Semi-Empirical Modeling of Non-Linear Dynamic Systems through Identification of Operating Regimes and Local Models

Tor A. Johansen and Bjarne A. Foss

Department of Engineering Cybernetics
Norwegian Institute of Technology
7034 Trondheim, Norway

Abstract. An off-line algorithm for semi-empirical modeling of non-linear dynamic systems is presented. The model representation is based on the interpolation of a number of simple local models, where the validity of each local model is restricted to an operating regime, but where the local models yield a complete global model when interpolated. The input to the algorithm is a sequence of empirical data and a set of candidate local model structures. The algorithm searches for an optimal decomposition into operating regimes, and local model structures. The method is illustrated using simulated and real data. The transparency of the resulting model and the flexibility with respect to incorporation of prior knowledge is discussed.

1 Introduction

The problem of identifying a mathematical model of an unknown system from a sequence of empirical data is a fundamental one which arises in many branches of science and engineering. The complexity of solving such a problem depends on many factors, such as a priori knowledge, quality and completeness of the data sequence, and required model form and accuracy.

A rich non-linear model representation based on patching together a number of simple local models into a complex global model is suggested in [30, 27, 13, 8, 9]. With this representation, the modeling problem is basically to decompose the operating range of the system into a set of operating regimes, the identification of simple local models within each regime, and the interpolation of the local models to get a global model. This is an example of the classical divide-and-conquer strategy, where a complex problem is decomposed into simple subproblems that can be solved independently, and whose solutions add up to solve the complex problem. In [9, 10] we have focused on the use of system knowledge for regime decomposition. The aim of the present paper is to report on a algorithm that automatically finds a decomposition and local models based on empirical data.

The paper is organized as follows: The problem is formulated in section 2, before an empirical modeling algorithm is developed in section 3. The algorithm is applied to simulated and real data in section 4, and in section 5 the role of prior knowledge and the transparency of the model are discussed. A comparison to related work follows, along with some concluding remarks.

2 Problem Formulation

We address the problem of identifying a model of an unknown non-linear system on the basis of a sequence of l input/output-pairs

$$\mathcal{D}_l = ((u(1), y(1)), (u(2), y(2)), ..., (u(l), y(l)))$$

where $u(t) \in R^r$ and $y(t) \in R^m$ are the input and output vectors of the system, respectively. We denote by $\mathcal{D}_t$ the subsequence of $\mathcal{D}_l$ containing data up to and including time $t \leq l$. First, consider static models

$$y(t) = f(u(t)) + e(t) \tag{1}$$

where $e(t) \in R^m$ is zero-mean noise, and f is an unknown function to be estimated. An approximation $\hat{f}$ to f suggests the predictor $\hat{y}(t|u(t)) = \hat{f}(u(t))$ which gives the prediction error

$$\varepsilon(t) = y(t) - \hat{y}(t|u(t)) \;=\; \left(f(u(t)) - \hat{f}(u(t))\right) + e(t)$$

Next, consider a stable dynamic system represented by the NARMAX (non-linear ARMAX) model

$$\begin{aligned} y(t) = f(y(t-1), ..., y(t-n_y), u(t-1), ..., u(t-n_u), \\ e(t-1), ..., e(t-n_e)) + e(t) \end{aligned} \tag{2}$$

where $e(t) \in R^m$ is zero-mean noise, and n_y, n_u, and n_e are non-negative integers. Given an approximation $\hat{f}$ to the function f, a one-step-ahead predictor $\hat{y}(t|\mathcal{D}_{t-1})$ can be formulated. The predictor and prediction error are defined by

$$\begin{aligned} \hat{y}(t|\mathcal{D}_{t-1}) &= \hat{f}(y(t-1), ..., y(t-n_y), u(t-1), ..., u(t-n_u), \varepsilon(t-1), ..., \varepsilon(t-n_e)) \\ \varepsilon(t) &= y(t) - \hat{y}(t|\mathcal{D}_{t-1}) \end{aligned}$$

The motivation behind this predictor is that while the noise sequence e is unknown, $\varepsilon(t) \rightarrow e(t)$ as $t \rightarrow \infty$, if $\hat{f} = f$ and the model is invertible.

Finally, we consider state-space models

$$x(t+1) = g(x(t), u(t)) + v(t) \tag{3}$$

$$y(t) = h(x(t)) + w(t) \tag{4}$$

where $x(t)$ is a state-vector, and $v(t)$ and $w(t)$ are zero-mean disturbance and noise vectors of appropriate dimensions. In this case, the model is defined by the functions g and h. Again, using approximations $\hat{g}$ and $\hat{h}$, it is possible to construct a one-step-ahead predictor $\hat{y}(t|\mathcal{D}_{t-1})$ using the extended Kalman-filter approach, e.g. [19]

$$\begin{aligned} \hat{x}(t|\mathcal{D}_{t-1}) &= \hat{g}(\hat{x}(t-1|\mathcal{D}_{t-2}), u(t-1)) + K(t-1)\varepsilon(t-1) \\ \hat{y}(t|\mathcal{D}_{t-1}) &= \hat{h}(\hat{x}(t|\mathcal{D}_{t-1})) \\ \varepsilon(t) &= y(t) - \hat{y}(t|\mathcal{D}_{t-1}) \end{aligned}$$

where $K(t)$ is the Kalman-filter gain matrix. This matrix will depend explicitly on the time, the functions $\hat{g}$ and $\hat{h}$, and the covariance matrices of the disturbance and noise sequences.

2.1 A Generalized Framework

In all these cases, we can write the model on the form

$$\eta(t) = f(\xi(t)) + e(t) \tag{5}$$

where $\eta(t) \in R^{\overline{m}}$ is a generalized output-vector, $\xi(t) \in R^{\overline{r}}$ is a generalized input-vector, and $e(t) \in R^{\overline{m}}$ is zero-mean noise. We denote the space $R^{\overline{r}}$ the input space. In the static model case (1), the input and output vectors equal the generalized input and output vectors. In the NARMAX case (2), the generalized input vector contains delayed input and output vectors in addition to delayed noise vectors, while the generalized output equals the system output. If noise terms $e(t-1), ..., e(t-n_e)$ are present, the generalized input vector is partially unknown and cannot be found exactly from the data $\mathcal{D}_t$. For state-space models (3)-(4), neither the generalized input nor the generalized output vectors can be found exactly, because they contain the unknown state vector. The purpose of the formulation (5) with the generalized input and output vectors is to write the model in a generic form with one unknown function f. The problem we address is to estimate this function, and since the function immediately gives the model equations, this also solves the system identification problem. Notice that the fact that the generalized inputs and outputs may not be exactly known does not complicate this problem too much, since the model parameters can still be estimated from the input/output data using a prediction error approach with the predictors described above [19].

2.2 Model Representation

In [30, 27, 13, 8, 9, 10] a non-linear model representation with good interpolation and extrapolation properties is described. It is based on the decomposition of the system's operating range into a number of smaller operating regimes, and the use of simple local models to describe the system within each regime. A global model is formed by interpolating the local models using smooth interpolation functions, that depend on the operating point.

We define the system's operating point at time t as $z(t) = (z_1(t), ..., z_d(t))^T \in Z = R^d$, where typically $d \leq \overline{r}$ and the operating space Z is a subspace or sub-manifold of the input space. It is assumed that ξ and z are related by a known bounded mapping H so that $z = H(\xi)$. Typically, Z and H are designed such that the operating point $z(t)$ characterizes different modes of behavior of the system under different operating conditions. The design of Z and H is discussed in more detail in section 4, and in [11]. Suppose Z is decomposed into N disjoint sets $\{Z_i\}_{i \in I_N}$ (regimes) so that

$$Z = \bigcup_{i \in I_N} Z_i$$

for some index set $I_N = \{i_1, ..., i_N\}$ with N elements. Assume that for each regime Z_i we have a local model structure defined by the function $\hat{f}_i(\xi; \theta_i)$ (parameterized by the vector θ_i) and a local model validity function $\rho_i(z) \geq 0$ which indicates the relative validity of the local model as a function of z. In addition to being smooth, ρ_i is designed to have the property that $\rho_i(z)$ is close to zero if $z \notin Z_i$. Furthermore, it is assumed that for all $z \in Z$ there exists an $i \in I_N$ so that $\rho_i(z) > 0$, to ensure completeness of the model. A global model can be formed as

$$\hat{f}(\xi) = \sum_{i \in I_N} \hat{f}_i(\xi; \theta_i) w_i(z) \tag{6}$$

$$w_i(z) = \rho_i(z) \Big/ \sum_{j \in I_N} \rho_j(z) \tag{7}$$

where the functions $\{w_i\}_{i \in I_N}$ are called interpolation functions. This representation is discussed in detail in [9], and it is shown that if the local model validity functions and operating space are adequately chosen, then any continuous function f can be uniformly approximated to an arbitrary accuracy on any compact subset of the input space using this representation. A model structure based on a decomposition into N regimes is written

$$\mathcal{M}_N = \left\{ \left(Z_i, \rho_i, \hat{f}_i \right) \right\}_{i \in I_N} \tag{8}$$

This is somewhat redundant, since there is a close (but not necessarily one-to-one) relationship between Z_i and ρ_i. With this representation, the modeling problem consists of the following subproblems:

1. Choose the variables with which to characterize the operating regimes, i.e. the operating space Z and mapping H.
2. Decompose Z into regimes, and choose local model structures.
3. Identify the local model parameters for all regimes.

In [9, 10, 11] it is demonstrated that in some cases, some coarse qualitative system understanding is sufficient to carry out this procedure. In the following sections we propose an algorithm that requires significantly less prior knowledge in order for us to decompose Z, choose local model structures, and construct interpolation functions.

2.3 Model Structure Identification Criteria

Let a model structure $\mathcal{M}$ of the form (8) be given. Notice that in a model structure, the model parameters $\theta^T = (\theta_{i_1}^T, ..., \theta_{i_N}^T)$ are considered unknown. The model structure $\mathcal{M}$ together with the admissible parameter set $\Theta_{\mathcal{M}}$ generate a model set $\{\mathcal{M}(\theta);\ \theta \in \Theta_{\mathcal{M}}\}$. In this section, we will discuss how different model structures can be compared using a sequence of empirical data to estimate their expected prediction performance. Let an unknown *future* data sequence be denoted $\mathcal{D}_t^\star$, and assume $\mathcal{D}_t^\star$ and $\mathcal{D}_l$ are uncorrelated. We introduce the notation

$$y(t) = y^\star(\mathcal{D}_{t-1}^\star) + e(t)$$
$$\varepsilon(t|\mathcal{M}, \theta) = y(t) - \hat{y}(t\,|\mathcal{D}_{t-1}^\star, \mathcal{M}, \theta)$$

where $y^\star(\mathcal{D}_{t-1}^\star)$ is the deterministic (predictable) component of the system output, $e(t)$ is the stochastic (unpredictable) component, and $\varepsilon(t|S,\theta)$ is the residual. Let $\hat{\theta}_{\mathcal{M}}$ be the parameter estimate that minimizes the prediction error criterion

$$J_{\mathcal{M}}(\theta) = \frac{1}{l}\sum_{t=1}^{l} \text{trace}\left(\varepsilon(t\,|\mathcal{M},\theta)\varepsilon^T(t\,|\mathcal{M},\theta)\right) \tag{9}$$

and let $E_{\mathcal{D}}$ and $E_{\mathcal{D}^\star}$ denote expectations with respect to $\mathcal{D}_l$ and $\mathcal{D}_t^\star$, respectively. The future prediction error is given by

$$\varepsilon^\star(t|\mathcal{M},\hat{\theta}_{\mathcal{M}}(\mathcal{D}_l)) = y^\star(\mathcal{D}_{t-1}^\star) - \hat{y}(t|\mathcal{D}_{t-1}^\star,\mathcal{M},\hat{\theta}_{\mathcal{M}}(\mathcal{D}_l)) + e(t)$$

where the dependence of $\mathcal{D}_l$ on $\hat{\theta}_{\mathcal{M}}$ has been written explicitly. The expected squared prediction error is defined by

$$\Sigma(\mathcal{M}) = E_{\mathcal{D}^\star}E_{\mathcal{D}}\left(\varepsilon^\star(t|\mathcal{M},\hat{\theta}_{\mathcal{M}}(\mathcal{D}_l))\right)\left(\varepsilon^\star(t|\mathcal{M},\hat{\theta}_{\mathcal{M}}(\mathcal{D}_l))\right)^T$$

Assuming $e^\star(t)$ is white noise that is uncorrelated with $\mathcal{D}_{t-1}^\star$ and $\mathcal{D}_l$, we get the following bias/variance decomposition of this expected squared prediction error

$$\begin{aligned}
\Sigma(\mathcal{M}) = E_{\mathcal{D}^\star}&\left(y^\star(\mathcal{D}_{t-1}^\star) - E_{\mathcal{D}}\hat{y}(t|\mathcal{D}_{t-1}^\star,\mathcal{M},\hat{\theta}_{\mathcal{M}}(\mathcal{D}_l))\right)\\
&\qquad\qquad\cdot\left(y^\star(\mathcal{D}_{t-1}^\star) - E_{\mathcal{D}}\hat{y}(t|\mathcal{D}_{t-1}^\star,\mathcal{M},\hat{\theta}_{\mathcal{M}}(\mathcal{D}_l))\right)^T\\
&+E_{\mathcal{D}^\star}E_{\mathcal{D}}\left(\hat{y}(t|\mathcal{D}_{t-1}^\star,\mathcal{M},\hat{\theta}_{\mathcal{M}}(\mathcal{D}_l)) - E_{\mathcal{D}}\hat{y}(t\left|\mathcal{D}_{t-1}^\star,\mathcal{M},\hat{\theta}_{\mathcal{M}}(\mathcal{D}_l)\right.)\right)\\
&\qquad\qquad\cdot\left(\hat{y}(t|\mathcal{D}_{t-1}^\star,\mathcal{M},\hat{\theta}_{\mathcal{M}}(\mathcal{D}_l)) - E_{\mathcal{D}}\hat{y}(t|\mathcal{D}_{t-1}^\star,\mathcal{M},\hat{\theta}_{\mathcal{M}}(\mathcal{D}_l))\right)^T\\
&+E_{\mathcal{D}^\star}\left(e(t)e^T(t)\right)
\end{aligned} \tag{10}$$

The first term is the squared systematic error (squared bias) caused by a too simple model structure. The second term is the squared random error (variance) that is present because the best model in the model set $\{\mathcal{M}(\theta);\ \theta \in \Theta_{\mathcal{M}}\}$ cannot in general be identified on the basis of the finite data sequence $\mathcal{D}_l$. Finally, the third term is the unpredictable component of the system output. Notice that the first term does not depend on the data $\mathcal{D}_l$, while the third term depends neither on the data $\mathcal{D}_l$, nor on the model structure $\mathcal{M}$. It is evident that a small bias requires a complex model structure, in general. On the other hand, a small variance requires a model structure that is simple, with few parameters compared to the number of observations l. The perfect model is characterized both by small bias and variance, and this appears to be impossible to achieve for a small l. This is known as the bias/variance dilemma, cf. Fig 1.

The model set will be based on a set of functions that can approximate any smooth function uniformly on a compact subset of the input space. This is obviously a desirable property of the model set, but also a cause for some problems. The richness implies that there will exist models in the model set that

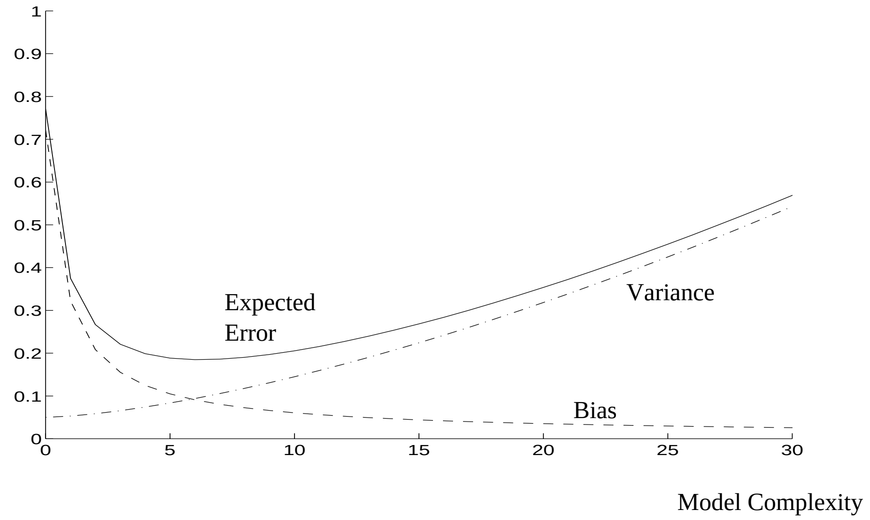

Fig. 1. Typical relationship between bias, variance and model structure complexity, when it is assumed that model structure complexity can be measured by a real number.

may make the bias arbitrarily small. However, the finite amount of data will give large variance for such models, and such a model will be fitted to represent not only the system, but also the particular realization of the noise. In other words, the model may give very good prediction of $\mathcal{D}_l$, but poor prediction capability when applied to $\mathcal{D}_t^\star$. This is known as over-fitting, and is caused by too many degrees of freedom in the model structure. It is therefore important that a model structure with an appropriate number of degrees of freedom is found, in the sense that it balances bias and variance. We will base the model structure identification algorithm on statistical criteria that have this property.

The mean square error (MSE) criterion is defined by

$$J_{MSE}(\mathcal{M}) = \mathrm{trace}(\Sigma(\mathcal{M}))$$

Minimizing J_{MSE} will lead to a parsimonious model structure, but with a finite sequence of data $\mathcal{D}_l$, the problem is ill-posed. The reason is simply that J_{MSE} cannot be computed since the probability distribution for the prediction error is unknown. An alternative would be to minimize the average squared residuals (ASR) criterion with respect to the model structure

$$J_{ASR}(\mathcal{M}) = \mathrm{trace}\left(\frac{1}{l}\sum_{t=1}^{l}\varepsilon(t|\mathcal{M},\hat{\theta}_{\mathcal{M}})\varepsilon(t|\mathcal{M},\hat{\theta}_{\mathcal{M}})^T\right)$$

For finite l, J_{ASR} may be a strongly biased estimate of J_{MSE}, since the prediction performance is measured using the same data as those to which the parameters are fitted, and the law of large numbers is not valid because this introduces a strong dependence. Hence, the use of J_{ASR} for structure identification will not, in general, lead to a parsimonious model. We will in the following present several criteria that are far better estimates of J_{MSE} than J_{ASR}.

If a separate data sequence $\mathcal{D}_l^\star$ (independent of $\mathcal{D}_l$) is known, an unbiased estimate of J_{MSE} can be found by computing the empirical average squared prediction error that results when the model fitted to the data $\mathcal{D}_l$ is used to predict the data $\mathcal{D}_l^\star$. This is the simplest and perhaps most reliable procedure, but suffers from the drawback that a significantly larger amount of data is required. Experiments and collection of data are major costs for many modeling problems. We therefore proceed with some alternatives that allow the data $\mathcal{D}_l$ to be reused in order to find good estimates of J_{MSE}. First we consider the final prediction error criterion (FPE) [1], given by

$$J_{FPE}(\mathcal{M}) = \frac{1 + p(\mathcal{M})/l}{1 - p(\mathcal{M})/l} \; J_{ASR}(\mathcal{M})$$

where $p(\mathcal{M})$ is the effective number of parameters (degrees of freedom) in the model structure. J_{FPE} is an estimate of J_{MSE}, and penalizes model complexity relative to the length of the available data sequence through the term $p(\mathcal{M})/l$. A major restriction is that the predictor is assumed to be linearly parameterized. A non-linear generalization is given in [17]. An alternative criterion can be formulated using cross-validation [28, 26]. The idea is to fit the parameters to different subsets of the data set, and test the prediction performance of the model structure on the remaining (presumed independent) data. Cross-validation may give a reasonable approximation to the use of independent data for selecting the model structure, at the cost of extra computations. The computational complexity can be considerably reduced if the predictor is a linear function of its parameters, or in general by using one of the approximate cross-validation criteria in [26] It is shown that the approximate criteria are asymptotically equivalent to FPE, as $l \to \infty$. Another approximation to cross-validation is the Generalized Cross Validation (GCV) criterion [4]

$$J_{GCV}(\mathcal{M}) = \frac{1}{(1 - p(\mathcal{M})/l)^2} \; J_{ASR}(\mathcal{M})$$

which is easily seen to be asymptotically equivalent to FPE, and also assumes linear parameterization of the predictor. Any one of these criteria can be applied with the structure identification algorithm we will present in the next section.

3 System Identification

Let a set of candidate local model structures $\mathcal{L} = \{L_1, L_2, ..., L_{N_L}\}$ be given. L_i is a parameterized function that defines a local model structure, cf. (8).

3.1 The Set of Model Structures Candidates

Assume the input- and output-samples in $\mathcal{D}_l$ are bounded. Then the system's operating range Z can be approximated by the d-dimensional box

$$Z_1 = \left[z_{1,1}^{\min}, z_{1,1}^{\max}\right] \times \cdots \times \left[z_{1,d}^{\min}, z_{1,d}^{\max}\right]$$

where $z(t) \in Z_1$ for all $t \in \{1, ..., l\}$, since H is a bounded mapping. Notice that the resulting model will extrapolate and can be applied for operating points outside Z_1. Next, we consider the problem of decomposing Z_1 into regimes.

Consider the possible decompositions of the set Z_1 into two disjoint subsets Z_{11} and Z_{12} with the property $Z_1 = Z_{11} \cup Z_{12}$. We restrict the possibilities by the constraint that the splitting boundary is a hyper-plane orthogonal to one of the natural basis-vectors of R^d, i.e.

$$Z_{11} = \{z \in Z_1 \mid z_{d_1} < \zeta_1\}$$
$$Z_{12} = \{z \in Z_1 \mid z_{d_1} \geq \zeta_1\}$$

for some dimension index $d_1 \in \{1, ..., d\}$ and splitting point $\zeta_1 \in \left[z_{1,d_1}^{\min}, z_{1,d_1}^{\max}\right]$. Local model validity functions for the two regimes are defined by the recursion

$$\rho_{11}(z) = \rho_1(z) b(z_{d_1} - \overline{z}_{11,d_1}; \lambda_{11})$$
$$\rho_{12}(z) = \rho_1(z) b(z_{d_1} - \overline{z}_{12,d_1}; \lambda_{12})$$

where $\overline{z}_{i,d_1} = 0.5\left(z_{i,d_1}^{\min} + z_{i,d_1}^{\max}\right)$ for $i \in \{11, 12\}$ is the center point of Z_i in the d_1-direction. The function $b(r;\lambda)$ is a scalar basis-function with scaling parameter λ, and the local model validity function associated with the regime Z_1 is $\rho_1(z) = 1$. The scaling parameters are chosen by considering the overlap between the local model validity functions. For $i \in \{11, 12\}$, we choose $\lambda_i = 0.5\gamma\left(z_{i,d_1}^{\max} - z_{i,d_1}^{\min}\right)$ where γ is a design parameter that typically takes on a value between 0.25 and 2.0. There will be almost no overlap when $\gamma = 0.25$, and large overlap when $\gamma = 2.0$. For each dimension index $d_1 \in \{1, ..., d\}$ we represent the interval $\left[z_{1,d_1}^{\min},\ z_{1,d_1}^{\max}\right]$ by a finite number of N_1 points uniformly covering the interval. Now $d_1, \zeta_1, L_{1,v}$, and $L_{2,w}$ define a new model structure, where the regime Z_1 is decomposed according to the dimension index d_1 at the point ζ_1, and the two local model structures are $L_{1,v}$ and $L_{2,w}$. Formally, the set of candidate model structures $\mathcal{S}_n$ with n regimes is given by

$$\mathcal{S}_1 = \{\{(Z_1, \rho_1, L_j)\};\ j \in \{1, 2, ..., N_L\}\}$$
$$\mathcal{S}_2 = \left\{\left\{\left(Z_{11}^i, \rho_{11}^i, L_j\right), \left(Z_{12}^i, \rho_{12}^i, L_k\right)\right\};\ i \in \{1, 2, ..., dN_1\},\ j, k \in \{1, 2, ..., N_L\}\right\}$$
$$\begin{aligned}\mathcal{S}_3 = \Big\{\big\{&\left(Z_{11}^i, \rho_{11}^i, L_j\right), \left(Z_{121}^m, \rho_{121}^m, L_k\right), \left(Z_{122}^m, \rho_{122}^m, L_n\right)\big\};\\ &i, m \in \{1, 2, ..., dN_1\},\ j, k, n \in \{1, 2, ..., N_L\}\Big\}\\ \cup \Big\{\big\{&\left(Z_{111}^m, \rho_{111}^m, L_k\right), \left(Z_{112}^m, \rho_{112}^m, L_n\right), \left(Z_{12}^i, \rho_{12}^i, L_j\right)\big\};\\ &i, m \in \{1, 2, ..., dN_1\},\ j, k, n \in \{1, 2, ..., N_L\}\Big\}\end{aligned}$$

$$\mathcal{S}_4 = \cdots$$

The model structure set is now $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3 \cup \cdots$. The model structure set is illustrated as a search tree in Fig. 2. Strictly speaking, the model structure set is not a tree, since different sequences of decompositions sometimes lead to the same model structure. However, we choose to represent it as a tree, for the sake of simplicity. Now the structure identification problem can be looked upon as a multi-step decomposition process, where at each step one regime from the previous step is decomposed into two sub-regimes. Such an approach will lead to a sequence of model structures $\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_n$ where the model structure $\mathcal{M}_{i+1}$ has more degrees of freedom than $\mathcal{M}_i$. Due to the normalization of the local model validity functions, the model set is usually not strictly hierarchical, in the sense that $\mathcal{M}_i$ cannot be exactly represented using $\mathcal{M}_{i+1}$. However, the increasing degrees of freedom define a hierarchical structure.

3.2 Basic Search Algorithm

The problem is now to search through the set $\mathcal{S}$ for the best possible model structure. The estimate of the parameters in the model structure $\mathcal{M}$ is defined by a prediction error criterion

$$\hat{\theta} = \arg\min_{\theta} J_{\mathcal{M}}(\theta) \tag{11}$$

where it has been assumed that the minimum exists. This can be ensured by restricting the parameters to a compact set. Now, the chosen structure identification criterion is written $J'(\mathcal{M})$. We define for a given $n \geq 1$

$$\mathcal{M}_n = \arg\min_{\mathcal{M} \in \mathcal{S}_n} J'(\mathcal{M}) \tag{12}$$

Consider the following extended horizon search algorithm, where the integer $n^\star \geq 1$ is called the search horizon.

Search Algorithm.

1. Start with the regime Z_1. Let $n = 1$.
2. At each step $n \geq 1$, find a sequence of decompositions $\mathcal{M}_n, \mathcal{M}_{n+1}, ..., \mathcal{M}_{n+n^\star}$ that solves the optimization problem
$$\min_{\mathcal{M} \in \mathcal{S}_{n+n^\star}} J'(\mathcal{M})$$
3. Restrict the search tree by keeping the decomposition that leads to $\mathcal{M}_{n+1}$ fixed for the future.
4. If
$$J'(\mathcal{M}_n) > \min_{k \in \{1,2,\ldots,n^\star\}} J'(\mathcal{M}_{n+k})$$
then increment n and go to 2. Otherwise, the model structure $\mathcal{M}_n$ is chosen.

Referring to Fig. 2, this algorithm will search the tree starting at the top (corresponding to one local model covering the whole operating space), and selecting a decomposition at each level through a sequence of "locally exhaustive" searches of depth $n^\star$. If $n^\star = 1$, this is a local search algorithm.

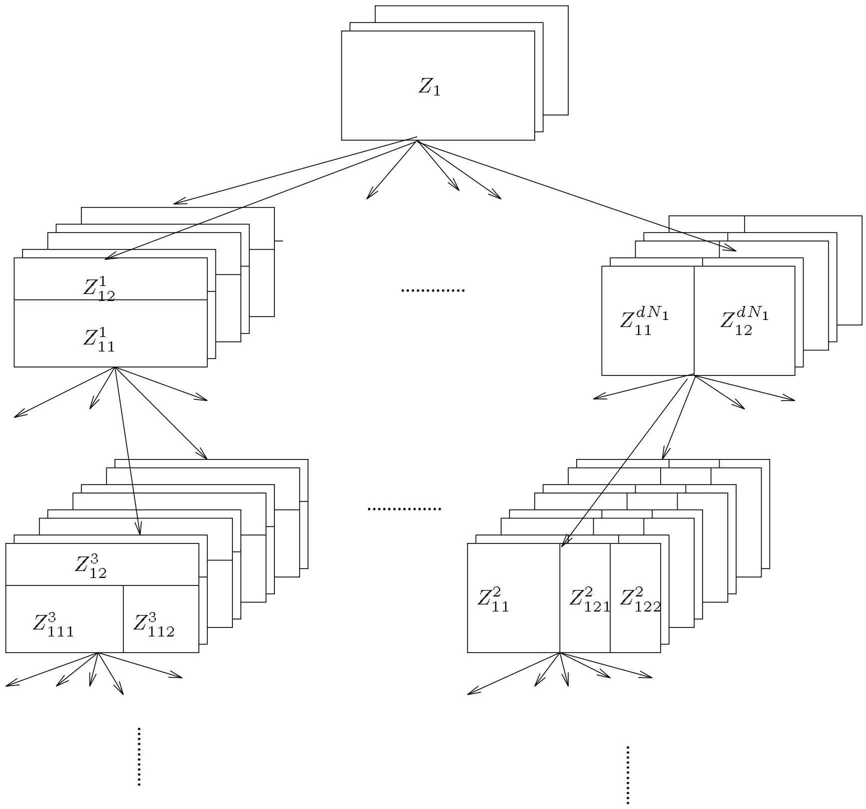

Fig. 2. Model structure search tree illustrating possible decompositions into regimes and choice of local model structures. Each level in the tree corresponds to the possible decompositions into one more regime than at the previous level, i.e. the model structure sub-sets $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \ldots$ The sub-set of model structures at each "super-node" in the tree corresponds to a fixed decomposition into regimes, but to different combinations of local model structures. The suggested algorithm will search this tree starting at the top (corresponding to one local model covering the whole operating space), and selecting a decomposition at each level through a sequence of "locally exhaustive" searches of depth $n^\star$.

3.3 Heuristic Search Algorithm

Clearly, the performance of the algorithm is expected to improve as $n^\star$ increases, but the computational complexity makes $n^\star > 3$ not feasible for any practical problem with 1994 desktop computer technology, even if the local model structure set $\mathcal{L}$ contains as few as one or two possibilities.

Example. Consider the problem of identifying a state-space model of the form $x(t+1) = f(x(t))$, where $\dim(x) = 5$, and we apply local models of the form

$$\begin{pmatrix} x_1(t+1) \\ x_2(t+1) \\ \vdots \\ x_5(t+1) \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_5 \end{pmatrix} + \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{15} \\ a_{21} & a_{22} & & a_{25} \\ \vdots & & \ddots & \vdots \\ a_{51} & a_{52} & \cdots & a_{55} \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_5(t) \end{pmatrix}$$

Each of the 30 parameters can be replaced by a structural zero, which gives a set of local linear model structure with $2^{30} \approx 10^9$ elements. On the other hand, even if there is only one possible local model structure that one can choose, the number of possible decompositions into no more than five regimes is

$$\#\mathcal{S}_1 + ... + \#\mathcal{S}_5 = 1 + dN_1 + 2(dN_1)^2 + 3!(dN_1)^3 + 4!(dN_1)^4$$

For $d = 2$ and $N_1 = 10$, this is approximately $4 \cdot 10^6$ candidate decompositions. With $n^\star < 4$, the model structure set is considerably reduced. In particular, $n^\star = 3$ gives about 10^5 decompositions among which to search, $n^\star = 2$ gives about 2500, while $n^\star = 1$ gives 80 candidate decompositions.

□

Because of the combinatorial nature of the model structure set, it is clearly of interest to implement some heuristics that cut down on the computational complexity without sacrifying too much of the optimality of the algorithm. As we have seen, the number of candidate decompositions at each step in the search may be large. To reduce the number of candidates, we suggest applying the following heuristics in the "locally exhaustive" search at the second step in the search algorithm:

Heuristic 1. *At each level in the search tree, proceed with only the most promising candidates.*

The best candidates are of course not known a priori, so there is always a possibility that this may lead to a sub-optimal model. We suggest proceeding with the best decomposition for each of the possible splitting dimensions. Instead of trying to find the best candidates, one can often more easily single out the "least promising candidate decompositions":

Heuristic 2. *Discard the candidate decompositions that give an increase in the criterion from one level in the search tree to the next.*

The number of remaining candidates will typically be larger than when using Heuristic 1, but the chance of discarding the optimal decomposition may be smaller. Some candidate decompositions may give rise to regimes where no substantial amount of data is available, and may therefore be classified a priori as not feasible:

Heuristic 3. *Discard candidate decompositions that lead to regimes with few data points relevant to this regime compared to the number of degrees of freedom in the corresponding local model structure and local model validity function.*

Counting the number of relevant data-points associated with each regime is controversial, since the interpolation functions overlap. We use the heuristic count $l_i = \sum_{t=1}^{l} w_i(z(t))$, which has the attractive property $\sum_{i \in I_N} l_i = l$.

Heuristic 4. *Use a (backward or forward) stepwise regression procedure to handle local model structure sets $\mathcal{L}$ of combinatorial nature [29].*

Related to the example above, one should start with no structural zeroes, and then add one structural zero at a time, choosing the one that gives the largest improvement in the prediction performance. This should give less than $30 + 29 + 28 + ... + 1 < 30^2$ candidate model structures, which is quite different from 2^{30}.

3.4 User Choices

The basis-function $b(r; \lambda)$ with scaling parameter λ has the purpose of providing a smooth interpolation between the local models. The basis-function is assumed to have the property $b(r; \lambda) \geq 0$ for all $r \in R$ and $b(r, \lambda) \to 0$ as $|r| \to \infty$. Typical choices are kernel-functions, like the unnormalized Gaussian $\exp(-r^2/2\lambda^2)$. It may appear that the choice of this function has significant impact on the model. However, it is our experience that the algorithm and model's prediction performance are quite insensitive with respect to this choice, and the specification of this function does not require any prior knowledge about the system. What is more important is the choice of λ, which is controlled by the user-specified parameter γ.

In order to compute the criterion J_{GCV} or J_{FPE}, the effective number of parameters p in the model structure must be known. If the choice of model structure is not based on the data, then the effective number of parameters is

$$p = \sum_{i \in I_N} \dim(\theta_i) \tag{13}$$

in the case of linear regression. However, the proposed algorithm for model structure identification makes use of the data $\mathcal{D}_l$ during the search. Hence, the p-value given by (13) will be too small. Counting the effective number of parameters in this case is controversial. We apply the heuristic

$$p = \kappa(N-1) + \sum_{i \in I_N} \dim(\theta_i)$$

where $\kappa \geq 0$ is a heuristic constant, which can be interpreted as a smoothing parameter, since a large κ will put a large penalty on model complexity, and will therefore give a smooth model. A typical choice of κ is between 0 and 4, cf. [6].

3.5 Statistical Properties

Consider the bias/variance decomposition (10). It has been shown in [12] that both the bias and variance will asymptotically (as $l \to \infty$) tend towards their smallest possible values, with probability one, provided

1. The parameter estimator is consistent, see [18] for conditions under which this holds.

2. The estimate J' of the expected squared prediction error used for model structure identification is consistent.
3. Global minima of the parameter and structure optimization problems are found with probability one.
4. The model set $\{\mathcal{M}(\theta); \mathcal{M} \in \mathcal{S}, \theta \in \Theta_{\mathcal{M}}\}$ can be covered by a finite ϵ-net.

It is known that the use of a separate validation data sequence for model structure identification gives a J' that satisfies the second requirement [12]. It is also known that FPE and AIC may be slightly biased [23].

Neither the parameter optimization nor the structure optimization algorithms need result in global minima. An attractive feature of the model structure set is that it appears to have not only multiple global minima, but also many close-to-optimal local minima. It is easy to see that the restriction of the search to any sub-tree of the model structure tree does not exclude any possible decompositions into regimes. The worst thing that can happen is that the number of decompositions may be somewhat larger than necessary, or alternatively that the partition may not be as fine as desired. Obviously, this leads to suboptimality for finite amount of data, but not necessarily so asymptotically.

The fourth condition is somewhat technical, but it does in general impose a restriction on the complexity of the model set. In practise, this is not a serious restriction, as discussed in [12].

4 Examples

4.1 A Simulated Fermentation Reactor

Consider the fermentation of glucose to gluconic acid by the micro-organism *Pseudomonas ovalis* in a well stirred batch reactor. The main overall reaction mechanism is described by

$$\begin{aligned}
\text{Cells} + \text{Glucose} + O_2 &\rightarrow \text{More Cells} \\
\text{Glucose} + O_2 &\xrightarrow{\text{Cells}} \text{Gluconolactone} \\
\text{Gluconolactone} + H_2O &\rightarrow \text{Gluconic Acid}
\end{aligned}$$

The production of gluconolactone is enzyme-catalysed by the cells. We use the following state-space model to simulate the "true system" [7]:

$$\begin{aligned}
\dot{\chi} &= \mu_m \frac{sc}{k_s c + k_o s + sc}\chi \\
\dot{p} &= k_p l \\
\dot{l} &= v_l \frac{s}{k_l + s}\chi - 0.91 k_p l \\
\dot{s} &= -\frac{1}{Y_s}\mu_m \frac{sc}{k_s c + k_o s + sc}\chi - 1.011 v_l \frac{s}{k_l + s}\chi \\
\dot{c} &= k_l a(c^\star - c) - \frac{1}{Y_o}\mu_m \frac{sc}{k_s c + k_o s + sc}\chi - 0.09 v_l \frac{s}{k_l + s}\chi
\end{aligned}$$

Table 1. Symbols and constants in the state-space simulation model for the fermenter.

Symbol	Description
χ	Cell concentration $[UOD/ml]$
p	Gluconic acid concentration $[g/l]$
l	Gluconolactone concentration $[g/l]$
s	Glucose concentration $[g/l]$
c	Dissolved oxygen concentration $[g/l]$
μ_m	$0.39\ h^{-1}$
k_s	$2.50\ g/l$
k_o	$0.55 \cdot 10^{-3}\ g/l$
k_p	$0.645\ h^{-1}$
v_l	$8.30 mg\ UOD^{-1}\ h^{-1}$
k_l	$12.80\ g/l$
$k_l a$	$150.0 - 200.0\ h^{-1}$
Y_s	$0.375\ UOD/mg$
Y_o	$0.890\ UOD/mg$
$c^\star$	$6.85 \cdot 10^{-3}\ g/l$

with initial conditions $\chi(0) = \chi_0$, $p(0) = 0$, $l(0) = 0$, $s(0) = s_0$ and $c(0) = c^\star$. The symbols and constants are defined in Table 1.

We simulated 10 hour batches using these equations, "measuring" all states with 0.5 hour intervals, and adding sequentially uncorrelated random noise with a signal-to-noise ratio of approximately 30 dB to the states. We collected two sets of data, by randomly varying the initial conditions χ_0, s_0 and the agitation speed (affecting $k_l a$). The first set contains data from 100 batches, and is used for system identification, while the second independent set is used for model testing. We define the following dimension-less normalized variables: $\chi_n = \chi/(3\ \mathrm{UOD}/ml)$, $p_n = p/(50\ g/l)$, $l_n = l/(15\ g/l)$, $s_n = s/(50\ g/l)$, $c_n = c/(0.01\ g/l)$, and the normalized state-vector $x = (\chi_n, p_n, l_n, s_n, c_n)^T$. We have specified only one possible local linear discrete-time state-space model structure $x(t+1) = a_i + A_i x(t)$ with 12 structural zeros in the A_i matrices. These zeros follow directly from the reaction mechanism, see also [10]. We observe that glucose and oxygen are rate-limiting and, consequently, expected to be the main contributors to the system's non-linearities. It follows that the operating point $z = (s_n,\ c_n)^T$ captures these non-linearities and characterizes the operating conditions of the process with respect to local linear models, see also [10].

Running the identification algorithm with $n^\star = 1$, using the FPE criterion with $\kappa = 1$, and Gaussian basis-function with $\gamma = 1$, results in a model with five local models, and root average squared one-step-ahead prediction error (PE) on the test data PE=0.0139, cf. Table 2. Restricting the number of local models to three, gives PE=0.0147, while one global linear model gives PE=0.0303. This clearly indicates that there exist significant non-linearities which have been captured by the two more complex models, and not by the linear model. The five regimes are illustrated in Fig. 3. Perhaps the most interesting and attractive

Table 2. Root average squared prediction performance for the model based on five operating regimes.

State	One-step-ahead prediction	Ballistic prediction
χ	0.0134	0.0341
p	0.0123	0.0211
l	0.0157	0.0357
s	0.0140	0.0204
c	0.0137	0.0315
Total	0.0139	0.0293

feature of the method is that the identified model can be interpreted in a natural way. The five regimes correspond to the following phases in the batch

1. Initial phase.
2. Growth phase, where only the amount of micro-organisms is limiting the rate of the reactions.
3. Oxygen supply is rate-limiting.
4. Glucose is rate-limiting.
5. No glucose left, termination.

This gives a high-level qualitative description of the system. More low-level quantitative details on e.g. reaction kinetics can be added by examining the parameters of the local models corresponding to each regime. A simulation (ballistic prediction) of a typical batch from the test set is shown in Fig. 4, using the model with five local models, and the identified global linear model for comparison.

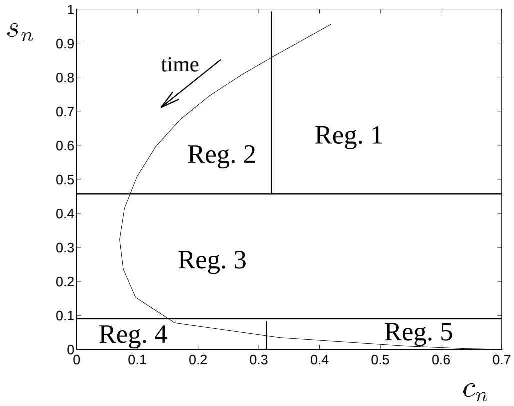

Fig. 3. The decomposition into five regimes using the simulated fermenter data, with a typical simulated system trajectory projected onto the (c_n, s_n)-plane.

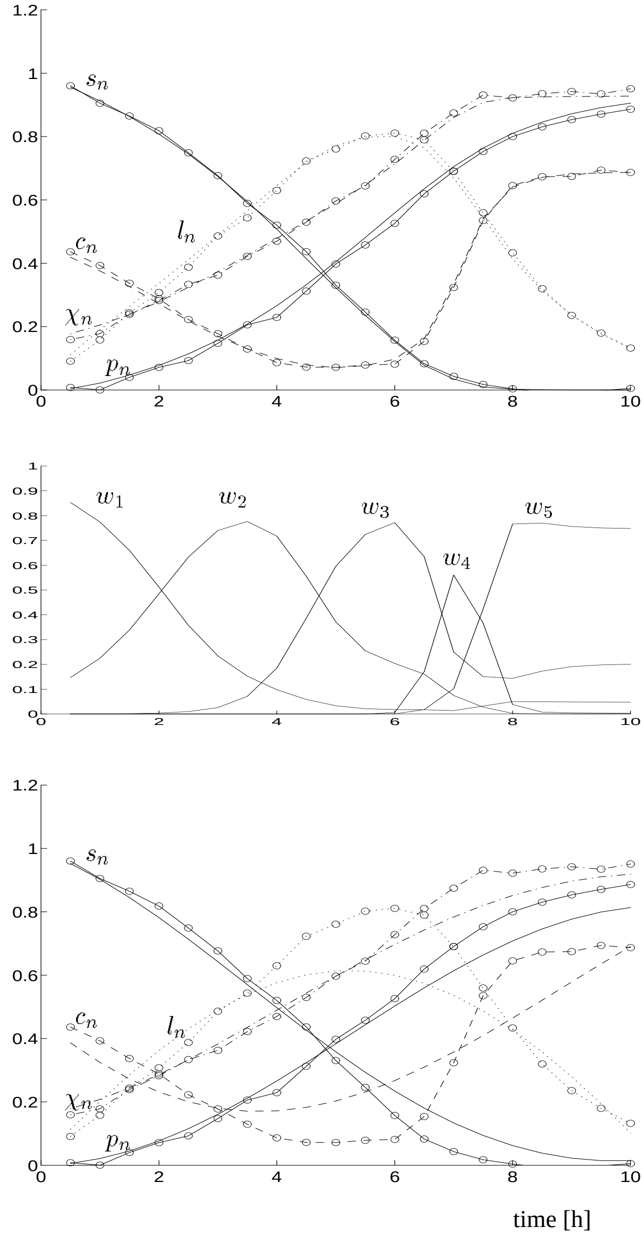

Fig. 4. Top: Trajectories with circles are generated by the simulated "true system", while the other trajectories are simulations of the model based on five local models. Middle: The relative weight of the various local models in the interpolation. Bottom: Simulation of the identified global linear model.

Clearly, the results favor the non-linear model. Comparison with similar models with three or four regimes, designed by hand on the basis of system knowledge [10] indicates that their performance are comparable. Although semi-empirical, the high transparency suggests that the identified model should not be viewed as a black-box.

The applied a priori knowledge is essentially the overall reaction mechanism, used for structuring the A_i matrices and for selection of the two variables used for characterizing the operating space. It must be noted that the algorithm has also been applied without this knowledge, i.e. with full A_i matrices and operating point $z = x$, resulting in only a slight decrease in the prediction accuracy of the identified model. In this case, it is interesting to observe that among the five components of the operating point, the algorithm chooses only χ and c for decomposition into regimes. As noted in [10], due to the batch nature of the process operation, the information in the state χ is highly redundant, given the information in s. This is also evident from Fig. 4, which clearly shows the collinearity between these variables. Hence, the algorithm is forced to make a somewhat arbitrary choice about which variable to use for decomposition, a fact that makes the interpretation of the model more difficult. This problem is also observed with the MARS algorithm [5]. This only emphasizes the important fact that the success of empirical modeling is heavily dependent on the information in the empirical data, and that data deficiencies should to the highest possible extent be compensated for using prior knowledge.

4.2 Modeling of a Hydraulic Manipulator

A data sequence $\mathcal{D}_{8000}$ logged from a hydraulic TR4000 robot [15] from ABB Trallfa Robotics A/S was used to find a model describing the inverse dynamics

$$\tau(t) = f(q(t), \dot{q}(t), \ddot{q}(t)) + e(t)$$

of a joint of this robot, where $\tau(t)$ is the control signal to the servo valve, $q(t)$ is joint position, and $e(t)$ is equation error. The joint position was logged at a sampling rate of 100 Hz while the robot was moving along a randomly generated trajectory. The joint velocity and acceleration was estimated by low-pass filtering and numerical differentiations. The prediction of an estimated linear model was subtracted from the data in order to emphasize the non-linearities. According to [15], the non-linearities are mainly due to variations in the momentum arm of the hydraulic cylinder, non-linear damping, and non-linear pressure gain characteristics due to varying flow-rates in the servo valve. In addition, 1000 independent samples were used for testing the model. A number of models were identified, based on the structure identification algorithm, a least squares parameter estimation algorithm, local linear model structure, and a Gaussian basis-function with $\gamma = 1$. The results are summarized and compared to the results in [16] (marked ‡), [3] (marked $^{+}$), and the MARS algorithm [6] in Table 3. The table shows that the structure identification algorithm is able to find an adequate model with a small number of parameters while maintaining the high

Table 3. Results of applying various identification algorithms on the hydraulic manipulator joint data. The NRMSE criterion is defined as the square root of the ratio of the average squared one-step-ahead prediction error to the variance of the output, using the independent test data.

Model	Comments	Num. Param.	NRMSE
ASMOD[‡]	Quadratic Spline Basis	561	15 %
MARS		155	16 %
Local linear	$\gamma = 1$, $n^\star = 2$, Heuristics 1, 2, and 3	80	17 %
Local linear	$\gamma = 1$, $n^\star = 1$, Heuristic 3	80	17 %
MARS		53	17 %
Local linear	$\gamma = 1, n^\star = 2$, Heuristics 1, 2 and 3	40	18 %
Local linear	$\gamma = 1, n^\star = 1$, Heuristic 3	40	19 %
RBF[+]	Gaussian radial basis-functions	112	19 %
MARS		16	20 %
ASMOD[‡]	Quadratic Spline Basis	48	20 %
Local linear	$\gamma = 1, n^\star = 2$, Heuristics 1, 2 and 3	20	23 %
NN[+]	Sigmoidal Neural Network (3-20-1)	101	23 %
Local linear	$\gamma = 1, n^\star = 1$, Heuristic 3	20	26 %
NN[+]	Sigmoidal Neural Network (3-5-1)	26	26 %

accuracy of the models found by MARS and ASMOD. In all cases, only the parameters corresponding to local model parameters or basis-function coefficients are counted in the table. Notice that according to the FPE criterion with $\kappa = 1$, the data sequence allows more degrees of freedom to be added to the identified model structures. This was not pursued due to the computational complexity involved. The operating point was chosen as $z = (q, \dot{q}, \ddot{q})$ although complementary identification experiments showed that $z = (\dot{q}, \ddot{q})$ was sufficient to capture most of the non-linearities, in particular when using no more than 40 parameters. Unfortunately, the large number of regimes makes the interpretation of the empirical model more difficult than in the previous example. The model should be viewed as a black box. This example mainly serves as a benchmark that shows that the accuracy achieved with the local modeling approach is comparable to some of the most popular empirical modeling algorithms from the literature.

5 Discussion

The amount of prior knowledge required with the proposed approach is quite reasonable. First of all, an operating point space Z is required. In many cases, it is possible to choose Z equal to a subspace or sub-manifold of the input space [9]. The design of Z need not be based solely on a priori knowledge, but can in addition consider the distribution of the data $\mathcal{D}_l$. Quite often, there are collinearities or correlations in the data, so that $\mathcal{D}_l$ can be embedded in a subspace or sub-manifold of considerably lower dimension than the input space. In that case, z need not be of higher dimension than this embedding. Some prior knowledge will often make it possible to reduce the dimension of z considerably. This is important, since it may both reduce the complexity of the model and improve its transparency, and also reduce the computational complexity for the empirical modeling algorithm considerably.

A set of local model structure candidates must be specified. If no a priori knowledge exists to support one choice over the other, one will typically choose local linear model structures of various orders and possibly with structural zeros as default, since linear models are well understood and possible to interpret. Moreover, a linear model will always be a sufficiently good approximation locally, provided the system is smooth, and the regimes are small enough. On the other hand, if there is substantial a priori knowledge available in terms of mechanistic local model structures, these can be included as illustrated in [11]. Such local model structures may for example be simplified mass- and energy-balances.

The purposes of a model can be diverse, e.g. system analysis, design, optimization, prediction, control, or diagnosis. In many applications, it is important that the model can be easily interpreted and understood in terms of the system mechanisms. With empirical models, which are often based on black-box model representations, this is often a hard or impossible task. However, the approach presented in this paper gives highly transparent empirical models because

- local models are simple enough to be interpreted,
- the operating regimes constitue a qualitative high-level description of the system that is close to engineering thinking.

Notice that the interpretation of local linear models as linearizations of the system at various operating points is valid only if the model validity functions do not overlap too much. This point is stressed in [20], and the use of local identification algorithms for each local model is one possibility that will improve the interpretability.

5.1 Related work

Local linear models are applied in [13] and [27] together with a clustering algorithm to determine the location of the local models. A parameterized regime description and a hierarchical estimator used to estimate the regime parameters simultaneously with the local model parameters is described in [8] and [14]. An algorithm based on local linear models and decomposition of regimes in which the system appears to be more complex than the model, is suggested in [21]. A model representation based on local polynomial models and smooth interpolation is proposed in [22]. The structure identification algorithm is based on an orthogonal regression algorithm that sequentially discards local model terms that are found to have small significance.

When the interpolation functions are chosen as the characteristic functions of the regime-sets Z_i, a piecewise linear model results. The resulting model will not be smooth, and may not even be continuous, which may be a requirement in some applications. Also, we have experienced that smooth interpolation between local linear models usually gives better model fit compared to a piecewise linear model with the same number of parameters. The local linear modeling approach combined with a fuzzy set representation of the regimes also leads to a model representation with interpolation between the local linear models [30]. In that case, it is the fuzzy inference mechanism that implicitly gives an interpolation. Structure identification algorithms based on clustering [2, 32, 31] and local search [29] have been proposed in this context. In this case, the ρ_i-functions are interpreted as membership functions for fuzzy sets.

The algorithm of Sugeno and Kang [29] is perhaps the closest relative to the present algorithm. The main difference is the extra flexibility, and effort is begin applied to find a closer to optimal model with the present algorithm. A statistical pattern recognition approach with multiple models leads to a similar representation based on a piecewise linear model and discriminant functions to represent the regime boundaries [24]. Finally, in [25] it is suggested a model representation with neural nets as local models and a structure identification algorithm based on pattern recognition. The pattern recognition algorithm will detect parts of the input space in which the model fit is inadequate, and refine the model locally.

With this large body of literature in mind, one may ask: What are the contributions and improvements represented by the present approach? We have attempted to take the most attractive features from the algorithms in the literature and combined these into one algorithm. We have emphasized interpretability of the resulting model, flexibility with respect to incorporation of prior knowledge, and a transparent modeling and identification process that is close to engineering thinking. The price we have to pay is a computer intensive algorithm. Some may also argue that the algorithm is too flexible and not completely automatic, and as a result it may be difficult to apply for inexperienced users. However, it is our view that real world applications require perhaps even more flexibility and a less automated approach.

6 Concluding Remarks

The proposed empirical modeling and identification algorithm is based on a rich non-linear model representation which utilizes local models and interpolation to represent a global model. With this representation, the semi-empirical modeling problem is solved using a structure identification algorithm based on a heuristic search for decompositions of the system's operating range into operating regimes. This algorithm is the main contribution of this work. We want to emphasize an important property of the modeling method and identification algorithm, namely the transparency of the resulting model. The transparency is linked with the possibility to interpret each of the simple local models independently, but more importantly with the fact that the identified regimes can often be interpreted in terms of the system behavior or mechanisms.

In general, the fundamental assumptions behind empirical modeling are that 1) the empirical data is not too contaminated by noise and other unmodeled phenomena, and 2) that the data set is complete in the sense that it contains a sufficient amount of information from all interesting operating conditions and system variables. Unfortunately, these assumptions are often not met in practical applications. The proposed algorithm should therefore be applied with care, and as a part of a computer aided modeling environment that allows flexible incorporation of prior knowledge, not as an automatic modeling algorithm. Moreover, one should undertake a study of the robustness of the algorithm with respect to contaminated, sparse, and incomplete data, in particular for high dimensional and otherwise complex modeling problems.

Acknowledgments

This work was partly supported by The Research Council of Norway under doctoral scholarship grant no. ST. 10.12.221718 given to the first author. We want to thank Dr. Tom Kavli and several of his colleagues at SINTEF-SI for valuable discussions and making the hydraulic manipulator data available.

References

1. H. Akaike. Fitting autoregressive models for prediction. *Ann. Inst. Stat. Math.*, 21:243–247, 1969.
2. J. C. Bezdek, C. Coray, R. Gunderson, and J. Watson. Detection and characterization of cluster substructure. II. Fuzzy c-varieties and complex combinations thereof. *SIAM J. Applied Mathematics*, 40:352–372, 1981.
3. M. Carlin, T. Kavli, and B. Lillekjendlie. A comparison of four methods for nonlinear data modeling. *Chemometrics and Int. Lab. Sys.*, 23:163–178, 1994.
4. P. Craven and G. Wahba. Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerical Math.*, 31:317–403, 1979.
5. R. D. De Veaux, D. C. Psichogios, and L. H. Ungar. A tale of two nonparametric estimation schemes: MARS and neural networks. In *Proc. 4th Int. Conf. Artificial Intelligence and Statistics*, 1993.
6. J. H. Friedman. Multivariable adaptive regression splines (with discussion). *The Annals of Statistics*, 19:1–141, 1991.
7. T. K. Ghose and P. Ghosh. Kinetic analysis of gluconic acid production by pseudomonas ovalis. *J. Applied Chemical Biotechnology*, 26:768–777, 1976.
8. R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.
9. T. A. Johansen and B. A. Foss. Constructing NARMAX models using ARMAX models. *Int. J. Control*, 58:1125–1153, 1993.
10. T. A. Johansen and B. A. Foss. State-space modeling using operating regime decomposition and local models. In *Preprints 12th IFAC World Congress, Sydney, Australia*, volume 1, pages 431–434, 1993.
11. T. A. Johansen and B. A. Foss. A dynamic modeling framework based on local models and interpolation – combining empirical and mechanistic knowledge and data. Submitted to Computers and Chemical Engineering, 1994.

12. T. A. Johansen and E. Weyer. Model structure identification using separate validation data - asymptotic properties. Submitted to the European Control Conference, Rome, 1995.
13. R. D. Jones and co-workers. Nonlinear adaptive networks: A little theory, a few applications. Technical Report 91-273, Los Alamos National Lab., NM, 1991.
14. M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. Technical Report 9301, MIT Computational Cognitive Science, 1993.
15. T. Kavli. Nonuniformly partitioned piecewise linear representation of continuous learned mappings. In *Proceedings of IEEE Int. Workshop on Intelligent Motion Control, Istanbul*, pages 115–122, 1990.
16. T. Kavli. ASMOD – An algorithm for adaptive spline modelling of observation data. *Int. J. Control*, 58:947–967, 1993.
17. J. Larsen. A generalization error estimate for nonlinear systems. In *Proc. IEEE Workshop on Neural Networks for Signal Processing, Piscataway, NJ*, pages 29–38, 1992.
18. L. Ljung. Convergence analysis of parametric identification methods. *IEEE Trans. Automatic Control*, 23:770–783, 1978.
19. L. Ljung. *System Identification: Theory for the User.* Prentice-Hall, Inc., Englewood Cliffs, NJ., 1987.
20. R. Murray-Smith. Local model networks and local learning. In *Fuzzy Duisburg*, pages 404–409, 1994.
21. R. Murray-Smith and H. Gollee. A constructive learning algorithm for local model networks. In *Proceedings of the IEEE Workshop on Computer-Intensive Methods in Control and Signal Processing, Prague, Czech Republic*, pages 21–29, 1994.
22. M. Pottmann, H. Unbehauen, and D. E. Seborg. Application of a general multimodel approach for identification of highly non-linear processes – A case study. *Int. J. Control*, 57:97–120, 1993.
23. R. Shibata. Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrica*, 63:117–126, 1976.
24. A. Skeppstedt, L. Ljung, and M. Millnert. Construction of composite models from observed data. *Int. J. Control*, 55:141–152, 1992.
25. E. Sørheim. A combined network architecture using ART2 and back propagation for adaptive estimation of dynamical processes. *Modeling, Identification and Control*, 11:191–199, 1990.
26. P. Stoica, P. Eykhoff, P. Janssen, and T. Söderström. Model-structure selection by cross-validation. *Int. J. Control*, 43:1841–1878, 1986.
27. K. Stokbro, J. A. Hertz, and D. K. Umberger. Exploiting neurons with localized receptive fields to learn chaos. *J. Complex Systems*, 4:603, 1990.
28. M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Royal Statistical Soc. B*, 36:111–133, 1974.
29. M. Sugeno and G. T. Kang. Structure identification of fuzzy model. *Fuzzy Sets and Systems*, 26:15–33, 1988.
30. T. Takagi and M. Sugeno. Fuzzy identification of systems and its application to modeling and control. *IEEE Trans. Systems, Man, and Cybernetics*, 15:116–132, 1985.
31. R. R. Yager and D. P. Filev. Unified structure and parameter identification of fuzzy models. *IEEE Trans. Systems, Man, and Cybernetics*, 23:1198–1205, 1993.
32. Y. Yoshinari, W. Pedrycz, and K. Hirota. Construction of fuzzy models through clustering techniques. *Fuzzy Sets and Systems*, 54:157–165, 1993.