quadratic relationship is present in the data.

Finally, the plots of $PLDF$ clearly illustrate that lag one and two are the most important lags and that other lags are, practically, non-significant. In conclusion, an appropriate model seems to be a non-linear autoregressive model containing lag one and two, i.e. a model of the type (3.19) with $k = 2$.

Estimating this model using local quadratic smoothers and a nearest neighbour bandwidth of 1.0 yields the results shown in Figure 3.7. The response for lag one seems to be nearly linear. This aspect should be further investigated. The results agree well with the results of Lin & Pourahmadi (1998).

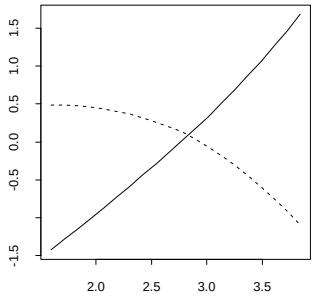

Figure 3.7: Non-linear additive autoregressive model for the $\log_{10}$-transformed Canadian lynx data ($\hat{f}_{21}(\cdot)$ solid, $\hat{f}_{22}(\cdot)$ dotted). The estimate of the constant term is 2.76 and the MSE of the residuals is 0.0414.

3.3.8 Lagged Cross Dependence

Given two time series $\{x_1, \ldots, x_N\}$ and $\{y_1, \ldots, y_N\}$ the Sample Cross Correlation Function between processes $\{X_t\}$ and $\{Y_t\}$ in lag k ($SCCF_{xy}(k)$) is an estimate of the correlation between X_{t-k} and Y_t. It is possible to generalize this in a way similar to the way LDF is constructed. Like $SCCF$ this generalization will be sensitive to autocorrelation, or lag dependence, in $\{X_t\}$ in general. For $SCCF$ this problem is (approximately) solved by prewhitening [Brockwell & Davis 1987, p. 402]. However, prewhitening is very dependent on the assumption of linearity, in that it relies on the impulse response function from the noise being independent of the level. For this reason, in the non-linear case, it is not possible to use prewhitening and the appropriateness of the generalization of $SCCF$ depends on $\{X_t\}$ being i.i.d.

3.3.9 Final Remarks

Most of the methodology presented in this paper is implemented in an S-PLUS library called LDF which can be downloaded from http://www.imm.dtu.dk/∼han/software.html.

Chapter 4
Cumulants and polyspectra

4.1 Introduction

Linearity and Gaussianity are highly related. This important fact is further discussed in section 4.2.

Some of the traditional tools for testing for non-linearity are based on tests in the frequency domain. For this reason the traditional power spectrum is generalized to higher order spectra in section 4.3. The higher order spectra is defined as a Fourier transformation of the auto covariance function to higher order moments using higher order cumulants for the process. Most frequently, however, only the bispectrum is considered. The bispectrum, including a method for estimating the bispectrum, is described in section 4.4. Section 4.5 contains non-parametric tests for non-linearity and non-Gaussianity based on the bispectrum. Finally, some parametric tests for determining the nonlinearity, in particular the Lagrange multiplier statistic are presented in section 4.6.

Aspects of model validation, which relates to the identification problem are described in Chapter 3.

4.2 Gaussianity and linearity

The following decomposition theorem is well known from linear time series analysis – see e.g. [Madsen 1989].

Theorem 4.1 (Wold's Theorem). *Let X_t be a zero mean second order stationary process. Then X_t can be expressed in the form*

$$X_t = U_t + V_t$$

where

- *$\{U_t\}$ and $\{V_t\}$ are uncorrelated processes.*
- *$\{U_t\}$ is non-deterministic with a one-sided representation $U_t = \sum_{i=0}^{\infty} a_i \eta_{t-i}$ where $a_0 = 1$ and $\sum_{i=1}^{\infty} a_i^2 < \infty$, $\{\eta_t\}$ is an uncorrelated sequence which is uncorrelated with $\{V_t\}$.*
- *The sequences $\{a_i\}$ and $\{\eta_t\}$ are uniquely determined.*
- *$\{V_t\}$ is deterministic.*

Proof. Omitted. ■

This theorem implies that *any Gaussian process conforms to a linear process*. Note that the converse is not necessarily true, i.e. not every linear process is Gaussian.

4.3 Polyspectra

4.3.1 Cumulants and polyspectra

Definition 4.2 (Joint Cumulants). The joint cumulant for the random variables $X_1, X_2, ..., X_k$ is defined by

$$C\{X_1, X_2, ..., X_k\} = \sum_{\nu} (-1)^{p-1}(p-1)! \mu_{\nu_1} \mu_{\nu_2} ... \mu_{\nu_p} \tag{4.1}$$

where μ_{ν_i} is the mean of the products of X's corresponding to the partition ν_i. The sum is over all partitions of the numbers $1, \cdots, k$ and p is the size of the partition. ▲

Example 4.1.

$$\begin{aligned} C\{X_1\} &= \mathrm{E}[X_1] && (4.2)\\ C\{X_1, X_2\} &= \mathrm{E}[X_1X_2] - \mathrm{E}[X_1]\,\mathrm{E}[X_2] && (4.3)\\ C\{X_1, X_2, X_3\} &= \mathrm{E}[X_1X_2X_3] - \mathrm{E}[X_1X_2]\,\mathrm{E}[X_3] - \mathrm{E}[X_1X_3]\,\mathrm{E}[X_2] \\ &\quad \mathrm{E}[X_1]\,\mathrm{E}[X_2X_3] + 2\,\mathrm{E}[X_1]\,\mathrm{E}[X_2]\,\mathrm{E}[X_3] && (4.4) \end{aligned}$$

♦

Remark. The cumulants are equal to the k'th order central moment for $k = 1, 2, 3$, but not for higher values of k. This means that $C\{X_1, X_2\} = \mathrm{Cov}\{X_1, X_2\}$, and the third-order cumulant is

$$C\{X_1, X_2, X_3\} = E[(X_1 - E[X_1])(X_2 - E[X_2])(X_3 - E[X_3])] \tag{4.5}$$

▼

Definition 4.3 (Moment Generating Function). The moment generating function for the random variables $X_1, X_2, ..., X_k$ is defined by

$$M(\theta_1, ..., \theta_k) = E\{\exp(\theta_1 X_1 + ... + \theta_k X_k)\} \tag{4.6}$$

▲

Definition 4.4 (Cumulant Generating Function). The cumulant generating function $K(\theta_1, ..., \theta_k)$ is defined by

$$K(\theta_1, ..., \theta_k) = \log M(\theta_1, ..., \theta_k) \tag{4.7}$$

▲

Some tedious (but straight forward) algebra shows that

$$\frac{\partial}{\partial\theta_1}...\frac{\partial}{\partial\theta_k}M(\theta_1, ..., \theta_k)|_{\theta_i=0, i=1,...,k} = E\{X_1...X_k\} \tag{4.8}$$

and

$$\frac{\partial}{\partial\theta_1}...\frac{\partial}{\partial\theta_k}K(\theta_1, ..., \theta_k)|_{\theta_i=0, i=1,...,k} = C\{X_1...X_k\}. \tag{4.9}$$

Definition 4.5 (Polyspectrum of a Stationary Process). Consider a stationary stochastic process $\{X_t\}$ and define

$$C(\tau_1, \tau_2, ..., \tau_{k-1}) = C\{X_t, X_{t+\tau_1}, X_{t+\tau_2}, ..., X_{t+\tau_{k-1}}\} \tag{4.10}$$

Now the k-th order polyspectrum for the process $\{X_t\}$ denoted by $f_X(\omega_1, \omega_2, \ldots, \omega_{k-1})$ is defined by the Fourier transform of $C\{X_t, X_{t+\tau_1}, X_{t+\tau_2}, ..., X_{t+\tau_{k-1}}\}$, i.e.

$$f_X(\omega_1, \omega_2, ..., \omega_{k-1}) = (\frac{1}{2\pi})^{k-1} \sum_{\tau_1=-\infty}^{\infty} ... \sum_{\tau_{k-1}=-\infty}^{\infty} C(\tau_1, \tau_2, ..., \tau_{k-1}) \exp\{-i(\omega_1\tau_1 + \omega_2\tau_2 + ... + \omega_{k-1}\tau_{k-1})\}. \quad (4.11)$$

Note that the second order polyspectrum equals the standard spectrum of a stationary process. ▲

Definition 4.6 (Time reversibility). A stationary process is said to be *time reversible* if for any integer n, and every $t_1, \cdots, t_n$ the joint distribution of $(X_{-t_1}, \cdots, X_{-t_n})$ is the same as that of $(X_{t_1}, \cdots, X_{t_n})$.

A stationary time series which is not time reversible is said to be *time irreversible*. ▲

For a *time reversible* process it follows that

$$C(\tau_1, \cdots, \tau_{k-1}) = C(-\tau_1, \cdots, -\tau_{k-1}), \quad (4.12)$$

which implies that $\mathrm{Im}\{f_X(\omega_1, \omega_2, ..., \omega_{k-1}) = 0\}$.

4.3.2 Polyspectra, non-Gaussianity and non-linearity

Theorem 4.7. *Consider a stationary Gaussian stochastic process $\{X_t\}$. Then all of its polyspectra of order higher than 2 vanish.*

Proof. Assume that the random variables $X_1, X_2, ..., X_k$, $k \geq 2$, have a Gaussian joint distribution with mean $\boldsymbol{m}$ and covariance $\boldsymbol{\Sigma}$. The characteristic function of the random variables is

$$\exp\{i\boldsymbol{m}^T\boldsymbol{t} - \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\} \quad (4.13)$$

Thus the moment generating function is given by

$$\exp\{\boldsymbol{m}^T\boldsymbol{\theta} + \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\Sigma}\boldsymbol{\theta}\} \quad (4.14)$$

This shows that the cumulant generating function is quadratic in $\boldsymbol{\theta}$ and its derivatives with order higher than two vanish. This completes the proof. ■

4.4 Bispectrum

Definition 4.8 (Bispectrum). The third order polyspectrum $f(\omega_1, \omega_2)$ of a stationary stochastic process is called bispectrum. ▲

By theorem 4.7 the bispectrum of a stationary Gaussian process is identically zero. This provides *a basis for a test for Gaussianity*.

4.4.1 Bispectra and Linearity

Suppose that the two stationary processes $\{X_t\}$ and $\{Y_t\}$ are *linearly* related through

$$X_t = \sum_{i=0}^{\infty} h_i Y_{t-i}. \tag{4.15}$$

Then

$$f_X(\omega_1, \omega_2) = H(-\omega_1 - \omega_2)H(\omega_1)H(\omega_2)f_Y(\omega_1, \omega_2), \tag{4.16}$$

where $H(\omega)$ is the frequency response function of the filter with impulse response $\{h_i\}$. The result follows directly from the definition of bispectrum.

If $\{Y_t\}$ is a white noise process then equation (4.15) defines a linear stationary process. Assume that a stationary process $\{X_t\}$ has the linear representation

$$X_t = \sum_{i=-\infty}^{\infty} a_i \epsilon_{t-i} \tag{4.17}$$

where $\{\epsilon_t\}$ is a sequence of independent, zero mean, identically distributed random variables (white noise). Denoting the third moment of ϵ_t by μ_3, we have

$$(\frac{\mu_3}{2\pi})^2 = \frac{f_X(\omega_1, \omega_2)}{H(\omega_1)H(\omega_2)H(-\omega_1 - \omega_2)} \tag{4.18}$$

and denoting the second moment of ϵ_t by μ_2,

$$f_X(\omega) = \frac{\mu_2}{2\pi}|H(\omega)|^2$$

Hence, under the linearity hypothesis, the quantity

$$\frac{|f_X(\omega_1, \omega_2)|^2}{|f_X(-\omega_1 - \omega_2)f_X(\omega_1)f_X(\omega_2)|} \tag{4.19}$$

is a constant for all ω_1 and ω_2. This provides *a basis for a test for linearity*. Furthermore, if the process is Gaussian, this constant is identically zero for all frequency pairs. Therefore, the estimation of the bispectrum from observed data is a fundamental step in one possible test for linearity and Gaussianity.

Example 4.2 (The bispectrum for an ARMA(1,1) process). This example illustrates how to calculate the bispectrum for an ARMA(1,1) process with non-Gaussian noise input.

Consider the ARMA(1,1) process

$$X_t + aX_{t-1} = \epsilon_t + b\epsilon_{t-1} \tag{4.20}$$

where $\{\epsilon_t\}$ is a sequence of identically distributed independent $\Gamma(k,\beta)$-distributed random variables.

The frequency response function is easily found as (see [Madsen 1995*a*])

$$H(\omega) = \frac{1 + be^{-i\omega}}{1 + ae^{-i\omega}} \tag{4.21}$$

When we include the above in equation (4.16) we get

$$f_X(\omega_1,\omega_2) = f_\epsilon(\omega_1,\omega_2)\frac{P(b,\omega_1,\omega_2)}{P(a,\omega_1,\omega_2)} \tag{4.22}$$

where the real part is given by

$$\begin{aligned}\mathrm{Re}[P(x,\omega_1,\omega_2)] &= 1 + x^3 + (x^2+x)\cos(\omega_1) + \\ &\quad (x^2+x)\cos(\omega_2) + (x^2+x)\cos(\omega_1+\omega_2)\end{aligned}$$

and the imaginary part by

$$\begin{aligned}\mathrm{Im}[P(x,\omega_1,\omega_2)] &= (x^2-x)\sin(\omega_1) + (x^2-x)\sin(\omega_2) + \\ &\quad (x-x^2)\sin(\omega_1+\omega_2).\end{aligned}$$

As $\{\epsilon_t\}$ is a sequence of independent random variables it follows directly from the definition of the bispectrum that

$$f_\epsilon(\omega_1,\omega_2) = \frac{\mu_3}{(2\pi)^2}, \tag{4.23}$$

where $\mu_3 = m(0,0)$. Since ϵ_t is Gamma distributed

$$f_t(\omega_1,\omega_2) = \frac{2k\beta^3}{(2\pi)^2} \tag{4.24}$$

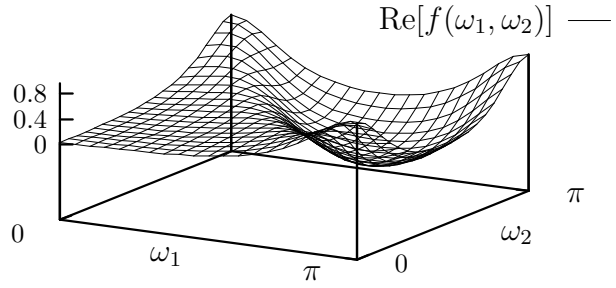

Figure 4.1: Real part of the theoretical bispectrum for an ARMA(1,1) process with Gamma distributed noise

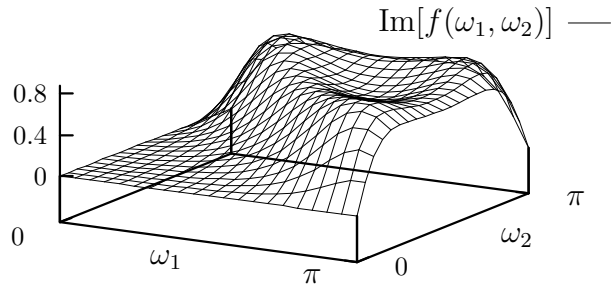

Figure 4.2: Imaginary part of the theoretical bispectrum for an ARMA(1,1) process with Gamma distributed noise

since $\mu_3 = 2k\beta^3$ for $X \sim \Gamma(k, \beta)$.

The theoretical bispectrum for the ARMA(1,1) process given by $a = 0.5$ and $b = -0.3$ is shown in Figure 4.2.

♦

4.4.2 Estimation of bispectra

Let $X_1, X_2, ..., X_N$ be realizations from a stationary (at least up to third order) process $\{X_t\}$ with third central moment $C(\tau_1, \tau_2)$. The natural estimate of $C(\tau_1, \tau_2)$ will be

$$\begin{aligned} \hat{C}(\tau_1, \tau_2) &= \frac{1}{N} \sum_{t=1}^{N-\gamma} (X_t - \bar{X})(X_{t+\tau_1} - \bar{X})(X_{t+\tau_2} - \bar{X}), \qquad (4.25) \\ \tau_1, \tau_2 &> 0, \\ \gamma &= \max(0, \tau_1, \tau_2). \end{aligned}$$

Then from the definition of bispectrum, we introduce the estimate $I(\omega_1, \omega_2)$:

$$\begin{aligned} I(\omega_1, \omega_2) &= \frac{1}{N(2\pi)^2} \{\sum_1^N (X_t - \bar{X}) \exp(-i\omega_1 t)\} \qquad (4.26) \\ &\quad \times \{\sum_1^N (X_t - \bar{X}) \exp(-i\omega_2 t)\} \qquad (4.27) \\ &\quad \times \{\sum_1^N (X_t - \bar{X}) \exp(i(\omega_1 + \omega_2) t)\} \qquad (4.28) \end{aligned}$$

This estimate is often called the *third order periodogram*. Exactly as in the periodogram case, this estimate is asymptotically unbiased but not consistent. Thus, it must be smoothed by choosing some suitable window.

4.4.3 Consistent estimate of bispectra

In order to obtain a consistent estimate of the bispectrum, the bispectrum is smoothed by using a kernel in the frequency domain. This is a direct generalization of periodogram smoothing for estimation of spectra and is closely related to the non-parametric methods discussed in Chapter 2. Recall that the smoothed periodogram estimate at a frequency is obtained by a weighted average of periodogram estimates at neighbor frequencies or equivalently by

$$\hat{f}_M(\omega) = \int_{-\pi}^{\pi} I(\theta) K_M(\theta - \omega) d\theta$$

where $K_M(\cdot)$ is a kernel with bandwidth determined by M, $\hat{f}_M(\omega)$ is the estimate of the spectrum $f(\cdot)$, and $I(\cdot)$ is the raw periodogram estimate.

We consider unit bandwidth kernels in two parameters $K_0(\theta_1, \theta_2)$ that satisfy:

$$K_0(\theta_1, \theta_2) = K_0(\theta_2, \theta_1) = K_0(\theta_1, -\theta_1 - \theta_2) = K_0(-\theta_1 - \theta_2, \theta_2) \quad (4.29)$$

$$\int_{\theta_2}\int_{\theta_1} K_0(\theta_1, \theta_2)d\theta_1 d\theta_2 = 1, \quad (4.30)$$

$$\int_{\theta_2}\int_{\theta_1} K_0^2(\theta_1, \theta_2)d\theta_1 d\theta_2 < \infty, \quad (4.31)$$

$$\int_{\theta_2}\int_{\theta_1} \theta_i^2 K_0(\theta_1, \theta_2)d\theta_1 d\theta_2 < \infty, i = 1, 2. \quad (4.32)$$

Exercise 4.1. Using the symmetry condition (4.29), show that

$$\frac{1}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \theta_i^2 K_0(\theta_1, \theta_2)d\theta_1 d\theta_2 = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \theta_1\theta_2 K_0(\theta_1, \theta_2)d\theta_1 d\theta_2, i = 1, 2.$$

Now denote the inverse Fourier transform of $K_0(\theta_1, \theta_2)$ by $\lambda_0(\tau_1, \tau_2)$, i.e.

$$\lambda_0(\tau_1, \tau_2) = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \exp(i\tau_1\theta_1 + i\tau_2\theta_2)K_0(\theta_1, \theta_2)d\theta_1 d\theta_2.$$

The function $\lambda_0(\tau_1, \tau_2)$ is a window in time domain (note that convolution in frequency domain corresponds to windowing in time domain and vice versa). We further assume that the first partial derivatives of $\lambda_0(\tau_1, \tau_2)$ with respect to τ_1 and τ_2 at $(0, 0)$ vanish but the second derivatives are different from zero. Such window functions are said to belong to the $C_K^{(2)}$ class. A kernel with bandwidth M, $K_M(\theta_1, \theta_2)$, is defined by

$$K_M(\theta_1, \theta_2) = M^2 K_0(M\theta_1, M\theta_2). \quad (4.33)$$

It can be shown that if M is chosen such that $M \to \infty$, where N is the number of samples, but $\frac{M^2}{N} \to 0$, then the smoothed estimate of the bispectrum given by

$$\hat{f}_M(\omega_1, \omega_2) = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} K_M(\theta_1 - \omega_1, \theta_2 - \omega_2)I(\theta_1, \theta_2)d\theta_1 d\theta_2 \quad (4.34)$$

is asymptotically unbiased and consistent. The following theorem may be used to assess the mean square error of the smoothed bispectrum estimate.

Theorem 4.9. *Let the second order partial derivatives of $f(\omega_1, \omega_2)$ exist and be finite, and let $C(\tau_1, \tau_2)$ satisfy the condition*

$$\sum_{\tau_1,\tau_2=-\infty}^{\infty} |\tau_i C(\tau_1, \tau_2)| < \infty, i = 1, 2.$$

Further, let $K_0(\cdot, \cdot) \in C_K^{(2)}$. Then the asymptotic bias of $\hat{f}(\omega_1, \omega_2)$ denoted by $b(\omega_1, \omega_2)$ is given by

$$b(\omega_1, \omega_2) = \frac{B_K}{M^2} D^{(2)}(\omega_1, \omega_2)$$

and the asymptotic variance term var$(\hat{f}(\omega_1,\omega_2))$ *by*

$$\frac{M^2}{N}\frac{V_2}{2\pi}f(\omega_1)f(\omega_2)f(\omega_1+\omega_2)$$

where

$$B_k = -\int_{\theta_2}\int_{\theta_1}\theta_1\theta_2 K_0(\theta_1,\theta_2)d\theta_1 d\theta_2, \tag{4.35}$$

$$V_2 = (2\pi)^2\int_{\theta_2}\int_{\theta_1}K_0^2(\theta_1,\theta_2)d\theta_1 d\theta_2, \tag{4.36}$$

$$D^{(2)}(\omega_1,\omega_2) = (\frac{\partial^2}{\partial\omega_1^2}+\frac{\partial^2}{\partial\omega_2^2}-\frac{\partial^2}{\partial\omega_1\omega_2})f(\omega_1,\omega_2). \tag{4.37}$$

Exercise 4.2. Let

$$M^*(\omega_1,\omega_2) = \arg\min_M MSE(\omega_1,\omega_2)$$

where the mean square error is given by

$$MSE(\omega_1,\omega_2) = |b(\omega_1,\omega_2)|^2 + \text{var}(\hat{f}(\omega_1,\omega_2)).$$

Show that

$$M^*(\omega_1,\omega_2) = M_R G(\omega_1,\omega_2)$$

where

$$M_R = (\frac{B_K}{\sqrt{V_2}})^{\frac{1}{3}},$$

$$G(\omega_1,\omega_2) = \left[2D^{(2)}(\omega_1,\omega_2)\sqrt{\frac{\pi N}{f(\omega_1)f(\omega_2)f(\omega_1+\omega_2)}}\right]^{\frac{1}{3}}.$$

Notice that G does not depend upon the choice of kernel. Further, show that the optimal value of MSE (with respect to M) is given by

$$2[\frac{1}{2\pi N}f(\omega_1)f(\omega_2)f(\omega_1+\omega_2)D^{(2)}(\omega_1,\omega_2)V_2B_k]^{-2/3},$$

which indicates that the term V_2B_K is instrumental in evaluating the appropriateness of the type of kernel to be employed among the class of $C_K^{(2)}$ kernels.

Exercise 4.3. Consider the one dimensional lag window $\lambda_0(\tau)$ satisfying

$$\lambda_0(0)=1, \lambda_0'(0)=0, \lambda_0''(0)\neq 0, \tag{4.38}$$

and the (unit bandwidth) kernel defined by

$$K_0(\theta_1,\theta_2) = \frac{1}{(2\pi)^2}\sum_{\tau_1=-\infty}^{\infty}\sum_{\tau_2=-\infty}^{\infty}\lambda_0(\tau_1,\tau_2)\exp(-i\tau_1\theta_1 - i\tau_2\theta_2),$$

where $\lambda_0(\tau_1, \tau_2) = \lambda_0(\tau_1)\lambda_0(\tau_2)\lambda_0(\tau_1 - \tau_2)$. Verify that condition (4.29) holds for this window. Moreover, verify that the first partial derivatives of $\lambda_0(\tau_1, \tau_2)$ with respect to τ_1 and τ_2 at $(0, 0)$ vanish but the second derivatives are different from zero. Repeat the above for the construction $\lambda_0(\tau_1)\lambda_0(\tau_2)$. Does the window defined by $\lambda_0(\tau_1)\lambda_0(\tau_2)$ belong to the $C_K^{(2)}$ class? Finally show that (4.38) is satisfied by (unit bandwidth) Daniell, Tukey-Hamming, Parzen, and Bartlett-Priestley windows (see e.g. [Madsen 1995*a*], Table 6.2).

Summary of Window and Bandwidth Selection:

- See [Madsen 1995*a*], pp. 178, Table 6.2. Choose a window and set $M = 1$ to obtain $\lambda_0(\cdot)$.

- Compute the three dimensional window $\lambda_0(\tau_1, \tau_2) = \lambda_0(\tau_1)\lambda_0(\tau_2)\lambda_0(\tau_1 - \tau_2)$. Choose M, noting that
  - If it is possible, M should be less than the square root of the sample size, N.
  - It should be smaller than the value of M used in estimating the spectral density.
  - Use (4.2) to calculate the optimal bandwidth. Once for some selection of kernel X, the optimal bandwidth M_X is found the M value for others, say, Parzen window M_{Par} is calculated according to

$$M_{Par} = \frac{M_{R,Parzen}}{M_{R,X}} M_X \tag{4.39}$$

    where $M_{R,\cdot}$ denotes the value of M_R for the corresponding window selection.

Example 4.3 (Estimation of bispectra). Consider a time series of length $N = 500$ which is generated by the ARMA(1,1) process with Gamma distributed noise considered in Example 4.2.

Figure 4.4 and 4.6 show the estimated bispectrum with $M = 10$ and $M = 18$, respectively. In both cases a Parzen window is used.

As expected a more smooth variation of the estimated bispectra is seen for $M = 10$ compared to $M = 18$.

A comparison with the theoretical bispectrum in Figure 4.2 indicates a reasonable agreement for the real part; but for the imaginary part the estimated bispectra seems

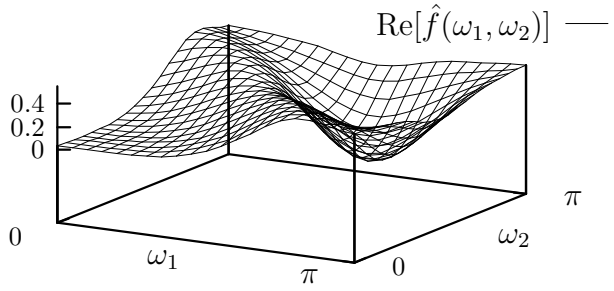

Figure 4.3: Real part of estimated bispectrum, Parzen window, $N = 500$, $M = 10$

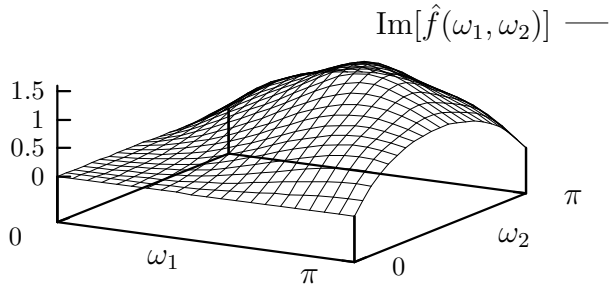

Figure 4.4: Imaginary part of estimated bispectrum, Parzen window, $N = 500$, $M = 10$

having trouble with detecting the local minima observed in the imaginary part of the theoretical bispectrum.

♦

This concludes our discussion of *non-parametric methods* for estimating the bispectrum of a process. If the data generating mechanism is known *a priori*, e.g. a bilinear model, it is also possible to explicitly compute the bispectrum for that model and evaluate the result for some estimated parameter values. This is a *parametric method*.

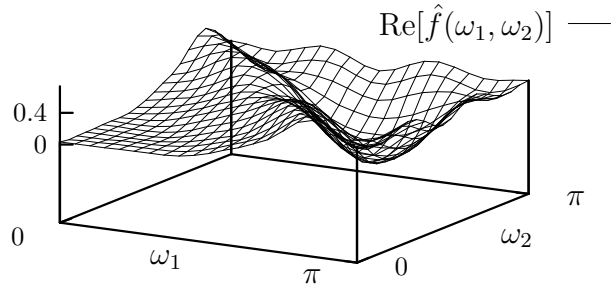

Figure 4.5: Real part of estimated bispectrum, Parzen window, $N = 500$, $M = 18$

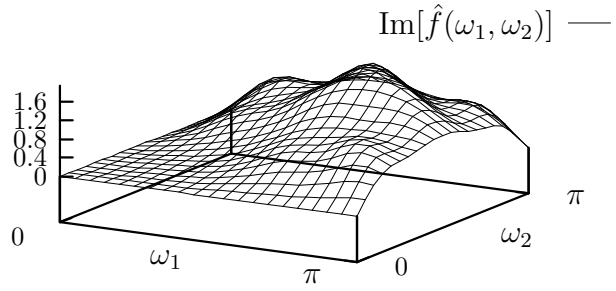

Figure 4.6: Imaginary part of estimated bispectrum, Parzen window, $N = 500$, $M = 18$

4.5 Tests using the bispectrum for Gaussianity and linearity

Statistical Test for Gaussianity (Symmetry)

Here we have to test the null hypothesis H_0: $f(\omega_i, \omega_j) = 0$ for all ω_i, ω_j. However, as a result of the following symmetry relations:

$$\begin{aligned} f(\omega_1,\omega_2) &= f(\omega_2,\omega_1) && (4.40)\\ &= f(\omega_1,-\omega_1-\omega_2) && (4.41)\\ &= f(-\omega_1-\omega_2,\omega_2) && (4.42)\\ &= f^*(\omega_2,\omega_1) && (4.43) \end{aligned}$$

it will be enough to evaluate the bispectrum at the points confined to the area with boundaries:

$$\begin{cases} \omega_2 - \omega_1 = 0 \\ \omega_2 + \frac{1}{2}\omega_1 = \pi \\ \omega_1 = 0 \end{cases}$$

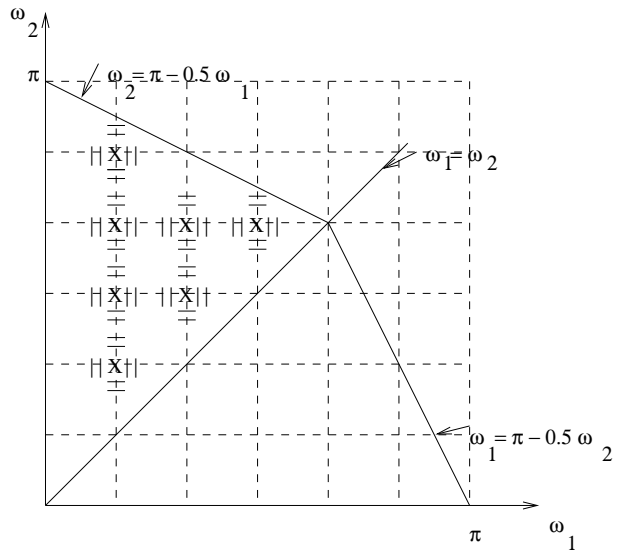

Figure 4.7: Choice of grid for P=7, n=9. The coarse grid points are marked by X and the fine grid points are marked by vertical or horizontal line pieces.

Our aim is to estimate a (complex) $P \times 1$ vector $\boldsymbol{\eta}$ where the i'th element is the bispectrum of the i'th node. In order to estimate $\boldsymbol{\eta}$, we estimate the bispectrum for the points corresponding to fine grid (see Figure 4.7). These estimates are then gathered in n vectors (each vector having P elements), each denoted by $\boldsymbol{\xi}_{(i)}$, $i = 1, ..., n$. Now estimate:

$$\hat{\boldsymbol{\eta}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\xi}_{(i)} \tag{4.44}$$

$$\boldsymbol{A} = \sum_{i=1}^{n} [\xi_{(i)} - \hat{\eta}][\xi_{(i)} - \hat{\eta}]^T \tag{4.45}$$

$$\hat{\boldsymbol{\Sigma}}_{\xi} = \frac{\boldsymbol{A}}{n} \tag{4.46}$$

Under the null hypothesis that $f(\omega_i, \omega_j) = 0$ for all ω_i, ω_j, the statistic F_1 defined by

$$F_1 = \frac{2(n-p)}{2p} T^2 \tag{4.47}$$

$$T^2 = n\hat{\boldsymbol{\eta}}^T \boldsymbol{A}^{-1} \hat{\boldsymbol{\eta}} \tag{4.48}$$

is distributed as a central F distribution with $(2P, 2(n-p))$ degrees of freedom. The latter is the complex generalization of Hotelling's T^2 test (see [Giri 1965, Khatri 1965]).

Test for Linearity

This test is very similar to the previous one, with the following modifications. The vector $\boldsymbol{\xi}_{(i)}$ is modified to $\boldsymbol{Y}_{(i)}$. The elements of the vector are now

$$\frac{|f(\omega_i, \omega_j)|^2}{f(\omega_i) f(\omega_j) f(\omega_i + \omega_j)} \tag{4.49}$$

instead of $f(\omega_i, \omega_j)$. Denote the counter-part of the vector $\boldsymbol{\eta}$ by $\boldsymbol{Z}$. Now estimate:

$$\hat{\boldsymbol{Z}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Y}_{(i)} \tag{4.50}$$

$$\boldsymbol{S} = \sum_{i=1}^{n} (\boldsymbol{Y}_{(i)} - \hat{\boldsymbol{Z}})(Y_{(i)} - \hat{\boldsymbol{Z}})^T \tag{4.51}$$

$$\hat{\boldsymbol{\Sigma}}_Y = \frac{\boldsymbol{S}}{n} \tag{4.52}$$

Define the statistic F_2 by

$$F_2 = \frac{n-Q}{Q} T^2 \tag{4.53}$$

$$T^2 = n \boldsymbol{\beta}^T \boldsymbol{S}_B^{-1} \boldsymbol{\beta} \tag{4.54}$$

$$\boldsymbol{\beta} = \boldsymbol{B} \hat{\boldsymbol{Z}} \tag{4.55}$$

$$\boldsymbol{S}_B = \boldsymbol{B} \hat{\boldsymbol{\Sigma}}_Y \boldsymbol{B}^T \tag{4.56}$$

where $Q = P - 1$ and $\boldsymbol{B}$ is a $Q \times P$ matrix defined by

$$\begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ & & \vdots & & \\ 0 & 0 & \dots & 1 & -1 \end{bmatrix}$$

Under the null hypothesis, the statistic F_2 is F distributed with $(Q, n-Q)$ degrees of freedom.

For a more comprehensive study of non-parametric tests in the frequency domain see [Subba Rao & Gabr 1984*b*].

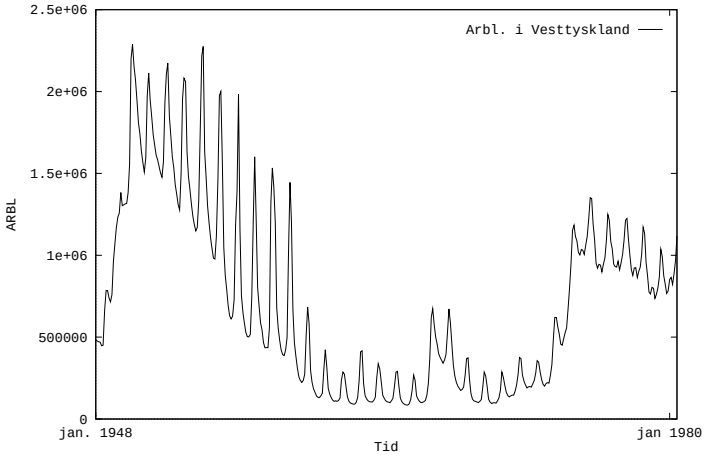

Figure 4.8: Monthly observations of unemployment in Germany.

There also exist other non-parametric tests. For example, a test procedure is suggested by Hinnich which is basically the same as Rao-Gabr's test with slight modifications. The finite Fourier transform of the sample is used to construct a consistent estimator of the bispectrum, instead of Rao-Gabr's standard windowed sample covariance method. For details see [Hinnich 1982].

Some comments on the use of bispectra

Mellentin (1992) has experimented with the estimation of bispectra and non-parametric tests for Gaussianity and non-linearity, see [Mellentin 1992]. He points out some draw backs of Rao-Gabr's test and their results gathered in [Subba Rao & Gabr 1984*b*]. He suggests among others that the test quantity in some cases is very sensitive to the test parameters, e.g. the distance between the points of the fine grid. Furthermore, he criticizes Rao-Gabr's choice of Daniell window in [Subba Rao & Gabr 1984*b*], since it in some cases results in a negative value of the spectrum in some frequency ranges. This obviously questions the credibility of the test results. He also suggests that one should not use the same window and truncating point for estimation of spectra and bispectra. Nevertheless, he believes that the tests in bispectra have the advantage that they are non-parametric and are not restricted to an *a priori* structure.

His experimentation with Hinnich's proposed test, as described in [Hinnich 1982] shows that the test results are superior to Rao-Gabr's procedure. However, Hinnich's test is computationally more demanding.

Example 4.4 (Tests for Gaussianity and linearity in German unemployment data). Mellentin (1992) has analyzed the German unemployment data series $\{Z_t\}$ in the period January 1948 - May 1980. The time series is shown in Figure 4.8.

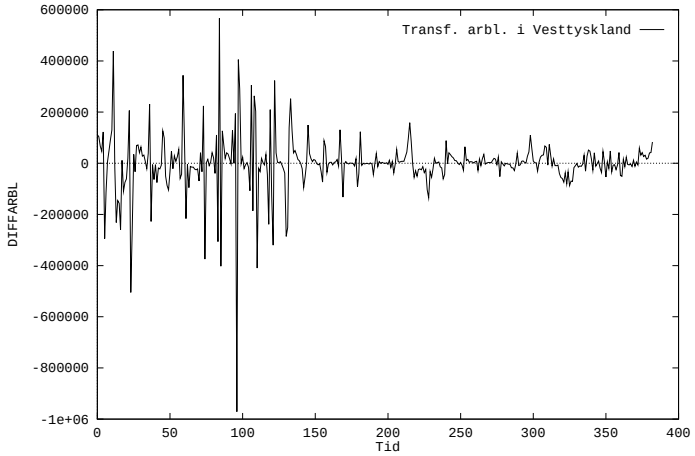

Figure 4.9: The time series DIFFARBL.

Test for ...	Test in [Subba Rao & Gabr 1984*b*]	Test in [Mellentin 1992]	95% quant. in F-dist. F(14,4)/F(6,3)
Gaussianity	349.9	399.1	5.89
Linearity	221.53	1084.7	8.94

Table 4.1: Tests in the bispectra for Gaussianity and linearity

In order to remove variations in the data, he performs the following transformation:

$$Y_t = (1 - B)(1 - B^{12})Z_t$$

The series $\{Y_t\}$ is denoted by DIFFARBL and shown in Figure 4.9.

Estimated bispectra for DIFFARBL are shown in Figures 4.4 and 4.4. In the latter figure normalized values of the bispectra are considered.

It is clearly seen that the bispectrum is non-zero. This means that the process in non-Gaussian. Considering the large number of extreme values in DIFFARBL this seems very reasonable.

The statistical non-parametric tests based on estimation of bispectra rejects the Gaussianity and linearity hypothesis.

Tests for Gaussianity and linearity are shown in Table 4.1 – the results are further discussed in [Mellentin 1992]. The tests are based on a Daniell window with $M = 20$ and $d = 12$ (Furthermore $k = 6$ and $r = 2$).

The statistical non-parametric tests based on estimation of bispectra rejects both the Gaussianity and the linearity hypothesis.

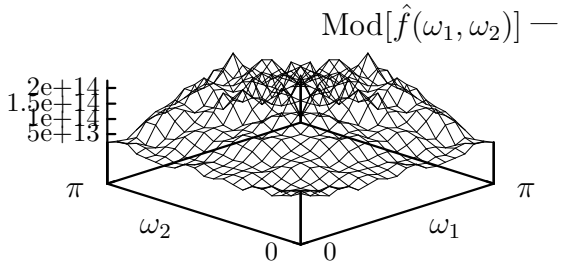

Figure 4.10: Modulus of bispectrum for DIFFARBL, Daniell, $M = 16$

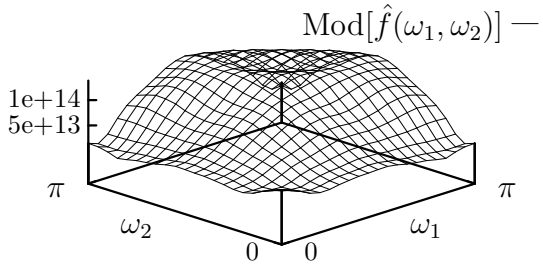

Figure 4.11: Modulus of bispectrum for DIFFARBL, Daniell, $M = 10$

Is is seen from Table 4.1 that [Subba Rao & Gabr 1984*b*] and [Mellentin 1992] obtains different results using the same data. However, in both cases the Gaussianity and the linearity is rejected. The differences illustrate the fact that tests using the non-parametric bispectra is sometimes rather doubtful. The problems is further discussed in [Mellentin 1992]. ♦

4.6 Parametric tests

Some parametric tests are based on the Lagrange Multiplier (LM) statistic – see section 4.7.2 for an introduction.

Problem Formulation: Suppose that the model can be written as

$$X_t = \beta_1^T X_{t-1} + f(\beta_2, X_{t-1}) + \epsilon_t \tag{4.57}$$

Partition the parameter vector $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = [\beta_1, \beta_2]$ where β_1 and β_2 parameterize the linear and non-linear parts, respectively. It is desired to test:

$$H_0 : \beta_2 = 0$$

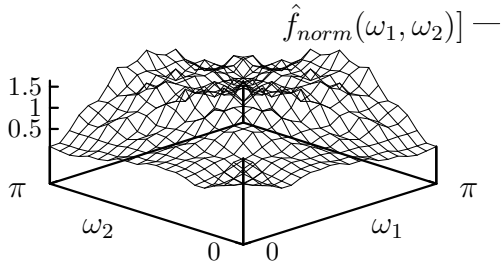

Figure 4.12: Norm. bispectrum for DIFFARBL, Parzen, $M = 15$

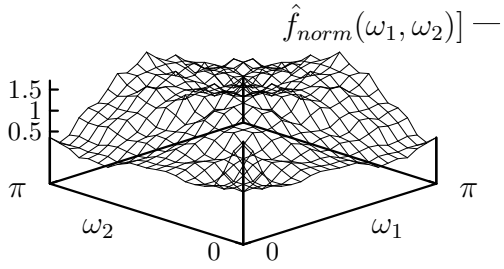

Figure 4.13: Norm. bispectrum for DIFFARBL, Parzen, $M = 20$

against the alternative:

$$H_1 : \beta_2 \neq 0$$

In general the procedure is:

1. Minimize
$$S(\beta) = \sum_{t=1}^{T} \epsilon_t^2(\boldsymbol{\beta}) \tag{4.58}$$
under the restriction $\beta_2 = 0$ and estimate β_1 denoted by $\hat{\beta}_1$. Denote the residuals under the null hypothesis by $\tilde{\epsilon}_t$.

2. Compute the efficient score
$$z_t(\beta) = \frac{\partial \epsilon_t}{\partial \beta} \tag{4.59}$$
for the hypothetical non-linear model. Evaluate $z_t(\boldsymbol{\beta})$ at $(\beta_1, \beta_2) = (\hat{\beta}_1, 0)$ and denote it by
$$\tilde{z}_t = [\tilde{z}'_{1t}, \tilde{z}'_{2t}]$$
where partitioning is performed in the obvious way.

3. Compute the statistic

$$LM = \tilde{\sigma}^{-2}(\sum \tilde{z}_{2t}\tilde{\epsilon}_t)'\{\sum \tilde{z}_{2t}\tilde{z}'_{2t} - \sum \tilde{z}_{2t}\tilde{z}'_{1t}(\sum \tilde{z}_{1t}\tilde{z}'_{1t})^{-1}\sum \tilde{z}_{1t}\tilde{z}'_{2t}\}^{-1}(\sum \tilde{z}_{2t}\tilde{\epsilon}_t) \tag{4.60}$$

where $\tilde{\sigma}^2$ is a maximum likelihood estimate for the variance of the residuals under H_0. In a standard situation, the statistic LM is asymptotically χ^2 distributed with degrees of freedom equal to the dimension of β_2.

For a further discussion about the LM test, see [Harvey 1989] and [Saikkonen & Luukkonen 1988].

4.7 Model validation

4.7.1 Introduction

There is no universally accepted procedure for modeling a system. In time series analysis, however, there are various "tools" to verify the validity of an assumed model or model structure. One of the most primary test procedures is test for linearity and Gaussianity that was thoroughly investigated in the section on identification of non-linear time-series. In this section, we are mostly concerned with the other end of the problem: Given a model and a set of observations, how to verify that the model is a "good" fit to the true system. The appropriateness of the fit is obviously a goal oriented quantity (if such quantity is defined at all).

It is common practice within the time-series community to define the "distance" between the model and the true system by some quadratic expression of the prediction errors (or residuals). Thus the appropriateness of a model is normally defined as its ability to provide a better prediction of the future behavior. In the following, we mention some widely accepted (or at least widely known) criteria. The interested reader may consult [Tong 1990*b*] and [Madsen 1995*a*] for details.

4.7.2 Tests for Structure Discrimination

Likelihood Ratio Test

The problem is to test the hypothesis

$$H_0 : \theta \in \mathcal{M}_0$$

against the alternative

$$H_1 : \theta \in \mathcal{M}_1$$

where $\boldsymbol{\theta}$ is the parameter vector and $\mathcal{M}_0 \subset \mathcal{M}_1$.

$$\Lambda = \mathcal{L}(\hat{\theta}_0)/\mathcal{L}(\hat{\theta}_1) \tag{4.61}$$

$$LR = -2\log\Lambda \in_{a.s.} \chi^2(s-r) \tag{4.62}$$

where $\mathcal{L}(\theta_i)$ $(i = 0, 1)$ is the maximized likelihood function under H_0 and H_1, respectively. s and r denote the dimensions of $\mathcal{M}_1$ and $\mathcal{M}_0$.

Lagrange Multiplier Test

If the subset $\mathcal{M}_0$ is defined as the r-dimensional subspace for which the s-dimensional parameter $\boldsymbol{\theta}$ is restricted by the vector of independent constraints $\boldsymbol{g}(\boldsymbol{\theta}) = 0$, one may apply the Lagrange Multiplier Test. The test quantity is

$$LM = N^{-1}\hat{\boldsymbol{d}}^T\hat{\boldsymbol{S}}^{-1}\hat{\boldsymbol{d}} \tag{4.63}$$

where

$$\boldsymbol{d} = \frac{\partial \log \mathcal{L}}{\partial \boldsymbol{\theta}} \tag{4.64}$$

and

$$\boldsymbol{S} = N^{-1}\frac{\partial^2 \log \mathcal{L}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} \tag{4.65}$$

and N is the number of observations. $\hat{\boldsymbol{d}}$ and $\hat{\boldsymbol{S}}$ are obtained by evaluating $\boldsymbol{d}$ and $\boldsymbol{S}$ at the *constrained* maximum likelihood estimate $\hat{\theta}$. Under H_0, the test quantity LM is asymptotically distributed as $\chi^2(s-r)$. LM is also know as the efficient-score statistics

Another variant of the test known as Wold's test is obtained by evaluating $\boldsymbol{d}$ and $\boldsymbol{S}$ at the unconstrained maximum likelihood estimate, see [Harvey 1989].

Bayesian Model Discrimination

We wish to test the null hypothesis that the true model is $\mathcal{M}_0$ against the alternative H_1 that the true model is $\mathcal{M}_1$. Let's denote the loss of choosing H_1 when H_0 is true by $L(H_0, H_1)$ and similarly define $L(H_1, H_0)$. If H_0 and H_1 are true with prior probabilities $P(H_0)$ and $P(H_1)$, then Bayes' rule yields that the posterior probabilities (given data $\mathcal{Y}$) become:

$$P(H_0|\mathcal{Y}) = P(\mathcal{Y}|H_0)P(H_0)$$

and similarly for H_1. Now, H_0 is accepted if

$$P(H_1|\mathcal{Y})L(H_1, H_0) < P(H_0|\mathcal{Y})L(H_0, H_1)$$

Combining this last inequality with the Bayes' rule yields that H_0 is accepted if:

$$\frac{P(\mathcal{Y}|H_0)}{P(\mathcal{Y}|H_1)} > \frac{P(H_1|\mathcal{Y})L(H_1, H_0)}{P(H_0|\mathcal{Y})L(H_0, H_1)}$$

This is the same as likelihood ratio test with the advantage that the critical values of the test are determined by the loss of making wrong decision and the prior probabilities.

4.7.3 An Overview of the Model Order Determination Criteria

AIC (Akaike Information Criterion) is given by

$$AIC = -2\log(\text{maximized likelihood}) + 2p$$

The minimizer of AIC determines the model order. A draw back of the criterion is its tendency to overestimate the order of the model for large data records. In order to relieve this problem another criterion called Bayesian Information Criterion (BIC) is often used:

$$BIC = -2\log(\text{maximized likelihood}) + p\log N \tag{4.66}$$

4.7.4 Model Validation and Residual analysis

The idea behind residual analysis is to figure out if the system's response is "close" to the predictions from the model. Under the assumption that the model is valid the residual sequence ϵ_i must be uncorrelated. Thus

$$\hat{\rho}_k = \frac{\sum_{k+1}^{N}(\epsilon_i - \bar{\epsilon})(\epsilon_{i-k} - \bar{\epsilon})}{\sum_{i=1}^{N}(\epsilon_i - \bar{\epsilon})^2} \tag{4.67}$$

will provide an estimate of the auto-correlation function of the residuals for lag $k = 0, 1, 2, ...$ It is well known that under the null hypothesis the sequence is white noise, $\sqrt{N}\hat{\rho}_k$, $k = 1, 2, ...$ is Gaussian with mean value zero and unit variance. Then in this case is easy to test the null hypothesis. There are other tests based on cumulative periodogram and tools involve investigating if the input and the residual sequence are correlated, see [Madsen 1995*a*].

Chapter 5

Parameter estimation in nonlinear systems

5.1 Introductory discussion

This chapter treats parameter estimation in nonlinear systems with discrete time. The systems are supposed to be described with models containing unknown parameters, given the structure determination and tests for nonlinearity that were discussed in the previous chapter. A couple of algorithms for parameter estimation will be discussed, keeping in mind some of important properties of the resulting parameter estimate that are to be sought, as e.g.

- asymptotic unbiasedness, i.e. no systematic error when the number of data points, (N), increased unboundedly,
- asymptotic consistency, i.e. the variance of the estimate goes to zero asymptotically with $N \to \infty$,
- asymptotic efficiency, i.e. minimal variance amongst unbiased estimates.

In addition we are interested in knowing the distribution of the estimates, if that is computable. Finally, we would like to say something about the convergence speed as function fo N.

The general problem will be appoached by the Maximum Likelihood method, and by a Prediction error technique, sometimes, cf. [Tong 1990*b*] called the conditional least-

squares method. Just as in the linear case these two methods will show some similarities, cf. e.g. [**?**] for a discussion. Due to the variety of nonlinear model representations, the properties of the parameter estimates can be expected to be very diverse, hence in addition to study the general principles mentioned, the prediction error method will also be discussed in connection with estimation in models with thresholds and in bilinear models.

The model selection and validation problems are also discussed and a couple of tests will be presented.

5.1.1 On a previous test

The Lagrange multiplier (LM) approach presented in the previous chapter gives a possibility to test possible extensions of a basic linear or nonlinear model with (extra) nonlinear terms without having to estimate the 'large' model, i.e. the model with full parametrization. The procedure saves computations and if the extended model indeed is an overfit the procedure also makes it possible to avoid numerical trouble. The idea is of course also applicable in the linear case, cf. [Bohlin 1978] where a test of this type is shown to give asymptotic maximum discrimination power. The test statistic, which is a function of the efficient score, $\frac{\partial \varepsilon}{\partial \beta}$ has under the nulhypothesis a χ^2 distribution with as many degrees of freedom as the number of parameters in the nonlinear part of the model.

Note the important property of this type of procedure being that it not only gives at test of nonlinearity but also gives an idea about the alternative.

5.1.2 On e.g. Hammerstein models

A Hammerstein model (introduced already in the mid-sixties) is an exemple of a nonlinear model with a regression structure, with a predictor like e.g.

$$\hat{y}(t|\theta) = \sum_i \theta_i^T \varphi_i(u^t, y^{t-1})$$

where the φ_i are arbitrary functions of past data. Look on this as a finite-dimensional parametrization of an unknown nonlinear function. The choice of nonlinear functions, i.e. φ_i depends on possible physics, intuition, data, etc and maybe tested as above.

The Hammerstein case is when we have

$$A(q^{-1})y(t) = B_1(q^{-1})u(t) + .. + B_p(q^{-1})u^p(t) + e(t)$$

where $e(t)$ is a noise sequence and the B:s are polynomials in the backward shift operator. It may be regarded as a simple form for Volterra model with a simple form of generalized transfer function, cf. section 1.2.1. By introducing the unknown parameters in the polynomials as θ and correspondingly for the output and input signals we get something of the sort mentioned above. The predictor thus contains an ARMAX part and a nonlinear part that depends on powers of the input signal. The parameter estimation may be based on e.g. a least-square approach, and have nice convergence properties, given a sufficiently exciting input signal, cf. e.g. [Söderström & Stoica 1988] or [Olbjer et al. 1994].

Note that in this case we have a static and continuous nonlinearity working on the input or the system. In cases when such a nonlinearity acting on an external signal is discontinuous or its derivative is discontinuous, such as e.g. when there is a saturation or a hysteresis before the linear part of the system, a new problem pops up, that of determining the point of change. That problem is however not solved in full generality, but it might be seen (partly) as a threshold problem, cf. below.

Nonlinear models of this polynomial type are common as first approximations to general nonlinear systems. The applicability of least-squares methods for parameter estimation makes the useful also in connection with adaptive control or adaptive signal processing. This is treated in the thesis, [Vadstrup 1988] in connection with adaptive control of nonlinear systems with static nonlinearities on the input. He defined what you might mean with an optimal predictor and showed that under suitable conditions you do indeed get an optimal predictor and optimal control also in the asymptotic adaptive case.

5.2 Maximum Likelihood in general

5.3 An approximative Maximum Likelihood method for nonlinear systems

Let the unknown system be decribed by

$$X_{t+1} = f(X_t, u_t, \theta) + g(X_t, u_t, \theta)w_t, \tag{5.1}$$

$$Y_t = h(X_t, u_t, \theta) + e_t. \tag{5.2}$$

Assume that the initial condition is Gaussian, and that the functions $f(), g()$ and $h()$ are continuous. Write the likelihood for the states, X_t, denoting the sample path of the

state variable by $\mathcal{X}$

$$\begin{aligned} L(\theta, \mathcal{X}^N) &= p_X(\mathcal{X}^N|\theta) = \\ &= \left[\prod^N p_X(X_t|X_{t-1}, \theta)\right] p_X(X_{t_0}|\theta). \end{aligned} \tag{5.3}$$

$p_X(X_{t_0}|\theta)$ is the density for the initial state. This equation is however most often not a possible starting point for achieving a working algorithm when estimating the unknown parameters, θ. The problems are e.g. that the distribution of the states X_t as a whole have to be computed for every time step and every considered candidate of the unknown parameter, θ. Furthermore, usually X is not observed.

However, by alternatively considering the observations, $\mathcal{Y}$, the likelihood can be written

$$\begin{aligned} L(\theta, \mathcal{Y}^N) &= p_Y(\mathcal{Y}^N|\theta) = \\ &= \left[\prod^N p_Y(y_t|\mathcal{Y}^{t-1}, \theta)\right] p_Y(y_{t_0}|\theta). \end{aligned} \tag{5.4}$$

where, the entire history of the output signal is used in the conditioning, since y_t gives an incomplete observation of the state vector. Again, the problem is the unknown densities. However, by considering the one-step prediction errors

$$\varepsilon_t(\theta) = y_t - h(\hat{X}_{t|t-1}, u_t, \theta) \tag{5.5}$$

and assuming Gaussianity of these variables, resting on the assumtion that the observation noise $e(t)$ is Gaussian, then the likelihood for the outputs $\mathcal{Y}$, which is expressed by using the prediction errors, can be completely characterized by the first and second moments in the conditional distributions, which gives

$$L(\theta, \mathcal{Y}^N) = -\frac{1}{2}\prod_t^N \left(\frac{1}{2\pi}\right)^{\frac{N}{2}} \det(R_{t|t-1})^{-\frac{1}{2}} \exp(-\frac{1}{2}\varepsilon_t^T R_{t|t-1}\varepsilon_t), \tag{5.6}$$

giving the loglikelihood

$$\log L(\theta, \mathcal{Y}^N) = -\frac{1}{2}\sum_t^N \left[\log\det(R_{t|t-1}) + \varepsilon_t^T R_{t|t-1}\varepsilon_t)\right] + \text{const}, \tag{5.7}$$

with $R_{t|t-1}$ being the conditional variance, i.e.

$$R_{t|t-1} = \text{Var}(y_t|\mathcal{Y}^{t-1}, \theta). \tag{5.8}$$

Now the approximative maximum likelihood estimate of the parameters are found by minimizing (5.7), giving the estimates $\hat{\theta}$. An estimate of the precision of the estimate may be obtained from the inverse Hessian obtained from the estimation routine.

The Hessian can be interpreted as an estimate of the information matrix. The properties of this approximative algorithm depends on the validity of the assumption about Gaussianity above. Hence no general statements can be made, but will depend on the nonlinear functions involved and on the distributions of the disturbances acting on the state and on the output.

5.4 Maximum likelihood estimation in nonlinear systems

The presentation here will be very brief, a more thorough discussion may be found e.g. in [Hall & Heyde 1980] where general properties of maximum likelihood estimation of parameters in stationary ergodic Markov chains is discussed.

The presentation here will first be in a handwaiving manner, then some very short arguments will be given towards a more strict presentation.

Start with a sample of data

$$\mathcal{Y}^N = (Y_1, \ldots, Y_N) \tag{5.9}$$

from a time series $\{Y_t\}$ with a probability density $P_N(y_1, \ldots, y_N; \theta)$ that depends on a parameter θ, with a true value θ_0. Start with the single parameter case.

Introduce the loglikelihood function $L_N(\theta) = \log P_N(y_1, \ldots, y_N, \theta)$, its gradient $\frac{dL_N(\theta)}{d\theta}$ and the increment of gradient as

$$u_i(\theta) = \frac{dL_i(\theta)}{d\theta} - \frac{dL_{i-1}(\theta)}{d\theta}, \tag{5.10}$$

which gives $\frac{dL_N(\theta)}{d\theta} = \sum u_i$. Assume that $L_0(\theta) = 0$ and that $E_\theta(\frac{dL_N(\theta)}{d\theta})^2 < \infty$.

Then the ML-solution, $\hat{(}\theta)$, to the problem is computed as a solution to the likelihood equation

$$\frac{dL_N(\theta)}{d\theta} = 0. \tag{5.11}$$

The estimate is iteratively computable via the Taylorexpansion of $\frac{dL_N(\theta)}{d\theta}$

$$0 = \sum^N u_i(\theta) - (\theta - \theta^*) I_N(\theta) + \text{remainder}, \tag{5.12}$$

where the expression $I_N(\theta)$, can be regarded as an approximation of Fisher's Information Matrix. It is computed from

$$I_N = \sum^N E(u_i^2 | \mathcal{Y}^{i-1}). \tag{5.13}$$

Now the Maximum Likelihood estimate attains a Gaussian distribution as the number of data points increases

$$[I_N(\theta_0)]^{\frac{1}{2}} (\hat{\theta} - \theta_0) \to \mathcal{N}(0, 1). \tag{5.14}$$

This estimate has a number of nice properties such as e.g. that any other consistent estimate of θ which has an asymptotic Gaussian distribution when normalized with $[I_N(\theta)]^{\frac{1}{2}}$ will have a larger estimation error variance.

This is indeed just browsing over the material. However the important message is that everything is as should be expected and including that we get an expression for the precision of the estimates, to be interpreted as the Fisher information for the estimation procedure.

Furthermore, as usual, the computation of the likelihoodfunction and the estimate of the information matrix might be the difficult and demanding parts of this solution of the estimation problem.

Remark. A (slightly) more precise discussion starts by defining a *martingale*. Let $\mathcal{B}^n$ be a sequence of growing σ-algebras and X_n being a random variable.

Definition 5.1. $(X_n, \mathcal{B}^n)$ is a *martingale* if

- X_n is measurable with respect to $\mathcal{B}^n$
- $E|X_n| < \infty$
- $E(X_n|\mathcal{B}^m) = X_m; \quad m < n$
  if $\geq X_m$ we have a *submartingale* and
  if $\leq X_m$ we have a *supermartingale*.

▲

An important property of this type of processes is that a limited submartingale is convergent!

Introduce the loglikelihood function and its gradient as above in equation(5.10 to arrive at

$$E_\theta(u_i|\mathcal{Y}^{i-1}) = 0$$

where $\mathcal{Y}^i$ is the σ-algebra generated by the Y_i sequence and $\{gL_i, \mathcal{Y}^i\}$ can be shown to be a square integrable martingale.

Hence $\{u_i, \mathcal{Y}^i\}$ is a square integrable martingale difference sequence and

$$I_N(\theta) = \sum_1^N E_\theta(u_i^2|\mathcal{Y}^{i-1})$$

expresses the conditional information on the data on θ ($\approx$ Fisher Information).

Then, under some conditions that are stated in [Tong 1990*b*] from [Hall & Heyde 1980], we get, as $N \to \infty$, that

$$[I_N(\theta)]^{\frac{1}{2}} (\hat{\theta}_N - \theta) \to \mathcal{N}(0, 1)$$

where $\hat{\theta}_N$ is a root of the likelihood equation

$$gL(\theta) = 0.$$

▼

This form of maximum likelihood estimates of parameters in nonlinear systems is indeed also efficient in a specifiable meaning. If S_N is another consistent estimate of the parameter θ, for which we have

$$[I_N(\theta_0)]^{\frac{1}{2}} (S_N - \theta_0) \to \mathcal{N}(0, \gamma^2(\theta_0)), \tag{5.15}$$

where $\gamma()$ is bounded and continuous in θ, then under suitable regularity conditions $\gamma^2(\theta_0) \geq 1$. Hence the ML estimate attains the optimal variance.

In case there are more than one parameter to estimate, the extension of this maximum likelihood algorithm essentially goes via a definition of the increment of the gradient of the likelihoodfunction

$$u(\theta) = \frac{\partial}{\partial\theta}(L_i(\theta) - L_{i-1}(\theta))$$

and a corresponding definition of the 'Fisher information'

$$I_N(\theta) = \sum_1^N E_\theta(u_i(\theta)u_i(\theta)^T|\mathcal{B}_{i-1})$$

leading to consistency and asymptotic normality.

Remark. To see the analogy with the linear case, start with the likelihood function for estimation in an AR(1) process written on regression form

$$P_N(x_1, \ldots, x_N, \theta) = \Pi^N p(x_i|x_{i-1}, \ldots, \theta) \cdot p(\text{initial values})$$

and find the loglikelihoodfunction as

$$L_N(\theta) = \cdots = \text{const} - \frac{1}{2\sigma^2}\sum^N \varepsilon^2(t,\theta)$$

and thus the increment of its gradient

$$u_i = gL_i - gL_{i-1} = \cdots = \frac{1}{\sigma^2}\varepsilon_i\varphi_i.$$

Hence

$$E_\theta(u_i|\mathcal{X}^{i-1}) = 0 \tag{5.16}$$

$$E_\theta(u_i^2|\mathcal{X}^{i-1}) = \cdots = \frac{1}{\sigma^2}\varphi_i^2 \tag{5.17}$$

which as desired, gives

$$\sum E_\theta(u_i^2|\mathcal{X}^{i-1}) = \frac{1}{\sigma^2}\sum \varphi_i^2$$

and the information matrix in the linear case is refound, cf. [**?**] or [**?**]! ▼

Example 5.1. Look at a nonlinear autoregression of first order, an NLAR(1) process, given as

$$X_t = \lambda(x_{t-1},\theta) + \varepsilon_t. \tag{5.18}$$

The loglikelihood function, given the initial values and assuming that the noise sequence is Gaussian, $\varepsilon_t \in \mathcal{N}(0,1)$, is

$$L_N(\theta) = -\frac{1}{2}\sum^N [x_i - \lambda(x_{i-1},\theta)]^2 + \text{constant} \tag{5.19}$$

which gives the increment of the gradient

$$u_i(\theta) = \varepsilon_i\frac{d\lambda(x_{i-1},\theta)}{d\theta}, \tag{5.20}$$

which indeed has zero conditional expectation. The information matrix is

$$I_N(\theta) = \sum^N E_\theta(u_i^2|\mathcal{X}^{i-1}) \tag{5.21}$$

$$= \sum^N \left(\frac{d\lambda(x_{i-1},\theta)}{d\theta}\right)^2 \tag{5.22}$$

$$\to NE\left(\frac{d\lambda}{d\theta}\right)^2, N\to\infty. \tag{5.23}$$

Asymptotically the parameter estimate attains a Gaussian distribution

$$\hat{\theta} - \theta \to \mathcal{N}(0, \left[NE \left(\frac{d\lambda}{d\theta} \right)^2 \right]^{-1}). \tag{5.24}$$

If, in particular $\lambda(x, \theta) = \theta\chi(x)$ we get

$$\hat{\theta}_N = \frac{\sum^N x_i \chi(x_{i-1})}{\sum^N \chi^2(x_{i-1})} \tag{5.25}$$

$$\text{Var}(\hat{\theta}_N) = \frac{1}{NE\chi^2(x_1)}. \tag{5.26}$$

Note that this in indeed a Hammerstein model, revisited. ♦

5.5 Prediction error methods

Prediction error methods in their general form, presented in connection with linear system identification, are applicable also in the nonlinear case. Again, refer to [Hall & Heyde 1980] and to [Tong 1990*b*] for details. The original reference to the prediction error technique is the nonlinear case is [Klimko & Nelson 1978], and the method is treated in depth also in [**?**].

Start by defining the lossfunction

$$Q_N(\theta) = \sum_1^N \left(x_j - \mathrm{E}_\theta(x_j | \mathcal{X}_{j-1}) \right)^2,$$

where the process x_t is predicted by its conditional mean $\mathrm{E}_\theta(x_j|\mathcal{X}_{j-1})$ given the previous observations that are collected in the information set (sigmaalgebra) $\mathcal{X}$. The parameter estimates are given by the solution to the set of equations

$$\frac{\partial Q_N}{\partial \theta_i} = 0; \; i = 1, \ldots, p.$$

We see that the one-step prediction errors for the nonlinear process enter into this formulation of the loss function, but the problem formulation allows different parametrizations e.g. in order to reflect the intended use of the model. Hence, if prediction models might be developped also for k-step predictions with $k > 1$.

In ordet to derive an estimation algorithm, a Taylor expansion of the loss function Q around θ_0 is performed, to get

$$\begin{aligned} Q_N(\theta) &= Q_N(\theta_0) + (\theta - \theta_0) D_\theta Q_N(\theta_0) \\ &+ \frac{1}{2}(\theta - \theta_0)^T V_N (\theta - \theta_0) \\ &+ \frac{1}{2}(\theta - \theta_0)^T T_N(\theta^*)(\theta - \theta_0) \end{aligned}$$

Here D is a gradient, V_N is the Hessian at θ_0 and

$$T_N(\theta^*) = D_\theta^2 Q_N(\theta^*) - V_N$$

and $D_\theta^2 Q_N(\theta^*)$ is the Hessian at θ^*.

Then, if

- $T_N(\theta^*)$ is small
- $(2N)^{-1} V_N \to V$
- $N^{-1} D_\theta Q_N(\theta_0) \to 0$,

there is a sequence of estimates $\{\hat{\theta}_N\}$ such that

$$\hat{\theta}_N \to \theta_0, a.s.$$

and for $\varepsilon > 0; \exists$ an event E with $P(E) > 1 - \varepsilon$ and an N_0 such that on E for $N > N_0$, $\hat{\theta}_N$ satisfies the (conditional) least squares equation

$$\frac{\partial Q_N}{\partial \theta_i} = 0; \ i = 1, \ldots, p.$$

and Q_N attains it relative minimum at $\hat{\theta}_N$.

Note that the discussion on "event E" can be seen as a way of treating the fact that the sequence of estimates does not necessarily stay inside the stationarity region during the iterative search, in particular when the estimate has to be based on few datapoints. When $N \to \infty$ the character of the loss function guarantees the estimate almost surely to stay inside the region in parameter space where stationarity as well as invertibility are guaranteed.

The precision computations done in the linear case carries over to the nonlinear formulation as well, and so does the asymptotic normality. Infact, it is possible to prove

$$N^{0.5}(\hat{\theta}_N - \theta_0) \to \mathcal{N}(0, P_\theta)$$

where $P_\theta = V^{-1}WV^{-1}$ and W stems from the convergence of

$$0.5N^{0.5}D_\theta Q(\theta_0) \to \mathcal{N}(0, W)$$

which is identical with the linear case (to be illustated in a remark below).

Furthermore, for any non-zero vector c of constants one can show a iterated logarithm law like

$$\limsup_{N\to\infty} \frac{N^{0.5}c^T(\hat{\theta}_N - \theta_0)}{(2\sigma^2 \log\log N)^{0.5}} = 1 \quad \text{a.s.}$$
$$\sigma^2 = c^T P_\theta c$$

giving the rate of convergence for $N^{0.5}(\hat{\theta}_N - \theta_0)$.

Example 5.2. Look, as in example (5.1) at estimation of a single parameter in a non-linear autoregression of first order, an NLAR(1) process, given as

$$x_t = \lambda(x_{t-1}, \theta) + \varepsilon_t, \tag{5.27}$$

where the noise sequence is assumed to be a sequence of independent random variables, with a non-specified distribution. The one-step prediction in this model is

$$\hat{x}_{t+1|t} = \lambda(x_t, \theta), \tag{5.28}$$

which gives the gradient of the loss function

$$\begin{aligned} \frac{dQ_N(\theta)}{d\theta} &= \frac{d}{d\theta}\sum^N (x_j - \lambda(x_{j-1}, \theta))^2, \\ &= -2\sum^N \varepsilon_j(\theta)\frac{d\lambda(x_j, \theta)}{d\theta} \end{aligned} \tag{5.29}$$

and its Hessian

$$\frac{d^2Q_N(\theta)}{d\theta^2} = -2\left(\sum^N \varepsilon_j(\theta)\frac{d^2\lambda(x_j,\theta)}{d\theta^2} - \sum^N (\frac{d\lambda(x_j,\theta)}{d\theta})^2\right). \tag{5.30}$$

Now, dividing by N and letting $N \to \infty$ means that the first term in the expression for the Hessian tends to an estimate of the correlation between the noise sequence and a function of previous information, which then tends to zero due to the assumptions on independent noise. This gives that

$$N^{-1}\frac{d^2Q_N(\theta)}{d\theta^2} \approx \frac{2}{N}\sum^N (\frac{d\lambda(x_j,\theta)}{d\theta})^2 \to 2V, \tag{5.31}$$

where

$$V = \lim_{N\to\infty}\frac{1}{N}\sum^N (\frac{d\lambda(x_j,\theta)}{d\theta})^2. \tag{5.32}$$

In the same way we consider the gradient of the lossfunction and define

$$\begin{aligned} W &= \lim_{N\to\infty} \frac{1}{N}\frac{1}{2}\frac{dQ_N(\theta)}{d\theta}(\frac{1}{2}\frac{dQ_N(\theta)}{d\theta}) \\ &= \frac{1}{N}(\sum^N \varepsilon_j(\theta)\frac{d\lambda(x_j,\theta)}{d\theta})(\sum^N \varepsilon_i(\theta)\frac{d\lambda(x_i,\theta)}{d\theta}) \\ &\approx \frac{1}{N}\sum^N (\frac{d\lambda(x_j,\theta)}{d\theta})^2 \to V, \end{aligned} \tag{5.33}$$

such that indeed, as in the previous example 5.1

$$\hat{\theta} - \theta \to \mathcal{N}(0, \left[NE\left(\frac{d\lambda}{d\theta}\right)^2 \right]^{-1}). \tag{5.34}$$

Hence in this case the estimate obtained by applying the prediction error method to the problem has the same asymptotic properties as the estimate gotten when using the maximum likelihood method when the noise was assumed to be Gaussian. Hence, **if** the noise is Gaussian, the prediction error method in this case will produce an efficient estimate. When the noise is **not** Gaussian, then we have seen consistency and asymptotic normality of the resulting estimate in the prediction error case. However, the efficiency question is then unsolved. ⧫

Remark. To be completed!
In the linear case...
It is also possible to include a discussion about the exact expression for the parameter covariance P_θ. It is then the same as in the case discussed in connection with linear systems. ▼

Notice, that in this case we have to compute the conditional expectation in order to predict x_t, or alternatively, we need to invert the process to compute the prediction error. The assumption on invertibility is thus inherent in the algorithm, and a test of invertibility might be a necessary part of the algorithm during the estimation procedure, as is the case in linear estimation in e.g. ARMA-processes, cf. [**?**]. Whether this is a difficult part of the algorithm or not, depends on the process structure. Note also, that we did not assume anything about e.g. how the noise enters into the system!

In chapter **??** the prediction error method will be applied in a practical case of modelling a district heating system, showing results from parameter estimation in linear as well as non-linear models.

5.6 Threshold models

Parameter estamation in threshold models contains some new features to be considered before applying e.g. the prediction error technique. In order to realize that, one may observe that the previous discussion on parameter estimation methods is based on the (not explicitly expressed) assumtion that the parameter estimate as a function of the random data is a continuous random variable. This is in particular manifested in the fact that the asymptotic distribution is found to be Gaussian.

In threshold models, like e.g. in the SETAR(2,1,1) model below

$$x_t = \begin{cases} \alpha_0 + \alpha_1 x_{t-1} + \varepsilon_{\alpha,t} & x_d \leq r \\ \beta_0 + \beta_1 x_{t-1} + \varepsilon_{\beta,t} & x_d > r \end{cases} \tag{5.35}$$

this is not the case. The unknown set of parameters includes in addition to the parameters in the two autoregressions also the threshold value, r, and the delay, d, between current time and the regime decision time. The delay time attains just non-negative integer values and the lossfunction, equation (5.5), is discontinuous at the observations and constant between them when considered as function of the threshold value. The sum of squared prediction errors thus obtains its minimum as function of the threshold value over an interval. Hence, the threshold estimate has to be given particular consideration when part of the estimation procedure, and the presentation below is based on [**?**], and is also presented in [Tong 1990*b*]. However, in cases when the structral parameters in the threshold model are known, the prediction error approach above is directly applicable for estimation of the parameters in the different regime models, giving consistent and asymptotically normal estimates.

Hence, assume that a SETAR-process is given as

$$x_t = \begin{cases} \alpha_0 + \alpha_1 x_{t-1} + .. + \alpha_p x_{t-p} + \sigma_\alpha \varepsilon_{\alpha,t} & x_d \leq r \\ \beta_0 + \beta_1 x_{t-1} + .. + \beta_p x_{t-p} + \sigma_\beta \varepsilon_{\beta,t} & x_d > r \end{cases} . \tag{5.36}$$

The parameters to be estimated by using a dataset with N observations are collected in the vector $\theta = (\alpha^T, \beta^T, r, d)$, and the estimates are collected in the corresponding vector $\widehat{\theta_N}$. The method demands the different regime models to be distinguishable, i.e. and the a upper limit for the delay to be known *a priori*. The estimates are found by minimization of the loss function, i.e.

$$Q_N(\theta) = \sum_1^N \left(x_j - \mathrm{E}_\theta(x_j|\mathcal{X}_{j-1})\right)^2, \tag{5.37}$$

where the minimization procedure results in estimates $\widehat{\alpha}_N(\widehat{r}_N, \widehat{d}_n)$ and $\widehat{\beta}_N(\widehat{r}_N, \widehat{d}_n)$ based on $\widehat{r}_N(\widehat{d}_n)$.

5.7 Validation

The model selection problem, considered as part of the validation contains a discussion about

- the Akaike criterion
- the Bayesian approach
- the crossvalidation
- the Hannan-Quinn approach
- the Rissanen approach

In addition to [**?**], [Tong 1990*b*], [Madsen 1995*b*], [Olbjer et al. 1994] the validation problem is treated in [Holst, Holst, Madsen & Melgaard 1992] for grey-box models and in [Hjorth 1994] with emphasis on computer intensive methods.

Chapter 6

A case study: Non–linear modeling of heat load in district heating systems

The present case study considers the problem of constructing a dynamic model for the heat load in a district heating (DH) system. The approach here will be explicitely to model the nonlinearities in the system.

Another approach is that of using a model, which actually is insufficient, for instance, a linear model, and applying adaptive estimation methods to let the insufficient model approximate the system locally.

A discussion of the advantages and drawbacks of the adaptive approach will be given here, by means of which a motivation for the investigation of the nonlinear modelling approach appears.

Instead of using the physical insight explicitly, the approach in this chapter will be to use knowledge about variables, which are known to have an influence on the heat load in DH systems, and to use statistical methods for determination of model orders.

It may be shown, cf. Sejling (1993), that the most important variables are the ambient air temperature, the supply temperature and the time and type of day.

For the purpose of elaborating on nonlinear dependences the investigation in this chapter will concentrate on forming a model using only these variables to explain the variations of the heat load.

Results will be given for four different model classes:

- General linear transfer function models.
- Transfer function models for second order expansions of explanatory variables.
- Smooth threshold transfer function models.
- Neural network models.

6.1 Data

The investigation is carried out using observations from the DH system in Esbjerg, Denmark. The energy is produced in a Combined Heat and Power plant (CHP), i.e. hot water for district heating as well as electricity is produced. Data consists of two separate periods of observations, hereafter referred to as Data A and Data B, respectively. The observations used are hourly averages of:

- Heat supply.
- Temperature of the supply water from the CHP plant in Esbjerg.
- Ambient air temperature at the CHP plant in Esbjerg.

The observations in Data A were sampled from July 6th until December 10th, 1989, and in Data B from December 28th, 1989 until April 11th, 1990. In these two periods there are no holidays apart from in the first few days of Data B, and the observations from these days are not used for estimation or validation. Note that the first period covers heat production during summer, autumn and beginning of the winter periods, while the second period is dominated by winter conditions.

Throughout the investigation Data A is used for estimation of model parameters, and Data B is used for validation of the models.

That is, the estimated models are applied on a different set of data to test the general applicability of the model estimate.

Data A is chosen for the estimation as this is the set of data, which has the largest variation range of the ambient air temperature.

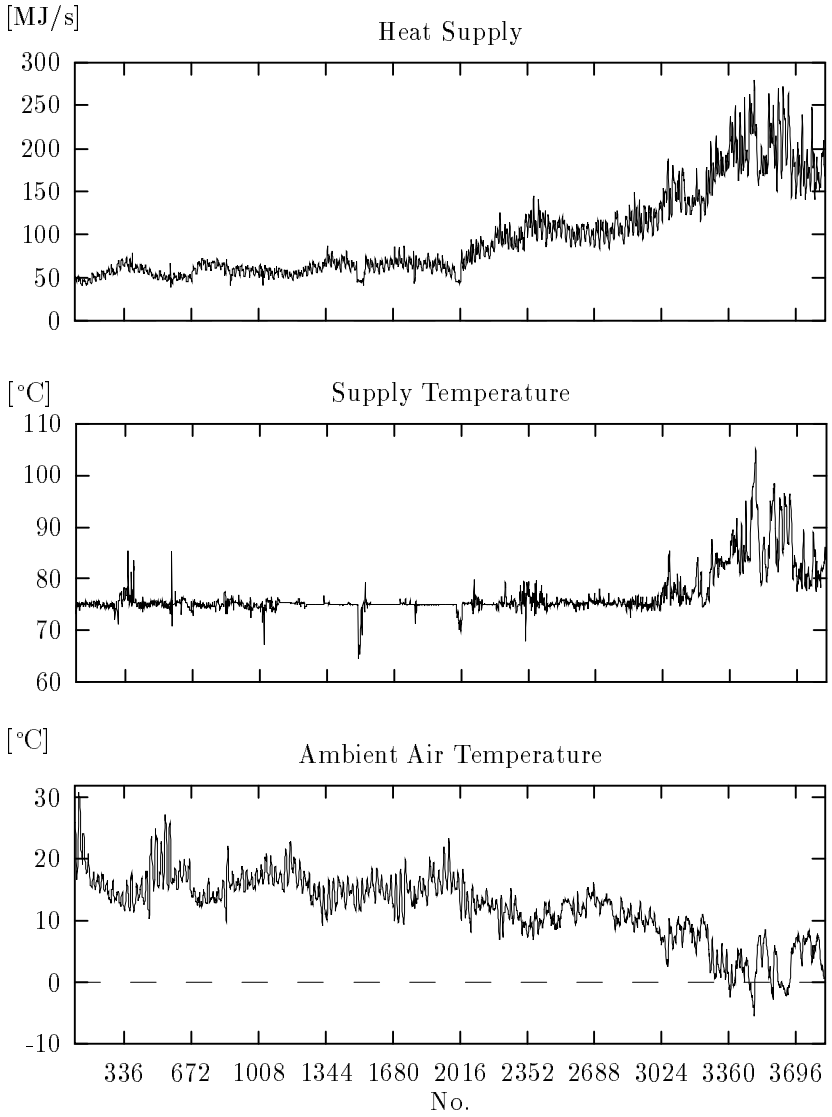

Figure 6.1: Heat supply, supply temperature and ambient air temperature in Esbjerg in Data A.

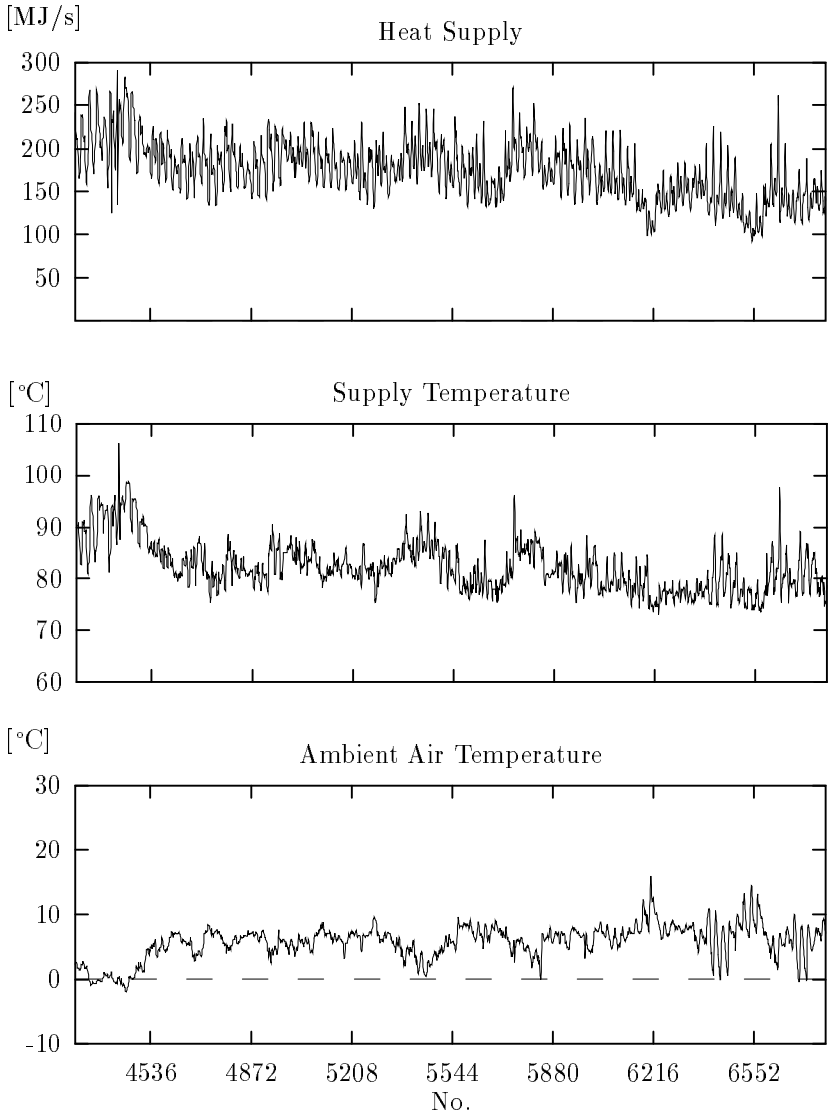

Figure 6.2: Heat supply, supply temperature and ambient air temperature in Esbjerg in Data B.

6.2 Estimation

All models considered can be written on the form:

$$y_t = g_\theta\big(\mathcal{Y}^{t-1}, \mathcal{U}^t\big) + e_t \tag{6.1}$$

with

$$\mathcal{Y}^{t-1} \equiv (y_{t-1}, y_{t-2}, \ldots) \tag{6.2}$$

and

$$\mathcal{U}^t \equiv (u_{1,t}, u_{1,t-1}, \ldots, u_{n_u,t}, u_{n_u,t-1} \ldots)\,. \tag{6.3}$$

That is, y_t is a nonlinear function $g_\theta(\cdot)$ of $(\mathcal{Y}^{t-1}, \mathcal{U}^t)$ superposed by a sequence of random variables, $\{e_t\}$, which are supposed to have zero mean, to have variance σ_e^2, to be independent and identically distributed (i.i.d.), and to be independent of $(\mathcal{Y}^{t-1}, \mathcal{U}^t)$. n_u is the number of external variables in the model.

Estimation method: The parameters in the models will be estimated by the prediction error method, which was discussed in Chapter 5, and see also e.g. Klimko & Nelson (1978) and Ljung (1995), i.e. by a minimization of the sum of squared one-step prediction errors.

This means that the estimate $\hat{\boldsymbol{\theta}}_N$ is found as the parameters minimizing the criterion

$$V(\boldsymbol{\theta}) = \frac{1}{N}\sum_{t=1}^{N} \frac{1}{2}\left(y_t - \hat{y}_t(\theta)\right)^2 = \frac{1}{N}\sum_{t=1}^{N} \frac{1}{2}\varepsilon_t(\theta)^2. \tag{6.4}$$

The one-step prediction $\hat{y}_t(\theta)$ of the process $\{y_t\}$ at time t is obtained by conditioning on the observed $\{y_t\}$ until time $t-1$ and $\{u_{i,t}\}, i \in [1; n_u]$ until time t, i.e. it is assumed that the external variables affecting the process at time t are known at that time. Hence, the prediction is given as the expected value of the process conditioned on these sets of observations for a given value of the parameter vector

$$\hat{y}_t(\theta) \equiv E\left[y_t \mid \boldsymbol{\theta}, \mathcal{Y}^{t-1}, \mathcal{U}^t\right]. \tag{6.5}$$

The minimization routine provides the numerical approximation to the Hessian, $\hat{\boldsymbol{H}}(\hat{\boldsymbol{\theta}}_N)$, obtained in the minimum of the criterion,

$$\hat{\boldsymbol{H}}(\hat{\boldsymbol{\theta}}_N) = \left.\frac{\partial^2 V(\boldsymbol{\theta})}{\partial \theta^2}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_N}, \tag{6.6}$$

and an estimate of the covariance matrix for the parameter estimate is given by

$$\widehat{\mathrm{Cov}}\left\{\hat{\boldsymbol{\theta}}_N\right\} = \frac{\widehat{\sigma_e^2}}{N}\left[\hat{\boldsymbol{H}}(\hat{\boldsymbol{\theta}}_N)\right]^{-1}, \tag{6.7}$$

where

$$\widehat{\sigma_e^2} = \frac{1}{N}\sum_{t=1}^{N}\varepsilon_t(\hat{\theta})^2 \tag{6.8}$$

is an estimate of the innovation variance.

6.3 Performance assessment

The one-step prediction qualities of the obtained model is adequately assessed by the empirical RMS-value of the prediction errors, i.e.

$$SS(\theta) = \sqrt{\frac{1}{t_2 - t_1 + 1}\sum_{t=t_1}^{t_2}\left(y_t - \hat{y}_t(\theta)\right)^2}. \tag{6.9}$$

Alternatively, in order to make comparisons between load modelling efforts in different systems possible, a relative prediction error may be computed, leading to the following quality measure

$$SS^{\%}(\theta) = \sqrt{\frac{1}{t_2 - t_1 + 1}\sum_{t=t_1}^{t_2}\left(\frac{y_t - \hat{y}_t(\theta)}{y_t}100\right)^2}, \tag{6.10}$$

which produces a relative measure of the prediction error in percent of the simultaneous signal value.

A more appropriate criterion for determining the model order is the Bayes Information Criterion (BIC), derived by Schwarz (1978), which apart from a constant is given by

$$BIC = N\log\widehat{\sigma_e^2} + n_p\log N, \tag{6.11}$$

where N is the number of observations used in the estimation, and n_p is the number of parameters in the model. $\widehat{\sigma_e^2}$ is the ML estimator, (6.8), of the innovation variance for normally distributed innovations. An estimate of the appropriate model is then be found by minimizing BIC with respect to model order and model parameterization, cf. Chapter 5.7.

6.4 Modelling using transfer functions

6.4.1 General Transfer Function Models

First only linear transfer function models are considered. The models in this class can be written on the form

$$A(q^{-1})y_t = \sum_{i=1}^{n_u} b_0^i \frac{B^i(q^{-1})}{F^i(q^{-1})} q^{-k_i} u_{i,t} + \frac{C(q^{-1})}{D(q^{-1})} e_t \tag{6.12}$$

with $\{u_{i,t}, i \in [1; n_u]\}$ being external processes and $\{e_t\}$ white noise with variance σ_e^2.

This class is in [Ljung 1987] referred to as *the general family of model structures* including a number of the commonly used model structures as special cases. For instance, the finite impulse response (FIR) model is obtained with all polynomials except the B^i ones being one, and the ARMAX model corresponds to case where the F^i and D polynomials are one.

Each of the polynomials A, B^i, F^i, C and D are polynomials in the back-shift operator. In the implementation they can all be specified as products of a number of lower order polynomials depending on the desired parameterization, i.e. each polynomial in (6.12) is of the following form

$$G(q^{-1}) \equiv \prod_{j=1}^{n_G} G_j(q^{-1}) \equiv \prod_{j=1}^{n_G} 1 + g_{j,1}q^{-1} + \cdots + g_{j,n_{G_j}} q^{-n_{G_j}}. \tag{6.13}$$

It is obvious that a transfer function model being linear in supply and ambient air temperature will not be sufficient in describing the dynamic dependence on these variables. However, the purpose of showing the results with this kind of model is to allow for comparison with subsequent non-linear models.

After estimation of a number of the most appropriate structures of the transfer functions from external variables and noise process, the model below was found

$$\begin{aligned}
\nabla_{24} p_t &= 3.57 \, \frac{1 - 0.91q^{-1}}{1 - 0.80q^{-1}} \, \nabla_{24} s_t \\
&- 1.62 \, \frac{1 - 0.97q^{-1}}{1 - 1.32q^{-1} + 0.33q^{-2}} \, \nabla_{24} a_t + \\
&\left[-2.87 \sin\left(\frac{2\pi t}{24}\right) - 0.34 \cos\left(\frac{2\pi t}{24}\right) + 4.14 \sin\left(\frac{2\pi t}{12}\right) + 2.76 \cos\left(\frac{2\pi t}{12}\right) \right] \nabla_{24} I_{A,t} \\
&+ \frac{1 + 0.05q^{-1} - 0.91q^{-24}}{1 - 1.10q^{-1} + 0.18q^{-2}} \, e_t.
\end{aligned} \tag{6.14}$$

The variables in the model are

- p_t : heat load.
- s_t : supply temperature.
- a_t : ambient air temperature.

$I_{A,t}$ is an indicator function defined as one if the current hour is on a day differing from typical work days and zero otherwise. ∇_{24} is the diurnal difference operator defined by $\nabla_{24}x_t = x_t - x_{t-24}$. This means that $\nabla_{24}I_{A,t}$ is equal to one on the first non-work day following a work day, and equal to minus one on the first work day following a non-work day. For all other days is it zero.

All the parameter estimates of this model were found to be significant on 5 % level using the Hessian approximation as an estimate of the parameter covariance matrix.

This model includes an additive diurnal profile, composed of sine and cosine functions with periods 24 and 12 hours, for weekend days and days classified as non-working days. A diurnal profile common for all days is, in this particular model, redundant due to the diurnal differencing of both heat load and external variables.

6.4.2 Second Order Nonlinear Models

The linearity of the transfer function model is evidently a restriction in the capability of the model to describe load variations and dependence on external variables. Both stationary and dynamic relations obviously depend on operation point and weather condition.

Now it is tried to improve the performance of the dynamic model (6.14) by introducing additional transfer functions with second order components.

As will appear, this extension implies, e.g. that stationary dependency relations between the supply temperature and the heat load varies linearly with the inverse mass flow, being low-pass filtered.

Likewise for the ambient air temperature the stationary dependence varies linearly with the low-pass filtered ambient air temperature.

This kind of models belongs to the class of *second order Volterra series expansions*.

The response of the system to changes in the supply temperature must clearly depend

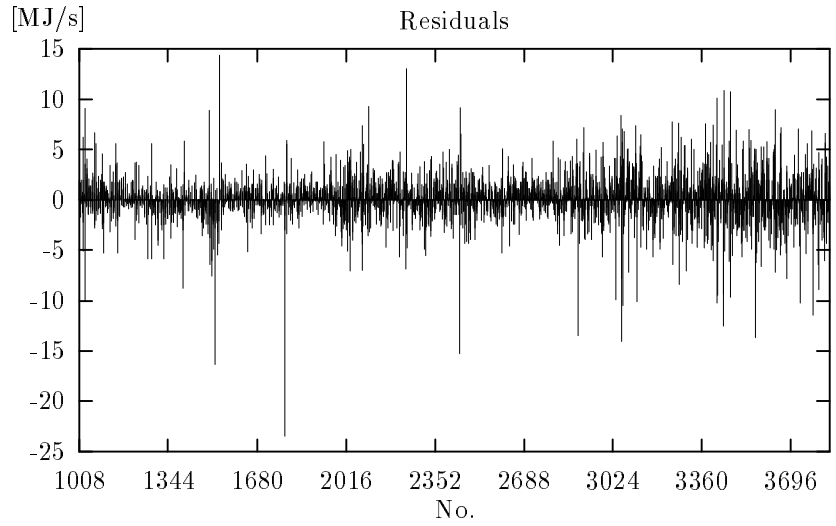

Figure 6.3: Residuals corresponding to model (6.14).

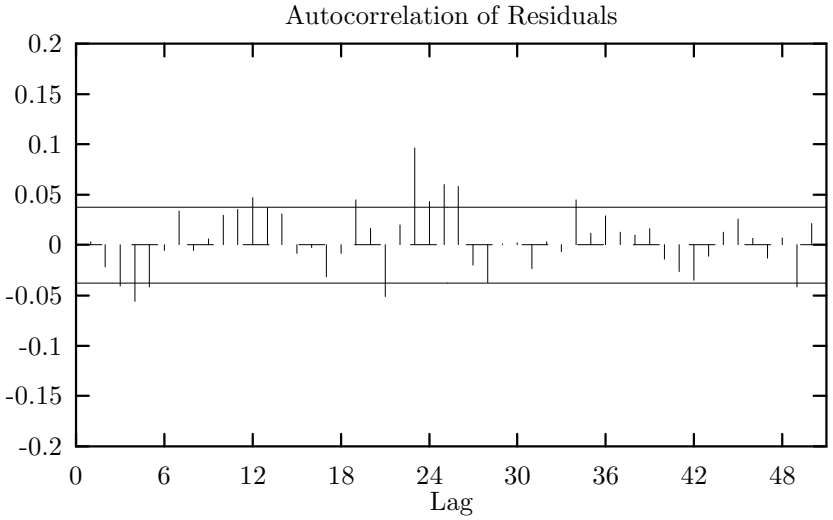

Figure 6.4: Autocorrelation of residuals corresponding to model (6.14).

on the mass flow, since the time delays through the system, in steady state, is proportional to the inverse of the mass flow.

Hence, $\dot{m}_{f,t}^{-1}$ is introduced in a supplementary transfer function from supply temperature to allow for an approximation of the mass flow dependent transfer function from the supply temperature to the heat load.

This approximation of the model implies a linear relation between the stationary gain and the low-pass filtered inverse mass flow.

As the result of estimation of a number of candidate models with second order transfer functions the following model gave the lowest value of the information criterion (BIC)

$$
\begin{aligned}
\nabla_{24} p_t \;=\; & \overset{(.04)}{6.30}\;\frac{1-\overset{(.01)}{1.05}\,q^{-1}}{1-\overset{(.02)}{0.83}\,q^{-1}}\,\nabla_{24} s_t - \\
& \overset{(.02)}{9.32}\;\frac{1-\overset{(.03)}{1.22}\,q^{-1}}{1-\overset{(.02)}{0.87}\,q^{-1}}\dot{m}_{f,t-1}^{-1}\,\nabla_{24} s_t - \\
& \overset{(.08)}{2.88}\;\frac{1-\overset{(.002)}{0.98}\,q^{-1}}{1-\overset{(.02)}{1.29}\,q^{-1}+\overset{(.02)}{0.30}\,q^{-2}}\,\nabla_{24} a_t + \\
& \overset{(.01)}{0.12}\;\frac{1+\overset{(.03)}{0.24}\,q^{-1}}{1-\overset{(.03)}{0.08}\,q^{-1}}a_{f,t}\,\nabla_{24} a_t + \\
& \left[-\overset{(.02)}{2.71}\sin\left(\frac{2\pi t}{24}\right)-\overset{(.02)}{0.18}\cos\left(\frac{2\pi t}{24}\right)+\overset{(.02)}{4.08}\sin\left(\frac{2\pi t}{12}\right)+\overset{(.02)}{2.54}\cos\left(\frac{2\pi t}{12}\right)\right]\nabla_{24} I_{A,t} \\
& + \;\frac{1+\overset{(.006)}{0.05}\,q^{-1}-\overset{(.005)}{0.91}\,q^{-24}}{1-\overset{(.02)}{1.08}\,q^{-1}+\overset{(.02)}{0.17}\,q^{-2}}\,e_t. \qquad (6.15)
\end{aligned}
$$

The filtered variables in this model are determined by exponential smoothing from

$$\dot{m}_{f,t}^{-1} = \frac{1-\hat{\alpha}_1}{1-\hat{\alpha}_1 q^{-1}}\frac{s_t - r_t}{p_t}, \quad \hat{\alpha}_1 = \overset{(.002)}{0.989}, \qquad (6.16)$$

where r_t is the return temperature, and

$$a_{f,t} = \frac{1-\hat{\alpha}_2}{1-\hat{\alpha}_2 q^{-1}}a_t, \quad \hat{\alpha}_2 = \overset{(.001)}{0.994}\,. \qquad (6.17)$$

$\dot{m}_{f,t}$ expresses, apart from a scale parameter, the low-pass filtered mass flow through the heat plant.

The parameters of the filters (6.16) and (6.17) were estimated simultaneously with the other parameters of the model. The filters implement a simple exponential smoothing with unit gain of the inverse mass flow and ambient air temperature, respectively.

Likewise, a transfer function from ambient air temperature with $a_{f,t}$-dependent gain is introduced to obtain an approximation to the dependence of heat load on ambient air temperature.

That is, the additional transfer function has the implication that the dynamic dependence on ambient air temperature is a function of the low-pass filtered ambient air temperature itself. For instance, this model can be given the interpretation that the stationary gain of the relation between heat load and ambient air temperature depends linearly on the level of the ambient air temperature (the low-pass filtered ambient air temperature).

The estimated variance for the residuals is reduced from 5.80 for the linear model to 4.94 for the above non-linear model.

6.4.3 Smooth Threshold Transfer Function Models

Now the model class considered is the *smooth threshold autoregressive* (STAR) model applied on the transfer function model.

The basic idea of using thresholds in general transfer function models is that hereby regimes are introduced in which one particular set of model components of the model is valid. Applying the principle of thresholds in model (6.12) gives the following general formulation of the model

$$A^{J_t}(q^{-1})y_t = \sum_{i=1}^{n_u} b_0^{iJ_t} \frac{B^{iJ_t}(q^{-1})}{F^{iJ_t}(q^{-1})} q^{-k_i^{J_t}} u_{i,t} + \frac{C^{J_t}(q^{-1})}{D^{J_t}(q^{-1})} e_t. \tag{6.18}$$

This means that the total model actually consists of a number of models, all with the structure of a linear transfer function model, where in every time step only one of the models is active. The value of J_t corresponds to the regime being active at time t determining what model structure and parameters are valid at this moment.

The smoothing is made by using a continuous and smooth function, and the logistic

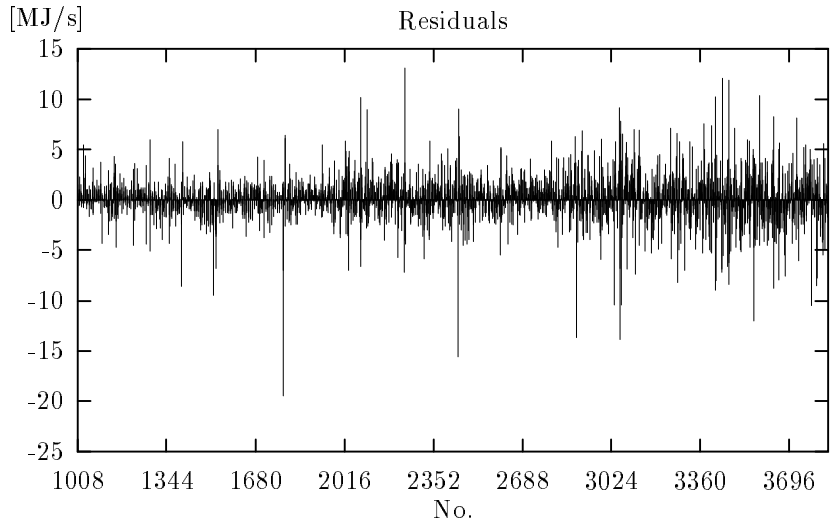

Figure 6.5: Residuals corresponding to model (6.15).

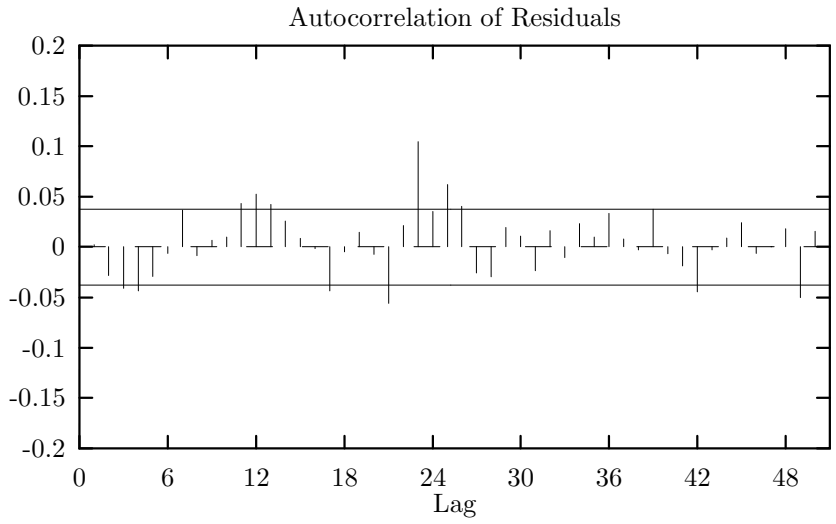

Figure 6.6: Autocorrelation of residuals corresponding to model (6.15).

distribution function is a relatively common choice,

$$F_{\alpha,\beta}(x) = \frac{1}{1+\exp\left(\frac{-(x-\alpha)}{\beta}\right)}. \tag{6.19}$$

Among a number of different models with smooth threshold gain dependence the following resulted in the lowest value of BIC. The standard deviations of the parameters estimates are marked in parenthesis.

$$\begin{aligned}
\nabla_{24} p_t \;=\; & \overset{(.1)}{5.42}\;\frac{1-\overset{(.02)}{1.01}\,q^{-1}}{1-\overset{(.03)}{0.81}\,q^{-1}}\,\nabla_{24}s_t - \\
& \frac{\overset{(.1)}{3.40}}{1+\exp(\overset{(.1)}{4.08}-\overset{(.1)}{15.57}\,\dot{m}^{-1}_{f,t-1})}\,\frac{1-\overset{(.04)}{1.21}\,q^{-1}}{1-\overset{(.05)}{0.85}\,q^{-1}}\,\nabla_{24}s_t + \\
& \left(\frac{\overset{(.1)}{1.19}}{1+\exp(\overset{(.1)}{19.14}-\overset{(.04)}{1.52}\,a_{f,t})}-\overset{(.1)}{1.98}\right)\,\frac{1-\overset{(.008)}{0.97}\,q^{-1}}{1-\overset{(.06)}{1.36}\,q^{-1}+\overset{(.06)}{0.37}\,q^{-2}}\,\nabla_{24}a_t + \\
& \left[-\overset{(.2)}{2.38}\sin\left(\frac{2\pi t}{24}\right)-\overset{(.2)}{0.27}\cos\left(\frac{2\pi t}{24}\right)+\overset{(.1)}{4.02}\sin\left(\frac{2\pi t}{12}\right)+\overset{(.1)}{2.60}\cos\left(\frac{2\pi t}{12}\right)\right]\nabla_{24}I_{A,t} \\
& + \frac{1+\overset{(.008)}{0.05}\,q^{-1}-\overset{(.009)}{0.92}\,q^{-24}}{1-\overset{(.02)}{1.08}\,q^{-1}+\overset{(.02)}{0.17}\,q^{-2}}\,e_t.
\end{aligned} \tag{6.20}$$

The filtered variables in this model are given by

$$\dot{m}^{-1}_{f,t} = \frac{1-\hat{\alpha}_1}{1-\hat{\alpha}_1 q^{-1}}\frac{s_t - r_t}{p_t}, \quad \hat{\alpha}_1 = \overset{(.006)}{0.974} \tag{6.21}$$

and

$$a_{f,t} = \frac{1-\hat{\alpha}_2}{1-\hat{\alpha}_2 q^{-1}} a_t, \quad \hat{\alpha}_2 = \overset{(.009)}{0.951}\ . \tag{6.22}$$

The estimated variance of the threshold model is 4.89 (little better than the second order model).

The transfer function from supply temperature is seen to consist of two separate transfer functions, one of which has a gain being dependent on the low pass filtered scaled inverse mass flow, lagged one hour. Hereby, the composite transfer function at time t includes parameters being function of $\dot{m}^{-1}_{f,t-1}$. $F_{\alpha,\beta}(\dot{m}^{-1}_{f,t-1})$ is introduced to allow for an approximation to a mass flow dependent transfer function from the supply temperature to the heat load.

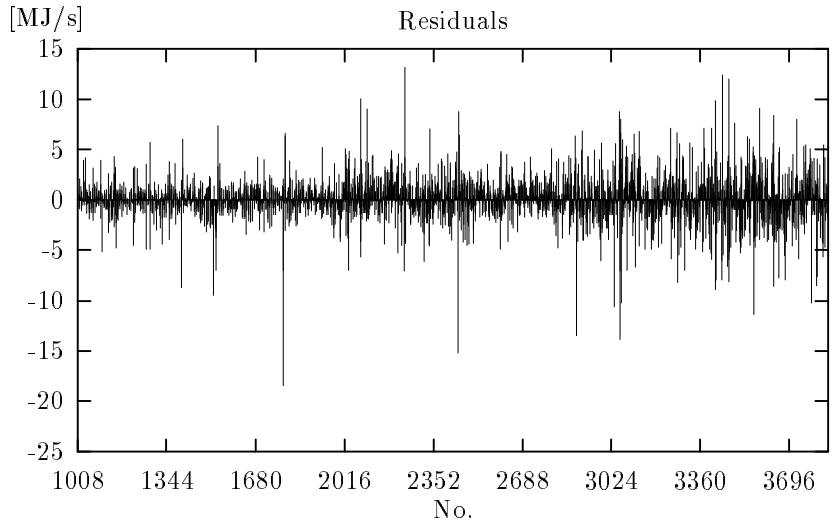

Figure 6.7: Residuals corresponding to model (6.20).

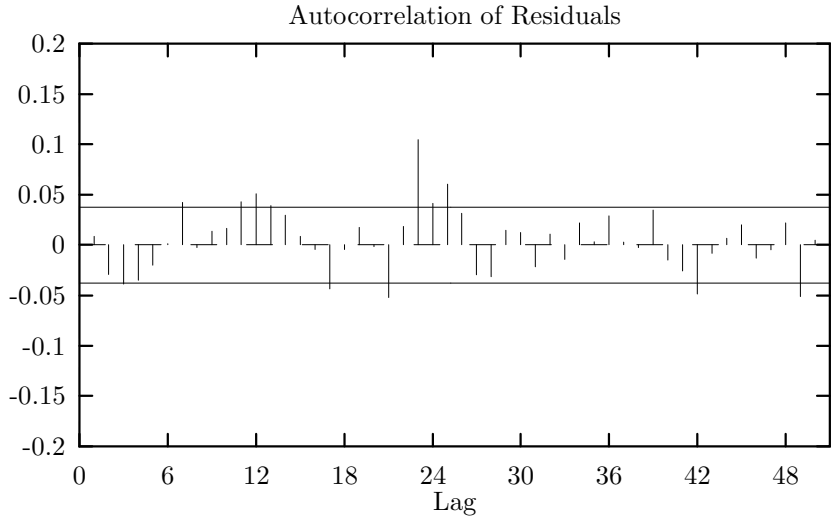

Figure 6.8: Autocorrelation of residuals corresponding to model (6.20).

6.4.4 Analysis of Modelling Results

In this section the results for the three model types considered previously are further analyzed.

First the DC-gain or stationary gain for the three models, $H_s(e^{i\omega})\mid_{\omega=0}$, in the transfer from supply temperature are compared. For the linear model the DC-gain is constant, $H_s(1) = 1.61$ MJ/^{o}Cs, whereas for the second order model there is a linear dependence between gain and $\dot{m}_f^{-1}$, i.e.

$$H_s(1) = 15.4\dot{m}_f^{-1} - 1.89 \text{ MJ/}^o\text{Cs}. \tag{6.23}$$

For the smooth threshold model the DC-gain is

$$H_s(1) = \frac{4.86}{1 + \exp(4.08 - 15.6\dot{m}_f^{-1})} - 0.389 \text{ MJ/}^o\text{Cs}. \tag{6.24}$$

For the observations in Data A $\dot{m}_f^{-1}$ is typically 0.5 ^{o}Cs/MJ in the summer period and between 0.2 and 0.3 ^{o}Cs/MJ in the coldest period. Hereby it is seen that both the threshold model and the second order model bear the, possibly erroneous, quality of postulating a large DC-gain for the summer observations.

In the transfer functions from ambient air temperature to load the stationary gain is found as the fraction between the stationary increase in load and an infinitesimal increase in ambient air temperature, when contributions from all other model components are set to zero. Thus, for the second order model the stationary gain becomes

$$H_a(1) = 0.163a - 9.48 \text{ MJ/}^o\text{Cs}, \tag{6.25}$$

while for the smooth threshold model it is

$$H_a(1) = \frac{5.17}{1 + \exp(19.1 - 1.52a)} - 8.62 \text{ MJ/}^o\text{Cs}. \tag{6.26}$$

For the linear model the gain is $H_a(1) = -7.35$ MJ/^{o}Cs.

These relations between stationary gain and ambient air temperature, shown in Figure 6.10, seem reasonable.

The threshold model shows a two-level gain with a smooth transition for ambient air temperature from 10^{o}C to 15^{o}C. The placement of the two levels is reasonable, and the ambient temperature where the transition takes place seems also to be reasonable.

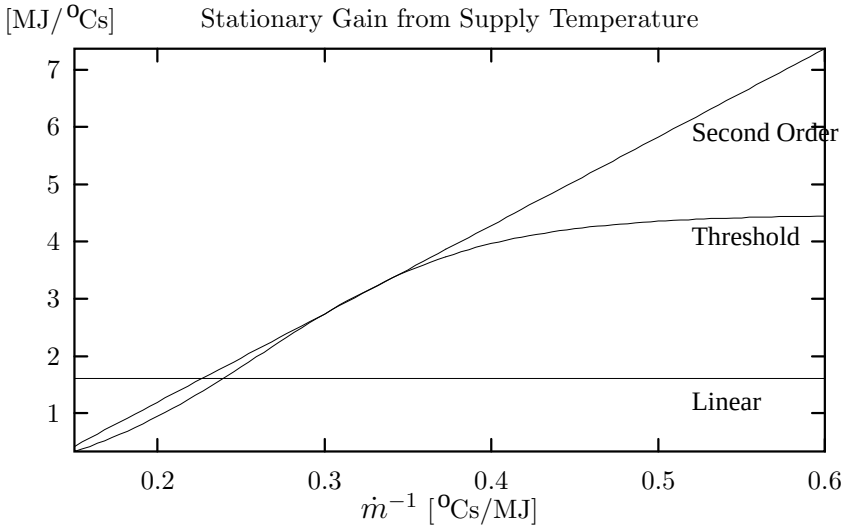

Figure 6.9: Stationary Gain in transfer function from supply temperature.

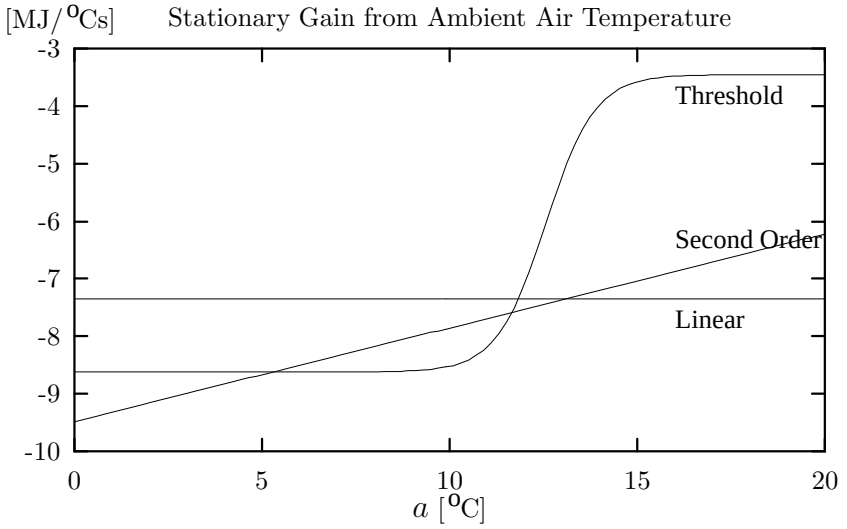

Figure 6.10: Stationary Gain in transfer function from ambient air temperature.

6.5 Modelling using neural network

Finally, artificial neural networks are considered as a class of model structures holding the capability of describing the existing nonlinear relations in the heat load dependence on supply temperature and ambient air temperature.

It has been proved that multi-layer networks hold the capability of approximating any continuous function. (Hornik, Stinchcombe and White, 1989).

This investigation considers the application of one particular choice of neural network in gaining a sufficient nonlinear predictor of the heat demand. This class of neural networks includes one, which previously has been successfully applied to, e.g. the sunspot series.

Architecture

The general structure of the network is shown on Figure 6.11, where the neurons are shown as circular nodes.

In this investigation only models with two layers of neurons are considered. The motivation for the decision of leaving out the models with only one layer of neurons is an intuitive belief that a two-layer model has a higher degree of flexibility in modelling nonlinearities than a one-layer model, for the same number of parameters.

The network model has n_I inputs, $u_{i,t}, i \in [1; n_I]$, and n_1 and n_2 neurons in the two intermediate layers, respectively. The network has one output corresponding to the dependent variable, the variations of which is the object of the modelling.

In the general structure each input is connected to each neuron in the first layer, each neuron in the first layer is connected to each neuron in the second layer and each neuron in the second layer is connected to the output. Each connection has an adjustable weight, $w_{\cdot}$.

The output and each neuron has an adjustable bias, $b_{\cdot}$.

The output from neuron j in the first layer at time t is

$$o_{j,t}^{(1)} = g\left(\sum_{i=1}^{n_I} w_{ij}^{(1)} u_{i,t} - b_j^{(1)}\right), \qquad j \in [1; n_1], \tag{6.27}$$

where $g(\cdot)$ is the activation rule of the neuron, see below.

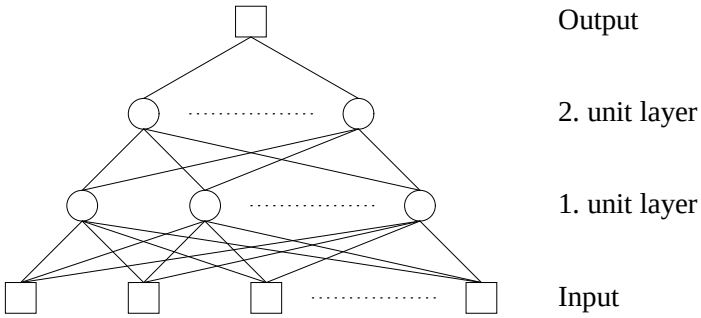

Figure 6.11: Architecture.

The output at time t of neuron k in the second layer is

$$o_{k,t}^{(2)} = g\left(\sum_{j=1}^{n_1} w_{jk}^{(2)} o_{j,t}^{(1)} - b_k^{(2)}\right), \qquad k \in [1; n_2], \tag{6.28}$$

and the output from the network is

$$o_t = \sum_{k=1}^{n_2} w_k^{(o)} o_{k,t}^{(2)} - b^{(o)}. \tag{6.29}$$

The total number of parameters (weights and biases) is

$$N_p = (n_I + 1)n_1 + (n_1 + 2)n_2 + 1. \tag{6.30}$$

Using a neural network as a model describing functional relations has the same character as a non-parametric approach [Ripley 1993, Ripley 1996]. The reason is that a neural network can be applied without using any physical knowledge about the system.

The neural network can simply be applied as a nonlinear approximation to the existing system relations in the same way as a traditional nonparametric method can be applied.

Furthermore, it is not possible to give any physical interpretation of the individual weights and biases in a neural network model.

Activation rule

The nonlinearity of the neural network is introduced in the activation rule of the neurons.

In this investigation the sigmoid, performing a smooth mapping $(-\infty;\infty) \rightarrow (0;1)$, given by

$$g_S(x) = \frac{1}{1+\exp(-ax)}, \tag{6.31}$$

is used.

Dynamics of computation

If the identity $(g(x) = x)$ is used as activation rule, the neural network degenerates to a linear filter. Thus, the network will be able to perform as a predictor obtained from an ARX-model, since for an ARX-model the one-step predictor is linear in both model parameters and regressor variables. The neural network will, however, contain superfluous parameters due to the intermediate layers rendering redundant in a linear filter.

Determining the weights and biases in a neural network is, in the neural network terminology, referred to as *training* of the net.

In statistics the process of determining the weights and biases is referred to as parameter estimation, the weights and biases being the parameters of the model.

It is assumed that the dependent variable can be described by the following model

$$y_t = o(u_t) + e_t. \tag{6.32}$$

Denoting the parameters of the model $\boldsymbol{\theta}$, the prediction of the network is given as

$$\hat{y}_t(\theta) = o_t(u_t;\theta). \tag{6.33}$$

The estimate of the neural network parameters, minimizing the prediction error variance, is thus found by minimization of the criterion function in (6.4).

In the section on general transfer function modelling the best performing model, rewritten to match the formulation in (6.32), would have the form

$$\begin{aligned} H_1(q^{-1})\nabla_{24}p_t &= H_2(q^{-1})\nabla_{24}s_t + H_3(q^{-1})\nabla_{24}a_t \\ &+H_1(q^{-1})\nabla_{24}\nu_{1,t} + H_1(q^{-1})\nabla_{24}\left(\nu_{2,t}I_{A,t}\right) + e_t. \end{aligned} \tag{6.34}$$

$\nu_{1,t}$ and $\nu_{2,t}$ are trigonometric diurnal profiles, the first one actually being superfluous due to the diurnal differencing.

The chosen number of inputs is 15. With the dependent variable being p_t the inputs are

$$\left\{ \begin{array}{lll} p_{t-1} & s_{t-24} & 1 - I_{A,t} \\ p_{t-2} & a_t & \sin(2\pi t/24) \\ p_{t-24} & a_{t-1} & \cos(2\pi t/24) \\ s_t & a_{t-24} & \sin(2\pi t/12) \\ s_{t-1} & I_{A,t} & \cos(2\pi t/12) \end{array} \right\}$$

For this selection of inputs, estimation were carried for $n_1 \in [2; 8]$ and $n_2 \in [2; 3]$.

In the estimation of the neural network models, the model estimate was continuously applied on the validation set (Data B) to test its prediction performance. The reason for doing this that in the field of training neural network models, it is recommended to apply the model, in the course of the estimation, on a test set, and stop the training when the performance on the set deteriorates.

The idea of this approach is that the neural network structure should be so wide that it for a certain set of weights and biases reaches the optimal prediction performance.

6.6 Model validation

Table 6.1, and figure 6.12 summarize the modelling results and show the standard deviation of the prediction errors when the estimated models were applied on the two datasets, Data A and Data B.

The important result is that, although the neural network model is superior in fitting Data A, the second order model and the smooth threshold model gives a smaller prediction error variance when used on Data B. One interpretation of this result might be that explicitely introducing physical characteristics of the system, as was done in the non-linear models, improves the possibility for the model to describe data outside the estimation set.

The linear transfer function model is still inferior to the others, but the difference from the neural network is not so evident.

Model	Estimation (Data A)		Validation (Data B)	# Parameters
	$\hat{\sigma}_e^2$	BIC	$\tilde{\sigma}_e^2$	
Linear	5.80	5104	11.29	15
Second Order	4.94	4695	9.59	23
Smooth Threshold	4.89	4672	9.89	25
Neural Network	3.92	4532	10.99	95

Table 6.1: Results from estimation and validation.

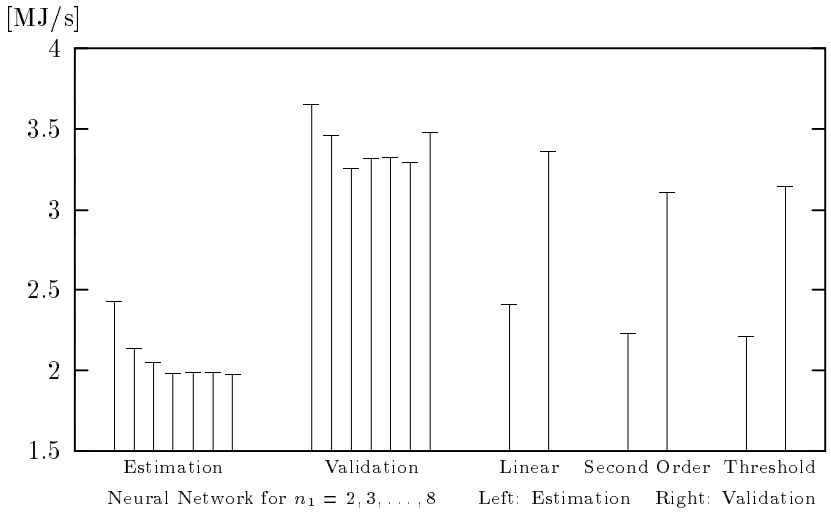

Figure 6.12: Bar diagram of $\hat{\sigma}_e$ for the estimation set and $\tilde{\sigma}_e$ for the validation set.

Chapter 7

State space models

7.1 Introduction

State space models are very useful for describing *non-stationary* or *time varying* systems. The reason is due to the (first order) *Markov property* of the state vector which implies that prediction and state estimation is fairly easy also for time varying systems – see e.g. [Madsen 1995*a*]. But also some *non-linear* systems can conveniently be described in state space form.

This chapter concentrates on state space models and in particular how to estimate parameters in state space models. It is assumed that the reader is familiar with the classical linear stochastic state space model. First a maximum likelihood method for estimating the embedded parameters in a linear state space model is described in Section 7.2.

Later on in Section 7.3 it is illustrated how the parameters of some doubly stochastic models can be estimated by using the above mentioned maximum likelihood method.

A traditional approach for estimating unknown parameters in state space models is to include those parameters in the state vector and then use a state filtering approach, which then (also) provides an estimate of the unknown parameters. An extended Kalman filter for that purpose is briefly described in Section 7.4.

The EM-method, which is a very useful method for parameter estimation in particular cases, as e.g. in the case of missing observations, is described in Section 7.5.

Finally, the class of so-called state dependent models (SDM) is introduced in Sec-

tion 7.6, and it is shown how many of the non-linear time series models can be written in state space form using the SDM. Estimation of parameters in state dependent models is briefly outlined.

7.2 Maximum likelihood estimators for state space models

Let us first consider a linear state space model and a procedure for finding ML estimates of the embedded parameters of the state space model. The estimates are, in general, found by numerical optimization.

The ML estimator for estimating parameters of, for instance, a multivariate ARMAX models will come out of the method as a result under stationary conditions, where the calculations can be tremendously simplified. This is shown in the final part of this section.

The state space approach described is useful also for multivariate ARMAX models with missing or aggregated output data. The state space method is used, for instance, in Splus [Statistical Sciences 1995].

Consider the linear stochastic state space model:

$$\boldsymbol{X}_t = \boldsymbol{A}_t \boldsymbol{X}_{t-1} + \boldsymbol{B}_t \boldsymbol{u}_{t-1} + \boldsymbol{e}_{1,t}, \tag{7.1}$$

$$\boldsymbol{Y}_t = \boldsymbol{C}_t \boldsymbol{X}_t + \boldsymbol{e}_{2,t}, \tag{7.2}$$

where $\{\boldsymbol{e}_{1,t}\}$ and $\{\boldsymbol{e}_{2,t}\}$ are mutually uncorrelated normal distributed white noise sequences with variance $\boldsymbol{\Sigma}_{1,t}$ and $\boldsymbol{\Sigma}_{2,t}$, respectively. Further we assume that $\dim \boldsymbol{Y} = m$ and that the model is global identifiable (will be discussed in a subsequent chapter).

The parameters are embedded in the matrices $\boldsymbol{A}_t$, $\boldsymbol{B}_t$, $\boldsymbol{C}_t$, $\boldsymbol{\Sigma}_{1,t}$ and $\boldsymbol{\Sigma}_{2,t}$, which might be *time varying*, as indicated by the notation.

Remark. We allow for a total *missing observation* of the output vector at time t by putting (using a sloppy notation):

$$\boldsymbol{\Sigma}_{2,t} = \infty \tag{7.3}$$

For simplicity we will, however, assume that m values are observed (possibly aggregated) if an observation is not totally missing. ▼

By this remark and the following approach it is possible to estimate parameters based on time series with missing observations. Let $\mathcal{Y}_t$ denote all observation up till and including time t.

Since $\{\boldsymbol{e}_{1,t}\}$ and $\{\boldsymbol{e}_{2,t}\}$ are normal distributed $\boldsymbol{X}_t|\mathcal{Y}_t$ is also normal distributed and entirely characterized by the mean:

$$\hat{\boldsymbol{X}}_{t|t} = \mathrm{E}\left[\boldsymbol{X}_t|\mathcal{Y}_t\right], \tag{7.4}$$

and variance:

$$\boldsymbol{\Sigma}^{xx}_{t|t} = \mathrm{V}\left[\boldsymbol{X}_t|\mathcal{Y}_t\right]. \tag{7.5}$$

The one-step predictions are:

$$\hat{\boldsymbol{X}}_{t+1|t} = \boldsymbol{A}_t\hat{\boldsymbol{X}}_{t|t} + \boldsymbol{B}_t\boldsymbol{u}_t \tag{7.6}$$

$$\boldsymbol{\Sigma}^{xx}_{t+1|t} = \boldsymbol{A}_t\boldsymbol{\Sigma}^{xx}_{t|t}\boldsymbol{A}_t^T + \boldsymbol{\Sigma}_{1,t} \tag{7.7}$$

$$\hat{\boldsymbol{Y}}_{t+1|t} = \boldsymbol{C}_t\hat{\boldsymbol{X}}_{t+1|t}. \tag{7.8}$$

At time $t+1$ we get the updated estimates by using a *Kalman filter* (see the note [Madsen 1992] about projections in Hilbert spaces) :

$$\begin{aligned}\hat{\boldsymbol{X}}_{t+1|t+1} &= \mathrm{E}\left[\boldsymbol{X}_{t+1}|\mathcal{Y}_t, \boldsymbol{Y}_{t+1}\right] \\ &= \hat{\boldsymbol{X}}_{t+1|t} + \boldsymbol{K}_{t+1}\left(\boldsymbol{Y}_{t+1} - \hat{\boldsymbol{Y}}_{t+1|t}\right),\end{aligned} \tag{7.9}$$

$$\boldsymbol{K}_{t+1} = \boldsymbol{\Sigma}^{xx}_{t+1|t}\boldsymbol{C}_t^T\left[\boldsymbol{C}_t\boldsymbol{\Sigma}^{xx}_{t+1|t}\boldsymbol{C}_t^T + \boldsymbol{\Sigma}_{2,t}\right]^{-1} \tag{7.10}$$

$$\boldsymbol{\Sigma}^{xx}_{t+1|t+1} = \boldsymbol{\Sigma}^{xx}_{t+1|t} - \boldsymbol{K}_{t+1}\boldsymbol{C}_t\boldsymbol{\Sigma}^{xx}_{t+1|t}. \tag{7.11}$$

Note that in case of a missing observation at time $t+1$ no updating of neither the mean nor the variance takes place.

Let us finally introduce the *one-step prediction error* (the *innovation*) and its covariance:

$$\tilde{\boldsymbol{Y}}_{t+1|t} = \boldsymbol{Y}_{t+1} - \hat{\boldsymbol{Y}}_{t+1|t}, \tag{7.12}$$

$$\begin{aligned}\boldsymbol{R}_{t+1} &= \mathrm{V}\left[\tilde{\boldsymbol{Y}}_{t+1|t}\right] = \mathrm{V}\left[\boldsymbol{Y}_{t+1}|\mathcal{Y}_t\right] = \boldsymbol{\Sigma}^{yy}_{t+1|t} \\ &= \boldsymbol{C}_t\boldsymbol{\Sigma}^{xx}_{t+1|t}\boldsymbol{C}_t^T + \boldsymbol{\Sigma}_{2,t}.\end{aligned} \tag{7.13}$$

Since the mean and variance of $\boldsymbol{Y}_{t+1}|\mathcal{Y}_t$ now are calculated we are able to state the conditional density function:

$$f\left(\boldsymbol{Y}_{t+1}|\mathcal{Y}_t\right) = \left[(2\pi)^m \det \boldsymbol{R}_{t+1}\right]^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\tilde{\boldsymbol{Y}}^T_{t+1|t}\boldsymbol{R}^{-1}_{t+1}\tilde{\boldsymbol{Y}}_{t+1|t}\right] \tag{7.14}$$

Let $\boldsymbol{\theta}$ denote the unknown parameters, and $\mathcal{Y}_{\mathcal{N}} = (\boldsymbol{Y}_{\mathcal{N}}, \cdots, \boldsymbol{Y}_0)$ all $\mathcal{N}$ available observations in the time series of length N ($\mathcal{N} < N$).

The *conditional likelihood function* (conditioned on $\boldsymbol{Y}_0$) is

$$\begin{aligned} L(\boldsymbol{\theta}; \mathcal{Y}_{\mathcal{N}}) &= f(\mathcal{Y}_{\mathcal{N}}|\boldsymbol{\theta}) && (7.15) \\ &= f(\boldsymbol{Y}_{\mathcal{N}}|\mathcal{Y}_{\mathcal{N}-1}, \boldsymbol{\theta}) f(\boldsymbol{Y}_{\mathcal{N}-1}|\mathcal{Y}_{\mathcal{N}-2}, \boldsymbol{\theta}) \cdots f(\boldsymbol{Y}_1|\boldsymbol{Y}_0, \boldsymbol{\theta}). && (7.16) \end{aligned}$$

By using the above calculated density function we obtain

$$L(\boldsymbol{\theta}; \mathcal{Y}_{\mathcal{N}}) = \prod_{i=1}^{\mathcal{N}} [(2\pi)^m \det \boldsymbol{R}_i]^{-\frac{1}{2}} \exp\left[-\frac{1}{2} \tilde{\boldsymbol{Y}}_{i|i-1}^T \boldsymbol{R}_i^{-1} \tilde{\boldsymbol{Y}}_{i|i-1}\right], \tag{7.17}$$

As starting values we can take $\hat{\boldsymbol{X}}_{0|0} = 0$ and $\boldsymbol{\Sigma}_{0|0}^{xx} = \alpha \boldsymbol{I}$ where α is 'large'.

The maximization of the likelihood function is equivalent to maximization of

$$\log L(\boldsymbol{\theta}; \mathcal{Y}_{\mathcal{N}}) = -\frac{1}{2} \sum_{i=1}^{\mathcal{N}} [\log \det \boldsymbol{R}_i + \tilde{\boldsymbol{Y}}_{i|i-1}^T \boldsymbol{R}_i^{-1} \tilde{\boldsymbol{Y}}_{i|i-1}] + \text{const.} \tag{7.18}$$

and the ML-estimate of $\boldsymbol{\theta}$ is the argument which maximizes this, i.e.

$$\hat{\boldsymbol{\theta}} = \arg\left\{\max_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}; \mathcal{Y}_{\mathcal{N}})\right\}. \tag{7.19}$$

For the maximization a numerical method must be used.

As an approximation of the covariance of the parameter estimates we can use

$$\mathrm{V}\left[\hat{\boldsymbol{\theta}}\right] \simeq -\boldsymbol{H}^{-1}, \tag{7.20}$$

where

$$\{\boldsymbol{H}\}_{ij} = \left.\frac{\partial^2 \log L(\boldsymbol{\theta}; \mathcal{Y}_{\mathcal{N}})}{\partial \theta_i \partial \theta_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{7.21}$$

Note that $\boldsymbol{H}$ is used as an approximation of the Fisher information matrix.

Using the above described method it is possible to estimated parameters in *time varying models* formulated as linear state space models. In Section 7.3 it is illustrated how the method can be used also for estimating parameters in a class of *non-linear models*, namely the doubly stochastic model.

7.2.1 ML estimates under stationary conditions

It is straight forward to seen that in a *stationary situation* the problem reduces to the following likelihood function

$$L\left(\boldsymbol{\theta};\mathcal{Y}_N\right) = \prod_{t=1}^{N}[(2\pi)^m \det \boldsymbol{\Sigma}]^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\tilde{\boldsymbol{Y}}^T_{t|t-1}\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{Y}}_{t|t-1}\right] \tag{7.22}$$

$$= [(2\pi)^m \det \boldsymbol{\Sigma}]^{N/2} \exp\left[-\frac{1}{2}\sum_{t=1}^{N}\tilde{\boldsymbol{Y}}_{t|t-1}\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{Y}}_{t|t-1}\right] \tag{7.23}$$

The assumption about stationarity implies that the matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are constant. This also means that missing observations are not allowed, cf. (7.3).

Compared to the notation above we now use the symbol $\boldsymbol{\Sigma}$ for the covariance of the innovation. Hence the problem is simplified to calculate the one-step prediction errors $\tilde{\boldsymbol{Y}}_{t|t-1} = \boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_{t|t-1}$, which is easily done for e.g. a multivariate ARMAX model.

The problem is most conveniently separated in the cases $\boldsymbol{\Sigma}$ known and unknown.

$\boldsymbol{\Sigma}$ known

In this case maximization of the likelihood function is equivalent to minimization of $S_1(\boldsymbol{\theta})$ where

$$S_1(\boldsymbol{\theta}) = \frac{1}{N}\sum_{t=1}^{N}\tilde{\boldsymbol{Y}}^T_{t|t-1}\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{Y}}_{t|t-1} \tag{7.24}$$

where $\tilde{\boldsymbol{Y}}_{t+1|t} = \boldsymbol{Y}_{t+1} - \hat{\boldsymbol{Y}}_{t+1|t}$ is the one-step prediction error.

Let us introduce the sample covariance of the predictions errors, i.e.

$$\boldsymbol{D}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{t=1}^{N}\tilde{\boldsymbol{Y}}_{t|t-1}\tilde{\boldsymbol{Y}}^T_{t|t-1} \tag{7.25}$$

then the *cost function* S_1 is written

$$S_1(\boldsymbol{\theta}) = \operatorname{tr}\boldsymbol{\Sigma}^{-1}\boldsymbol{D}(\boldsymbol{\theta}) \tag{7.26}$$

However, if the distribution assumption is postponed, the method is called a *prediction error method* where the cost function is given by S_1. Remember that it is assumed that $\boldsymbol{\Sigma}$ is known.

Σ unknown

In this case the maximization of the likelihood function above is equivalent to minimization of

$$S(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = \frac{1}{2} N \log\det \boldsymbol{\Sigma} + \frac{1}{2} \sum_{t=1}^{N} \tilde{\boldsymbol{Y}}_{t|t-1}^T \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{Y}}_{t|t-1} \tag{7.27}$$

Differentiating with respect to $\boldsymbol{\Sigma}$ gives

$$\frac{\partial S}{\partial \boldsymbol{\Sigma}} = \frac{N}{2} \boldsymbol{\Sigma}^{-1} - \frac{1}{2} \boldsymbol{\Sigma}^{-1} (\sum_{t=1}^{N} \tilde{\boldsymbol{Y}}_{t|t-1} \tilde{\boldsymbol{Y}}_{t|t-1}^T) \boldsymbol{\Sigma}^{-1} \tag{7.28}$$

which equals zero for

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_{t=1}^{N} \tilde{\boldsymbol{Y}}_{t|t-1} \tilde{\boldsymbol{Y}}_{t|t-1}^T = \boldsymbol{D}(\boldsymbol{\theta}) \tag{7.29}$$

This means that for any given value of $\boldsymbol{\theta}$ then (7.29) minimizes (7.27). Therefore the problem can be reduced by imposing the constraint $\boldsymbol{\Sigma} = \boldsymbol{D}(\boldsymbol{\theta})$. By substituting this constraint into the above cost function gives

$$S_c(\boldsymbol{\theta}) = \frac{1}{2} \log\det \boldsymbol{D}(\boldsymbol{\theta}) + \frac{1}{2} \sum_{t=1}^{N} \tilde{\boldsymbol{Y}}_{t|t-1}^T \boldsymbol{D}^{-1}(\boldsymbol{\theta}) \tilde{\boldsymbol{Y}}_{t|t-1} \tag{7.30}$$

or

$$\begin{aligned} S_c(\boldsymbol{\theta}) &= \frac{1}{2} \log\det \boldsymbol{D}(\boldsymbol{\theta}) + \frac{1}{2} \operatorname{tr} \left[\left(\sum_{t=1}^{N} \tilde{\boldsymbol{Y}}_{t|t-1} \tilde{\boldsymbol{Y}}_{t|t-1}^T \right) \boldsymbol{D}^{-1}(\boldsymbol{\theta}) \right] \\ &= \frac{1}{2} \log\det \boldsymbol{D}(\boldsymbol{\theta}) + \frac{N}{2} \operatorname{tr} I_m \\ &= \frac{1}{2} \log\det \boldsymbol{D}(\boldsymbol{\theta}) + \frac{Nm}{2} \end{aligned}$$

Since the last term is constant the ML estimate is found by minimizing the *cost function*

$$S_2(\boldsymbol{\theta}) = \log\det \boldsymbol{D}(\boldsymbol{\theta}) \tag{7.31}$$

Again this is a *prediction error method* if the assumption about the distribution must be postponed, where, in the present case, the cost function is given by S_2.

7.3 Estimation of some doubly stochastic models

In this section it is outlined how the estimation method described in Section 7.2 can be used to estimate the parameters in some doubly stochastic models. The principles outlined can be generalized to other non-linear models.

Consider the *non-linear* process:

$$\begin{aligned} Y_t &= \Phi_t Y_{t-1} + \epsilon_t && (7.32) \\ \Phi_t - \mu &= \phi(\Phi_{t-1} - \mu) + \zeta_t && (7.33) \end{aligned}$$

where $\{\epsilon_t\}$ and $\{\zeta_t\}$ are mutually uncorrelated Gaussian white noise processes with variances σ_ϵ^2 and σ_ζ^2, respectively. This model is some times referred to as an AR(1)-AR(1) model, since both the model for Y_t and the embedded model for the parameter variation is an AR(1) model.

The parameters of the non-linear model are $(\phi, \mu, \sigma_\epsilon^2, \sigma_\zeta^2)$.

By choosing a different parameterization of the model it can, however, be written as a linear state space model. The model for the parameter variation can be written

$$\Phi_t = \phi\Phi_{t-1} + \mu(1-\phi) + \zeta_t$$

Hence, by introducing $\delta = \mu(1-\phi)$ the variation of both $\{Y_t\}$ and $\{\Phi_t\}$ can be described by the linear state space model

$$\begin{aligned} \begin{pmatrix} \Phi_t \\ \delta_t \end{pmatrix} &= \begin{pmatrix} \phi & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \Phi_{t-1} \\ \delta_{t-1} \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} \zeta_t && (7.34) \\ Y_t &= (Y_{t-1}, 0) \begin{pmatrix} \Phi_t \\ \delta_t \end{pmatrix} + \epsilon_t && (7.35) \end{aligned}$$

Note that the new parameter is introduced as a state variable, and that there is a unique relation between the new parameters $(\phi, \delta, \sigma_\epsilon^2, \sigma_\zeta^2)$ and the parameters of the model (7.32).

The principle outlined above is readily generalized. It is, for instance, rather easy to see that the following doubly stochastic model (an AR(1)-AR(2) model)

$$\begin{aligned} Y_t &= \Phi_t Y_{t-1} + \epsilon_t && (7.36) \\ \Phi_t - \mu &= \phi_1(\Phi_{t-1} - \mu) + \phi_2(\Phi_{t-1} - \mu) + \zeta_t && (7.37) \end{aligned}$$

can be written as the following linear state space model

$$\begin{pmatrix} \Phi_t \\ \Phi_{t-1} \\ \delta_t \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \Phi_{t-1} \\ \Phi_{t-2} \\ \delta_{t-1} \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \zeta_t$$

$$Y_t = (Y_{t-1}, 0, 0) \begin{pmatrix} \Phi_t \\ \Phi_{t-1} \\ \delta_t \end{pmatrix} + \epsilon_t$$

where $\delta = \mu(1 - \phi_1 - \phi_2)$.

As soon as the doubly stochastic model is formulated as a linear state space model, then the method for parameter estimation described in Section 7.2 can be used.

It is left as an exercise for the reader to generalize the principles to other doubly stochastic models.

Example 7.1 (Doubly stochastic process - AR(1)-AR(1)). The above outlined class of doubly stochastic processes provides a rather flexible description of non-linear phenomena.

Consider as an example a doubly stochastic process with two different parameter sets as described in Table 7.1.

Parameter	ϕ	μ	σ_ζ^2	σ_ϵ^2
True values No. 1	0.85	0.80	0.10^2	4.0^2
True values No. 2	0.95	0.80	0.10^2	4.0^2

Table 7.1: Parameters used for simulations of a doubly stochastic process.

Simulations of the process are shown on Figure 7.1 and 7.2 for the two set of parameters.

It is noticed that large values of the process is obtained in periods where the system describing the embedded parameter variation is unstable. ♦

By considering for instance a second order parameter model, as the AR(1)-AR(2) model, with complex poles in the transfer function from the noise to the value of the parameter, it is possible to obtain a process which tends to show large values repeatably with a certain period of time.

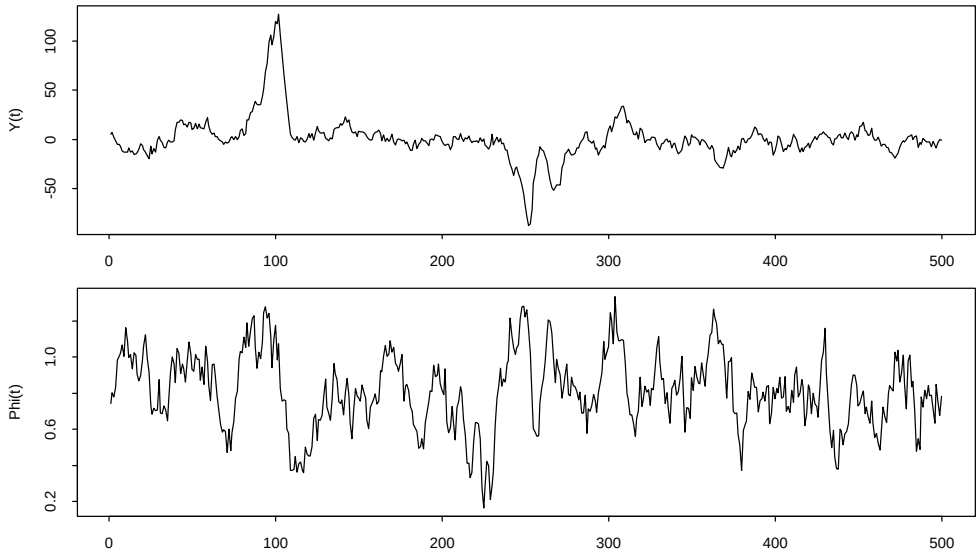

Figure 7.1: A simulation of a doubly stochastic process - $\phi = 0.85$

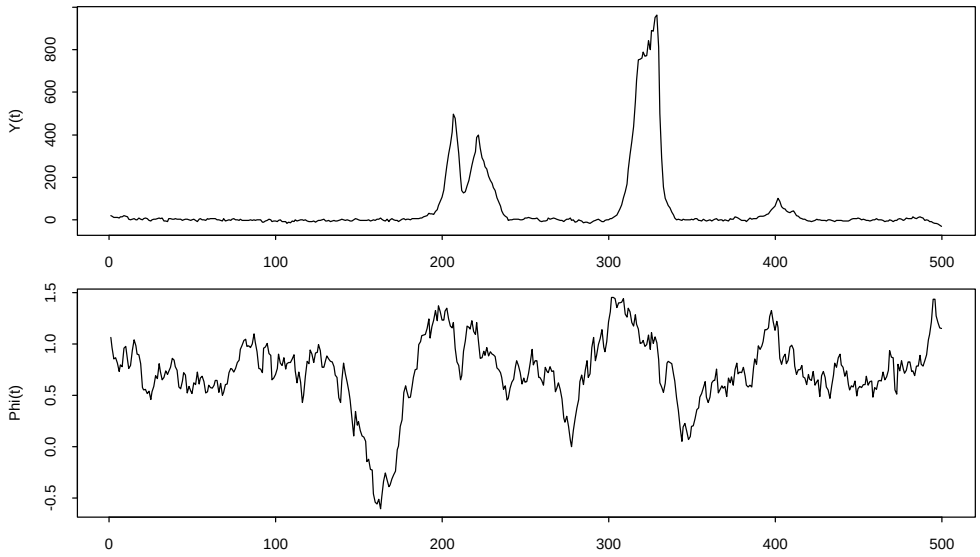

Figure 7.2: A simulation of a doubly stochastic process - $\phi = 0.95$.

7.4 Extended Kalman filter used for parameter estimation

This section describes an often used principle for estimating parameters in state space models, namely to *include the unknown parameters in the state vector* and then use a method for state estimation. The resulting state space model is in general a non-linear state space model, and the extended Kalman filter is then most often used for the state estimation.

Consider the following state space model

$$\begin{aligned} \boldsymbol{X}_{t+1} &= \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{X}_t + \boldsymbol{B}(\boldsymbol{\theta})\boldsymbol{u}_t + \boldsymbol{v}_t && (7.38)\\ \boldsymbol{Y}_t &= \boldsymbol{C}(\boldsymbol{\theta})\boldsymbol{X}_t + \boldsymbol{e}_t && (7.39) \end{aligned}$$

where the parameter vector $\boldsymbol{\theta}$ *is unknown*.

By including ($\boldsymbol{\theta} = \boldsymbol{\theta}_t$) in a new state vector

$$\boldsymbol{Z}_t = \begin{pmatrix} \boldsymbol{X}_t \\ \boldsymbol{\theta}_t \end{pmatrix}$$

this leads to the following non-linear state space model

$$\begin{aligned} \boldsymbol{Z}_{t+1} &= f(\boldsymbol{Z}_t, \boldsymbol{u}_t) + \begin{pmatrix} \boldsymbol{v}_t \\ 0 \end{pmatrix} && (7.40)\\ \boldsymbol{Y}_t &= h(\boldsymbol{Z}_t) + \boldsymbol{e}_t && (7.41)\\ & && (7.42) \end{aligned}$$

where

$$\begin{aligned} f(\boldsymbol{Z}_t, \boldsymbol{u}_t) &= \begin{pmatrix} \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{X}_t + \boldsymbol{B}(\boldsymbol{\theta})\boldsymbol{u}_t \\ \boldsymbol{\theta}_t \end{pmatrix} \\ h(\boldsymbol{Z}_t) &= \boldsymbol{C}(\boldsymbol{\theta})\boldsymbol{X}_t \end{aligned}$$

Now a state estimation will provide a simultaneous estimation of the original state vector and the parameter vector $\boldsymbol{\theta}$.

7.4.1 A version of the Extended Kalman Filter (EKF)

Consider the *non-linear*, discrete time system

$$\begin{aligned} \boldsymbol{X}_{t+1} &= f(t, \boldsymbol{X}_t) + \boldsymbol{v}_t && (7.43)\\ \boldsymbol{Y}_t &= h(t, \boldsymbol{X}_t) + \boldsymbol{e}_t && (7.44) \end{aligned}$$

where

$$\mathrm{E}[\boldsymbol{v}\boldsymbol{v}^T] = \boldsymbol{R}_{1,t} \quad \mathrm{E}[\boldsymbol{e}\boldsymbol{e}^T] = \boldsymbol{R}_{2,t} \quad \mathrm{E}[\boldsymbol{v}\boldsymbol{e}^T] = \boldsymbol{0}$$

Then the EKF estimate of $\boldsymbol{X}_{t+1}$ given $\boldsymbol{Y}_0, \boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_t$ is obtained by

$$\hat{\boldsymbol{X}}_{t+1|t} = \boldsymbol{f}(t, \hat{\boldsymbol{X}}_t) + \boldsymbol{K}_t[\boldsymbol{Y}_t - h(t, \hat{\boldsymbol{X}}_{t|t-1})] \tag{7.45}$$

where

$$\boldsymbol{K}_t = \boldsymbol{F}(t, \hat{\boldsymbol{X}}_{t|t-1})\boldsymbol{P}_t\boldsymbol{H}^T(t, \hat{\boldsymbol{X}}_{t|t-1})[\boldsymbol{H}(t, \hat{\boldsymbol{X}}_{t|t-1})\boldsymbol{P}_t\boldsymbol{H}^T(t, \hat{\boldsymbol{X}}_{t|t-1}) + \boldsymbol{R}_{2,t}]^{-1} \tag{7.46}$$

$$\boldsymbol{P}_{t+1} = \boldsymbol{F}(t, \hat{\boldsymbol{X}}_{t|t-1})\boldsymbol{P}_t\boldsymbol{F}^T(t, \hat{\boldsymbol{X}}_{t|t-1}) + \boldsymbol{R}_{1,t} \tag{7.47}$$

$$- \boldsymbol{K}_t[\boldsymbol{R}_{2,t} + \boldsymbol{H}(t, \hat{\boldsymbol{X}}_{t|t-1})\boldsymbol{P}_t\boldsymbol{H}^T(t, \hat{\boldsymbol{X}}_{t|t-1})]\boldsymbol{K}_t^T \tag{7.48}$$

and

$$\boldsymbol{F}(t, \hat{\boldsymbol{X}}) = \frac{\partial}{\partial \boldsymbol{X}} f(t, \boldsymbol{X})|_{\boldsymbol{X}=\hat{\boldsymbol{X}}} \tag{7.49}$$

$$\boldsymbol{H}(t, \hat{\boldsymbol{X}}) = \frac{\partial}{\partial \boldsymbol{X}} h(t, \boldsymbol{X})|_{\boldsymbol{X}=\hat{\boldsymbol{X}}} \tag{7.50}$$

This is called a *discrete-discrete extended Kalman filter* since both the system equation and the observation equation are given in discrete time.

The basic idea of the extended Kalman filter is thus to linearize about each state estimate. This is different from the *linearized Kalman filter* or *perturbation Kalman filter* where a global linearization of the non-linear model is used.

In Chapter 8 the extended Kalman filter and other higher order filters will be described in more detail.

7.4.2 Some convergence results

The following references deals with the convergence of the state estimation when the EKF is used for parameter estimation.

L. Ljung (1977): *Reduced an investigation of the convergence of EKF to investigate the stability of some ODE.*

L. Ljung (1979): *Showed that divergence can be traced down to miss-specification of the state noise covariance.*

D. Wiberg (1990): *By considering an extension of the EKF to include third order moments, he proved that the EKF estimates, with probability one, converge to some ML estimates.*

New paper: *"If an estimation of the Kalman Gain matrix is included, then the EKF estimates converge to the true values."*

With these new results the EKF may become an attractive way of estimating parameters in non-linear dynamic models.

7.5 The EM algorithm

The EM algorithm is proposed by [Dempster, Laird & Rubin 1977]. It is an iterative method for computing ML estimates of parameters in a model. For some problems the likelihood function for a complete data set is much more easy to find than the likelihood function for the data at hand, which is assumed to be a subset of the complete data set. The idea is then iteratively to find an estimate of the complete likelihood function given the actual data and the current parameter estimate. This estimate of the complete likelihood function is then maximized which results in an updated parameter estimate, which again is used in improving the estimate of the complete likelihood function.

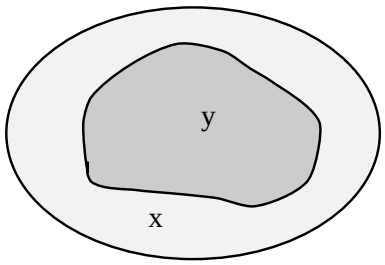

Figure 7.3: Complete (x) and incomplete data (y).

The algorithm contains thus two steps per iteration, the **E**xpectation step for estimating the complete likelihood function conditioning on the available data and the latest parameter estimates, and the **M**aximization step for finding the optimal parameter values in the current iteration.

Given the 'complete data' $\boldsymbol{x}$, the 'incomplete data' $\boldsymbol{y}$ and a mapping $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$. Then the density function of $\boldsymbol{x}$ is $g(\boldsymbol{x}|\boldsymbol{\theta})$ with unknown parameters $\boldsymbol{\theta}$ and the density function of $\boldsymbol{y}$ is

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \int_{\mathcal{X}(\boldsymbol{y})} g(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x}$$

The parameters $\boldsymbol{\theta}$ are to be estimated using the ML method by maximizing $f(\boldsymbol{y}|\boldsymbol{\theta})$ over $\boldsymbol{\theta}$.

Here it is important to note that the basic idea is that it is normally much easier to maximize $g(\boldsymbol{x}|\boldsymbol{\theta})$, i.e. the complete-data likelihood than the incomplete-data likelihood $f(\boldsymbol{y}|\boldsymbol{\theta})$. However, that is not possible due to the lack of complete data. The next point is then to find a method to use the 'easiness', but base the estimation on the incomplete set of data.

As mentioned the optimization of the complete likelihood g represents for many problems an easier task than optimization of the incomplete summary f. However, since $\boldsymbol{x}$

is unobservable, we replace $\log g(\boldsymbol{x}|\boldsymbol{\theta})$ by its conditional mean given $\boldsymbol{y}$ and the current fit $\boldsymbol{\theta}^{(p)}$ (the E step) and maximize the resulting function (the M step).

Hence, let $k(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$ be the conditional density of $\boldsymbol{x}$ given $\boldsymbol{y}$ and $\boldsymbol{\theta}$, i.e.

$$k(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) = \frac{g(\boldsymbol{x}|\boldsymbol{\theta})}{f(\boldsymbol{y}|\boldsymbol{\theta})}.$$

where it has been used that $\boldsymbol{y}$ is a subset of $\boldsymbol{x}$.

Then the log likelihood is (while replacing with the conditional means)

$$\begin{aligned} L(\boldsymbol{\theta}) &= \log f(\boldsymbol{y}|\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) \\ Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) &= \mathrm{E}\left[\log g(\boldsymbol{x}|\boldsymbol{\theta})|\boldsymbol{y},\boldsymbol{\theta}^{(p)}\right] \\ H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) &= \mathrm{E}\left[\log k(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})|\boldsymbol{y},\boldsymbol{\theta}^{(p)}\right]. \end{aligned}$$

Now we can define the EM-iteration

- *E-step* : Compute $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) = \mathrm{E}_p[\log L(\boldsymbol{x},\boldsymbol{\theta}|y_1,\ldots,y_n,\boldsymbol{\theta}^{(p)}]$,
- *M-step* : Choose $\boldsymbol{\theta}^{(p+1)}$ to maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)})$.

That is the EM algorithm. The EM algorithm thus finds the maximum of the likelihood function iteratively.

The notation E_p denotes expectation with respect to the distribution that can be computed using the p'th estimate.

From [Efron 1982] Efron: Maximum Likelihood and Decision Theory, Annals of Stat, **10**, pp 340-356, 1982:

Some properties:

i) First of all
$$L(\boldsymbol{\theta}^{(p)}) \leq L(\boldsymbol{\theta}^{(p+1)})$$
i.e. any EM sequence $\{\boldsymbol{\theta}^{(p)}\}$ increases the likelihood, i.e. $L(\boldsymbol{\theta}^{(p)})$, if bounded above, converges to L^*.

ii) If Q is continuous in both arguments then L^* is a stationary value of L and if $\boldsymbol{\theta}$ converges to some $\boldsymbol{\theta}^*$, then $\boldsymbol{\theta}^*$ is a stationary point.

iii) If [ii], under conditions on Q and if $\| \boldsymbol{\theta}^{(p+1)} - \boldsymbol{\theta}^{(p)} \| \to 0, p \to \infty$ this point is a local maximum (under assumption that they are isolated).

iv) If $L(\boldsymbol{\theta})$ is unimodal (control by selecting different starting points for the estimation) over the whole parameter space, and if the partial derivative of Q with respect to the first argument is continuous in both arguments, i.e. in both $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{(p)}$, then $\boldsymbol{\theta}^{(p)}$ converges to the unique maximizer $\boldsymbol{\theta}^*$ of $L(\boldsymbol{\theta})$.

-i) It is sensitive to the choice of initial value.

-ii) It might end up in a saddle point.

-iii) It might take some time to get along, i.e. sometimes it is a little slow.

Example 7.2 (EM-estimation of parameters in a NLAR(1) model). Consider the *Non-Linear AR(1)-model (NLAR(1))*

$$Y_t = h(Y_{t-1}, \theta) + \epsilon_t \tag{7.51}$$

where $\{\epsilon_t\}$ is i.i.d. N(0,1).

Assume that the available observations are

$$\mathcal{Y}_N = (Y_1, Y_2, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_N) \tag{7.52}$$

i.e. Y_j is missing.

The i'th iteration of the EM-algorithm is

- *E-step* : Find $\mathrm{E}[Y_j|\mathcal{Y}_N, \theta^{(i)}] = Y_j^{(i)}, \mathrm{E}[Y_j^2|\mathcal{Y}_N, \theta^{(i)}] = (Y_j^2)^{(i)}, \mathrm{E}[h(Y_j)|\mathcal{Y}_N, \theta^{(i)}] = h(Y_j)^{(i)}, \mathrm{E}[h^2(Y_j)|\mathcal{Y}_N, \theta^{(i)}] = h^2(Y_j)^{(i)}$.

  In the evaluation it can be used that (for the NLAR(1)-model)

  $$p(Y_j|\mathcal{Y}_N) = \frac{p(Y_j|Y_{j-1})p(Y_{j+1}|Y_j)}{p(Y_{j+1}|Y_{j-1})} \tag{7.53}$$

  Note also that $p(Y_{j+1}|Y_j) = \phi(Y_{j+1} - h(Y_j))$ where ϕ is the density function of N(0,1).

- *M-step* : Minimize with respect to θ the 'sum of squares'

  $$\sum_{t=1}^{N} Y_t^2 - 2\sum_{t=1}^{N} Y_t h(Y_{t-1}) + \sum_{t=1}^{N} h^2(Y_{t-1}) \tag{7.54}$$

  where we replace $Y_j, Y_j^2, h(Y_j), h^2(Y_j)$ by the values found under the E-step. This gives $\theta^{(i+1)}$.

♦

7.5.1 Estimation of parameters in a state space model

Given the system

$$\begin{aligned} \boldsymbol{x}_t &= \boldsymbol{\Phi}\boldsymbol{x}_{t-1} + \boldsymbol{v}_t;\ t = 1, \ldots, N \\ \boldsymbol{y}_t &= \boldsymbol{M}_t\boldsymbol{x}_t + \boldsymbol{e}_t. \end{aligned}$$

where $\{\boldsymbol{M}_t\}$ is assumed known.

Suppose that the input and output noises as well as the initial values are stochastic processes with Gaussian distributions with parameters $(0, \boldsymbol{R}_1)$, $(0, \boldsymbol{R}_2)$ and $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, respectively. The states as well as the parameters in the initial value and noise distributions and the parameters in the $\boldsymbol{\Phi}$ matrix are to be estimated.

The complete data set does in this case contain $\{\boldsymbol{x}_i\}_0^N$ and $\{\boldsymbol{y}_i\}_1^N$. Since everything is Gaussian, the log likelihood function for the complete data is

$$\begin{aligned} 2\log L = &-\log\det\boldsymbol{\Sigma} - (\boldsymbol{x}_0 - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_0 - \boldsymbol{\mu}) - \\ &- n\log\det\boldsymbol{R}_1 - \sum_1^N (\boldsymbol{x}_t - \boldsymbol{\Phi}\boldsymbol{x}_{t-1})^T\boldsymbol{R}_1^{-1}(\boldsymbol{x}_t - \boldsymbol{\Phi}\boldsymbol{x}_{t-1}) - \\ &- n\log\det\boldsymbol{R}_2 - \sum_1^N (\boldsymbol{y}_t - \boldsymbol{M}_t\boldsymbol{x}_t)^T\boldsymbol{R}_2^{-1}(\boldsymbol{y}_t - \boldsymbol{M}_t x_t) \end{aligned} \quad (7.55)$$

E-step :

The first step is to find the conditional mean for this function given the actually measured data $\{\boldsymbol{y}_i\}_1^N$. The parameters that are to be estimated are collected in

$$\boldsymbol{\theta}^{(r)} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Phi}, \boldsymbol{R}_1, \boldsymbol{R}_2)^T.$$

The conditional expectation is

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \mathrm{E}_r(\log L|y_1, \ldots, y_N).$$

Note, that E_r denotes expectation with respect to the distribution that can be computed using the r'th estimate.

Introducing (notice that the usual $\hat{x}$-notion is skipped)

$$\boldsymbol{x}_{t|s} = \mathrm{E}(\boldsymbol{x}_t|y_1, \ldots, y_s)$$

and correspondingly for the covariance $\boldsymbol{P}_{t|s}$ we get

$$\begin{aligned} 2Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = &- \log\det\boldsymbol{\Sigma} - \mathrm{tr}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{P}_{0|N} + (\boldsymbol{x}_{0|N} - \boldsymbol{\mu})(\cdot)^T)\} - \\ &- N\log\det\boldsymbol{R}_1 - \mathrm{tr}\{\boldsymbol{R}_1^{-1}(\boldsymbol{C} - \boldsymbol{B}\boldsymbol{\Phi}^T - \boldsymbol{\Phi}\boldsymbol{B}^T + \boldsymbol{\Phi}\boldsymbol{A}\boldsymbol{\Phi}^T)\} - \\ &- N\log\det\boldsymbol{R}_2 - \mathrm{tr}\{\boldsymbol{R}_2^{-1}\sum_1^N[(\boldsymbol{y}_t - \boldsymbol{M}_t\boldsymbol{x}_{t|N})(\cdot)^T + \boldsymbol{M}_t\boldsymbol{P}_{t|N}\boldsymbol{M}_t^T]\} \end{aligned} \quad (7.56)$$

where

$$\begin{aligned} \boldsymbol{A} &= \sum_1^N(\boldsymbol{P}_{t-1|N} + \boldsymbol{x}_{t-1|N}\boldsymbol{x}_{t-1|N}^T) \\ \boldsymbol{B} &= \sum_1^N(\boldsymbol{P}_{t|N}^{t,t-1} + \boldsymbol{x}_{t|N}\boldsymbol{x}_{t-1|N}^T) \\ \boldsymbol{C} &= \sum_1^N(\boldsymbol{P}_{t|N} + \boldsymbol{x}_{t|N}\boldsymbol{x}_{t|N}^T). \end{aligned}$$

where $\boldsymbol{P}_{t|N}^{t,t-1}$ is the covariance between $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-1}$ given all N observations.

Kalman smoother estimates The values $\boldsymbol{x}_t^N$, $\boldsymbol{P}_t^N$ and $\boldsymbol{P}_{t,t-1}^N$ are thus the expectation and covariance of $\boldsymbol{x}_t$ given all the N observations. These quantities are calculated using the *Kalman smoother* estimates which can be obtained recursively by using the *backward recursions* for $t = N, N-1, \ldots, 1$

$$\boldsymbol{x}_{t-1|N} = \boldsymbol{x}_{t-1|t-1} + \boldsymbol{J}_{t-1}(\boldsymbol{x}_{t|N} - \boldsymbol{x}_{t|t-1}) \quad (7.57)$$

$$\boldsymbol{J}_{t-1} = \boldsymbol{P}_{t-1|t-1}\boldsymbol{\Phi}^T(\boldsymbol{\Phi}\boldsymbol{P}_{t-1|t-1}\boldsymbol{\Phi}^T + \boldsymbol{R}_1)^{-1} \quad (7.58)$$

where $\boldsymbol{P}_{t-1|t-1}$ is the covariance obtained by using an ordinary (forward) Kalman filter.

The covariance of the smoothed estimator satisfies

$$\boldsymbol{P}_{t-1|N} = \boldsymbol{P}_{t-1|t-1} + \boldsymbol{J}_{t-1}(\boldsymbol{P}_{t|N} - (\boldsymbol{\Phi}\boldsymbol{P}_{t-1|t-1}\boldsymbol{\Phi}^T + \boldsymbol{R}_1))\boldsymbol{J}_{t-1}^T \quad (7.59)$$

M-step :

If we now define the $(r+1)$'st parameter estimate as the value of $\boldsymbol{\theta}$ that maximizes the

conditional expectation $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$, we get in this case a closed solution which is

$$\begin{aligned}
\boldsymbol{\Phi}^{(r+1)} &= \boldsymbol{B}\boldsymbol{A}^{-1} \\
\boldsymbol{R}_1^{(r+1)} &= N^{-1}(\boldsymbol{C} - \boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T) \\
\boldsymbol{R}_2^{(r+1)} &= N^{-1}\sum_1^N [(\boldsymbol{y}_t - \boldsymbol{M}_t\boldsymbol{x}_{t|N})(\cdot)^T + \boldsymbol{M}_t\boldsymbol{P}_{t|N}\boldsymbol{M}_t^T]
\end{aligned}$$

The initial values are just given as a single replicate. Choose $\boldsymbol{\mu}^{(r+1)} = \boldsymbol{x}_0^N$ and $\boldsymbol{\Sigma}$ reasonable.

See [Shumway 1988] for further details.

- Hence we have closed the loop with the $(r+1)$'st estimate. Iterate until convergence.
- We have simultaneously computed filtered and smoothed estimates of the states and estimates of the parameters in the system.
- Note that the Kalman filter is included in the EM-algorithm for computation of the conditional estimates. Compare with the previous described ML method for estimating parameters in state space models, where the likelihood function was optimized directly! (using a numerical method). One of the things that we then got, but we didn't get in the algorithm as it is presented here is the precision of the parameter estimates. If that is to be included in this algorithm, it has to be extended.

7.6 State-dependent models (SDM)

State-dependent models are proposed by [Priestley 1980]. A good overview including some examples is found in [Priestley 1988].

In Chapter 1 we introduced the general non-linear model,

$$X_t = h'(\epsilon_t, \epsilon_{t-1}, \dots) \tag{7.60}$$

which clearly is "infinite dimensional", since infinitely many past variables are used. The idea behind the state-dependent models is to reduce this to a "finite dimensional" form.

In a parsimonious finite description it might be a good idea to consider previous X and ϵ-values simultaneously.

Hence consider:

$$X_t = h(X_{t-1}, \cdots, X_{t-p}, \epsilon_{t-1}, \cdots, \epsilon_{t-q}) + \epsilon_t \tag{7.61}$$

Note that the function h in (7.61) is not, of course, the same function as that which appears in Chapter 1.

Define the "state vector"

$$\tilde{\boldsymbol{x}}_t = (\epsilon_{t-q+1}, \cdots, \epsilon_t, X_{t-p+1}, \cdots, X_t)^T$$

Priestly (1980, 1988) uses the term "state vector" although the definition does not in general lead to a *minimal realization*[1]. It is for instance well-known that for an ARMA(p, q) model the minimal realization has a state vector of dimension $\max(p, q+1)$, whereas the state vector $\tilde{\boldsymbol{x}}_t$ has the dimension $(q+p)$.

After a first-order expansion of h about some fixed time point we obtain the *State-Dependent Model (SDM) of order (p,q)*:

$$X_t + \sum_{i=1}^{p} \phi_i(\tilde{\boldsymbol{x}}_{t-1}) X_{t-i} = \mu(\tilde{\boldsymbol{x}}_{t-1}) + \epsilon_t + \sum_{i=1}^{q} \psi_i(\tilde{\boldsymbol{x}}_{t-1}) \epsilon_{t-i} \tag{7.62}$$

[1]Recall that a realization of a state space model (represented by the triplet of matrices ($\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$)) is minimal iff the corresponding system is completely controllable and completely observable

(Note that this is also a *doubly stochastic model* - see Chapter 1).

The SDM model has an appealing interpretation as a locally linear ARMA model in which the parameters ϕ, ψ and μ all depend on the "state" of the process, $\tilde{\boldsymbol{x}}$, at time $(t-1)$.

By selecting appropriate forms for the coefficients, the state-dependent model contains, as special cases, many of the well-known non-linear models described in Chapter 1.

Example 7.3 (Some non-linear models on SDM form). Let us show by examples how some of the well-known non-linear models are written as state-dependent model.

- *ARMA models* : μ, $\{\phi_i\}$ and $\{\psi_i\}$ constants.
- *Bilinear models* : μ, $\{\phi_i\}$ constants, and

  $$\psi_i(\tilde{\boldsymbol{x}}_{t-1}) = c_i + \Sigma_{j=1}^{m} b_{ji} X_{t-j} \quad ; \quad i = 1, \cdots, q \tag{7.63}$$

  By selecting q properly and by putting some of c_i or b_{ji} equal to zero the general bilinear model (1.34) is obtained.
- *Threshold models* :

  $$\begin{aligned} \{\psi_i\} &= 0 \\ \mu_t &= a_0^{(J_t)} \\ \phi_i(\tilde{\boldsymbol{x}}_{t-1}) &= -a_i^{(J_t)} \quad \text{if} \quad X_{t-d} \in R^{(J_t)} \end{aligned}$$

  Note, for the threshold model, the ϕ are step-functions which are constrained to depend on just one component of the state vector, namely X_{t-d}.
- *Exponential AR models* : For $\mu = 0$, $\{\psi_i\} = 0$ and

  $$\phi_i(\tilde{\boldsymbol{x}}_{t-1}) = -(\alpha_i + \beta_i e^{-\delta X_{t-1}^2}) \quad i = 1, \cdots, p \tag{7.64}$$

  then (7.62) reduces to the EXPAR model.

♦

7.6.1 A formal state space description of the SDM

The "state vector"

$$\tilde{\boldsymbol{x}}_t = (\epsilon_{t-q+1}, \cdots, \epsilon_t, X_{t-p+1}, \cdots, X_t)^T$$

together with the innovation, ϵ_{t+1}, determine completely the evolution from time t to time $t+1$.

A formal state space representation is

$$\begin{aligned} \tilde{\boldsymbol{x}}_{t+1} &= \boldsymbol{\mu}(\tilde{\boldsymbol{x}}_t) + \boldsymbol{F}(\tilde{\boldsymbol{x}}_t)\tilde{\boldsymbol{x}}_t + \tilde{\boldsymbol{\epsilon}}_{t+1} \\ X_t &= \boldsymbol{C}\tilde{\boldsymbol{x}}_t \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\mu}(\tilde{\boldsymbol{x}}_{t-1}) &= (0, \cdots, 0 \vdots 0, \cdots, \mu(\tilde{\boldsymbol{x}}))^T \\ \tilde{\boldsymbol{\epsilon}}_t &= \epsilon_t(0, \cdots, 1 \vdots 0, \cdots, 1)^T \\ \boldsymbol{C} &= (0, \cdots, 0, 1) \end{aligned}$$

and

$$\boldsymbol{F}(\tilde{\boldsymbol{x}}_t) = \left[\begin{array}{ccccc:ccccc} 0 & 1 & 0 & \cdots & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & & & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 & 0 & \cdots & 1 \\ \psi_q & \psi_{q-1} & & \cdots & \psi_1 & -\phi_p & -\phi_{p-1} & & \cdots & -\phi_1 \end{array}\right] \tag{7.65}$$

It is obvious from the discussion above that the state dependent model (7.62) represents a *local linearization* of the more general non-linear model (7.61), and the model may be interpreted as representing the tangent plane to the surface $h(\tilde{\boldsymbol{x}})$.

Hence we may introduce the "state dependent transfer function"

$$H(z) = \frac{1 + \sum_{i=1}^{q} \psi_i(\tilde{\boldsymbol{x}}_{t-1}) z^{-i}}{1 + \sum_{i=1}^{p} \phi_i(\tilde{\boldsymbol{x}}_{t-1}) z^{-i}} \tag{7.66}$$

7.6.2 Estimation of state-dependent models

In order to estimate a state-dependent model functional forms of the parameters $\mu(\tilde{\boldsymbol{x}}), \phi_i(\tilde{\boldsymbol{x}}), \psi_i(\tilde{\boldsymbol{x}})$ are needed.

For example we may assume a linear function of the state vector:

$$\phi_i(\tilde{\boldsymbol{x}}_t) = \phi_i^{(0)} + \tilde{\boldsymbol{x}}_t{}^T \boldsymbol{\alpha}_i \tag{7.67}$$

$$\psi_i(\tilde{\boldsymbol{x}}_t) = \psi_i^{(0)} + \tilde{\boldsymbol{x}}_t{}^T \boldsymbol{\beta}_i \tag{7.68}$$

This form is appropriate for bilinear models (with $\boldsymbol{\alpha}_i = 0$); but it is hardly reasonable to suppose that all non-linear models can be approximated with the above linear approximation. However, as long as ϕ and ψ are reasonable "smooth" functions of $\tilde{\boldsymbol{x}}_t$ it would be quite reasonable to use the linear functions as a locally linear approximation. Hence α and β are themselves state-dependent.

The basic solution is then to include α and β in the "state vector" and allow them to vary smoothly in time and to adapt the underlying variation. The natural approach is to suggest a *random walk* variation for α and β.

The estimation procedure is then based on an algorithm which is closely related to the Kalman filter – see [Priestley 1988] for details.

Chapter 8

Stochastic processes and state filtering

8.1 Introduction – Physical modelling

In this chapter we shall start considering continuous time modelling of physical systems. The essence of the chapter is an attempt to answer the question on how to incorporate the deterministic physical knowledge in the model while extending it also to include a description of the disturbances.

The basic deterministic and physically based model of a real process or system is assumed to be described by a set of non-linear differential equation. In most practical applications, one should also include a stochastic part in the equations. There are a number of arguments for including such a stochastic model component. Clearly, noise is an intrinsic part of any real world system, e.g. as measurement disturbances. The stochastic part may also represent disturbances or inputs to the system which are not measured or known, or it may represent unmodelled dynamics of the system.

The chapter then will start by presenting a seemingly possible way of creating the stochastic extension, which however will fail, but will lead to the introduction of the Wiener process (Brownian motion), and the definition of a stochastic differential equation (sde).

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, t)\, dt + \boldsymbol{G}(\boldsymbol{x}_t, t)\, d\boldsymbol{\beta}_t \tag{8.1}$$

$\boldsymbol{\beta}_t$ is standard Brownian motion (Wiener process), to be defined shortly. The solution of such an equation is defined via a stochastic integral which is introduced and discussed and the solution to the sde is presented.

The presentation leads to a state space formulation of the system description. Most often the state variables are unmeasurable and only a function of them is available. The computation of an estimate of the state is a filtering problem, the solution of which is exactly expressable but not always computationally achievable. Hence there is a demand for approximative filters and in the end of the chapter, a number of such approximations are presented.

The contents of the chapter are in various parts, but not always as a whole, treated in the literature. The stochastic differential equations as such are discussed in e.g. [Gard 1988], [Kloeden & Platen 1992] and [Øksendal 1989], the filtering problem is the topic of e.g. [Åström 1970], [Jazwinski 1970] and [Maybeck 1982].

8.1.1 A first attempt to a stochastic differential equation

A seemingly very natural and heuristic extension of the differential equation

$$\frac{d\boldsymbol{x}_t}{dt} = \boldsymbol{f}(x_t, t) \tag{8.2}$$

to treating also a stochastic disturbance, is to add on a stochastic signal $\boldsymbol{v}(\boldsymbol{x}_t, t)$, which is supposed to have zero mean, in order to get

$$\frac{d\boldsymbol{x}_t}{dt} = \boldsymbol{f}(x_t, t) + \boldsymbol{v}(\boldsymbol{x}_t, t). \tag{8.3}$$

Let us suppose that the equation (8.3) is a state space model, which implies that $\boldsymbol{v}(\boldsymbol{x}_t, t)$ and $\boldsymbol{v}(\boldsymbol{y}_s, s)$ are independent $\forall \boldsymbol{x}_t, \boldsymbol{y}_s$ when $t \neq s$ (otherwise the future states would depend on the way the current states were obtained). If it furthermore is assumed that $\boldsymbol{v}(\boldsymbol{x}_t, t)$ is continuous in mean square sense, i.e.

$$\mathrm{E}[\boldsymbol{v}(\boldsymbol{x}_{t+h}, t+h) - \boldsymbol{v}(\boldsymbol{x}_t, t)] \to \mathbf{0}; \qquad h \to 0$$

and has a finite variance $\forall t$, then it is possible to show, see [Åström 1970] that

$$\mathrm{E}[v^2(\boldsymbol{x}_t, t)] = 0 \quad ! \tag{8.4}$$

Hence, it is not possible to keep both the assumption on finite variance and independence, and if the final aim is to achieve a state space model, the assumption on finite variance of the derivative has to be sacrified. The construction of a stochastic differential equation then starts with the Wiener process.

8.1.2 The Wiener process

A Wiener process is a Gaussian stochastic process $\{\boldsymbol{\beta}_t, t \geq 0\}$. The mathematical properties which defines it are given by:

(i) $\boldsymbol{\beta}_0 = 0$ w.p.1

(ii) The increments $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$, $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$, $\cdots$, $\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n-1}$, of the process, for any partitioning of the time interval $0 \leq t_0 < t_1 < \cdots < t_n < \infty$ are mutually independent.

(iii) The increment $\boldsymbol{\beta}_t - \boldsymbol{\beta}_s$ for any $0 \leq s < t$ is Gaussian with mean and covariance respectively

$$\mathrm{E}[\boldsymbol{\beta}_t - \boldsymbol{\beta}_s] = 0 \; ; \qquad \mathrm{V}[\boldsymbol{\beta}_t - \boldsymbol{\beta}_s] = \sigma^2 |t - s| \tag{8.5}$$

Some other characteristics of the Wiener process are e.g. (cf. [Maybeck 1982]):

(i) The Wiener process is a Markov process with respect to $\mathcal{A}_t$ ($\mathcal{A}_s$ is defined as the σ-algebra generated by $\boldsymbol{\beta}_\cdot$ in the time interval $[0, s]$) i.e.

$$p(\boldsymbol{\beta}_t | \mathcal{A}_s) = p(\boldsymbol{\beta}_t | \boldsymbol{\beta}_s) \tag{8.6}$$

for $0 \leq s \leq t$.

(ii) It is furthermore a martingale with respect to $\mathcal{A}_t$, i.e. $\mathrm{E}[|\boldsymbol{\beta}_t|] < \infty$, and for all $0 \leq s < t$,

$$\mathrm{E}[\boldsymbol{\beta}_t | \mathcal{A}_s] = \boldsymbol{\beta}_s \qquad \text{w.p.1} \tag{8.7}$$

(iii) The sample paths of the process are continuous with probability one, but they are nowhere differentiable w.p.1.

Hence, the Wiener process is a nonstationary stochastic process that starts in 0 and has independent, stationary and Gaussian increments. By considering the process in $[0, t]$, we see that the variance of the process increases linearly with t, and that the covariance function fulfills

$$r_{\boldsymbol{\beta}}(s, t) = \sigma^2 \min(s, t). \tag{8.8}$$

The continuity of the process follows from the continuity of the covariance function in (8.8) on the diagonal $s = t$. The non-differentiability of the process, e.g. in mean square sense can be seen from

$$\mathrm{E}\left[\frac{(\boldsymbol{\beta}_{t+h} - \boldsymbol{\beta}_t)}{h}\right]^2 = \frac{\sigma^2 h}{h^2} = \frac{\sigma^2}{h} \to \infty, \qquad h \to 0.$$

The covariance function for the 'derivative' of the process is

$$r_{\boldsymbol{\beta}'}(s,t) = \frac{\partial^2 r(s,t)}{\partial s \partial t} = -\delta(s-t) \tag{8.9}$$

where $\delta(t)$ is a Dirac pulse, hence the covariance function for the 'derivative' is defined only as a generalised function, i.e. in distribution sense. Using it in this sence, it is possible to compute e.g. the spectral density of the 'derivative' of the Wienerprocess to show that is has a constant spectral density for all frequencies and thus has infinite variance. Note that this generalised process is discontinuous everywhere and that it at time t is independent of the process at all other time instants $s \neq t$. This is the process that is closest to the concept 'continuous time white noise'. It is now to be used in constructing a stochastic differential equation.

8.1.3 A stochastic differential equation

One possible way of computing the solution to a deterministic differential equation is to use finite differences, Eulers method being the most simple one. Conversely, the deterministic differential equation may be defined from a difference equation formulation by letting the timestep tend to zero. If the same procedure is applied in the stochastic case, we have to start with stochastic *difference* equation

$$\boldsymbol{x}_{t+h} - \boldsymbol{x}_t = h\boldsymbol{f}(\boldsymbol{x}_t, t) + \boldsymbol{G}(\boldsymbol{x}_t, t)(\boldsymbol{\beta}_{t+h} - \boldsymbol{\beta}_t) \tag{8.10}$$

where the now introduced stochastic process $\{\boldsymbol{\beta}_t\}$ is supposed to be a Wiener process. Note in particular, that drift and diffusion expressions

$$\begin{aligned} \mathrm{E}\left[\boldsymbol{x}_{t+h} - \boldsymbol{x}_t | \boldsymbol{x}_t\right] &= h\boldsymbol{f}(\boldsymbol{x}_t, t) \quad \text{– drift} & (8.11)\\ \mathrm{V}\left[\boldsymbol{x}_{t+h} - \boldsymbol{x}_t | \boldsymbol{x}_t\right] &= h\boldsymbol{G}^2(x_t, t) \quad \text{– diffusion} & (8.12) \end{aligned}$$

where it clearly is impossible to divide by h and let it tend to zero, since the diffusion expression depends linearly on h, due to the properties of the Wiener process. However, it is possible to let the timestep tend to zero in equation (8.10), which gives us the desired expression for the *stochastic differential equation*, cf. equation (8.1),

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, t)dt + \boldsymbol{G}(\boldsymbol{x}_t, t)d\boldsymbol{\beta}_t, \tag{8.13}$$

with $\{\boldsymbol{\beta}_t\}$ being a Wiener process, $\mathrm{E}[d\boldsymbol{\beta}_t]^2 = dt$.

8.2 Stochastic differential equations and stochastic integrals

The stochastic differential equation in equation (8.13) is defined using the Wiener process, hence the derivative of the solution $\boldsymbol{x}_t$ does not exist as a normal function. Thus, the interpretation of the solution remains to be discussed.

Again, compare with the deterministic case, where the solution to an ordinary differential equation may be computed using a pathwise integration. Hence consider the stochastic differential equation

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, t)\, dt + \boldsymbol{G}(\boldsymbol{x}_t, t)\, d\boldsymbol{\beta}_t \tag{8.14}$$

where $\boldsymbol{f}(\boldsymbol{x}_t, t)$ is the drift coefficient, $\boldsymbol{G}(x_t, t)$ is the diffusion coefficient, and $\boldsymbol{\beta}_t$ is a standard Wiener process, representing the source of noise in the system. A solution to (8.14) would formally have the form

$$\boldsymbol{x}_t = \boldsymbol{x}_0 + \int_0^t \boldsymbol{f}(\boldsymbol{x}_s, s)\, ds + \int_0^t \boldsymbol{G}(\boldsymbol{x}_s, s)\, d\boldsymbol{\beta}_s \tag{8.15}$$

The first of these integrals. i.e. the integral of the drift, can be intrepreted as a standard integral defined e.g. in the Riemann sense.

The second integral is however more difficult to handle, since as discussed previously, the sample paths of the Wiener process have unbounded variation. The problem is seen if we try to approximate the integral with a sum

$$\sum_{i=0}^{N-1} G(x_{\tau_i}, \tau_i) \left(\boldsymbol{\beta}_{t_{i+1}} - \boldsymbol{\beta}_{t_i}\right), \qquad t_i \leq \tau_i < t_{i+1} \tag{8.16}$$

and define the integral as the mean square limit of this rectangular approximation for all partitions $t_0 < t_1 < \cdots < t_N = t$ as the maximum step size $\max_i(t_{i+1} - t_i) \to 0$. The result clearly depends on the choice of τ_i, since x_{τ_i} to a varying extent is correlated with $\boldsymbol{\beta}_{t_i}$ In fact, varying choices of τ_i lead to various integral concepts, cf. e.g. [Gard 1988]. Here we choose to have

$$\tau_i = t_i \tag{8.17}$$

which gives the approximating summation

$$\sum_{i=0}^{N-1} \boldsymbol{G}(\boldsymbol{x}_{t_i}, t_i) \left(\boldsymbol{\beta}_{t_{i+1}} - \boldsymbol{\beta}_{t_i}\right), \tag{8.18}$$

the *Itô stochastic integral*. An existence and uniqueness theorem, proved using successive approximations, holds for (8.18) when the drift and diffusion coefficients satisfy Lipschitz and bounded growth conditions, see [Gard 1988, Kloeden & Platen 1992].

This, so defined integral has a couple of interesting properties, such as having the mean value zero and giving the covariance between two integrals as the Riemann integral of the mean value of the product of the integrands.

Hence, the solution to the stochastic differential equation is now well defined, and we may continue to compute functions of the solution to the stochastic differential equation, such as its moments.

8.2.1 Itô Calculus

Defining the Itô integral, one should notice that the ordinary calculus does not apply any more. For example, if $y_t = h(t, x_t)$ with x_t a solution to (8.1) and h a sufficiently smooth function we obtain

$$dy_t = (\frac{\partial h}{\partial t} + f\frac{\partial h}{\partial x} + \frac{1}{2}G^2\frac{\partial^2 h}{\partial x^2})\,dt + \frac{\partial h}{\partial x}G\,d\boldsymbol{\beta}_t \tag{8.19}$$

8.3 Solving stochastic differential equations

Let the model be described by the vector Itô stochastic differential equation

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta}, t)dt + \boldsymbol{G}(\boldsymbol{x}_t, \boldsymbol{\theta}, t)d\boldsymbol{\beta}_t \tag{8.20}$$

with $\{\boldsymbol{\beta}_t\}$ being a standard Wiener process. The observations $\boldsymbol{y}_k$ are taken at discrete time instants, t_k

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_k, \boldsymbol{u}_k, \boldsymbol{\theta}, t_k) + \boldsymbol{e}_k \tag{8.21}$$

where $\{\boldsymbol{e}_k\}$ is a Gaussian white noise process independent of $\boldsymbol{\beta}_t, \forall t, k$, and $\boldsymbol{e}_k \sim N(\boldsymbol{0}, \boldsymbol{S}(\boldsymbol{\theta}, \boldsymbol{t_k}))$. (The super- and subscript k is shorthand notation for t_k).

Condition 1 *(Itô Equation).* Suppose the real functions $\boldsymbol{f}$ and $\boldsymbol{G}$, and initial condition $\boldsymbol{x}_0$, satisfy the following conditions. $\boldsymbol{f}$ and $\boldsymbol{G}$ satisfy uniform Lipschitz conditions in $\boldsymbol{x}$:

$$\|\boldsymbol{f}(\boldsymbol{x}_2, t) - \boldsymbol{f}(\boldsymbol{x}_1, t)\| \leq K\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|,$$
$$\|\boldsymbol{G}(\boldsymbol{x}_2, t) - \boldsymbol{G}(\boldsymbol{x}_1, t)\| \leq K\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|.$$

$\boldsymbol{f}$ and $\boldsymbol{G}$ satisfy Lipschitz conditions in t on $[t_0, T]$:

$$\|\boldsymbol{f}(\boldsymbol{x}, t_2) - \boldsymbol{f}(\boldsymbol{x}, t_1)\| \leq K\|t_2 - t_1\|,$$
$$\|\boldsymbol{G}(\boldsymbol{x}, t_2) - \boldsymbol{G}(\boldsymbol{x}, t_1)\| \leq K\|t_2 - t_1\|.$$

The initial condition $\boldsymbol{x}_0$ is a random variable with $\mathrm{E}[\|\boldsymbol{x}_0\|^2] < \infty$, independent of $\{\boldsymbol{\beta}_t, t \in [t_0, T]\}$.

If condition 1 is satisfied then $\{\boldsymbol{x}_t\}$ is a Markov process, and, in the mean square sense, is uniquely determined by the initial condition $\boldsymbol{x}_0$, see [Jazwinski 1970].

8.3.1 State Transition Distribution

The evolution of the probability density $p(\boldsymbol{x}_t|\mathcal{Y}^k),\ t \in [t_k, t_{k+1}[$ is described by *Kolmogorov's forward equation* or the *Fokker-Planck* equation:

$$dp(\boldsymbol{x}_t|\mathcal{Y}^k) = L(p)\, dt \qquad t \in [t_k, t_{k+1}[\tag{8.22}$$

where

$$L(\,\cdot\,) = -\sum_{i=1}^{n} \frac{\partial(\,\cdot\, f_i)}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^{n} \frac{\partial^2(\,\cdot\,(\boldsymbol{G}\boldsymbol{G}^T)_{ij})}{\partial x_i x_j} \tag{8.23}$$

is the forward diffusion operator, see [Jazwinski 1970, Maybeck 1982]. This is the evolution *between* observations. The initial condition, at t_k, is $p(\boldsymbol{x}_k|\mathcal{Y}^k)$. We assume this density exists and is once continuously differentiable with respect to t and twice with respect to $\boldsymbol{x}$.

8.3.2 Updating Through Bayes' rule

Now, we also assume that the observation mapping $\boldsymbol{h}$, given in (8.26), is continuous in $\boldsymbol{x}$ and in t and bounded for each t_k w.p.1.

Integrating the Kolomogorov's equation yields $p(\boldsymbol{x}_{k+1}|\mathcal{Y}^k)$. When a new measurement is available at t_{k+1}, the distribution should be updated through Bayes' rule yielding $p(\boldsymbol{x}_{k+1}|\mathcal{Y}^{k+1})$. This is used in turn as the new initial condition for Kolomogorov's equation, ...

The updating equation is therefore

$$p(\boldsymbol{x}_k|\mathcal{Y}^k) = \frac{p(\boldsymbol{y}_k|\boldsymbol{x}_k)p(\boldsymbol{x}_k|\mathcal{Y}^{k-1})}{\int_{[\boldsymbol{x}_k]} p(\boldsymbol{y}_k|\boldsymbol{\xi})p(\boldsymbol{\xi}|\mathcal{Y}^{k-1})\, d\boldsymbol{\xi}} \tag{8.24}$$

Note that the denomitor is simply $p(\boldsymbol{y}_k|\mathcal{Y}^{k-1})$. Now, we have the analytical tool to solve the state filtering problem.

8.4 Exact state filtering

8.4.1 What is a State Filtering Problem?

Consider again

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta}, t)dt + \boldsymbol{G}(\boldsymbol{x}_t, \boldsymbol{\theta}, t)d\boldsymbol{\beta}_t \tag{8.25}$$

with the observation equation

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_k, \boldsymbol{u}_k, \boldsymbol{\theta}, t_k) + \boldsymbol{e}_k \tag{8.26}$$

$\boldsymbol{x}_t$ is the state of a physical system. We do not have access to the state, i.e. we can not directly measure it. However, we can directly measure a function of the state (contaminated with noise). One can raise three questions in connection with estimating the state of the system at time t based on measurements until time t_k. These questions are labeled

- Interpolation problem if $t < t_k$.
- Filtering problem if $t = t_k$.
- Prediction problem if $t > t_k$.

What we have achieved so far is that by defining a suitable integral (Itô integral), we presented a solution to the transition distribution of the state (Kolomogorov's equation). Then we used the observation equation and the Bayes' rule to update the distribution after each measurement. Thus, we have the whole posterior probability distribution of the state. We often like to reduce the distribution to some numbers that are much easier to work with. A very common approach is to look upon the first and second moments of the posterior distribution. In fact, it can be shown that choosing the posterior mean of the state as an estimate for the value of the state, we have minimized a Bayesian risk. More precisely, we have the following theorem:

Theorem 8.1 (Optimal State Filtering). *The estimate that minimizes* $\mathrm{E}[(\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t)^T \boldsymbol{W}(\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t)]$*, where* $\boldsymbol{W} \geq 0$ *is some positive definite weight matrix and* $\hat{\boldsymbol{x}}_t$ *a functional on* $\mathcal{Y}^k$ *for* $t \geq t_k$ *is the conditional mean.*

Proof. See [Jazwinski 1970]. ■

Let $\mathrm{E}^k[\cdot]$ denote the conditional mean of $[\cdot]$ given $\mathcal{Y}^{t_k}$. Also, let $\hat{\boldsymbol{x}}_{t|k}$ and $\boldsymbol{P}_{t|k}$ denote the conditional first and second moments given $\mathcal{Y}^{t_k}$, respectively. Then we have the theorem:

Theorem 8.2. *Assume the conditions required for the derivation of the conditional density in (8.22) and (8.24). Between observations, the conditional mean and covariance satisfy*

$$
\begin{aligned}
d\hat{\boldsymbol{x}}_{t|k}/dt &= \hat{\boldsymbol{f}}(\boldsymbol{x}_t, t) && (8.27)\\
d\boldsymbol{P}_{t|k}/dt &= \widehat{\boldsymbol{x}_t \boldsymbol{f}^T} - \hat{\boldsymbol{x}}_{t|k}\hat{\boldsymbol{f}}^T + \widehat{\boldsymbol{f}\,\boldsymbol{x}_t^T} && (8.28)\\
&- \hat{\boldsymbol{f}}\,\hat{\boldsymbol{x}}_{t|k}^T + \widehat{\boldsymbol{G}\boldsymbol{G}^T} && (8.29)
\end{aligned}
$$

for $t \in [t_k, t_{k+1}[$, where $\widehat{[\,\cdot\,]} = \mathrm{E}^k[\,\cdot\,]$. When a new observation arrives at t_k we have

$$
\begin{aligned}
\hat{\boldsymbol{x}}_{k|k} &= \frac{\mathrm{E}^{k-1}(\boldsymbol{x}_k p(\boldsymbol{y}_k|\boldsymbol{x}_k))}{\mathrm{E}^{k-1}(p(\boldsymbol{y}_k|\boldsymbol{x}_k))} && (8.30)\\
\boldsymbol{P}_{k|k} &= \frac{\mathrm{E}^{k-1}(\boldsymbol{x}_k \boldsymbol{x}_k^T p(\boldsymbol{y}_k|\boldsymbol{x}_k))}{\mathrm{E}^{k-1}(p(\boldsymbol{y}_k|\boldsymbol{x}_k))} - \hat{\boldsymbol{x}}_{k|k}\hat{\boldsymbol{x}}_{k|k}^T && (8.31)
\end{aligned}
$$

Predictions $\hat{\boldsymbol{x}}_{t|k}$ and $\boldsymbol{P}_{t|k}$, with $t > t_k$, based on $\mathcal{Y}^k$, also satisfy (8.27) and (8.29).

Proof. Is found e.g. in [Maybeck 1982]. ■

The theorem provides the complete solution to the state filtering problem. However, in order to obtain a computationally realizable filter, some approximations must be made. First, we restrict our attention to a special but at the same time very important class of models, namely, linear models.

Kalman Filter: The exact solution to the state filtering problem for linear dynamic models, i.e. models described by stochastic differential equations of the form

$$
\begin{aligned}
d\boldsymbol{x}_t &= \boldsymbol{A}(\boldsymbol{u}_t, \boldsymbol{\theta}, t)\,\boldsymbol{x}_t\,dt + \boldsymbol{B}(\boldsymbol{u}_t, \boldsymbol{\theta}, t)\,\boldsymbol{u}_t\,dt + \boldsymbol{G}(\boldsymbol{\theta}, t)\,d\boldsymbol{\beta}_t\\
\boldsymbol{y}_k &= \boldsymbol{C}(\boldsymbol{u}_k, \boldsymbol{\theta}, t_k)\,\boldsymbol{x}_k + \boldsymbol{D}(\boldsymbol{u}_k, \boldsymbol{\theta}, t_k)\,\boldsymbol{u}_k + \boldsymbol{e}_k
\end{aligned}
$$

is given by Kalman Filter.

The Kalman Filter is described by the following equations:

$$
\begin{aligned}
\hat{\boldsymbol{x}}_{k|k} &= \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k(\boldsymbol{y}_k - \hat{\boldsymbol{y}}_{k|k-1}) && (8.32)\\
\boldsymbol{P}_{k|k} &= \boldsymbol{P}_{k|k-1} - \boldsymbol{K}_k \boldsymbol{R}_{k|k-1} \boldsymbol{K}_k^T && (8.33)\\
\boldsymbol{K}_k &= \boldsymbol{P}_{k|k-1}\boldsymbol{C}^T \boldsymbol{R}_{k|k-1}^{-1} && (8.34)
\end{aligned}
$$

The formulas for prediction of mean and covariance of the state-vector and observations are given by,

$$\begin{aligned} d\hat{\boldsymbol{x}}_{t|k}/dt &= \boldsymbol{A}\,\hat{\boldsymbol{x}}_{t|k} + \boldsymbol{B}\,u_t, && (8.35)\\ d\boldsymbol{P}_{t|k}/dt &= \boldsymbol{A}\,\boldsymbol{P}_{t|k} + \boldsymbol{P}_{t|k}\boldsymbol{A}^T + \boldsymbol{G}\boldsymbol{G}^T, && (8.36)\\ t &\in [t_k, t_{k+1}[&& (8.37) \end{aligned}$$

$$\begin{aligned} \hat{\boldsymbol{y}}_{k+1|k} &= \boldsymbol{C}\,\hat{\boldsymbol{x}}_{k+1|k} + \boldsymbol{D}\,\boldsymbol{u}_{k+1} && (8.38)\\ \boldsymbol{R}_{k+1|k} &= \boldsymbol{C}\,\boldsymbol{P}_{k+1|k}\boldsymbol{C}^T + \boldsymbol{S} && (8.39) \end{aligned}$$

The initial conditions are $\hat{\boldsymbol{x}}_{1|0} = \mu_0$ and $P_{1|0} = V_0$. The dependencies of time and external input of the matrices in the Kalman filter equations have been suppressed for convenience.

8.5 Approximate filters

8.5.1 Linearized Kalman Filter

In the general non-linear case, a very common approach is based on linearization of the functions around a nominal trajectory. Thus, we will be able to use the Kalman filter derived for the linear model. The matrices are calculated by $\boldsymbol{A}(\boldsymbol{u}_t, \boldsymbol{\theta}, t) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\big|_{\boldsymbol{x}=\boldsymbol{x}^*}$, $\boldsymbol{B}(\boldsymbol{u}_t, \boldsymbol{\theta}, t) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}\big|_{\boldsymbol{x}=\boldsymbol{x}^*}$ etc., where $\boldsymbol{x}^*$ is some reference signal. There different ways to calculate the reference trajectory. One way is integrating the state equation after setting the noise equal to zero. This method, called the *linearized Kalman filter*, will only converge if the noise levels are sufficiently small.

8.6 Extended Kalman Filter (EKF)

A better choice for the linearization trajectory, is to use the current estimate of the state. Linearizing about it at every sampling time and applying a Kalman filter to the resulting linearized model yields the algorithm known as *extended Kalman filter*. Performance improvement for the extended Kalman filter may be obtained by local iterations (over a single sample period) on nominal trajectory redefinition and subsequent relinearization. If we iterate on the equations for the measurement update, by replacing equation (8.32) with the following iterator

$$\boldsymbol{\eta}_i = \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k(\boldsymbol{y}_k - \boldsymbol{h}(\boldsymbol{\eta}_{i-1}, t_k) - \boldsymbol{C}\,(\hat{\boldsymbol{x}}_{k|k-1} - \boldsymbol{\eta}_{i-1})) \qquad (8.40)$$

with $\boldsymbol{C} = (\partial \boldsymbol{h}/\partial \boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{\eta}_{i-1}}$, and $\boldsymbol{K}_k = \boldsymbol{K}_k(\boldsymbol{\eta}_{i-1})$, iterated for $i = 1, \cdots, l$, starting with $\boldsymbol{\eta}_0 = \hat{\boldsymbol{x}}_{k|k-1}$, and terminating with the result $\hat{\boldsymbol{x}}_{k|k} = \boldsymbol{\eta}_l$, we have the algorithm called *iterated extended Kalman filter* – see Figure 8.1.

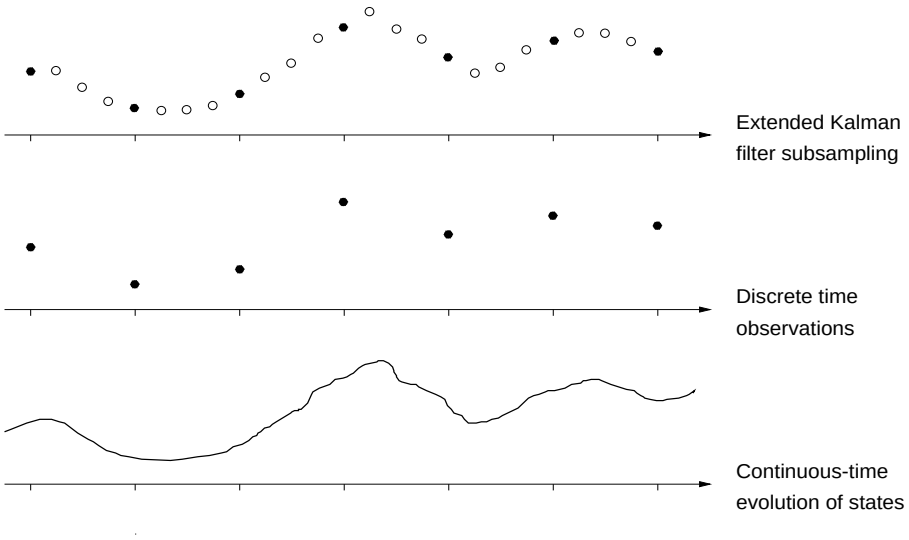

Figure 8.1: The continuous time and discrete time scales considered for the iterated extended Kalman filter

8.7 Statistically linearized filter

In this approach, we consider approximating $\boldsymbol{f}(\boldsymbol{x}_t, t)$ by a linear approximation of the form

$$\boldsymbol{f}(\boldsymbol{x}_t, t) = \boldsymbol{f}_0(t) + \boldsymbol{F}(t)\,\boldsymbol{x}_t + \boldsymbol{\epsilon}_t \tag{8.41}$$

which has the minimum mean square error

$$J = \mathrm{E}[\boldsymbol{\epsilon}_t^T \boldsymbol{W} \boldsymbol{\epsilon}_t | \mathcal{Y}^{k-1}] \tag{8.42}$$

for all $t \in [t_{k-1}, t_k[$, where $\boldsymbol{W} \geq 0$ is a weighting matrix. Calculating the partial derivatives of (8.42), with respect to $\boldsymbol{f}_0(t)$ and $\boldsymbol{F}(t)$ and setting them to zero yields, using the notation of Theorem 8.2

$$\boldsymbol{f}_0(t) = \hat{\boldsymbol{f}} - \boldsymbol{F}(t)\,\hat{\boldsymbol{x}}_t \tag{8.43}$$

$$\boldsymbol{F}(t) = (\widehat{\boldsymbol{f}\boldsymbol{x}_t^T} - \hat{\boldsymbol{f}}\hat{\boldsymbol{x}}_t^T)\,\boldsymbol{P}_{t|t}^{-1} \tag{8.44}$$

with $\boldsymbol{P}_{t|t}$ being the conditional covariance of $\boldsymbol{x}_t$.

Similarly for the measurement equation:

$$\boldsymbol{h}(\boldsymbol{x}_k, t_k) \cong \boldsymbol{h}_0(t_k) + \boldsymbol{H}(t_k)\,\boldsymbol{x}_k \tag{8.45}$$

with the coefficients statistically optimized to get

$$\boldsymbol{h}_0(t_k) = \hat{h} - \boldsymbol{H}(t_k)\,\hat{\boldsymbol{x}}_{k-1} \tag{8.46}$$

$$\boldsymbol{H}(t_k) = (\widehat{\boldsymbol{h}\boldsymbol{x}_k^T} - \hat{\boldsymbol{h}}\hat{\boldsymbol{x}}_{k-1}^T)\,\boldsymbol{P}_{k|k-1}^{-1} \tag{8.47}$$

The issue now is the computing of $\hat{\boldsymbol{f}}$, $\hat{\boldsymbol{h}}$, $\boldsymbol{F}(t)$ and $\boldsymbol{H}(t_k)$. They all depends upon the conditional probability density function of $\boldsymbol{x}$, which is generally not available. We therefore assume the density is Gaussian.

We obtain the *statistically linearized filter*, with the following equations for the time propagation

$$d\hat{\boldsymbol{x}}_{t|k}/dt = \hat{\boldsymbol{f}}(\boldsymbol{x}_{t|k}, t), \tag{8.48}$$

$$d\boldsymbol{P}_{t|k}/dt = \boldsymbol{F}(t)\boldsymbol{P}_{t|k} + \boldsymbol{P}_{t|k}\boldsymbol{F}^T(t) + \boldsymbol{G}\boldsymbol{G}^T \tag{8.49}$$

$$t \in [t_k, t_{k+1}[\tag{8.50}$$

with $\boldsymbol{F}(t)$ given by (8.44), and the conditional expectations involved calculated assuming $\boldsymbol{x}_t$ to be Gaussian with mean $\hat{\boldsymbol{x}}_{t|k}$ and covariance $\boldsymbol{P}_{t|k}$. The measurement update at time t_k is given by

$$\boldsymbol{K}_k = \boldsymbol{P}_{k|k-1}\boldsymbol{H}^T(t_k)[\boldsymbol{H}(t_k)\boldsymbol{P}_{k|k-1}\boldsymbol{H}^T(t_k) + \boldsymbol{S}]^{-1} \tag{8.51}$$

$$\hat{\boldsymbol{x}}_{k|k} = \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k(\boldsymbol{y}_k - \hat{\boldsymbol{h}}) \tag{8.52}$$

$$\boldsymbol{P}_{k|k} = \boldsymbol{P}_{k|k-1} - \boldsymbol{K}_k\boldsymbol{H}(t_k)\boldsymbol{P}_{k|k-1} \tag{8.53}$$

The conditional expectations are calculated as if $\boldsymbol{x}_k$ were Gaussian with mean $\hat{\boldsymbol{x}}_{k|k-1}$ and covariance $\boldsymbol{P}_{k|k-1}$. The equations for the gain and covariance are the same as those for the EKF, but with $\boldsymbol{F}(t)$ replacing $(\partial \boldsymbol{f}/\partial \boldsymbol{x})\big|_{\boldsymbol{x}=\hat{\boldsymbol{x}}_{t|k}}$ and $\boldsymbol{H}(t_k)$ replacing $(\partial \boldsymbol{h}/\partial \boldsymbol{x})\big|_{\boldsymbol{x}=\hat{\boldsymbol{x}}_{k|k-1}}$.

8.7.1 Suggestions for Further Reading

For a more detailed treatment of the filtering problem see [Maybeck 1982] and [Jazwinski 1970]. In [Åström 1970] the stochastic control problem is the main issue. The book is rather old, as is [Jazwinski 1970], but both of them are very well written and worth reading.

The extended Kalman filter has been used also as a tool for estimating unknown parameter in a state space representation, via extension of the state with the unknown parameters. This problem is in [Ljung 1979] shown to have a biased solution in most implementations, due to the neglected parameter dependence of the gain matrix in the Kalman filter. Two solutions to that problem are given in the same reference, the first is computationally demanding and simply introduces and treats the parameter dependence in the gain matrix. The second solution considers only the stationary

Kalman gain and extends the state also with this gain. It is then shown that if properly structured, the resulting algorithm is of prediction error type, giving e.g. consistent estimates.

The EKF approach to parameter estimation has furthermore been treated by Wiberg and DeWolf in a series of papers, using an averaging approach, see e.g.[DeWolf & Wiberg 1993, Wiberg & DeWolf 1993, Wiberg, Ljungkvist & Powell 1994]. They derive comparatively complicated but tractable recursive parameter estimation algoritms that are consistent also for estimation of the parameters in the noise covariance matrices. An essential extension that is done in their algorithms compared to standard extended Kalman filters, is the retention of also the third order moment in the parameter and covariance update, which essentially takes care of the at least transially nonsymmetric distribution of the parameter estimation errors.

In [Melgaard 1994], the application of filtering and prediction to estimation of physical parameters of a dynamic system is extensively investigated, and in [Hendricks 1992] parameter estimation in an engine model using physical modelling and extended Kalman methodology is extensively discussed.

Chapter 9

Modelling using stochastic differential equations

9.1 Introduction

Often non-linearities and non-stationarities are most conveniently modelled by considering continuous time stochastic models.

By that approach information from physics is more easily used. The modelling technique described in this chapter is called *grey box modelling*, since it provides good tools for combining information from physics with information from the data.

This chapter describes in detail a *maximum likelihood method* for estimation of linear stochastic differential equations in state space form. Also the estimation of a class of non-linear differential equations is considered. The state space formulation is adequate for modelling some non-stationary time series.

Furthermore the *maximum aposteriori estimation technique* (MAP), the *prediction error method* (PEM) (or *non-linear least squares*), and the *indirect prediction error method* (IPEM) are considered. Finally, some methods for filtering the residuals along with the estimation and some methods for making the estimation robust are described.

At the end of this chapter several cases of modelling physical systems using stochastic differential equatins are presented.

9.2 The continous time linear stochastic state space model

The deterministic part of a model considered is frequently simply a system of ordinary differential equations.

In matrix notation the equations can be concatenated in *the deterministic linear state space model in continuous time*

$$\frac{d\boldsymbol{X}}{dt} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{B}\boldsymbol{U} \tag{9.1}$$

where $\boldsymbol{X}$ is the state vector and $\boldsymbol{U}$ is the input vector.

The dynamical behavior of the system is characterized by the matrix $\boldsymbol{A}$, and $\boldsymbol{B}$ is a matrix, which specify how the input signals enter the system.

To describe the deviation between (9.1) and the true variation of the states an additive noise term is introduced.

Then we have the stochastic differential equation

$$d\boldsymbol{X} = \boldsymbol{A}\boldsymbol{X}dt + \boldsymbol{B}\boldsymbol{U}dt + d\boldsymbol{w}(t) \tag{9.2}$$

where the stochastic process $\boldsymbol{w}(t)$ is assumed to be a process with independent increments. (9.2) is *the stochastic linear state space model in continuous time*.

There are many reasons for introducing a noise term:

- *Modeling approximations*. For instance the dynamics, as described by the matrix $\boldsymbol{A}$ in (9.2) might be an approximation to the true system.
- *Unrecognized and unmodeled inputs*. Some variables, which are not considered, such as e.g. wind speed, may affect the system.
- *Measurements of the input are noise corrupted*. In this cases the measured input is regarded as the actual input to the system, and the deviation from the true input is described by $\boldsymbol{w}(t)$.

In the general case we assume that only a linear combination of the states is measured, and if we introduce $\boldsymbol{Y}$ to denote the measured or recorded variables we can write

$$\boldsymbol{Y}(t) = \boldsymbol{C}\boldsymbol{X}(t) + \boldsymbol{D}\boldsymbol{U}(t) + \boldsymbol{e}(t) \tag{9.3}$$

where $\boldsymbol{C}$ is a constant matrix, which specifies which linear combination of the states that actually are measured, and $\boldsymbol{D}$ is a constant matrix, which accounts for input variables which directly affect the output.

The equation is for obvious reasons called *the measurement equation*.

The term $\boldsymbol{e}(t)$ is the

- *Measurement error*. The sensors that measure the output signals are affected by noise and drift.

In the following it is assumed that $w(t)$ and $e(t)$ are mutually uncorrelated, which seems to be quite reasonable.

As an example of a model belonging to the class (9.2) consider the following model describing the heat dynamics for a building

$$\begin{aligned}\begin{bmatrix} dT_m \\ dT_i \end{bmatrix} &= \begin{bmatrix} \frac{-1}{r_i c_m} & \frac{1}{r_i c_m} \\ \frac{1}{r_i c_i} & -\left(\frac{1}{r_a c_i} + \frac{1}{r_i c_i}\right) \end{bmatrix} \begin{bmatrix} T_m \\ T_i \end{bmatrix} dt + \\ &\quad \begin{bmatrix} 0 & 0 & A_w p/c_m \\ 1/(r_a c_i) & 1/c_i & A_w(1-p)/c_i \end{bmatrix} \begin{bmatrix} T_a \\ \phi_h \\ \phi_s \end{bmatrix} dt + \begin{bmatrix} dw_m(t) \\ dw_i(t) \end{bmatrix} \quad (9.4) \\ \boldsymbol{T}_r(t) &= (0\ 1)\boldsymbol{T}(t) + \boldsymbol{e}(t) \quad (9.5)\end{aligned}$$

The states of the model are given by the temperature T_m of a large heat accumulating medium with the heat capacity c_m, and by the temperature T_i of the room air and possibly the inner part of the walls with the capacity c_i. The term r_i is the resistance against heat transfer between the room air and the large heat accumulating medium, while r_a is the resistance against heat transfer from the room air to the ambient air with the temperature T_a.

The input energy is supplied by the electrical heaters ϕ_h and the solar radiation which penetrates through the windows facing south $A_w\phi_s$, where A_w is the effective window area. The effective window area is the window area corrected for shade effects, and absorption and reflection by the triple glazed windows. Note that only the indoor air temperature is measured.

9.2.1 Identifiability of State Space Models

A good reference is the tutorial paper by **?**.

It is a crucial question whether the parameters of a specified state space model can be identified. If a non identifiable model is specified, the methods for estimation will not converge.

The identifiability concept has to do with the problem whether the identification procedure will yield a unique value of the parameters.

One aspect of the problem has to do with the experimental conditions, whether the dataset is informative enough (*persistent exciting*), to distinguish between different models.

The other aspect has to do with the model structure, i.e. if different sets of parameters will give equal models, given that the dataset is informative enough. The latter topic will be discussed here.

The problem of *structural identifiability* arises from the fact that for a given transfer function model corresponds, in general, a continuum of state space models. It is therefore necessary to put restrictions on the structure of the state space model in order to provide a unique relation between the parameters of the state space model and the transfer function.

Consider a continuous time state space model

$$\dot{\boldsymbol{X}}(t) = \boldsymbol{A}\boldsymbol{X}(t) + \boldsymbol{B}\boldsymbol{U}(t) + \boldsymbol{e}_1(t) \tag{9.6}$$

$$\boldsymbol{Y}(t) = \boldsymbol{C}\boldsymbol{X}(t) + \boldsymbol{D}\boldsymbol{U}(t) + \boldsymbol{e}_2(t) \tag{9.7}$$

where the state vector $\boldsymbol{X}(t)$ has dimension n, the input $\boldsymbol{U}(t)$ has dimension m and the output $\boldsymbol{Y}(t)$ has dimension s, $\boldsymbol{e}_1(t)$ and $\boldsymbol{e}_2(t)$ are mutually independent white noise processes with variances $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ of dimension n and s respectively. The part of the output, that cannot be predicted from past data is $\boldsymbol{\epsilon}(t) = \boldsymbol{Y}(t) - \hat{\boldsymbol{Y}}(t)$. This quantity, denoted by $\boldsymbol{\epsilon}(t)$ is called the *innovation*.

By rewriting (9.6) and (9.7) we obtain the *innovation form* of the state space description

$$\dot{\hat{\boldsymbol{X}}}(t) = \boldsymbol{A}\hat{\boldsymbol{X}}(t) + \boldsymbol{B}\boldsymbol{U}(t) + \boldsymbol{K}\boldsymbol{\epsilon}(t) \tag{9.8}$$

$$\boldsymbol{Y}(t) = \boldsymbol{C}\hat{\boldsymbol{X}}(t) + \boldsymbol{D}\boldsymbol{U}(t) + \boldsymbol{\epsilon}(t) \tag{9.9}$$

Introducing p as the differentiation operator the corresponding transfer function model has the form

$$\boldsymbol{Y}(t) = \boldsymbol{G}(p)\boldsymbol{U}(t) + \boldsymbol{H}(p)\boldsymbol{\epsilon}(t) \tag{9.10}$$

where $\boldsymbol{\epsilon}(t)$ is a white noise process of dimension s with variance $\boldsymbol{\Sigma}$.

The transfer functions are determined as

$$\begin{aligned} \boldsymbol{G}(p) &= \boldsymbol{C}(p\boldsymbol{I}-\boldsymbol{A})^{-1}\boldsymbol{B}+\boldsymbol{D} && (9.11)\\ \boldsymbol{H}(p) &= \boldsymbol{C}(p\boldsymbol{I}-\boldsymbol{A})^{-1}\boldsymbol{K}+\boldsymbol{I} && (9.12)\end{aligned}$$

where $\boldsymbol{K}$, the continuous time Kalman gain, is given as

$$\boldsymbol{K} = \boldsymbol{P}\boldsymbol{C}'\boldsymbol{\Sigma}_2^{-1} \tag{9.13}$$

and $\boldsymbol{P}$ is the solution to the steady state Riccati equation :

$$\boldsymbol{0} = \boldsymbol{A}\boldsymbol{P}+\boldsymbol{P}\boldsymbol{A}'+\boldsymbol{\Sigma}_1-\boldsymbol{P}\boldsymbol{C}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{C}\boldsymbol{P} \tag{9.14}$$

The variance of the noise process $\boldsymbol{\epsilon}(t)$ is given by

$$\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{P}\boldsymbol{C}'+\boldsymbol{\Sigma}_2 \tag{9.15}$$

From the observations of $\boldsymbol{U}(t)$ and $\boldsymbol{Y}(t)$ we are able to observe the transfer matrices $\boldsymbol{G}(p)$ and $\boldsymbol{H}(p)$ with elements of the form

$$\boldsymbol{G}_{ij}(p) = \frac{b_u p^u + b_{u-1}p^{u-1}+\cdots+b_0}{a_v p^v + a_{v-1}p^{v-1}+\cdots+a_0} \tag{9.16}$$

where $a_v = 1$.

By comparing the parameterization of (9.11) with elements of the form (9.16) it is possible to check if the parameters of a specified model is identifiable.

An example: The problem of identifiability is illustrated by an example. Consider a system (9.6) and (9.7) with matrices:

$$\begin{aligned} \boldsymbol{A} &= \begin{bmatrix} -a & a \\ b & -(b+c)\end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} 0 \\ d\end{bmatrix}, \quad \boldsymbol{C} = \begin{bmatrix} 0 & 1\end{bmatrix}, \\ \boldsymbol{D} &= \begin{bmatrix} 0\end{bmatrix} \end{aligned} \tag{9.17}$$

The transfer function $G(p)$ given as (9.11) now becomes

$$G(p) = \frac{d(p+a)}{p^2+(a+b+c)p+ac} \tag{9.18}$$

This should be compared to what we are actually able to observe, namely

$$G(p) = \frac{b_1 p + b_0}{p^2 + a_1 p + a_0} \tag{9.19}$$

That is, we are able to observe b_0, b_1, a_1 and a_2. If (9.18) and (9.19) are compared, it is seen that based upon these values, we are able to identify a, b, c and d. Hence the model is identifiable for the parameters of the matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$.

For the noise model $H(p)$ given by (9.12), there a more possibilities for parameterization. One possibility is a direct parameterization of the innovations form, i.e. directly estimation of the kalman gain

$$\boldsymbol{K} = \begin{bmatrix} k_1 \\ k_2 \end{bmatrix} \tag{9.20}$$

by the calculations in (9.12), this gives the noise model

$$H(p) = \frac{p^2 + (a+b+c+k_2)p + (k_1 b + k_2 a + ac)}{p^2 + (a+b+c)p + ac} \tag{9.21}$$

Comparing this model with the model of the kind (9.19) it is seen, that it is possible to find k_1, k_2 and the variance of the innovations $\boldsymbol{\Sigma}$, because $G(p)$ in (9.18) was identifiable. The total model including the noise model is in this case identifiable.

Another way to parameterize the noise model is by the matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. These matrices are connected to $\boldsymbol{K}$ and $\boldsymbol{\Sigma}$ through (9.13), (9.14) and (9.15) in a fairly complicated manner. Since the dimensions of $\boldsymbol{K}$ is $n \times s$ and $\boldsymbol{\Sigma}$ has $s(s+1)/2$ elements, where n is the order of the system and s is number of outputs from the system, $ns + s(s+1)/2$ is the maximum number of parameters which can be identified for $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. For the considered example we are able to identify 3 parameters for the noise model this could be done by estimating $\boldsymbol{\Sigma}_2$, which is 1 parameter and the diagonal of $\boldsymbol{\Sigma}_1$, which is 2 parameters and then fix $\Sigma_{1,12} = 0$. With this parameterization of the noise, the model is still identifiable.

Consider now the following slightly modified version of (9.17), namely the matrices

$$\boldsymbol{A} = \begin{bmatrix} -a & b \\ c & -d \end{bmatrix}, \qquad \boldsymbol{B} = \begin{bmatrix} 0 \\ e \end{bmatrix} \tag{9.22}$$

the transfer function corresponding to this model becomes

$$G(p) = \frac{e(p+a)}{p^2 + (a+d)p + (ad - cb)} \tag{9.23}$$

By comparison with (9.19) is is seen that c and b can not both be identified; but only the product.

9.3 A maximum likelihood method for parameter estimation in continous time stochastic state space models

9.3.1 The Model

It is assumed that the model of the dynamics is described by the stochastic differential equation

$$d\boldsymbol{X} = \boldsymbol{A}\boldsymbol{X}dt + \boldsymbol{B}\boldsymbol{U}dt + d\boldsymbol{w}(t) \tag{9.24}$$

where $\boldsymbol{X}$ is m dimensional, and the m dimensional stochastic process $\boldsymbol{w}(t)$ is assumed to be a process with independent increments.

With the purpose of calculating the likelihood function, $\boldsymbol{w}(t)$ is restricted to be a Wiener-process with the incremental covariance $\boldsymbol{R}_1^c(t)$.

The measured variables are written

$$\boldsymbol{Y}(t) = \boldsymbol{C}\boldsymbol{X}(t) + \boldsymbol{D}\boldsymbol{U}(t) + \boldsymbol{e}(t) \tag{9.25}$$

The term $\boldsymbol{e}(t)$ is the measurement error. It is assumed that $\boldsymbol{e}(t)$ is normally distributed white noise with zero mean and variance $\boldsymbol{R}_2$. Furthermore, it is assumed that $\boldsymbol{w}(t)$ and $\boldsymbol{e}(t)$ are mutually independent.

9.3.2 From Continuous to Discrete Time

Since it is assumed that the system is described by the stochastic differential equation (9.24), it is possible analytically to perform an integration, which under some assumptions exactly specifies the system evolution between discrete time instants.

The discrete time model corresponding to the continuous time model (9.24) is obtained by integrating the differential equation through the sample interval $[t, t+\tau)$.

Thus the sampled version of (9.24) can be written as

$$\boldsymbol{X}(t+\tau) = e^{\boldsymbol{A}(t+\tau-t)}\boldsymbol{X}(t) + \int_t^{t+\tau} e^{\boldsymbol{A}(t+\tau-s)}\boldsymbol{B}\boldsymbol{U}(s)ds + \int_t^{t+\tau} e^{\boldsymbol{A}(t+\tau-s)}d\boldsymbol{w}(s) \tag{9.26}$$

Under the assumption that $\boldsymbol{U}(t)$ is constant in the sample interval the sampled version

can be written as the following discrete time model in state space form

$$\boldsymbol{X}(t+\tau) = \boldsymbol{\phi}(\tau)\boldsymbol{X}(t) + \boldsymbol{\Gamma}(\tau)\boldsymbol{U}(t) + \boldsymbol{v}(t;\tau) \tag{9.27}$$

where

$$\boldsymbol{\phi}(\tau) = e^{\boldsymbol{A}\tau}\ ; \quad \boldsymbol{\Gamma}(\tau) = \int_0^\tau e^{\boldsymbol{A}s}\boldsymbol{B}\,ds\ ; \quad \boldsymbol{v}(t;\tau) = \int_t^{t+\tau} e^{\boldsymbol{A}(t+\tau-s)}\,d\boldsymbol{w}(s) \tag{9.28}$$

On the assumption that $\boldsymbol{w}(t)$ is a Wiener process, $\boldsymbol{v}(t;\tau)$ becomes normally distributed white noise with zero mean and covariance

$$\boldsymbol{R}_1(\tau) = E\left[\boldsymbol{v}(t;\tau)\boldsymbol{v}(t;\tau)^T\right] = \int_0^\tau \boldsymbol{\phi}(s)\boldsymbol{R}_1^c\boldsymbol{\phi}(s)^T ds \tag{9.29}$$

If the sampling time is constant (equally spaced observations), the stochastic difference equation can be written

$$\boldsymbol{X}(t+1) = \boldsymbol{\phi}\boldsymbol{X}(t) + \boldsymbol{\Gamma}\boldsymbol{U}(t) + \boldsymbol{v}(t) \tag{9.30}$$

where the time scale now is transformed such that the sampling time becomes equal to one time unit.

The problem of calculating the exponential of a matrix is a major task - see [Madsen & Melgaard 1991].

9.3.3 Maximum Likelihood Estimates

In the following it is assumed that the observations are obtained at regularly spaced time intervals, and hence that the time index t belongs to the set $\{0, 1, 2, ..., N\}$. N is the number of observations.

We further introduce

$$\mathcal{Y}(t) = [\boldsymbol{Y}(t), \boldsymbol{Y}(t-1), \ldots, \boldsymbol{Y}(1), \boldsymbol{Y}(0)]^T \tag{9.31}$$

i.e. $\mathcal{Y}(t)$ is a matrix containing all the observations up to and including time t. Finally, let $\boldsymbol{\theta}$ denote a vector of all the unknown parameters of the *embedded continous time model* in (9.24) – (9.25) - including the unknown variance and covariance parameters in $\boldsymbol{R}_1^c$ and $\boldsymbol{R}_2$.

The likelihood function is the joint probability density of all the observations assuming that the parameters are known, i.e.

$$\begin{aligned} L^T(\boldsymbol{\theta}; \mathcal{Y}(N)) &= p(\mathcal{Y}(N)|\boldsymbol{\theta}) \\ &= p(\boldsymbol{Y}(N)|\mathcal{Y}(N-1), \boldsymbol{\theta}) p(\mathcal{Y}(N-1)|\boldsymbol{\theta}) \\ &= \left(\prod_{t=1}^{N} p(\boldsymbol{Y}(t)|\mathcal{Y}(t-1), \boldsymbol{\theta}) \right) p(\boldsymbol{Y}(0)|\boldsymbol{\theta}) \end{aligned} \tag{9.32}$$

where successive applications of the rule $P(A \cap B) = P(A|B)P(B)$ is used to express the likelihood function as a product of conditional densities.

Since both $\boldsymbol{v}(t)$ and $\boldsymbol{e}(t)$ are normally distributed the conditional density is also normal. The normal distribution is completely characterized by the mean and the variance.

Hence, in order to parameterize the conditional distribution, we introduce the conditional mean and the conditional variance as

$$\hat{\boldsymbol{Y}}(t|t-1) = E[\boldsymbol{Y}(t)|\mathcal{Y}(t-1), \boldsymbol{\theta}] \qquad \text{and} \tag{9.33}$$

$$\boldsymbol{R}(t|t-1) = V[\boldsymbol{Y}(t)|\mathcal{Y}(t-1), \boldsymbol{\theta}] \tag{9.34}$$

respectively. It is noticed that (9.33) is the one-step prediction and (9.34) the associated variance. Furthermore, it is convenient to introduce the one-step prediction error (or innovation)

$$\boldsymbol{\epsilon}(t) = \boldsymbol{Y}(t) - \hat{\boldsymbol{Y}}(t|t-1) \tag{9.35}$$

Using (9.32) - (9.35) the conditional likelihood function (conditioned on $\boldsymbol{Y}(0)$) becomes

$$L(\boldsymbol{\theta}; \mathcal{Y}(N)) = \prod_{t=1}^{N} \left((2\pi)^{-m/2} \det \boldsymbol{R}(t|t-1)^{-1/2} \exp(-\tfrac{1}{2} \boldsymbol{\epsilon}(t)^T \boldsymbol{R}(t|t-1)^{-1} \boldsymbol{\epsilon}(t)) \right) \tag{9.36}$$

where m is the dimension of the $\boldsymbol{Y}$ vector. Most frequently the logarithm of the conditional likelihood function is considered. It is written

$$\log L(\boldsymbol{\theta}; \mathcal{Y}(N)) = -\tfrac{1}{2} \sum_{t=1}^{N} \left(\log \det \boldsymbol{R}(t|t-1) + \boldsymbol{\epsilon}(t)^T \boldsymbol{R}(t|t-1)^{-1} \boldsymbol{\epsilon}(t) \right) + \text{const} \tag{9.37}$$

The conditional mean $\hat{\boldsymbol{Y}}(t|t-1)$ and the conditional variance $\boldsymbol{R}(t|t-1)$ can be calculated recursively by using *a Kalman filter*. This is described in section 9.7.1.

The maximum-likelihood estimate (ML-estimate) is the set $\hat{\boldsymbol{\theta}}$, which maximizes the likelihood function. Since it is not possible analytically to optimize the likelihood

function, a numerical method has to be used. A reasonable method is the quasi-Newton method.

An estimate of the uncertainty of the parameters is obtained by the fact that the ML-estimator is asymptotically normally distributed with mean $\boldsymbol{\theta}$ and variance

$$\boldsymbol{D} = \boldsymbol{H}^{-1} \tag{9.38}$$

where the matrix $\boldsymbol{H}$ is given by

$$\{h_{lk}\} = -E\left[\frac{\partial^2}{\partial\theta_l\partial\theta_k}\log L(\boldsymbol{\theta};\mathcal{Y}(N))\right] \tag{9.39}$$

An estimate of $\boldsymbol{D}$ is obtained by equating the observed value with its expectation and applying

$$\{h_{lk}\} \approx -\left(\frac{\partial^2}{\partial\theta_l\partial\theta_k}\log L(\boldsymbol{\theta};\mathcal{Y}(N))\right)\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{9.40}$$

The above equation is thus used for estimating the variance of the parameter estimates. The variances serves as a basis for calculating t-test values for test under the hypotesis that the parameter is equal to zero. The correlation between the estimates is readily found based on the variance matrix.

9.4 Maximum aposteriori estimate

This method gives a possibility to include partial *prior knowledge about the parameters* in the estimation. The prior knowledge is expressed by some *prior probability density*, and the *Bayesian approach* is used for mixing the prior information with the information in the data.

Compared to classical estimation, the Bayesian approach gives a conceptually different treatment of the parameter estimation problem. In the Bayesian approach the parameter itself is considered as a random variable with the *prior probability density*

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}) \tag{9.41}$$

Then based on observations of other random variables, the data set $\boldsymbol{y}^N = [\boldsymbol{y}_N, \boldsymbol{y}_{N-1}, \cdots, \boldsymbol{y}_1, \boldsymbol{y}_0]$, which is correlated with the parameters, we may infer information about its value. Suppose the conditional probability density of the observations, given $\boldsymbol{\theta}$, (i.e. the *likelihood function*) is

$$p(\boldsymbol{y}^N|\boldsymbol{\theta}), \tag{9.42}$$

Then by combining the prior information with the sample information, using Bayes' theorem, it is possible to form the posterior information for $\boldsymbol{\theta}$, i.e. the conditional probability density given the observations

$$p(\boldsymbol{\theta}|\boldsymbol{y}^N) = \frac{p(\boldsymbol{y}^N|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y}^N)} \propto p(\boldsymbol{y}^N|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{9.43}$$

It is seen, that the *posterior probability density* is proportional to the likelihood function multiplied by the prior probability density.

From the posterior probability density, *different point estimates* of $\boldsymbol{\theta}$ can be determined, for instance the value where the probability density attains its maximum. This is called the *maximum a posteriori estimate (MAP)*, i.e.

$$\hat{\boldsymbol{\theta}}_{map} = \arg\max_{\theta\in\Omega} p(\boldsymbol{y}^N|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{9.44}$$

over an admissible parameter set Ω.

The maximum likelihood estimator (ML) selects the parameter that maximizes the likelihood of the data,

$$\hat{\boldsymbol{\theta}}_{ml} = \arg\max_{\theta\in\Psi} p(\boldsymbol{y}^N|\boldsymbol{\theta}) \tag{9.45}$$

If the prior distribution in (9.44) is uniform then the MAP estimator is equivalent to an ML estimator that is restricted to $\Psi = \Omega$. So the ML estimator corresponds to the MAP estimator when the prior information is diffuse or non-informative, with the constraint $\Psi = \Omega$.

Example 9.1 (MAP estimates). Assume that data can be described by:

$$\boldsymbol{y}_k = E(\boldsymbol{y}_k|\boldsymbol{y}^{k-1}, \boldsymbol{\theta}) + \boldsymbol{\epsilon}_k(\boldsymbol{\theta}) \tag{9.46}$$

where the sequence of innovations $\{\boldsymbol{\epsilon}_k(\boldsymbol{\theta})\}$ is a stochastic process with the property $E(\boldsymbol{\epsilon}_k(\boldsymbol{\theta})|\boldsymbol{y}^{k-1}) = \mathbf{0}$. If we assume the innovations are mutually independent and have the probability density function $p(\boldsymbol{\epsilon}_k(\boldsymbol{\theta})|\boldsymbol{\theta})$, we obtain the following likelihood function of the data:

$$p(\boldsymbol{y}^N|\boldsymbol{\theta}) = \prod_{k=1}^{N} p(\boldsymbol{\epsilon}_k(\boldsymbol{\theta})|\boldsymbol{\theta}) \tag{9.47}$$

If the innovations are assumed to be Gaussian with zero mean, and covariance $\boldsymbol{R}_k(\boldsymbol{\theta})$, we have

$$p(\boldsymbol{y}^N|\boldsymbol{\theta}) \propto \prod_{k=1}^{N} (\det \boldsymbol{R}_k(\boldsymbol{\theta}))^{-1/2} \exp(-\frac{1}{2}\boldsymbol{\epsilon}_k(\boldsymbol{\theta})^T \boldsymbol{R}_k^{-1}(\boldsymbol{\theta})\boldsymbol{\epsilon}_k(\boldsymbol{\theta})) \tag{9.48}$$

The maximizing argument of (9.48) with relation to $\boldsymbol{\theta}$ thus yields the ML estimate, $\hat{\boldsymbol{\theta}}_{ml}$. Let the prior distribution function for $\boldsymbol{\theta}$ be given by a normal distribution with mean $\boldsymbol{\mu} = \boldsymbol{\mu}_\theta$ and covariance $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_\theta$. Then

$$p(\boldsymbol{\theta}) \propto \exp(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)^T \boldsymbol{\Sigma}_\theta^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)) \tag{9.49}$$

Following (9.43), the conditional pdf for $\boldsymbol{\theta}$ given the data $\boldsymbol{y}^N$, that is the posterior pdf for $\boldsymbol{\theta}$, is given by

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}^N) \quad \propto \quad & \exp(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)^T \boldsymbol{\Sigma}_\theta^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)) \times \\ & \prod_{k=1}^{N} (\det \boldsymbol{R}_k(\boldsymbol{\theta}))^{-1/2} \exp(-\frac{1}{2}\boldsymbol{\epsilon}_k(\boldsymbol{\theta})^T \boldsymbol{R}_k^{-1}(\boldsymbol{\theta})\boldsymbol{\epsilon}_k(\boldsymbol{\theta})) \end{aligned} \tag{9.50}$$

The argument that maximizes (9.50) is the MAP estimate, $\hat{\boldsymbol{\theta}}_{map}$. The difference between ML and MAP is clearly seen from (9.48) and (9.50). As the prior standard deviation increases, $\boldsymbol{\Sigma}_\theta \to \infty$, the prior pdf approaches a uniform and non-informative distribution, and the MAP estimate becomes equivalent with the ML estimate. ♦

The advantage of considering MAP estimators instead of ML estimators is of course that our prior information about the parameters of the model is incorporated in the final estimate of the parameters. We may then be able to reach the same level of confidence of the final estimates, from a shorter experiment. Or *non-identifiable models* may become identifiable, due to the external information.

The problematic part of the MAP estimation approach, is the specification of the prior pdf. The prior information can either be generated from samples of past data, e.g. from ML estimation, in which case the asymptotic distribution of the estimates is known. The prior information can also be specified from introspection, or theoretical considerations, and in this case the prior information may differ for individual persons, because they have different beliefs. One should also consider the sensibility of the final estimate from the specification of a wrong prior pdf. Often the prior pdf represents both kinds of prior information, data-based and non data-based.

The Bayesian approach is quite appealing in relation to *grey box identification*, where the idea is to utilize the prior information about the system, in the identification. The MAP estimator is also called the grey box estimator by Tulleken (1993).

It can be shown that asymptotically (when the number of data observations $N \to \infty$) the MAP estimator converge with probability one to the conventional ML estimator, provided the admissible model set contains the the true parameters. That is, the MAP estimator is asymptotically unbiased and efficient. This is a direct consequence of the

fact that the likelihood will asymptotically become Gaussian distributed, around the true parameters, with vanishing covariance matrix. This also means that the influence from the prior distribution is vanishing with increasing number of observations in the likelihood function, see [Tulleken 1993].

9.5 The prediction error method

The prediction error methods described in Section 5.5 is formulated in such a way that the description is valid also for estimation of parameters in stochastic differential equations using discrete time observations.

Hence, we consider the lossfunction – or the norm

$$V_N(\boldsymbol{\theta}) = \sum_1^N \left(x_j - \mathrm{E}_\theta(x_j|\mathcal{B}_{j-1})\right)^2. \tag{9.51}$$

and the *PEM estimate* is given as

$$\hat{\boldsymbol{\theta}}_N = \arg\min_{\theta \in D_M} V_N(\boldsymbol{\theta}) \tag{9.52}$$

That is, the basic PEM estimator minimizes the sum of squared one-step prediction errors, given by

$$\begin{aligned} \boldsymbol{\epsilon}(t_k, \boldsymbol{\theta}) &= \boldsymbol{y}_k - \hat{\boldsymbol{y}}_k & (9.53)\\ \hat{\boldsymbol{y}}_k &= \boldsymbol{g}(t_k, \boldsymbol{\theta}, \boldsymbol{y}^{k-1}, \boldsymbol{u}^{k-1}) & (9.54) \end{aligned}$$

where $\boldsymbol{g}(t_k, \boldsymbol{\theta}, \boldsymbol{y}^{k-1}, \boldsymbol{u}^{k-1})$ is a deterministic function of old data and the parameters.

9.5.1 Filtering the prediction errors

The approach may be even further specialized as described in the following. The idea is to choose some norm that measures the size of $\boldsymbol{\epsilon}$, and then find the parameter vector $\hat{\boldsymbol{\theta}}_N$ that minimizes this norm.

Let the prediction error sequence be filtered through a stable linear filter $\boldsymbol{L}(q)$,

$$\boldsymbol{\epsilon}_f(t_k, \boldsymbol{\theta}) = \boldsymbol{L}(q)\boldsymbol{\epsilon}(t_k, \boldsymbol{\theta}) \tag{9.55}$$

for all $k \in \{1, 2, \cdots, N\}$. Then use the following norm,

$$V(\boldsymbol{\theta}, \boldsymbol{z}^N) = \frac{1}{N}\sum_{k=1}^N l(t_k, \boldsymbol{\theta}, \boldsymbol{\epsilon}_f(t_k, \boldsymbol{\theta})) \tag{9.56}$$

where $l(\cdot)$ is a scalar valued function. The estimate $\hat{\boldsymbol{\theta}}_N$ is chosen as the minimizing value of (9.56), i.e.

$$\hat{\boldsymbol{\theta}}_N = \arg\min_{\theta \in D_M} V(\boldsymbol{\theta}) \tag{9.57}$$

Following Ljung (1987) methods that corresponds to the approach of (9.57) are called *prediction error methods*. This approach contains as special cases a number of known methods, like the *least squares* and *maximum likelihood* methods. The different methods apply by specific choices of the prefilter $\boldsymbol{L}(q)$ and the norm $l(\cdot)$.

The effect of the filter $\boldsymbol{L}$ is easy to understand in a frequency domain interpretation of the criterion (9.57). $\boldsymbol{L}$ acts like a frequency weighting of the criterion, i.e. if $\boldsymbol{L}$ is a low pass filter then the criterion will only be little affected by high frequency disturbances. By another choice of the prefilter it is possible to remove slow drift terms in the data. It should be noted that the inclusion of the filter is to allow extra freedom in dealing with the properties of the prediction errors. The filter can be considered as a part of the model, by changing the predictor. If we consider a model of the form

$$\boldsymbol{y}_k = \boldsymbol{G}(q, \boldsymbol{\theta})\boldsymbol{u}_k + \boldsymbol{H}(q, \boldsymbol{\theta})\boldsymbol{e}_k \tag{9.58}$$

then the effect of prefiltering the prediction errors according to (9.55) is identical to changing the noise model from $\boldsymbol{H}(q, \boldsymbol{\theta})$ to

$$\boldsymbol{H}_L(q, \boldsymbol{\theta}) = \boldsymbol{L}^{-1}(q)\boldsymbol{H}(q, \boldsymbol{\theta}) \tag{9.59}$$

see [Ljung 1987].

9.6 An indirect prediction error method

The so-called *indirect prediction error method (IPEM)*, suggested by Söderström, Stoica & Friedlander (1991), is in fact a two-step estimation method. It is applicable to situations where the parameters of a model of interest (M_1) are functions of parameters in a larger model (M_2) which are relatively easy to estimate. In the first step the parameters of the model M_2 is estimated, and in the second step estimates of M_1 are found using the functional relation between the set of parameters:

$$\text{measured data} \overset{(PEM)}{\longrightarrow} \text{parameters in } M_2 \longrightarrow \text{parameters in } M_1$$

One typical application of the IPEM method is the situation where M_1 corresponds to the parameters of a stochastic differential equation and M_2 is the parameters of a

discrete time counterpart, i.e. the corresponding difference equation. An example of this application is shown in [Händel 1994].

Another well-known example is where M_2 corresponds to the parameters of a high order autoregressive model and M_1 to an ARMA model. This is for instance considered by [Wahlberg 1987] and [Wahlberg 1989].

Let θ denote the parameters corresponding to M_1 and α the parameters corresponding to M_2. The PEM estimates are given by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{t=1}^{N} \boldsymbol{\epsilon}_1^T(t,\boldsymbol{\theta})\boldsymbol{\epsilon}_1(t,\boldsymbol{\theta}) \tag{9.60}$$

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \sum_{t=1}^{N} \epsilon_2^T(t,\boldsymbol{\alpha})\epsilon_2(t,\boldsymbol{\alpha}) \tag{9.61}$$

where $\boldsymbol{\epsilon}_1(t,\boldsymbol{\theta})$ is the prediction error

$$\boldsymbol{\epsilon}_1(t,\boldsymbol{\theta}) = \boldsymbol{y}(t) - \hat{\boldsymbol{y}}(t \mid 1;\boldsymbol{\theta})$$

in the model structure M_1 and similarly for $\boldsymbol{\epsilon}_2(t,\alpha)$.

We will impose the following two assumptions on the model structures M_1 and M_2 and the data.

A1. The structures are nested

$$M_1 \subset M_2 \tag{9.62}$$

so that there exists a smooth map

$$\boldsymbol{\alpha}(\boldsymbol{\theta}) : S_1 \to S_2; S_1 \subset \mathbb{R}^{\dim \boldsymbol{\theta}}, S_2 \subset \mathbb{R}^{\dim \boldsymbol{\alpha}}$$

such that

$$\boldsymbol{\epsilon}_2(t,\boldsymbol{\alpha}(\boldsymbol{\theta})) = \boldsymbol{\epsilon}_1(t,\boldsymbol{\theta}). \tag{9.63}$$

Furthermore, $\dim \boldsymbol{\theta} \leq \dim \boldsymbol{\alpha}$ and S_1 is an open (nonzero measure) set in $\mathbb{R}^{\dim \boldsymbol{\theta}}$ such that $\boldsymbol{\epsilon}_1(t,\boldsymbol{\theta})$ is a stationary process for any $\boldsymbol{\theta} \in S_1$. We assume that $\partial\boldsymbol{\alpha}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ has full rank over S_1.

A2. Both structures give parameter identifiability. This means that there exist *unique*, "true", parameter vectors $\boldsymbol{\alpha}^\star$ and $\boldsymbol{\theta}^\star$ such that

$$\boldsymbol{\epsilon}_1(t,\boldsymbol{\theta}^\star) = \epsilon_2(t,\boldsymbol{\alpha}^\star) \text{ is white noise of zero mean and variance } \lambda^2. \tag{9.64}$$

Note that (9.64) and assumption A1 imply

$$\boldsymbol{\alpha}^\star = \boldsymbol{\alpha}(\boldsymbol{\theta}^\star). \tag{9.65}$$

Assumption A1 is solely a condition on M_1 and M_2, while A2 involves also the properties the data ("the system"). Assumption A2 may seem strong. The requirement that the quantity in (9.64) is white noise is neither needed for derivation of the estimator nor the interpretations. It is needed only for the ML interpretation used in the analysis of asymptotic properties in [Söderström et al. 1991].

The algorithm : Assume PEM estimates $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ and an estimate of the corresponding covariance matrix, $\hat{\boldsymbol{P}}_{\boldsymbol{\alpha}}$, are given. Assume furthermore that the relation between $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ is given by

$$\alpha(\theta) \tag{9.66}$$

and that the Jacobian $\boldsymbol{J} = \partial\boldsymbol{\alpha}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ has full rank over S_1.

Given the initial estimates (guess) $\hat{\boldsymbol{\theta}}_i$ the iterative equations are

$$\begin{aligned}
\boldsymbol{J}(i) &= \left(\partial\boldsymbol{\alpha}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_i}\right)^T && (9.67)\\
\hat{\boldsymbol{P}}_{\boldsymbol{\theta}}(i) &= (\boldsymbol{J}(i)\hat{\boldsymbol{P}}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{J}^T(i))^{-1} && (9.68)\\
\hat{\boldsymbol{\theta}}_{i+1} &= \hat{\boldsymbol{\theta}}_i + \hat{\boldsymbol{P}}_{\boldsymbol{\theta}_i}\boldsymbol{J}(i)\hat{\boldsymbol{P}}_{\boldsymbol{\alpha}}^{-1}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}(\hat{\boldsymbol{\theta}}(i))) && (9.69)
\end{aligned}$$

A stop criterion is

$$||\hat{\boldsymbol{\theta}}_{i+1} - \hat{\boldsymbol{\theta}}_i|| < \delta$$

where δ is some small number. Due to the iterative structure the algorithm may be used for on-line implementations together with a recursive PEM.

An example of using the algorithm is given in [Händel 1994].

9.7 Robust norms

From the previous sections it has been shown that for minimizing the variance of the estimated parameters the optimal choice of the scalar norm l for the general PEM criterion (9.56) is

$$l(\epsilon_t(\theta)) = -\log p(\epsilon_t(\theta)) \tag{9.70}$$

where $p(\epsilon_t(\theta))$ is the probability density function of the innovations (the ML criterion). The problem is that $p(\epsilon_t(\theta))$ may not be known for the true innovations. If we assume

that the density is Gaussian, but the true density has thicker tails, this may have a large influence on the estimation, cf. [Martin & Yohai 1985]. It is convenient to consider the innovations as being generated from an outliers model, with the following distribution of the innovations

$$P(\epsilon_t) = (1-\gamma)N(0,\sigma^2) + \gamma G\ , \qquad 0 < \gamma < 1 \tag{9.71}$$

where G is some symmetric outlier generating distribution and γ is a small fraction.

In order to evaluate the influence of the outliers on the estimates when using a given norm l, it is valuable to rewrite the expression for the covariance matrix of the parameters (**??**). Under the assumption that $\mathcal{S} \in \mathcal{M}$ and that the limiting parameters equals the true ones, $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$, the expression for the covariance of the parameters can be written,

$$\boldsymbol{P}_\theta = \kappa(l)\,(\bar{W}''(\boldsymbol{\theta}_0))^{-1} \tag{9.72}$$

where the covariance of the parameters is scaled by the scalar

$$\kappa(l) = \mathrm{E}_\epsilon[\psi^2(\epsilon_t(\boldsymbol{\theta}_0))]\ /\ (\mathrm{E}_\epsilon[\psi'_\epsilon(\epsilon_t(\boldsymbol{\theta}_0))])^2 \tag{9.73}$$

where $\psi(\epsilon_t) = l'_\epsilon(\epsilon_t)$ (derivation is with respect to ϵ) and expectation should be taken with respect to the true distribution of the innovations. Consider for simplicity the scalar case of Gaussian ML with the given variance 1, then (9.70) is $l(\epsilon_t) = \frac{1}{2}\epsilon_t^2$ (neglecting the constant term). For the ideal case where the true distribution of the innovations is Gaussian we have $\kappa(l) = 1$. If on the other hand the true distribution is given by (9.71) with a $\gamma > 0$ this scalar can become much larger than 1, cf. [Ljung 1987, pp. 397–398]. This means, that even if the fraction of outliers is very small the variance of the parameters can be much deteriorated when using the norm (9.70). Thus, the norm (9.70) is very sensitive to the true distribution of the innovations. This is of course not desirable because the true distribution is usually unknown. In order to make the criterion more *robust* to the unknown variations of the distribution of the innovations, it is suggested to modify (9.70) in such a way that $\psi(x)$ in (9.73) behaves like x for small values of x, then saturates, and even tend to zero as x increases (redescending). One choice of a ψ-function, is called *Huber's ψ–function*

$$\psi_H(x) = \begin{cases} x & |x| \le c_H \\ c_H\ \mathrm{sgn}(x) & |x| > c_H \end{cases} \tag{9.74}$$

If one wants protection against extremely heavy-tailed distributions, a redescending psi-function can be used. A frequently used type of redescending function is *Tukey's bisquare* function

$$\psi_B(x) = \begin{cases} x\,(1 - x^2/c_B^2)^2 & |x| \le c_B \\ 0 & |x| > c_B \end{cases} \tag{9.75}$$

see [Martin & Yohai 1985] for a discussion of the psi-functions and robustness.

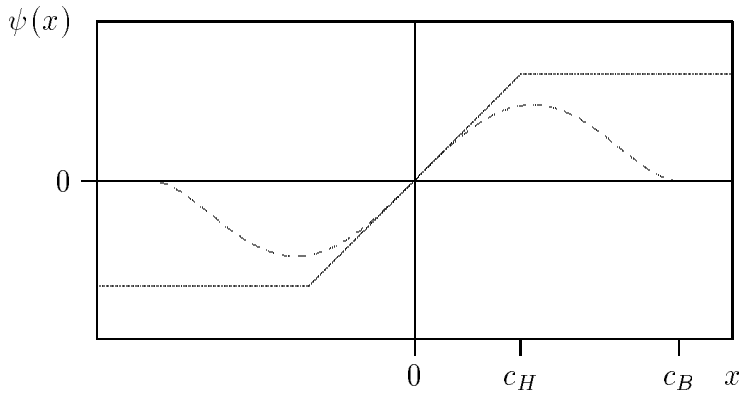

Figure 9.1: Huber's psi-function and Tukey's bisquare function.

When off-line methods are of greatest interest we have the possibility of repeating the estimation on the data set. Hence, we are able to detect and handle outliers in the data set, e.g. in connection with the model validation. Thus for off-line methods it may not be necessary to be protected against very heavy-tailed distributions, but it is still useful to modify the estimation criterion according to e.g. Huber's psi-function and with a c_H value that is not too small, e.g. $c_H = 3\,\hat{\sigma}$. When such a robust norm is used there will only be a small increase of $\kappa(l)$ when outliers are present and also when the true distribution is Gaussian (no outliers) there is only a small loss of optimality compared to the original norm (9.70). The price of a small increase of the variance for the nominal case is thus worth paying to have robustness against small variations in the true distribution of the innovations.

9.7.1 Kalman Filter

For linear models the Kalman filter provides the exact solution for the filtering problem discussed previously. The equations for the Kalman filter are given in a previous chapter but repeated below. The routine calculates the conditional mean $\hat{\boldsymbol{y}}_{k|k-1}$ and the conditional variance $\boldsymbol{R}_{k|k-1}$ to be used for the estimation. The Kalman filter performs the optimal (minimum variance) linear updating and prediction of the state variables. Finally, some numerical aspects regarding the implementation are discussed.

If the model is time invariant the model can be discretized before filtering. If the time dependency is weak compared to the dominating eigenvalues of the system, this implementation of the Kalman filter may also be used for time varying systems, by discretizing the continuous model at each sampling instant, assuming that $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{G}$ are constant within the sampling interval.

For the discrete time state space model the equations for *updating* the estimate of the state $\boldsymbol{x}$ becomes (with a slightly changed notation)

$$\hat{\boldsymbol{x}}_{k|k} = \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k(\boldsymbol{y}_k - \hat{\boldsymbol{y}}_{k|k-1}) \tag{9.76}$$

$$\boldsymbol{P}_{k|k} = \boldsymbol{P}_{k|k-1} - \boldsymbol{K}_k\boldsymbol{R}_{k|k-1}\boldsymbol{K}_k^T \tag{9.77}$$

where $\boldsymbol{K}_k$ is given by

$$\boldsymbol{K}_k = \boldsymbol{P}_{k|k-1}\boldsymbol{C}^T\boldsymbol{R}_{k|k-1}^{-1} \tag{9.78}$$

The formulas for *prediction* becomes

$$\hat{\boldsymbol{x}}_{k+1|k} = \boldsymbol{\Phi}\hat{\boldsymbol{x}}_{k|k} + \boldsymbol{\Gamma}\boldsymbol{u}_k \tag{9.79}$$

$$\hat{\boldsymbol{y}}_{k+1|k} = \boldsymbol{C}\hat{\boldsymbol{x}}_{k+1|k} + \boldsymbol{D}\boldsymbol{u}_{k+1} \tag{9.80}$$

$$\boldsymbol{P}_{k+1|k} = \boldsymbol{\Phi}\boldsymbol{P}_{k|k}\boldsymbol{\Phi}^T + \boldsymbol{\Lambda} \tag{9.81}$$

$$\boldsymbol{R}_{k+1|k} = \boldsymbol{C}\boldsymbol{P}_{k+1|k}\boldsymbol{C}^T + \boldsymbol{S} \tag{9.82}$$

The formulas require some initial values, which describes the prior knowledge about the states of the system in terms of the prior mean and variance $\hat{\boldsymbol{x}}_{1|0} = \boldsymbol{\mu}_0$ and $\boldsymbol{P}_{1|0} = \boldsymbol{V}_0$. The matrix $\boldsymbol{P}_{k+1|k}$ is the variance of the one-step prediction of the state, $\boldsymbol{x}$, of the system.

Numerical Aspects It is well known that the Kalman filter in some situations is numerically unstable. The problems arise when some of the variances, because of rounding errors, become non-positive definite. Therefore careful handling of the equations for the variances (9.77), (9.78), (9.81) and (9.82) is needed in order to numerically stabilize the Kalman filter. Since all variances should be symmetric and positive definite, it is desirable to use their Cholesky factorization. One possibility is to use

the $\boldsymbol{LDL}^T$-factorization, i.e. the square root free Cholesky decomposition, where $\boldsymbol{L}$ is unit lower matrix and $\boldsymbol{D}$ is diagonal.

An equation for updating a factorized matrix is

$$\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{G}\boldsymbol{D}_g\boldsymbol{G}^T \tag{9.83}$$

where $\tilde{\boldsymbol{A}}$ is known from other considerations to be positive definite, and $\boldsymbol{D}_g$ is a diagonal matrix. Thus, it is necessary to compute a unit lower triangular matrix $\tilde{\boldsymbol{L}}$ and a diagonal matrix $\tilde{\boldsymbol{D}}$ with $\tilde{d}_i > 0$ such that

$$\tilde{\boldsymbol{A}} = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{L}}^T \tag{9.84}$$

Algorithms to solve this problem are outlined in [Madsen & Melgaard 1991]. It is obvious that equation (9.81) and (9.82) easily are brought into this form. Equation (9.77) can be rewritten as

$$\begin{aligned}
\boldsymbol{P}_{k|k} &= \boldsymbol{P}_{k|k-1} - \boldsymbol{K}_k\boldsymbol{R}_{k|k-1}\boldsymbol{K}_k^T \quad \Leftrightarrow \\
\boldsymbol{P}_{k|k} &= \boldsymbol{P}_{k|k-1} - \boldsymbol{P}_{k|k-1}\boldsymbol{C}^T\boldsymbol{R}_{k|k-1}^{-1}\boldsymbol{C}\boldsymbol{P}_{k|k-1}^T \quad \Leftrightarrow \\
\tilde{\boldsymbol{L}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{L}}^T &= \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T - \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T\boldsymbol{C}^T(\boldsymbol{L}_r\boldsymbol{D}_r\boldsymbol{L}_r^T)^{-1}\boldsymbol{C}(\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T)^T \quad \Leftrightarrow \\
\tilde{\boldsymbol{L}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{L}}^T &= \boldsymbol{L}(\boldsymbol{D} - \boldsymbol{G}\boldsymbol{D}_r^{-1}\boldsymbol{G}^T)\boldsymbol{L}^T
\end{aligned} \tag{9.85}$$

where

$$\boldsymbol{G} = \boldsymbol{D}\boldsymbol{L}^T\boldsymbol{C}^T\boldsymbol{L}_r^{-T} \tag{9.86}$$

The expression $(\boldsymbol{D} - \boldsymbol{G}\boldsymbol{D}_r^{-1}\boldsymbol{G}^T)$ in (9.85) are in the form (9.83), and can thus be solved for the factors $\bar{\boldsymbol{L}}\bar{\boldsymbol{D}}\bar{\boldsymbol{L}}^T$, and we have $\tilde{\boldsymbol{L}} = \boldsymbol{L}\bar{\boldsymbol{L}}$ and $\tilde{\boldsymbol{D}} = \bar{\boldsymbol{D}}$.

This implementation of the Kalman filter is able to handle the multiple-input, multiple-output case with a high degree of accuracy and stability, see e.g. [Bierman 1977].

9.8 Estimation of a class of non–linear models

9.8.1 A Class of Non-Linear Models

Now let the model be described by the stochastic differential equation

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta}, t)dt + \boldsymbol{G}(\boldsymbol{\theta}, t)d\boldsymbol{\beta}_t \tag{9.87}$$

with $\boldsymbol{\beta}$ being a standard Wiener process. The observations $\boldsymbol{y}_k$ are taken at discrete time instants, t_k

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_k, \boldsymbol{u}_k, \boldsymbol{\theta}, t_k) + \boldsymbol{e}_k \tag{9.88}$$

where $\boldsymbol{e}$ is a Gaussian white noise process independent of $\boldsymbol{\beta}$, and $\boldsymbol{e}_k \sim N(\boldsymbol{0}, \boldsymbol{S}(\boldsymbol{\theta}, t_k))$.

9.8.2 Extended Kalman Filter

In this case the extended Kalman filter is used as a first order approximative filter. Being linearized about $\hat{\boldsymbol{x}}_t$ the state and covariance propagation equations have a structure similar to the Kalman filter propagation equations for linear systems. Hence, we are able to reuse the numerical stable routines implemented for the Kalman filter from the previous section.

The necessary modifications of the equations in the previous section are the following. The matrix $\boldsymbol{C}$ is the linearization of the measurement equation,

$$\boldsymbol{C}(\hat{\boldsymbol{x}}_{k|k-1}, \boldsymbol{u}_k, \boldsymbol{\theta}, t_k) = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}\Big|_{x=\hat{x}_{k|k-1}}, \tag{9.89}$$

and $\boldsymbol{A}$ is the linearization of the system equation,

$$\boldsymbol{A}(\hat{\boldsymbol{x}}_t, \boldsymbol{u}_t, \boldsymbol{\theta}, t) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\Big|_{x=\hat{x}_t}, \tag{9.90}$$

and $\boldsymbol{\Phi}$ is the discrete system matrix calculated as a transformation of $\boldsymbol{A}$, see the following section. The prediction of the output, Equation 9.80, is replaced by

$$\hat{\boldsymbol{y}}_{k+1|k} = \boldsymbol{h}(\hat{\boldsymbol{x}}_{k+1|k}, \boldsymbol{u}_{k+1}, \boldsymbol{\theta}, t_{k+1}) \tag{9.91}$$

The formulas for prediction of mean and covariance of the state-vector are normally given by,

$$\begin{aligned} d\hat{\boldsymbol{x}}_{t|k}/dt &= \boldsymbol{f}(\hat{\boldsymbol{x}}_{t|k}, \boldsymbol{u}_t, \boldsymbol{\theta}, t)) \ , \qquad t \in [t_k, t_{k+1}[& (9.92)\\ d\boldsymbol{P}_{t|k}/dt &= \boldsymbol{A}(\hat{\boldsymbol{x}}_{t|k}, \boldsymbol{u}_t, \boldsymbol{\theta}, t)\, \boldsymbol{P}_{t|k} + \boldsymbol{P}_{t|k}\boldsymbol{A}^T(\hat{\boldsymbol{x}}_{t|k}, \boldsymbol{u}_t, \boldsymbol{\theta}, t) \\ &\quad + \boldsymbol{G}(\boldsymbol{\theta}, t)\boldsymbol{G}^T(\boldsymbol{\theta}, t) \ , \qquad t \in [t_k, t_{k+1}[& (9.93) \end{aligned}$$

where $\boldsymbol{A}$ is given by (9.90).

In order to make the integration of (9.92) and (9.93) computational feasible and numerically stable for stiff systems, the time interval $[t_k, t_{k+1}[$ is sub sampled and the equations are linearized about the state estimate at the given sub sampling time. For the state propagation the equation becomes

$$\begin{aligned} d\hat{\boldsymbol{x}}_t/dt &= \boldsymbol{f}(\hat{\boldsymbol{x}}_j) + \boldsymbol{A}(\hat{\boldsymbol{x}}_j)\{\hat{\boldsymbol{x}}_t - \hat{\boldsymbol{x}}_j\} \ , \qquad t \in [t_j, t_{j+1}[& (9.94)\\ &= \boldsymbol{A}(\hat{\boldsymbol{x}}_j)\, \hat{\boldsymbol{x}}_t + \{\boldsymbol{f}(\hat{\boldsymbol{x}}_j) - \boldsymbol{A}(\hat{\boldsymbol{x}}_j)\, \hat{\boldsymbol{x}}_j\} \ , \ t \in [t_j, t_{j+1}[& (9.95) \end{aligned}$$

where $[t_j, t_{j+1}[$ is one of the sub intervals of the sampling interval $[t_k, t_{k+1}[$, assuming the sampling interval has been divided in n_s sub intervals. In these derivations only the state dependency is mentioned for clarity. Equation 9.95 is a linear ordinary differential equation which has the exact solution

$$\begin{aligned} \hat{\boldsymbol{x}}_{j+1} &= \hat{\boldsymbol{x}}_j + (e^{\boldsymbol{A}(\hat{\boldsymbol{x}}_j)\tau_s} - \boldsymbol{I})(\boldsymbol{A}(\hat{\boldsymbol{x}}_j)^{-1}\boldsymbol{f}(\hat{\boldsymbol{x}}_j)) & (9.96)\\ &= \hat{\boldsymbol{x}}_j + (\boldsymbol{\Phi}_s(\hat{\boldsymbol{x}}_j) - \boldsymbol{I})(\boldsymbol{A}(\hat{\boldsymbol{x}}_j)^{-1}\boldsymbol{f}(\hat{\boldsymbol{x}}_j)) & (9.97) \end{aligned}$$

where $\tau_s = t_{j+1} - t_j = \tau/n_s$, and τ is the sampling time. Efficient algorithms to calculate the matrix exponential in (9.96) is described in the next section. Correspondingly the equation for the state covariance becomes

$$\boldsymbol{P}_{j+1} = \boldsymbol{\Phi}_s(\hat{\boldsymbol{x}}_j)\boldsymbol{P}_j\boldsymbol{\Phi}_s(\hat{\boldsymbol{x}}_j)^T + \boldsymbol{\Lambda}_s(\hat{\boldsymbol{x}}_j) \;, \tag{9.98}$$

which is similar to (9.81). The algorithm to solve (9.92) and (9.93) is use $\hat{\boldsymbol{x}}_{k|k}$ and $\hat{\boldsymbol{P}}_{k|k}$ as starting values for (9.97) and (9.98) and then perform n_s iterations of (9.97) and (9.98) simultaneously. This algorithm has the advantage of being numerically stable for stiff systems and still computationally efficient, since the fast and stable routines of the linear Kalman filter can be used.

9.8.3 Discretizing the Model and Estimation

When using the Kalman filter from the previous sections we need to discretize the model either once for a given set of parameters, for time invariant models or at each sampling instant, or more frequently, for time invariant models and for the extended Kalman filter. Hence we need to calculate the following quantities

$$\begin{aligned} &\boldsymbol{\Phi}(\tau) = e^{\boldsymbol{A}\tau} \;; \qquad \boldsymbol{\Gamma}(\tau) = \int_0^\tau e^{\boldsymbol{A}s}\boldsymbol{B}\, ds \;; \\ &\boldsymbol{\Lambda}(\tau) = \int_0^\tau \boldsymbol{\Phi}(s)\boldsymbol{G}\boldsymbol{G}^T\boldsymbol{\Phi}(s)^T\, ds \end{aligned} \tag{9.99}$$

where τ is the sampling time, i.e. we need to calculate the exponential of a matrix and integrals involving the matrix exponential.

The likelihood function is as previously, i.e. the changes for non-linear and non-stationary models is that the model is discretized for every sample and the extended Kalman filter are used for state filtering and prediction.

9.9 Case 1: A model for heat dynamics of a building

The thermal characteristic of buildings is frequently approximated by a simple network with resistors and capacitances, see for instance [Hammersten, van Hattem, Bloem & Colombo 1988] or [Madsen 1985].

In this section, such a (simplified) lumped parameter model for the heat dynamics of a simple building, called a test cell is presented.

The dominating heat capacity of the test cell is located in the outer wall. For such buildings, the model with two time constants shown in Figure 9.2 is frequently found adequate.

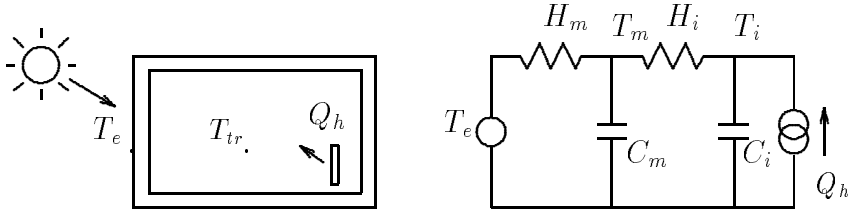

Figure 9.2: *A model with two time constants of the test building and the equivalent electrical network.*

The states of the model are given by the temperature, T_i, of the indoor air and possibly inner part of the walls with heat capacity C_i, and by the temperature, T_m, of the heat accumulating medium, with the heat capacity C_m. H_i is the transmittance of the heat transfer between the room air and the walls, while H_m is the heat transmittance between the inner part of the walls and the external surface of the walls. The input to the system is the heat supply, Q_h, and the outdoor surface temperature, T_e. By considering the outdoor *surface* temperature instead of the outdoor *air* temperature as the input, the effect of solar radiation is automatically taken into account.

In state space form the model is written,

$$\begin{bmatrix} dT_i \\ dT_m \end{bmatrix} = \begin{bmatrix} -H_i/C_i & H_i/C_i \\ H_i/C_m & -(H_i+H_m)/C_m \end{bmatrix} \begin{bmatrix} T_i \\ T_m \end{bmatrix} dt + \begin{bmatrix} 0 & 1/C_i \\ H_m/C_m & 0 \end{bmatrix} \begin{bmatrix} T_e \\ Q_h \end{bmatrix} dt + \begin{bmatrix} dw_i(t) \\ dw_m(t) \end{bmatrix} . \quad (9.100)$$

An additive noise term is introduced to describe deviations between the model and the true system. The term can also be considered as noise on the input signals.

Hence, the model of the heat dynamics is given by the (matrix) stochastic differential equation

$$d\boldsymbol{T} = \boldsymbol{A}\boldsymbol{T}dt + \boldsymbol{B}\boldsymbol{U}dt + d\boldsymbol{w}(t) \ , \tag{9.101}$$

where $\boldsymbol{w}(t)$ is assumed to be a Wiener process with incremental covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{1,i}^2 & 0 \\ 0 & \sigma_{1,m}^2 \end{bmatrix} \ . \tag{9.102}$$

The measured air temperature is naturally encumbered with some measurement errors, and hence the measurement equation is written

$$T_{tr}(t) = [1 \ 0] \begin{bmatrix} T_i \\ T_m \end{bmatrix} + e(t) \ , \tag{9.103}$$

where $e(t)$ is the measurement error, assumed to be normally distributed with zero mean and variance σ_2^2.

The following values of the parameters have been estimated in an earlier experiment on a test cell : $H_i = 55.29$ W/K, $H_m = 13.86$ W/K, $C_i = 325.0$ Wh/K, $C_m = 387.8$ Wh/K, $\sigma_{1,i}^2 = 0.00167$ K^2, $\sigma_{1,m}^2 = 0.00978$ K^2, and $\sigma_2^2 = 0.00019$ K^2. Corresponding to these parameters, the time constants of the system are $\tau_1 = 3.03$ hours and $\tau_2 = 54.28$ hours.

Using these parameters for the model, we are able to simulate the sample paths of the system. By simulating several sample paths from the system and estimating the parameters of the model from each sample path one can validate the estimation procedure. By considering several simulated sequences this investigation considers both the mean values and the variances of the estimated parameters. For the simulation study the following input and output signals of the model have been used : T_e is measured surface temperature, from a Danish test building, Q_h the heat supply, is a PRBS (pseudo-random binary sequence) with the number of stages, $n = 6$ and smallest switching interval T_{prbs} = 8 hours (see [Godfrey 1980]), switching between 0 W and 300 W, and T_{tr} is the indoor room temperature, simulated from the specified model.

In this study the sampling time is fixed to $T_{sampl} = 20$ minutes, and the length of each experiment is 21 days, which is equal to 1512 observations per simulated series. We have simulated 50 equal series, but with different realizations of the noise sequences.

In table 9.1 and Table 9.2 the results are summarized using CTLSM [Melgaard & Madsen 1993], which is a tool for maximum likelihood estimation of multivariate stochastic differential equations with a linear or non-linear state space formulation.

In the tables results are summarized for the parameters of the model and for two physical characteristics, which can be calculated as functions of the model parameters. The

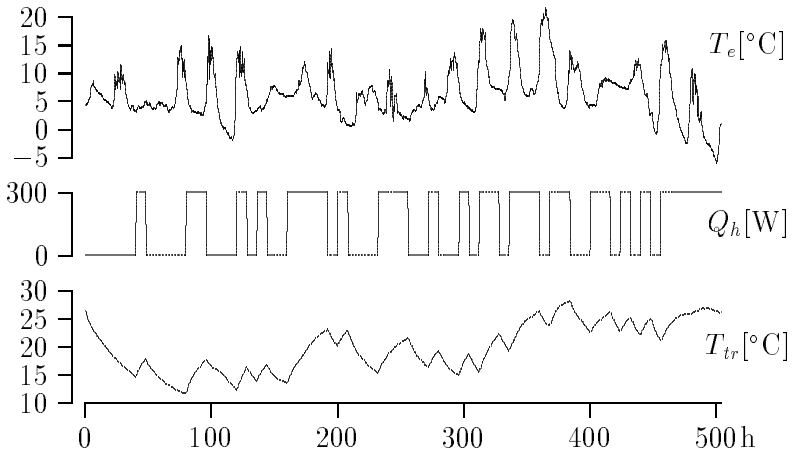

Figure 9.3: *The signals in the simulations: the outdoor surface temperature, the controlled heat power and the indoor air temperature.*

UA-value is the overall heat transmittance and CI is the internal heat capacity of the building. For all parameters, the mean of the estimated values are given, which can be compared to the simulated values. Also the empirical variance and the mean of the estimated variances of the parameters are given. A comparison of these values will indicate if the method is able to estimate the right uncertainty of the parameters.

Parameter	x_{sim}	$\bar{x}$	s_x^2	$\bar{s^2}$	F-stat.	$\|t\|$-stat.
H_i [W/K]	55.290	55.279	2.1645	2.3772	0.9105	0.053
H_m [W/K]	13.860	13.867	0.0360	0.0390	0.9244	0.261
C_i [Wh/K]	325.00	325.0	9.5667	10.746	0.8902	0.011
C_m [Wh/K]	387.78	387.33	162.26	159.89	1.0148	0.250
$\sigma_{1,i}^2$ [K^2]	1.670e-3	1.650e-3	2.185e-8	2.475e-8	0.8825	0.975
$\sigma_{1,m}^2$ [K^2]	9.780e-3	9.554e-3	8.650e-7	1.109e-6	0.7803	1.716
σ_2^2 [K^2]	1.900e-4	1.912e-4	7.956e-10	7.446e-10	1.0686	0.308
UA [W/K]	11.082	11.084	0.0112	0.0111	1.0114	0.125
CI [MJ/K]	2.2862	2.2846	1.4553e-3	1.3815e-3	1.0535	0.304

Table 9.1: *Results from estimation of $n_e = 50$ series, using 1-step predictions in the criterion. The mean autocorrelation of residuals is, $\bar{\rho}(1) = -0.001$.*

In table 9.1 the results are obtained using 1-step predictions in the criterion for maximum likelihood. When estimating the true model, which is the case here, using 1-step predictions is the optimal choice [Stoica & Nehorai 1989].

On the other hand if the true model is not contained in the model set, it can be advantageous to filter the residuals before estimating the parameters. One way of doing this is to use a k-step prediction-horizon in the criterion for some $k > 1$. This kind of filtering works like a low-pass filter and will put more weight on the low frequency part of the dynamics. In the considered case the physical characteristic parameters UA- and CI-values represent low frequency dynamics. The columns of the tables are: x_{sim}, the simulated values, $\bar{x}$, the mean of the estimated values, s_x^2 is the empirical variance of the estimated parameters, $\bar{s^2}$ is the mean of the estimated variance of the parameters, F-stat. is an F statistic given by $Z_F = s_x^2/\bar{s^2}$ and $|t|$-stat. is a t statistic given by $Z_t = |\bar{x} - x_{sim}|/(s_x\sqrt{n_e})$.

In order to verify if the variance of the parameters provided by the estimation tool is equal to the empirical variance, one wish to test the hypotheses

$$\begin{aligned} H_0 &: \quad s_x^2 = \bar{s^2} \\ H_1 &: \quad s_x^2 \neq \bar{s^2} \ . \end{aligned}$$

Under H_0 we have in this case that $Z_F \sim F(49, \infty)$. The critical set for this test is $\{z < F(49,\infty)_{\alpha/2} \vee z > F(49,\infty)_{1-\alpha/2}\}$ on level α. By choosing $\alpha = 0.1$ we obtain the critical set $\{z < 0.74 \vee z > 1.35\}$. It is seen from Table 9.1 that we cannot reject H_0 for any parameter on the chosen level.

Another test is performed in order to verify that the estimated parameters are not biased. The following hypotheses are tested

$$\begin{aligned} H_0 &: \quad \bar{x} = x_{sim} \\ H_1 &: \quad \bar{x} \neq x_{sim} \ . \end{aligned}$$

Under H_0 the distribution of the test statistic is $Z_t \sim t(49)$. The critical set for this test is $\{z > t(49)_{1-\alpha/2}\}$ on level α. For $\alpha = 0.1$, the critical set is $\{z > 2.0\}$, thus from Table 9.1 we cannot reject H_0 for any of the parameters on the chosen level.

In table 9.2 the results from estimation using a 4-step prediction-horizon in the criterion are shown. In this ideal case where the true model is contained by the model set there is no difference in the estimated parameters which is also seen from the tables, they are still unbiased. A problem, though, when "filtering" the residuals in this way is that the residuals are no longer white noise. From the tables we have that the mean autocorrelation of residuals are $\bar{\rho}(1) = 0.712$ when using the criterion based on 4-step predictions against $\bar{\rho}(1) = -0.001$ for 1-step predictions. When the residuals are autocorrelated a number of statistical tests for model validation are no longer valid. Another problem in table 9.2 is that the uncertainties of the parameters are underestimated when the autocorrelation of residuals is not taken into account in the calculation of the uncertainties. Wahlberg ([Wahlberg & Ljung 1986]) has shown, that asymptotically, it is only the prediction horizon itself, that affects the weighting in the frequency

Parameter	x_{sim}	$\bar{x}$	s_x^2	$\bar{s^2}$	F-stat.	$\|t\|$-stat.
H_i [W/K]	55.290	55.339	2.2259	0.6154	3.617	0.232
H_m [W/K]	13.860	13.863	0.0383	0.0099	3.877	0.109
C_i [Wh/K]	325.00	324.8	9.486	3.3338	2.842	0.481
C_m [Wh/K]	387.78	387.20	163.74	47.685	3.4338	0.3210
$\sigma_{1,i}^2$ [K^2]	1.670e-3	1.616e-3	4.158e-8	1.620e-8	2.5665	1.880
$\sigma_{1,m}^2$ [K^2]	9.780e-3	9.562e-3	8.680e-7	4.156e-7	2.0884	1.656
σ_2^2 [K^2]	1.900e-4	2.072e-4	3.700e-9	4.145e-9	0.8926	1.997
UA [W/K]	11.082	11.084	0.0115	0.0028	4.1290	0.107
CI [MJ/K]	2.2862	2.2837	1.469e-3	3.954e-4	3.7159	0.460

Table 9.2: *Results from estimation of $n_e = 50$ series, using 4-step predictions in the criterion. The mean autocorrelation of residuals is, $\bar{\rho}(1) = 0.712$.*

domain, and not how it is split up into sampling interval times number of predicted sampling instants. This means that, asymptotically, we could obtain the same results as using a k-step prediction, by using a proper anti-aliasing filter followed by a new sampling of the data and then estimate with a one-step prediction criterion. In this way one could avoid the autocorrelated residuals.

9.10 Case 2: A non-linear model for the heat dynamics of a building

The case considered is related to a CEC research project on identification of thermal characteristics of building components tested in situ. In Figure 9.4 is shown a rough sketch of a test building used for testing the building components which is placed as the south wall of the building. In this case the wall to be tested consists of a lightweight insulated sandwich panel, provided with a double glazed window. Also in Figure 9.4 is mentioned the main measured quantities, which includes $T_{tr,a}$, $T_{tr,s}$: internal air temperature and surface temperature, respectively, $T_{sr,a}$: the air temperature of the service room, $T_{e,s}$, $T_{e,a}$: the external surface and air temperature of the test building, respectively, q_h : the internal heat supply, and $G_{dif,v}$, $G_{dir,v}$: global vertical diffuse and direct solar radiation on south, respectively. The total vertical global radiation is $G_v = G_{dif,v} + G_{dir,v}$. The solar angle of incidence, θ_{inc}, is calculated from a physical model based on the ground location and the time of the year.

The procedure is first to perform a calibration experiment with a thick opaque insu-

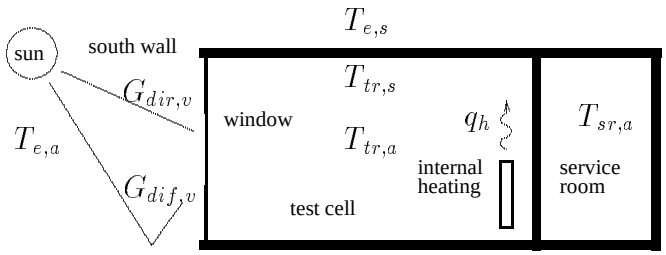

Figure 9.4: *A cross section of the test building with an installed south wall, here a lightweight component with a double glazed window in the middle.*

lation panel replacing the test component and subsequently identify the model of the test cell from this experiment. Then a second experiment is performed with the test component installed, and by including the information obtained from the calibration experiment it is possible to extract information about the test component separately. One of the questions to ask about this procedure is *how to incorporate the information from previous experiments in the estimation on the new data in a proper way*. One possibility, and the usual approach, is to fix the previously estimated parameters in the total model of the test cell and test component and estimate the remaining parameters. Due to the rather high correlation between some of the inputs to the model, mainly between the climate series, it is necessary that some of the parameters are fixed, in order for the system to be identifiable. The parameters to fix have been estimated from the calibration experiment, with an associated uncertainty. If the uncertainty is small there is no problem with this approach, but this is not always the case. Therefore a Bayesian approach is proposed, where the prior information is specified as a prior distribution function, and hence accounts for the associated uncertainty of the parameters. Since the parameters from the calibration experiment are ML estimates they are approximately normally distributed. Thus the prior distribution is determined uniquely from the estimated parameters and their associated covariance matrix. Then MAP estimates for the model are obtained by maximizing (9.50) as in Example 9.1.

The model of the test building is mainly build up as an R-C network, of heat resistors and heat capacitors by analogy with an electrical network. This is basically a linear model which is a lumped model of the equations for heat diffusion. The main structure of the model is shown in Figure 9.6.

The top branch of the model in Figure 9.6 represents the east wall, west wall, floor and roof test cell. The middle branch is the partition wall between the service room and the test cell. In the calibration experiment the parameters of these to branches of the model are estimated. The south wall is then replaced by a calibration wall, which is highly insulated and homogeneous. Thus, the heat flow through this wall can be measured by a heat-flow meter during the calibration experiment. In the calibration experiment

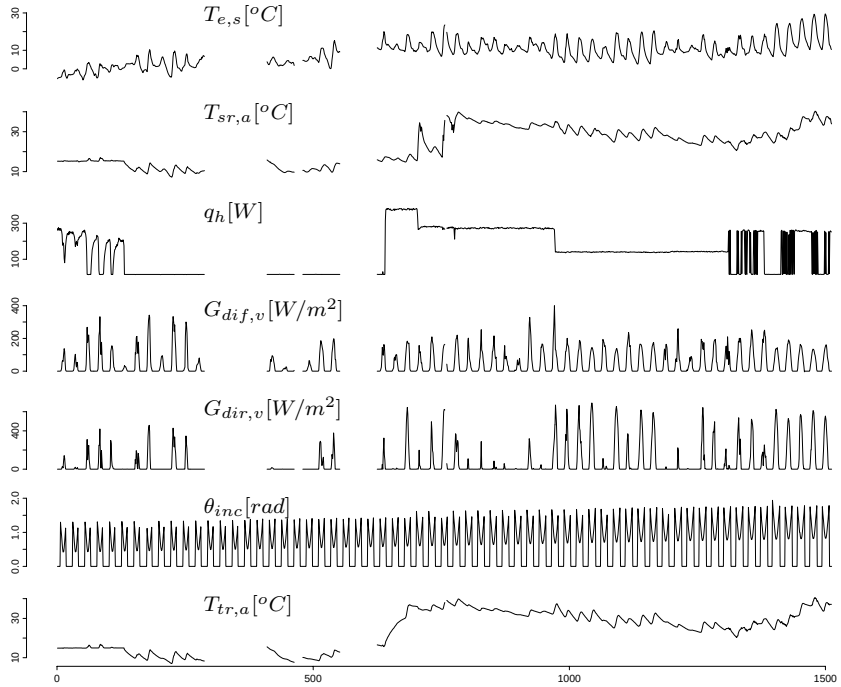

Figure 9.5: *Some of the data series from the experiment. The unit on the horizontal axis is hours. The holes in the series are missing observations due to different errors during data logging of the experiment.*

there is no solar radiation entering the internal test cell surface (the arrow on the top branch of the model). The parameters of these two branches of the model can then either be fixed, or included as a prior distribution function, when a MAP estimator is used for estimation based on the new experiment.

The bottom branch of the network model in Figure 9.6 describes the heat transfer through the test wall. The parameter H_{psc1} is fixed to some prior expected value, expressing the prior expected conductance between the middle of the test cell and the internal surface on the south wall. $f(G_{dif,v}, G_{dir,v})$ is a nonlinear function describing the solar transmittance through the window. The direct radiation is known to have a transmittance which is dependant upon the angle of incidence of the radiation, due to reflection by the window. A theoretical expression for the direct transmittance with relation to losses by reflections for different numbers of glass covers is used as the basis for estimating the nonlinear relation for a real window, which also includes a frame etc.

$$f(G_{dif,v}, G_{dir,v}) = A_{diff} G_{dif,v} + A_{dir} F(\theta_{inc}, \alpha_{eff}) G_{dir,v} \tag{9.104}$$

where A_{diff} and A_{dir} are constants (to be estimated). $F(\theta_{inc}, \alpha_{eff})$ is the theoreti-

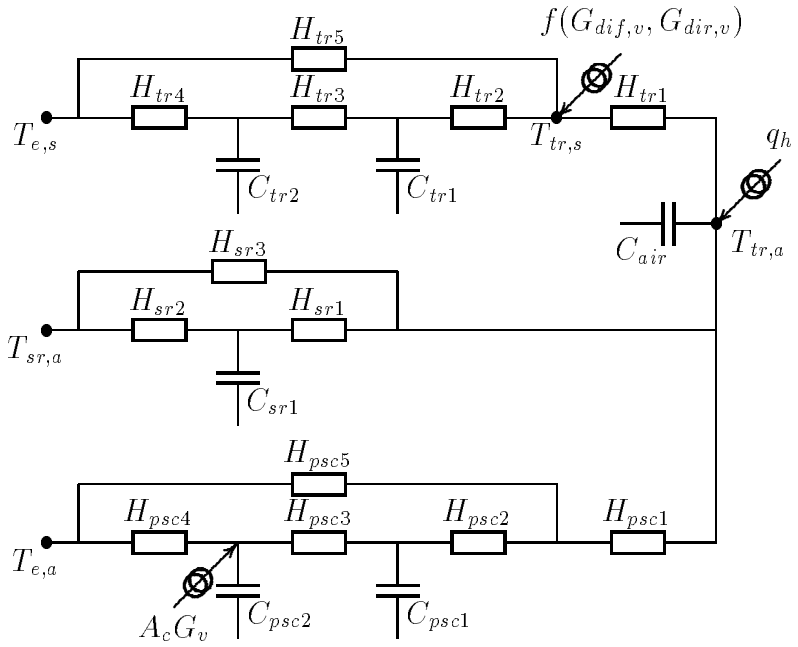

Figure 9.6: Network model of the test building with the installed test component.

cal function for the direct transmittance dependent on the angle of incidence and for α_{eff} number of glass covers. The expression is plotted in Figure 9.7. Theoretically the shape of the curve, in this case, should correspond to double glazing, but due to unmodelled boundary conditions, (the frame etc.), the shape of the curve is estimated from the data. The estimated number, α_{eff}, may be considered as the effective number of glass covers, which can take any positive real number.

The nonlinear model is compared to a linear model, which was used earlier in the PASSYS project [Wouters 1993]. For the linear model, the expression for the solar transmittance is given by

$$f(G_{dif,v}, G_{dir,v}) = A_{tot}(G_{dif,v} + G_{dir,v}) \ , \tag{9.105}$$

where A_{tot} is some constant.

In addition to the linear R-C network in Figure 9.6 and the nonlinear expression for the direct transmittance shown in Figure 9.7, the total model of the system also contains a model for the noise. The noise model accounts for the measurement noise on the inputs and outputs of the system, and un-modelled dynamics. The inputs to the model are

$$\boldsymbol{u}^T = (T_{e,s} \ \ T_{sr,a} \ \ T_{e,a} \ \ q_h \ \ G_{dif,v} \ \ G_{dir,v} \ \ \theta_{inc})^T \ . \tag{9.106}$$

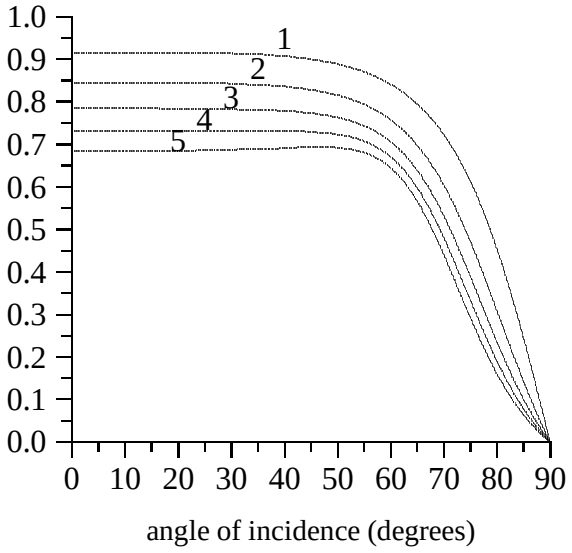

Figure 9.7: *Theoretical direct transmittance w.r.t. losses by reflections for different number of glass covers.*

The outputs of the model are

$$\boldsymbol{y}^T = (T_{tr,a} \;\; T_{tr,s})^T \; . \tag{9.107}$$

The whole model can be put in the form of (8.32) and (8.32), described in Section **??**, with an additive noise term on the state vector and an additive measurement noise on the output equation. We are then able to estimate the parameters of the model, using e.g. CTLSM, see [Melgaard & Madsen 1993].

Results

From the overall model a number of characteristic physical parameters are calculated, these include: $UA_{tr,e}$, steady state overall thermal transmission coefficient between test room and outdoor surfaces, $UA_{tr,sr}$, the steady state overall thermal transmission coefficient between test room and service room, UA_{psc}, steady state overall thermal transmission coefficient for the test component, gA_{psc}, steady state overall solar transmittance, or total solar heat gain factor of the test component, which is the ratio of heat entering the test-cell caused by solar radiation on the component, divided by the intensity of incident solar radiation on the component, $CI_{tr,e}$, internal test room heat capacity, i.e. the amount of heat which goes into the test room-envelope as a result of a change from one steady state situation to the same steady state situation except for the indoor temperature being raised by 1 K, $CI_{tr,sr}$, similar for the partition to service room, and CI_{psc}, similar for the test component.

Four situations of the estimation have been considered and compared. Fixed/ML versus MAP/ML estimates have been compared, and models with or without the nonlinear function for the solar transmittance have been compared. The main results from the four situations are shown in Table 9.10 below. Beyond the estimated values of the

	linear		non-linear	
Parameter	fixed/ML	MAP/ML	fixed/ML	MAP/ML
$UA_{tr,e}$	7.826	8.265	7.826	8.294
$[W/K]$	(0.096)	(0.066)	(0.096)	(0.088)
$CI_{tr,e}$	2.045	1.976	2.043	2.000
$[MJ/K]$	(0.045)	(0.015)	(0.045)	(0.023)
$UA_{tr,sr}$	1.001	2.096	1.001	2.170
$[W/K]$	(0.148)	(0.237)	(0.148)	(0.242)
$CI_{tr,sr}$	0.274	0.350	0.274	0.355
$[MJ/K]$	(0.024)	(0.006)	(0.024)	(0.009)
UA_{psc}	6.047	5.885	5.917	5.735
$[W/K]$	(0.044)	(0.055)	(0.049)	(0.094)
CI_{psc}	0.136	0.153	0.116	0.119
$[MJ/K]$	(0.006)	(0.006)	(0.009)	(0.011)
gA_{psc}	0.573	0.594	0.612	0.633
$[m^2]$	(0.006)	(0.008)	(0.004)	(0.005)
α_{eff}	-	-	3.391	3.401
	-	-	(0.811)	(0.340)
mean p.e.	-0.0076	0.0016	-0.0054	0.0004
var p.e.	0.0151	0.0122	0.0131	0.0105

Table 9.3: *Main results from the estimations.*

characteristic parameters and their associated standard deviation in brackets, is also listed the mean and variance of the one-step prediction errors, for the internal surface temperature, $T_{tr,s}$, for the different cases.

It should be mentioned that in all cases the estimated physical characteristics are reasonable compared to the expected theoretical values, except for $UA_{tr,sr}$ in both fixed/ML cases, in which the values are too low compared to the theoretical values. The theoretical values, however, only covers one-dimensional heat loss; i.e. thermal bridges etc. are not taken into account. The gA values cannot be compared directly, because the linear cases consider some mean value of the solar aperture, whereas in

the non-linear cases it is an angle dependant function, (the numbers are specified for 25% diffuse radiation, and 45° angle of incidence for the direct radiation).

The difference in the estimated physical parameters from linear to non-linear model is small, except for the gA value; but the inclusion of the non-linear part has increased the capability of the model for prediction, the variance of the prediction error has decreased about 13 % for both the ML and MAP case. The larger difference is between the ML and MAP approach. In the pure ML approach, a number of the parameters have been fixed to the expected value estimated from the calibration experiment, whereas in the MAP approach, all parameters are estimated with a prior distribution determined by the calibration experiment. It is seen, that the physical characteristics which are mainly determined by the calibration experiment, i.e. $UA_{tr,e}$ and especially $UA_{tr,sr}$ have changed significantly from their prior expected value to their posterior expected value. This means, that despite the fact that the inputs of exterior climate variables to the model are rather high correlated, there is still enough information in the new experiment to change these physical characteristics. Especially for $UA_{tr,sr}$ it was only possible to obtain vague information from the calibration experiment, due to a badly excited input signal, $T_{sr,a}$, for that part of the model. This fact was also confirmed by visually plotting the inputs for the calibration experiment. In the new experiment this input signal has a better excitation. Still, though, this signal does not have a very good excitation, compare Figure 9.5, for large parts of the experiment $T_{sr,a} = T_{tr,a}$.

It has been shown that the proposed method and model are superior to the earlier approach for modelling this system. This includes the use of MAP estimation as a means of including the prior information about the system, and the use of a nonlinear model for the solar transmission through the window.

A plot of the measured and simulated output from the model gives an indication of the performance of the model, see Figure 9.8.

Even though the model seems to perform well with respect to simulated output, it is still not perfect, which is clearly seen from a cumulative periodogram of the residuals (the one-step prediction errors). A plot of the cumulative periodogram for the residuals of the two outputs of the model is given in Figure 9.9 below.

From the cumulative periodogram it is seen that none of the residual sequences will pass a white noise test. This is especially true for $T_{tr,s}$. A closer look at the plot reveals that the curve for $T_{tr,s}$ has some well defined steps. These are located at the frequency of daily cycle and the higher harmonics of this frequency. This could indicate that the model of the transmitted solar radiation hitting the internal surface of the test cell (floor and wall) need further improvements. This may not be possible with the current measurement setup. Either more temperature sensors are needed at the inside surface of the test cell (where the sun hits the floor) or it should be realized that $T_{tr,s}$ can not be

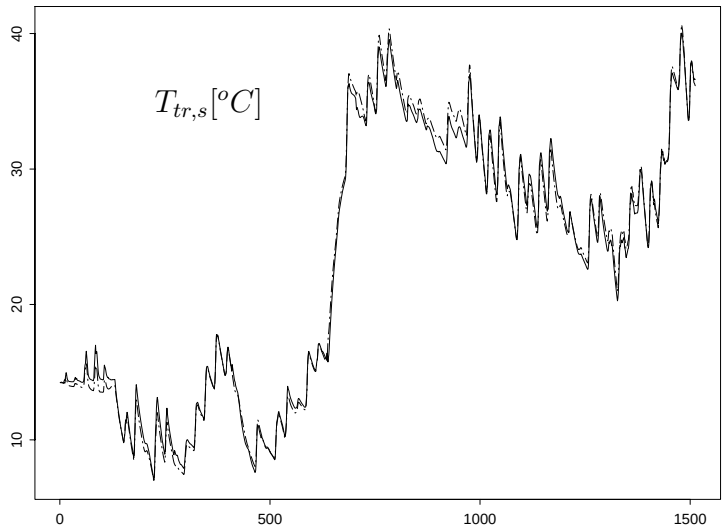

Figure 9.8: *Plot of $T_{tr,s}$, measured (solid) and simulated from the proposed model (dashed). The unit on the horizontal axis is hours.*

modelled by a one-dimensional model of the test cell, when the test wall has a window.

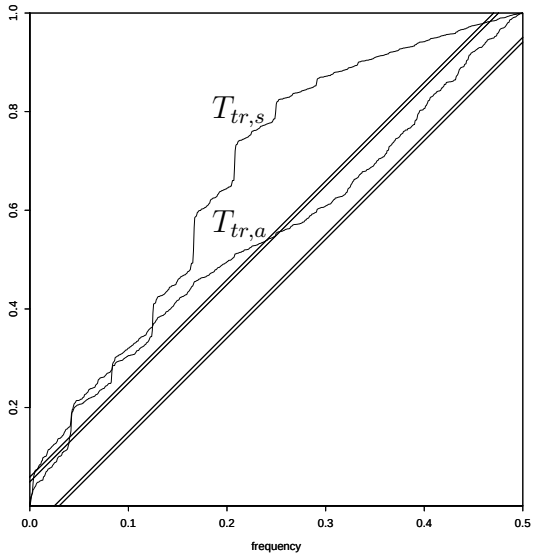

Figure 9.9: *The cumulative periodogram of the residuals of the two outputs of the model. The 95 % and 99 % confidence limits for a white noise test are depicted.*

9.11 Case 3: Estimation of a non-linear model — Predator/prey relations

In order to illustrate the non-linear estimation and specifically the extended Kalman filter, an example using a nonlinear model is considered.

The example is a classical type of nonlinear system describing the temporal oscillations of a population of predators and their prey in a localized geographic region. The prey population is denoted by $N_1(t)$ and the predator population by $N_2(t)$. It is assumed that the growth rate of prey, in the absence of predators, is $aN_1(t)$, while the predator multiplication rate is $cN_1(t)N_2(t)$. Further, is is assumed that the loss rate of prey is proportional to the numbers of prey and predators, i.e. loss rate of prey equals $bN_1(t)N_2(t)$, while the loss rate of predators equals their death rate $dN_2(t)$.

Putting together these assumptions, the dynamics of the predator - prey interaction are described by the deterministic Lotka-Volterra equations

$$\dot{N}_1(t) = aN_1(t) - bN_1(t)N_2(t) \,, \tag{9.108}$$
$$\dot{N}_2(t) = cN_1(t)N_2(t) - dN_2(t) \,, \tag{9.109}$$

where a, b, c and d all are positive constants. The point $N_1^* = d/c$, $N_2^* = a/b$ is the only nontrivial equilibrium of the Lotka-Volterra system. If the initial condition $(N_1(0), N_2(0)) \neq (N_1^*, N_2^*)$, the trajectory in the phase–planeof the system is a closed orbit as depicted in Figure 9.10. In the figure, the trajectory is moving anti-clockwise. Thus, for any given initial condition, the populations of predator and prey will oscillate

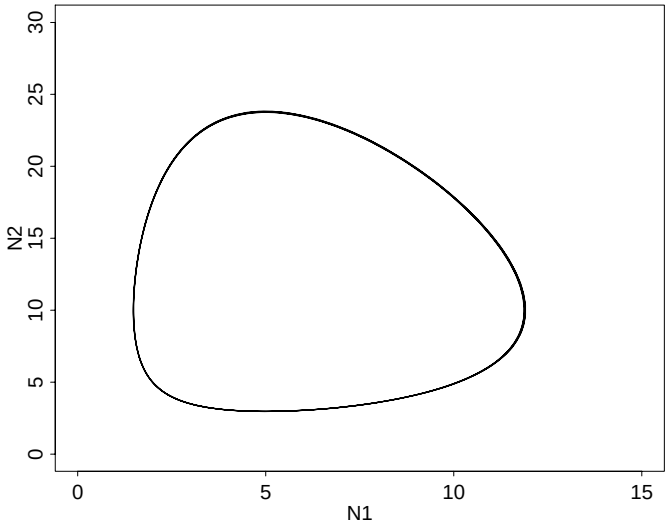

Figure 9.10: *Phase-plane of the deterministic Lotka-Volterra system.*

cyclically. Neither species will die out, nor will they grow indefinitely. Furthermore, except for the improbable initial state (N_1^*, N_2^*), the populations will not remain constant.

The deterministic Lotka-Volterra model is modified in order to give a more likely picture of a real world system. It is assumed that the rate of change of each population is influenced by an additive random term representing the effect of other factors not included in the model, such as other predators in the food chain and weather variables. This results in a stochastic Lotka-Volterra model

$$dN_1(t) = (aN_1(t) - bN_1(t)N_2(t))\,dt + \sigma_1 d\beta_1(t)\ , \tag{9.110}$$
$$dN_2(t) = (cN_1(t)N_2(t) - dN_2(t))\,dt + \sigma_2 d\beta_2(t)\ , \tag{9.111}$$

where $\beta_1(t)$ and $\beta_2(t)$ are mutually independent standard Wiener processes and σ_1 and σ_2 are positive constants giving the levels of the noise. If we are able to observe the populations, by measuring the system at given time instants, the measurements are assumed to by influenced by a random error, i.e. the measurement equation is

$$N_{1,m}(t_k) = N_1(t_k) + \sigma_{1,m} e_1(t_k)\ , \tag{9.112}$$
$$N_{2,m}(t_k) = N_2(t_k) + \sigma_{2,m} e_2(t_k)\ , \tag{9.113}$$

where $e_1(t_k)$ and $e_2(t_k)$ are standard Gaussian white noise and $\sigma_{1,m}$ and $\sigma_{2,m}$ are positive constants giving the levels of the noise. An example of the measurements of such a system is simulated, using stochastic simulation, and the timeseries of measured populations are shown in Figure 9.11.

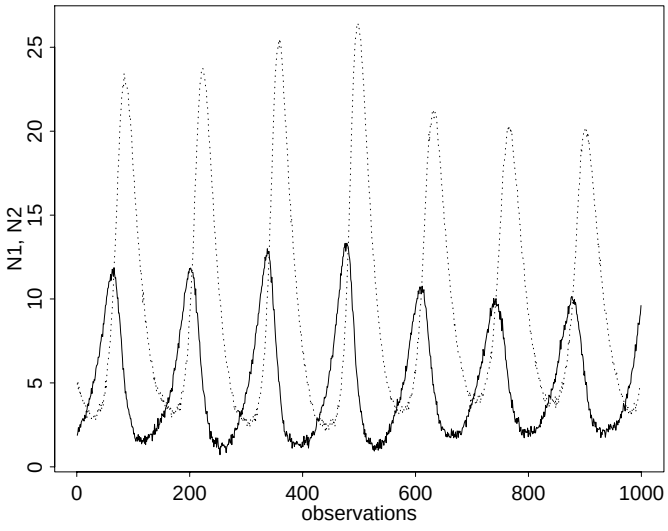

Figure 9.11: *Timeseries plot of the populations $N1$ and $N2$ for the stochastic Lotka-Volterra system.*

In this example the following dimensionless variables were chosen, $a = 10$, $b = 1$, $c = 2$, $d = 10$, $\sigma_1^2 = \sigma_2^2 = 0.3$, $\sigma_{1,m}^2 = \sigma_{2,m}^2 = 0.03$, and the sampling time, $T_s = 0.005$, and initial populations, $(N_1(0), N_2(0)) = (2, 5)$. A phase-plane of the same realization as in figure 9.11 is shown in Figurelv-stoc2. The qualitative structure

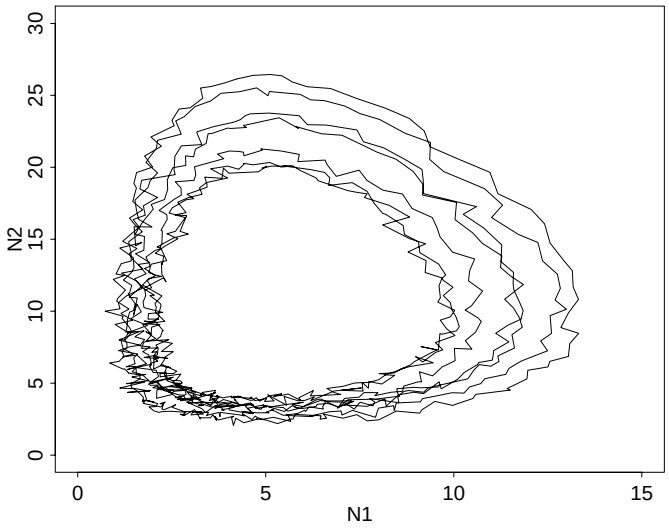

Figure 9.12: *Phase-plane of the stochastic Lotka-Volterra system.*

of the trajectory is the same as for the deterministic system and it is still a stable oscillating system.

50 sequences of stochastic independent realizations of the system was simulated. Each sequence with the length of 500 observations of the populations. In Table 9.4 the results from estimation of the parameters of the model from the 50 series are summarized. For all parameters, the mean of the estimated values are given, which can be compared to the simulated values. Also the empirical variance and the mean of the estimated variances of the parameters are given. A comparison of these values will indicate if the tool is able to estimate the right uncertainty of the parameters. The columns of the table are : x_{sim}, the simulated values, $\bar{x}$, the mean of the estimated values, s_x^2 is the empirical variance of the estimated parameters, $\bar{s^2}$ is the mean of the estimated variance of the parameters, F-stat. is an F statistic given by $Z_F = s_x^2/\bar{s^2}$ and $|t|$-stat. is a t statistic given by $Z_t = |\bar{x} - x_{sim}|/(s_x\sqrt{n_e})$.

In order to verify if the variance of the parameters provided by the estimation tool is equal to the empirical variance, one wish to test the hypotheses

$$\begin{aligned} H_0 &: \quad s_x^2 = \bar{s^2} \\ H_1 &: \quad s_x^2 \neq \bar{s^2} \ . \end{aligned}$$

Under H_0 we have in this case that $Z_F \sim F(49, \infty)$. The critical set for this test is $\{z \in \mathbb{R} | z \in \mathbb{R} | z < F(49, \infty)_{\alpha/2} \vee z > F(49, \infty)_{1-\alpha/2}\}$ on level α. By choosing

Parameter	x_{sim}	$\bar{x}$	s_x^2	$\bar{s^2}$	F-stat.	$\|t\|$-stat.
a	10.000	9.985	8.568e-3	9.589e-3	0.8936	1.151
b	1.0000	0.9990	6.453e-5	8.267e-5	0.7805	0.914
c	2.0000	2.0012	1.253e-4	1.095e-4	1.1441	0.758
d	10.000	10.001	4.756e-3	4.481e-3	1.0614	1.446
$N_1(0)$	2.0000	2.0066	5.426e-3	5.888e-3	0.9215	0.630
$N_2(0)$	5.0000	5.0175	7.893e-3	5.902e-3	1.3373	1.393
σ_1^2	0.3000	0.2909	2.883e-3	3.529e-3	0.8169	1.196
σ_2^2	0.3000	0.2896	9.011e-3	7.981e-3	1.1290	0.773
$\sigma_{1,m}^2$	0.03000	0.02966	3.229e-6	3.855e-6	0.8377	1.346
$\sigma_{2,m}^2$	0.03000	0.03047	4.593e-6	4.184e-6	1.0978	1.534

Table 9.4: *Results from estimation of the $n_e = 50$ series from the stochastic Lotka-Volterra system.*

$\alpha = 0.1$ we obtain the critical set $\{z \in \mathbb{R} | z < 0.74 \vee z > 1.35\}$. It is seen from Table 9.4 that we cannot reject H_0 for any parameter on the chosen level.

Another test is performed in order to verify that the estimated parameters are unbiased. The following hypotheses are tested

$$\begin{aligned} H_0 &: \quad \bar{x} = x_{sim} \\ H_1 &: \quad \bar{x} \neq x_{sim} \ . \end{aligned}$$

Under H_0 the distribution of the test statistic is $Z_t \sim t(49)$. The critical set for this test is $\{z \in \mathbb{R} | z > t(49)_{1-\alpha/2}\}$ on level α. For $\alpha = 0.1$, the critical set is $\{z \in \mathbb{R} | z > 2.0\}$, thus from Table 9.4 we cannot reject H_0 for any of the parameters on the chosen level.

Chapter 10

Tracking time-varitions in the system description

10.1 Introduction

Non-stationarity in non-linear or linear systems, may show up as timevariation in the basic or an approximating system description. In such situations, the parameter or functional estimation may be done recursively. Thus algorithms for system tracking can be used e.g.

- when the system is timevarying, or in some other respect is nonstationary,
- when the system is supervised with some kind of event detection algorithm, e.g. change-of-mode detection.
- when the system is adaptively controlled, predicted etc. and the new e.g. control input have to be based on the latest parameter estimates,
- when the computer space required for the computations is limited, or when the available time for computing a new estimate is short.

In algorithms for recursive parameter estimation, the parameter estimates are updated at each sampling instant using the latest measurement. General recursive versions of on-line methods can be derived, e.g. via minimization of a function of $\boldsymbol{\theta}$, the unknown vector of parameters, as e.g.

$$V_t(\boldsymbol{\theta}) = 0.5 \sum_{1}^{t} \ell(\varepsilon_i(\boldsymbol{\theta}), i, \boldsymbol{\theta}), \tag{10.1}$$

where ℓ is a function of the prediction error

$$\varepsilon_t(\boldsymbol{\theta}) = y_t - \widehat{y}_{t|t-1}(\boldsymbol{\theta}), \tag{10.2}$$

as in the off-line methods that previously have been studied. This topic is discussed in general books on system identification, see e.g. [Ljung 1995] and [Söderström & Stoica 1988] and in more specialized opii on recursive estimation, see e.g. [Ljung & Söderström 1983] or [Young 1984]. Also [Goodwin & Sin 1984] contains an extensive treatment of recursive estimation algorithms, in particular in connection with adaptive signal processing.

10.2 Recursive least squares

The generic recursive estimation algorithm is the Recursive Least Squares (RLS) algorithm which, with proper choice of initial values, is an exact reformulation of the off-line least squares algorithm. Hence the sequence of estimates obtained from RLS is except for the effects of initial conditions of the algorithm identical to the estimates from the off-line algorithm when applied to the same set of data.

It is customary to refer to Kalman, [**?**], as the very early presentation of the recursive version of the least squares method. However, the algorithm was around also a little before that, it may be found also in a paper by Plackett from 1950, [Plackett 1950].

Consider the following model of a linear system with discrete time

$$Y_t + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} = \omega_1 U_{t-1} + \cdots + \omega_m U_{t-m} + \epsilon_t, \tag{10.3}$$

where $\{Y_t\}$ is the output signal, $\{\epsilon_t\}$ is white noise, uncorrelated with the external signal $\{U_t\}$. Introduce the vectors of regressors and of unknown parameters, $\boldsymbol{X}_t$ and $\boldsymbol{\theta}$ respectively,

$$\boldsymbol{X}_t^T = (-Y_t, \cdots, -Y_{t-p}, U_{t-1}, \cdots, U_{t-m}) \quad \text{and} \tag{10.4}$$

$$\boldsymbol{\theta}^T = (\phi_1, \cdots, \phi_p, \omega_1, \cdots, \omega_s). \tag{10.5}$$

The model (10.3) may now be written as a linear regression

$$Y_t = \boldsymbol{X}_t^T \boldsymbol{\theta} + \epsilon_t. \tag{10.6}$$

The *on-line least squares estimate* or *recursive least squares estimate* of the parameters $\boldsymbol{\theta}$ is found as the recursively computed solution to

$$\widehat{\boldsymbol{\theta}}_t = \arg\min_{\boldsymbol{\theta}} S_t(\boldsymbol{\theta}), \tag{10.7}$$

where

$$S_t(\boldsymbol{\theta}) = \sum_{s=1}^{t} (Y_s - \boldsymbol{X}_s^T \boldsymbol{\theta})^2. \tag{10.8}$$

The off-line solution is easily found to be

$$\widehat{\boldsymbol{\theta}}_t = \boldsymbol{R}_t^{-1} \boldsymbol{h}_t, \tag{10.9}$$

$$\boldsymbol{R}_t = \sum_{s=1}^{t} \boldsymbol{X}_s \boldsymbol{X}_s^T \quad \text{and} \tag{10.10}$$

$$\boldsymbol{h}_t = \sum_{s=1}^{t} \boldsymbol{X}_s Y_s. \tag{10.11}$$

Both $\boldsymbol{R}_t$ and $\boldsymbol{h}_t$ may be written recursively by using the previously computed values $\boldsymbol{R}_{t-1}$ and $\boldsymbol{h}_{t-1}$. The *updating* formulas are

$$\begin{aligned} \boldsymbol{R}_t &= \boldsymbol{R}_{t-1} + \boldsymbol{X}_t \boldsymbol{X}_t^T && (10.12) \\ \boldsymbol{h}_t &= \boldsymbol{h}_{t-1} + \boldsymbol{X}_t Y_t. && (10.13) \end{aligned}$$

The expressions (10.9) and (10.12), (10.13) do not show the evolution of the sequence of parameter estimates, but it is easy to derive an expression for the new parameter estimate, $\widehat{\boldsymbol{\theta}}_t$, in terms of the old estimate, $\widehat{\boldsymbol{\theta}}_{t-1}$, of $\boldsymbol{\theta}$.

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_t &= \boldsymbol{R}_t^{-1} \boldsymbol{h}_t = \boldsymbol{R}_t^{-1} \left[\boldsymbol{h}_{t-1} + \boldsymbol{X}_t Y_t \right] \\ &= \boldsymbol{R}_t^{-1} \left[\boldsymbol{R}_{t-1} \widehat{\boldsymbol{\theta}}_{t-1} + \boldsymbol{X}_t Y_t \right] \\ &= \boldsymbol{R}_t^{-1} \left[\boldsymbol{R}_t \widehat{\boldsymbol{\theta}}_{t-1} - \boldsymbol{X}_t \boldsymbol{X}_t^T \widehat{\boldsymbol{\theta}}_{t-1} + \boldsymbol{X}_t Y_t \right] \\ &= \widehat{\boldsymbol{\theta}}_{t-1} + \boldsymbol{R}_t^{-1} \boldsymbol{X}_t \left[Y_t - \boldsymbol{X}_t^T \widehat{\boldsymbol{\theta}}_{t-1} \right]. \end{aligned}$$

Note that the last part of the expression shows the change of the parameter estimates that is caused by the one-step prediction error obtained at time t using the parameter estimates from $t-1$. The vector $\boldsymbol{R}_t^{-1} \boldsymbol{X}_t$ thus may be interpreted as a gain vector. In case of scalar output the update of the parameter estimates thus falls along a line in parameter space, determined by the gain vector. Consequently, it is important that the gain vector during the estimation traverses the whole parameter space in order to make it possible to estimate all parameters. This may be interpreted as a persistantly of excitation criterion (cf. citeLLg95) on the external signal, and will be treated in the upcoming discussion on selective forgetting algorithms, cf. Section **??** as well as in the subsequent chapter on experiment design, cf. Chapter 12.

In summary we have the *RLS algorithm* for parameter estimation in dynamic models:

$$\widehat{\boldsymbol{\theta}}_t = \widehat{\boldsymbol{\theta}}_{t-1} + \boldsymbol{R}_t^{-1} \boldsymbol{X}_t \left[Y_t - \boldsymbol{X}_t^T \widehat{\boldsymbol{\theta}}_{t-1} \right] \quad (10.14)$$

$$\boldsymbol{R}_t = \boldsymbol{R}_{t-1} + \boldsymbol{X}_t \boldsymbol{X}_t^T. \quad (10.15)$$

In order to avoid the calculation of the inverse of $\boldsymbol{R}_t$ in every step, introduce

$$\boldsymbol{P}_t = \boldsymbol{R}_t^{-1} \quad (10.16)$$

and use the matrix inversion lemma cf. [**?**], supposing the inverses exist.

$$[\boldsymbol{A} + \boldsymbol{BCD}]^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1} \boldsymbol{B} \left[\boldsymbol{D} \boldsymbol{A}^{-1} \boldsymbol{B} + \boldsymbol{C}^{-1} \right]^{-1} \boldsymbol{D} \boldsymbol{A}^{-1} \quad (10.17)$$

with $\boldsymbol{A} = \boldsymbol{R}_{t-1}$, $\boldsymbol{B} = \boldsymbol{D}^T = \boldsymbol{X}_t$ and $\boldsymbol{C} = 1$ to obtain

$$\boldsymbol{P}_t = \boldsymbol{P}_{t-1} - \frac{\boldsymbol{P}_{t-1} \boldsymbol{X}_t \boldsymbol{X}_t^T \boldsymbol{P}_{t-1}}{1 + \boldsymbol{X}_t^T \boldsymbol{P}_{t-1} \boldsymbol{X}_t}. \quad (10.18)$$

If the gain matrix

$$\boldsymbol{K}_t = \boldsymbol{R}_t^{-1}\boldsymbol{X}_t = \boldsymbol{P}_{t-1}\boldsymbol{X}_t - \frac{\boldsymbol{P}_{t-1}\boldsymbol{X}_t\boldsymbol{X}_t^T\boldsymbol{P}_{t-1}\boldsymbol{X}_t}{1+\boldsymbol{X}_t^T\boldsymbol{P}_{t-1}\boldsymbol{X}_t}$$

or

$$\boldsymbol{K}_t = \frac{\boldsymbol{P}_{t-1}\boldsymbol{X}_t}{1+\boldsymbol{X}_t^T\boldsymbol{P}_{t-1}\boldsymbol{X}_t} \tag{10.19}$$

is introduced, the updating of the estimate of $\boldsymbol{\theta}$ can be written

$$\widehat{\boldsymbol{\theta}}_t = \widehat{\boldsymbol{\theta}}_{t-1} + \boldsymbol{K}_t\left[Y_t - \boldsymbol{X}_t^T\widehat{\boldsymbol{\theta}}_{t-1}\right]. \tag{10.20}$$

Hence, summarizing, the Recursive Least Squares algorithm may be written as

$$\widehat{\boldsymbol{\theta}}_t = \widehat{\boldsymbol{\theta}}_{t-1} + \boldsymbol{K}_t\left[Y_t - \boldsymbol{X}_t^T\widehat{\boldsymbol{\theta}}_{t-1}\right] \tag{10.21}$$

$$\boldsymbol{K}_t = \frac{\boldsymbol{P}_{t-1}\boldsymbol{X}_t}{1+\boldsymbol{X}_t^T\boldsymbol{P}_{t-1}\boldsymbol{X}_t} \tag{10.22}$$

$$\boldsymbol{P}_t = \boldsymbol{P}_{t-1} - \frac{\boldsymbol{P}_{t-1}\boldsymbol{X}_t\boldsymbol{X}_t^T\boldsymbol{P}_{t-1}}{1+\boldsymbol{X}_t^T\boldsymbol{P}_{t-1}\boldsymbol{X}_t}. \tag{10.23}$$

These difference equations have to be given initial values $\widehat{\boldsymbol{\theta}}_0$ and $\boldsymbol{P}_0$. The choice of such initial values is discussed in Section **??**. Note that the algorithm is based on the assumption that the parameter vector $\boldsymbol{\theta}$ is independent of time which e.g. means that $\boldsymbol{P}_t$ decreases towards zero as $1/t$ (as long as the system is properly excited). This is quite consistent with producing exactly the same estimate as off-line least squares estimation, apart from the effects of the choice of initial values.

The RLS algorithm for parameter estimation in the system (10.3), may also be given a Kalman filter interpretation, cf. Section 10.6.

10.2.1 Time-varying parameters

The parameters in the system, equation (10.3) may be described as time varying. If such time-variations are known, they shall preferably be included in the estimation algorithm, if it e.g. may be assumed that a parameter varies periodically, the model should reflect the periodicity, leaving e.g. the amplitude of the variations to be estimated. This explicite modelling of the time-variations of the parameters will be further treated in Section 10.6. However in the by far most prevalent cases, the variations are unknown, and have to be handled in an ad hoc manner. In such cases, the observed time variations might be considered as generated by changing operating conditions

for the system, in which case the connotation of $\boldsymbol{\theta}$ as a parameter may be doubted. However it is a commonplace notation in the literature and it will be used also here.

The estimation of time varying parameters in the RLS context is accomplished via extension of the basic algorithm with either of the two dominating techniques, linear or exponential forgetting. The linear forgetting is closely related to modelling of the parameter variations, and hence further discussed in Section 10.6.

10.2.2 Recursive estimation using a forgetting factor

The *forgetting factor* or *exponential forgetting* technique is formally a simple extension of the basic RLS algorithm with an ad hoc measure to handle time-varying parameter via discounting of old prediction errors in the loss to be minimized. Hence in this respect the technique is similar to exponential smoothing (cf. [Madsen 1995*b*]), but with a more general description of the underlying system. The properties of the algorithm with forgetting factor is however drastically different, the most important difference being that the sequence of estimates no longer converges to a point in parameter space, but instead can be described as a stochastic process. This process can be shown to be non-Gaussian and the parameter estimation error will in general have a non-zero mean, cf. [Arvastson, Olsson & Holst 1995], [Lindoff & Holst 1995*b*] and [Lindoff & Holst 1995*a*].

In order to derive an algorithm for exponential forgeting in least squares estimation, let us consider the following *weighted least squares estimator*

$$\widehat{\boldsymbol{\theta}}_t = \arg\min_{\boldsymbol{\theta}} S_t(\boldsymbol{\theta}), \tag{10.24}$$

where

$$S_t(\boldsymbol{\theta}) = \sum_{s=1}^{t} \beta(t,s)(Y_s - \boldsymbol{X}_s^T\boldsymbol{\theta})^2. \tag{10.25}$$

Assume that the sequence of weights satisfies

$$\beta(t,s) = \lambda(t)\beta(t-1,s); \qquad 1 \le s \le t-1 \tag{10.26}$$

$$\beta(t,t) = 1, \tag{10.27}$$

which means that

$$\beta(t,s) = \prod_{j=s+1}^{t} \lambda(j), \tag{10.28}$$

i.e. the weight in the computation of the parameter estimate at time t of the squared residual at time s is computed as the product all the intermediate weighting factors.

The solution this weighted least squares problem is then

$$\widehat{\boldsymbol{\theta}}_t = \boldsymbol{R}_t^{-1}\boldsymbol{h}_t, \tag{10.29}$$

where

$$\boldsymbol{R}_t = \sum_{s=1}^{t} \beta(t,s)\boldsymbol{X}_s\boldsymbol{X}_s^T, \quad \boldsymbol{h}_t = \sum_{s=1}^{t} \beta(t,s)\boldsymbol{X}_s Y_s, \tag{10.30}$$

and the *updating* equations are

$$\boldsymbol{R}_t = \lambda(t)\boldsymbol{R}_{t-1} + \boldsymbol{X}_t\boldsymbol{X}_t^T \tag{10.31}$$

$$\boldsymbol{h}_t = \lambda(t)h_{t-1} + \boldsymbol{X}_t Y_t. \tag{10.32}$$

As previously the updating equation for the estimate of the parameter vector $\boldsymbol{\theta}$ may be found by using the updates of $\boldsymbol{R}_t$ and $\boldsymbol{h}_t$:

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_t &= \boldsymbol{R}_t^{-1}\boldsymbol{h}_t = \boldsymbol{R}_t^{-1}\left[\lambda(t)\boldsymbol{h}_{t-1} + \boldsymbol{X}_t Y_t\right] \\
&= \boldsymbol{R}_t^{-1}\left[\lambda(t)\boldsymbol{R}_{t-1}\widehat{\boldsymbol{\theta}}_{t-1} + \boldsymbol{X}_t Y_t\right] \\
&= \boldsymbol{R}_t^{-1}\left[\boldsymbol{R}_t\widehat{\boldsymbol{\theta}}_{t-1} - \boldsymbol{X}_t\boldsymbol{X}_t^T\widehat{\boldsymbol{\theta}}_{t-1} + \boldsymbol{X}_t Y_t\right] \\
&= \widehat{\boldsymbol{\theta}}_{t-1} + \boldsymbol{R}_t^{-1}\boldsymbol{X}_t\left[Y_t - \boldsymbol{X}_t^T\widehat{\boldsymbol{\theta}}_{t-1}\right].
\end{aligned} \tag{10.33}$$

Thus, in summary we have *the RLS method with a forgetting structure*

$$\widehat{\boldsymbol{\theta}}_t = \widehat{\boldsymbol{\theta}}_{t-1} + \boldsymbol{R}_t^{-1}\boldsymbol{X}_t\left[Y_t - \boldsymbol{X}_t^T\widehat{\boldsymbol{\theta}}_{t-1}\right] \tag{10.34}$$

$$\boldsymbol{R}_t = \lambda(t)\boldsymbol{R}_{t-1} + \boldsymbol{X}_t\boldsymbol{X}_t^T. \tag{10.35}$$

If $\lambda(t) = \text{const} = \lambda$, with $0 < \lambda < 1$, then λ is called *the forgetting factor*. The squared errors in equation (10.25) are then weighted exponentially, and we may introduce the *effective number of observations* as

$$T_0 = \sum_{i=0}^{\infty}\lambda^i = \frac{1}{1-\lambda}. \tag{10.36}$$

Typical values for λ ranges from 0.95 to 0.999. Often λ is chosen as a result of experimentation with some reasonable alternatives. However, it may also be possible to minimize the loss function above also as a function of the forgetting factor, i.e.

$$\lambda^* = \arg\min_{\lambda}\left[\min_{\boldsymbol{\theta}} S_N(\boldsymbol{\theta})\right]. \tag{10.37}$$

A further analysis along this track shows that the optimal value of the forgetting factor is a well-defined compromize between the rate of variation of the parameters and the covariance matrix for the regressor vector, see [**?**] for further details.

10.2.3 Convergence properties of RLS with exponential forgetting

Since the recursive version of the least squares algorithm without forgetting factor, apart from the effect of the choice of initial values for the difference equations in equation (10.2.2) is a direct reformulation of the off-line version of the LS algorithm, the convergence charactristics are the same. However, the introduction of the forgetting factor make the an asymptotic analysis more complicate. In case of estimation of a constant parameter, the analysis leads to the convergence results being expressed in terms of possibly stationary distributions of the sequence of parameter estimates. In case of first order autoregression, the limiting distribution can be exactly computed, cf. [**?**]. For higher order systems the analysis is confined to first and second moments in the limiting distribution, to be discussed below.

Consider the *first order system* with constant parameter driven by Gaussian white noise

$$\begin{aligned} y_t &= a y_{t-1} + e_t; \qquad y_0 = 0 \\ e_t &\in N(0, \sigma^2), \end{aligned} \tag{10.38}$$

and suppose that the parameter is to be estimated using recursive least squares estimation with exponential forgetting. The resulting distribution is non-Gaussian and has to be computed numerically. It depends on the parameter to be estimated, a, and the forgetting factor, λ. In figure the time evolution of the density for the estimate of the parameter in the model above is shown. The density is obviously not Gaussian, furthermore its skewness increases with decreasing values of the forgetting factor, λ, but the limiting distribution does not seem to have any mass outside the unit circle.

The estiamates given by the forgetting factor algorithm are biased, and in figure 10.2 the exact asymptotic bias is shown for varying forgetting factor and true parameter value. Note that the bias is negative, showing that the estimated system is assumed to be faster than the true one.

By using a Taylor expansion of the expressions for the parameter estimate, approximative asymptotic expressions for the bias ($\beta(\hat{a})$) and variance ($V(\hat{a})$) for the parameter estimates may be computed to give

$$\beta(\hat{a}) = -\frac{2a\lambda(1-\lambda)(1-a^2)}{(1-a^2\lambda)(1+\lambda)} \tag{10.39}$$

$$V(\hat{a}) = \frac{(1-\lambda)(1-a^2)}{1+\lambda}. \tag{10.40}$$

A comparison between the exact and the approximative expressions for the bias for two different values of the a-parameter is shown in figure 10.3, indicating that the quality of the approximation improves with growing forgetting factor.

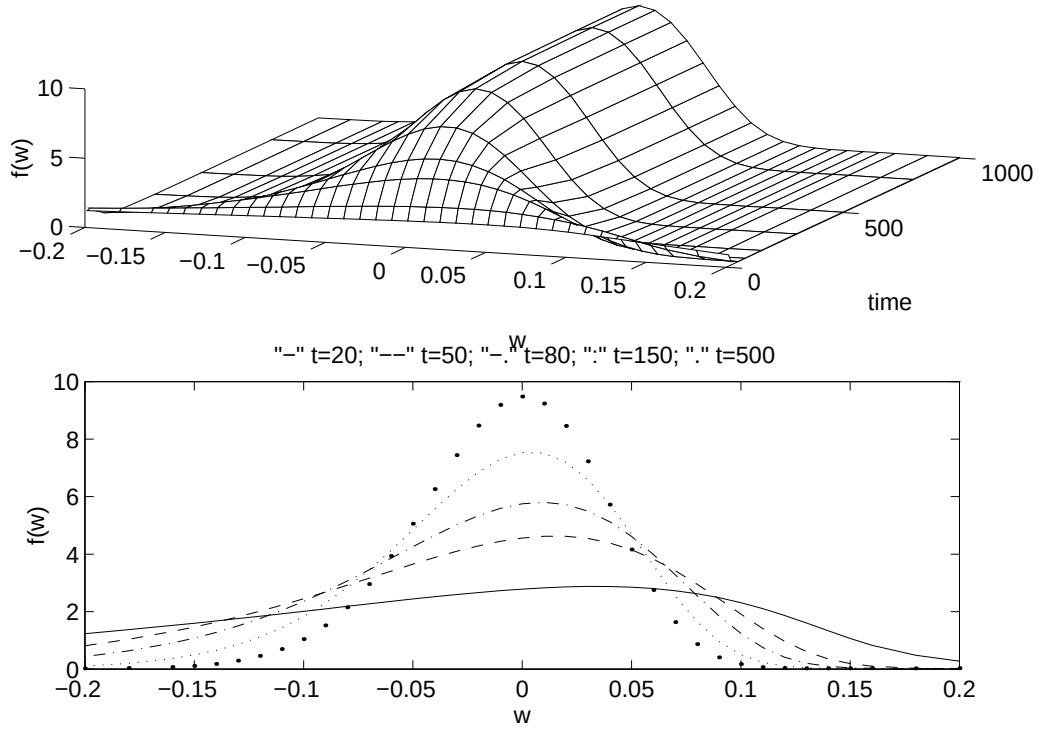

Figure 10.1: Density function for the estimates of the a-parameter in the AR(1)-process in equation 10.38 for varying forgetting factor, λ, and varying estimation time.

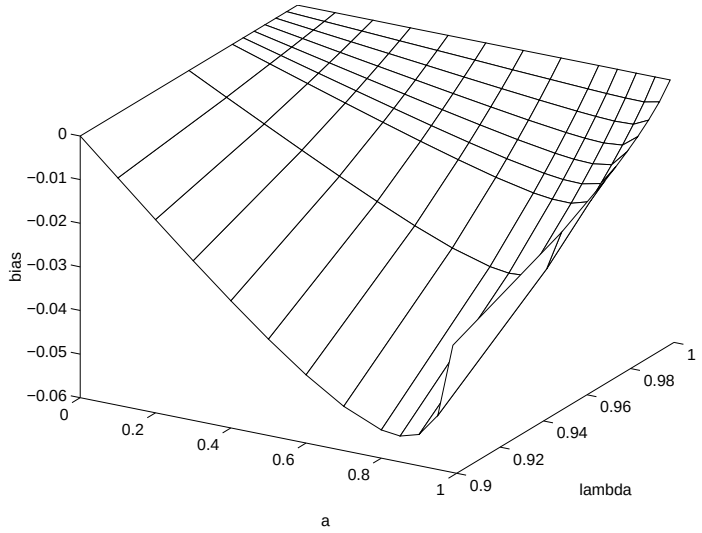

Figure 10.2: Exact asymptotic bias for the estimate of the a-parameter in the AR(1)-process in equation 10.38 for varying forgetting factor, λ, and varying true parameter value.

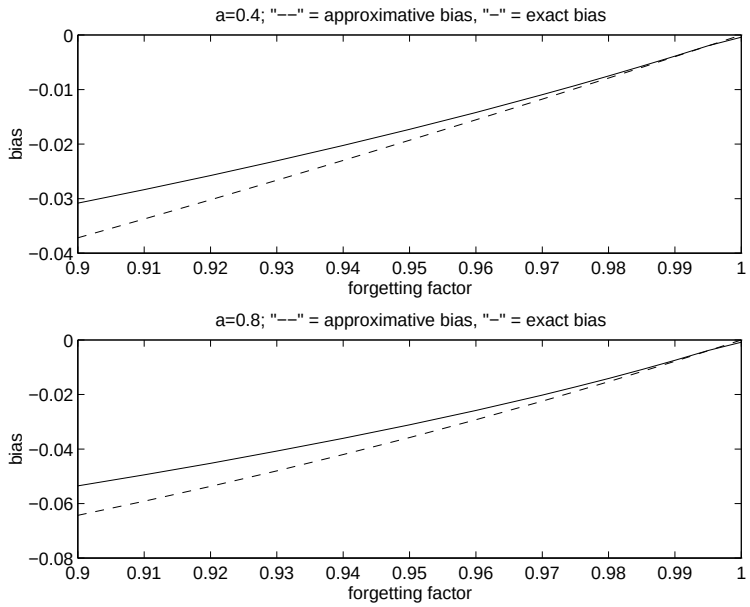

Figure 10.3: Exact and approximative asymptotic bias for the estimate of the a-parameter in the AR(1)-process in equation 10.38 for varying forgetting factor, λ, and varying true parameter value.

The bias and variance of the parameter estimates are also computable for higher order AR- or ARX-processes, cf. [] and [].

Haer skall vi diskutera de resulat som finns i Scand J of Stat och fraan Int J of Adaptive Control and Signal Processing

10.2.4 Variable forgetting

A slightly different procedure leading to a time-varying forgetting factor is based on the assumption that the 'loss function' should be kept constant, i.e. the time-varying forgetting factor, λ_t should be chosen from

$$\begin{aligned} S_t &= \lambda_t S_{t-1} + \varepsilon_t^2 \\ S_t(\widehat{\boldsymbol{\theta}}_t) &= S_{t-1}(\widehat{\boldsymbol{\theta}}_{t-1}) = \cdots = S_0 \end{aligned}$$

It is customary to write $S_0 = N_0\sigma^2$, where N_0 is interpreted as an equivalent horizon, during which the parameters may be assumed to be reasonably constant. In practice N_0, and hence S_0 is a tuning factor.

The idea was proposed in the paper [Fortesque, Kershenbaum & Ydstie 1981]. The forgetting structure which satisfies the above is given by the solution to a second order

equation. The interesting solution is

$$\lambda(t) \simeq 1 - \frac{\epsilon_t^2}{S_0\left[1 + \boldsymbol{X}_t^T \boldsymbol{P}(t-1)\boldsymbol{X}_t\right]}. \tag{10.41}$$

Note the feedback from the prediction error, which means that in case of large prediction errors, interpreted as bad tracking of the parameters, old observations are given decreasing weight, and the last observations determines the parameter estimates. Hence, an algorithm with this type of changing forgetting factor is in principle able to follow parameter variations with different rates of change. However, in case of very fast parameter changes, detector algorithms should be preferred, cf. Section **??**. Furthermore, in case of varying rate of time-variation of the parameters, this method leads to changing forgetting in the entire parameter space, which might be deteriorating the performance of the algorithm in cases when the parameters are not varying homogeneously, or in cases when the excitation of the parameter estimation is nonhomogeneous. Such cases will be treated below, in connection with selective forgetting.

10.2.5 Selective forgetting

10.3 Recursive pseudo-linear regression (RPLR)

Until now the models considered have been linear in the parameters. In the following methods which allow for non-linearity in the parameters will be formulated. These methods are all based on the least squares criterion and a Newton-Raphson solution. In the classical control theory literature, where the methods were first developed, only linear transfer function models (ARMAX, Box-Jenkins, etc.) were considered. In the following a more general treatment, which also allows for non-linear models, will be given.

Let us first consider *pseudo-linear models*, where the one-step prediction can be written

$$\widehat{Y}_{t|t-1}(\boldsymbol{\theta}) = \boldsymbol{X}_t^T(\boldsymbol{\theta})\boldsymbol{\theta} \tag{10.42}$$

This will be called a *pseudo-linear regression*, and the formulation makes it possible to include also e.g. the ARMA model.

At time t the parameter estimate is

$$\widehat{\boldsymbol{\theta}}_t = \arg\min_{\boldsymbol{\theta}} S_t(\boldsymbol{\theta}) \tag{10.43}$$

where

$$\begin{aligned} S_t(\boldsymbol{\theta}_t) &= \sum_{s=1}^{t} \beta(t,s)\epsilon_s^2(\boldsymbol{\theta}_t) = \sum_{s=1}^{t} \beta(t,s)(Y_s - \widehat{Y}_{s|s-1}(\boldsymbol{\theta}_t))^2 \\ &= \sum_{s=1}^{t} \beta(t,s)(Y_s - \boldsymbol{X}_s^T(\boldsymbol{\theta}_t)\boldsymbol{\theta}_t)^2 \qquad (10.44) \\ &= \lambda(t)S_{t-1}(\boldsymbol{\theta}_t) + (Y_t - \boldsymbol{X}_t^T(\boldsymbol{\theta}_t)\boldsymbol{\theta}_t)^2 \qquad (10.45) \end{aligned}$$

Let us first assume that S_t is quadratic in $\boldsymbol{\theta}$.

If S_t is quadratic in $\boldsymbol{\theta}$ then the minimum is obtained after a single Newton-Raphson iteration:

$$\widehat{\boldsymbol{\theta}}_t = \widehat{\boldsymbol{\theta}}_{t-1} - \left[\boldsymbol{H}_t(\widehat{\boldsymbol{\theta}}_{t-1})\right]^{-1} \nabla_{\boldsymbol{\theta}} S_t(\widehat{\boldsymbol{\theta}}_{t-1}) \qquad (10.46)$$

where $\boldsymbol{H}$ is the Hessian matrix, which under the assumption is independent of $\boldsymbol{\theta}$.

$$\boldsymbol{H}_t = \lambda(t)\boldsymbol{H}_{t-1} + 2\boldsymbol{X}_t\boldsymbol{X}_t^T \qquad (10.47)$$

and since $\nabla S_{t-1}(\widehat{\boldsymbol{\theta}}_{t-1}) = \mathbf{0}$ it follows that

$$\nabla S_t(\widehat{\boldsymbol{\theta}}_{t-1}) = -2\boldsymbol{X}_t\left[Y_t - \boldsymbol{X}_t^T\widehat{\boldsymbol{\theta}}_{t-1}\right] \qquad (10.48)$$

By putting $\boldsymbol{R}_t = \frac{1}{2}\boldsymbol{H}_t$ we find (again) the RLS algorithm.

Using the recursions without the assumption that S_t is quadratic in $\boldsymbol{\theta}$ we obtain *the recursive pseudo-linear regression*:

$$\widehat{\boldsymbol{\theta}}_t = \widehat{\boldsymbol{\theta}}_{t-1} + \boldsymbol{R}_t^{-1}\boldsymbol{X}_t\left[Y_t - \boldsymbol{X}_t^T\widehat{\boldsymbol{\theta}}_{t-1}\right] \qquad (10.49)$$

$$\boldsymbol{R}_t = \lambda(t)\boldsymbol{R}_{t-1} + \boldsymbol{X}_t\boldsymbol{X}_t^T \qquad (10.50)$$

In the control theory literature the pseudo-linear regression (PLR) and the recursive pseudo-linear regression (RPLR) are related to the so-called *L-structure model*:

$$\gamma(B)Y_t = \frac{\omega(B)}{\rho(B)}u_t + \frac{\theta(B)}{\phi(B)}\epsilon_t \qquad (10.51)$$

10.3.1 Recursive extended least squares (RELS)

For ARMAX models, i.e. models of the form

$$\phi(B)Y_t = \omega(B)u_t + \theta(B)\epsilon_t \tag{10.52}$$

the RPLS procedure is also known as *recursive extended least squares* (RELS). In the non-recursive version the method is called extended least squares (ELS).

It can be shown, see [Ljung & Söderström 1983] that a sufficient condition for ELS to converge to the true estimates is that

$$\mathrm{Re}\left[\frac{1}{\boldsymbol{\theta}_0(\mathrm{e}^{i\omega})}\right] \geq \frac{1}{2}, \quad \forall\omega, \quad -\pi \leq \omega \leq \pi \tag{10.53}$$

It can also be shown that there are stationary and invertible systems for which the RPLS do not converge, examples may be found in [Holst 1979] and [Stoica, Holst & Söderström 1982].

Example 10.1 (ARMAX model in pseudo linear form). Let us illustrate how the ARMAX model is written in the pseudo linear form

Introduce

$$\boldsymbol{\theta} = (\phi_1, \cdots, \phi_p, \omega_1, \cdots, \omega_s, \theta_1, \cdots, \theta_q)^T \tag{10.54}$$

Then the one-step predictions can be written as

$$\widehat{Y}_{t|t-1}(\boldsymbol{\theta}) = [1 - \phi(B)]Y_t + \omega(B)u_t + [\theta(B) - 1]\epsilon_t(\boldsymbol{\theta}) \tag{10.55}$$

where the prediction error

$$\epsilon_t(\boldsymbol{\theta}) = Y_t - \widehat{Y}_{t|t-1}(\boldsymbol{\theta}) \tag{10.56}$$

depends on $\boldsymbol{\theta}$.

By introducing

$$\boldsymbol{X}_t(\boldsymbol{\theta}) = (-Y_{t-1}, \cdots, -Y_{t-p}, u_{t-1}, \cdots u_{t-s}, \epsilon_{t-1}(\boldsymbol{\theta}), \cdots, \epsilon_{t-q}(\boldsymbol{\theta}))^T \tag{10.57}$$

the one-step prediction is written on the pseudo-linear form

$$\widehat{Y}_{t|t-1}(\boldsymbol{\theta}) = \boldsymbol{X}_t^T(\boldsymbol{\theta})\boldsymbol{\theta} \tag{10.58}$$

And the parameters can now be estimated using the RPLR method. ♦

10.4 Recursive prediction error method (RPEM)

The formulation will be kept in a form which allow for recursive estimation of non-linear models. In Section 10.5 results and comments related to the typical control theory literature on the subject will be provided.

It is assumed that

$$Y_t = \widehat{Y}_{t|t-1} + \epsilon_t \tag{10.59}$$

where $\widehat{Y}_{t|t-1}$ is the one-step prediction provided by the (possibly non-linear) model.

Again the estimate at time t is

$$\widehat{\boldsymbol{\theta}}_t = \arg\min S_t(\boldsymbol{\theta})$$

where

$$S_t(\boldsymbol{\theta}_t) = \sum_{s=1}^{t} \beta(t,s)\epsilon_s^2(\boldsymbol{\theta}_t) = \sum_{s=1}^{t} \beta(t,s)(Y_s - \widehat{Y}_{s|s-1}(\boldsymbol{\theta}_t))^2 \tag{10.60}$$

The gradient is

$$\nabla_{\boldsymbol{\theta}} S_t(\boldsymbol{\theta}_t) = -2\sum_{s=1}^{t} \beta(t,s)\boldsymbol{\psi}_s(\boldsymbol{\theta}_t)\epsilon_s(\boldsymbol{\theta}_t) \tag{10.61}$$

where

$$\boldsymbol{\psi}_s(\boldsymbol{\theta}_t) = \nabla_{\boldsymbol{\theta}} \widehat{Y}_{s|s-1}(\boldsymbol{\theta}_t). \tag{10.62}$$

Correspondingly the Hessian is

$$\boldsymbol{H}_t(\boldsymbol{\theta}) = 2\sum_{s=1}^{t} \beta(t,s)\boldsymbol{\psi}_s(\boldsymbol{\theta})\boldsymbol{\psi}_s^T(\boldsymbol{\theta}) - 2\sum_{s=1}^{t} \beta(t,s)\nabla_{\boldsymbol{\theta}}\boldsymbol{\psi}_s(\boldsymbol{\theta})\epsilon_s(\boldsymbol{\theta}) \tag{10.63}$$

Close to the true value of $\boldsymbol{\theta}$ the last term is approximately zero.

Hence

$$\begin{aligned} \boldsymbol{H}_t(\boldsymbol{\theta}) &= 2\lambda(t)\sum_{s=1}^{t-1} \beta(t-1,s)\boldsymbol{\psi}_s(\boldsymbol{\theta})\boldsymbol{\psi}_s^T(\boldsymbol{\theta}) + 2\boldsymbol{\psi}_t(\boldsymbol{\theta})\boldsymbol{\psi}_t^T(\boldsymbol{\theta}) \\ &= \lambda(t)\boldsymbol{H}_{t-1}(\boldsymbol{\theta}) + 2\boldsymbol{\psi}_t(\boldsymbol{\theta})\boldsymbol{\psi}_t^T(\boldsymbol{\theta}) \end{aligned} \tag{10.64}$$

Again using Newton-Raphson:

$$\widehat{\boldsymbol{\theta}}_t = \widehat{\boldsymbol{\theta}}_{t-1} - \left[\boldsymbol{H}_t(\widehat{\boldsymbol{\theta}}_{t-1})\right]^{-1} \nabla_{\boldsymbol{\theta}} S_t(\widehat{\boldsymbol{\theta}}_{t-1}) \tag{10.65}$$

If follows that

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} S_t(\widehat{\boldsymbol{\theta}}_t) &= -2\lambda(t)\sum_{s=1}^{t-1}\beta(t-1,s)\boldsymbol{\psi}_s(\boldsymbol{\theta}_t)\epsilon_s(\boldsymbol{\theta}_t) - 2\boldsymbol{\psi}_t(\boldsymbol{\theta}_t)\epsilon_t(\boldsymbol{\theta}_t) \\ &= -2\lambda(t)\nabla_{\boldsymbol{\theta}} S_{t-1}(\boldsymbol{\theta}_t) - 2\boldsymbol{\psi}_t(\boldsymbol{\theta}_t)\epsilon_t(\boldsymbol{\theta}_t) \qquad (10.66)\end{aligned}$$

Let us assume that $\widehat{\boldsymbol{\theta}}_{t-1}$ is minimizing S_{t-1}, then

$$\nabla_{\boldsymbol{\theta}} S_t(\widehat{\boldsymbol{\theta}}_{t-1}) = -2\boldsymbol{\psi}_t(\widehat{\boldsymbol{\theta}}_{t-1})\epsilon_t(\widehat{\boldsymbol{\theta}}_{t-1}) \qquad (10.67)$$

Again putting $\boldsymbol{R}_t = \frac{1}{2}\boldsymbol{H}_t$ we obtain *the RPEM algorithm*

$$\epsilon_t(\widehat{\boldsymbol{\theta}}_{t-1}) = Y_t - \widehat{Y}_{t|t-1}(\widehat{\boldsymbol{\theta}}_{t-1}) \qquad (10.68)$$

$$\widehat{\boldsymbol{\theta}}_t = \widehat{\boldsymbol{\theta}}_{t-1} + \boldsymbol{R}_t^{-1}(\widehat{\boldsymbol{\theta}}_{t-1})\boldsymbol{\psi}_t(\widehat{\boldsymbol{\theta}}_{t-1})\epsilon_t(\widehat{\boldsymbol{\theta}}_{t-1}) \qquad (10.69)$$

$$\boldsymbol{R}_t(\widehat{\boldsymbol{\theta}}_{t-1}) = \lambda(t)\boldsymbol{R}_{t-1}(\widehat{\boldsymbol{\theta}}_{t-1}) + \boldsymbol{\psi}_t(\widehat{\boldsymbol{\theta}}_{t-1})\boldsymbol{\psi}_t^T(\widehat{\boldsymbol{\theta}}_{t-1}) \qquad (10.70)$$

In the above equations we have not stated how $\boldsymbol{\psi}_t(\boldsymbol{\theta})$ is calculated. For linear models $\boldsymbol{\psi}_t(\boldsymbol{\theta})$ is obtained by a linear filtering of the observed signals. This will be illustrated in the subsection below considering the ARMAX model.

In the control theory literature the RPEM is related to the L-structure model (10.51).

10.4.1 Recursive maximum likelihood (RML)

For ARMAX models, i.e. models of the form

$$\phi(B)Y_t = \omega(B)u_t + \theta(B)\epsilon_t \qquad (10.71)$$

the RPEM procedure is also known in the engineering and control theory literature as *recursive maximum likelihood* (RML).

The following example outlines in detail the RML procedure. Furthermore it illustrates how $\boldsymbol{\psi}_t(\boldsymbol{\theta})$ is calculated.

Example 10.2 (Recursive Maximum Likelihood (RML)). Consider the ARMAX model (10.71), which alternatively can be written

$$Y_t = \frac{\omega(B)}{\phi(B)}u_t + \frac{\theta(B)}{\phi(B)}\epsilon_t = H_1(B)u_t + H_2(B)\epsilon_t$$

That is

$$\begin{aligned}\widehat{Y}_{t|t-1} &= H_1(B)u_t + [H_2(B)-1]\,\epsilon_t \\ &= H_1(B)u_t + \left[1 - H_2^{-1}(B)\right] H_2(B)\epsilon_t \\ &= H_1(B)u_t + \left[1 - H_2^{-1}(B)\right] [Y_t - H_1(B)u_t] \\ &= H_2^{-1}H_1(B)u_t + \left[1 - H_2^{-1}(B)\right] Y_t \end{aligned} \tag{10.72}$$

Hence it follows that

$$\widehat{Y}_{t|t-1} = \frac{\omega(B)}{\theta(B)}u_t + \left[1 - \frac{\phi(B)}{\theta(B)}\right] Y_t \tag{10.73}$$

or

$$\theta(B)\widehat{Y}_{t|t-1} = \omega(B)u_t + [\theta(B) - \phi(B)]Y_t \tag{10.74}$$

Now we see that

$$\theta(B)\frac{\partial \widehat{Y}_{t|t-1}}{\partial \phi_k} = -B^k Y_t \tag{10.75}$$

$$\theta(B)\frac{\partial \widehat{Y}_{t|t-1}}{\partial \omega_k} = B^k u_t \tag{10.76}$$

$$B^k\widehat{Y}_{t|t-1} + \theta(B)\frac{\partial \widehat{Y}_{t|t-1}}{\partial \theta_k} = B^k Y_t \tag{10.77}$$

Since $Y_t - \widehat{Y}_{t|t-1} = \epsilon_t$ we are able to collect the above in

$$\theta(B)\boldsymbol{\psi}_t(\boldsymbol{\theta}) = \boldsymbol{X}_t(\boldsymbol{\theta}) \tag{10.78}$$

where

$$\boldsymbol{X}_t(\boldsymbol{\theta}) = (-Y_{t-1}, \cdots, -Y_{t-p}, u_{t-1}, \cdots, u_{t-s}, \epsilon_{t-1}(\boldsymbol{\theta}), \cdots, \epsilon_{t-q}(\boldsymbol{\theta}))^T$$

♦

10.5 Summary and comments on least squares recursive methods

Obviously all of the methods considered so far consider the *least squares criterion*

$$S_t(\boldsymbol{\theta}_t) = \sum_{s=1}^{t} \epsilon_s^2(\boldsymbol{\theta}_t) = \sum_{s=1}^{t} (Y_s - \widehat{Y}_{s|s-1}(\boldsymbol{\theta}_t))^2 \tag{10.79}$$

and *the model* is one of the following:

$$\begin{aligned} Y_t &= \boldsymbol{X}_t^T\boldsymbol{\theta} \text{ (linear model)} \\ Y_t &= \boldsymbol{X}_t^T(\boldsymbol{\theta})\boldsymbol{\theta} \text{ (pseudo-linear model)} \\ Y_t &= X_t(\boldsymbol{\theta}) \text{ (non-linear model)} \end{aligned}$$

Finally, for all methods the minimization can be considered as obtained by a single *Newton-Raphson iteration*

$$\widehat{\boldsymbol{\theta}}_t = \widehat{\boldsymbol{\theta}}_{t-1} - \left[\boldsymbol{H}_t(\widehat{\boldsymbol{\theta}}_{t-1})\right]^{-1} \nabla_{\boldsymbol{\theta}} S_t(\widehat{\boldsymbol{\theta}}_{t-1}) \tag{10.80}$$

Let us now illustrate the difference between linear and non-linear models:

10.5.1 Linear model

If the model is linear then

$$\nabla_{\boldsymbol{\theta}} S_t(\boldsymbol{\theta}) = -2\sum_{s=1}^{t} \boldsymbol{X}_s \epsilon_s(\boldsymbol{\theta}) \tag{10.81}$$

$$\boldsymbol{H}_t(\boldsymbol{\theta}) = \boldsymbol{H}_t = 2\sum_{s=1}^{t} \boldsymbol{X}_s \boldsymbol{X}_s^T \tag{10.82}$$

and the minimum is obtained using only one iteration no matter the value of $\widehat{\boldsymbol{\theta}}_{t-1}$. Traditionally, $\widehat{\boldsymbol{\theta}}_{t-1}$ is put equal to zero, and then $\epsilon_s = Y_s$ in (10.81).

Notice that by writing the sums recursively we have now obtained the RLS, RPLR and RELS algorithm. In the non-recursive version the algorithms are called LS, PLR and ELS, respectively.

10.5.2 Non-linear model

If the model is pseudo-linear or non-linear in θ then

$$\nabla_{\boldsymbol{\theta}} S_t(\boldsymbol{\theta}) = 2\sum_{s=1}^{t} \nabla_{\boldsymbol{\theta}} \epsilon_s(\boldsymbol{\theta}) \epsilon_s(\boldsymbol{\theta}) \tag{10.83}$$

$$= -2\sum_{s=1}^{t} \boldsymbol{\psi}_s(\boldsymbol{\theta}) \epsilon_s(\boldsymbol{\theta}) \tag{10.84}$$

$$\boldsymbol{H}_t(\boldsymbol{\theta}) = 2\sum_{s=1}^{t} \boldsymbol{\psi}_s(\boldsymbol{\theta}) \boldsymbol{\psi}_s^T(\boldsymbol{\theta}) - 2\sum_{s=1}^{t} \nabla_{\boldsymbol{\theta}} \boldsymbol{\psi}_s(\boldsymbol{\theta}) \epsilon_s(\boldsymbol{\theta}) \tag{10.85}$$

$$\approx 2\sum_{s=1}^{t} \boldsymbol{\psi}_s(\boldsymbol{\theta}) \boldsymbol{\psi}_s^T(\boldsymbol{\theta}) \tag{10.86}$$

where

$$\boldsymbol{\psi}_s(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \widehat{Y}_{s|s-1}(\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}} \epsilon_s(\boldsymbol{\theta}) \tag{10.87}$$

In general several Newton-Raphson iteration is needed. However, close to the true value of $\boldsymbol{\theta}$ it is reasonable to assume that S_t is approximately quadratic, and hence only a single iteration is used in practice.

Notice that by writing the sums recursively the RPEM and the RML algorithms are readily obtained.

10.6 Model based parameter tracking

A different approach to handling of time-varying parameters, the model based on parameter tracking, starts with the observation that the RLS estimation algorithm in equations (10.21), (10.22) and (10.23) can be given a Kalman filter interpretation.

Assume we are given the system

$$\begin{aligned} \boldsymbol{\theta}_t &= \boldsymbol{\theta}_{t-1} \\ \boldsymbol{Y}_t &= \boldsymbol{X}_t^T \boldsymbol{\theta} + \boldsymbol{e}_t; \quad \mathrm{V}[\boldsymbol{e}_t] = \boldsymbol{\Sigma}_{e,t} \end{aligned}$$

Here, $\boldsymbol{\theta}_t$ is interpreted as the state vector and $\boldsymbol{X}_t^T$ as the output matrix, which in this representation is time-varying and determined by output and external signals that are known at time $t-1$. Since the Kalman filter computes conditional means and variances at time t given information at time $t-1$, the regression vector for the output calculation is known, and the resulting filter is

Time Update :

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{t|t-1} &= \widehat{\boldsymbol{\theta}}_{t-1|t-1} \\ \boldsymbol{P}_{t|t-1} &= \boldsymbol{P}_{t-1|t-1}\end{aligned}$$

Measurement Update :

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{t|t} &= \widehat{\boldsymbol{\theta}}_{t|t-1} + \boldsymbol{K}_t\boldsymbol{\epsilon}_t \\ \boldsymbol{\epsilon}_t &= \boldsymbol{Y}_t - \boldsymbol{X}_t^T\widehat{\boldsymbol{\theta}}_{t|t-1} \\ \boldsymbol{K}_t &= \boldsymbol{P}_{t|t-1}\boldsymbol{X}_t(\boldsymbol{X}_t^T\boldsymbol{P}_{t|t-1}\boldsymbol{X}_t + \boldsymbol{\Sigma}_{e,t})^{-1} \\ \boldsymbol{P}_{t|t} &= \boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1}\boldsymbol{X}_t(\boldsymbol{X}_t^T\boldsymbol{P}_{t|t-1}\boldsymbol{X}_t + \boldsymbol{\Sigma}_{e,t})^{-1}\boldsymbol{X}_t^T\boldsymbol{P}_{t|t-1} \quad \text{or} \\ \boldsymbol{P}_{t|t}^{-1} &= \boldsymbol{P}_{t|t-1}^{-1} + \boldsymbol{X}_t\boldsymbol{\Sigma}_{e,t}^{-1}\boldsymbol{X}_t^T\end{aligned}$$

The residual is

$$\begin{aligned}\boldsymbol{\epsilon}_{r,t} &= \boldsymbol{Y}_t - \boldsymbol{X}_t^T\widehat{\boldsymbol{\theta}}_{t|t} \\ &= \boldsymbol{\Sigma}_{e,t}(\boldsymbol{X}_t^T\boldsymbol{P}_{t|t-1}\boldsymbol{X}_t + \boldsymbol{\Sigma}_{e,t})^{-1}\boldsymbol{\epsilon}_t.\end{aligned}$$

Note that by writing $\widehat{\boldsymbol{\theta}}_{t|t}$ or $\widehat{\boldsymbol{\theta}}_{t|t-1}$ and $\boldsymbol{P}_{t|t}$ or $\boldsymbol{P}_{t|t-1}$ as $\widehat{\boldsymbol{\theta}}_t$ and $\boldsymbol{P}_t$ respectively, and assuming scalar output, we recover the equations on page 223. Also note that by comparing these equations above with equation (10.10), clearly the matrix $\boldsymbol{P}_t$ in the RLS-algorithm for scalar output may be interpreted as a normalized covariance matrix for the parameter estimation errors, where the normalization factor is the variance of the measurement noise.

In case of time varying parameters this model based on tracking is extended also to cover modelling of the parameter variations. These models should reflect (partially known) variations as well as ad hoc assumptions, cf. eg. [Young 1984] for some examples of these kind of models. The most straightforward extension from the assumption about constancy of the parameters, is to assume these to perform a random walk, rendering the following state space model:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \boldsymbol{v}_t; \quad \mathrm{V}[\boldsymbol{V}_t] = \boldsymbol{\Sigma}_{v,t} \tag{10.88}$$

$$\boldsymbol{Y}_t = \boldsymbol{X}_t^T\boldsymbol{\theta}_t + \boldsymbol{e}_t; \quad \mathrm{V}[\boldsymbol{e}_t] = \boldsymbol{\Sigma}_{e,t} \tag{10.89}$$

Using the Kalman filter on predictive form, we obtain the following algorithm:

$$\widehat{\boldsymbol{\theta}}_{t+1|t} = \widehat{\boldsymbol{\theta}}_{t|t-1} + \boldsymbol{K}_t\left(\boldsymbol{Y}_t - \boldsymbol{X}_t^T\widehat{\boldsymbol{\theta}}_{t|t-1}\right) \tag{10.90}$$

$$\boldsymbol{K}_t = \frac{\boldsymbol{P}_{t|t-1}\boldsymbol{X}_t}{\left(\boldsymbol{X}_t^T\boldsymbol{P}_{t|t-1}\boldsymbol{X}_t + \boldsymbol{\Sigma}_{e,t}\right)} \tag{10.91}$$

$$\boldsymbol{P}_{t+1|t} = \boldsymbol{P}_{t|t-1} + \boldsymbol{\Sigma}_{v,t} - \frac{\boldsymbol{P}_{t|t-1}\boldsymbol{X}_t\boldsymbol{X}_t^T\boldsymbol{P}_{t|t-1}}{\left(\boldsymbol{X}_t^T\boldsymbol{P}_{t|t-1}\boldsymbol{X}_t + \boldsymbol{\Sigma}_{e,t}\right)}, \tag{10.92}$$

where starting values can be selected according to our prior knowledge about the parameter values, as discussed in Section **??**. This way of treating the time-variations is named *linear forgetting*.

Remark.
- The forgetting factor approaches are best suited for identifying non-stationary systems with slowly varying parameters, whereas
- The state space approach can be used for identifying non-stationary systems with more rapidly changing parameters.

▼

Example 10.3 (Relation between air temperature and solar radiation). Consider a model for the relation between the air temperature T and the net radiation at the surface of the ground R. The model is based on basic physical considerations, and it shows that explicit modelling of the parameter variations might be possible and also proves to be fruitful.

By considering basic physics, the following model is found adequate ([**?**]):

$$\phi(B)(T_t - \mu_T) = \omega(B)(R_t - \mu_R) + \epsilon_t \tag{10.93}$$

where the order of ϕ is 3 and the order of ω is 1.

In order to obtain a model which is linear in the parameters we rewrite the model

$$\phi(B)T_t = \omega(B)R_t + d + \epsilon_t \tag{10.94}$$

where $d = \phi(1)\mu_T - \omega(1)\mu_R$.

Now we are able to use the RLS procedure, since the model can be written

$$T_t = \boldsymbol{X}_t^T\boldsymbol{\theta} + \epsilon_t \tag{10.95}$$

where

$$\begin{aligned}\boldsymbol{X}_t &= (-T_{t-1}, -T_{t-2}, -T_{t-3}, u_{t-1}, u_{t-2}, 1)^T\\ \boldsymbol{\theta} &= (\phi_1, \phi_2, \phi_3, \omega_0, \omega_1, d)^T\end{aligned}$$

♦

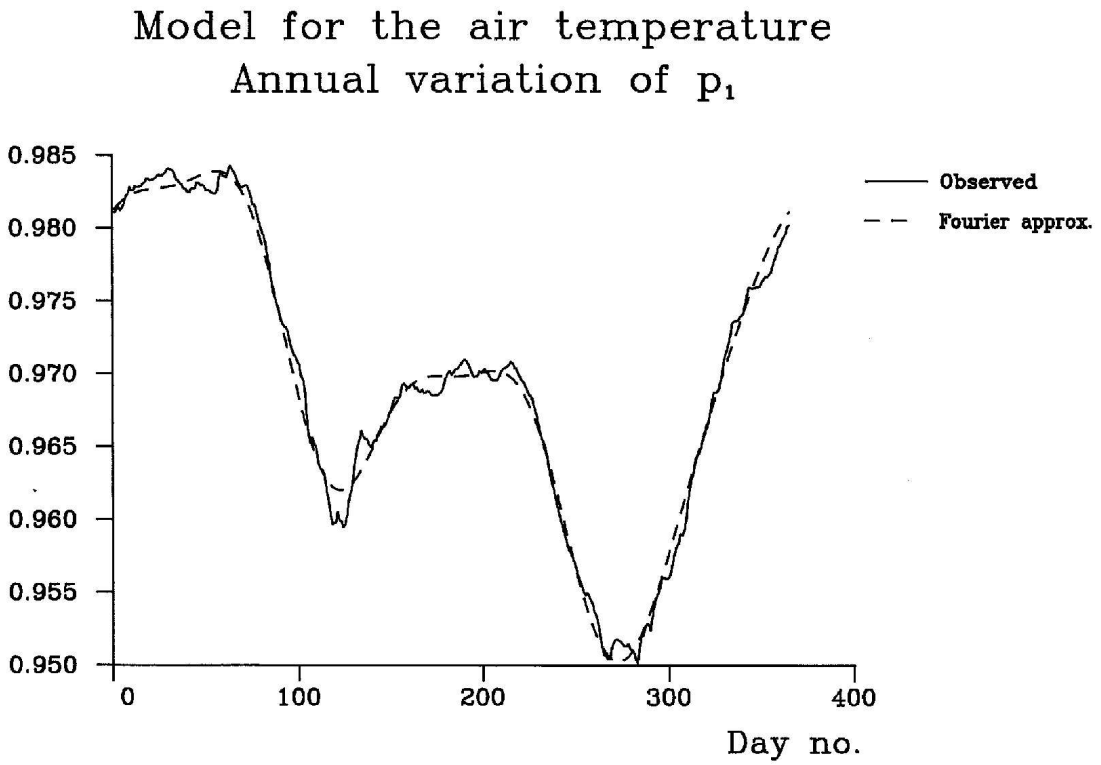

Figure 10.4: Recursive estimation of the parameters in a transfer function model for the outdoor air temperature

10.7 Tracking time-varying coefficient-functions

10.7.1 Introduction

The conditional parametric ARX-model (CPARX-model) is a non-linear model formulated as a linear ARX-model in which the parameters are replaced by smooth, but otherwise unknown, functions of one or more explanatory variables. These functions are called *coefficient-functions*, cf. [**?**] for an extended discussion. For on-line applications it is advantageous to allow the estimates of these coefficient functions be modified as data become available. Furthermore, because the system may change slowly over time, observations should be down-weighted as they become older by using some recursive estimation method with forgetting. Hence, the method to be discussed in this section is a combination of recursive least squares with exponential forgetting, cf. Section 10.2.2 and locally weighted polynomial regression, cf. Section **??**. Here *adaptive estimation* is used to denote, that old observations are down-weighted, i.e. in the sense of *adaptive in time*.

Non-adaptive recursive estimation of a regression function is a related problem, which has been studied by **?** using kernel methods and by **?** using local polynomial regression. Since these methods are non-adaptive a key aspect considered in these papers is how to decrease the bandwidth as new observations become available. This problem do not arise for adaptive estimation since old observations are down-weighted and eventually disregarded as part of the algorithm, even though the combined effect of choice of fortgetting factor and bandwidth has to be considered.

10.7.2 Conditional parametric models and local polynomial estimates

When using a conditional parametric model to model the response y_s the explanatory variables are split in two groups. One group of variables $\mathbf{x}_s$ enter globally through coefficients depending on the other group of variables $\mathbf{u}_s$, i.e.

$$y_s = \mathbf{x}_s^T \boldsymbol{\theta}(\mathbf{u}_s) + e_s, \tag{10.96}$$

where $\boldsymbol{\theta}(\cdot)$ is a vector of coefficient-functions to be estimated and e_s is the noise term. Note that $\mathbf{x}_s$ may contain lagged values of the response. The dimension of $\mathbf{x}_s$ can be quite large, but the dimension of $\mathbf{u}_s$ must be low (1 or 2) for practical purposes [**?**, pp. 83-84].

The functions $\boldsymbol{\theta}(\cdot)$ in (10.96) are estimated at a number of distinct points, *fitting points*, by approximating the functions using polynomials and fitting the resulting linear model

locally to each of these points. To be more specific let $\mathbf{u}$ denote a particular fitting point. Let $\theta_j(\cdot)$ be the j'th element of $\boldsymbol{\theta}(\cdot)$ and let $\mathbf{p}_{d(j)}(\mathbf{u})$ be a column vector of terms in the corresponding d:th-order polynomial evaluated at $\mathbf{u}$. If for instance $\mathbf{u} = [u_1\ u_2]^T$ then $\mathbf{p}_2(\mathbf{u}) = [1\ u_1\ u_2\ u_1^2\ u_1u_2\ u_2^2]^T$. Furthermore, let $\mathbf{x}_s = [x_{1,s} \dots x_{p,s}]^T$. With

$$\mathbf{z}_s^T = \left[x_{1,s}\mathbf{p}_{d(1)}^T(\mathbf{u}_s) \dots x_{j,s}\mathbf{p}_{d(j)}^T(\mathbf{u}_s) \dots x_{p,s}\mathbf{p}_{d(p)}^T(\mathbf{u}_s)\right] \tag{10.97}$$

and

$$\boldsymbol{\phi}_u^T = [\boldsymbol{\phi}_{u,1}^T \dots \boldsymbol{\phi}_{u,j}^T \dots \boldsymbol{\phi}_{u,p}^T], \tag{10.98}$$

where $\boldsymbol{\phi}_{u,j}$ is a column vector of local coefficients at $\mathbf{u}$ corresponding to $x_{j,s}\mathbf{p}_{d(j)}(\mathbf{u}_s)$. The linear model

$$y_s = \mathbf{z}_s^T\boldsymbol{\phi}_u + e_s;\ \ i = 1, \dots, N, \tag{10.99}$$

is then fitted locally to $\mathbf{u}$ using weighted least squares (WLS), i.e.

$$\hat{\boldsymbol{\phi}}(\mathbf{u}) = \underset{\boldsymbol{\phi}_u}{\operatorname{argmin}} \sum_{s=1}^{N} w_u(\mathbf{u}_s)(y_s - \mathbf{z}_s^T\boldsymbol{\phi}_u)^2, \tag{10.100}$$

for which a unique closed-form solution exists provided the matrix with rows $\mathbf{z}_s^T$ corresponding to non-zero weights has full rank. The weights are assigned as

$$w_u(\mathbf{u}_s) = W\left(\frac{||\mathbf{u}_s - \mathbf{u}||}{\hbar(\mathbf{u})}\right), \tag{10.101}$$

where $||\cdot||$ denotes the Euclidean norm, $\hbar(\mathbf{u})$ is the bandwidth used for the particular fitting point, and $W(\cdot)$ is a weighting function taking non-negative arguments. In the examples below we have used

$$W(u) = \begin{cases} (1-u^3)^3, & u \in [0;1) \\ 0, & u \in [1;\infty) \end{cases} \tag{10.102}$$

i.e. the weights are between 0 and 1. The elements of $\boldsymbol{\theta}(\mathbf{u})$ are estimated by

$$\hat{\theta}_j(\mathbf{u}) = \mathbf{p}_{d(j)}^T(\mathbf{u})\,\hat{\boldsymbol{\phi}}_j(\mathbf{u});\ \ j = 1, \dots p, \tag{10.103}$$

where $\hat{\boldsymbol{\phi}}_j(\mathbf{u})$ is the WLS estimate of $\boldsymbol{\phi}_{u,j}$. The estimates of the coefficient-functions obtained as outlined above are called *local polynomial estimates*. For the special case where all coefficient-functions are approximated by constants we use the term local constant estimates.

If $\hbar(\mathbf{u})$ is constant for all values of $\mathbf{u}$ it is denoted a fixed bandwidth. If $\hbar(\mathbf{u})$ is chosen so that a certain fraction α of the observations fulfill $||\mathbf{u}_s - \mathbf{u}|| \leq \hbar(\mathbf{u})$ then α is denoted a nearest neighbour bandwidth. A bandwidth specified according to the nearest neighbour principle is often used as a tool to vary the actual bandwidth with the local density of the data.

Interpolation is used for approximating the estimates of the coefficient-functions for other values of the arguments than the fitting points. This interpolation should only have marginal effect on the estimates. Therefore, it constraints the number and placement of the fitting points. If a nearest neighbour bandwidth is used it is reasonable to select the fitting points according to the density of the data as it is done when using *k-d* trees [**?**, Section 8.4.2]. However, here the approach is to select the fitting points on an equidistant grid and ensure that several fitting points are within the (smallest) bandwidth so that linear interpolation can be applied safely.

10.7.3 Adaptive estimation

In the previous section it was shown that local polynomial estimation can be viewed as local constant estimation in a model (10.99) derived from the original model (10.96). For simplicity the adaptive estimation method is described as a generalization of exponential forgetting. However, more general forgetting methods could also be used.

Using exponential forgetting and assuming observations at time $s = 1, \ldots, t$ are available, the adaptive least squares estimate of the parameters $\boldsymbol{\phi}$ relating the explanatory variables $\mathbf{z}_s$ to the response y_s using the linear model $y_s = \mathbf{z}_s^T \boldsymbol{\phi} + e_s$ is found as

$$\hat{\boldsymbol{\phi}}_t = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \sum_{s=1}^{t} \lambda^{t-s} (y_s - \mathbf{z}_s^T \boldsymbol{\phi})^2, \tag{10.104}$$

where $0 < \lambda < 1$ is the forgetting factor.The estimate can be seen as a local constant approximation in the direction of time. This suggests that the estimator may also be defined locally with respect to some other explanatory variables $\mathbf{u}_t$. If the estimates are defined locally to a fitting point $\mathbf{u}$, the adaptive estimate corresponding to this point can be expressed as

$$\hat{\boldsymbol{\phi}}_t(\mathbf{u}) = \underset{\boldsymbol{\phi}_u}{\operatorname{argmin}} \sum_{s=1}^{t} \lambda^{t-s} w_u(\mathbf{u}_s) (y_s - \mathbf{z}_s^T \boldsymbol{\phi}_u)^2, \tag{10.105}$$

where $w_u(\mathbf{u}_s)$ is a weight on observation s depending on the fitting point $\mathbf{u}$ and $\mathbf{u}_s$ as discussed above.

Remark. In order to illustrate the estimator and its relations to non-parametric regression, consider a special case where $\mathbf{x}_s$ in (10.96) is 1 for all s and $d(1)$ is chosen to be 0. Then it follows from (10.97) that $\mathbf{z}_s = 1$ for all s, and

$$\hat{\phi}_t(\mathbf{u}) = \frac{\sum_{s=1}^{t} \lambda^{t-s} w_u(\mathbf{u}_s) y_s}{\sum_{s=1}^{t} \lambda^{t-s} w_u(\mathbf{u}_s)}. \tag{10.106}$$

For $\lambda = 1$ this is a kernel estimator of $\phi(\cdot)$ in $y_s = \phi(\mathbf{u}_s) + e_s$,why (10.106) is called an adaptive kernel estimator of $\phi(\cdot)$ and the estimator (10.105) may be called an adaptive

local constant estimator of the coefficient-functions $\boldsymbol{\phi}(\cdot)$ in the conditional parametric model $y_s = \mathbf{z}_s^T \boldsymbol{\phi}(\mathbf{u}_s) + e_s$. ▼

Recursive formulation

Following the same arguments as above in Section 10.2.2, the adaptive estimates (10.105) can be found recursively as

$$\hat{\boldsymbol{\phi}}_t(\mathbf{u}) = \hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u}) + w_u(\mathbf{u}_t)\mathbf{R}_{u,t}^{-1}\mathbf{z}_t \left[y_t - \mathbf{z}_t^T \hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u}) \right] \tag{10.107}$$

and

$$\mathbf{R}_{u,t} = \lambda \mathbf{R}_{u,t-1} + w_u(\mathbf{u}_t)\mathbf{z}_t\mathbf{z}_t^T \, . \tag{10.108}$$

Hnece, the previously discussed numerical procedures implementing recursive least squares estimation with exponential forgetting in linear models can be applied, by replacing $\mathbf{z}_t$ and y_t in these procedures with $\mathbf{z}_t\sqrt{w_u(\mathbf{u}_t)}$ and $y_t\sqrt{w_u(\mathbf{u}_t)}$, respectively. Note that $\mathbf{z}_t^T\hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u})$ is a predictor of y_t locally with respect to $\mathbf{u}$ and for this reason it is used in (10.107). To predict y_t a predictor like $\mathbf{z}_t^T\hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u}_t)$ is appropriate.

Modified updating formula

When $\mathbf{u}_t$ is far from the particular fitting point $\mathbf{u}$ it is clear from (10.107) and (10.108) that $\hat{\boldsymbol{\phi}}_t(\mathbf{u}) \approx \hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u})$ and $\mathbf{R}_{u,t} \approx \lambda\mathbf{R}_{u,t-1}$, i.e. old observations are down-weighted without new information becoming available (which is exactly analogue to the covariance blowup problem in Section 10.2.5). Hence, this may result in abruptly changing estimates if $\mathbf{u}$ is not visited regularly, since the matrix $\mathbf{R}$ is decreasing exponentially in this case. Thus, it is suggested to modify (10.108) to ensure that past are observations are weighted down only when new information becomes available, i.e.

$$\mathbf{R}_{u,t} = \lambda v(w_u(\mathbf{u}_t);\lambda)\mathbf{R}_{u,t-1} + w_u(\mathbf{u}_t)\mathbf{z}_t\mathbf{z}_t^T, \tag{10.109}$$

where $v(\cdot\,;\lambda)$ is a nowhere increasing function on $[0;1]$ fulfilling $v(0;\lambda) = 1/\lambda$ and $v(1;\lambda) = 1$. Note that this requires that the weights span the interval ranging from zero to one, as is done e.g. when using the weighting function in equation 10.102. Here, we consider only the linear function $v(w;\lambda) = 1/\lambda - (1/\lambda - 1)w$, for which (10.109) becomes

$$\mathbf{R}_{u,t} = (1 - (1-\lambda)w_u(\mathbf{u}_t))\mathbf{R}_{u,t-1} + \; w_u(\mathbf{u}_t)\mathbf{z}_t\mathbf{z}_t^T. \tag{10.110}$$

It is reasonable to denote

$$\lambda_{eff}^u(t) = 1 - (1-\lambda)w_u(\mathbf{u}_t) \tag{10.111}$$

the *effective forgetting factor* for point $\mathbf{u}$ at time t. Note, however, that λ^u_{eff} then will depend on $\mathbf{u}$ as well as on time, and in case of $\mathbf{u}$ being stochastic, the also the effective forgetting factor is a stochastic variable.

When using (10.109) or (10.110) it is ensured that $\mathbf{R}_{u,t}$ can not become singular if the process $\{\mathbf{u}_t\}$ moves away from the fitting point for a longer period. However, the process $\{\mathbf{z}_t\}$ should be persistently excited as for linear ARX-models. In this case, given the weights, the estimates define a global minimum corresponding to (10.105).

Nearest neighbour bandwidth

Assume that $\mathbf{u}_t$ is a stochastic variable and that the pdf $f(\cdot)$ of $\mathbf{u}_t$ is known and constant over t. Based on a nearest neighbour bandwidth the actual bandwidth can then be calculated for a number of fitting points $\mathbf{u}$ placed within the domain of $f(\cdot)$ and used to generate the weights $w_u(\mathbf{u}_t)$. The actual bandwidth $\hbar(\mathbf{u})$ corresponding to the point $\mathbf{u}$ will be related to the nearest neighbour bandwidth α by

$$\alpha = \int_{\mathbb{D}_u} f(\boldsymbol{\nu}) d\boldsymbol{\nu}, \tag{10.112}$$

where $\mathbb{D}_u = \{\boldsymbol{\nu} \in \mathbb{R}^d \mid ||\boldsymbol{\nu} - \mathbf{u}|| \leq \hbar(\mathbf{u})\}$ is the neighbourhood, d is the dimension of $\mathbf{u}$, and $||\cdot||$ is the Euclidean norm. In applications the density $f(\cdot)$ is often unknown. However, $f(\cdot)$ can be estimated from data, e.g. by the empirical pdf.

Effective number of observations

In order to select an appropriate value for α the effective number of observations used for estimation must be considered. In analogy with the concept of effective number of observations in the pure exponential forgetting case, cf. equation 10.36, and considering λ^u_{eff} as a random variable,

$$\tilde{\eta}_u = \frac{1}{1 - E[\lambda^u_{eff}(t)]} = \frac{1}{(1-\lambda)E[w_u(\mathbf{u}_t)]} \tag{10.113}$$

is a lower bound on the effective number of observations (in the direction of time) corresponding to a fitting point $\mathbf{u}$, in general (10.113) can be considered an approximation, cf. [**?**]. When selecting α and λ it is then natural to require that the number of observations within the bandwidth, i.e. $\alpha\tilde{\eta}_u$, is sufficiently large to justify the complexity of the model and the order of the local polynomial approximations.

As an example consider $u_t \sim N(0,1)$ and $\lambda = 0.99$ where the effective number of observations within the bandwidth, $\alpha\tilde{\eta}_u$, is displayed in Figure 10.5. It is seen that $\alpha\tilde{\eta}_u$

depends strongly on the fitting point u but only moderately on α. When investigating the dependence of $\alpha\tilde{\eta}_u$ on λ and α it turns out that $\alpha\tilde{\eta}_u$ is almost solely determined by λ. In conclusion, for the example considered, the effective forgetting factor $\lambda^u_{eff}(t)$ will be affected by the nearest neighbour bandwidth, so that the effective number of observations within the bandwidth will be strongly dependent on λ, but only weakly dependent on the nearest-neighbour bandwidth, α. The ratio between the rate at which the weights on observations goes to zero in the direction of time and the corresponding rate in the direction of u_t will be determined by α.

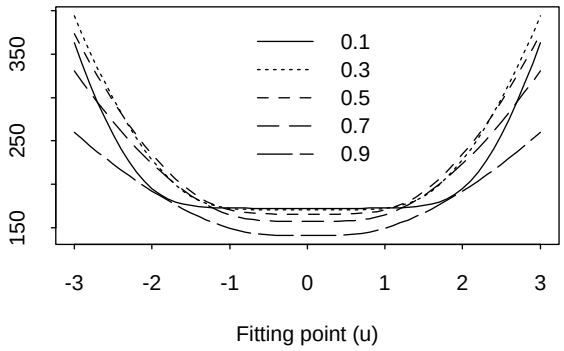

Figure 10.5: *Effective number of observations within the bandwidth (*$\alpha\tilde{\eta}_u(u)$*) for* $\alpha = 0.1, \ldots, 0.9$ *and* $\lambda = 0.99$*.*

As mentioned above and as illustrated by figure 10.5 the effective number of observations behind each of the local approximations depends on the fitting point. This is contrary to the non-adaptive nearest neighbour method, cf. Section 10.7.2.However, if λ is chosen to be non-constant, but instead is varied with the fitting point as $\lambda(\mathbf{u}) = 1 - 1/(T_0 E[w_u(\mathbf{u}_t)])$ then $\tilde{\eta}_u = T_0$. Thus, in that case the effective number of observations within the bandwidth is constant across fitting points. Furthermore, T_0 can be interpreted as the memory time constant. However, we do not further persue this idea, and in the remaining part of the section, λ is not varied across fitting points.

The choice of optimal bandwidth selection in non-parametric function estimation is a commonly treated problem, cf. e.g. [**?**] for a discussion of the non-adaptive case. In the current algorithm optimal values of the forgetting factor as well as of the bandwidth could be considered, cf. [**?**] for a treatment of this topic.

In this section local polynomial approximations in the direction of time is not considered. Such a method is proposed for usual ARX-models by **?**. This method can be combined with the method described here and will result in local polynomial approximations where cross-products between time and the conditioning variables ($\mathbf{u}_t$) are excluded.

Summary of the method

The algorithm will briefly summarized in this section. It is assumed that at each time step t measurements of the output y_t and the two sets of inputs $\mathbf{x}_t$ and $\mathbf{u}_t$ are received.

Prior to the application of the algorithm a number of fitting points $\mathbf{u}^{(i)};\ i = 1, \ldots, n_{fp}$ in which the coefficient-functions are to be estimated have to be selected. Furthermore the bandwidth associated with each of the fitting points $\hbar^{(i)};\ i = 1, \ldots, n_{fp}$ and the degrees of the approximating polynomials $d(j);\ j = 1, \ldots, p$ have to be selected for each of the p coefficient-functions. For simplicity the degree of the approximating polynomial for a particular coefficient-function will be fixed across fitting points. The forgetting factor λ in equation (10.110) has to be selected. Finally, initial estimates of the coefficient-functions in the model corresponding to local constant estimates, i.e. $\hat{\boldsymbol{\phi}}_0(\mathbf{u}^{(i)})$, must be chosen, as will the initial values of the matrices $\mathbf{R}_{u^{(i)},0}$. In case of no other information, it is common practice to initialize the coefficient functions with zero, and the $\mathbf{R}$-matrices with $\epsilon\mathbf{I}$, where ϵ is a small positive number.

In the following description of the algorithm it will be assumed that $\mathbf{R}_{u^{(i)},t}$ is non-singular for all fitting points. In practice we would just stop updating the estimates if the matrix becomes singular. Under the assumption mentioned the algorithm can be described as:

For each time step t: Loop over the fitting points $\mathbf{u}^{(i)};\ i = 1, \ldots, n_{fp}$ and for each fitting point:

- Construct the explanatory variables corresponding to local constant estimates using (10.97):
  $\mathbf{z}_t^T = [x_{1,t}\mathbf{p}_{d(1)}^T(\mathbf{u}_t) \ldots x_{p,t}\mathbf{p}_{d(p)}^T(\mathbf{u}_t)]$.
- Calculate the weight using (10.101) and (10.102):
  $w_{u^{(i)}}(\mathbf{u}_t) = (1 - (||\mathbf{u}_t - \mathbf{u}^{(i)}||/\hbar^{(i)})^3)^3$, if $||\mathbf{u}_t - \mathbf{u}^{(i)}|| < \hbar^{(i)}$ and zero otherwise.
- Find the effective forgetting factor using (10.111):
  $\lambda_{eff}^{(i)}(t) = 1 - (1 - \lambda)w_{u^{(i)}}(\mathbf{u}_t)$.
- Update $\mathbf{R}_{u^{(i)},t-1}$ using (10.110):
  $\mathbf{R}_{u^{(i)},t} = \lambda_{eff}^{(i)}(t)\mathbf{R}_{u^{(i)},t-1} + w_{u^{(i)}}(\mathbf{u}_t)\mathbf{z}_t\mathbf{z}_t^T$.
- Update $\hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u}^{(i)})$ using (10.107):
  $\hat{\boldsymbol{\phi}}_t(\mathbf{u}^{(i)}) = \hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u}^{(i)}) + w_{u^{(i)}}(\mathbf{u}_t)\mathbf{R}_{u^{(i)},t}^{-1}\mathbf{z}_t\left[y_t - \mathbf{z}_t^T\hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u}^{(i)})\right]$.
- Calculate the updated local polynomial estimates of the coefficient-functions using (10.103):
  $\hat{\theta}_{jt}(\mathbf{u}^{(i)}) = \mathbf{p}_{d(j)}^T(\mathbf{u}^{(i)})\,\hat{\boldsymbol{\phi}}_{j,t}(\mathbf{u}^{(i)});\ \ j = 1, \ldots p$

The algorithm could also be implemented using the matrix inversion lemma.

10.7.4 Simulation and application

The simulations below are performed using the non-linear model

$$y_t = a(t, u_{t-1})y_{t-1} + b(t, u_{t-1})x_t + e_t, \tag{10.114}$$

where $\{x_t\}$ is the input process, $\{u_t\}$ is the process controlling the coefficients, $\{y_t\}$ is the output process, and $\{e_t\}$ is a white noise standard Gaussian process. The coefficient-functions are simulated as

$$a(t,u) = 0.3 + (0.6 - \frac{1.5}{N}t)\ \exp\left(-\frac{(u - \frac{0.8}{N}t)^2}{2(0.6 - \frac{0.1}{N}t)^2}\right)$$

and

$$b(t,u) = 2 - \exp\left(-\frac{(u + 1 - \frac{2}{N}t)^2}{0.32}\right),$$

where $t = 1, \ldots, N$ and $N = 5000$, i.e. $a(t,u)$ ranges from -0.6 to 0.9 and $b(t,u)$ ranges from 1 to 2. The functions are displayed in Figure 10.6. As indicated by the figure both coefficient-functions are based on a Gaussian density in which the mean and variance varies linearly with time.

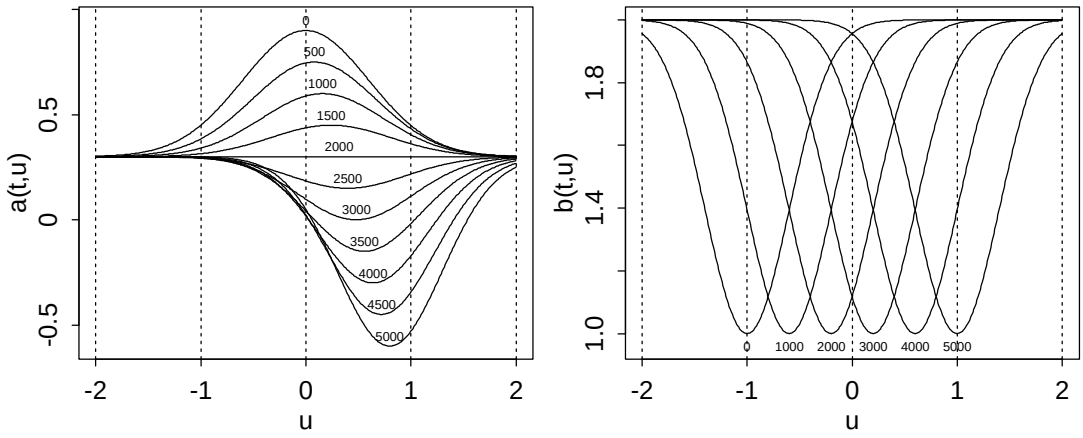

Figure 10.6: *The time-varying coefficient-functions plotted for equidistant points in time as indicated on the plots.*

Local linear ($d(j) = 1$ for all j) adaptive estimates of the functions $a()$ and $b()$ are then found using the proposed procedure with the model

$$y_t = a(u_{t-1})y_{t-1} + b(u_{t-1})x_t + e_t. \tag{10.115}$$

In all cases initial estimates of the coefficient-functions are set to zero and during the initialization the estimates are not updated, for the fitting point considered, until ten observations have received a weight of 0.5 or larger.

Highly correlated input processes

In the simulation presented in this section a strongly correlated $\{\mathbf{u}_t\}$ process is used and also the $\{\mathbf{x}_t\}$ process is quite strongly correlated. This allows us to illustrate various aspects of the method. For less correlated series the performance is much improved. The data are generated using (10.114) where $\{x_t\}$ and $\{u_t\}$ are zero mean $AR(1)$-processes with poles in 0.9 and 0.98, respectively. The variance for both series is one and the series are mutually independent. In Figure 10.7 the data are displayed. Based on these data adaptive estimation in (10.115) is performed using nearest neighbour bandwidths, which are calculated assuming a standard Gaussian distribution for u_t.

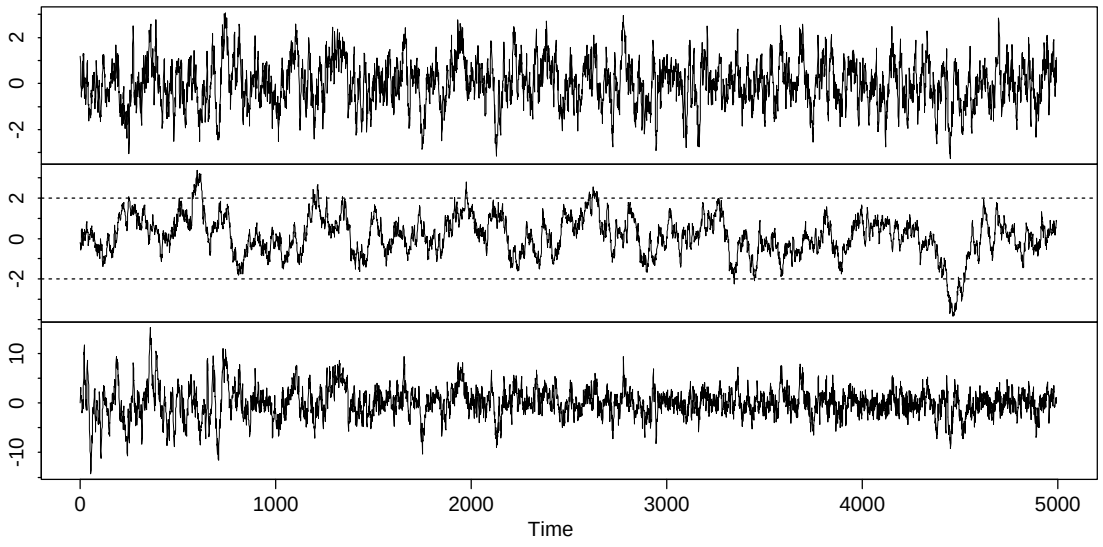

Figure 10.7: *Simulated output (bottom) when x_t (top) and u_t (middle) are $AR(1)$-processes.*

The results obtained using the modified updating formula (10.110) are displayed for fitting points $u = -2, -1, 0, 1, 2$ in Figures 10.8 and 10.9. For the first 2/3 of the period the estimates at $u = -2$, i.e. $\hat{a}(-2)$ and $\hat{b}(-2)$, only gets updated occasionally. This is due to the correlation structure of $\{u_t\}$ as illustrated by the realization displayed in Figure 10.7.

For both estimates the bias is most pronounced during periods in which the true coefficient-

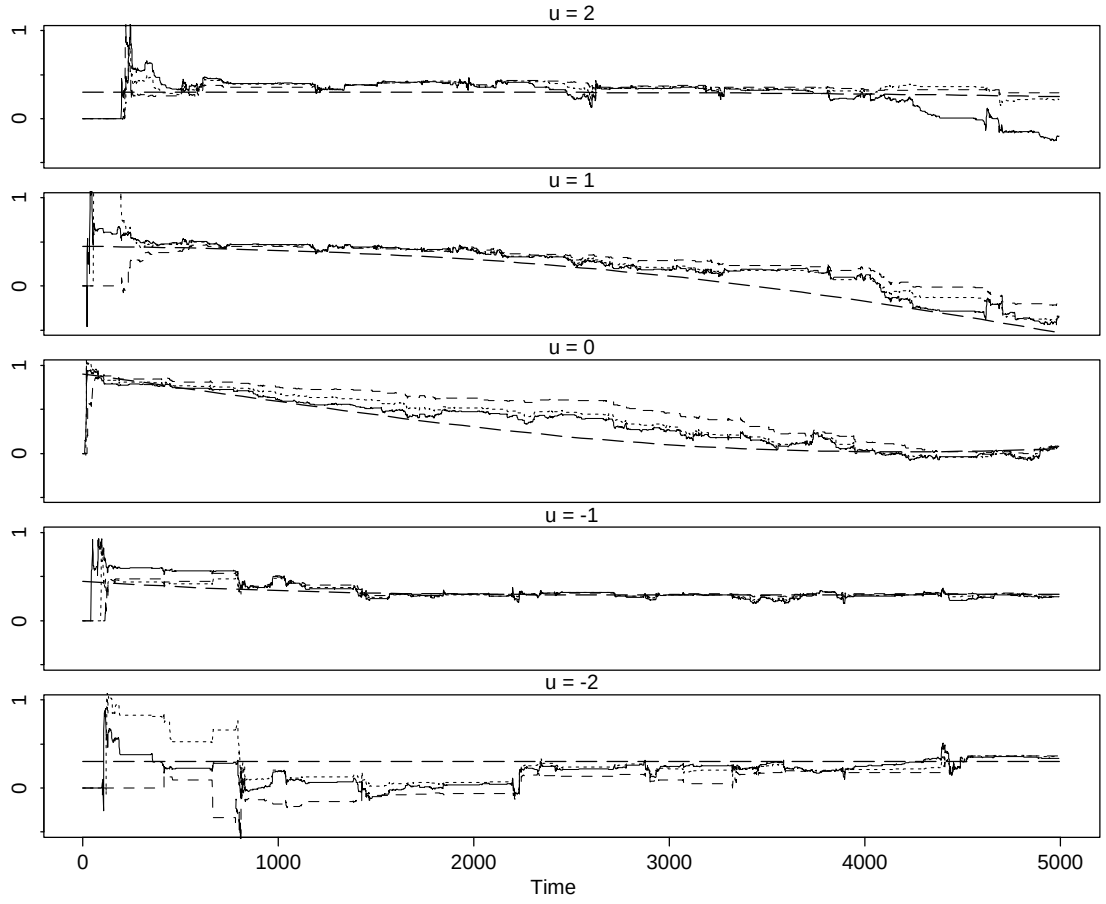

Figure 10.8: *Adaptive estimates of $a(u)$ using local linear approximations and nearest neighbour bandwidths 0.3 (dashed), 0.5 (dotted), and 0.7 (solid). True values are indicated by smooth dashed lines.*

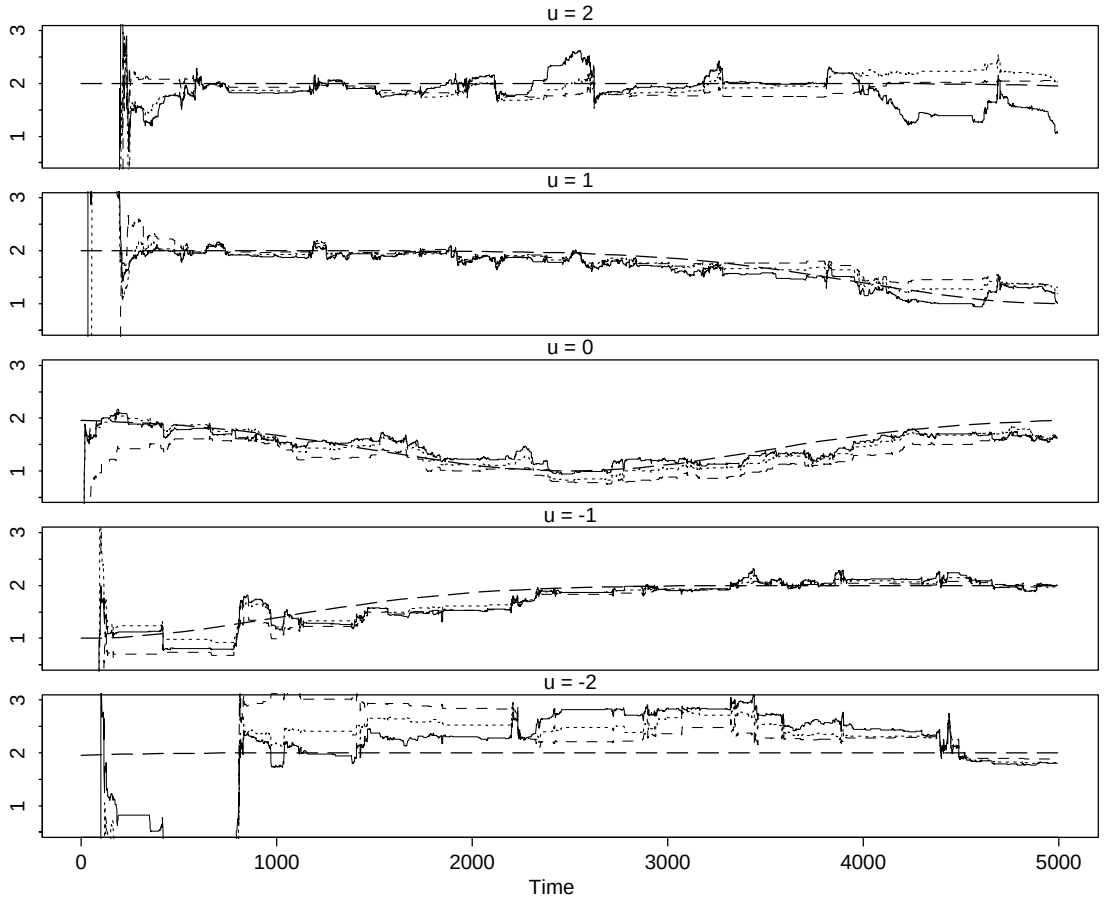

Figure 10.9: *Adaptive estimates of* $b(u)$ *using local linear approximations and nearest neighbour bandwidths 0.3 (dashed), 0.5 (dotted), and 0.7 (solid). True values are indicated by smooth dashed lines.*

function changes quickly for values of u_t near the fitting point considered. This is further illustrated by the true functions in Figure 10.6 and it is, for instance clear that adaption to $a(t,1)$ is difficult for $t > 3000$. Furthermore, $u = 1$ is rarely visited by $\{u_t\}$ for $t > 3000$, see Figure 10.7. In general, the low bandwidth ($\alpha = 0.3$) seems to result in large bias, presumably because the effective forgetting factor is increased on average, cf. Section 10.7.3. Similarly, the high bandwidth ($\alpha = 0.7$) result in large bias for $u = 2$ and $t > 4000$. A nearest neighbour bandwidth of 0.7 corresponds to an actual bandwidth of approximately 2.5 at $u = 2$ and since most values of u_t are below one, it is clear that the estimates at $u = 2$ will be highly influenced by the actual function values for u near one. From Figure 10.6 it is seen that for $t > 4000$ the true values at $u = 1$ is markedly lower that the true values at $u = 2$. Together with the fact that $u = 2$ is not visited by $\{u_t\}$ for $t > 4000$ this explains the observed bias at $u = 2$, see Figure 10.10.

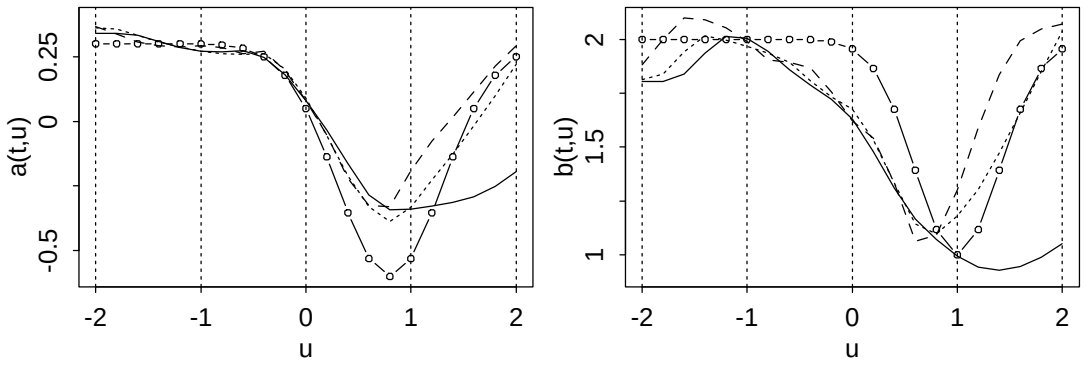

Figure 10.10: *Adaptive estimates for the example considered in Section 10.7.4 at* $t = 5000$ *for* $\alpha = 0.3$ *(dashed),* 0.5 *(dotted),* 0.7 *(solid). True values are indicated by circles and fitting points ranging from -2 to 2 in steps of 0.2 are used.*

Abrupt changes in input signals

One of the main advantages of the modified updating formula (10.110) over the normal updating formula (10.108) is that it does not allow fast changes in the estimates at fitting points which has not been visited by the process $\{u_t\}$ for a longer period. Actually, it is possible that, using the normal updating formula will result in a nearly singular $\mathbf{R}_t$.

To illustrate this aspect 5000 observations are simulated using the model (10.114). The sequence $\{x_t\}$ is simulated as a standard Gaussian $AR(1)$-process with a pole in 0.9.

Furthermore, $\{u_t\}$ is simulated as an iid process where

$$u_t \sim \begin{cases} N(0,1), & t = 1,\ldots,1000 \\ N(3/2, 1/6^2), & t = 1001,\ldots,4000 \\ N(-3/2, 1/6^2), & t = 4001,\ldots,5000 \end{cases}$$

To compare the two methods of updating, i.e. (10.108) and (10.110), a fixed λ is used in (10.110) across the fitting points and the effective forgetting factors are designed to be equal. If $\tilde{\lambda}$ is the forgetting factor corresponding to (10.108) it can be varied with u as

$$\tilde{\lambda}(u) = E[\lambda^u_{eff}(t)] = 1 - (1-\lambda)E[w_u(u_t)],$$

where $E[w_u(u_t)]$ is calculated assuming that u_t is standard Gaussian, i.e. corresponding to $1 \leq t \leq 1000$. A nearest neighbour bandwidth of 0.5 and $\lambda = 0.99$ are used, which results in $\tilde{\lambda}(0) = 0.997$ and $\tilde{\lambda}(\pm 2) = 0.9978$.

The corresponding adaptive estimates obtained for the fitting point $u = -1$ are shown in Figure 10.11. The figure illustrates that for both methods the updating of the estimates stops as $\{u_t\}$ leaves the fitting point $u = -1$. Using the normal updating (10.108) of $\mathbf{R}_t$ its value is multiplied by $\tilde{\lambda}(-1)^{3000} \approx 0.00015$ as $\{u_t\}$ returns to the vicinity of the fitting point. This results in large fluctuations of the estimates, starting at $t = 4001$. As opposed to this, the modified updating (10.110) does not lead to such fluctuations after $t = 4000$.

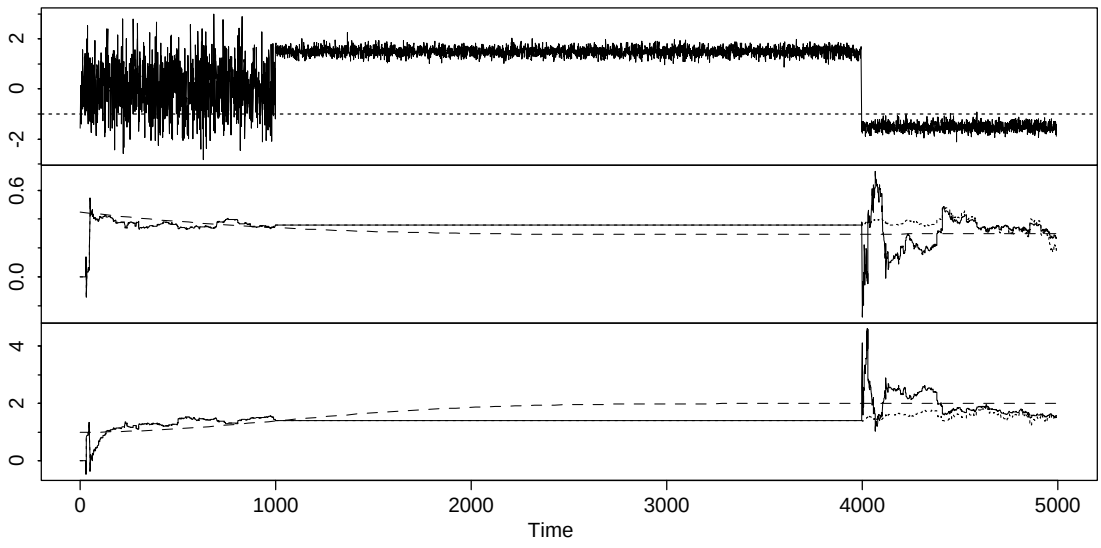

Figure 10.11: *Realization of $\{u_t\}$ (top) and adaptive estimates of $a(-1)$ (middle) and $b(-1)$ (bottom), using the normal updating formula (solid) and the modified updating formula (dotted). True values are indicated by dashed lines.*

An application

In [**?**] this class of models is used in relation to district heating systems to model the non-linear dynamic response of network temperature on supply temperature and flow at the plant. A particular feature of district heating systems is, that the response on supply temperature variations depends on the flow. This is modelled by describing the relation between temperatures by an ARX(30)-model in which the coefficients depend on the flow, hence the dimension of the vector $\mathbf{x}_s$ is 30 and there is just one variable $\mathbf{u}_s$ as introduced above in Section 10.7.2.

In this application the modified update of the $\mathbf{R}_s$-matrix is nessecary, since if, for instance, we wish to adaptively estimate the stationary relation between the heat consumption of a town and the ambient air temperature then $\{u_t\}$ contains an annual fluctuation and at some geographical locations the transition from, say, warm to cold periods may be quite fast. In such a situation the normal updating formula (10.108) will, essentially, forget the preceding winter during the summer, allowing for large changes in the estimate at low temperatures during some initial period of the following winter, implying the need for the modifications discussed in order not to obtain too fluctuating estimates. As in the last simulation case above, it is possible that the normal updating formula will result in a nearly singular $\mathbf{R}_t$.

Chapter 11

Prediction

11.1 Basic methodology – Ideas

The basic idea is to use old information about a process and its external influences to compute future values of the process, which might include computing also future values of the externals.

The most convenient way of condensing the past is in a model, and we have seen various types of models previously, linear and nonlinear timeseries with discrete or continuous time. The models have been parametric or nonparametric, however, here we will concentrate on the parametric case and refer to the previous Chapter 2.14 for nonparametric predictions.

Start off with a linear ARMAX-process,

$$Y_t + \phi_1 Y_{t-1} ..... \tag{11.1}$$

The prediction idea, if formulated verbally, is to take data as they are now and use as a predictor of the process the mean value of the previous responses to this type of data.

In time series this is formulated as the conditional mean

$$\hat{Y} = E(Y_{t+k}|\mathcal{I}^T). \tag{11.2}$$

$\mathcal{I}^T$ denotes the information that is available at time t. In this case it consists of old measurements of the inputs and outputs. The conditional mean is known to give the optimal predictor for a large number of criteria and distributions, cf. [Åström 1970] or

[Madsen 1995*b*]. The same basic idea is however used also in connection with non-parametric regressions as seen in Chapter 2.14, for neural networks and in prediction in chaotic systems, cf [**?**].

The criterion which is use to designate the 'optimal' predictor is usually the mean square prediction error, and then the optimal predictor is given by the conditional mean. However, the conditional mean solves the problem also for general symmetric, nondecreasing criteria when the distribution is symmetric and nonincreasing, cf. [Åström 1970].

If we go back to the ARMAX model above, the optimal predictor is given by

..... (11.3)

This representation of the optimal predictor is iterative in the number of steps, but of course the k-step predictor could be expressed directly as a filter from $\mathcal{I}^T$

.... (11.4)

The basic system is however often timevarying, which may be treated using a estimation technique that allows for parameter variation, i.e. to use recursive estimation, cf. Chapter 10. As was shown there an alternative is to find the physical origin to the timevariations (if possible) and include it in a (nonlinear) model. Usually, both approaches have to be used.

11.1.1 Modeling of the Power Load in Kulladal, Malmö

This example is more thoroughly treated in [Arvastson & Holst 1995]. The load on a district heating network is to be modelled and predicted using physical knowledge as well as stochastic modelling. Data are collected from Kulladal, a residential part of Malmö. It is about 1 % of the total district heating network in town and the maximum heat demand is around 15 MW. The net is shown in figure 11.1, which is taken from [Andersson, Frederiksen & Lundell 1993].

11.2 External signals

- they are needed
- but which are representative, (validation)
- location

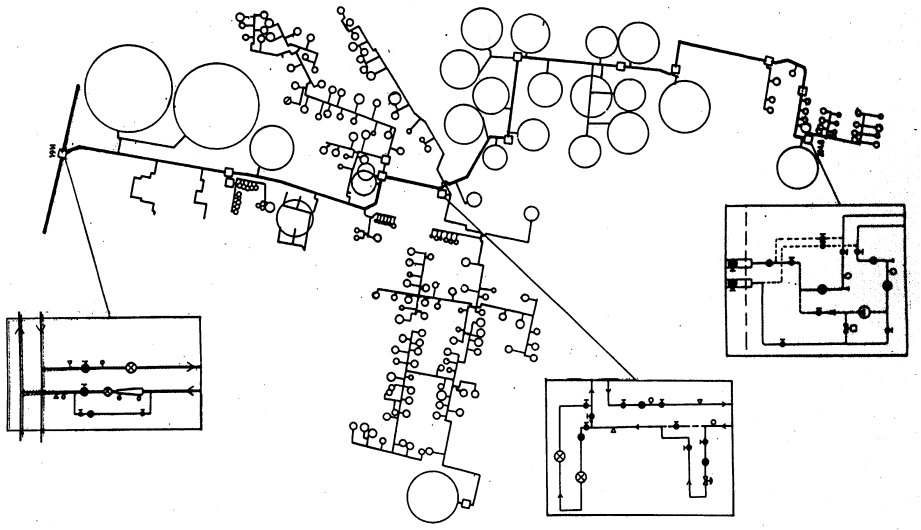

Figure 11.1: The Kulladal net, the area of the circles corresponds to the heat load in the different buildings. The three valve chambers where the measurement equipment are located are also shown, the one to the left connects the Kulladal net with the rest of the net in Malmö. Figure from Andersson et al, [XXXX].

- also the external signals need to be predicted,
  short term – timeseries
  long term – other sources (DHI), pictures, indicators far away

11.2.1 Prediction of Electric Power Load in Southern Sweden

11.3 Precision of forecasts

11.4 Holidays and gross errors

11.4.1 Robustness

11.5 Interpolation

11.6 Event prediction

11.7 On the control problem

Chapter 12
Design of experiments

12.1 Introduction

Experiment design is concerned with the choice of experiments that yield the best model of the system. The choice of the best experiments is therefore influenced by the concept of the appropriateness of the model and this is normally dictated by the intended application of the model. This chapter will focus on *optimal experiment design for system modeling* with a special attention to parameter estimation, which is the most common usage of experiment design. The criteria of optimality should be formulated in terms of the interesting physical characteristics of the system.

The purpose of experiment design is then to ensure that the information (mostly about the parameters) provided by the experiment is maximized. The design of experiments for system modeling includes choice of the *sequence of input signals, sampling time* and *presampling filters*.

Any design must take into account the constraints on the allowable experimental conditions. Some typical constraints that might be met in practice are amplitude constraints on inputs, outputs or internal signals; power constraints on signals; limited total experimental time or limited number of samples. Further more, the available measuring and/or actuating equipment introduces constraints like least possible sampling interval, highest possible accuracy, etc.

A different example of experiment design, which, however, has not received much attention in the literature, is the problem of finding an optimal input sequence for *fault detection*. In this case a suitable choice of input signal can improve the statistical properties of the detection test involved – see e.g. [Sadegh, Madsen & Holst 1996].

This chapter will focus on experiment design for linear systems. The experiment design for nonlinear systems offers a sequence of added problems in addition to the ones for linear systems. The reason is that the most appropriate tools and methods for optimal experiment design are specified in the frequency domain, which can not be generally used for nonlinear systems, even though a generalization of the transfer function is available via the Volterra kernel methods.

There is a substantial literature on optimal experiment design. In the books by Fedorov (1972) and Pazman (1986), among others, the theory and mathematical treatment of optimal experiment design in general is considered.

Optimal design for linear dynamic systems are specifically considered in the books by Goodwin & Payne (1977) and Zarrop (1979). Design of experiments for identification of transfer functions in the frequency domain has been considered by e.g. Yuan & Ljung (1985), who discuss the use of prior information.

Prior information about the system and the intended use of the model are both inevitable terms in connection with optimal design of experiments. Especially for physical models there are often prior information about the model available. This prior information should also be used in the design of experiments. This can be done either in a classical way using only prior expected values of the parameters or in a Bayesian approach, by incorporating prior distributions of the parameters for the design of optimal experiments, see [Melgaard, Sadegh, Madsen & Holst 1993] and [Sadegh, Melgaard, Madsen & Holst 1994].

12.2 Identifiability

Identifiability is a concept that addresses the problem of whether the given identification procedure will yield a unique value of the parameters. One aspect of the problem has to do with the experimental conditions, e.g. whether the data set is informative enough (persistently exciting) to distinguish between different models for a given estimation method. This is considered in Section 12.2.2. Another aspect has to do with the model structure, i.e. if different sets of parameters will give equal models, given that the data is informative. This qualitative aspect of experiment design is considered in Section 12.2.1.

12.2.1 Structural Identifiability

Structural identifiability has to do with the parameterization and structure of the model. Hence, the questions of structural identifiability can be answered before the data have been collected on the system.

The concept provides a necessary condition to recover the unknown characteristics of the system from measured data. Structural identifiability is previously considered in Section 9.2.1, but will briefly be mentioned here for completeness.

There is an extensive literature on the subject. The reader may consult the review paper by Walter & Pronzato (1990) for the current state of art in both structural identifiability and experiment design. Ljung & Glad (1994) have recently reported new algebraic methods for investigating the identifiability of rather general non-linear model structures. Nevertheless, the literature on structural identifiability of non-linear models is sparse.

Let us first illustrate the problem by considering a simple example. More examples are given in Section 9.2.1.

Example 12.1 (Structural identifiability - simple case). As an example, consider the very simple static model

$$y_t = (p_1 + p_2)u_t + \epsilon_t \tag{12.1}$$

where $\{u_t\}$, $\{y_t\}$, and $\{\epsilon_t\}$ are input, output, and noise sequences. It is obvious that no matter how long time the data are gathered, we won't be able to estimate p_1 and p_2. We'll only be able to estimate $p_1 + p_2$. This example also underlines the role of prior knowledge. If we for example know (from physics or any other source) that p_1 and p_2 satisfy the relation

$$c_1 p_1 + c_2 p_2 = 0 \tag{12.2}$$

then the structure will be identifiable. ♦

Definition 12.1 (Structural identifiability). The parameter θ_i is said to be *structurally globally identifiable* if for almost any $\boldsymbol{\theta}^*$

$$M(\boldsymbol{\theta}) = M(\boldsymbol{\theta}^*) \Rightarrow \theta_i = \theta_i^*. \tag{12.3}$$

It is called *structurally locally identifiable* if for almost any $\boldsymbol{\theta}^*$ there exits a neighborhood $\upsilon(\boldsymbol{\theta}^*)$ such that if $\boldsymbol{\theta} \in \upsilon(\boldsymbol{\theta}^*)$, then the implication (12.3) is true. ▲

The problem of structural identifiability is of importance for both linear and nonlinear models, but there are some specific properties which differs for the two types. In this context it is relevant to speak of two kinds of nonlinearities of the model, one is

nonlinear *in the inputs*, which means that the output does not satisfy the superposition principle concerning the inputs:

$$\boldsymbol{y}_M(\boldsymbol{\theta}, \lambda\,\boldsymbol{u}_1 + \mu\,\boldsymbol{u}_2, t) \neq \lambda\,\boldsymbol{y}_M(\boldsymbol{\theta}, \boldsymbol{u}_1, t) + \mu\,\boldsymbol{y}_M(\boldsymbol{\theta}, \boldsymbol{u}_2, t) \tag{12.4}$$

for some real λ and μ. The other kind of nonlinearity, is when the model is nonlinear *in the parameters*, which means that the output from the model does not satisfy the superposition principle concerning the parameters:

$$\boldsymbol{y}_M(\lambda\,\boldsymbol{\theta}_1 + \mu\,\boldsymbol{\theta}_2, \boldsymbol{u}, t) \neq \lambda\,\boldsymbol{y}_M(\boldsymbol{\theta}_1, \boldsymbol{u}, t) + \mu\,\boldsymbol{y}_M(\boldsymbol{\theta}_2, \boldsymbol{u}, t) \tag{12.5}$$

Most often phenomenological models are nonlinear in the parameters.

Remark. It should be remarked that only models that are nonlinear in the parameters can have structurally locally identifiable parameters which are not at the same time structurally globally identifiable, see [Walter & Pronzato 1990]. The outputs of models that are linear in the parameters can be written as

$$\boldsymbol{y}_M(t, \boldsymbol{\theta}, \boldsymbol{u}) = \boldsymbol{\phi}(t, \boldsymbol{u})^T\boldsymbol{\theta} \tag{12.6}$$

and

$$M(\boldsymbol{\theta}) = M(\boldsymbol{\theta}^*) \Leftrightarrow \boldsymbol{\phi}(t, \boldsymbol{u})^T\boldsymbol{\theta} = \boldsymbol{\phi}(t, \boldsymbol{u})^T\boldsymbol{\theta}^* \tag{12.7}$$

This set of linear equations has a unique solution if the columns of $\boldsymbol{\phi}$ are linearly independent, or else it has an infinite number of solutions. ▼

Remark. Ljung & Glad (1994) have shown that testing global identifiability is equivalent to the possibility to express the model structure as a linear regression

$$\boldsymbol{\psi}(t) = \boldsymbol{\phi}(t)^T\boldsymbol{\theta}\ . \tag{12.8}$$

This is a generalization of (12.6) where $\boldsymbol{\psi}(t)$ and $\boldsymbol{\phi}(t)$ are matrices, entirely formed from past values of $\boldsymbol{y}(t)$, $\boldsymbol{u}(t)$ and their derivatives. [Ljung & Glad 1994] treat this problem for models that have polynomial nonlinearities with respect to the parameters and the inputs. ▼

In the following when the model is only referred to as linear or nonlinear this means with respect to the inputs.

Linear Models

The problem of structural identifiability of linear state space models arises from the fact that for a given transfer function model corresponds, in general, a continuum of state space models. It is therefore necessary to put restrictions on the structure of the

state space model in order to provide a unique relation between the parameters of the state space model and the transfer function.

In this section the descriptions are kept in continuous time even though our observations are in discrete time for any practical applications. It is obvious that the sampling of the observations (inputs and outputs) may deteriorate the identifiability properties of the model, but concerning only the structural identifiability it will have no influence.

Nonlinear Models

Testing for structurally identifiability in a nonlinear model is slightly more complicated. One method consist of linearizing the model around some equilibrium point and then apply the method for linear models. A second approach uses a series expansion of the output either in the time domain or in the time and input domain [Walter & Pronzato 1990]. $M(\boldsymbol{\theta}) = M(\boldsymbol{\theta}^*)$ then implies that the coefficients of the series should be equal, which, as in the linear case, yields a set of equations in $\boldsymbol{\theta}$ parameterized by $\boldsymbol{\theta}^*$.

12.2.2 Persistence of Excitations

Persistence of excitations is another necessary condition for estimating the parameters of a model. Again a very simple example is used to illustrate the problem.

Example 12.2. Persistent excitations. Consider the simple model

$$y_t = p_1 u_t + p_2 u_{t-1} + \epsilon_t \tag{12.9}$$

with the same notation as before. Assume that the DC signal $u_t = 1$, $t = ..., -1, 0, 1, ...$ is applied. It is obvious that again we'll only be able to estimate the sum $p_1 + p_2$ and not each of the parameters, even though the model is structurally identifiable. ♦

Definition 12.2. Persistency of excitation. An input signal u fulfill the condition of *persistent excitations* of order n if the (Yule-Walker) matrix of the autocorrelations from lag 0 until lag $(n-1)$ is positive definite. ▲

It is very easy to see that this condition is reasonable by considering a least squares estimation of the parameters in the model

$$y_t = p_1 u_1 + \cdots + p_n u_n + \epsilon_t \tag{12.10}$$

Both structural identifiability and persistence of excitations can be analyzed through studying *Fisher's information matrix* whose features are the main topic of experimental design.

12.3 Design of optimal experiments

12.3.1 Preliminaries

The basis for experiment design is maximizing a function that is a measure of goodness of the experiments. The classical experiment design which is thoroughly investigated in [Fedorov 1972] is based on maximizing a scalar function of Fisher's information matrix, which is defined by

Definition 12.3 (Fisher's Information Matrix). Fisher's Information Matrix is defined by

$$\boldsymbol{M}_F = \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{\theta}} \left\{ (\frac{\partial \log p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{u})}{\partial \boldsymbol{\theta}})^T (\frac{\partial \log p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{u})}{\partial \boldsymbol{\theta}}) \right\} \tag{12.11}$$

where $\boldsymbol{Y}$ denotes the output, $\boldsymbol{u}$ the input, $p(\cdot)$ density functions, and finally, $\boldsymbol{\theta}$ denotes the parameter vector. ▲

The main reason for choosing a scalar function of the information matrix is the well known Cramer-Rao inequality which, subject to certain regularity conditions, states that a lower limit of *the covariance of every unbiased efficient estimator of $\boldsymbol{\theta}$ is asymptotically given by the inverse of the information matrix.* More precisely ([Rao 1973])

$$V(\hat{\boldsymbol{\theta}}) \geq \boldsymbol{M}_F^{-1} \tag{12.12}$$

Example 12.3. As an introduction and motivate the further study, consider the simple model (12.9) again. The model can be written in regression form as

$$y_t = \boldsymbol{\phi}_t^T \boldsymbol{\theta} + \epsilon_t$$

where $\phi_t^T = (u_t, u_{t-1})$ and $\theta^T = (p_1, p_2)$. Defining $\boldsymbol{Y}$ to be the data vector and the matrix $\boldsymbol{\Phi}$ with the i'th row to be $\boldsymbol{\phi}_i$, the least squares estimate of the parameter $\boldsymbol{\theta}$ is given by the so called normal equation :

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T Y \tag{12.13}$$

See [Madsen 1995*a*].

It is left as an exercise to show that the Fishers information matrix is proportional to $\boldsymbol{\Phi}^T \boldsymbol{\Phi} = \sum_t \boldsymbol{\phi}_t \boldsymbol{\phi}_t^T$. Let's solve the experiment design problem for the model (12.9).

If $\mathbf{\Phi}^T\mathbf{\Phi}$ is singular, it is obvious that the parameters can not be estimated (verify that the DC input makes the information matrix singular). As a criterion for designing the optimal inputs, we can maximize the determinant of the information matrix. Notice that

$$\sum_t \boldsymbol{\phi}_t\boldsymbol{\phi}_t^T = \begin{bmatrix} \sum_t u_t^2 & \sum_t u_t u_{t-1} \\ \sum_t u_t u_{t-1} & \sum_t u_{t-1}^2 \end{bmatrix}$$

If the experiment time is sufficiently large and the inputs are restricted to the class of ergodic signals, the information matrix is proportional to

$$\begin{bmatrix} \gamma_{uu}(0) & \gamma_{uu}(1) \\ \gamma_{uu}(1) & \gamma_{uu}(0) \end{bmatrix} \tag{12.14}$$

Assume further that the input power is restricted to be not more than C. It is easily verified that any input signal with power C and $\gamma_{uu}(1) = 0$ maximizes the determinant of the information matrix and is therefore optimal. An example of such optimal input is white noise with variance C. Another input with the same optimality feature is the signal $u_t = \cos(\frac{\pi}{2}t)$, $t = ..., -1, 0, 1, ...$ (verify that!). The latter signal has the pleasant property that it is periodic. The question is whether this holds more generally, i.e. whether it is possible to find optimal sequences that are periodic. The answer is (surprisingly) yes. ♦

In the following, we consider optimality criteria and the fundamental theorems in more detail.

12.3.2 Optimality Criteria

As mentioned above the quantitative study of the experiment design implies that a measure of the information achieved from an experiment is needed. It is common practice, cf. [Goodwin & Payne 1977], to select a performance measure related to the expected accuracy of the parameter estimates to be obtained, in general the covariance of the parameters. When the estimator is asymptotically efficient, the rationale for using the Fisher information matrix as a suitable characterization of the asymptotic parameter uncertainty is obtained.

It should be remarked that an experiment satisfying $\det \boldsymbol{M}_F(\boldsymbol{\theta}) \neq 0$ is called informative, this ensures local identifiability for the model parameters. Designing an experiment by minimizing a suitable criterion

$$J(\boldsymbol{\xi}) = \phi(\boldsymbol{M}_F(\boldsymbol{\theta}, \boldsymbol{\xi})) \tag{12.15}$$

can thus be seen as maximizing a measure of identifiability, where ϕ is a scalar function and $\boldsymbol{\xi}$ is the design.

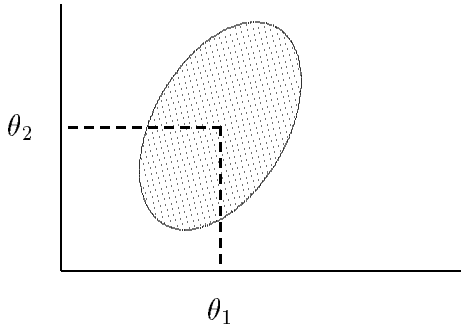

Figure 12.1: D-optimal design minimizes the volume of the confidence ellipsoid.

Local Design

A number of different standard measures of information has been studied. For a review of the properties of the different measures of information see [Pazman 1986] and [Walter & Pronzato 1990]. A large number of these standard measures are particular cases of the general L_k-class of the optimality criteria. The criterion belonging to this class is defined by a function of the form

$$\phi(\boldsymbol{M}_F) = \begin{cases} [n^{-1} \operatorname{tr}(\boldsymbol{V}\boldsymbol{M}_F^{-1}\boldsymbol{V}^T)^k]^{1/k} & \text{if } \det \boldsymbol{M}_F \neq 0 \\ \infty & \text{if } \det \boldsymbol{M}_F = 0 \end{cases} \tag{12.16}$$

where $k > 0$ and $\boldsymbol{V}$ is a nonsingular $n \times n$ matrix. The local optimal design refers to the case where the criterion is minimized for a given value of the parameters $\boldsymbol{\theta}^*$, say. This value is often chosen as the expected mean $\boldsymbol{\theta}^* = \mathrm{E}_\theta[\boldsymbol{\theta}]$ calculated from the prior distribution of the parameters $p(\boldsymbol{\theta})$.

D-optimality

The most studied criterion is the *D-optimality* criterion, which is defined by

$$\phi(\boldsymbol{M}_F(\boldsymbol{\xi})) = -\log \det \boldsymbol{M}_F(\boldsymbol{\xi}) \; . \tag{12.17}$$

The criterion is obtained for $\boldsymbol{V} = \boldsymbol{I}$ and $k \to 0$ in (12.16). Thus this design minimizes the generalized variance of the parameter estimates. The criterion has a geometrical interpretation: The asymptotic confidence regions for the maximum likelihood estimate of $\boldsymbol{\theta}$ are ellipsoids, and a D-optimal experiment thus minimizes the volume of these ellipsoids – see Figure 12.3.2. An important property of D-optimal experiments is their independence of the scaling of parameters, due to the geometrical property of the determinant.

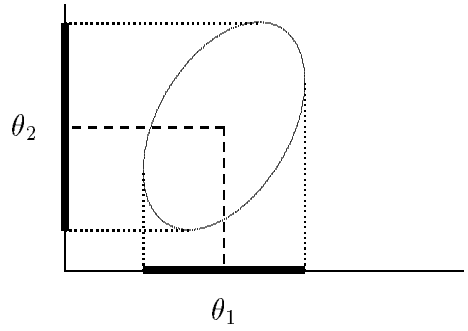

Figure 12.2: C-optimal design minimizes the relative precision of the single parameters, without paying attention to the correlation structure.

L-optimality

Criteria that are linear in the inverse Fisher information matrix (= dispersion matrix of the best linear estimator) are obtained from (12.16) by setting $k = 1$. The particular choices $\boldsymbol{V} = \boldsymbol{I}$ and $\boldsymbol{V} = \mathrm{diag}(\theta_i^{-1}, i = 1, \cdots, n)$ respectively, correspond to *A-* and *C-optimality*. A-optimal experiments minimize the sum of variances of $\boldsymbol{\theta}$. C-optimal experiments are related to the relative precision of the estimates, actually they summarize $1/t_i^2$, where t_i is the t-statistic of the ith parameter. This criterion is also independent of the parameter scale. The *A-* and *C-optimality* criteria does not take the correlation between the parameters into account. From a pragmatic point of view such criteria should be seen with suspicion because they can lead to design of experiments in which the parameters are unidentifiable, i.e. $\det \boldsymbol{M}_F = 0$ (see [Goodwin & Payne 1977]).

E-optimality

The *E-optimality* criterion is obtained from (12.16) by setting $\boldsymbol{V} = \boldsymbol{I}$ and $k \to \infty$, this corresponds to minimizing the maximum eigenvalue of $\boldsymbol{M}_F^{-1}$,

$$\phi(\boldsymbol{M}_F(\boldsymbol{\xi})) = \lambda_{\max}(\boldsymbol{M}_F(\boldsymbol{\xi})^{-1}) \tag{12.18}$$

Geometrically this can be understood as minimizing the maximum diameter of the asymptotic confidence ellipsoids for the parameters, because the semi-axes of the ellipsoids are directed as the eigenvectors of $\boldsymbol{M}_F^{-1}$ with lengths proportional to the eigenvalues of the matrix. In other words, an E-optimal design aims at improving the most uncertain region of the parameter space and making the confidence region as spherical as possible. By using $\boldsymbol{V} = \mathrm{diag}(\theta_i^{-1}, i = 1, \cdots, n)$, the criterion is independent

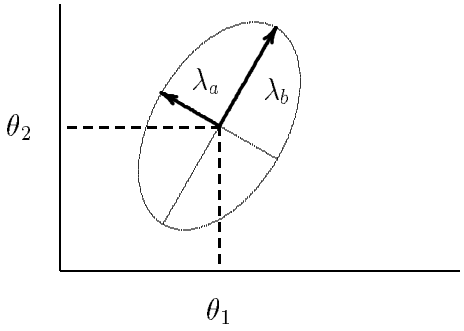

Figure 12.3: E-optimal design minimizes the maximum eigenvalue of the inverted information matrix, thus making the confidence region of the parameters as spherical as possible.

of the scaling of parameters. The geometrical interpretation is still valid, but for the *normalized* inverted information matrix.

Physical Measures

Other criteria for optimality than the standard ones can be specified. Specifically when dealing with physical models nonstandard criteria may be of interest, since the physical parameters of greatest interest might not be directly entering into the system description but rather through some transformation of these parameters. Assume that the parameters of interest are described by the functional relation $\boldsymbol{f}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the parameters of the model. It is important to take this transformation into account in the design of experiments. The asymptotic covariance of $\boldsymbol{f}$ is calculated by using Gauss' formula

$$\boldsymbol{M}_{F,f}^{-1} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}^T} \boldsymbol{M}_F^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}} \ . \tag{12.19}$$

Any of the standard measures of information can now be applied for $\boldsymbol{M}_{F,f}$. In [Melgaard, Madsen & Holst 1992] a physical model of the thermal characteristics of a building is considered. The classical measures of information all lead to the same optimal design of input signal, a certain PRBS sequence. It turned out, however, that the main interest is not focussed on the individual parameters of the model, but on the overall heat transmittance and internal heat capacity of the building. These physical characteristics are given as a function of the model parameters. An optimal design considering this application specific measure of information leads to an input signal, which is a step. This design turns out to be much different from the original design, (more weight on low frequencies), but it reflects the demands of the building physicists.

12.3.3 Bayesian Design

In general the designs of the previous sections are dependent upon the unknown parameter values. Any design will and shall be dependent upon the prior information available about the system to be identified. From a statistical point of view it is obvious, that one way of incorporating this prior information is to use a Bayesian procedure. In this approach the prior knowledge of the parameters are expressed via their the prior distribution function, which expresses the partial information on the system available prior to an experiment.

In the Bayesian approach a loss function $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ is defined, which describes the consequences of obtaining $\hat{\boldsymbol{\theta}}$, when $\boldsymbol{\theta}$ is the true parameter vector. This function is then used as a basis for obtaining the optimal experiment design. Prior to the experiment the expected performance is

$$J_1 = E_{\boldsymbol{\theta},\boldsymbol{Y}}[L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})]. \tag{12.20}$$

This criterion may be optimized directly with respect to the allowable experimental conditions. By making suitable assumptions and using truncated Taylor expansion J_1 can be simplified to:

$$\begin{aligned} J_1 &= \mathrm{E}_{\boldsymbol{\theta},\boldsymbol{Y}}[L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})] && (12.21)\\ &= \mathrm{E}_{\boldsymbol{\theta}}\,\mathrm{E}_{\boldsymbol{Y}|\boldsymbol{\theta}}[L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})] && (12.22)\\ &\simeq \mathrm{E}_{\boldsymbol{\theta}}\,\mathrm{E}_{\boldsymbol{Y}|\boldsymbol{\theta}}[L(\boldsymbol{\theta}, \boldsymbol{\theta}) + \frac{\partial L}{\partial \hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \\ &\quad \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T(\frac{\partial^2 L}{\partial \hat{\boldsymbol{\theta}}^2})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})] && (12.23)\\ &= \mathrm{E}_{\theta}[L(\boldsymbol{\theta}, \boldsymbol{\theta}) + \frac{1}{2}\,\mathrm{tr}\{(\partial^2 L/\partial \hat{\boldsymbol{\theta}}^2)\boldsymbol{M}_F^{-1}\}]\ , \end{aligned}$$

where $\boldsymbol{M}_F$ is given by (12.11) and it is assumed that $\hat{\boldsymbol{\theta}}$ is an efficient unbiased estimator. Optimizing this criterion is thus equivalent to optimizing a criterion of the form $\mathrm{E}_{\theta}[\mathrm{tr}\{\boldsymbol{W}\boldsymbol{M}_F^{-1}\}]$, where $\boldsymbol{W} = \partial^2 L/\partial \hat{\boldsymbol{\theta}}^2$ is a weight matrix, see [Goodwin & Payne 1977]. Generalizing this criterion yields

$$J_2 = E_{\theta}(\phi(\boldsymbol{M}_F(\boldsymbol{\theta}, \boldsymbol{\xi})))\ , \tag{12.24}$$

where expectation is taken over the prior distribution of $\boldsymbol{\theta}$, and ϕ is suitable scalar function. In general ϕ can be of the form considered in the previous sections for local designs. As pointed out by [Walter & Pronzato 1990] there are several average optimality criteria (12.24) corresponding to one local design (12.15). This is due to the fact that integration is *not* a commutative operator with nonlinear functions, e.g. inversion and logarithm. This implies that considering the maximization of $\phi = \det \boldsymbol{M}_F$

leads to a different design than considering the minimization of $\phi = (\det \boldsymbol{M}_F)^{-1}$ in the criterion (12.24). For a local design these functions result in the same design.

The following simple example provides an example of setting up a reasonable loss function L.

Example 12.4 (Example of a loss function). Consider the very simple system $y_t = \theta u_t + \epsilon_t$, $\theta \neq 0$, where $\{\epsilon_t\}$ is a sequence of i.i.d. random variables. Assume that the modeling goal is to design a controller to keep the output as close as possible to the reference signal $y_{ref} = 1$. The minimum power of $(y_t - y_{ref})$ is clearly obtained by choosing $u_t = 1/\theta$. However, the true parameter θ is not known and it is assumed that it must be estimated during an open loop identification experiment prior to the application of the controller.

It is desirable to derive an optimization criterion for designing the best identification input. The extra output error introduced by using the unbiased efficient estimate $\hat{\theta}$ in the control computation is given by $(\hat{\theta} - \theta)/\hat{\theta}$ which contributes to the power of the error signal $(y_t - y_{ref})$ by $[(\hat{\theta} - \theta)/\hat{\theta}]^2$. Thus, it is reasonable to define $L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$, and $E_\theta E\{L(\theta, \hat{\theta})|\theta\} = E_\theta\{\text{var}(\hat{\theta})\}$ where $\text{var}(\hat{\theta})$ denotes the variance of the estimate $\hat{\theta}$. This variance is asymptotically given by the lower bound of the Cramér-Rao inequality, i.e. the inverse of the information matrix. In this simple case the information matrix is independent of θ, which leads to optimization of M_F^{-1}. ♦

12.4 Fundamental concepts and tools

Thus, we have related a problem formulation based on decision theoretic considerations to one concerning maximization of the mean (with respect to θ) of a scalar function of the information matrix. Throughout the rest of the chapter, we consider optimization problems of the form $E_\theta\{\phi(M)\}$, where $\phi(\cdot)$ is a scalar function of its matrix valued argument. The optimization criterion is expressed in a Bayesian form which takes into account the fact that the prior information about θ is partial.

However, assume for the moment, and for the rest of this section, that the Bayesian criterion is approximated by replacing θ by its prior mean, θ_0, resulting in the non-Bayesian criterion $\phi(M)|_{\theta_0}$. We will return back to the Bayesian criterion later on, in Section 12.9.

Definition 12.4 (ϕ-function). Consider the mapping

$$\phi : \mathcal{M} \to \mathbb{R}.$$

If $\phi(\cdot)$ is non-increasing, i.e.

$$M_1 - M_2 \quad \text{nonnegative definite} \Rightarrow \phi(M_1) \leq \phi(M_2),$$

it is called a ϕ-function. ▲

Definition 12.5 (ϕ-optimality). Consider a ϕ-function defined over $\mathcal{M}$. Then an input measure ξ^* is called ϕ-optimal if

$$\phi(M(\xi^*)) \leq \phi(M(\xi)), \quad \forall \xi \in \mathcal{X}$$

where $\mathcal{X}$ is the set of all admissible input measures and $M(\xi)$ is the average information matrix corresponding to the measure ξ. See [Silvey 1980]. ▲

Example 12.5. $\phi(M) = -\log\det(M)$ is an example of a ϕ-function. This choice of $\phi(\cdot)$ is called D-optimality in the literature (see e.g. [Fedorov 1972]). ♦

Example 12.6. Another interesting choice is the D_s-optimality. Partition the parameter vector as $\theta = (\theta_1^\top, \theta_2^\top)^\top$ where θ_1 is the parameter of interest. Denote the corresponding information matrix by M and partition

$$M = \left\{ \begin{array}{cc} M_{11} & M_{12} \\ M_{12}^\top & M_{22} \end{array} \right\}$$

where M_{11} corresponds to the parameter of interest θ_1 and M_{22} corresponds to θ_2. Using the matrix inversion formula, the lower bound on the asymptotic covariance of $\hat{\theta}_1$ is given by $(M_{11} - M_{12}M_{22}^{-1}M_{12}^\top)^{-1}$, which leads to the ϕ-function corresponding to D_s-optimality

$$\phi(M) = -\log\det(M_{11} - M_{12}M_{22}^{-1}M_{12}^\top).$$

♦

Many interesting practical design criteria satisfy the condition of Definition 12.4 and they are also *convex* on $\mathcal{M}$.

Example 12.7. Consider the D-optimality criterion again. The fact that [Fedorov 1972]

$$\det(\beta M_1 + (1-\beta)M_2) > [\det(M_1)]^\beta [\det(M_2)]^{(1-\beta)}, 0 < \beta < 1,$$

implies that $\phi(\cdot)$ is convex on $\mathcal{M}$. ♦

From now on, we assume that the ϕ function is convex. Hence, solving for optimal input is equivalent to finding a point in the convex set $\mathcal{M}$ that minimizes the convex function $\phi(\cdot)$.

Theorem 12.6. *ϕ-optimal designs exist comprising not more than $p(p+1)/2+1$ sinusoidal components. Furthermore, if the ϕ-function is strictly decreasing (i.e., nonnegative definite non-zero $M_1 - M_2$ yields $\phi(M_1) < \phi(M_2)$), the upper bound on the number of sinusoidal components that suffice for optimality is $p(p+1)/2$.*

PROOF*: See [Goodwin & Payne 1977] and [Silvey 1980].*

The theorem implies that optimal design measures can be sought among line spectra with $l \leq p(p+1)/2+1$ or $l \leq p(p+1)/2$.

12.4.1 Algorithms for Constructing Optimal Design

One way of constructing optimal designs is by direct optimization. Theorem 12.6 implies that it is sufficient to write each candidate point in $\mathcal{M}$ as

$$\sum_{i=1}^{N_B} \alpha_i \tilde{M}(\omega_i),$$

where $N_B = p(p+1)/2+1$ and search over a $2N_B$ dimensional space to find the α's and ω's that minimize ϕ subject to

$$\sum_{i=1}^{N_B} \alpha_i = 1, \quad \forall i : \alpha_i > 0.$$

If $\phi(\cdot)$ is strictly decreasing then $N_B = p(p+1)/2$.

The other algorithm which is more conceptual is inspired by [Silvey 1980], and based on the following theorem. We start by defining the Fréchet derivative

Definition 12.7. The Fréchet derivative of a ϕ-function at M_1 in the direction of M_2 is

$$F_\phi(M_1, M_2) = \lim_{\gamma \to 0^+} \frac{1}{\gamma}[\phi\left((1-\gamma)M_1 + \gamma M_2\right) - \phi(M_1)].$$

▲

Theorem 12.8. *Assume that a ϕ-function is convex and differentiable at $M(\xi^*)$. Then $\xi^* \in \mathcal{X}$ is ϕ-optimal iff $F_\phi(M(\xi^*), \tilde{M}(\omega)) \geq 0$ for all $\omega \in [0, \pi/\Delta T]$.*

PROOF*: Necessity: The fact that $\phi(\cdot)$ attains its minimum at $M(\xi^*)$ implies*

$$\phi(M((1-\gamma)\xi^* + \gamma\xi)) - \phi\left(M(\xi^*)\right) \geq 0,$$

for all $\gamma \in [0,1]$ and all $\xi \in \mathcal{X}$. But, note that

$$(1-\gamma)M(\xi^*) + \gamma M(\xi) = M((1-\gamma)\xi^* + \gamma\xi).$$

Take ξ to be any single frequency design, i.e. $M(\xi) = \tilde{M}(\omega)$, $\omega \in [0, \pi/\Delta T]$. Then, the necessity follows from the definition of the Fréchet derivative. Sufficiency: Each $M(\xi) \in \mathcal{M}$ is written as

$$M(\xi) = \sum_{i=1}^{l} \alpha_i \tilde{M}(\omega_i)$$

where $l \leq p(p+1)/2+1$. Since $\phi(\cdot)$ is differentiable at $M(\xi^*)$

$$F_\phi(M(\xi^*), M(\xi)) = \sum_i \alpha_i F_\phi(M(\xi^*), \tilde{M}(\omega_i))$$

where this last equality follows from the fact that the Fréchet derivative is linear in the second argument, see e.g. [Rockafellar 1970]. Then $F_\phi(M(\xi^*), \tilde{M}(\omega_i)) \geq 0$ for all ω_i implies that

$$F_\phi\left(M(\xi^*), M(\xi)\right) \geq 0.$$

However, convexity of $\phi(\cdot)$ implies that for any $M_1, M_2 \in \mathcal{M}$

$$\frac{1}{\gamma}[\phi((1-\gamma)M_1 + \gamma M_2) - \phi(M_1)]$$

is a non-decreasing function of γ, see [Whittle 1971]. Hence, setting $\gamma = 1$ yields

$$F_\phi\left(M(\xi^*), M(\xi)\right) \leq \phi(M(\xi)) - \phi(M(\xi^*)),$$

and the sufficiency follows.

Remark. The above theorem can be used to devise simple numerical algorithms for optimization. Consider $\phi(M) = -\log\det(M)$. For a square non-singular matrix M, we have $\partial \log\det M/\partial M = M^{-\top}$ (see e.g. [Goodwin & Payne 1977]), then

$$\begin{aligned} F_\phi\{M(\xi^*), \tilde{M}(\omega)\} &= \lim_{\beta\to 0^+} \frac{\partial}{\partial\beta}\{-\log\det[(1-\beta)M(\xi^*) + \beta\tilde{M}(\omega)]\} \\ &= -\text{trace}([M(\xi^*)]^{-1}[\tilde{M}(\omega) - M(\xi^*)]) \\ &= p - \text{trace}([M(\xi^*)^{-1}\tilde{M}(\omega)]). \end{aligned}$$

Theorem 12.8 then implies that ξ^* is optimal iff

$$v(\xi^*, \omega) = \text{trace}([M(\xi^*)^{-1}\tilde{M}(\omega)]) \leq p, \quad \forall\omega \in [0, \pi/\Delta T].$$

The function $v(\xi^*, \omega)$ is called response dispersion, see [Goodwin & Payne 1977]. The algorithm then simply consists of starting with an initial measure and updating according to

$$\xi_{k+1} = (1-\beta_k)\xi_k + \beta_k\xi_k^\circ$$

where ξ_k° is a single frequency measure with the frequency of support $\omega_k^\circ = \arg\max_\omega v(\xi_k, \omega)$. Since $\xi_{k+1} = \xi_k$ for $\beta_k = 0$ and

$$\lim_{\beta_k\to 0^+} \frac{\partial}{\partial\beta_k}\{-\log\det M(\xi_{k+1})\} = p - v(\xi_k, \omega_k^\circ) < 0$$

for any non-optimal ξ_k, this update leads to a decrease in the value of the criterion for sufficiently small β_k. [Gaffke & Mathar 1992] show that if $\{\beta_k\}$ is chosen such that $\sum_k \beta_k \to \infty$ and $\beta_k \to 0$ as $k \to \infty$, the algorithm converges to the optimum. Notice that the generic technique can be similarly applied to all differentiable convex ϕ functions. ▼

12.5 Design of optimal inputs

We will now evaluate the information matrix in order to determine optimal designs of input signals. Consider for simplicity a scalar discrete time model of the form:

$$y_t = G_1(q)u_t + G_2(q)\epsilon_t \tag{12.25}$$

where $\{u_t\}$ and $\{y_t\}$ are the input and output sequences, respectively, and $\{\epsilon_t\}$ is a sequence of Gaussian random variables with variance σ^2. G_1 and G_2 are transfer functions in the shift-operator. In the following an expression for M_F is derived for the model (12.25). Define ϵ_t as the residual sequence, then:

$$\epsilon_t = G_2^{-1}(q)[y_t - G_1(q)u_t] \ . \tag{12.26}$$

By using this expression in the definition of Fishers information matrix (12.11), we obtain

$$M_F = \mathrm{E}_{Y|\theta}\left\{\frac{1}{\sigma}\sum_{t=1}^{N}(\frac{\partial\epsilon_t}{\partial\boldsymbol{\theta}})(\frac{\partial\epsilon_t}{\partial\boldsymbol{\theta}})^T\right\} + \frac{N}{2\sigma^2}(\frac{\partial\sigma}{\partial\boldsymbol{\theta}})(\frac{\partial\sigma}{\partial\boldsymbol{\theta}})^T \ , \tag{12.27}$$

[Goodwin & Payne 1977] (page 131). Differentiating expression (12.26) and using superposition yields:

$$\frac{\partial\epsilon_t}{\partial\boldsymbol{\theta}} = (\frac{\partial\bar{\epsilon}_t}{\partial\boldsymbol{\theta}}) + (\frac{\partial\tilde{\epsilon}_t}{\partial\boldsymbol{\theta}}) \ , \tag{12.28}$$

where

$$\frac{\partial\bar{\epsilon}_t}{\partial\boldsymbol{\theta}} = -G_2^{-1}(q)(\frac{\partial G_1(q)}{\partial\boldsymbol{\theta}})u_t \ , \tag{12.29}$$

$$\frac{\partial\tilde{\epsilon}_t}{\partial\boldsymbol{\theta}} = -G_2^{-1}(q)(\frac{\partial G_2(q)}{\partial\boldsymbol{\theta}})\epsilon_t \ . \tag{12.30}$$

After substituting these expressions in (12.27) and assuming no feedback in the system ($\{u_t\}$ is independent of $\{\epsilon_t\}$), we get:

$$M_F = \frac{1}{\sigma}\sum_{t=1}^{N}(\frac{\partial\bar{\epsilon}_t}{\partial\boldsymbol{\theta}})(\frac{\partial\bar{\epsilon}_t}{\partial\boldsymbol{\theta}})^T + M_c \ , \tag{12.31}$$

where M_c does not depend upon the choice of input signal.

Now make the following simplifying assumptions :

1. The experiment time (i.e. N) is large.
2. The input $\{u_t\}$ is restricted to the class admitting a spectral representation with spectral distribution function $F(\omega),\ \omega \in [-\pi, \pi]$.

3. The allowable input power is constrained.

Since N is large it is more convenient to work with the average information matrix, which gives the following:

$$\bar{M}_F = \lim_{N\to\infty} \frac{1}{N} M_F = \frac{1}{\pi} \int_0^\pi (\tilde{M}_F(\omega) + \bar{M}_c)\, d\xi(\omega) \ , \tag{12.32}$$

where $\xi(\omega)$ is defined by

$$d\xi(\omega) = \begin{cases} \frac{1}{2} dF(\omega) \text{ for } \omega = 0, \omega = \pi \\ dF(\omega) \text{ for } \omega \in]0, \pi[\end{cases}$$

and

$$\tilde{M}_F(\omega) = \mathrm{Re}\left[\frac{1}{\sigma} [\frac{\partial G_1(e^{j\omega})}{\partial \boldsymbol{\theta}}] G_2^{-1}(e^{j\omega}) \times G_2^{-1}(e^{-j\omega}) [\frac{\partial G_1(e^{-j\omega})}{\partial \boldsymbol{\theta}}]^T \right] \tag{12.33}$$

and

$$\bar{M}_c(\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} [\frac{\partial G_2(e^{j\omega})}{\partial \boldsymbol{\theta}}] G_2^{-1}(e^{j\omega}) G_2^{-1}(e^{-j\omega}) [\frac{\partial G_2(e^{-j\omega})}{\partial \boldsymbol{\theta}}]^T d\omega + \frac{1}{2\sigma^2} (\frac{\partial \sigma}{\partial \boldsymbol{\theta}}) (\frac{\partial \sigma}{\partial \boldsymbol{\theta}})^T \ . \tag{12.34}$$

It should be remarked that both $\tilde{M}_F(\omega)$ and $\bar{M}_c(\omega)$ are dependent upon the parameters $\boldsymbol{\theta}$. This means that if a local design is used then only $\tilde{M}_F(\omega)$ need to be considered. On the other hand if a Bayesian design like (12.24) is used then both $\tilde{M}_F(\omega)$ and $\bar{M}_c(\omega)$ must be evaluated.

The power restriction of the input signal can be formulated as

$$P_u = \frac{1}{\pi} \int_0^\pi d\xi(\omega) = 1 \ . \tag{12.35}$$

We are now ready to give the following theorem, which states that it is always possible to find an optimal input comprising a finite number of sinusoids.

Theorem 12.9. *For any power constrained design $\xi_1(\omega)$ with corresponding $n \times n$ average information matrix $\bar{M}_F(\xi_1)$, there always exists a power constrained design $\xi_2(\omega)$ which is piecewise constant with at most $n(n+1)/2 + 1$ discontinuities and $\bar{M}_F(\xi_2) = \bar{M}_F(\xi_1)$. For the design criterion $J = -\log\det \bar{M}_F$, optimal designs exist comprising not more than $n(n+1)/2$ sinusoids.*

Proof. Only an outline of the proof is given. It can be shown that the set of all average information matrices corresponding to input power constrained designs is the convex hull of the set of all average information matrices corresponding to single frequency designs. Hence, from Caratheodory's theorem (see [Fedorov 1972]) the result for the first part of the theorem follows. For a convex function of the information matrix the optimal design is a boundary point of the convex hull, therefore one less sinusoidal component is needed. For the complete proofs see [Fedorov 1972, Goodwin & Payne 1977]. ■

12.5.1 Bayesian Approach

We now turn to the Bayesian approach, considering the generalized criterion (12.24). Assume that the prior knowledge of the system is given in terms of a prior distribution of the parameters. Hence we are able to evaluate the expectation of the considered criterion with respect to the prior distribution of the parameters cf. (12.24), instead of simply evaluating the criterion at some fixed values of the parameters. In general the resulting design will be different for the two approaches, cf. the example following. In the Bayesian case it is possible to prove a theorem similar to theorem 12.9.

Theorem 12.10. *Using $J = \mathrm{E}_\theta\{-\log\det \bar{M}\}$ as criterion, optimal designs exist comprising not more than $n(n+1)/2$ sinusoidal components.*

Proof. The criterion may be written as

$$J = \int_{\Omega_\theta} -\log\det \bar{M}(\xi(\omega), \boldsymbol{\theta}))\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}\ , \tag{12.36}$$

where $\Omega_\theta \subset \mathbb{R}^n$ is assumed to be a closed and bounded interval of the parameters and $p(\boldsymbol{\theta})$ is the prior probability density of the parameters. The mean-value theorem states, that for all $\xi(\omega)$ there exists a $\boldsymbol{\theta}^* \in \Omega_\theta$ such that

$$J = K(-\log\det \bar{M}(\xi(\omega), \boldsymbol{\theta}^*)\, p(\boldsymbol{\theta}^*)) \tag{12.37}$$

for some constant K which is independent of $\xi(\omega)$ and $\boldsymbol{\theta}^*$. It is seen that (12.37) has the form of the criterion considered in the previous theorem. The difference is that now $\boldsymbol{\theta}^*$ depends upon $\xi(\omega)$. But since (12.32) is still valid, we conclude that the set of all average information matrices is the convex hull of all average information matrices corresponding to single frequency designs, and the proof follows immediately. ■

Since in (12.37) $\boldsymbol{\theta}^*$ depends upon $\xi(\omega)$ this is a more complicated optimization problem than the previously considered. In a practical application, though, it is not necessary to actually find $\boldsymbol{\theta}^*$, one would use the result of the theorem and apply it to the

criterion (12.36) directly. From the proof it is readily seen that in the general case (12.24) it is also possible to find an optimal design comprising a finite number of sinusoids, as long as ϕ is a continuous convex function of the information matrix.

The criteria discussed so far have all been based on ML estimation of the parameters. If instead a MAP estimator is used to estimate the unknown parameters the criteria for optimality must be changed accordingly. The following theorem establishes a relation between the posterior covariance of the parameters and Fishers information matrix when a MAP estimator is used.

Theorem 12.11. *Assume that the prior knowledge about the model parameters is embodied in a Gaussian distribution with covariance matrix $\mathbf{\Sigma}_{pre}$. Also assume that a MAP estimator is used to estimate the unknown parameters based on sampled observations for a model brought into the regression form*

$$\boldsymbol{y}_t = \boldsymbol{\varphi}_t^T \boldsymbol{\theta} + \boldsymbol{\epsilon}_t$$

where $\{\boldsymbol{\epsilon}_t\}$ is a sequence of Gaussian random variables with known covariance, uncorrelated with $\{\boldsymbol{\varphi}_t\}$. Then the posterior covariance matrix $\mathbf{\Sigma}_{post}$ is given by

$$\mathbf{\Sigma}_{post}^{-1} = \mathbf{\Sigma}_{pre}^{-1} + \boldsymbol{M}_F = \mathbf{\Sigma}_{pre}^{-1} + N\bar{\boldsymbol{M}}_F \tag{12.38}$$

$\boldsymbol{M}_F$ is Fishers information matrix, $\bar{\boldsymbol{M}}_F$ the average information matrix, and N the length of experiment.

Proof. See [Sadegh et al. 1994]. ■

In the following, the concept of information is related to Lindley's measure of average information. In this way, we are able to formulate design criteria also based on MAP estimators. First some definitions are needed.

Definition 12.12. The entropy of a random variable X having probability density function $p(X)$ is defined as

$$H_x = -E_X[\log p(X)] \; . \tag{12.39}$$

▲

Definition 12.13. Lindley's measure of the average amount of information provided by an experiment ξ with data y and parameters θ is defined as

$$J(\xi) = H_\theta - E_y[H_{\theta|y}] \; . \tag{12.40}$$

▲

Now the relation between Fishers information matrix and Lindley's measure of average information can be established via the following theorem [Sadegh et al. 1994].

Theorem 12.14. *With the same assumptions as in Theorem 12.11, maximizing Lindley's measure of the average amount of information, $J(\xi)$, is equivalent to solving the optimization problem*

$$\begin{array}{l} min_{\xi} J \\ J = -E_{\theta}[\log\det\{\boldsymbol{\Sigma}_{pre}^{-1} + \boldsymbol{M}_F\}] \end{array} \tag{12.41}$$

Proof. The estimation is regarded as a means by which further information about the system parameters is provided. Since MAP is the mode of the posterior distribution, the maximum amount of information with respect to Lindley's measure is obtained by MAP. Since the prior distribution of the parameters and the distribution of observations given θ is Gaussian, the posterior distribution of the parameters is also Gaussian. Denote the posterior mean and covariance by $\bar{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{post}$. From (12.38) we have

$$\boldsymbol{\Sigma}_{post}^{-1} = \boldsymbol{\Sigma}_{pre}^{-1} + \boldsymbol{M}_F$$

From Definition 12.13

$$\begin{aligned} J(\xi) &= -\,\mathrm{E}_{\theta}\{\log p(\boldsymbol{\theta})\} + E_y E_{\theta|y}\{\log p(\boldsymbol{\theta}|\boldsymbol{y})\} && (12.42) \\ &= -\,\mathrm{E}_{\theta}\{\log p(\boldsymbol{\theta})\} + \mathrm{E}_y\,\mathrm{E}_{\theta|y}\{-\frac{n}{2}\log(2\pi)\} \\ &\quad -\,\mathrm{E}_y\,\mathrm{E}_{\theta|y}\left\{\frac{1}{2}\log det\Sigma_{post} + \frac{1}{2}(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})^T\boldsymbol{\Sigma}_{post}^{-1}(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})\right\} && (12.43) \end{aligned}$$

As all the other terms obviously are constants, we only focus on the last two terms

$$\begin{aligned} &E_y E_{\theta|y}\{\frac{1}{2}\log\det\boldsymbol{\Sigma}_{post}\} \\ &\qquad = E_{\theta}E_{y|\theta}\{\frac{1}{2}\log\det\boldsymbol{\Sigma}_{post}\} && (12.44) \\ &\qquad = E_{\theta}\{\frac{1}{2}\log\det\boldsymbol{\Sigma}_{post}\} && (12.45) \\ &\qquad = -E_{\theta}\{\frac{1}{2}\log\det\boldsymbol{\Sigma}_{post}^{-1}\} && (12.46) \end{aligned}$$

The other term can be written as

$$\begin{aligned} &E_y E_{\theta|y}\{\frac{1}{2}(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})^T\boldsymbol{\Sigma}_{post}^{-1}(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})\} \\ &\qquad = E_y E_{\theta|y}\{\frac{1}{2}\,\mathrm{tr}\,\boldsymbol{\Sigma}_{post}^{-1}(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})^T\} && (12.47) \\ &\qquad = E_y\{\frac{1}{2}\,\mathrm{tr}\,\boldsymbol{\Sigma}_{post}^{-1}E_{\theta|y}[(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})^T]\} && (12.48) \\ &\qquad = E_y\{\frac{n}{2}\} = \frac{n}{2} && (12.49) \end{aligned}$$

where n is the number of parameters. Now using (12.38) establishes the theorem. ■

An approximation of the mean value in (12.41) is obtained by setting the parameters equal to their prior mean values. This approximation, which corresponds to a local design, simplifies the computations considerably. Thus depending on the estimation method, ML or MAP, and the choice of local or average criterion the following criteria would be of interest:

$$J_1 = -[\log\det(\bar{M}_F)]_{\theta=\mathrm{E}\{\theta\}} \tag{12.50}$$

$$J_2 = -[\log\det(N\bar{M}_F + \Sigma_{pre}^{-1})]_{\theta=\mathrm{E}\{\theta\}} \tag{12.51}$$

$$J_3 = -\,\mathrm{E}_\theta[\log\det(\bar{M}_F)] \tag{12.52}$$

$$J_4 = -\,\mathrm{E}_\theta[\log\det(N\bar{M}_F + \Sigma_{pre}^{-1})] \tag{12.53}$$

These criteria demonstrate different levels of including partial prior information about parameter values. Optimization with respect to J_1 results in designs which are strongly dependent upon the prior information. The dependence is even more pronounced in J_2. It may therefore be wise to perform a sensitivity analysis, i.e. determine the sensitivity of the design to changes in the parameters, when using these criteria. An alternative is to choose an average criterion like J_3 and J_4. Table 12.1 summaries the relation between the choice of estimators and the optimization criterion, expressed in both a local and average form. Applications with these criteria are found in e.g. [Melgaard et al. 1993, Sadegh et al. 1994].

	ML	MAP
Local design	J_1	J_2
Bayesian/average	J_3	J_4

Table 12.1: Summary of the optimality criteria

12.6 Sampling time and presampling filters

When estimating the parameters in a continuous time model from sampled data, the parameters will in general depend on the input to the system, the sampling instants and the presampling filter. The aims of designing optimal sampling instants and presampling filters is to avoid loss of information from the data due to sampling. Consider the following scalar continuous time system

$$y(s) = G(s)\,u(s) + \eta(s)\ , \qquad s = j\omega\ , \tag{12.54}$$

where u and y are the input and output, respectively, and η denotes colored measurement noise having spectral density $\psi(\omega)$ for all $\omega \in]-\infty, \infty[$.

Assume that the input signal spectrum is band limited to $[-\omega_h, \omega_h]$. The case of uniform sampling interval is considered. Suppose that the output is sampled faster than the Nyquist rate for ω_h, i.e.

$$\omega_s > 2\,\omega_h \tag{12.55}$$

where $\omega_s = 2\pi\, f_s$ and f_s is the sampling frequency. Such a sampler does not distort the part of the output spectrum arising from the input. The part of the output spectrum arising from the noise will, however, be distorted due to aliasing. The aliased noise spectrum is a superposition of different parts of the original spectrum

$$\psi_s(\omega) = \psi(\omega) + \sum_{r=1}^{\infty}(\psi(\omega + r\omega_s) + \psi(\omega - r\omega_s)) \tag{12.56}$$

for $\omega \in\,]-\omega_s/2, \omega_s/2[$. From expression (12.56) it is seen that $\psi_s(\omega) - \psi(\omega)$ is positive definite. It follows that the experiment with sampled data cannot be better than the corresponding experiment with continuous observations, see [Goodwin & Payne 1977]. However by inclusion of a suitable presampling filter, equality can be achieved.

Assume the presampling filter has the transfer function F with the ideal property

$$|F(j\omega)| = 0\ , \qquad \text{for } \omega \le -\omega_s/2 \text{ or } \omega \ge \omega_s/2\ , \tag{12.57}$$

and invertible otherwise. By including the presampling filter, the filtered and sampled noise spectrum $\psi_{fs}(\omega)$ is

$$\psi_{fs}(\omega) = \psi_f(\omega) + \sum_{r=1}^{\infty}(\psi_f(\omega + r\omega_s) + \psi_f(\omega - r\omega_s)) \tag{12.58}$$

and $\psi_f(\omega)$ is the filtered noise spectrum given by

$$\psi_f(\omega) = F^T(j\omega)\psi(\omega)F(-j\omega)\ . \tag{12.59}$$

For a presampling filter satisfying (12.57), $\psi_{fs}(\omega)$ reduces to $\psi_f(\omega)$. This shows that the specified ideal presampling filter F eliminates the information loss due to sampling.

The approach discussed so far for optimal design of sampling time and presampling filter has considered the average information matrix per unit time. This is useful for design of experiments with a fixed experiment *time*, but without constraints on the number of samples. This leads to separate design of optimal input followed by choice of sampling time and presampling filter according to (12.55) and (12.57). However, if the total number of samples is constrained, then it is more appropriate to consider the average information matrix per *sample*. If the input spectrum is band limited to

$[-\omega_h, \omega_h]$ and the sampling time and presampling filter is chosen according to (12.55) and (12.57) as before, then the average information matrix per sample is given by

$$\tilde{M}_F = \bar{M}_F / f_s \ , \tag{12.60}$$

where $\bar{M}_F$ is the average information matrix per unit time and f_s is the sampling rate. Optimizing a criterion based on this expression leads to a joint optimal design of input and sampling time for experiments with constrained number of samples. Such a design will in general lead to a compressed optimal input spectrum with a lower sampling rate and hence increased experiment time compared with the continuous observation case [Goodwin & Payne 1977].

12.7 Use of different criteria

In the following a simple example is given on the use of some of the different criteria for optimal design of input signal. The optimal design of sampling time is not covered by this example, a fixed sampling time is used. Consider a first order stochastic differential equation with discrete observations:

$$dx(t) = -ax(t)\,dt + bu(t)\,dt + dw(t) \ , \tag{12.61}$$

$$y_k = x_k + e_{2,k} \ , \tag{12.62}$$

where $w(t)$ is a Wiener process with $\mathrm{E}[dw(t)] = 0$, $\mathrm{V}[dw(t)^2] = r\,dt$, and $e_{2,k}$ is a Gaussian white noise process with $\mathrm{V}[e_{2,k}] = r_2$. $dw(t)$ and $e_{2,k}$ are assumed to be independent; subscript k is shorthand for t_k. We are interested in the set of parameters $\theta = [a, b]^T$, and to find an optimal input signal for the system. Since the inputs are assumed to be constant within a sampling interval, the corresponding discrete time model is written:

$$x_{k+1} = \alpha x_k + \beta u_k + e_{1,k} \ , \tag{12.63}$$

where α, β and $\mathrm{V}[e_{1,k}]$ depends on the sampling time τ by the following expressions:

$$\alpha = e^{-a\tau} \ , \qquad \beta = \int_0^\tau e^{-as} b\,ds = \frac{b}{a}(1 - e^{-a\tau}) \ ,$$

$$V(e_{1,k}) = \int_0^\tau e^{-as} r e^{-as}\,ds = \frac{r}{2a}(1 - e^{-2a\tau}) \ .$$

The discrete time model is brought into the innovations form

$$\hat{x}_{k+1} = \alpha \hat{x}_k + \beta u_k + K\epsilon_k \ , \tag{12.64}$$

$$y_k = \hat{x}_k + \epsilon_k \ , \tag{12.65}$$

where

$$V(\epsilon_k) = \bar{P} + r_2 \ , \tag{12.66}$$

$$K = \alpha\bar{P}(\bar{P} + r_2)^{-1} \ , \tag{12.67}$$

and $\bar{P}$ is the positive semidefinite solution of the stationary Ricatti equation,

$$\bar{P} = \alpha^2\bar{P} + r_1 - (\alpha\bar{P})^2(\bar{P} + r_2)^{-1} \ . \tag{12.68}$$

The innovations form of the model (12.64) and (12.65) may be rewritten as:

$$y_k = \frac{\beta q^{-1}}{1 - \alpha q^{-1}} u_k + \frac{1 + (K - \alpha)q^{-1}}{1 - \alpha q^{-1}} \epsilon_k \ . \tag{12.69}$$

Since equation (12.69) is of the form (12.25) we are now ready to formulate the average Fishers information matrix for the problem. Considering a power constrained input in the frequency domain we use equations (12.32), (12.33) and (12.34).

By considering the criterion $\min_\xi J_1 = \min_\xi(-\log\det \bar{M}_F)$, Theorem 12.9 states that it suffices to look for designs comprising no more than 3 sinusoids for this example,

$$\bar{M}_F(\xi) = \sum_{k=1}^{3} \mu_k \bar{M}_F(\omega_k) \ , \tag{12.70}$$

where $\mu_k \geq 0$ and $\sum \mu_k = 1$. $\bar{M}_F(\omega_k)$ can be separated into $\tilde{M}_F(\omega_k)$ and $\bar{M}_c$ according to (12.33) and (12.34). It should be remarked, that when a local design is considered, only $\bar{M}_F(\omega_k)$ need to be calculated, but when a Bayesian design is used the total $\bar{M}_F(\omega_k)$ must be calculated. For the calculation of the matrices in this example the computer algebra language, MAPLE, was used. Hence the optimization problem is formulated and some general available software for solving nonlinearly constrained minimization can be applied.

It turns out, that a single sinusoid is sufficient for optimality in the considered case. A plot of this optimal input frequency for different values of a is shown below for $b = 1.0$, $r = 1$, $r_2 = 0.1$ and $\tau = 1$.

Now we would like to compare this traditional local design of experiments with the Bayesian approach for a given prior probability density of the parameters. Hence we consider the criterion $\min_\xi J_2 = \min_\xi E_\theta(-\log\det \bar{M}_F)$. It is assumed that parameter a is uniformly distributed in the interval $[0.5, 1.1]$ and the other parameters are fixed to the same values as above. The optimal design in this case is

$$\omega_2^* = 0.4575 \ , \qquad J_2(\omega_2^*) = 1.1951 \ . \tag{12.71}$$

This should be compared to the previous design using the expected value of the parameters, $\mathrm{E}[a] = 0.8$, which has the optimal design, according to J_1

$$\omega_1^* = 0.6606 \ , \qquad J_1(\omega_1^*) = 0.6107 \ , \tag{12.72}$$

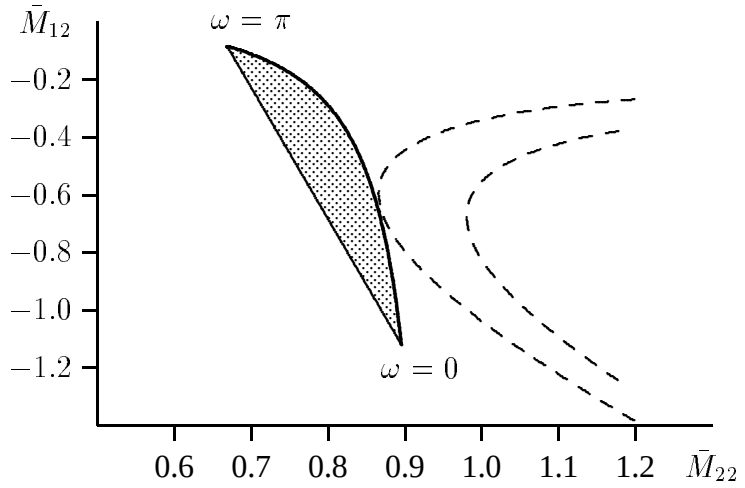

Figure 12.4: *Convex hull of single frequency designs for the example and contours of constant* $\det \bar{M}_F$ *(increasing to the right).*

hence, here the two approaches give a different design. The average information when a Bayesian design is used is smaller than the information from the local design, when the used $\mathrm{E}[a]$ is true.

Instead of optimizing the average performance with respect to the prior distribution one might be interested in a design which optimize the worst possible design. A plot of the performance of single frequency designs for different values of a is given below.

Consider the minimax criterion

$$\min_{\xi} J_3 = \min_{\xi}(\max_{a\in\Theta} - \log\det \bar{M}_F) \tag{12.73}$$

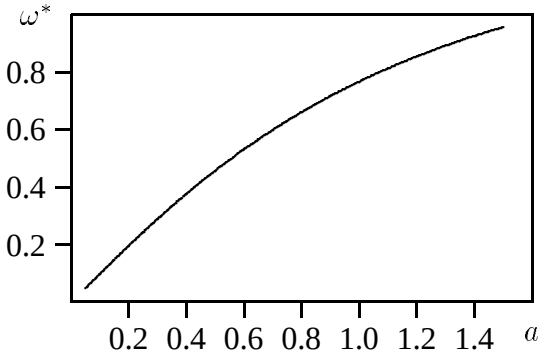

Figure 12.5: Optimal design of input frequency ω^* for different values of a.

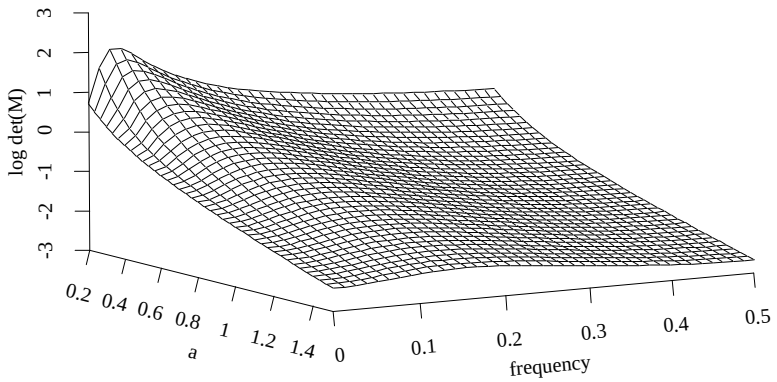

Figure 12.6: The performance of single frequency designs for different values of a. The frequency axis has the scale $f = 2\pi\omega$.

Minimizing this criterion for the example with $a \in [0.5, 1.1]$ and the other parameters as above gives the optimal design

$$\omega_3^* = 0.8128 \ , \qquad J_3(\omega_3^*) = 1.3741 \ , \tag{12.74}$$

which is seen to differ from the previous designs. The value of the criterion expresses the worst possible performance from this design depending on the true value of a.

12.8 Experiment Design and Physical Models

As mentioned previously physical significance and partial prior information are among the key features of gray-box models. Physical significance of the model is typically equivalent to nonlinearity of the model parameterization. Even, if attention is restricted to linear systems, the parameterization should be allowed to be nonlinear. In the present section, we briefly discuss the design of optimal experiments for models with physical significance. All the assumptions of the previous sections are also valid here. In the next section, we study the impact of the other feature of gray-box models, namely, the availability of partial prior information.

Suppose that a physical parameter θ_p is related to θ through $\theta = F(\theta_p)$. Then if F is differentiable, a simple application of the chain rule and Definition **??** yield that

$$M_p = (\frac{\partial F(\theta_p)}{\partial \theta_p})^\top M(\frac{\partial F(\theta_p)}{\partial \theta_p}),$$

where M_p is the average information matrix corresponding to parameterization θ_p. Once the single frequency average information matrices $\tilde{M}(\omega)$ for the parameterization θ are computed, it is possible to readily compute the single frequency average

information matrices for the reparameterized model using the above formula. However, it is not readily clear how a solution for the parameterization θ translates to a solution for the parameterization θ_p, and in general, the only way to find the optimal inputs for the reparameterized model is to solve a new optimization problem.

The other point is that models in continuous-time are more interesting in a gray-box setting since the physical knowledge is normally expressed as differential equations. Fortunately, only minor modifications are needed for using the developed theory to continuous-time design problems. Assume that a linear continuous-time model is given by

$$y(t) = G_1(\delta)u(t) + G_2(\delta)\epsilon(t)$$

where $\epsilon(t)$ is continuous-time white noise. Here, δ is the differentiation operator, $u(t)$ and $y(t)$ are input and output signals, respectively, and G_1 and G_2 are transfer functions. All the discussions concerning experimental design for discrete-time models remain valid if the following modifications are made [Goodwin & Payne 1977]:

- replace $e^{\jmath\omega\Delta T}$ by $\jmath\omega$,
- the integral limits are obtained by letting $\Delta T \to 0$.

The convexity of the information matrix and the rest of the theory remain intact.

Example 12.8 (Design for Estimation of the Heat Dynamics of a Building). In this example, we investigate the design of optimal experiments for identification of a physical model that describes the heat transfer dynamics of a building, see [Madsen & Holst 1995] for a further discussion of this problem. The model is

$$\begin{aligned}\begin{bmatrix} dT_m \\ dT_i \end{bmatrix} &= \begin{bmatrix} \frac{-1}{R_i C_m} & \frac{1}{R_i C_m} \\ \frac{1}{R_i C_i} & -(\frac{1}{R_a C_i} + \frac{1}{R_i C_i}) \end{bmatrix} \begin{bmatrix} T_m \\ T_i \end{bmatrix} dt \\ &+ \begin{bmatrix} 0 & 0 & \frac{A_w p_s}{C_m} \\ \frac{1}{R_a C_i} & \frac{1}{C_i} & \frac{A_w(1-p_s)}{C_i} \end{bmatrix} \begin{bmatrix} T_a \\ Q_h \\ Q_s \end{bmatrix} dt + \begin{bmatrix} dw_m(t) \\ dw_i(t) \end{bmatrix}\end{aligned} \tag{12.75}$$

where the following physical parameters are to be considered:

- R_i – resistance between the room air and the large heat accumulating medium,
- R_a – resistance between inner walls and the ambient air,
- C_i – capacity of the small heat accumulating medium (the indoor air and the inner part of the walls, see [Madsen & Holst 1995]),

- C_m – capacity of the large heat accumulating medium.

The indoor air temperature is measured, and both process and measurement noise terms are considered. The remaining entities in the model (12.75) are

- T_i – temperature of the small heat accumulating medium (indoor air temperature),
- T_m – temperature of the large heat accumulating medium,
- A_w – effective window area,
- p_s – part of the solar radiation which reaches the large heat accumulating medium in the building,
- T_a – ambient air temperature,
- Q_h – heat supplied by heater (controlled input),
- Q_s – heat supplied by solar radiation,
- $w_m(t), w_i(t)$ – elements of a vector Wiener process introduced in order to describe the disturbances and errors in the system model.

For the input design we only consider the influence on the temperatures in the building caused by the heater, Q_h, i.e. the effects of solar radiation, Q_s, and ambient air temperature, T_a, are neglected. Since only the indoor temperature, T_i, is measured the state-space model can be transformed into a single-input single-output transfer function model where the input is Q_h. We define the parameter vector $(R_i, R_a, C_i, C_m, K_1, K_2, \lambda)^\top$ where K_1, K_2, and λ are Kalman gains in the state estimator and the covariance of the innovation process respectively (see [Madsen & Holst 1995]). We assume that the prior mean of the parameter is given by $(1, 1, 1, 4, 0.2, 0.5, 0.01)^\top$.
The D-optimal input, $u_o(t)$ comprises sinusoidal components with frequencies

$$\omega_1 = 0.088 \quad , \quad \omega_2 = 1.407,$$

and corresponding power shares

$$\alpha_1 = 0.5 \quad , \quad \alpha_2 = 0.5.$$

This design is composed of a low frequency and a high frequency component corresponding to the large and small capacities of the heat accumulating media. Using the D_s-optimality criterion with the interesting parameter being the small capacitance C_i, the optimal input $u_{o1}(t)$ comprises sinusoids with

$$\omega_{11} = 0.143 \quad , \quad \omega_{21} = 2.644,$$

and corresponding power shares

$$\alpha_{11} = 0.052 \quad , \quad \alpha_{21} = 0.948.$$

Notice the considerable shift towards high frequencies.

The design with C_m as the interesting parameter is the signal $u_{o2}(t)$ with frequencies

$$\omega_{12} = 0.085 \quad , \quad \omega_{22} = 0.755,$$

and corresponding power shares

$$\alpha_{12} = 0.812 \quad , \quad \alpha_{22} = 0.188.$$

Notice the considerable shift towards low frequencies.

We then apply these input signals and estimate the parameters 100 times (cross-sections) for each input. The results are gathered in Table 12.2 where the numbers are obtained by averaging over the relevant values for the 100 cross-sections. While $u_o(t)$ is the D-optimal signal for estimating all the parameters (minimizes the determinant of the covariance matrix of the estimate), the other D_s-optimal inputs provide most information about the interesting parameters, C_i and C_m, respectively. The estimation for each cross-section is based on 5000 samples where the sampling time is $\Delta T = 0.05$.

	$u_o(t)$	$u_{o1}(t)$	$u_{o2}(t)$
$\log\det(\cdot)$	-33.13	-32.92	-32.85
$\log\sigma^2_{\hat{C}_i}$	-6.17	-7.42	-4.98
$\log\sigma^2_{\hat{C}_m}$	-0.11	0.80	-0.86

Table 12.2: Comparison between D-optimal and D_s-optimal designs. In the table, $\det(\cdot)$ denotes the determinant of the covariance of the estimate $\hat{\theta}_p$, $\sigma^2_{\hat{C}_i}$ the variance of the estimate $\hat{C}_i$, and $\sigma^2_{\hat{C}_m}$ the variance of the estimate $\hat{C}_m$.

♦

12.9 Solution for Bayesian Criteria

As discussed earlier, the optimization criterion for input design is in general a function of the unknown parameter value. In this section, we consider optimization of the Bayesian criteria, i.e. criteria of the form

$$E_\theta\{\phi(M)\} \tag{12.76}$$

which take care of the fact that the prior knowledge about the parameter value is partial.

In this connection, it is of interest to study if the theory of the previous sections can be fully or partially employed. Especially, it is of interest to investigate whether or not:

- the Bayesian criteria admit the use of standard experimental design algorithms,
- the upper bound on the number of sinusoidal components that suffice for optimality (Theorem 12.6) still exists.

Attention will be restricted to discrete-time systems. The extension for continuous-time systems may be accomplished in the same way as in Section 12.8. Focus on the first question.

Theorem 12.15. *Fix the input frequency set*

$$\Omega = \{\omega_i | i = 1, ..., l, \omega_i \in [0, \pi/\Delta T]\}.$$

Define

$$f(\alpha) = E_\theta\{\phi(M(\xi))\}$$

where $\phi(\cdot)$ is convex and $\alpha = (\alpha_1, ..., \alpha_l)^\top$ gives the power share of the power restricted input measure $\xi \in \mathcal{X}$ on Ω. An input measure ξ^* characterized by α^* minimizes the Bayesian criterion in equation (12.76) among all other power restricted input measures whose frequencies are contained in Ω if and only if

$$\lim_{\beta \to 0^+} \frac{\partial}{\partial \beta} f((1-\beta)\alpha^* + \beta\alpha^{(i)}) \geq 0$$

where $\alpha^{(i)}$ is a unit vector with 1 in the entry i.

Proof. The key observation is that when $\phi(\cdot)$ is a convex function of M then $f(\cdot)$ is a convex function of α. To see why, take any two power shares α and α' and denote their corresponding measures by ξ and ξ'. Then for any $0 < \beta < 1$

$$\begin{aligned} f((1-\beta)\alpha + \beta\alpha') &= E_\theta\{\phi((1-\beta)M(\xi) + \beta M(\xi'))\} \\ &\leq (1-\beta)E_\theta\{\phi(M(\xi))\} + \beta E_\theta\{\phi(M(\xi'))\} \\ &= (1-\beta)f(\alpha) + \beta f(\alpha'). \end{aligned}$$

As a result of input power restriction

$$\sum_{i=1}^{l} \alpha_i = 1, \quad \alpha_1, ..., \alpha_l \geq 0,$$

the set of all feasible α's denoted by $\mathcal{A}$ is a convex set. $\mathcal{A}$ is in fact the convex hull of the points $\alpha^{(i)}$, $i = 1, ..., l$. The rest of the proof follows in a way similar to the proof of Theorem 12.8 after replacing $\mathcal{M}$ by $\mathcal{A}$ and $\phi : \mathcal{M} \to \mathbb{R}$ by $f : \mathcal{A} \to \mathbb{R}$. ■

Remark. Here, we use Theorem 12.15 to derive an algorithm for constructing optimal input measure for the criterion $E_\theta\{-\log\det(M)\}$. In fact, we wish to find the optimal power weight α^* on a *given* set of frequencies Ω. This set can for example be a fine grid on the frequency axis. Similar to Remark 12.4.1

$$\lim_{\beta\to 0^+} \frac{\partial}{\partial\beta} f((1-\beta)\alpha^* + \beta\alpha^{(i)}) =$$
$$\lim_{\beta\to 0^+} \frac{\partial}{\partial\beta} E_\theta\{-\log\det[(1-\beta)M(\xi^*) + \beta\tilde{M}(\omega_i)]\} =$$
$$p - E_\theta\{\text{trace}([M(\xi^*)^{-1}\tilde{M}(\omega_i)])\}$$

where ξ^* denotes the input measure characterized by α^*. Then, ξ^* is optimum if and only if

$$v_B(\xi^*, \omega_i) = \text{trace}(E_\theta\{[M(\xi^*)^{-1}\tilde{M}(\omega_i)]\}) \leq p.$$

Thus, start with an initial power share on Ω and update exactly as in Remark 12.4.1 noting that $\omega_k^\circ = \arg\max_{\omega\in\Omega} v_B(\xi_k, \omega)$. The proof of convergence is analogous to the case described in Remark 12.4.1, see [Gaffke & Mathar 1992]. Again the generic technique can be applied to optimization of the criteria of equation (12.76) for all differentiable convex ϕ functions. ▼

Remark. In Theorem 12.15 and Remark 12.9, the frequency set Ω is fixed. They state a necessary and sufficient condition for the optimal power weights on Ω and an algorithmic search to find these weights. However, Theorem 12.8 and Remark 12.4.1 required no fixed Ω. What was fixed was actually the convex set $\mathcal{M}$ where the ϕ-function was defined and that was exactly the convexity of $\mathcal{M}$ which led to the theorem on the upper bound on the number of sinusoidal components (Theorem 12.6). When solving for Bayesian criteria, this convex set is no longer fixed since each point on the boundary of $\mathcal{M}$ is a function of the parameter and the parameter is not fixed any longer. Hence, the answer to the second question, in general nonlinear regression cases, is unfortunately negative. The reason is that no fixed convex hull of information matrices corresponding to single frequency design measures exist. However, the numerical search algorithms in many cases yield measures which essentially comprise finitely many sinusoidal components, see the examples below. ▼

Example 12.9. Assume that the following system is given

$$y_t = \frac{bq^{-1}}{1 + aq^{-1}} u_t + \epsilon_t$$

where the unknown parameter is $\theta = (a, b)^\top$ and the covariance of the noise sequence is known.

Assume that the prior information about a is embedded in a Gaussian distribution with mean $m_a = 0.8$ and variance $\sigma_a^2 = (0.08)^2$. Fix $b = 1$ and use the algorithm given in Remark 12.9 to find the D-optimal power weights on a fine grid on the ω axis denoted by Ω. The set Ω is obtained by dividing [2.95,3.19] into 20 intervals of equal length. This procedure yields the design ξ^*, essentially comprising frequencies

$$\omega_1 = 3, \quad \omega_2 = 3.01, \quad \omega_3 = 3.14,$$

and power shares

$$\alpha_1 = 0.1079, \quad \alpha_2 = 0.7377, \quad \alpha_3 = 0.1544,$$

respectively.

To confirm that this design is indeed optimal, $v_B(\xi^*, \omega)$ is plotted as a function of $\omega \in \Omega$ in Figure 12.7. Since $v_B(\xi^*, \omega) \leq p$, Theorem 12.15 suggests that ξ^* is optimal.

If also a is fixed in the design computation at $a = 0.8$, the design measure consists of a single frequency component $\omega^* = 3.01$.

♦

Example 12.10. Now consider a second order continuous-time model with $G_1(s) = 1/[(s + a)(s + b)]$ and $G_2(s) = 1$. The unknown parameter is $\theta = (a, b)^\top$ and the noise is wide band white with known covariance. Fix $b = 2$ and assume that the prior information about a is embedded in a Gaussian distribution with mean $m_a = 1$ and variance $\sigma_a^2 = 0.4^2$. Similar to the previous example, we find the D-optimal measure ξ^* with

$$\omega_1 = 0, \quad \omega_2 = 0.43,$$

and

$$\alpha_1 = 0.0218, \quad \alpha_2 = 0.9782.$$

The frequency grid Ω is obtained by dividing [0,1.5] to intervals of length 0.1 and 0.01 adapted to the area around 0.5.

A plot of $v_B(\xi^*, \omega)$ as a function of $\omega \in \Omega$ is given in Figure 12.8 which confirms that ξ^* is indeed optimal ($v_B(\xi^*, \omega) \leq p$).

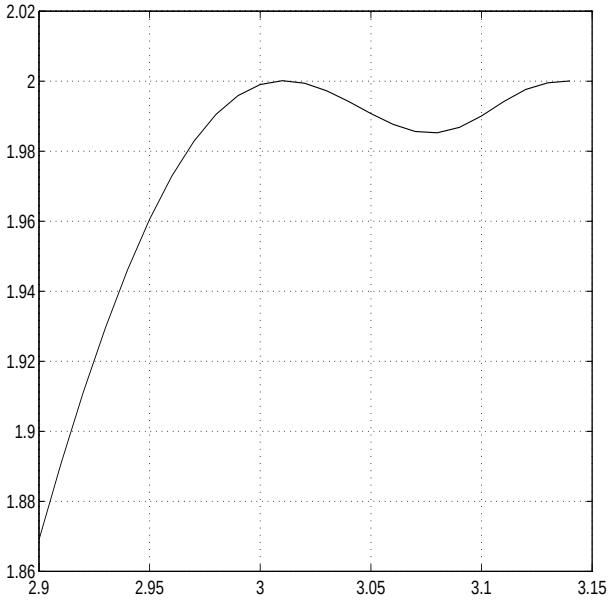

Figure 12.7: $v_B(\xi^*, \omega)$ *as a function of* $\omega \in \Omega$ *in Example 5.2.*

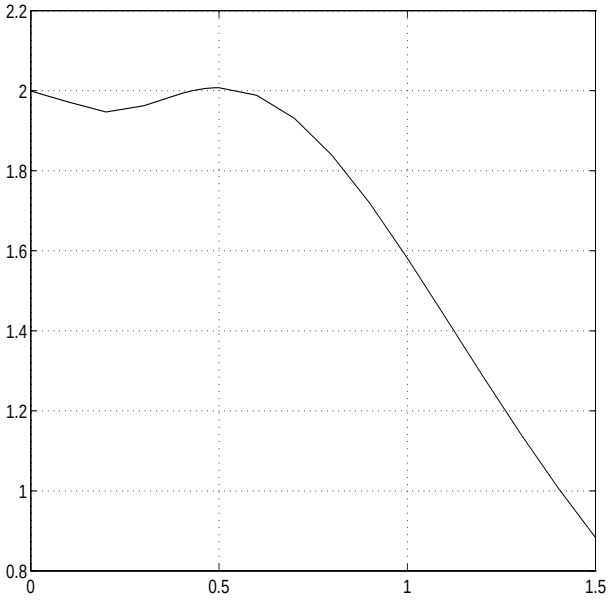

Figure 12.8: $v_B(\xi^*, \omega)$ *as a function of* $\omega \in \Omega$ *in Example 5.3.*

Fixing $a = 1$ yields a single frequency design measure with $\omega^* = 0.49$.

♦

12.9.1 Conclusion

The design of optimal input signals for identification of gray-box models has been studied. Physical significance and partial prior information are identified as the key features of gray-box models. Considering them as part of the experimental design procedure necessitates extensions of the design theory, which are studied. Probability distributions are used as means for expressing partial prior information.

The connection between design for a model with some input-output parameterization and a physical model is conceptually as simple as computing the information matrix for the reparameterized model.

When a Bayesian criterion is considered as a means of taking the whole prior distribution of the parameter into account, the convex theory of experimental design which leads to the theorem on finite number of support points for the design measure can not be applied directly. The design problem is then solved over a finite grid of frequencies. The constructive, simple experimental design algorithms for the non-Bayesian criterion can be generalized to this class of problems.

Examples are included throughout the text to illustrate the concepts.

Bibliography

Anderson-Sprecher, R. (1994), 'Model comparisons using r^2', *Amer. Statist.* **48**, 113–117.

Andersson, H., Frederiksen, S. & Lundell, A. (1993), System temperatures in a sector of the Malmö network, report from a field measurement project, *in* '26th UNICHAL-Congress', Paris. 13 pp.

Arvastson, L. & Holst, J. (1995), A physically based stochastic model for the power load on a district heating network, *in* '5th International Symposium on Automation of District Heating Systems', Helsinki, Finland.

Arvastson, L., Olsson, H. & Holst, J. (1995), Parameter distribution in exponential forgetting, Technical report, Department of Mathematical Statistics, Lund Institute of Technology.

Åström, K. (1970), *Introduction to Stochastic Control Theory*, Academic Press, New York.

Bierman, G. (1977), *Factorization methods for discrete sequential estimation*, Mathematics in science and engineering, Academic Press, New York.

Bohlin, T. (1978), 'Maximum-power validation of models without higher-order fitting', *Automatica* **14**, 137–146.

Box, G. & Jenkins, G. (1976), *Time Series Analysis: Forecasting and Control*, Holden Day, San Fransisco.

Brockett, R. (1977), Convergence of Volterra series on infinite intervals and bilinear approximations, *in* V. Lakshmikathan, ed., 'Nonlinear Systems and Applications', Academic Press, New York, pp. 39–46.

Brockwell, P. & Davis, R. (1987), *Time Series: Theory and Methods*, Springer-Verlag, Berlin.

Chen, R. & Tsay, R. S. (1993), 'Functional-coefficient autoregressive models', *JASA* **88**(421), 298–308.

Cleveland, W. & Devlin, S. (1988), 'Locally weighted regression: An approach to regression analysis by local fitting', *JASA* **83**, 596–610.

Dempster, A., Laird, M. & Rubin, D. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *J.R. Stat. Soc.* **B39**, 1–38.

DeWolf, D. & Wiberg, D. (1993), 'An ordinary differential equation technique for continuous-time parameter estimation', *IEEE Transactions on Automatic Control* **AC-38**, 514–528.

Efron (1982), 'Maximum likelihood and decision theory', *Annals of Stat.* **10**, 340–356.

Efron, B. & Tibshirani, R. (1993), *An Introduction to the Bootstrap*, Chapman & Hall.

Engle, R. F. (1982), 'Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation', *Econometrica* **50**(4), 987–1007.

Engle, R. F., Lilien, D. M. & Robins, R. P. (1987), 'Estimating time varying risk premia in the term strucre: The arch-m model', *Econometrica* **55**(2), 391–407.

Ezekiel, M. & Fox, K. (1959), *Methods of Correlation and Regression Analysis*, Wiley, New York.

Fedorov, V. (1972), *Theory of optimal experiments*, Academic Press, New York.

Fortesque, T., Kershenbaum, L. & Ydstie, B. (1981), 'Implementation of self-tuning regulators with variable forgetting factors', *Automatica* **18**, 224–238.

Gaffke, N. & Mathar, R. (1992), 'On a class of algorithms from experimental design theorem', *Optimization* **24**, 91–126.

Gard, T. (1988), *Introduction to Stochastic Differential Equations*, Marcel Dekker, New York.

Giri, N. (1965), 'On the complex analouges of t^2 and r^2 tests', *Ann. Math. Statist.* **36**, 664–670.

Godfrey, K. (1980), 'Correlation methods', *Automatica* **16**, 527–534.

Goodwin, G. & Payne, R. (1977), *Dynamic system identification. Experiment design and data analysis*, Academic Press, New York.

Goodwin, G. & Sin, K. (1984), *Adaptive Filtering, Prediction and Control*, Prentice Hall, Englewood Cliffs, N.J.

Granger, C. (1983), *Applied Time Series Analysis of Economic Data*, US. Dept. Commerce, Washington, chapter Forecasting White Noise, pp. 308–314.

Hall, P. & Heyde, C. (1980), *Martingale limit theory and its applications*, Academic Press, New York.

Hammersten, S., van Hattem, D., Bloem, H. & Colombo, R. (1988), 'Passive solar component testing with identification methods', *Solar Energy* **41**(1), 5–13.

Händel, P. (1994), Indirect estimation of thermal networks, *in* 'IFAC World Congress', Vol. 3, IFAC, Sydney, pp. 415–418.

Härdle, W. (1990), *Applied nonparametric regression*, Cambridge University Press, New York.

Harvey, A. (1989), *Forecasting, structural time series models and the Kalman filter*, Cambridge Press.

Hastie, T. & Tibshirane, R. (1993), 'Varying-coefficient models', *Journal of the Royal Statistical Society, Series B* **55**, 757–796.

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, Chapman & Hall.

Hendricks, E. (1992), Identification and estimation of nonlinear systems using physical modelling, PhD thesis, Institute of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark.

Hinnich, M. (1982), 'Testing for gaussianity and linearity of a stationary time series', *Journal of Time Series Analysis* **3**(3), 169–176.

Hjorth, J. (1994), *Computer Intensive Statistical Methods*, Chapman & Hall, London.

Holst, J. (1979), Local convergence of some recursive stochastic algorithms, *in* 'IFAC symposium on identification and system parameter estimation, Darmstadt', pp. 1139–1146.

Holst, J., Holst, U., Madsen, H. & Melgaard, H. (1992), Validation of grey-box models, *in* '4th IFAC international symposium on Adaptive Signals in Control and Signal Processing, Grenoble', pp. 407–414.

Holst, U., Lindgren, G., Holst, J. & Thuvesholmen, M. (1994), 'Recursive estimation in switching autoregressions with a Markov regime', *Journal of Time Series Analysis* **15**, 489–506.

Jazwinski, A. (1970), *Stochastic Processes and Filtering Theory*, Academic Press, New York.

Kendall, M. & Stuart, A. (1961), *The Advanced Theory of Statistics*, Vol. 2, Charles Griffin & Company Limited, London.

Khatri, C. G. (1965), 'Classical statistical analysis based on a certain multivariate complex gaussian distribution', *Ann. Math. Statist.* **36**, 98–107.

Klimko, L. & Nelson, P. (1978), 'On conditional least square estimation for stochastic processes', *Annals of Statistics* **6**, 629–642.

Kloeden, P. & Platen, E. (1992), *Numerical Solution of Stochastic Differential Equations*, Springer Verlag, Berlin.

Kushner, H. (1971), *Introduction to Stochastic Control*, Holt, Reinhart and Winston, New York.

Lin, T. & Pourahmadi, M. (1998), 'Nonparametric and non-linear models and data mining in time series; a case-study on the Canadian lynx data', *Appl. Statist.* **47**, 187–201.

Lindoff, B. & Holst, J. (1995*a*), Bias and covariance on the RLS-estimator with exponential forgetting in vector autoregressions, Technical report, Department of Mathematical Statistics, Lund Institute of Technology.

Lindoff, B. & Holst, J. (1995*b*), Distribution of the RLS-estimator in a time-varying AR(1)-process, *in* '5th IFAC symposium on Adaptive Systems in Control and Signal Processing'.

Ljung, L. (1979), 'Asymptotic behaviour of the extended Kalman filter as a parameter estimator for linear systems', *IEEE Transactions on Automatic Control* **AC-24**, 36–50.

Ljung, L. (1987), *System Identification – Theory for the User*, Prentice Hall, Englewood Cliffs.

Ljung, L. (1995), *System Identification – Theory for the User, 2.nd Edition*, Prentice Hall, Englewood Cliffs.

Ljung, L. & Glad, T. (1994), 'On global identifiability for arbitrary model parameterizations', *Automatica* **30**(2), 265–276.

Ljung, L. & Söderström, T. (1983), *Theory and Practice of Recursive Identification*, MIT press, Cambridge, Massachusetts.

Madsen, H. (1985), Statistically determined dynamical models for climate processes, Ph.d. thesis, IMSOR, DTU, 2800 Lyngby.

Madsen, H. (1989), *Tidsrækkeanalyse, Arbejdsudgave*, 1 edn, IMM, DTU, DK-2800 Lyngby.

Madsen, H. (1992), 'Projection and separation in hilbert spaces', Notes in Ph.D. course in multivariate system identification.

Madsen, H. (1995*a*), *Tidsrækkeanalyse*, 2 edn, IMM, DTU, DK-2800 Lyngby.

Madsen, H. (1995*b*), *Time Series Analysis (in Danish)*, 2nd edn, Institute of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark.

Madsen, H. & Holst, J. (1995), 'Estimation of continuous-time models for the heat dynamics of a building', *Energy and Buildings* **22**, 67–79.

Madsen, H. & Melgaard, H. (1991), The mathematical and numerical methods used in ctlsm - a program for ml-estimation in stochastic, continuous time dynamical models, Technical Report 7, IMSOR, DTU, 2800 Lyngby.

Madsen, H., Nielsen, T. & Søgaard, H. T. (1996), *Control of Supply Temperatures in District Heating Systems*, Department of Mathematical Modelling, Technical University of Denmark, Lyngby.

Martin, R. & Yohai, V. (1985), *Robustness in time series and estimating arma models*, Vol. 5 of *Handbook of Statistics*, Elsevier Science Publishers, Amsterdam, chapter 119-155.

Maybeck, P. (1982), *Stochastic Models, Estimation and Control, Vol 1, 2 & 3*, Academic Press, New York.

McKeague, I. & Zhang, M.-J. (1994), 'Identification of nonlinear time series from first order cumulative characteristics', *The Annals of Statistics* **22**(1), 495–514.

Melgaard, H. (1994), Identification of Physical Models, PhD thesis, Institute of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark.

Melgaard, H. & Madsen, H. (1993), Ctlsm – continuous time linear stochastic modeling (version 2.6), Technical Report 1, IMSOR, DTU, 2800 Lyngby.

Melgaard, H., Madsen, H. & Holst, J. (1992), Methods for parameter estimation in stochastic differential equations, *in* 'Proceedings of the 1st Workshop on Stochastic Numerics', Institut für Angewandte Analysis und Stochastik, Berlin, pp. 53–64.

Melgaard, H., Sadegh, P., Madsen, H. & Holst, J. (1993), Experiment design for grey-box models, *in* 'Preprints of the 12th IFAC World Congress', Vol. 2, IFAC, pp. 145–148.

Mellentin, G. (1992), Ikke-lineær tidsrækkeanalyse, Master's thesis, IMM, DTU, DK-2800 Lyngby.

Moran, A. (1953), 'The statistical analysis of the sunspot and lynx cycles', *J. Animal Ecol.* **18**, 115–116.

Nadaraya, E. (1964), 'On estimating regression', *Theory Prob. Appl.* **10**, 186–190.

Nielsen, H. & Madsen, H. (2001), 'A generalization of some classical time series tools', *Computational Statistics and Data Analysis* .

Nielsen, H., Nielsen, T. & Madsen, H. (1997), ARX-models with parameter variations estimated by local fitting, *in* 'System Identification Conference', IFAC.

Øksendal, B. (1989), *Stochastic Differential Equations – An Introduction with Applications*, Springer Verlag, Berlin.

Olbjer, L., Holst, U. & Holst, J. (1994), *Time Series Analysis (in Swedish)*, 4th edn, KF-Sigma, Lund, Sweden.

Palsson, O., Madsen, H. & Søgaard, H. (1994), 'Generalized predictive control for non-stationary systems', *Automatica* **30**, 1991–1997.

Pazman, A. (1986), *Foundations of optimum experimental design*, Mathematics and Its Applications, D. Reidel Publ., Dordrecht.

Pham Dinh Tuan (1985), 'Bilinear markovian representation and bilinear models', *Stochastic Processes Appl.* **20**, 295–306.

Pham Dinh Tuan (1986), 'The mixing property of bilinear and generalised random coefficient autoregressive models', *Stochastic Processes Appl.* **23**, 291–300.

Plackett, R. (1950), 'Some theorems in least squares', *Biometrika* **37**, 149.

Priestley, M. (1980), 'State-dependent models: a general approach to non-linear time series analysis', *J. Time Series Anal.* **1**, 47–71.

Priestley, M. (1988), *Non-linear and Non-stationary Time Series Analysis*, Academic Press, London.

Rao, C. (1973), *Linear statistical inference and its applications*, Wiley, New York.

Ripley, B. (1993), Statistical aspects of neural networks, *in* O. Barndorff-Nielsen, J. Jensen & W. Kendall, eds, 'Networks and Chaos - Statistical and Probabilistic Aspects', Chapman & Hall, pp. 48–123.

Ripley, B. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press.

Robinson, P. (1983), 'Nonparametric estimators for time series', *Journal of time series analysis* **4**(3), 185–207.

Rockafellar, R. (1970), *Convex Analysis*, Princeton University Press.

Rydén, T. (1993), Parameter estimation for Markov modulated Poisson processes and Overload control of SPC switches, PhD thesis, Dept of Mathematical Statistics, Lund Institute of Technology, Lund, Sweden.

Sadegh, P., Madsen, H. & Holst, J. (1996), 'Optimal input design for fault detection and diagnosis', *Send to Signal Processing* .

Sadegh, P., Melgaard, H., Madsen, H. & Holst, J. (1994), On the usefulness of grey-box information when doing experiment design for system modelling and identification, *in* 'IFAC SYSID '94', Vol. 3, pp. 219–224.

Saikkonen, P. & Luukkonen, R. (1988), 'Lagrange multiplier tests for testing non-linearities in time series models', *Scandinavian Journal of Statistics* **15**, 55–68.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics* **6**, 461–464.

Sejling, K. (1993), Modelling and Prediction on Load in District Heating Systems, PhD thesis, Institute of Mathematical Statistics and Operations Research, Technical University of Denmark, Lyngby, Denmark.

Shumway, R. (1988), *Applied Statistical Time Series Analysis*, 1 edn, Prentice Hall, London.

Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York.

Silvey, S. D. (1980), *Optimal Design*, Chapman & Hall, London.

Søgaard, H. & Madsen, H. (1991), On-line estimation of time-varying delays in district heating systems, *in* 'Proc. of the 1991. European Simulation Conference', pp. 619–624.

Söderström, T. & Stoica, P. (1988), *System Identification*, Prentice Hall International, New York.

Söderström, T., Stoica, P. & Friedlander, B. (1991), 'An indirect prediction error method for system identification', *Automatica* **27**(1), 183–188.

Statistical Sciences (1995), *S-PLUS guide to Statistical and Mathematical Analysis, Version 3.3*, StatSci - MathSoft, Seattle.

Stoica, P., Holst, J. & Söderström, T. (1982), 'Eigenvalue location of certain matrices arising in convergence analysis problems', *Automatica* **18**, 487–491.

Stoica, P. & Nehorai, A. (1989), 'On multistep prediction error methods for time series models', *Journal of Forecasting* **8**, 357–368.

Subba Rao, T. & Gabr, M. (1984*a*), *An Introduction to Bispectral Analysis and Bilinear Time Series models*, Lecture Notes in Statistics, Vol 24, Springer, New York.

Subba Rao, T. & Gabr, M. (1984*b*), *An introduction to bispectral analysis and bilinear time series models*, Springer Verlag, Berlin.

Thuvesholmen, M. (1994), Recursive estimation and segmentation in autoregressive processes with Markov regime, Technical report, 1994:E4, Dept of Mathematical Statistics, Lund University, Lund, Sweden.

Tjøstheim, D. (1986), 'Some doubly stochastic time series models', *J. Time Series Anal* **7**, 51–72.

Tjöstheim, D. (1994), 'Non-linear time series: A selective review', *Scandinavian Journal of Statistics* **21**, 97–130.

Tong, H. (1983), *Threshold Models in Non-Linear Time Series Analysis*, Springer, Heidelberg.

Tong, H. (1990*a*), *Non-linear Time Series*, Oxford Univ. Press, Oxford.

Tong, H. (1990*b*), *Non-linear Time Series – A Dynamic System Approach*, Oxford Science Publishers, Oxford.

Tukey, J. (1961), Curves as parameters and touch estimation, *in* 'Proc. Fourth Berkeley Symp. Math. Statist. Probab.', Vol. 1, Univ. California Press, Berkeley, pp. 681–684.

Tulleken, H. (1993), 'Grey-box modelling and identification using physical knowledge and bayesian techniques', *Automatica* **29**(2), 285–308.

Vadstrup, P. (1988), Adaptive control of nonlinear systems, PhD thesis, Servolaboratory, Technical University of Denmark, Lyngby, Denmark.

Volterra, V. (1930), *Theory of Functionals*, Blackie, London.

Wahlberg, B. (1987), On the identification and appproximation of linear systems, Doctoral dissertation, Linköping University, Linköping, Sweden.

Wahlberg, B. (1989), 'Estimation of arma models via high order autoregressive approximations', *J. Time Series Anal.* **10**, 283–299.

Wahlberg, B. & Ljung, L. (1986), 'Design variables for bias distribution in transfer function estimation', *IEEE Transactions on Automatic Control* **31**(2), 134–144.

Walter, E. & Pronzato, L. (1990), 'Qualitative and quantitative experiment design for phenomenological models – a survey', *Automatica* **26**(2), 195–213.

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, Wiley, New York.

Whittle, P. (1971), *Optimization Under Constraints*, Wiley-Interscience, New York.

Wiberg, D. & DeWolf, D. (1993), 'A convergent approximation to the continuous-time optimal estimator', *IEEE Transactions on Automatic Control* **AC-38**, 529–545.

Wiberg, D., Ljungkvist, D. & Powell, T. (1994), Towards a useful approximation of the optimal recursive parameter estimator, *in* '10th IFAC symposium on system identification', Vol. 2, pp. 685–690.

Wiener, N. (1958), *Non-linear Problems in Random Theory*, MIT Press, Cambridge, MA.

Wouters, P. (1993), Passys - final issue, *in* T. C. Steemers, ed., 'Commission of the European Communities – Directorate General XII for Science, Research and Development', DG XII, EU.

Young, P. (1984), *Recursive estimation and time-series analysis*, Springer, Berlin.

Yuan, Z.-D. & Ljung, L. (1985), 'Unprejudiced optimal open loop input design for identification of transfer functions', *Automatica* **21**(6), 697–708.

Zarrop, M. B. (1979), *Optimal Experimenexperiment design for dynamic system identification*, Springer-Verlag, Berlin.

Index