Submitted to International Journal of Control

Constructing NARMAX models using ARMAX models

May 5, 1992, Revised April 20, 1993

Tor A. Johansen and Bjarne A. Foss
Department of Engineering Cybernetics
Norwegian Institute of Technology
N-7034 Trondheim

Abstract

This paper outlines how it is possible to decompose a complex non-linear modelling problem into a set of simpler linear modelling problems. Local ARMAX models valid within certain operating regimes are interpolated to construct a global NARMAX (non-linear NARMAX) model. Knowledge of the system behavior in terms of operating regimes is the primary basis for building such models, hence it should not be considered as a pure black-box approach, but as an approach that utilizes a limited amount of a priori system knowledge. It is shown that a large class of non-linear systems can be modelled in this way, and indicated how to decompose the systems range of operation into operating regimes. Standard system identification algorithms can be used to identify the NARMAX model, and several aspects of the system identification problem is discussed and illustrated by a simulation example.

1 Introduction

Modelling complex systems using first principles is in many cases resource demanding. In some cases our system knowledge is so limited that detailed modelling is difficult. In other cases, the instrumentation and logged data from the system are so sparse or noisy that it is difficult to identify a large number of unknown physical parameters in the model. Examples of this are found in e.g. metallurgical and biochemical process industry.

In some cases, resources can be saved by using black-box models describing the input/output behavior of the system. Such models represents the controllable and observable part of the system. The structure and parameters of black-box models has in general no direct interpretation in terms of the physical properties of the system. The ARMAX model is a well known linear input/output model representation. The NARMAX (Nonlinear ARMAX) model representation is an extension of the linear ARMAX model, and represents the system by a nonlinear mapping of past inputs, outputs and noise terms to future outputs. In this paper we discuss how NARMAX models can be represented, and in particular discuss how a NARMAX model can be constructed from a set of ARMAX models.

We will concentrate on non-linear systems that are working in several operating regimes, because systems that normally work within one operating regime may in many cases be adequately described by a linear model. There are numerous examples of systems that must work in several operating regimes, including most batch processes (Rippin 1989). Apart from normal operating conditions, the control system may also have to take care of startup and shutdown, operation during maintenance and faulty operation, which obviously lead to different operating regimes.

Traditionally, the problem with multiple operating regimes is solved by non-linear first principles models covering several operating regimes, gain scheduling, or simply by manual or rule-based control of the system when operating outside the normal operating regimes. From an engineering point of view, it may seem appealing to decompose the modelling problem into a set of simpler modelling problems. This is exactly what we propose here: First the system operation is decomposed into a set of operating regimes that are assumed to cover the full range of operation we want our model to cover. Next, for each operating regime we design a simple (typically linear) local model. It is usually not natural to define the operating regimes as crisp sets, since there will usually be a smooth transition from one regime to another, not a jump. Hence, it makes sense to interpolate the local models in a smooth fashion to get a global model. The interpolation is such that the local model that is assumed to be the best model at the current operating point will be given most weight in the interpolation, while neighboring local models may be given some weight, and local models corresponding to distant operating regimes will not contribute to the global model at that operating point. To do the smooth interpolation at a given operating point, we need to know which of the local models describe the system well around that operating point. For that purpose, to each local model we associate a local model validity function, i.e. a function that indicates the relative validity of the local models at a given operating point.

The use of local linear models without interpolation, i.e. piecewise linear models, have been suggested by several authors, including Skeppstedt, Ljung & Millnert (1992), Skeppstedt (1988), Hilhorst, van Amerongen & Löhnberg (1991), Billings & Voon (1987), and Tong & Lim (1980). A related technique is the use of splines (Friedman 1991) for representing dynamics models (Psichogios, De Veaux & Ungar 1992). Splines are also local models, but unlike piecewise linear models, there are constraints that enforces smoothness on the boundaries between the local models. Different variations of interpolating memories (Tolle, Parks, Ersü, Hormel & Militzer 1992, Lane, Handelman & Gelfand 1992, Omohundro 1987) is a related local modelling technique, where a number of input/output-pairs of the system is memorized and interpolated to give a model. Our approach can be thought of as a generalization of these techniques, since we interpolate local models.

This paper is organized as follows: First, in section 2, we present a model representation based on local models. Then we discuss the approximation capabilities of this representation, and show that a large class of non-linear systems can be represented. The notion of operating regimes is introduced and we present a general result guiding the choice of operating point vector. Thereafter, we discuss some practical aspects of modelling using local models in section 3, and some aspects of system identification in section 4. In section 5, the concepts are illustrated by a simulation example, and section 6 contains some discussions and conclusions.

2 Model Representation

The NARMAX model representation

$$y(t) = f(y(t-1), ..., y(t-n_y), u(t-1), ..., u(t-n_u), e(t-1), ..., e(t-n_e)) + e(t) \quad (1)$$

is shown by Leontaritis & Billings (1985) and Chen & Billings (1989) to represent a large class of discrete-time nonlinear dynamic systems. Here $y(t) \in Y \subset R^m$ is the output vector, $u(t) \in U \subset R^r$ is the input vector, and $e(t) \in E \subset R^m$ is equation error. We introduce the $(m(n_y + n_e) + rn_u)$-dimensional information vector

$$\psi(t-1) \quad = \quad [y^T(t-1), ..., y^T(t-n_y), u^T(t-1), ..., u^T(t-n_u), e^T(t-1), ..., e^T(t-n_e)]^T$$

where $\psi(t-1)$ is in the set $\Psi = Y^{n_y} \times U^{n_u} \times E^{n_e}$. This enables us to write equation (1) in the form

$$y(t) = f(\psi(t-1)) + e(t) \quad (2)$$

Provided that necessary smoothness conditions on $f : \Psi \to Y$ are satisfied, a general way of representing functions is by series expansions. Using a 1st order Taylor-series expanded about the systems equilibrium point yields a standard ARMAX model. Second-order Taylor-expansions are possible, while higher-order Taylor-expansions are not very useful in practice because the number of parameters in the model increases drastically with the expansion order, and because of the poor extrapolation and interpolation capabilities of higher-order polynomials. Splines offers a solution to this problem, but the approximation in higher dimensional spaces may be difficult due to the smoothness constraints on the boundaries. Chen, Billings & Grant (1990*a*) have proposed to use a sigmoidal neural network expansion, Billings & Voon (1987) uses a piecewise linear model, and Chen, Billings, Cowan & Grant (1990*b*) have applied a radial basis function expansion as a means of representing f. A generic model representation based on *local* models was introduced in (Johansen & Foss 1992*b*, Johansen & Foss 1992*a*), inspired by work by Stokbro, Hertz & Umberger (1990) and Jones et al. (1991). We will in the following study this model representation in detail.

Approximation using local models and interpolation

Given a set of functions $\{\tilde{\rho}_i : \Psi \to [0,1]\}_{i=0}^{N-1}$, the following equation is trivially true

$$f(\psi) = \frac{\sum_{i=0}^{N-1} f(\psi)\tilde{\rho}_i(\psi)}{\sum_{i=0}^{N-1} \tilde{\rho}_i(\psi)} \quad (3)$$

assuming that at any point $\psi \in \Psi$, not all $\tilde{\rho}_i$ vanish. We will assume the function $\tilde{\rho}_i$ is chosen such that it is *localized* in a subset $\Psi_i \subset \Psi$. This means that $\tilde{\rho}_i(\psi)$ is significantly larger than zero for $\psi \in \Psi_i$, and $\tilde{\rho}_i(\psi)$ is very close to zero for $\psi \notin \Psi_i$. Since $\tilde{\rho}_i$ is close to zero for $\psi \notin \Psi_i$, we can substitute f on the right-hand side in (3) with a $\hat{f}_i$ that is a good approximation only in Ψ_i

$$\hat{f}(\psi) = \frac{\sum_{i=0}^{N-1} \hat{f}_i(\psi)\tilde{\rho}_i(\psi)}{\sum_{i=0}^{N-1} \tilde{\rho}_i(\psi)} \quad (4)$$

Introducing the normalized functions $\tilde{w}_i : \Psi \to [0,1]$ defined by

$$\tilde{w}_i(\psi) = \frac{\tilde{\rho}_i(\psi)}{\sum_{j=0}^{N-1} \tilde{\rho}_j(\psi)} \quad (5)$$

gives the approximation

$$\hat{f}(\psi) = \sum_{i=0}^{N-1} \hat{f}_i(\psi)\tilde{w}_i(\psi) \tag{6}$$

In this equation, we can interpret $\tilde{w}_i$ as a function that gives a value close to 1 in parts of Ψ where the function $\hat{f}_i$ is a good approximation to f, and close to zero elsewhere. By definition of $\tilde{w}_i$ we know that $\sum_{i=0}^{N-1} \tilde{w}_i(\psi) = 1$ for all $\psi \in \Psi$, and we call the functions $\tilde{w}_i$ *interpolation functions* because they are used to interpolate *local models* $\hat{f}_i$. We call $\hat{f}_i$ a local model since it is assumed to be an accurate description of the true f locally (where $\tilde{\rho}_i$ is not close to zero).

The set of all functions of the form (6) with local models of polynomial order p and smooth interpolation functions is denoted

$$\tilde{\mathcal{F}}_p \quad = \quad \left\{ \hat{f} : \Psi \to Y \mid \hat{f}(\psi) = \sum_{i=0}^{N-1} \hat{f}_i(\psi)\tilde{w}_i(\psi) \right\}$$

At the extreme, 0th order Taylor-expansions of f about $\psi_i \in \Psi_i$ may be used to define $\hat{f}_i$:

$$\hat{f}_i(\psi) = f(\psi_i) = \theta_i \tag{7}$$

where θ_i is a parameter vector. Such a simple local model is closely related to an interpolating memory (Tolle et al. 1992) and requires a large number of interpolation functions, since this means that the value $f(\psi_i)$ is extrapolated locally. This case is in fact identical to neural networks with localized receptive fields, (Moody & Darken 1989, Stokbro et al. 1990). Considering $\{\tilde{w}_i\}_{i=0}^{N-1}$ as a set of basis-functions, the method is also similar to radial basis-function expansions (Broomhead & Lowe 1988), for the following reason: If the functions $\tilde{\rho}_i$ are chosen as local radial functions, the normalized function $\tilde{w}_i$ defined by (5) will not be radial in general, but it will qualitatively have much the same shape and features as $\tilde{\rho}_i$, except near the boundary of Ψ.

A 1st order Taylor-expansion of f about ψ_i provides better extrapolation and interpolation than the 0th order expansion (7). Assuming the 1st derivative of f exists, the local models are given by

$$\hat{f}_i(\psi) = f(\psi_i) + \nabla f(\psi_i)(\psi - \psi_i) = \theta_i + \Theta_i(\psi - \psi_i) \tag{8}$$

where θ_i is a parameter vector and Θ_i is a parameter matrix. Observe that (8) is actually an ARMAX model resulting from a linearization about ψ_i. Both the Weighted Linear Maps of Stokbro et al. (1990), Stokbro (1991) and Stokbro & Umberger (1990) and the Connectionist Normalized Linear Spline Networks of Jones et al. (1991) and Jones, Lee, Barnes, Flake, Lee, Lewis & Qian (1989) uses a 1st order expansion locally. This representation makes it possible to build a NARMAX model by interpolating between a set of ARMAX models.

Higher order local models can of course also be used. Furthermore, there is no requirement that all the local models should have the same structure. Some of the local models may be based on first principles modelling, while others may be generic black-box models. Johansen & Foss (1992*c*) uses this approach to integrate first principles models with neural network type models.

Approximation Properties

It seems reasonable that the approximation can be made arbitrary good by choosing a sufficient number of local models. This is indeed the case, as illustrated in the following.

We use the following norm to measure the approximation accuracy

$$||f - \hat{f}||_\infty \quad = \quad \sup_{\psi \in \Psi} ||f(\psi) - \hat{f}(\psi)||_2$$

where $||\cdot||_2$ denotes Euclidian norm.

The $(p+1)$th derivative of the vector function f at the point ψ is denoted by $\nabla^{p+1} f(\psi)$. Assume f is continuously differentiable $p+1$ times, and $\{\hat{f}_i\}_{i=0}^{N-1}$ are local models equal to the first p terms of the Taylor-series expansion of f about ψ_i. For any $\psi \in \Psi$, we have

$$f(\psi) - \hat{f}(\psi) \quad = \quad \sum_{i=0}^{N-1} (f(\psi) - \hat{f}_i(\psi)) \tilde{w}_i(\psi)$$

If we assume $||\nabla^{p+1} f(\psi)|| < M$ for all $\psi \in \Psi$, where $||\cdot||$ denotes the induced operator norm, we obtain by Taylors theorem

$$||f(\psi) - \hat{f}(\psi)||_2 \quad < \quad \sum_{i=0}^{N-1} \frac{M}{(p+1)!} ||\psi - \psi_i||_2^{p+1} \; \tilde{w}_i(\psi)$$

In order to ensure that this norm is smaller than an arbitrary $\epsilon > 0$, we must ensure that for any $\psi \in \Psi$ the following condition holds

$$\sum_{i=0}^{N-1} ||\psi - \psi_i||_2^{p+1} \tilde{\rho}_i(\psi) < \epsilon \frac{(p+1)!}{M} \sum_{i=0}^{N-1} \tilde{\rho}_i(\psi) \tag{9}$$

Defining the set of functions $\{g_i : \Psi \to R\}_{i=0}^{N-1}$ by

$$g_i(\psi) \quad = \quad ||\psi - \psi_i||_2^{p+1} - \epsilon \frac{(p+1)!}{M}$$

and rewriting (9) gives the following condition that must hold for any $\psi \in \Psi$

$$\sum_{i=0}^{N-1} g_i(\psi) \tilde{\rho}_i(\psi) < 0 \tag{10}$$

or equivalently if we divide (10) by $\sum_{i=0}^{N-1} \tilde{\rho}_i(\psi)$

$$\sum_{i=0}^{N-1} g_i(\psi) \tilde{w}_i(\psi) < 0 \tag{11}$$

The problem is now to find the conditions on N and the functions $\{\tilde{\rho}_i\}_{i=0}^{N-1}$ to ensure that equation (10) holds for any given $\epsilon > 0$. A geometric interpretation of (10) is given in Figure 1. Certainly, this equation holds if the negative contribution of one term $g_i(\psi)\tilde{\rho}_i(\psi)$ in (10) dominates the (possibly positive) contribution of all other terms. A necessary condition is $g_i(\psi)\tilde{\rho}_i(\psi) \to 0$ as $||\psi||_2 \to \infty$. This will certainly be ensured if we choose $\tilde{\rho}_i$ as an exponential or Gaussian function.

Notice that the *shape* of the g_i-functions are fixed and given by the specifications. We are, however, free to choose the location and number N of local models. Let us choose the set $\{\psi_i\}_{i=0}^{N-1}$ so large and "sufficiently dense in Ψ" that at least one of the functions $\{g_i\}_{i=0}^{N-1}$ will be negative at any $\psi \in \Psi$. Then the functions $\{g_i\}_{i=0}^{N-1}$ are fixed, and we must choose the $\tilde{\rho}_i$-functions such that (10) hold.

This can be done in several ways. In the limit when the width of the $\tilde{\rho}_i$-functions go to zero the interpolation functions $\tilde{w}_i$ will approach step-functions as shown in Figure 2. The model will then approach a piecewise constant model if $p = 0$, a piecewise linear model if $p = 1$, etc. In this limit, at any $\psi \in \Psi$ there will exist a j such that

$$\tilde{w}_i(\psi) = \begin{cases} 1 & \text{If } i = j \\ 0 & \text{If } i \neq j \end{cases}$$

By the choice of $\{\psi_i\}_{i=0}^{N-1}$ we know that $g_j(\psi) < 0$, and since $\tilde{w}_i(\psi) = 0$ for $i \neq j$, (11) will hold. We can now provide a result for the case when Ψ is a bounded set:

Theorem 1 *Suppose given any integer $p \geq 0$, and suppose f has continuous $(p+1)$-th derivative. If Ψ is bounded, then for any $\epsilon > 0$ there is a $\hat{f} \in \tilde{\mathcal{F}}_p$ (with finite N, which may depend on ϵ) such that*

$$||f - \hat{f}||_\infty < \epsilon \tag{12}$$

Proof: Since $\nabla^{p+1} f(\psi)$ is continuous, it is bounded (by $M < \infty$) on Ψ. Since Ψ is bounded, a finite N is sufficient to ensure that one g_i-function is negative at any point. Since N is finite, we do not have to go to the limit and make $\{\tilde{w}_i\}_{i=0}^{N-1}$ step-functions, but can stop when we are sufficiently close. Then $\{\tilde{\rho}_i\}_{i=0}^{N-1}$ can be chosen as smooth functions such that (11) holds. Since $\psi \in \Psi$ was arbitrary, the theorem is proved.

□

This is an existence theorem. However, the proof is constructive and gives indications on how to construct the approximator. In order to use this proof to formulate an upper bound on the approximation error, we introduce the following definition of distance between sets, similar to the Haussdorf metric:

Definition 1 *Assume A and B are two nonempty subsets of a vector space. Then the distance between the sets is defined as*

$$\mathcal{D}(A, B) = \inf_{a \in A} \sup_{b \in B} ||a - b||_2$$

□

The crux in the proof of Theorem 1 is that at any point $\psi \in \Psi$ one of the g_i-functions is negative and that the $\tilde{\rho}_i$-functions are chosen such that at any point $\psi \in \Psi$, a negative term $g_i(\psi)\tilde{\rho}_i(\psi)$ will dominate the sum (10). At least one g_i-function will be negative at any $\psi \in \Psi$ if the following condition holds

$$\mathcal{D}(\{\psi_i\}, \Psi) \leq \left(\epsilon \frac{(p+1)!}{M} \right)^{\frac{1}{p+1}} \tag{13}$$

If the set $\{\psi_i\}$ is dense in Ψ, this distance will be zero. The term "sufficiently dense" used informally above means that the set $\{\psi_i\}_{i=0}^{N-1}$ should be chosen such that (13) holds for the given ϵ.

Theorem 2 *Suppose given an integer $p \geq 0$. If Ψ is bounded and f has bounded $(p+1)$-th derivative, i.e. $||\nabla^{p+1} f(\psi)|| \leq M$ for all $\psi \in \Psi$, then for any $\hat{f} \in \tilde{\mathcal{F}}_p$ with finite N and sufficiently narrow functions $\{\tilde{\phi}_i\}_{i=0}^{N-1}$, an upper bound on the approximation error is given by*

$$||f - \hat{f}||_\infty \leq \frac{M}{(p+1)!} \left(\mathcal{D}(\{\psi_i\}, \Psi) \right)^{p+1} \tag{14}$$

Proof: (13) will hold for ϵ equal to the right-hand side of 14). From the previous discussion, it is evident that (11) holds for any $\psi \in \Psi$. Hence, $||f - \hat{f}||_\infty$ will be bounded by ϵ, and the result follows.

□

This is under the condition that the $\tilde{\rho}_i$-functions are chosen narrow. This bound is conservative, meaning that for $\tilde{\rho}_i$-functions that are not too narrow and not too wide, one may expect better accuracy. However, if the functions $\{\tilde{\rho}_i\}_{i=0}^{N-1}$ are not narrow, the result does not hold.

From (14) we see that if the polynomial order p of the local models is increased, then the accuracy will improve. If Ψ is not bounded and $M > 0$, N must be infinite in order to guarantee a bounded error.

If f does not satisfy the smoothness conditions in Theorem 1, the proof obviously does not hold. If, however, f is such that it can be approximated arbitrary well by a sufficiently smooth function, then we can show that f can be approximated arbitrary well by interpolating local models. In patricular we have:

Corollary 1 *The results of Theorem 1 also holds if the smoothness assumption on f is relaxed to assuming only continuity. In other words, the set $\tilde{\mathcal{F}}_p$ is dense in the set of continuous functions from Ψ into Y.*

Proof: By the Weierstrass approximation theorem, e.g. (Stromberg 1981), for any $\epsilon > 0$ there exists a polynomial $\tilde{f}$ such that $||f - \tilde{f}||_\infty \leq \epsilon/2$. By theorem 1, $\tilde{f}$ can be approximated by a $\hat{f} \in \tilde{\mathcal{F}}_p$ on the bounded set Ψ such that $||\tilde{f} - \hat{f}||_\infty < \epsilon/2$. Using the triangle inequality we get $||f - \hat{f}||_\infty < \epsilon$.

□

Example 1

Assume $p = 1$, i.e. the local models are ARMAX models. Then (14) can be written

$$||f - \hat{f}||_\infty \leq \frac{M}{2}(\mathcal{D}(\{\psi_i\}, \Psi))^2 = \epsilon \tag{15}$$

If f and ψ are scalars, M is a bound on the second derivative of f, in other words a bound on the curvature. If the system is linear, then $M = 0$ and one local linear model is sufficient to make an arbitrary good global model (of course). M indicates the nonlinearity of the function, and we expect ϵ to increase with increasing M, i.e. increasing nonlinearity, which is indeed the case as indicated by (15). However, using the upper bound M gives a conservative result since the system may behave more linearly in some regions than others. Hence, we need not have high density of local models where the system does not have strong nonlinear behavior.

□

Example 2

With a simple example we illustrate the use of Theorem 2. Consider the function $f : [0, 2] \rightarrow R$ given by $f(\psi) = \psi^2 + 1$. Assume that we have two local linear models located

at $\psi_0 = 0.5$ and $\psi_1 = 1.5$. Then $\mathcal{D}(\{0.5, 1.5\}, [0, 2]) = 0.5$, $p = 1$ and $M = 2$. Theorem 2 predicts the bound $\epsilon = 0.25$ on the approximation accuracy. As shown by Figure 3, this bound is exact when using infinitly narrow functions $\tilde{\rho}_i$, i.e. a piecewise linear approximation. The reason for this is that $M = f''(\psi) = 2$ for all ψ, hence there is no regions where f is "less nonlinear". As we shall see later, better approximations can be achieved using well-chosen $\tilde{\rho}_i$-functions. From this figure we also see that the local linear models are not chosen as a first order Taylor-expansion, but chosen on the basis of e.g. a least squares regression, improvement might also be achieved.

□

Since the system function f can be approximated arbitrary well, we are able to make arbitrary good prediction on a finite horizon if there is no noise, provided the intial values are correct and the inputs and outputs are such that they give vectors ψ that remain in Ψ (Polycarpou, Ioannou & Ahmed-Zaid 1992). However, it is well known that the solution to some difference equations are sensitive with respect to intial values or modelling errors. Examples of such systems are chaotic or unstable systems.

Operating regimes

In the rest of this paper we will usually assume $p = 1$, i.e. we use linear ARMAX models locally to build a non-linear NARMAX model. In the representation (6) the interpolation functions $\{\tilde{w}_i\}_{i=0}^{N-1}$ are defined on the set Ψ. This is a subset of the information-space. If the information-space has high dimension (as it ofte has), the *curse of dimensionallity* problem arises. This problem was first described by Bellman (1961) is essentially that the number of local models needed to uniformly cover a region of this space increases exponentially with the dimension of the space. In practise, uniform coverage it usually not necessary, but the problem is still severe. In some cases the interpolation functions may be defined on a space of smaller dimension. This is our motivation for introducing the terms operating regime and operating point. First, we define Φ to be the *set of operating points*. Motivated by the fact that we want to model a nonlinear system with a set of linear models, it is convenient to define an operating regime as a subset of Φ where the system behaves approximately linearly.

Definition 2 *An operating regime is a set of operating points $\Phi_i \subset \Phi$ where the system behaves approximately linearly.*

A *model validity function* $\rho_i : \Phi \to [0, 1]$ is smooth and satisfies $\rho_i(\phi) \approx 1$ for $\phi \in \Phi_i$, and goes to zero outside Φ_i. The interpolation functions $w_i : \Phi \to [0, 1]$ are now defined as

$$w_i(\phi) \quad = \quad \frac{\rho_i(\phi)}{\sum_{j=0}^{N-1} \rho_j(\phi)}$$

assuming that at every operating point $\phi \in \Phi$, not all model validity functions ρ_i vanish.

In many cases there will exist a function $H : \Psi \to \Phi$ such that at any t will $\phi(t) = H(\psi(t))$. The function H will typically be a projection, i.e. Φ will be in a space of lower dimension than Ψ. In cases where the operating point is calculated on the basis of filtered or estimated quantities, the relationship between $\psi(t)$ and $\phi(t)$ is more complex, and must be described by an operator $\mathcal{H}$. This may be the case when ϕ is estimated using a recursive algorithm or a recursive filter to depress noise. Although very important, this complicates the analysis

considerably, and we will not consider this case here, but leave it as a topic for future research.

To summarize, the representation we address at this stage is

$$\hat{y}(t) = \hat{f}(\psi(t-1)) = \sum_{i=0}^{N-1} \hat{f}_i(\psi(t-1))w_i(\phi(t-1)) \tag{16}$$

where the local models

$$\hat{f}_i(\psi(t-1)) = \theta_i + \Theta_i(\psi(t-1) - \psi_i) \tag{17}$$

are ARMAX models. We define the set

$$\mathcal{F}_p \quad = \quad \left\{ \hat{f} : \Psi \to Y \mid \hat{f}(\psi) = \sum_{i=0}^{N-1} \hat{f}_i(\psi)w_i(\phi) \right\}$$

where p is the polynomial order of $\hat{f}_i$, the interpolation functions $\{w_i\}_{i=0}^{N-1}$ are smooth, and $\phi = H(\psi)$. Now we want to state some general results regarding the transform H from the information vector to the operating point vector. In general, f can be written as an affine function of some of its argument. We rearrange the elements of ψ into $\psi^T = [\psi_L^T \ \psi_N^T]$ such that

$$f(\psi) = f(\psi_L, \psi_N) = f_1(\psi_N) + f_2(\psi_N)\psi_L \tag{18}$$

Assume Ψ_L and Ψ_N are the subsets of the information-space corresponding to ψ_L and ψ_N respectively. $f_1 : \Psi_N \to R^m$ and $f_2 : \Psi_N \to R^{m \times m}$ are non-linear vector- and matrix-valued functions, respectively. Our principal result guiding the choice of ϕ is the following, which indicates that ϕ must be chosen such that it captures the systems non-linearities:

Theorem 3 *Assume f given in (18) is continuous, and Ψ is bounded. Then for any $\epsilon > 0$ there is a $\hat{f} \in \mathcal{F}_1$ with $\phi = \psi_N$, and finite N such that $||f - \hat{f}||_\infty < \epsilon$.*

Proof: Fix an arbitrary $\psi \in \Psi$ such that $\psi^T = [\psi_L^T \ \psi_N^T]$.

$$\begin{aligned}
||f(\psi) - \hat{f}(\psi)||_2 &= ||f(\psi_N, \psi_L) - \sum_{i=0}^{N-1} \hat{f}_i(\psi_N, \psi_L)w_i(\psi_N)||_2 \\
&= || \sum_{i=0}^{N-1} (f_1(\psi_N) + f_2(\psi_N)\psi_L - \hat{f}_i(\psi_N, \psi_L))w_i(\psi_N)||_2 \\
&= || \sum_{i=0}^{N-1} (f_1(\psi_N) - \hat{f}_{Ni}(\psi_N) + f_2(\psi_N)\psi_L - \hat{f}_{Li}(\psi_L))w_i(\psi_N)||_2
\end{aligned}$$

In the last line we split the linear function $\hat{f}_i : \Psi \to R^m$ into two linear functions $\hat{f}_{Ni} : \Psi_N \to R^m$ and $\hat{f}_{Li} : \Psi_L \to R^m$. Now we choose $\hat{f}_{Li}(\psi_L) = \beta_i\psi_L$ where β_i is a not yet specified constant parameter matrix. Then we have

$$\begin{aligned}
||f(\psi) - \hat{f}(\psi)||_2 &\leq || \sum_{i=0}^{N-1} (f_1(\psi_N) - \hat{f}_{Ni}(\psi_N) + (f(\psi_N) - \beta_i)\psi_L)w_i(\psi_N)||_2 \\
&\leq || \sum_{i=0}^{N-1} (f_1(\psi_N) - \hat{f}_{Ni}(\psi_N))w_i(\psi_N)||_2 + || \sum_{i=0}^{N-1} (f_2(\psi_N) - \beta_i)\psi_L w_i(\psi_N)||_2 \\
&\leq ||f_1(\psi_N) - \sum_{i=0}^{N-1} \hat{f}_{Ni}(\psi_N)w_i(\psi_N)||_2 + ||\psi_L||_2 \cdot ||f_2(\psi_N) - \sum_{i=0}^{N-1} \beta_i w_i(\psi_N)||_2
\end{aligned}$$

The first term in this equation can be made arbitrary small by Corollary 1 with $p = 1$ since $\hat{f}_{Ni}$ is linear. Since Ψ is bounded, the second term can be made arbitrary small by the same corollary with $p = 0$ through the choice of β_i. Hence, for any $\epsilon > 0$ we can make $||f(\psi) - \hat{f}(\psi)||_2 < \epsilon$ and since ψ is arbitrary we get $||f - \hat{f}||_\infty < \epsilon$.

□

Using the same notation as before, the attainable approximation error is bounded by

$$||f - \hat{f}||_\infty \leq \frac{M}{2}\left((\mathcal{D}(\{\phi_i\}, \Phi))^2 + 2||\Psi_L||\mathcal{D}(\{\phi_i\}, \Phi)\right) = \epsilon \tag{19}$$

where

$$||\Psi_L|| \quad = \quad \sup_{\psi_L \in \Psi_L} ||\psi_L||_2$$

The motivation for introducing the operating point ϕ is that in many cases this vector may be of a significantly lower dimension than ψ. With a fixed N the first term in (19) will be significantly smaller than the corresponding term

$$\tfrac{M}{2}(\mathcal{D}(\{\psi_i\}, \Psi))^2$$

However, the second term in (19) will make the error increase, but in most cases when Φ is of smaller dimension than Ψ, the approximation (16)-(17) will give better accuracy than (6), (8). Another important fact is that a low dimension of Φ makes it easier to partition the set into operating regimes.

Example 3

For example, if f is linear in the control variables $u(t-1)$, then ϕ need not contain any $u(t-1)$-terms. If we have the system

$$y(t) \quad = \quad f(y(t-1), u(t-1)) = f_1(y(t-1)) + f_2(y(t-1))u(t-1)$$

we can choose $\phi(t-1) = y(t-1)$ without loosing accuracy in the approximation.

□

We now generalize this result for local expansions of (polynomial) order p. We split ψ into two parts and rearrange $\psi^T = [\psi_L^T \ \psi_H^T]$ such that

$$f(\psi) = f(\psi_L, \psi_H) = f_{H1}(\psi_H) + f_{H2}(\psi_H) f_L(\psi_L) \tag{20}$$

where $f_L : \Psi_L \to R^m$ is of polynomial order less than or equal to p, $f_{H1} : \Psi_H \to R^m$, and $f_{H2} : \Psi_H \to R^{m \times m}$ may be of higher order.

Theorem 4 *Suppose f given in (20) is continuous and Ψ is a bounded set. Then for any $\epsilon > 0$ there is a $\hat{f} \in \mathcal{F}_p$ with $\phi = \psi_H$, and finite N such that $||f - \hat{f}||_\infty < \epsilon$.*

Proof: The proof follows the same idea as the proof of Theorem 3, but requires some tedious notation, and is therefore ommitted.

□

Some Comparisions

Using local linear models we can write the model representation (16) - (17) as

$$\begin{aligned}\hat{y}(t) &= \sum_{i=0}^{N-1}(\theta_i + \Theta_i(\psi(t-1)-\psi_i))w_i(\phi(t-1)) \\ &= \left(\sum_{i=0}^{N-1}(\theta_i - \Theta_i\psi_i)w_i(\phi(t-1))\right) + \left(\sum_{i=0}^{N-1}\Theta_i w_i(\phi(t-1))\right)\psi(t-1) \\ &= \theta(\phi(t-1)) + \Theta(\phi(t-1))\psi(t-1)\end{aligned}$$

This means that the non-linear model can be written as an apparently linear model, where the parameters are dependent on the operating point. Priestley (1981) introduced *State-dependent models* which can be written

$$\hat{y}(t) = \theta(x(t-1)) + \Theta(x(t-1))\psi(t-1) \tag{21}$$

where x is the "state-vector", θ is a state-dependent vector, and Θ is a state-dependent matrix. In general $x = \psi$ was suggested, but it was also observed that this might be redundant, so a simpler vector may be used to describe the parameter dependence. The present approach with $x = \phi$ has obvious similarities. Billings & Voon (1987) discusses the use of models with *signal-dependent parameters*, which are similar to (21) this $x = \omega$, where $\omega(t)$ is the auxilliary signal. In (Billings & Voon 1987) polynomials was used to define the dependence of the parameters on the auxilliary signal, i.e. $\theta(\omega(t))$ and $\Theta(\omega(t))$ are polynomials in $\omega(t)$. A similar approach was proposed by Cyrot-Normand & Mien (1980). Our approach is also similar, but system knowledge is applied to choose the ρ_i-functions, which again defines $\theta(\phi(t))$ and $\Theta(\phi(t))$. The *Threshold AR Model* by Tong & Lim (1980) can also be written in the form (21) with $x(t-1) = y(t-1)$

$$\begin{aligned}\theta(y(t-1)) &= \begin{cases}\theta_1 & \text{If } y(t-1) \in Y_1 \\ \theta_2 & \text{If } y(t-1) \in Y_2\end{cases} \\ \Theta(y(t-1)) &= \begin{cases}\Theta_1 & \text{If } y(t-1) \in Y_1 \\ \Theta_2 & \text{If } y(t-1) \in Y_2\end{cases}\end{aligned}$$

where $Y = Y_1 \cup Y_2$. Here the parameters are switched between two possible parameter sets, and the decision is based on the value of $y(t-1)$. The resulting model is a piecewise linear model and related to our approach if $\phi(t-1) = y(t-1)$ and the interpolation functions are step-functions.

The notion of operating points and model validity functions offers a complementary method for parameterizing the state-dependence of the parameters given in (Priestley 1988).

Takagi & Sugeno (1985) suggested a fuzzy logic based technique for combining in a smooth fashion a set of linear models into a global model. It turns out that if the operating regimes Φ_i are viewed as *fuzzy sets* with membership functions equal to the model validity functions, then inference on a rulebase of the form

$$\text{IF} \quad \phi(t-1) \in \Phi_i \quad \text{THEN} \quad \hat{y}(t) = \hat{f}_i(\psi(t-1))$$

gives a resulting global model of the same form as the one analysed in the present paper, provided the fuzzy operations are properly defined. This suggests the use of fuzzy sets and rules as a means of defining the operating regimes and local model validity functions.

This is appealing since this gives a direct method of representing the empirical knowledge the engineers and operators have about the system and local models.

A related non-linear modelling approach is radial basis-functions (RBF), (Powell 1987), (Broomhead & Lowe 1988). Using RBF's, a non-linear function may be modelled as

$$\hat{f}(\psi) = \sum_{i=0}^{N-1} \theta_i r_i(||\psi - \psi_i||)$$

where $r_i : R^+ \rightarrow R$ is typically chosen as a Gaussian function. The relationship between some of these approaches is best illustrated by an example.

Example 4

We consider again the function in Example 2, and the following 4 modelling approaches:

1. Two piecewise linear models, as Example 2, centered at $\psi_0 = 0.5$ and $\psi_1 = 1.5$. This may also be interpreted as a Thresholded AR model.
2. We choose Gaussian model validity functions

$$\rho_i(\phi) = \exp\left(-\frac{1}{2}\left(\frac{\phi - \phi_i}{\Sigma_i}\right)^2\right) \tag{22}$$

   with $\phi_0 = 0.5, \phi_1 = 1.5, \Sigma_i = 0.5^2$, and use 2 local linear models.
3. 5 local 0th order models centered at $\psi_0 = 0, \psi_1 = 0.5, \psi_2 = 1, \psi_3 = 1.5, \psi_4 = 2$, and Gaussian model validity functions with $\Sigma_i = 0.5^2$,
4. A radial basis-function expansion with 5 Gaussian basis-functions centered at $\psi_0 = 0, \psi_1 = 0.5, \psi_2 = 1, \psi_3 = 1.5, \psi_4 = 2$, and $\Sigma_i = 0.5^2$.

Linear regression is used to estimate the model parameters, and the results are shown in Figure 4. By comparing Figure 4a with 4b, it is obvious that interpolating local linear models using well chosen model validity functions can improve the accuracy compared to piecewice linear models.

Notice that f now is defined on $[-1, 3]$, while data on $[0, 2]$ is used for parameter estimation. The extrapolation capabilities can thus be evaluated, and we see that the local linear approximations give 1st order extrapolation, as would be expected, while the local 0th order models give 0th order extrapolation. The RBF approach does not given any extrapolation at all, since all basis-functions go to 0. A feed-forward neural net with one hidden layer (of sigmoidal basis-functions) would give an extrapolation qualitatively similar to the 0th order models in figure 4c. As we see, there are fundamental differences concerning the extrapolation capabilities.

□

3 Modelling

The representation (16-17) is appealing since ARMAX models are simple. We represent a complex nonlinear system with a number of simple linear systems. A piecewise linear

model will have the same features, but unless we enforce the model to be continuous on the boundary between the local models, the resulting global model may be discontinuous, which may be undesirable in some cases. Enforcing continuity poses restrictions on the parameter space, giving lost representational power of the model class. The same problem arises to an even larger extent using e.g. cubic splines. Unlike the piecewise linear model, the model (16) will be smooth when the model validity functions are chosen as smooth functions. In practice, there are at least 4 ways a NARMAX model can be constuced using local ARMAX models and interpolation:

1. First we choose a set of operating regimes Φ_i that correspond to the normal equilibrium points and major transient operating regimes of the system in question. This means that we partition the set of operating points into parts where be believe that the system will behave linearly. Then we perform experiments on the system and identify a local ARMAX model $\hat{f}_i$ for each operating regime, using cost indices corresponding to each local model

$$J_i = \frac{1}{m_i} \sum_{t=1}^{m_i} \Big(y(t) - \hat{f}_i(\psi(t-1))\Big)^T \Lambda_i \Big(y(t) - \hat{f}_i(\psi(t-1))\Big) \tag{23}$$

   where m_i is the number of data-points for regime Φ_i, and Λ_i is a scaling matrix. Then the local models are integrated using system knowledge to choose sensible model validity functions $\{\rho_i\}_{i=0}^{N-1}$ such that the set Φ is covered, as shown in figure 5. The choice of these functions will strongly influence the accuracy of the global model, since the local ARMAX models are identified before the functions ρ_i are chosen.

2. Instead of choosing the model validity functions $\{\rho_i\}_{i=0}^{N-1}$ empirically, we may try to find optimal validity functions after we have found the local ARMAX models. Choosing fixed structures for the functions $\{\rho_i\}_{i=0}^{N-1}$ is necessary to make the problem finite-dimensional. Keeping the parameters of the identified ARMAX models fixed, we may search for optimal parameters of $\{\rho_i\}_{i=0}^{N-1}$ using the same data used for identifying the ARMAX models, and a global performance index. This procedure leads to a two-step optimization procedure.

3. A more direct approach is to first choose the model validity functions corresponding to the operating regimes Φ_i using system knowledge. Keeping the model validity functions fixed, we minimize the global index

$$J = \frac{1}{m} \sum_{t=1}^{m} \Big(y(t) - \hat{f}(\psi(t-1))\Big)^T \Lambda \Big(y(t) - \hat{f}(\psi(t-1))\Big) \tag{24}$$

   with respect to all parameters in the local models. Here m is the number of data-points available, and Λ is a suitable scaling matrix. Now the shape of the model validity functions will be taken into consideration when finding the optimal parameters for the local models. This has two side effects: First, the accuracy of the global model is less sensitive to the choise of $\{\rho_i\}_{i=0}^{N-1}$. Second, the local models (17) are influenced by the user-specified functions $\{\rho_i\}_{i=0}^{N-1}$. Hence, they are no longer linearizations of f about $\{\psi_i\}_{i=0}^{N-1}$.

4. An obvious improvement to methods 2 and 3 would be to search for ARMAX and local model validity function parameters simultaneously. This leads to a complex non-linear programming problem.

Required A Priori System Knowledge

When building models, different kinds of system knowledge must be available. For ARMAX models, one must first estimate the dominant time-constants of the system to choose the sampling interval. Second, the ARMAX model order must be chosen and the structure of the system disturbances must be known in order to select the MA-part of the model.

When building NARMAX models, in addition it is necessary to find a suitable structure for the system function f. First principles knowledge can be applied here, if available. We have proposed a generic structure that do not *require* first principles modelling. It is, however, not a completely black-box approach, as some limited system knowledge must be included. In order to use the local modelling approach introduced here, a priori knowledge in terms of operating regimes must be available. One must be able to estimate operating regions in which the system will behave approximately linearly.

In general, the technique with local models and interpolation may be used in an elegant fashion to integrate first principles models with black-box models, since it is completely feasible that some of the local models may be derived based on first principles, while others may be black box models (Johansen & Foss 1992*c*). In operating regimes where the dominating physical pheomena are well understood and possible to model, and in operating regimes where the data is so spares that black-box modelling is not possible, it makes sense to use local first principles models. In the remaining regimes, black-box models can be constructed, and the proposed technique with operating regime decomposition can be applied to integrate the different models. Of course, in regimes where we have limited data and limited knowledge, modelling is impossible, and the best we can hope for is some reasonable extrapolation of the neighboring local models into those regimes.

Defining the operating point in a suitable manner is important. If the local models are linear, we have shown in Theorem 3 that the operating point must capture the systems non-linearities. Given a set of data from the system, different tests for linear reationships between some inputs and the outputs is of great interest (Haber 1985). In the case with signal-dependent piecewise linear models, it is observed by Billings & Voon (1987) that the model may be input-sensitive. This must certainly be expected if the data used for identification does not cover the full range of operation. Input sensitivity and biased models may also be the result if the operating point vector is not suitably chosen, i.e. only some of the non-linearities are captured by the operating point.

4 Identification

First we consider identification of local model parameters based on the local cost indices (23), and second we consider the global cost index (24). Finally, we consider identification of local model validity function parameters and model structure identification.

Identifying local model parameters using local cost indices

The prediction error at time t for the local model $\hat{f}_i$ is defined to be

$$\epsilon_i(t) \quad = \quad y(t) - \hat{f}_i(\psi(t-1))$$

The cost index J_i associated with the local model can be written as

$$J_i = \frac{1}{m_i} \sum_{t=1}^{m_i} \epsilon_i^T(t) \Lambda_i \epsilon_i(t) \tag{25}$$

Consider the local ARMAX model (17). The local models are parameterized by a vector θ_i and a matrix Θ_i. Since all local models (17) are linear functions of the parameters, the representation is basically linear in the parameters, and standard identification methods can be applied, e.g. (Söderström & Stoica 1988).

Assume first the noise $\epsilon(t)$ is sequentially uncorrelated. Since the information vector $\psi(t-1)$ do not contain noise terms $e(t-1), ..., e(t-n_e)$, the model can be written on the linear regression form

$$\hat{f}_i(\psi(t-1)) = \varphi_i^T(t-1)\overline{\theta}_i \tag{26}$$

where $\overline{\theta}_i$ is a parameter vector and $\varphi_i(t-1)$ is a regression matrix.

The parameters can be estimated using the least squares (LS) method. The regression matrix, $\varphi_i(t-1)$, is a matrix of computable or measurable quantities not depending on the parameter vector and not correlated with $e(t)$. The least squares estimate minimizing (25) can be written as

$$\hat{\overline{\theta}}_i = \left(\frac{1}{m_i} \sum_{t=1}^{m_i} \varphi_i(t) \Lambda_i \varphi_i(t)^T \right)^{-1} \left(\frac{1}{m_i} \sum_{t=1}^{m_i} \Lambda_i \varphi_i(t) y(t) \right) \tag{27}$$

In the general case when delayed noise, $e(t-1), e(t-2), ..., e(t-n_e)$, is included in $\psi(t)$, we use the prediction error (PE) method. Since the noise is assumed not to be measurable we do not know the values of $e(t-1), ..., e(t-n_e)$. If the model matches the true system, then $\epsilon_i(t) = e(t)$. Now $\epsilon_i(t-1)$ depends on the parameters $\overline{\theta}_i$ since it is the prediction error. Since $\varphi_i(t)$ depends on $\epsilon_i(t-1)$ and hence $\overline{\theta}_i$, we may conclude that the predictor (26) is no longer linear in the parameters. Hence it is not possible to find a simple analytic solution like (27). The cost indices (25) must be minimized numerically using e.g. the Newton-Raphson algorithm

$$\hat{\overline{\theta}}_i^{(k+1)} = \hat{\overline{\theta}}_i^{(k)} - \alpha_k \left(\nabla^2 J_i \left(\hat{\overline{\theta}}_i^{(k)} \right) \right)^{-1} \nabla J_i \left(\hat{\overline{\theta}}_i^{(k)} \right)$$

or the Gauss-Newton algorithm, which is based on simplified calculations of the inverse Hessian matrix.

Both the LS-method and PE-method can be formulated recursively. The RPE-algorithm is

$$\hat{\overline{\theta}}_i(t) = \hat{\overline{\theta}}_i(t-1) + K_i(t)\epsilon_i(t) \tag{28}$$

$$K_i(t) = P_i(t)\Psi(t)\Lambda_i \tag{29}$$

$$P_i(t) = P_i(t-1) - P_i(t-1)\Psi(t) \left(\Lambda_i^{-1} + \Psi^T(t) P_i(t-1) \Psi(t) \right)^{-1} \Psi^T(t) P_i(t-1) \tag{30}$$

with $\Psi(t) = -\frac{\partial \epsilon_i}{\partial \overline{\theta}_i}(t)$.

Identifying local model parameters using a global cost index

We define the global prediction error to be

$$\epsilon(t) = y(t) - \hat{f}(\psi(t-1))$$

The global cost index (24) can be rewritten as

$$J = \frac{1}{m}\sum_{t=1}^{m} \epsilon^T(t)\Lambda\epsilon(t) \tag{31}$$

The LS- and PE-methods can be formulated in the same manner as above. When the identification is performed on-line, some rather heuristic modification may be desirable. As pointed out in (Johansen & Foss 1992*b*), only the parameters of those local models that are assumed to be valid in the current operating regime should be updated. This is easily accomplished by assuring that the model validity function ρ_i is exactly zero for operating points where the local model is not valid. In practice, however, we would like the functions ρ_i to be smooth, i.e. they should go fast to zero instead of being exactly zero. This may cause problems when the system is operating within one operating regime for a long time. Then all the other local models will be updated slightly each time-step, and information about other operating regimes will "leak out", in particular if forgetting factors are used.

The heuristics we propose to eliminate this problem is to update only the parameters of those local models satisfying $w_i(\phi(t)) > \delta$, where typically $\delta \in [0.1, 0.5]$. It is a danger that parameters being an active part of the global model may never be updated. This requires that the system is excitated such that the parameters of all local models will be updated from time to time.

Identifying model validity function parameters

In general, the local model validity function parameters will enter the equations for the prediction error nonlinearly. In particular, if these parameters are to be identified simulateously with the local model parameters, we get a complex non-linear programming problem. We will not discuss this problem here, but refer to the vast litterature on non-linear programming, e.g. (Gill, Murray & Wright 1981). We will like to point out that most of our simulations so far indicate that a rough empirical choise of model validity functions combined with local model identification based on a global cost index in most cases gives better results than the use of local indices and subsequent optimization of the local model validity functions.

Identifying model structure

We have suggested how knowledge about operating regimes may be used to decompose the modelling problem into the problem of building simple ARMAX models corresponding to each operating regime. Together with model validity functions, this gives a model structure where only the parameters are unknown. In some cases, such knowledge may not be available and we need methods to find an adequate model structure, i.e., the number N of local models, and the structure of each local model. From the parsimony principle we know that the best model is the model with fewest parameters able to describe the system adequately. There are several theoretical frameworks that deal with this problem.

1. The prediction error using a good model should be uncorrelated with past prediction errors and inputs (and future inputs if the system operates in open loop). There are several correlation-based test, see eg. (Billings & Voon 1986).

2. If we consider the expected error criterion $\tilde{J}$ on *future data*, we obtain (Söderström & Stoica 1988)

$$\tilde{J} = (1 + P/m)J \tag{32}$$

This depends on the number of parameters P and the numbers of data points m used for identification. This is also related to the Akaike Information Criterion, (Akaike 1969)

In general J will decrease when more parameters are introduced in the model, e.g. when new local models are added. However, P will increase, and at some stage the increment in $1 + P/m$ will be larger than the decrease in J, if m is kept fixed. The index $\tilde{J}$ will then increase, indicating that the quality of the model decreases. It is therefore important to keep the number of local models at a minimum, and use as simple local models as possible.

Both these methods can be implemented off-line by an exhaustive search. Not all of them are suited for on-line structure identification, however. It requires large amounts of computer power to test more than 2-3 model hypotheses simultaneously. Generally *a batch of data* is required to perform any statistical test with some significance level. A large prediction error at some time sample may be due to noise, parameter errors or inadequate model structure. Collecting a batch of data over some time will reduce the impact of noise. If the batch is so large that the parameter estimator is expected to converge during the batch, the impact of parameter errors may also be insignificant. Hence, if the batch is large and the prediction error is biased or correlated, we can infer that the model structure is not good. In our approach with local models, we must decide which local models are the cause for the mismatch. This can be found by collecting statistics locally for every local model. The problem of on-line structure identification is discussed in (Johansen & Foss 1992*c*).

Time-varying systems

In case of time varying parameters the RPE-formulation (28)-(30) must be modified. This is usually done by introducing a method for artificially increasing the covariance matrix estimate P so it will not go to zero. The most common schemes are linear increase, which correspond to the Kalman filter, exponential increase, and covariance resetting.

At one time-step we get the information vector $\psi(t)$. The information vector is transformed to the vector $w(t) = [w_0(\phi(t)) \cdots w_{N-1}(\phi(t))]^T$. This vector correspond to the direction in *interpolation-space* where we get information. The interpolation-space consists of N dimensions, one for each local model. If forgetting is to be avoided, we should only update models along the direction in the model space where we get new information. This means that we should only forget if we know that new information will arrive to compensate for the forgotten information. At the operating regime level, this is rather simple in our case since by construction the components of $w(t)$ will be close to zero when the information is not relevant for the corresponding local models. Hence, by thresholding the parameter update, only the local models we get new information about will be updated. This leads to a small set of local models to be updated at each time-step. Within each local model, techniques based on the same type of reasoning can be applied to only update in the directions in parameters-space where we get new information, (Fortescue, Kershenbaum & Ydstie 1981, Sælid, Egeland & Foss 1985, Parkum, Poulsen & Holst 1990).

5 Simulation example

We will use the proposed approach to identify a model of a Continuous Stirred Tank Reactor (CSTR) where a first order, exothermic chemical reaction $A \rightarrow B$ takes place. The system is described by the following mass- and energy-balance

$$V\frac{d}{dt}c_A = c_{Ai}q_i - c_A q_o - V r_A \tag{33}$$

$$\rho c_p V\frac{d}{dt}T = \rho c_p T_i q_i - \rho c_p T q_o + Q - \Delta H_r V r_A \tag{34}$$

$$r_A = k_0 c_A \exp\left(-\frac{E_A}{R}(\frac{1}{T} - \frac{1}{T_R})\right) \tag{35}$$

with symbols as defined in Table 1.

Symbol	Value	Unit	Description
T	350-400	K	Reactor temperature
T_i	310	K	Inlet temperature
T_R	350	K	Reference temperature
c_A	ca. 1	$kmol/m^3$	Concentration of A in reactor
c_{Ai}	10	$kmol/m^3$	Inlet concentration of A
E_A	70000	$kJ/kmol$	Activation energy
R	8.314	$kJ/(Kkmol)$	Gas constant
k_0	0.042	$1/min$	Frequency factor
ρ	1.0	kg/l	Ave. density in reactor
c_p	4.0	$kJ/(Kkg)$	Ave. heat capacity in reactor
V	10.0	m^3	Reactor volume
ΔH_r	-90000	$kJ/kmol$	Reaction energy
q_i	ca. 50-500	l/min	Inlet flow
q_o	ca. 50-500	l/min	Outlet flow
Q	ca. -6 - 3	MW	External power from heat exchanger
r_A		$kmol/l$	Reaction rate

Table 1: Symbols.

Inputs to the system are the power Q and the inlet flow q_i. Outputs are temperature T and concentration c_A. We define the vectors $y = [5 \cdot 10^{-4} \cdot c_A \quad 5 \cdot 10^{-3} \cdot T]^T$ and $u = [50 \cdot q_i \quad 10^{-7} \cdot Q]^T$. The system is simulated as a continuous-time system. Two independent sequences $\{(y(t), u(t))\}_{t=1}^{1000}$ are collected by sampling the system every 2 minutes, see figure 6. One set is used for parameter identification only, and the other is used for model validation only.

The reactor is open-loop unstable. There are therefore two stabilizing single-loop PI-controllers. The reference signal to the controllers are LP-filtered white noise signals, ensuring the input is rich.

In our example we choose the information vector $\psi^T(t-1) = [y^T(t-1) \; u^T(t-1)]$. We have performed 4 simulations, and a least squares algorithm is used in all cases:

Simulation 1

First, an ARMAX model

$$\begin{aligned}\hat{y}(t) &= \varphi^T(t-1)\theta \qquad (36)\\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ y_1(t-1)-y_1^\star & 0 \\ y_2(t-1)-y_2^\star & 0 \\ 0 & y_1(t-1)-y_1^\star \\ 0 & y_2(t-1)-y_2^\star \\ u_1(t-1)-u_1^\star & 0 \\ u_2(t-1)-u_2^\star & 0 \\ 0 & u_1(t-1)-u_1^\star \\ 0 & u_2(t-1)-u_2^\star \end{pmatrix}^T \cdot \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \\ \theta_7 \\ \theta_8 \\ \theta_9 \\ \theta_{10} \end{pmatrix}\end{aligned}$$

is fitted using linear regression on the data. The points $[y_1^\star\ y_2^\star]^T$ and $[u_1^\star\ u_2^\star]^T$ are the points the ARMAX model is linearized about, and are chosen as mean values of the respective signals.

Simulation 2

Second, a NARMAX model with two local ARMAX models (36) is fitted. If c_{Ai} and T_i are constant, the model (33)-(35) can be written

$$\frac{d}{dt}y(t) = f(y(t),u(t)) = f_1(y(t)) + f_2(y(t))u(t) \qquad (37)$$

Discretizing by the Euler method gives

$$y(t) = y(t-1) + \Delta t(f_1(y(t-1)) + f_2(y(t-1))u(t-1)) \qquad (38)$$

Comparing this with (18), Theorem 3 indicates that the operating point $\phi = y$ captures the nonlinearities. In order to keep the example simple, we observe that the system exhibit strongest non-linear behavior with respect to temperature. As a first approximation, it is therefore reasonable to choose $\phi(t) = y_2(t) = 5 \cdot 10^{-4} \cdot T(t)$. The 2 local model validity functions are chosen as Gaussians (22) with $\phi_0 = 1.8$ and $\phi_1 = 1.9$ (corresponding to 360 K and 380 K). These values are found by examining the operating range the data span. The size of the operating regimes are estimated to be $\Sigma_i = 0.05$ (corresponding to 10 K). The identification of the two ARMAX models is performed separately using local cost indices (23).

Simulation 3

The same models structure as simulation 2, but a global cost index (24) is used to simultaneously identify the parameters of the two ARMAX models.

Simulation 4

Finally, we use 4 local ARMAX models and a global cost index (24). The local model validity functions are still Gaussians, centered at $\phi_0 = 1.775, \phi_1 = 1.825, \phi_2 = 1.875$, and

$\phi_3 = 1.925$ corresponding to $355, 365, 375, 385$ K respectively. The width parameter is chosen as $\Sigma_i = 0.02$, corresponding to 4 K.

Results

The results are summarized in table 2 (all the results are using the identified model for prediction on the data independent of the data used for identification). We see that all NARMAX approaches give significantly better results than the ARMAX approach. As would be expected, better results are achieved with a global performance index than local indices, and increasing the number of local models also improve on the model accuracy.

One-step-ahead prediction errors are shown in figure 7. The curves indicates that the prediction error is considerably reduced using the NARMAX models compared to the ARMAX models.

The covariance functions for the prediction errors are

$$E[(\epsilon(t) - \overline{\epsilon})(\epsilon(t+\tau) - \overline{\epsilon})^T] = \begin{pmatrix} r_{11}(\tau) & r_{12}(\tau) \\ r_{21}(\tau) & r_{22}(\tau) \end{pmatrix} \tag{39}$$

Estimates of the autocorrelation functions are shown in figure 8. It is well known that an unbiased model gives an autocorrelation function equal to a δ-function (This is not a sufficient condition. A more detailed model validation should also consider $E[(u(t) - \overline{u})(\epsilon(t+\tau) - \overline{\epsilon})^T]$ and higher order covariances (Billings & Voon 1986)). However, we may conclude from the curves that the NARMAX models strongly improve on the model accuracy compared to the ARMAX models. We may not expect a perfect model because the model structure is different than the true system: First, there is fundamental structural difference between the state-space system and the model based on local models and interpolation. The low number of local models limits the accuracy. Second, there is a structure error introduced by sampling the continuous system. Third, the simplified choise of operating point introduces a nonsystematic error.

We have also done simulations using optimization of the Σ_i-parameters in the local model validity functions after the local ARMAX parameters was identified using local cost indices. This gave substantial improvements, but the results was not as good as using a global index with prechosen (not optimal) Σ_i-parameters. Including $y_1(t)$ in the operating point vector also gave some improvements, but not as much as one may have expected. Hence, we can conclude that the choise of operating point is sensible.

6 Discussions and Conclusions

As discussed in the introduction, this model representation may be useful when first principles modelling is resource demanding or does not give satisfactory results. Today, ARMAX models are no doubt the most widely used black-box type model in industry. Neural networks have over the past few years gained considerable popularity, at least academically. The most popular feed-forward type networks can be used to build black box NARMAX models, typical examples can be found in (Chen et al. 1990*a*, Nguyen & Widrow 1990). Common to most types of neural networks is that the representation is very complex and difficult to understand and validate, which may explain why so few industrial applications are reported.

Simulation	Parameters	Variance: $E\left[(\epsilon(t)-\overline{\epsilon})(\epsilon(t)-\overline{\epsilon})^T\right]$
Simulation 1: ARMAX model	$\theta = \begin{pmatrix} 0.5055 \\ 1.8533 \\ 0.6256 \\ -1.9255 \\ 0.0755 \\ 1.4010 \\ 0.8599 \\ -0.1189 \\ -0.0336 \\ 0.1758 \end{pmatrix}$	$\begin{pmatrix} 9.40 & -1.77 \\ -1.77 & 0.34 \end{pmatrix} \cdot 10^{-4}$
Simulation 2: NARMAX model with 2 local ARMAX models. The local models are fitted separately using local prediction errors.	$\theta_1 = \begin{pmatrix} 0.5985 \\ 1.7812 \\ 0.7470 \\ -1.3276 \\ 0.0514 \\ 1.2817 \\ 0.9842 \\ -0.1043 \\ -0.0486 \\ 0.1721 \end{pmatrix}$ $\theta_2 = \begin{pmatrix} 0.3931 \\ 1.9235 \\ 0.4694 \\ -2.9493 \\ 0.1069 \\ 1.5959 \\ 0.8765 \\ -0.1875 \\ -0.0410 \\ 0.1890 \end{pmatrix}$	$\begin{pmatrix} 2.10 & -0.40 \\ -0.40 & 0.09 \end{pmatrix} \cdot 10^{-4}$
Simulation 3: NARMAX model two local ARMAX models. The local models are fitted using the global prediction error.	$\theta_1 = \begin{pmatrix} 0.5771 \\ 1.7859 \\ 0.9096 \\ -1.4806 \\ 0.0185 \\ 1.294 \\ 1.0534 \\ -0.0549 \\ -0.0534 \\ 0.1620 \end{pmatrix}$ $\theta_2 = \begin{pmatrix} 0.3865 \\ 1.9255 \\ 0.3176 \\ -3.3892 \\ 0.1381 \\ 1.6952 \\ 0.8493 \\ -0.2342 \\ -0.0418 \\ 0.1983 \end{pmatrix}$	$\begin{pmatrix} 0.96 & -0.19 \\ -0.19 & 0.04 \end{pmatrix} \cdot 10^{-4}$
Simulation 4: NARMAX model with 4 local NARMAX models. The global prediction error is used.	$\theta_1 = \begin{pmatrix} 0.6185 \\ 1.7528 \\ 0.8624 \\ -1.9076 \\ 0.0276 \\ 1.1933 \\ 1.0209 \\ -0.0573 \\ -0.0467 \\ 0.1626 \end{pmatrix}$ $\theta_2 = \begin{pmatrix} 0.5659 \\ 1.8131 \\ 0.7364 \\ -1.7990 \\ 0.0538 \\ 1.3712 \\ 0.9867 \\ -0.1131 \\ -0.0492 \\ 0.1737 \end{pmatrix}$ $\theta_3 = \begin{pmatrix} 0.4643 \\ 1.8839 \\ 0.5436 \\ -2.7944 \\ 0.0918 \\ 1.5668 \\ 0.9428 \\ -0.1685 \\ -0.0505 \\ 0.1847 \end{pmatrix}$ $\theta_4 = \begin{pmatrix} 0.2971 \\ 1.9696 \\ 0.1970 \\ -4.0579 \\ 0.1642 \\ 1.8275 \\ 0.8471 \\ -0.2624 \\ -0.0433 \\ 0.2042 \end{pmatrix}$	$\begin{pmatrix} 0.38 & -0.08 \\ -0.08 & 0.02 \end{pmatrix} \cdot 10^{-4}$

Table 2: Results for ARMAX and NARMAX model fitting.

Decomposing the system operation into several operating regimes and using local ARMAX models to describe each operating regime is appealing for several reasons:

- It is possible to integrate several kinds of system knowledge, including local first principles models, with a black-box type model. ARMAX models are well understood and widely used in industry, and is hence a convenient basis for building NARMAX models.
- The class of systems that can be represented is large, and a linear parameterization of the model is sufficient.
- The concept is straightforward, and the model structure is easy to understand. This is important, since the model structure can be easily validated. In addition, some validation may be performed by validating each local model separately.
- Describing a system by means of operating regimes is common practise in engineering. Integrating the models for each operating regime by interpolation have not been common so far, but seems to be a straightforward way to build models that are valid within several operating regimes. The interpolation may improve the accuracy of the model, compared to piecewise linear models. In addition, the smoothness of the model is an inherent property.

A fundamental problem with all local modelling methods is the curse of dimensionallity problem (Bellman 1961, Moody & Darken 1989, Tolle et al. 1992, Friedman 1991). In our case, we have shown that this problem may sometimes be reduced considerably, since the operating point may be of lower dimension than the information vector.

To summarize, we have investigated how non-linear systems can be modelled using NARMAX models based on local ARMAX models. The primary result is that given a sufficient number of local models and well defined operating regimes, the system function can be approximated to arbitrary accuracy. In practice, noise and the amount of data available will limit the attainable accuracy. Standard identification algorithms can easily be applied since the proposed representation will be linear in the parameters. The empirical choice of model validity function may however complicate the problem. Several practical aspects of building such models are outlined and illustrated by a simulation example.

The approach falls somewhat between first principles modelling and pure black-box modelling. Available local first principles models as well as a priori knowledge in terms of operating regimes, can be incorporated in the model.

Acknowledgements

This work was supported by the Royal Norwegian Council for Scientific and Industrial Research (NTNF) under doctoral scholarship grant no. ST. 10.12.221718 given to the first author.

References

Akaike, H. (1969), 'Fitting autoregressive models for prediction', *Ann. Inst. Stat. Math.* **21**, 243–247.

Bellman, R. E. (1961), *Adaptive Control Processes*, Princeton Univ. Press.

Billings, S. A. & Voon, W. S. F. (1986), 'Correlation based model validity tests for non-linear models', *Int. J. Control* **44**(1), 235–244.

Billings, S. A. & Voon, W. S. G. (1987), 'Piecewise linear identification of non-linear systems', *Int. J. Control* **46**, 215–235.

Broomhead, D. S. & Lowe, D. (1988), 'Multivariable functional interpolation and adaptive networks', *Complex Systems* **2**, 321–355.

Chen, S. & Billings, S. A. (1989), 'Representation of non-linear systems: the NARAMX model', *Int. J. Control* **49**(3), 1013–1032.

Chen, S., Billings, S. A. & Grant, P. M. (1990*a*), 'Non-linear system identification using neural networks', *Int. J. Control* **51**(6), 1191–1214.

Chen, S., Billings, S. A., Cowan, C. F. N. & Grant, P. (1990*b*), 'Practical identification of NARMAX models using radial basis functions', *Int. Journal of Control* **52**(6), 1327–1350.

Cyrot-Normand, D. & Mien, H. D. V. (1980), Non-linear state-affine identification methods: Application to electrical power plants, *in* 'Proc. IFAC Symposium on Automatic Control in Power Generation, Distribution and Protection', pp. 449–462.

Fortescue, T. R., Kershenbaum, L. S. & Ydstie, B. E. (1981), 'Implementation of self-tuning regulators with variable forgetting factors', *Automatica* **17**, 831–835.

Friedman, J. H. (1991), 'Multivariable adaptive regression splines', *The Annals of Statistics* **19**, 1–141.

Gill, P., Murray, W. & Wright, M. (1981), *Practical optimization*, Academic Press, Inc.

Haber, R. (1985), Nonlinearity tests for dynamics processes, *in* '7th IFAC Symp. on Identification and System Parameter Estimation', pp. 409–414.

Hilhorst, R. A., van Amerongen, J. & Löhnberg, P. (1991), Intelligent adaptive control of mode-switch processes, *in* 'Proc. IFAC Internatinal Symposium on Intelligent Tuning and Adaptive Control, Singapore'.

Johansen, T. A. & Foss, B. A. (1992*a*), 'A NARMAX model representation for adaptive control based on local models', *Modeling, Identification, and Control* **13**(1), 25–39.

Johansen, T. A. & Foss, B. A. (1992*b*), Nonlinear local model representation for adaptive systems, *in* 'Proceeding of the Singapore Int. Conf. on Intelligent Control and Instrumentation', Vol. 2, pp. 677–682.

Johansen, T. A. & Foss, B. A. (1992*c*), Representing and learning unmodeled dynamics with neural network memories, *in* 'Proceedings of the American Control Conference, Chicago, Il.', Vol. 3.

Jones, R. D., Lee, Y. C., Barnes, C. W., Flake, G. W., Lee, K., Lewis, P. S. & Qian, S. (1989), Function approximation and time series prediction with neural networks, Technical Report 90-21, Los Alamos National Lab., New Mexico.

Lane, S. H., Handelman, D. A. & Gelfand, J. J. (1992), 'Theory and development of higher-order CMAC neural networks', *IEEE Control System Magazine* **12**(2), 23–30.

Leontaritis, I. J. & Billings, S. A. (1985), 'Input-output parametric models for non-linear systems', *Int. Journal of Control* **41**(2), 303–344.

Jones et al., R. D. (1991), Nonlinear adaptive networks: A little theory, a few applications, Technical Report 91-273, Los Alamos National Lab., New Mexico.

Sælid, S., Egeland, O. & Foss, B. (1985), 'A solution to the blow-up problem in adaptive controllers', *Modeling, Identification and Control* **6**(1), 39–56.

Söderström, T. & Stoica, P. (1988), *System Identification*, Prentice Hall.

Moody, J. & Darken, C. J. (1989), 'Fast learning in networks of locally-tuned processing units', *Neural Computation* **1**, 281–294.

Nguyen, D. H. & Widrow, B. (1990), 'Neural networks for self-learning control systems', *IEEE Control Systems Magazine* **10**(3), 18–23.

Omohundro, S. M. (1987), 'Efficient algorithms with neural network behavior', *Complex Systems* **1**, 273–347.

Parkum, J. E., Poulsen, N. K. & Holst, J. (1990), Selective forgetting in adaptive procedures, *in* 'Proc. 11th IFAC World Congress, Tallin, Estonia'.

Polycarpou, M. M., Ioannou, P. A. & Ahmed-Zaid, F. (1992), Neural networks and on-line approximators for discrete-time nonlinear system identification, Submitted to IEEE Trans. Control Systems Technology.

Powell, M. J. D. (1987), Radial basis function approximations to polynomials, *in* '12th Biennal Numerical Analysis Conference, Dundee', pp. 223–241.

Priestley, M. B. (1981), *Spectral Analysis and Time Series*, Academic Press.

Priestley, M. B. (1988), *Non-linear and Non-stationary Time Series Analysis*, Academic Press.

Psichogios, D. C., De Veaux, R. D. & Ungar, L. H. (1992), Nonparametric system identification: A comparison of MARS and neural nets, *in* 'Proc. American Control Conference, Chicago, Il.'.

Rippin, D. W. T. (1989), Control of batch processes, *in* 'Proceedings DYCORD+ 89, August, Maastrict, The Netherlands', pp. 115–125.

Skeppstedt, A. (1988), Construction of Composite Models from large data-sets, PhD thesis, University of Linköping.

Skeppstedt, A., Ljung, L. & Millnert, M. (1992), 'Construction of composite models from observed data', *Int. J. Control* **55**(1), 141–152.

Stokbro, K. (1991), Predicting chaos with weighted maps, Technical Report 91/10 S, NORDITA, Copenhagen.

Stokbro, K. & Umberger, D. K. (1990), Forecasting with weighted maps, *in* 'Proc. 1990 Workshop on Nonlinear Modeling and Forecasting, Santa Fe Institute'.

Stokbro, K., Hertz, J. A. & Umberger, D. K. (1990), Exploiting neurons with localized receptive fields to learn chaos, Preprint 28, Niels Bohr Institute and NORDITA, Copenhangen. Submitted to Journal of Complex Systems.

Stromberg, K. R. (1981), *An Introduction to Classical Real Analysis*, Wadsworth, Inc., Belmont, Ca.

Takagi, T. & Sugeno, M. (1985), 'Fuzzy identification of systems and its application to modeling and control', *IEEE Trans. Systems, Man, and Cybernetics* **15**, 116–132.

Tolle, H., Parks, P. C., Ersü, E., Hormel, M. & Militzer, J. (1992), 'Learning control with interpolating memories – general idead, design lay-out, theoretical approaches and practical applications', *Int. J. Control* **56**, 291–317.

Tong, H. & Lim, K. S. (1980), 'Threshold autoregression, limit cycles and cyclical data', *J. Royal Stat. Soc. B* **42**, 245–292.

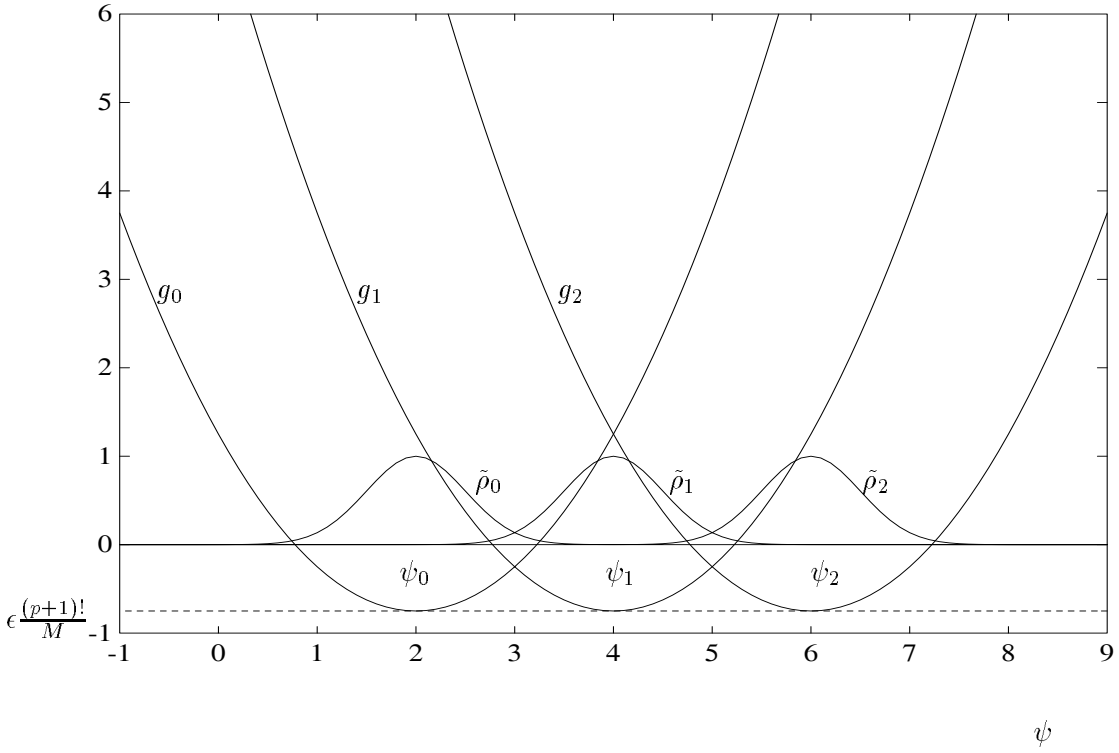

Figure 1: A geometric interpretation of the constraints on $\tilde{\rho}_i$

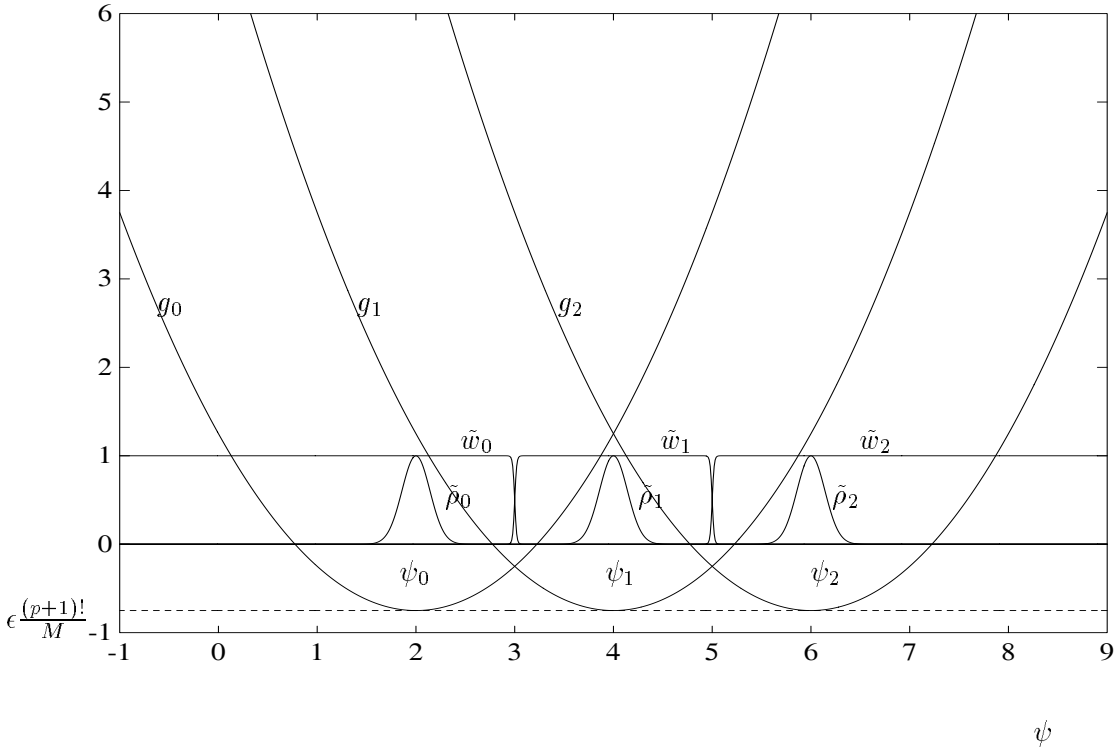

Figure 2: Situation when the width of $\tilde{\rho}_i$ goes to zero.

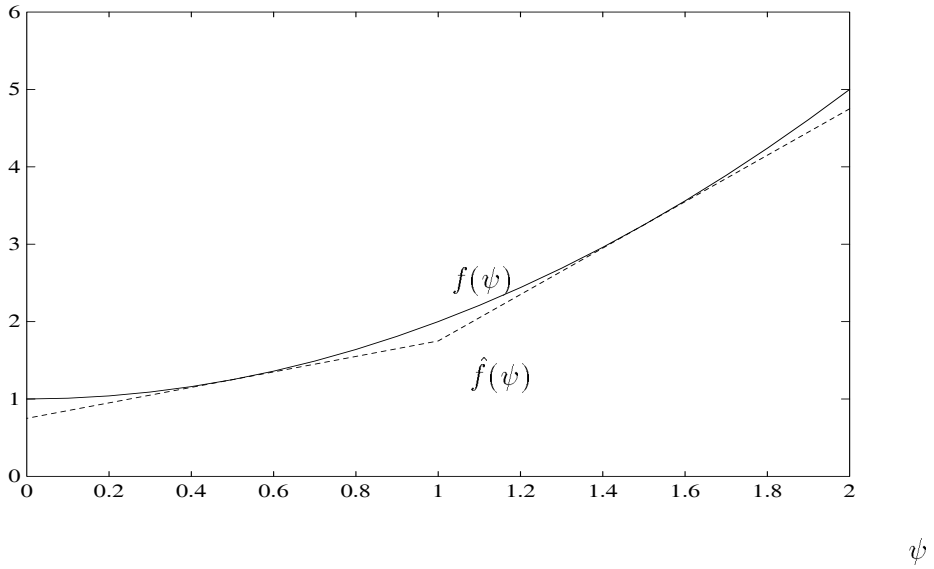

Figure 3: Approximation of $f(\psi) = \psi^2 + 1$ using two local linear models.

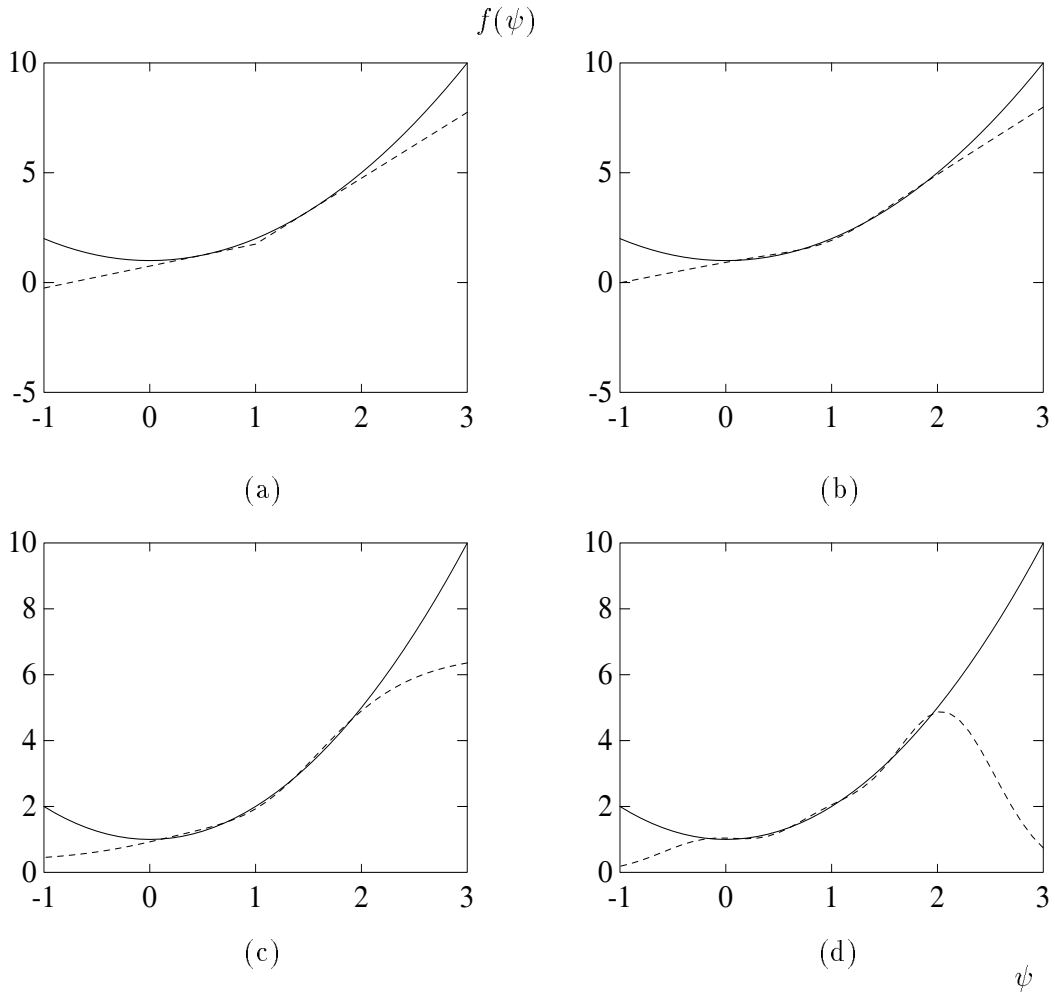

Figure 4: a) Approximation of $f(\psi) = \psi^2 + 1$ using a piecewise linear model with two local linear models. b) Approximation using two local linear models and Gaussian model validity functions. c) Approximation using 5 local 0th order expansions (constant) d) Approximation using a radial basis-function expansion with 5 Gaussian basis-functions.

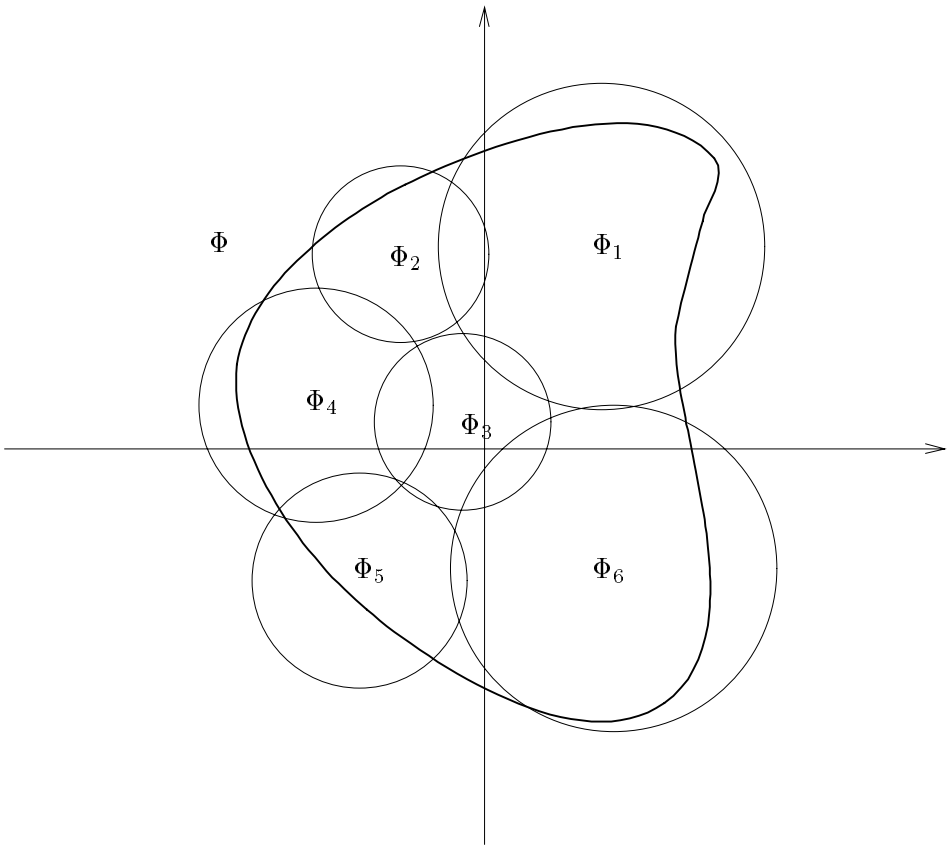

Figure 5: The set Φ of operating points is covered with local models. The figure shows lines where the model validity functions ρ_i are constant.

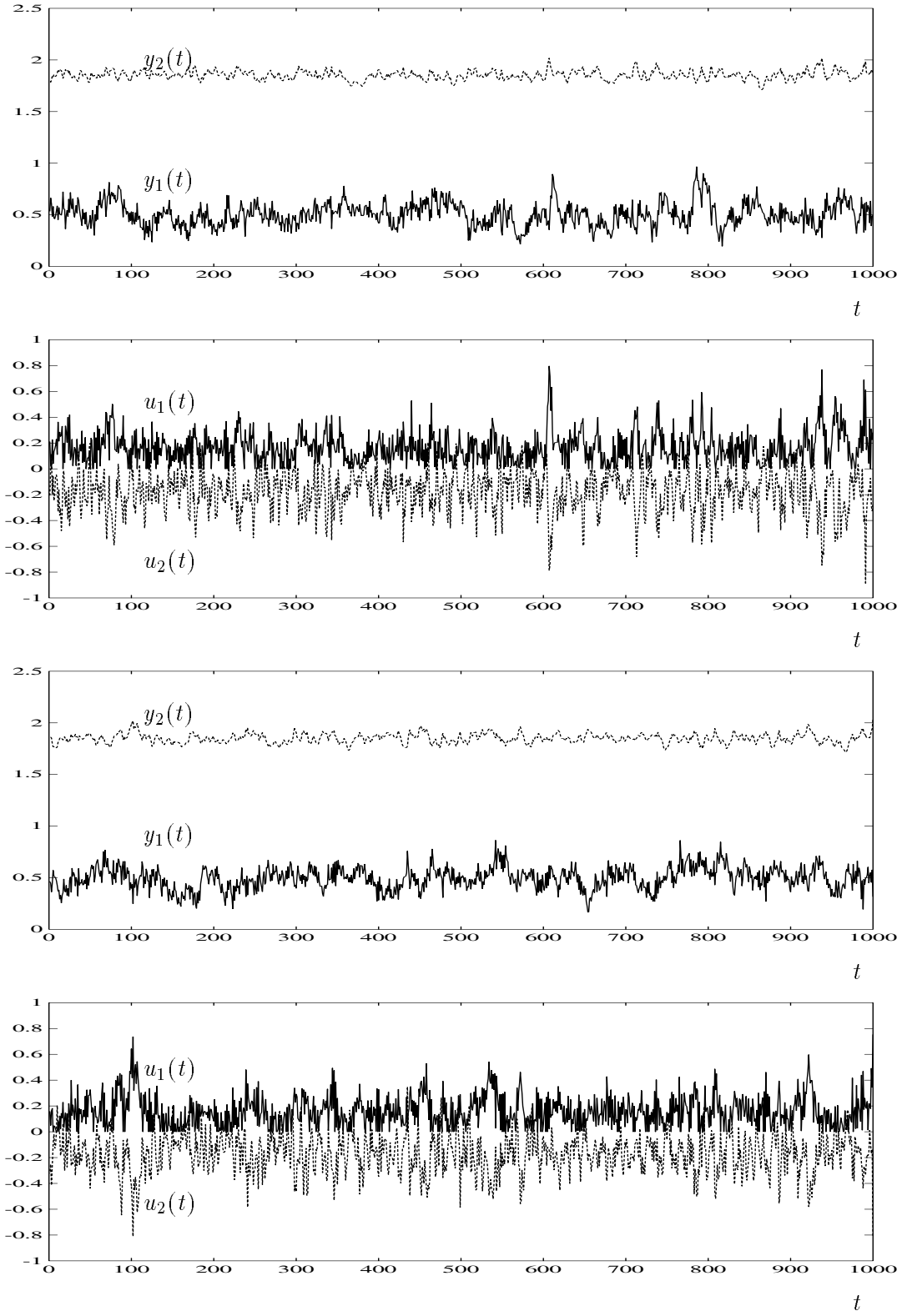

Figure 6: The two first curves show output and input data used for identification, while the two last curves show data used for validation.

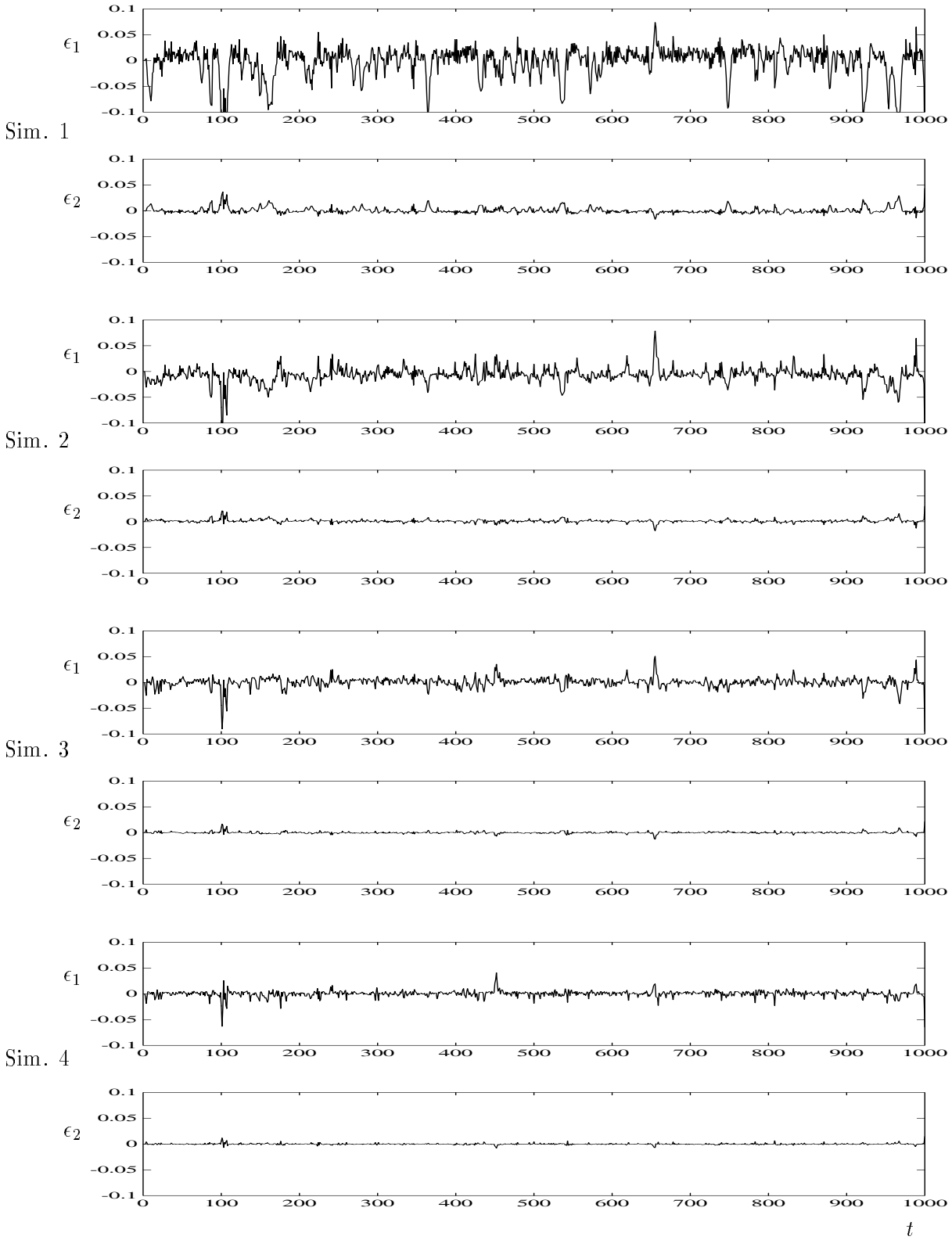

Figure 7: Prediction errors on the independent validation data for the resulting models of the 4 simulations. Here $\epsilon_1 = y_1 - \hat{y}_1$, and $\epsilon_2 = y_2 - \hat{y}_2$.

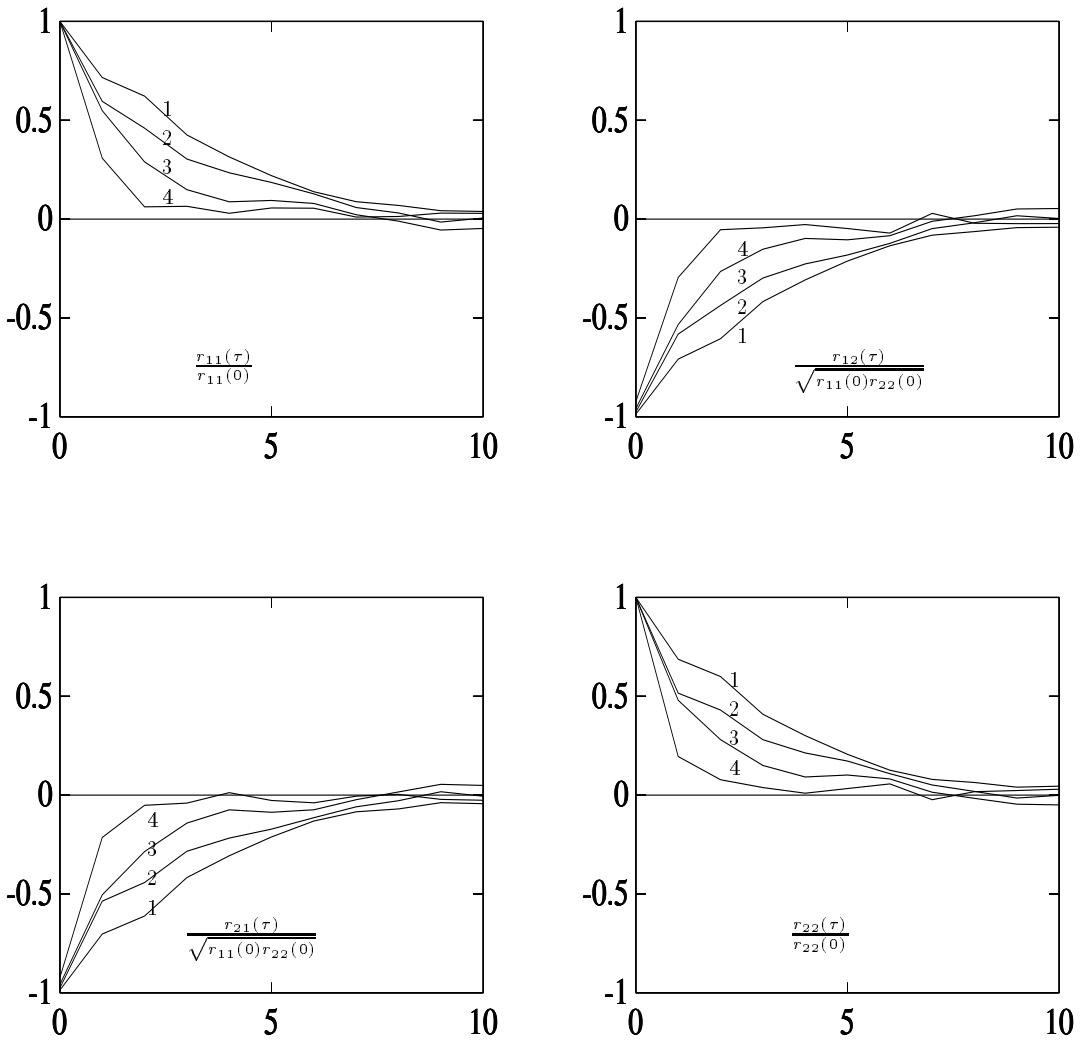

Figure 8: The curves show the autocorrelation functions for the prediction errors, for the resulting models of the 4 simulations.