Identification of Non-linear Systems using Empirical Data and Prior Knowledge – An Optimization Approach

Tor A. Johansen[1]

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

Abstract

The choice of a parametric model structure in empirical and semi-empirical non-linear modeling is usually viewed as an important and critical step. However, it is known that by augmenting the least-squares identification criterion with a term that imposes penalty on the non-smoothness of the model, an optimal non-parametric model can be found explicitly. The optimal non-parametric model will depend on the particular form of the penalty, which can be looked upon as a priori knowledge, or the desired properties of the model. In this paper these results are extended in several directions, i) we show how useful types of prior knowledge other than smoothness can be included as a term in the criterion or as a constraint, and how this influences the optimal model, ii) dynamic models and a general prediction error penalty are considered, iii) we present a practical numerical procedure for the identification of a close to optimal semi-parametric model. The numerical approach is motivated by the difficulty of deriving the optimal non-parametric model if there are complicated constraints or penalty terms in the criterion. Finally, iv) we discuss determination of the appropriate model complexity through selecting weights on the different penalties on the basis of empirical data. Since the optimal non-parametric model is identified using an augmented optimization criterion, and all prior knowledge and desired model properties may be specified through the augmented optimization criterion, it is the choice of this criterion that is critical in this approach. This elevates the empirical and semi-empirical modeling problems to a more transparent and engineering-friendly level than is achieved by directly specifying a parametric model structure. Hence, our results shed some more light on modern non-linear empirical modeling approaches like radial basis-functions, splines and neural networks. In addition to providing a fundamental insight into the role of different kinds of prior knowledge, this high-level formulation is also attractive from a practical point of view, as it lies closer to engineering thinking, and less guesswork is required. The flexibility and power of this approach is illustrated with a semi-realistic simulation example.

[1] Present address: SINTEF Automatic Control, N-7034 Trondheim, Norway. Email: Tor.Arne.Johansen@regtek.sintef.no.

1 Introduction

Determining a model from a finite sample of observations without any prior knowledge about the system is an ill-posed problem, in the sense that a unique model may not exist, or it may not depend continuously on the observations (Tikhonov & Arsenin 1977). Indeed, without any prior knowledge except the observations, nothing is known about the system behavior in states between the observations. Hence, even if the data are not corrupted by noise, the model can behave arbitrarily at an operating point not exactly captured by the observations, and it should be clear that such a model will not be unique, nor will it depend continuously on the data. The problem is essentially that the finite amount of data does not impose sufficient constraints on the model set. Fortunately, a minimum of prior knowledge will in general provide the necessary constraints, and there exists a theory that will aid reformulation of the problem into a well-posed one, namely regularization theory (Tikhonov & Arsenin 1977). The prior knowledge or assumption that is often used, is that the system has certain smoothness properties. This will constrain the system behavior in a neighborhood of each observation from changing abruptly, and some interpolation and extrapolation of the observations can be justified. Such an assumption is reasonable for a large class of real world systems, but certainly not for all. The smoothness assumption can be incorporated explicitly by the use of a parametric or non-parametric model, and a penalty on non-smoothness in the identification criterion (Tikhonov & Arsenin 1977, O'Sullivan 1986, Wahba 1990). The penalty will reduce the effective number of parameters (degrees of freedom) in the model to an amount less than the number of parameters being fitted. However, the smoothness assumption is usually made implicitly, through the choice of a parameterized model structure that is simple in the sense that it has few unknown parameters compared to the number of observations. In this case, the dimension of the model space is reduced a priori such that the parameter identification problem is a well-posed one.

It is usually desirable to effectively employ all the available prior knowledge and empirical data. Increased amounts of prior knowledge will make the identification problem better conditioned, since more constraints are imposed on the model set. If the prior knowledge is correct, this will in general lead to a better model that is robust against a deficient or incomplete data set. Herein lies the motivation for our work, and we will show that the regularization framework is a convenient tool for combining prior knowledge and empirical data. In particular, we will allow prior knowledge, empirical data, and desired properties of the model in the forms of:

- Smoothness of the model behavior.
- Partially or completely known models.
- Constraints on model structure, variables and behavior, like stability and linearity.
- Empirical data (steady-state, time series).

The various pieces of knowledge or properties can be locally valid under certain operating conditions only, or globally valid. Moreover, varying levels of accuracy, completeness, and reliability of the knowledge can be handled.

The unifying framework we propose here is an optimization formulation of the joint modeling and identification problem, where the different pieces of knowledge will enter as penalty terms or constraints in the optimization formulation. This formulation of the problem has been inspired by Thompson & Kramer (1994), which discussed how different kinds of knowledge can be used to structure the model and optimization criterion, but in a significantly less general framework than we employ here. Tulleken (1993) has also suggested the use of constraints derived from system knowledge to improve the model, in the context of linear system identification. Another source of inspiration is (Girosi, Jones & Poggio 1994), see also (Girosi, Jones & Poggio 1995), which analyzes how different non-smoothness penalties lead to different basis-functions in a series expansion of the model. The present work can be viewed as a unification and extension of these approaches. Contrary to the static model formulation in (Girosi et al. 1994, Thompson & Kramer 1994), we formulate the problem for dynamical systems described by MISO NARMAX (non-linear ARMAX) models. It should be noted that the problem can be easily reformulated for the general MIMO case for virtually any model representation, including state-space models. However, it is difficult to derive useful general results for the more general case.

The paper is organized as follows. In section 2 the problem is formulated mathematically and some simple motivating examples are given. Next, in section 3, the exact solution to a simplified modeling problem is found explicitly using function space methods, and this solution is discussed in detail, before numerical procedures that solve the general problem are discussed in section 4. In section 5, we discuss how to tune the parameters in the optimization criterion and present alternative criteria and procedures for this. Section 6 contains a fairly complete semi-realistic simulation example that illustrates the power of the approach, and also the effect of different kinds of prior knowledge and empirical data on the model. The paper ends with a discussion and some conclusions.

2 Problem Formulation

In this work we study the identification of NARMAX models (Leontaritis & Billings 1985) of the form

$$M(f): \qquad y(t) = f(\psi(t-1)) + e(t) \tag{1}$$

where $\psi(t-1) = (y(t-1), ..., y(t-n_y), u^T(t-1), ..., u^T(t-n_u), e(t-1), ..., e(t-n_e))^T$ is the information vector. The model is "parameterized" by the function $f \in \mathcal{F}$. Notice that the model $M(f)$ should be viewed as non-parametric, since f is an infinite-dimensional "parameter". For simplicity we assume the system output $y(t)$ is a scalar, while the system input $u(t) \in R^r$ may be r-dimensional. Let $\psi(t) \in \Psi \subset R^n$, where Ψ is a compact subset in which we are interested in modeling the system's input/output behavior. Furthermore, let $\mathcal{F}$ be an inner product space containing smooth functions on Ψ

$$\mathcal{F} \quad = \quad \{f \in L_2(\Psi) \mid f \text{ is smooth}\}$$

The smoothness in the definition of $\mathcal{F}$ should be interpreted as existence and continuity of sufficiently high-order derivatives of the functions in $\mathcal{F}$. For the purpose of model identification, a sequence of l input/output observations

$$((u(1), y(1)), (u(2), y(2)), ..., (u(l), y(l)))$$

may be available. Moreover, we allow for a possibly inaccurate default model M_a to be available a priori, together with soft equality constraints $Qf \approx q$, where $Q : \mathcal{F} \rightarrow \mathcal{Q}$ is an operator, $\mathcal{Q}$ is an inner product space, and $q \in \mathcal{Q}$. Furthermore, we allow hard constraints $Hf \leq 0$ and $Pf = 0$, where $H : \mathcal{F} \rightarrow \mathcal{H}$ and $P : \mathcal{F} \rightarrow \mathcal{P}$ are operators, and $\mathcal{H}$ and $\mathcal{P}$ are inner product spaces. In addition, an operator $S : \mathcal{F} \rightarrow \mathcal{S}$ may be given, where $\mathcal{S}$ is an inner product space and $||Sf||$ indicates the non-smoothness of f. For example, S may be a differential operator, since large high-order derivatives are intuitively symptoms on "oscillations" or non-smooth behavior.

We are now in position to informally state the problem addressed in this work, namely to find the function $f \in \mathcal{F}$ that defines the NARMAX model $M(f)$ that is most consistent with the a priori knowledge and empirical data, in a well defined sense. For simplicity, we will assume the integer parameters n_y, n_u and n_e in the model are given.

Finding the best $f \in \mathcal{F}$ can be formulated as an optimization problem in the inner product space $\mathcal{F}$. The prior knowledge and empirical data are used to formulate an optimization criterion that penalizes

1. Mismatch between model prediction and the empirical data.
2. Non-smoothness of the model.
3. Mismatch between model and the default model.
4. Violation of the soft constraints.

In addition, we allow hard equality and inequality constraints to the optimization problem. We choose the criterion

$$J(f) \quad = \quad \frac{1}{l}\sum_{t=1}^{l} \varepsilon^2(t; f) + \lambda||Sf||^2 + \gamma\mathcal{D}^2(M(f), M_a) + \beta||Qf - q||^2 \tag{2}$$

subject to the constraints

$$Hf \le 0 \qquad \text{and} \qquad Pf = 0 \tag{3}$$

where λ, γ, and β are non-negative constants, and the norms are induced by possibly weighted inner products in the various spaces. $\mathcal{D}(M(f), M_a)$ measures the distance between the model $M(f)$ and the default model M_a, and $\varepsilon(t; f)$ can for example be the residual with a one-step-ahead predictor defined by

$$\varepsilon(t; f) \;=\; y(t) - f(y(t-1), ..., y(t-n_y), u(t-1), ..., u(t-n_u), \varepsilon(t-1; f), ..., \varepsilon(t-n_e; f))$$

or with a multi-step-ahead prediction error. It must be assumed that this model on innovation form is asymptotically stable. Obviously, the prior knowledge and the choice of constants λ, γ and β will have great influence on the model. In particular, the problem of choosing these constants will be discussed in detail in section 5.

Before we proceed with the solution to this optimization problem, it will be illustrative to motivate this optimization formulation with examples of some prior knowledge and desired properties of the model that can be specified within this framework. A more complete example is provided in section 6.

Example 1. Default model.

Consider a default model M_a of the form

$$\left.\begin{array}{c} \dot{x} = f_a(x, u) + v \\ y = g_a(x) + w \end{array}\right\} \quad M_a$$

The purpose of the default model is to provide a basic model for use duringoperating conditions when there are no empirical data or other more relevant prior knowledge available. The model M_a can be used for one-step-ahead predictions, and a predictor can be formulated using an extended Kalman-filter, e.g. (Söderström & Stoica 1988)

$$\begin{aligned} \hat{x}(t|t) &= \hat{x}(t|t-1) + K(t)\left(y(t) - g_a(\hat{x}(t|t-1))\right) \\ \hat{x}(t+1|t) &= \int_{\tau=t}^{t+1} f_a(\hat{x}(\tau|t), u(\tau))d\tau \\ \hat{y}(t+1|M_a, t) &= g_a(\hat{x}(t+1|t)) \end{aligned}$$

where the Kalman-filter gain $K(t)$ is computed using the assumption that $v(t)$ and $w(t)$ are zero-mean white noise processes with known variance. A multi-step-ahead predictor can be formulated in a similar manner.

The distance $\mathcal{D}(M(f), M_a)$ between the models $M(f)$ and M_a can be defined in several ways. Since the two models are based on fundamentally different representations (NARMAX and continuous-time state-space), there is no direct way of comparing these two models in this case. Consequently, the natural approach is to compare them indirectly, by comparing their prediction performance when both are used to predict the response of the system to different inputs from different initial states. Hence, we suggest

$$\mathcal{D}^2(M(f), M_a) \;=\; \int_{x(t)} \int_{u(t)} \left(\hat{y}(t+1|M(f), t) - \hat{y}(t+1|M_a, t)\right)^2 \rho(t) dx(t) du(t)$$

where $\rho(t)$ is a weight-function that penalizes mismatch between the two models as a function of the operating condition at each time instant t.

□

Example 2. Linear noise model.

Suppose we want to restrict the model (1) to the set of models of the form

$$\begin{aligned} y(t) \;=\; & f'(y(t-1), ..., y(t-n_y), u(t-1), ..., u(t-n_u)) \\ & + c_1 e(t-1) + ... + c_{n_e} e(t-n_e) + e(t) \end{aligned}$$

where $c_1, c_2, ..., c_{n_e}$ are unknown constants. This gives a linear noise model that may be desirable because it simplifies estimation, analysis and application of the model. This restriction can be represented as a constraint

$$\frac{\partial^2}{\partial e(t-1)^2} f \;=\; \cdots \;=\; \frac{\partial^2}{\partial e(t-n_e)^2} f \;=\; 0$$

which can be written as a linear operator equation $Pf = 0$, by defining $\mathcal{P} = L_2^{n_e}(R^n)$ and

$$P = \begin{pmatrix} \frac{\partial^2}{\partial e(t-1)^2} \\ \vdots \\ \frac{\partial^2}{\partial e(t-n_e)^2} \end{pmatrix}$$

□

Example 3. Known linear model.

Next, suppose the origin is an equilibrium point for the system. A linearized model is given by

$$y(t) = \nabla_\psi^T f(0)\psi(t-1) + e(t) \tag{4}$$

Now, suppose a linear (ARMAX) model

$$A(z^{-1})y(t) = B(z^{-1})u(t) + C(z^{-1})e(t) \tag{5}$$

of the system behavior near the origin is known. Penalty on deviations of the linearization of the non-linear model from this linear model can be formulated as a soft constraint of the form $Qf \approx q$. The distance between the two linear models (4) and (5) can be measured using different metrics, for example in the frequency domain, time domain, or some model space. Perhaps the simplest is to measure their difference in the parameter space. Then we choose $\mathcal{Q} = R^n$, $Qf = \nabla_\psi f(0)$, and q is the Riesz representation of (5) (written on the form $y(t) = q^T\psi(t-1) + e(t)$). A more robust model space metric would involve comparison of poles, zeros, and gain.

Notice that this formulation is very different from a default model M_a on the form $y(t) = f_a(\psi(t-1)) = q^T\psi(t-1)$ where deviations from the default model will in general be penalized globally over the full range of operating conditions, while in this example the penalty is defined only in a single operating point. Of course, the non-smoothness penalty will indirectly make this penalty effective in a neighborhood of this operating point.

If parts of a linear model is known, like an approximate pole valid in some operating regime, this knowledge can be incorporated in the same way.

□

Example 4. Steady-state mass balance.

Consider a system which has two input flows with rates $u_1(t)$ and $u_2(t)$, and one output flow with rate $y(t)$. If we keep the input flow rates fixed at some values $u_1(t) = \overline{u}_1$ and $u_2(t) = \overline{u}_2$, and the system reaches a steady-state $\overline{y}$, then we may conclude that the following steady state mass balance holds:

$$\overline{u}_1 + \overline{u}_2 - \overline{y} = 0$$

Clearly, it may be desirable that the model have the same property. Let Pf be the steady-state output of the system defined by f, then $Pf \in \mathcal{P} = L_2(R^2)$ is a function of the steady-state inputs $\overline{u}_1$ and $\overline{u}_2$. P will in general be a non-linear operator that can be implemented by solving the difference equation exactly or approximated by a recursion that will be stopped when sufficiently close to convergence.

□

Example 5. Stability.

Another useful kind of prior knowledge is that the system is open-loop stable, or more precisely, that certain equilibrium points are asymptotically stable. If we seek a model of the form

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

then local asymptotic stability of an equilibrium point $(\overline{y}, \overline{u})$ is guaranteed provided the following inequality constraints are satisfied

$$-1 \ < \ \frac{\partial f}{\partial y(t-1)}(\overline{y}, \overline{u}) \ < \ 1$$

□

The choice of inner product on the different inner product spaces is of great importance. For example, the region of validity or relevance of a default model M_a of the NARMAX form (1), defined by a function $f_a \in L_2(\Psi)$, can be represented by a weighted L_2 inner product

$$\langle f_1, f_2 \rangle_\rho \quad = \quad \int_{\psi \in \Psi} f_1(\psi) f_2(\psi) \rho(\psi) d\psi$$

where ρ is a strictly positive weight function that integrates to one, and can be interpreted as our a priori knowledge about the validity of the default model as a function of the operating point, since we may define

$$\mathcal{D}^2(M(f), M_a) \ = \ ||f - f_a||_\rho^2 \quad = \quad \int_{\psi \in \Psi} (f(\psi) - f_a(\psi))^2 \rho(\psi) d\psi$$

where $|| \cdot ||_\rho$ is the induced norm. A similar weighting may also be useful for the constraints, since they may be more or less relevant under different operating conditions. One may also represent knowledge about the smoothness of the system behavior as a function of the operating point by a weighted inner product. Alternatively, the weight can be interpreted as the relative desired accuracy of the model as a function of the operating point, if the weights are similar for all inner products. A useful class of function spaces are Sobolev spaces (Adams 1975), which are subspaces of $L_2(\Psi)$ with inner products like

$$< f_1, f_2 > \quad = \quad \int_{\psi \in \Psi} f_1(\psi) f_2(\psi) d\psi + \int_{\psi \in \Psi} \langle \nabla_\psi f_1(\psi), \nabla_\psi f_2(\psi) \rangle \, d\psi$$

if an objective is to make the Jacobian of the model to be a close approximation to the Jacobian of the system. This may be of particular interest if the model is used for control system design, where it is often the accuracy of the Jacobian of f that is most important.

3 Optimization based on function space methods

In this section we will study the solution to the problem of minimizing the functional (2) in the function space $\mathcal{F}$, using calculus of variations. The purpose of this section is to give material for a discussion of some interesting features of this approach, rather than developing practical algorithms, which is the topic of section 4.

For simplicity, we will assume that there are no soft or hard constraints, the inner products are not weighted, $n_e = 0$ (we want to find an NARX model), and the default model is of the same form as the model we are seeking (i.e. an NARX model). Hence, the default model is defined by a function $f_a \in L_2(\Psi)$, which allows the simple distance metric $\mathcal{D}^2(M(f), M_a) = ||f - f_a||^2$. In other words, we consider a criterion

$$J(f) \quad = \quad \frac{1}{l} \sum_{t=1}^{l} \varepsilon^2(t; f) + \lambda ||Sf||^2 + \gamma ||f - f_a||^2$$

Like Girosi et al. (1994), we choose to represent the penalty on non-smoothness in the spatial frequency domain using the Fourier transform

$$\tilde{f}(\nu) \;=\; \int_{\psi \in \Psi} f(\psi) e^{j\psi^T \nu} d\psi$$

where $\tilde{f}$ is the Fourier transform of $f \in \mathcal{F}$, and ν is an n-dimensional vector that is interpreted as spatial frequency (frequency is here the inverse of distance in the information space, not time). Since f is continuous, and Ψ has compact support, this transform is well defined, and the inverse transform

$$f(\psi) \;=\; \frac{1}{(2\pi)^n} \int_{\nu \in R^n} \tilde{f}(\nu) e^{-j\psi^T \nu} d\nu$$

will be unique. On $L_2(\Psi)$ we apply the standard inner product

$$< f_1, f_2 > \;=\; \int_{\nu \in R^n} \tilde{f}_1(\nu) \tilde{f}_2^{\star}(\nu) d\nu$$

where the superscript in $\tilde{f}^{\star}(\nu)$ denotes complex conjugate. We restrict our attention to operators S of the form (Girosi et al. 1994)

$$(Sf)(\psi) \;=\; \frac{1}{(2\pi)^n} \int_{\nu \in R^n} \frac{1}{\tilde{G}'(\nu)} \tilde{f}(\nu) e^{-j\psi^T \nu} d\nu \tag{6}$$

In other words, $\tilde{f}$ is filtered by $\left(\tilde{G}'(\nu)\right)^{-1}$, which will be chosen as a symmetric filter that amplifies high-frequency energy, since this gives penalty on high-frequency energy in f, which is an intuitive measure of non-smoothness. The induced norm is

$$||Sf||^2 \;=\; \int_{\nu \in R^n} \frac{\left|\tilde{f}(\nu)\right|^2}{\tilde{G}(\nu)} d\nu$$

where $\tilde{G}(\nu) = \left|\tilde{G}'(\nu)\right|^2$.

The case when $\gamma = 0$ is analyzed in detail by Girosi et al. (1994). Applying the mathematical tools in (Wahba 1990, Madych & Nelson 1990, Dyn 1989), it is shown that the global minimum of J can be represented as

$$f(\psi) \;=\; \sum_{t=1}^{l} \alpha_t G(\psi - \psi(t-1)) + \sum_{j=1}^{d_G} \beta_j \xi_j^G(\psi) \tag{7}$$

where $\alpha_1, \alpha_2, ..., \alpha_l$ and $\beta_1, \beta_2, ..., \beta_{d_G}$ are real constants, and the set of linearly independent functions $\{\xi_1^G, \xi_2^G, ..., \xi_{d_G}^G\}$ satisfies

$$\frac{1}{\tilde{G}(\nu)} \tilde{\xi}_j^G(\nu) \;=\; 0 \tag{8}$$

The constant parameters $\alpha_1, \alpha_2, ..., \alpha_l$ and $\beta_1, \beta_2, ..., \beta_{d_G}$ can be found explicitly by substituting (7) back into the functional J and minimizing this as a function of this finite number of parameters, cf. (Wahba 1990, Girosi et al. 1994).

Observe that the ξ_j^G-terms in the model appear because certain functions will be "invisible" to the non-smoothness penalty, and such terms can be added to the solution without imposing any cost. For example, if $\tilde{G}'(\nu) = (j\nu)^{-2}$ then $\tilde{G}(\nu) = (-j\nu)^{-4} = \nu^{-4}$, and $(\tilde{G}(\nu))^{-1}$ is an operator that differentiates four times. Hence, $\xi_1^G, ..., \xi_{d_G}^G$ is a basis for the space of polynomials with order no greater than three. As a curiosity, the resulting basis functions are in this case cubic splines (Wahba 1990).

Next, we show that the case $\gamma > 0$ can be treated using the same technique. First, we rewrite the criterion $J(f)$ in terms of $\tilde{f}$ as

$$\begin{aligned} J(f) &= \frac{1}{l}\sum_{t=1}^{l}\left(y(t) - \frac{1}{(2\pi)^n}\int_{\nu\in R^n}\tilde{f}(\nu)e^{-j\nu^T\psi(t-1)}d\nu\right)^2 \\ &\quad +\lambda\int_{\nu\in R^n}\frac{\tilde{f}(\nu)\tilde{f}^\star(\nu)}{\tilde{G}(\nu)}d\nu + \gamma\int_{\nu\in R^n}(\tilde{f}(\nu)-\tilde{f}_a(\nu))(\tilde{f}^\star(\nu)-\tilde{f}_a^\star(\nu))d\nu \end{aligned}$$

J is a strictly convex functional that must have a unique global minimum. Using calculus of variations (e.g. (Luenberger 1969)), we first find the Gateaux variation

$$\begin{aligned} \delta J(f;\Delta f) &= -2\frac{1}{l}\sum_{t=1}^{l}(y(t)-f(\psi(t-1)))\frac{1}{(2\pi)^n}\int_{\nu\in R^n}\Delta\tilde{f}(\nu)e^{-j\nu^T\psi(t-1)}d\nu \\ &\quad +\lambda\int_{\nu\in R^n}\frac{\tilde{f}(\nu)\Delta\tilde{f}^\star(\nu)+\tilde{f}^\star(\nu)\Delta\tilde{f}(\nu)}{\tilde{G}(\nu)}d\nu \\ &\quad +\gamma\int_{\nu\in R^n}\left(\Delta\tilde{f}(\nu)\left(\tilde{f}^\star(\nu)-\tilde{f}_a^\star(\nu)\right)+\Delta\tilde{f}^\star(\nu)\left(\tilde{f}(\nu)-\tilde{f}_a(\nu)\right)\right)d\nu \end{aligned}$$

for an arbitrary perturbation $\Delta f \in \mathcal{F}$. It is straightforward to show that all terms are real, and it follows after some manipulation that

$$\begin{aligned} \delta J(f;\Delta f) &= 2\int_{\nu\in R^n}\mathrm{Re}\left(\Delta\tilde{f}(\nu)\left(-\frac{1}{(2\pi)^n}\frac{1}{l}\sum_{t=1}^{l}(y(t)-f(\psi(t-1)))e^{-j\nu^T\psi(t-1)}\right.\right. \\ &\quad \left.\left. +\lambda\frac{\tilde{f}^\star(\nu)}{\tilde{G}(\nu)}+\gamma\left(\tilde{f}^\star(\nu)-\tilde{f}_a^\star(\nu)\right)\right)\right)d\nu \end{aligned}$$

A sufficient condition for the Gateaux variation to be zero for all $\Delta f \in \mathcal{F}$ is that f satisfy the Euler-Lagrange equation

$$\frac{1}{(2\pi)^n}\frac{1}{l}\sum_{t=1}^{l}(y(t)-f(\psi(t-1)))e^{-j\nu^T\psi(t-1)} = \lambda\frac{\tilde{f}^\star(\nu)}{\tilde{G}(\nu)}+\gamma\left(\tilde{f}^\star(\nu)-\tilde{f}_a^\star(\nu)\right) \tag{9}$$

for all $\nu \in R^n$. Defining

$$\begin{aligned} \alpha_t &= \frac{1}{(2\pi)^n l\gamma}\left(y(t)-f(\psi(t-1))\right) \\ \tilde{K}(\nu) &= \tilde{G}(\nu)\left(\tilde{G}(\nu)+\lambda/\gamma\right)^{-1} \end{aligned}$$

we get

$$\tilde{f}(\nu) = \tilde{K}(\nu)\tilde{f}_a(\nu)+\sum_{t=1}^{l}\alpha_t\tilde{K}(\nu)e^{j\nu^T\psi(t-1)} \tag{10}$$

If the linearly independent set of functions $\{\xi_1^K,\xi_2^K,...,\xi_{d_K}^K\}$ satisfies

$$\frac{1}{\tilde{K}(\nu)}\tilde{\xi}_j^K(\nu) = 0 \tag{11}$$

we may add to the solution f any term in $\mathrm{span}\{\xi_1^K,\xi_2^K,...,\xi_{d_K}^K\}$ without violating the Euler-Lagrange equation (9). Hence, by transforming back into the original coordinates, we get the model

$$f(\psi) = (K * f_a)(\psi)+\sum_{t=1}^{l}\alpha_t K(\psi-\psi(t-1))+\sum_{j=1}^{d_K}\beta_j\xi_j^K(\psi) \tag{12}$$

We will later argue that $\tilde{K}(\nu)$ will always be a low-pass filter. Hence, $K * f_a$ is a smoothed variant of f_a. The fundamental differences between the cases $\gamma = 0$ and $\gamma > 0$ is interesting. When $\gamma > 0$, the model (12) is the smoothed default model f_a in addition to some smooth terms that compensate for mismatch between the smoothed default model and the observed data. Another major difference is that with $\gamma = 0$, the basis-functions are translations of $G(\psi)$, while they are translations of $K(\psi)$ for $\gamma > 0$. As one should expect, $K \to G$ as $\gamma \to 0$. We will in the following illustrate this difference with some examples.

Example: Relation between non-smoothness penalty and basis-functions.

Let us look at some typical choices for $\tilde{G}'(\nu)$, and look at the corresponding basis-functions $G(\psi)$ and $K(\psi)$ in the one-dimensional case ($n = 1$). As discussed, we must choose $\tilde{G}'(\nu)$ such that $(\tilde{G}(\nu))^{-1}$ is a symmetric filter that amplifies high-frequency energy.

1. Let us first choose $\tilde{G}'(\nu) = (j\nu)^{-1}$

$$\tilde{G}(\nu) = \frac{1}{\nu^2}, \qquad G(\psi) = -\frac{1}{2}|\psi|$$
$$\tilde{K}(\nu) = \frac{\sigma^2}{1+\sigma^2\nu^2}, \qquad K(\psi) = \frac{1}{2\sigma}e^{-|\psi|/\sigma}$$

   where $K(\psi)$ is characterized by the parameter $\sigma = \sqrt{\lambda/\gamma}$. A large penalty on non-smoothness (large λ) will give a soft basis-function K that will extrapolate the data points widely, while a small λ will give basis-functions that are sharp spikes. It is interesting to observe that while $K(\psi)$ is a local basis-function (a bump centered at the origin), $G(\psi)$ is a global basis-function. The reason for this will be discussed below. A model representation that utilizes the basis-function G is the hinging hyper-plane model representation of Breiman (1993). Also notice that S is the differential operator, and $\tilde{G}(\nu) = -(j\nu)^{-2}$ differentiates twice. Then $d_G = 2$, and the basis $\xi_1^G(\nu) = 1$ and $\xi_2^G(\nu) = \psi$ is a possible choice.

2. Next, consider the extreme case when we impose infinite penalty on high frequencies

$$\tilde{G}(\nu) = \begin{cases} 0, & |\nu| > \overline{\nu} \\ 1, & |\nu| \le \overline{\nu} \end{cases}, \qquad G(\psi) = \frac{\pi}{\overline{\nu}}\mathrm{sinc}(\overline{\nu}\psi)$$
$$\tilde{K}(\nu) = \begin{cases} 0, & |\nu| > \overline{\nu} \\ \frac{\sigma^2}{1+\sigma^2}, & |\nu| \le \overline{\nu} \end{cases}, \qquad K(\psi) = \frac{\pi\sigma^2}{\overline{\nu}(1+\sigma^2)}\,\mathrm{sinc}(\overline{\nu}\psi)$$

   which gives basis-functions $K(\psi)$ and $G(\psi)$ which are equal, except for a factor, and are local oscillating bumps centered at the origin. Notice that (7) and (11) are satisfied for all functions that have Fourier transform equal to zero for frequencies above $\overline{\nu}$. This is an infinite-dimensional space of band-limited function, and d_K and d_G are infinite. The basis-function G is a reproducing kernel for the Hilbert space of band-limited functions, and such kernels have been suggested as basis-functions in empirical modeling by Mosca (1970).

□

From the definition of $\tilde{K}(\nu)$ we immediately get the approximation

$$\tilde{K}(\nu) \quad \approx \quad \begin{cases} \frac{\gamma}{\lambda}\tilde{G}(\nu), & \text{if } \left(\tilde{G}(\nu)\right)^{-1} > \frac{\gamma}{\lambda} \\ 1, & \text{if } \left(\tilde{G}(\nu)\right)^{-1} < \frac{\gamma}{\lambda} \end{cases} \tag{13}$$

The first example illustrates the case when $\left(\tilde{G}(\nu)\right)^{-1} \to \infty$ as $\nu \to \infty$ and $\left(\tilde{G}(\nu)\right)^{-1} \to 0$ as $\nu \to 0$. While the filter $\tilde{G}(\nu)$ amplifies low-frequency energy and suppresses high-frequency energy, $\tilde{K}(\nu)$ will let low-frequency energy pass while it suppresses high-frequency energy. The amplification of low-frequency energy is essentially why the impulse response $G(\psi)$ is a global basis-function, while $K(\psi)$ is a local basis-functions.

In the second example, $\left(\tilde{G}(\nu)\right)^{-1} \to \infty$ as $\nu \to \infty$, while $\left(\tilde{G}(\nu)\right)^{-1} \to 1$ as $\nu \to 0$. This leads to $\tilde{K}(\nu) \approx \tilde{G}(\nu)$ for all ν, and since $\tilde{G}(\nu)$ does not amplify low-frequency energy, both $K(\psi)$

and $G(\psi)$ will be local basis-functions. We conclude that it is the low-frequency characteristic of $\left(\tilde{G}(\nu)\right)^{-1}$ that determines whether $G(\psi)$ is local or global. From (13) it follows that as long as $\left(\tilde{G}(\nu)\right)^{-1}$ does not amplify low-frequency energy, $K(\psi)$ will be a local basis-function. For all sensible choices of $\tilde{G}'(\nu)$, this will be the case. This emphasizes the interpretation that the data-dependent part of the model will compensate for the mismatch between the observed data and the smoothed default model, and this data dependent part will only have influence on the total model locally in a neighborhood of each observation. Intuitively, this makes sense, and a similar hybrid parametric model structure has been suggested on the basis of engineering considerations by Kramer, Thompson & Phagat (1992), see also (Thompson & Kramer 1994). The main difference is that the default model is not smoothed, and that the basis functions are chosen in an ad hoc manner.

Solving the modeling problem directly in the function space gives a mathematically very elegant solution. However, an exact mathematical solution can be found easily only for rather simple problems not involving complicated constraints, non-linear operators, a complicated default model, complicated inner products, or a complicated multidimensional geometry of Ψ. This may appear to seriously limit the applicability of the optimization approach, but we shall see in the next section that one can, as an alternative, study solutions based on approximating the optimal non-parametric model closely with a finite dimensional semi-parametric approximation. This is a more powerful practical approach, involving numerical computation suitable for a computer.

4 Optimization based on parameterized approximations

With a suitable parameterization of f, the problem of directly minimizing (2) with respect to these parameters will be computationally feasible in many cases. First, we must choose a rich representation for $f(\psi;\theta)$. This representation can be almost arbitrary since it is of minor importance with the augmented optimization criterion approach. What is required is only that the optimal non-parametric model can be approximated closely. As we shall see later, the effective number of parameters (degrees of freedom) in this model will be controlled through the choice of λ, γ, and β. Hence, the number of parameters in the semi-parametric representation of f can be arbitrarily large. Of course, if prior knowledge is available and suggests a particular parameterized structure for f, this should indeed be used. However, in this paper we assume that no prior knowledge of this kind is available. It is tempting to choose a linearly parameterized representation $f(\psi;\theta) = \theta^T\varphi(\psi)$, like a look-up table with interpolation. The reason is simply that this will make the optimization criterion quadratic for a significant number of problems. However, in cases when the dimension of ψ is high such representations may not be well suited, because the required number of parameters will typically grow exponentially with the dimension n if the accuracy of the approximation is required to be uniform. This is known as the curse of dimensionality. Since the effective number of parameters can be tuned by other means, the implication of this is purely a computational problem. Hence, non-linearly parameterized representations based on low-dimensional projections, like neural nets, may be more effective from a computational point of view in these cases.

With a parameterized model structure it is straightforward to reformulate the infinite-dimensional optimization problem (2) into the finite dimensional problem

$$\hat{J}(\theta) \quad = \quad \frac{1}{l}\sum_{t=1}^{l}\varepsilon^2(t; f(\cdot;\theta)) + \lambda||Sf(\cdot;\theta)||^2 + \gamma\mathcal{D}^2(M(f(\cdot;\theta)), M_a) + \beta||Qf(\cdot;\theta) - q||^2$$

subject to

$$Hf(\cdot;\theta) \leq 0, \qquad Pf(\cdot;\theta) = 0$$

If all operators operating on f are redefined as operators operating on the parameter vector θ, we get the optimization problem

$$\hat{J}(\theta) \quad = \quad \frac{1}{l}\sum_{t=1}^{l}\varepsilon^2(t;\theta) + \lambda||\hat{S}(\theta)||^2 + \gamma\mathcal{D}^2(M(f(\cdot;\theta)), M_a) + \beta||\hat{Q}(\theta) - q||^2 \qquad (14)$$

subject to

$$\hat{H}(\theta) \leq 0, \qquad \hat{P}(\theta) = 0 \tag{15}$$

Notice that the ranges of the operators $\hat{H}, \hat{P}, \hat{S}$ and $\hat{Q}$ are still of possibly infinite dimension. A taxonomy for such problems, together with possible solution methods, is:

- Quadratic criterion, and no constraints. This gives a set of linear algebraic equations that can easily be solved.
- Quadratic criterion, linear finite-dimensional constraints. Solved using quadratic programming.
- Non-quadratic criterion, no constraints. This is a non-linear programming problem that can be solved iteratively.
- Non-quadratic criterion, finite-dimensional constraints. Constrained non-linear programming problem.
- Infinite-dimensional constraints gives a semi-infinite programming problem.

An in-depth discussion of various practical optimization algorithms that take advantage of the particular structure of the problem is outside the scope of this paper. Instead, we refer to the vast literature on general-purpose algorithms, e.g. (Luenberger 1984, Gill, Murray & Wright 1981, Tanaka, Fukushima & Ibaraki 1988), and software tools like the MATLAB Optimization Toolbox (Grace 1990), and will here discuss in depth only the simple case when f is linearly parameterized, i.e.

$$f(\psi;\theta) = \sum_{n=1}^{N} \theta_i \varphi_n(\psi)$$

where $\{\varphi_n\}_{n=1}^N$ is a set of linearly independent smooth basis-functions defined on Ψ. Furthermore, we assume there are no constraints, and the default model is an NARMAX model defined by the function $f_a \in L_2(\Psi)$. We use the distance metric $\mathcal{D}^2(M(f), M_a) = ||f - f_a||^2$. While this is perhaps the simplest possible special case that remains interesting, it contains enough complexity to support a discussion of the various problems involved in solving the general optimization problem.

In this case we can easily compute the solution to the optimization problem explicitly. We define the vectors and matrices

$$\begin{aligned}
Y = (Y_t), &\qquad Y_t = y(t), \quad t = 1, ..., l \\
\Phi = (\Phi_{ti}), &\qquad \Phi_{ti} = \varphi_j(\psi(t)), \quad t = 0, ..., l-1, \quad i = 1, ..., N \\
\overline{V} = (\overline{V}_i), &\qquad \overline{V}_i = < \varphi_i, f_a >, \quad i = 1, ..., N \\
\overline{W} = (\overline{W}_i), &\qquad \overline{W}_i = < \varphi_i, q >, \quad i = 1, ..., N \\
\overline{S} = (\overline{S}_{ij}), &\qquad \overline{S}_{ij} = < S\varphi_i, S\varphi_j >, \quad i = 1, ..., N, \quad j = 1, ..., N \\
\overline{R} = (\overline{R}_{ij}), &\qquad \overline{R}_{ij} = < \varphi_i, \varphi_j >, \quad i = 1, ..., N, \quad j = 1, ..., N \\
\overline{Q} = (\overline{Q}_{ij}), &\qquad \overline{Q}_{ij} = < Q\varphi_i, Q\varphi_j >, \quad i = 1, ..., N, \quad j = 1, ..., N
\end{aligned}$$

Suppose S and Q are linear operators. Then we can reformulate the criterion as

$$\begin{aligned}
\hat{J}(\theta) &= \frac{1}{l} < \Phi\theta - Y, \Phi\theta - Y > + \lambda \left\langle \sum_{n=1}^{N} \theta_n S\varphi_n, \sum_{n=1}^{N} \theta_n S\varphi_n \right\rangle \\
&\quad + \gamma \left\langle \sum_{n=1}^{N} \theta_n \varphi_n - f_a, \sum_{n=1}^{N} \theta_n \varphi_n - f_a \right\rangle + \beta \left\langle \sum_{n=1}^{N} \theta_n Q\varphi_n - q, \sum_{n=1}^{N} \theta_n Q\varphi_n - q \right\rangle \\
&= \frac{1}{l}\theta^T \Phi^T \Phi \theta - 2\frac{1}{l}\theta^T \Phi^T Y + \frac{1}{l} Y^T Y + \lambda \theta^T \overline{S} \theta \\
&\quad + \gamma \left(\theta^T \overline{R} \theta - 2\theta^T \overline{V} + < f_a, f_a > \right) + \beta \left(\theta^T \overline{Q} \theta - 2\theta^T \overline{W} + < q, q > \right)
\end{aligned}$$

It follows that $\theta = (\theta_1, ..., \theta_N)^T$ is a global minimum of $\hat{J}$, provided that

$$\left(\frac{1}{l} \Phi^T \Phi + \lambda \overline{S} + \gamma \overline{R} + \beta \overline{Q} \right) \theta = \frac{1}{l} \Phi^T Y + \gamma \overline{V} + \beta \overline{W} \tag{16}$$

This set of equations will never be over-determined, so a solution will always exist. Moreover, it will be unique under very reasonable conditions, since all terms on the left-hand side of (16) are positive semi-definite, and uniqueness follows if at least one of them is positive definite. One possibility is that $\text{rank}(\Phi^T \Phi) = N$, which is related to the classical condition that one needs at least as many observations as unknown parameters. However, the introduction of prior knowledge allows us to relax this condition and use an over-parameterized approximate representation of the model. For example, a default model will always make the problem well-posed. In addition, there should be some restrictions on the choice of basis-functions $\{\varphi_n\}_{n=1}^N$. These restrictions are related to the fact that the operators S and Q may have null-spaces that may make certain functions "disappear" from the equations that determine the parameters, giving singular matrices $\overline{S}$ and $\overline{Q}$. A typical example is a differential operator that will map polynomials of sufficiently low order to the zero function.

It is evident that the computational complexity may be a considerable problem even for solving (16). The major computational cost involved is computing the inner products. In general, the basis-functions or operators may not always be so nice that this can be done analytically, and numerical integration suffers from the curse of dimensionality. Hence, the computational complexity may grow exponentially with the dimension of the information space. Of course, with non-linear parameterization, operators, or constraints, the problem is more apparent, and the computational complexity will reduce the applicability of the approach.

5 Tuning the criterion

In section 2 we formulated the semi-empirical joint modeling and identification problem as an optimization problem, and in sections 3 and 4 we discussed the solution to this problem. However, it is the choice of criterion that is the most challenging and interesting aspect in this optimization approach. While the choice of criterion terms (expect the non-smoothness penalty) are inherently problem specific, the problem of tuning of the criterion parameters (λ, γ, β) is a more generic problem, and is the topic of this section. From a more general perspective, the problem we discuss in this section is basically to decide on how much we trust the empirical data and the various pieces of prior knowledge relative to each other. Obviously, a lot can be said about this general problem, but here, space does not permit the in-depth discussion this problem deserves. We will focus on the case when we always require a smooth model, and give precedence to the empirical data relative to the prior knowledge for operating conditions when there are conflicts between the empirical data and the prior knowledge. On the other hand, we always rely on the prior knowledge under operating conditions when data is lacking.

If the available empirical data sequence is viewed as the realization of some stochastic process, and a model is identified on the basis of this data, then the expected squared prediction error can be decomposed into a systematic component (bias) and a random component (variance). The systematic component is due to the fact that the prior knowledge and empirical data set can be incorrect or incomplete. The random component is due to unpredictable noise, but most importantly caused by the fact that a finite number of observations is in general not sufficient to identify the best model parameters. The best model has a complexity that gives the best balance between bias and variance, cf. Fig. 1. It is therefore clear that the model complexity should depend on the information contents in the empirical data sequence, and the prior knowledge available. We will later see that the complexity of a model in some cases can be explicitly related to the effective number of parameters in the model, a quantity that depends on (λ, γ, β). Hence, it makes sense to let (λ, γ, β) depend on the empirical data. Assume that we want to find criterion parameters (λ, γ, β) such that the model minimizing J will be the one that has the best expected squared prediction error possible. It is easy too see that the performance of the model will indeed depend on these parameters. For example, a too small λ, γ or β may lead to over-fitting and poor performance when extrapolating, since too much emphasis will be given to the data. Likewise, a too large λ, γ or β may give too little emphasis on the empirical data, and a model that is too biased under operating conditions where the prior knowledge is incorrect or incomplete.

This suggests a hierarchical optimization approach where the criterion parameters (λ, γ, β) are optimized in an outer loop using a criterion $V(\lambda,\gamma,\beta)$ that reflects the expected squared prediction error on future data, cf. Fig 2. In a stochastic framework, exact computation of this criterion requires the joint probability distribution for all stochastic variables in the model to be known. Clearly, this is unrealistic and we therefore consider some practical alternatives based on empirical data, and mainly assume ergodicity of the stochastic processes.

First, suppose another independent data sequence

$$((u(l+1),y(l+1)),(u(l+2),y(l+2)),...,(u(l+L),y(l+L)))$$

with L input/output pairs is available. Then an estimate of $V(\lambda,\gamma,\beta)$ is given by

$$V_S(\lambda,\gamma,\beta) \quad = \quad \frac{1}{L}\sum_{t=l+1}^{L}\varepsilon^2(t;\theta(\lambda,\gamma,\beta))$$

where the prediction error $\varepsilon(t;\theta(\lambda,\gamma,\beta))$ is with a parameter vector $\theta(\lambda,\gamma,\beta)$ that solves the problem (14)-(15). In other words, an independent data sequence is used to test the prediction performance of the identified model, which is reasonable if this data sequence reflects the underlying distribution of the future data. This method is simple and has a theoretical foundation (Johansen & Weyer 1995), but suffers from the major drawback that a larger amount of empirical data is required.

As an alternative, to overcome this drawback, one can re-use the original data sequence using cross-validation. The cross-validation criterion is an estimate of V given by

$$V_{CV}(\lambda,\gamma,\beta) \quad = \quad \frac{1}{l}\sum_{t=1}^{l}\varepsilon^2(t;\theta^{-t}(\lambda,\gamma,\beta))$$

where $\theta^{-t}(\lambda,\gamma,\beta)$ is the parameter vector that solves the problem (14)-(15), when modified such that there is no weight on the residual at time t. This means that l models are fitted by removing one residual at a time from the criterion, and the prediction performance of each model is tested on the single observation that was removed from the criterion. It should be noted that cross-validation lacks a strong theoretical justification, in particular for correlated data sequences. More details can be found in e.g. (Wahba 1990, Stoica, Eykhoff, Janssen & Söderström 1986).

As a third alternative, when $\varepsilon(t;\theta)$ depends linearly on θ, a suitable criterion is a modified version of the Final Prediction Error criterion by Akaike (1969):

$$V_{FPE}(\lambda,\gamma,\beta) \quad = \quad \frac{1+d_1(\lambda,\gamma,\beta)/l}{1-d_1(\lambda,\gamma,\beta)/l}\; V_0(\theta(\lambda,\gamma,\beta))$$

where $d_1(\lambda,\gamma,\beta) = \mathrm{tr}(A(\lambda,\gamma,\beta)A(\lambda,\gamma,\beta))$ is interpreted as the effective number of parameters in the model (degrees of freedom), and

$$\begin{aligned} A(\lambda,\gamma,\beta) &= \Phi^T\Phi\left(\Phi^T\Phi+\lambda l\overline{S}+\gamma l\overline{R}+\beta l\overline{Q}\right)^{-1} \qquad (17)\\ V_0(\theta) &= \frac{1}{l}\sum_{t=1}^{l}\varepsilon^2(t;\theta) \end{aligned}$$

When the residuals depend non-linearly on θ, the FPE criterion for regularized models by Larsen & Hansen (1994) can be applied

$$V_{FPER}(\lambda,\gamma,\beta) \quad = \quad \frac{1+d_2(\lambda,\gamma,\beta)/l}{1-2d_3(\lambda,\gamma,\beta)/l+d_2(\lambda,\gamma,\beta)/l}\; V_0(\theta(\lambda,\gamma,\beta))$$

where

$$\begin{aligned} d_2(\lambda,\gamma,\beta) &= \mathrm{tr}\left(\mathcal{A}(\lambda,\gamma,\beta)\right)\\ d_3(\lambda,\gamma,\beta) &= \mathrm{tr}\left(\mathcal{A}(\lambda,\gamma,\beta)\mathcal{A}(\lambda,\gamma,\beta)\right)\\ \mathcal{A}(\lambda,\gamma,\beta) &= \frac{\partial^2}{\partial\theta^2}V_0(\theta(\lambda,\gamma,\beta))\left(\frac{\partial^2}{\partial\theta^2}V_0(\theta(\lambda,\gamma,\beta))+\frac{\partial^2}{\partial\theta^2}\hat{\Omega}(\theta(\lambda,\gamma,\beta))\right)^{-1} \qquad (18)\\ \hat{\Omega}(\theta) &= \lambda||\hat{S}(\theta)||^2+\gamma\mathcal{D}^2(M(f(\cdot;\theta)),M_a)+\beta||\hat{Q}(\theta)-q||^2 \end{aligned}$$

Next, consider the case when the parameters are constrained. Let the parameter estimate that solves (14)-(15) be a regular point denoted $\hat{\theta}$, and let c be the number of linearly independent equality and active inequality constraints at $\hat{\theta}$. Let us decompose θ into a c-dimensional sub-vector θ_2' and an $(N-c)$-dimensional sub-vector θ_1' such that $\theta = (\theta_1', \theta_2')$ and in a neighborhood of $\hat{\theta}$ there exists a function g such that the constraints are satisfied on the manifold defined by $\theta_2' = g(\theta_1')$. The existence of such a decomposition and function g is guaranteed by the Implicit Function Theorem, e.g. (Luenberger 1984). In other words, in a neighborhood of $\hat{\theta}$, the optimization problem can be reformulated as an unconstrained problem on an $(N-c)$-dimensional subspace of the parameter space. When regularization is not applied, it follows directly that the effective number of parameters equals $N-c$. However, in the general case, one must explicitly know the "implicit function" g in order to substitute $\theta = (\theta_1', g(\theta_1'))$ into $\hat{J}(\theta)$. It is easy to see that a linear approximation to g that is exact in the point $\hat{\theta}$, will be sufficient. Fortunately, such a linearization can usually be found from the equations for the constraints, but we do not want to go into the technical details here.

Clearly, there will be situations when the automatic procedure for choosing (λ, γ, β) suggested in this section will lead to spurious results. A trivial example is when the given default model is significantly wrong over the whole range of operating conditions where there are empirical data available, but adequate for other operating conditions. Then it is optimal to choose $\gamma = 0$, unless the inner product is chosen to give zero weight on the model near all the data points, in which case γ can be arbitrary. Hence, one reason why the algorithm may fail is in general that the future data distribution will be significantly different from the distribution of the empirical data used for identification. As discussed before, the automatic choice of λ, γ, and β is based on the assumption that the empirical data are more reliable than the prior knowledge. Such an assumption must be made in order to resolve conflicts between the different pieces of knowledge and data, and is controversial both in the philosophy of science, and in engineering practise. Still, even if the algorithm may fail when searching for γ or β, the choice of λ may still be sensible, since smoothness may be a uniform property of the system, i.e. uniformly true for all operating conditions, while the default model and constraints may be incorrect under some operating conditions. In summary, we do not recommend a completely automated choice of tuning parameters, in particularly not γ and β.

The estimated relative weights on the different terms in the criterion may be interpreted as indicators on the relative correctness or relevance of the corresponding prior knowledge or empirical data. As we have discussed above, such interpretations should be made carefully, as it is possible that knowledge that is incorrect or irrelevant for all the operating conditions we have empirical data available for, will be adequate for other operating conditions. One must always have the information contents in the data set in mind. However, if there are large amounts of data available, and the data covers all operating conditions, such interpretations may be valuable.

Finally, it should be mentioned that it is possible is include the order parameters n_y, n_u, and n_e together with λ, γ, and β in the search.

6 Simulation Example: pH neutralization tank

The purpose of this simulation example is to illustrate the effect of imprecise, incomplete, and incorrect prior knowledge and empirical data on the identified model, and how the given framework can handle the different kinds of prior knowledge and empirical data.

The system we will study is a pH neutralization tank. The system has one input (the base flow rate $u(t)$), one output (the measured pH $y(t)$), and no disturbances or noise. There are three influent streams and one effluent stream:

- Influent acid stream Q_1 (HNO_3)
- Influent buffer stream Q_2 ($NaHCO_3$)
- Influent base stream Q_3 ($NaOH$ and traces of $NaHCO_3$)
- Effluent stream Q_4

We use a model developed by Hall & Seborg (1989) to simulate the "true system", see Appendix A.

The available prior knowledge and empirical data is supposed to be as follows:

1. **Data sequence** from a limited region of operation, cf. Fig 3a. The sampling interval is 15 s, and the sequence contains 500 observations.
2. **Steady-state data** from 5 operating points, namely pH $\in \{5.03, 5.78, 6.31, 7.03, 9.15\}$.
3. **Simplified mass-balance model** where the buffer stream is neglected, cf. Appendix B.
4. **pH range**, in other words the fact that $0 \leq \text{pH} \leq 14$.
5. **Smoothness** of the system behavior.
6. **Open-loop stability** of the system.

We will study the problem of identifying a model of the NARX form

$$y(t) \quad = f(y(t-1), u(t-1)) + e(t) \tag{19}$$

The interesting operating range is defined by $0 \leq y \leq 14$ and $0 \leq u \leq 40$ ml/s. Hence, $\Psi = [0, 14] \times [0, 40]$. Notice that the available data sequence only covers a small part of Ψ, cf. Fig. 4a, while the model will be partially validated using a data sequence that covers a much larger part of Ψ, cf. Figs. 3b and 4b. The unknown non-linear function f is represented by a 2-dimensional lookup-table (θ_{ij}) with linear B-spline interpolation on this domain. The choice of representation is of minor importance, and the lookup-table is chosen because of its simplicity. The 390 entries in the lookup-table are considered as unknown parameters.

We choose a non-smoothness penalty of the form

$$\theta^T \overline{S} \theta \quad \approx \quad \int_{\psi \in \Psi} \left|\left| \frac{\partial^2}{\partial \psi^2} f(\psi) \right|\right|_F^2 d\psi$$

which is based on an easy-to-compute finite-difference approximation of the Hessian

$$\frac{\partial^2 f(\psi_{ij})}{\partial \psi^2} \quad \approx \quad \begin{pmatrix} \frac{\theta_{i+1,j} - 2\theta_{i,j} + \theta_{i-1,j}}{\Delta_1^2} & \frac{\theta_{i+1,j+1} + \theta_{i,j} - \theta_{i+1,j} - \theta_{i,j+1}}{\Delta_1 \Delta_2} \\ \frac{\theta_{i-1,j-1} + \theta_{i,j} - \theta_{i-1,j} - \theta_{i,j-1}}{\Delta_1 \Delta_2} & \frac{\theta_{i,j+1} - 2\theta_{i,j} + \theta_{i,j-1}}{\Delta_2^2} \end{pmatrix}$$

where $f(\psi_{ij}) = \theta_{ij}$ and the matrix norm applied is the Frobenius norm $||\cdot||_F$. Furthermore, $\Delta_1 = 1.0$ and $\Delta_2 = 1.6$ are the uniform discretization intervals in the lookup-table. One motivation behind this choice is that penalty on the Hessian can be interpreted as penalty on deviations from a linear model structure. This is reasonable, since a linear model structure is perhaps the simplest possible that may be adequate. A perhaps more convincing motivation is that small curvature intuitively is equivalent to smoothness.

In the following we will identify five semi-parametric models based on different combinations of the available knowledge and data, and as an alternative model structure we try a parametric neural network model for comparison.

6.1 Model 1. Knowledge: Data sequence, Smoothness, Stability

Open loop stability is represented as the constraint in the problem

$$\hat{J}(\theta) \quad = \quad \frac{1}{l} \sum_{t=1}^{l} (y(t) - \varphi^T(\psi(t-1))\theta)^2 + \lambda \theta^T \overline{S} \theta$$

subject to

$$|\theta_{i,j} - \theta_{i-1,j}| \quad \leq \quad \Delta_1, \qquad \text{for all} \quad i, j$$

This optimization problem is solved using quadratic programming (MATLAB function **qp**). Notice that the use of a lookup-table with linear interpolation allows us to reduce the stability constraint to a finite number of constraints on the parameters. The generalized FPE criterion, with the active constraints taken into account when computing the effective number of parameters, is minimized using a line search algorithm (MATLAB function **fmin**) to find the smoothing parameter $\lambda = 3.8 \cdot 10^{-4}$ corresponding to $d_1 = 17.1$ effective parameters. Notice that the prior knowledge about open loop stability is of great importance, since otherwise the model will be unstable for a wide range of λ values, because one-step-ahead prediction residuals are applied in the identification criterion.

6.2 Model 2. Knowledge: Data sequence, pH range, Smoothness

Notice that bounded pH range automatically implies BIBO-stability of the model. Here we solve the minimization problem

$$\hat{J}(\theta) = \frac{1}{l}\sum_{t=1}^{l}(y(t) - \varphi^T(\psi(t-1))\theta)^2 + \lambda\theta^T\overline{S}\theta$$

subject to

$$0 \leq \theta_{ij} \leq 14, \qquad \text{for all } \; i,j$$

Similar to Model 1, a minimization of the generalized FPE criterion gives $\lambda = 5.6 \cdot 10^{-5}$, which corresponds to $d_1 = 21.7$ effective parameters.

6.3 Model 3. Knowledge: Simplified mass-balance

No identification is needed, cf. the model in Appendix B, hereafter denoted M_a. Notice that this model contains very limited knowledge about the chemistry in the tank, since the buffer is neglected. Hence, the prior knowledge assumed is significantly less than the "true system", and the model is considerably simpler.

6.4 Model 4. Knowledge: Data sequence, Simplified mass-balance, Smoothness

The default model M_a (Model 3) is stable, so the stability constraint is not required. The criterion is now

$$\hat{J}(\theta) = \frac{1}{l}\sum_{t=1}^{l}(y(t) - \varphi^T(\psi(t-1))\theta)^2 + \lambda\theta^T\overline{S}\theta + \gamma D(M(f(\cdot;\theta)), M_a)$$

where

$$D(M(f(\cdot;\theta)), M_a) = \frac{1}{1000}\sum_{t=1}^{1000}(y_a(t) - \varphi^T(\psi_a(t-1))\theta)^2\rho(t)$$

and we have defined $\psi_a(t) = (y_a(t), u_a(t))^T$, and the data sequence $((u_a(1), y_a(1))$,$(u_a(2), y_a(2))$, ..., $(u_a(1000), y_a(1000)))$ is generated by simulating the model M_a. The input sequence is chosen to cover as much of the operating range of the process as possible. The weighting function is defined as $\rho(t) = 1$. In this case a search for λ and γ using the generalized FPE criterion fails. The reason is simply that the default model is significantly wrong for all the operating conditions captured by the sequence of empirical data available. Hence, the optimization gives $\gamma \approx 0$. A more intelligent choice of weighting function $\rho(t)$ that takes into account the inaccuracy of the default model for intermediate pH values, does not resolve the problem. Instead, we have chosen $\lambda = 10^{-3}$ and $\gamma = 0.05$, which gives $d_1 = 13.5$. This is motivated by a visualization of the model (actually the function f), which shows that this choice gives a smooth transition to the default model outside the range from which data is available.

6.5 Model 5. Knowledge: Steady-state data, Simplified mass-balance, Smoothness

Here we solve the minimization problem

$$\hat{J}(\theta) \quad = \quad \lambda \theta^T \overline{S} \theta + \gamma D(M(f(\cdot;\theta)), M_a) + \beta \frac{1}{5} \sum_{i=1}^{5} (y_{\text{steady-state}}(u_i) - y_i)^2$$

where $D(M(f(\cdot;\theta)), M_a)$ is the same as above. Since there is no data sequence available to support the choice of tuning parameters, we make the choice $\lambda = 0.005, \gamma = 0.05, \beta = 0.25$, again by consulting a visualization of f and giving a larger emphasis on the steady-state data relative to the default model. The steady-state model output is computed by simulating the model until a steady state is reached. The model's steady-state output is a non-linear function of the steady-state input, and the steady-state data is represented as five soft constraints. This optimization problem is solved using the MATLAB function **fminu**.

6.6 Models 6 and 7. Sigmoidal Neural Network

For comparison, we have also identified a feed-forward neural network model of the NARX form (19). This is an example of an a priori choice of a parametric model structure. The only reason why we have chosen this particular model structure is its popularity as a "general purpose" parametric model structure. The parameters are identified with the MATLAB Neural Network Toolbox (Demuth & Beale 1993), using the back-propagation algorithm with momentum term, the identification data sequence and a least squares criterion. Depending on the number of hidden nodes, the identification algorithm tuning-parameters, and initial parameter estimates, the result vary considerably. We have chosen to present two identified models with the same structure (10 hidden nodes with sigmoidal non-linearities, corresponding to 41 parameters), the same identification algorithm tuning-parameters, but different (small random) initial parameter estimates.

6.7 Results

The prediction performance of the models is illustrated by ballistic simulations using the validation data input sequence in Fig. 5. Moreover, the major non-linearity in this process is the steady-state response, or titration curve, which is shown for the different models in Fig. 6.

These curves clearly show that different aspects of the system behavior are captured, depending on the prior knowledge applied. The performance of Model 1 is poor at high and low pH values, since no empirical data and only weak prior knowledge like smoothness and stability is relevant under these conditions. Restricting the pH range gives some improvement, cf. Model 2. Likewise, the simplified mass balance that makes up Model 3 gives reasonable predictions only at high and low pH values. Using this model as a default model combined with the empirical data sequence from the intermediate pH range gives a model with reasonable predictions over the full pH range, cf. Model 4. It is also interesting to observe that the default model combined with a small amount of steady-state data from the intermediate pH range, but no dynamic data, gives qualitatively the same prediction results, cf. Model 5.

A comparison with the neural network models (Models 6 and 7) is also quite interesting. The two neural network models perform equally well on predicting the identification data, but perform considerably different on the validation data and have quite different titration curves. It should be mentioned that these differences are typical, and more pronounced differences (on the validation data) appear when the model structure and identification algorithm tuning-parameters change as well. Notice that it is most fair to compare Models 6 and 7 with Model 1, since they are both based on approximately the same amount of prior knowledge.

It seems reasonable to conclude that this example illustrates that the incorporation of increasing amounts of prior knowledge may improve the quality of the identified model. Furthermore, the optimization approach seems more adequate than the application of a black-box model structure like a neural network in this application, since the empirical data does not cover the operating range sufficiently.

This example also shows that the automatic choice of criterion tuning parameters (λ, γ, β) using empirical data is feasible only if the available data covers the interesting operating conditions sufficiently well, as one should expect. In particular, the algorithm gives reasonable choices for Models 1 and 2, but fails for Model 4.

7 Discussion

There are at least two major drawbacks to this approach:

1. The computational complexity may limit its applicability. This problem will in particular be apparent when the dimension of the information space is high, or there are non-linear operators or parameterizations involved. The problem may be reduced with the aid of tailored optimization algorithms, solving the problem analytically as far as possible, and trying to simplify the various elements involved. A library of flexible and computationally effective building blocks for representing various pieces of prior knowledge and information would be useful.

2. The resulting model need not be transparent, in the sense that it cannot easily be interpreted and analyzed, since it is non-parametric or semi-parametric. The only possibility is to interpret the model qualitatively in terms of the prior knowledge and empirical data applied. However, this may be too coarse for many applications. We would like to stress that all black-box model structures suffer from the same drawback, perhaps to an even larger extent, since the prior knowledge is usually applied in a less transparent manner.

The major advantage of the approach is its flexibility with respect to incorporation of prior knowledge in various forms. This flexibility can possibly be extended and improved in several directions:

1. As mentioned in the introduction, the extension to the MIMO-case is conceptually straightforward. Also, extensions to other model representations than NARMAX models, like discrete-time and continuous-time state-space models, follow directly, since (2)-(3) holds for any model parameterized by a function f. Moreover, the formulation can be extended to include multiple unknown functions, as well as a mixture of real, integer and function parameters.

2. A more accurate model (and better interpretation of estimated criterion tuning-parameters) may be found if the terms in the criterion are split into different terms that are valid under different operating conditions.

3. As mentioned before, a library of building blocks would be useful. Such a library might contain various model space metrics, non-smoothness operators, operators for representing steady-state data, stability etc.

4. More work is needed to develop practical engineering procedures for the coding of prior knowledge into the criterion and constraints. One can imagine a high-level expert system based user interface that automatically codes prior knowledge like "We have good confidence in the default model when the pH is either high or low, but no confidence in it for intermediate pH values" into the optimization criterion. This elevates the modeling problem to an even higher level where the optimization formulation is hidden for the user.

It is commonly stated that the main benefit of "modern" non-linear empirical modeling approaches like radial basis-functions, wavelets, and neural networks is that essentially no prior knowledge is required. This is to some extent true, but one must remember that the price to pay may be uncertainty about what set of basis-functions to apply, and a more or less ill-conditioned identification problem, which will typically lead to poor extrapolation capabilities of the model, cf. the simulation example in section 6. Unfortunately, it has also been common practise that the lack of transparency of such black-box approaches has led to the available prior knowledge being ignored. However, simple regularization methods like stopped training (Sjöberg & Ljung 1992*b*, Sjöberg, McKelvey & Ljung 1993), default models (Thompson & Kramer 1994, Su, Bhat & McAvoy 1992, Kramer et al. 1992, Johansen & Foss 1992, Sjöberg & Ljung 1992*a*), constraints (Joerding & Meador 1991, Thompson & Kramer 1994, Suykens, De Moor & Vandewalle 1994), penalty on parameter magnitude (Weigend, Huberman & Rumelhart 1990), and smoothness regularization (Bishop 1991) has recently been suggested. The relationship to Bayesian estimation is also being pursued (MacKay 1991, Williams 1995). The optimization framework presented here will be useful for regularizing such complex parameter identification problems, and is useful as a complementary technique applied together with parametric approaches such as neural networks, in order to reduce the sensitivity with respect to the a priori choice of model structure.

8 Conclusions

We have formulated the semi-empirical modeling problem as an optimization problem on a function space. In addition to the standard penalty on mismatch between model prediction and the empirical data, the criterion also penalizes mismatch between the model and various pieces of prior knowledge. Furthermore, prior knowledge can enter as constraints in the optimization problem. We have solved a simplified version of the optimization problem explicitly in the function space, and discussed finite-dimensional semi-parametric approximation to the general non-parametric solution. The approach is interpreted in the context of regularization, and it is shown how different pieces of prior knowledge will influence the effective number of parameters in the model. The problem of weighting the various pieces of knowledge and data is approached similarly to the problem of choosing a regularization parameter, and some criteria are suggested.

Our work is an attempt to provide a unified framework for incorporation of prior knowledge in non-linear system identification. It is primarily based on (Thompson & Kramer 1994, Girosi et al. 1994). In particular, the framework provides an attractive alternative to the direct specification of a parametric model structure, since the formulation of an optimization criterion is clearly a more transparent and engineering-friendly procedure that requires less guesswork.

Appendix A

Here we describe the experimentally verified pH neutralization tank model developed by Hall & Seborg (1989). It is based on the assumptions of perfect mixing, constant density, fast reactions and completely soluble ions. Only the following chemical reactions are modeled

$$
\begin{aligned}
H_2O &\rightleftarrows OH^- + H^+ \\
H_2CO_3 &\rightleftarrows HCO_3^- + H^+ \\
HCO_3^- &\rightleftarrows CO_3^{2-} + H^+
\end{aligned}
$$

because HNO_3 is a strong acid and NaOH is a strong base. Chemical reaction invariants for the process are (Gustafsson & Waller 1983)

$$
\begin{aligned}
W_a &= [H^+] - [OH^-] - [HCO_3^-] - 2[CO_3^{2-}] \\
W_b &= [H_2CO_3] + [HCO_3^-] + [CO_3^{2-}]
\end{aligned}
$$

Using the equilibrium equations $K_{a1} = [HCO_3^-][H^+][H_2CO_3]^{-1}$, $K_{a2} = [CO_3^{2-}][H^+][HCO_3^-]^{-1}$, and $K_w = [H^+][OH^-]$, an implicit equation for $[H^+]$ is found

$$
W_a = [H^+] - \frac{K_w}{[H^+]} - W_b \frac{K_{a1}/[H^+] + 2K_{a1}K_{a2}/[H^+]^2}{1 + K_{a1}/[H^+] + K_{a1}K_{a2}/[H^+]^2}
$$

Solving this equation for $[H^+]$, we can find pH $= -\log_{10}[H^+]$. A total mass balance of the tank gives

$$
A\dot{h} = Q_1 + Q_2 + Q_3 - c\sqrt{h - h_0}
$$

where c is a valve constant, A is the tank cross-section area, h is the freely varying tank level, and h_0 the vertical distance from the bottom of the tank to the outlet. Component balances gives

$$
\begin{aligned}
hA\dot{W}_a &= Q_1(W_{a1} - W_a) + Q_2(W_{a2} - W_a) + Q_3(W_{a3} - W_a) \\
hA\dot{W}_b &= Q_1(W_{b1} - W_b) + Q_2(W_{b2} - W_b) + Q_3(W_{b3} - W_b)
\end{aligned}
$$

where W_{ai} and W_{bi} are chemical reaction invariants of the i-th stream. The variables are defined in Table 1.

Appendix B

Here we describe a very simple model of the tank, based on neglecting the buffer stream in the model. Hence, the streams are

- Influent acid stream Q_1 (HNO_3)
- Influent base stream Q_3 (NaOH)
- Effluent stream Q_4

The only reaction considered in this model is

$$H_2O \rightleftarrows OH^- + H^+$$

A reaction invariant is $W = [H^+] - [OH^-]$, and an implicit equation for $[H^+]$ is

$$W = [H^+] - K_w/[H^+] \tag{20}$$

The mass balances are

$$
\begin{aligned}
A\dot{h} &= Q_1 + Q_3 - c\sqrt{h - h_0} && (21) \\
hA\dot{W} &= Q_1(W_1 - W) + Q_3(W_3 - W) && (22)
\end{aligned}
$$

Acknowledgments

This work was supported by The Research Council of Norway under doctoral scholarship grant no. ST. 10.12.221718. The author is grateful to Mike Thompson at the Massachusetts Institute of Technology, Aage V. Sørensen and Prof. Bjarne A. Foss at the Norwegian Institute of Technology for their comments on various parts of this work. Moreover, this work has benefited from discussions with Jan Larsen at the Technical University of Denmark.

References

Adams, R. A. (1975), *Sobolev spaces*, Adademic Press, New York.

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

Bishop, C. (1991), 'Improving the generalization properties of radial basis function neural networks', *Neural Computation* **3**, 579–588.

Breiman, L. (1993), 'Hinging hyperplanes for regression, classification, and function approximation', *IEEE Trans. Information Theory* **39**, 999–1013.

Demuth, H. & Beale, M. (1993), *Neural Network Toolbox User's Guide (MATLAB)*, The MathWorks, Inc.

Dyn, N. (1989), Interpolation and approximation by radial and related functions, *in* C. K. Chui, L. L. Schumaker & J. D. Ward, eds, 'Approximation Theory IV', Vol. 1, Academic Press, Inc., pp. 211–234.

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

Girosi, F., Jones, M. & Poggio, T. (1994), Priors, stabilizers and basis functions: From regularization to radial, tensor and additive splines, Technical Report AI Memo 1430, MIT, Cambridge.

Girosi, F., Jones, M. & Poggio, T. (1995), 'Regularization theory and neural network architectures', *Neural Computation* **7**, 219–269.

Grace, A. (1990), *Optimization Toolbox User's Guide (MATLAB)*, The MathWorks, Inc.

Gustafsson, T. K. & Waller, K. V. (1983), 'Dynamic modeling and reaction invariant control of pH', *Chemical Engineering Science* **38**, 389–398.

Hall, R. C. & Seborg, D. E. (1989), Modelling and self-tuning control of a multivariable pH neutralization process. Part I: Modelling and multiloop control, *in* 'Proceedings American Control Conference, Pittsburgh', Vol. 2, pp. 1822–1827.

Joerding, W. H. & Meador, J. L. (1991), 'Encoding a prior information in feedforward networks', *Neural Networks* **4**, 847.

Johansen, T. A. & Foss, B. A. (1992), Representing and learning unmodeled dynamics with neural network memories, *in* 'Proceedings of the American Control Conference, Chicago, Il.', pp. 3037–3043.

Johansen, T. A. & Weyer, E. (1995), Model structure identification using separate validation data - asymptotic properties. To be presented at the European Control Conference, Rome.

Kramer, M. A., Thompson, M. L. & Phagat, P. M. (1992), Embedding theoretical models in neural networks, *in* 'Proceedings American Control Conference, Chicago, Il.', pp. 475–479.

Larsen, J. & Hansen, L. K. (1994), Generalization performance of regularized neural network models, *in* 'Proc. IEEE Workshop on Neural Networks for Signal Processing, Ermioni, Greece'.

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

Luenberger, D. G. (1969), *Optimization by Vector Space Methods*, John Wiley.

Luenberger, D. G. (1984), *Introduction to Linear and Nonlinear Programming, 2nd Ed.*, Addison-Wesley, Inc., Reading, MA.

MacKay, D. J. C. (1991), Bayesian Methods for Adaptive Models, PhD thesis, California Institute of Technology, Pasadena, CA.

Madych, W. R. & Nelson, S. A. (1990), 'Multivariate interpolation and conditionally positive definite functions. II', *Mathematics of Computation* **54**, 211–230.

Mosca, E. (1970), System identification by reproducing kernel hilbert space methods, *in* 'Preprints 2nd IFAC Symposium on Identification and Process Parameter Estimation, Prague', p. Paper 1.1.

O'Sullivan, F. (1986), 'A statistical perspective on ill-posed inverse problems', *Statistical Science* **1**, 502–527.

Sjöberg, J. & Ljung, L. (1992*a*), A comment of "leakage" in adaptive algorithms, *in* 'Preprints IFAC Symposium on Adaptive Systems in Control and Signal Processing, Grenoble, France'.

Sjöberg, J. & Ljung, L. (1992*b*), Overtraining, regularization, and searching for minimum in neural networks, *in* 'Preprints IFAC Symposium on Adaptive Systems in Control and Signal Processing, Grenoble, France', pp. 669–674.

Sjöberg, J., McKelvey, T. & Ljung, L. (1993), On the use of regularization in system identification, *in* 'Preprints 12th IFAC World Congress, Sydney', Vol. 7, pp. 381–386.

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

Stoica, P., Eykhoff, P., Janssen, P. & Söderström, T. (1986), 'Model-structure selection by cross-validation', *Int. J. Control* **43**, 1841–1878.

Su, H.-T., Bhat, N. & McAvoy, T. J. (1992), Integrated neural networks with first principles models for dynamic modeling. In Preprints IFAC DYCORD+ '92, College Park, Maryland.

Suykens, J. A. K., De Moor, B. L. R. & Vandewalle, J. (1994), 'Static and dynamic stabilizing neural controllers, applicable to transition between equilibrium points', *Neural Networks* **7**, 819–831.

Tanaka, Y., Fukushima, M. & Ibaraki, T. (1988), 'A comparative study of several semi-infinite nonlinear programming algorithms', *European Journal of Operational Research* **36**, 92–100.

Thompson, M. L. & Kramer, M. A. (1994), 'Modeling chemical processes using prior knowledge and neural networks', *AIChE J.* **40**, 1328–1340.

Tikhonov, A. N. & Arsenin, V. Y. (1977), *Solutions of Ill-posed Problems*, Winston, Washington DC.

Tulleken, H. J. A. F. (1993), 'Grey-box modelling and identification using physical knowledge and bayesian techniques', *Automatica* **29**, 285–308.

Wahba, G. (1990), *Spline Models for Observational Data*, SIAM, Philadelphia.

Weigend, A. S., Huberman, B. A. & Rumelhart, D. E. (1990), 'Predicting the future: A connectionist approach', *International Journal of Neural Systems* **1**, 193–209.

Williams, P. M. (1995), 'Bayesian regularization and pruning using a Laplace operator', *Neural Computation* **7**, 117–143.

Table 1: Symbols, constants and variables used in models.

Symbol	Variable	Nominal value
A	Tank area	207 cm^2
h	Tank level	14 cm
h_0	Tank outlet level	5 cm
Q_1	Acid flow-rate	16.6 ml/s
Q_2	Buffer flow-rate	0.55 ml/s
Q_3	Base flow-rate	15.6 ml/s
$[HNO_3]_1$	Acid concentration in acid stream	0.003 molar
$[NaHCO_3]_3$	Buffer concentration in base stream	0.00005 molar
$[NaOH]_3$	Base concentration in base stream	0.003 molar
$[NaHCO_3]_2$	Buffer concentration in buffer stream	0.03 molar
c	Valve constant	8 $ml/s\sqrt{cm}$
pK_{a1}	$-\log_{10} K_{a1}$	6.35
pK_{a2}	$-\log_{10} K_{a2}$	10.33
pK_w	$-\log_{10} K_w$	14.00
W_{a1}	$[HNO_3]_1$	0.003 molar
W_{a2}	$-[NaHCO_3]_2$	-0.03 molar
W_{a3}	$-[NaHCO_3]_3 - [NaOH]_3$	-0.00305 molar
W_{b1}		0
W_{b2}	$[NaHCO_3]_2$	0.03 molar
W_{b3}	$[NaHCO_3]_3$	0.00005 molar
W_1	$[HNO_3]_1$	0.003 molar
W_2		0
W_3	$-[NaOH]_3$	-0.003 molar

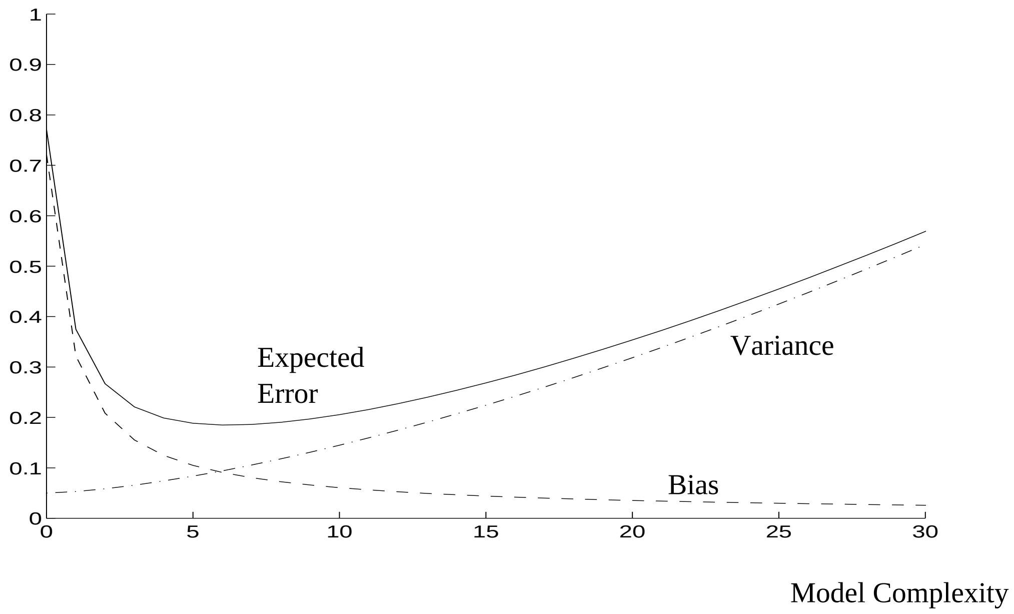

Figure 1: Typical relationship between model bias, variance and model complexity, when it is assumed that model complexity can be measured by a real number and the data sequence used for identification has fixed length. Of course, the relationship is greatly simplified, since two different models of exactly the same complexity may have different bias and variance.

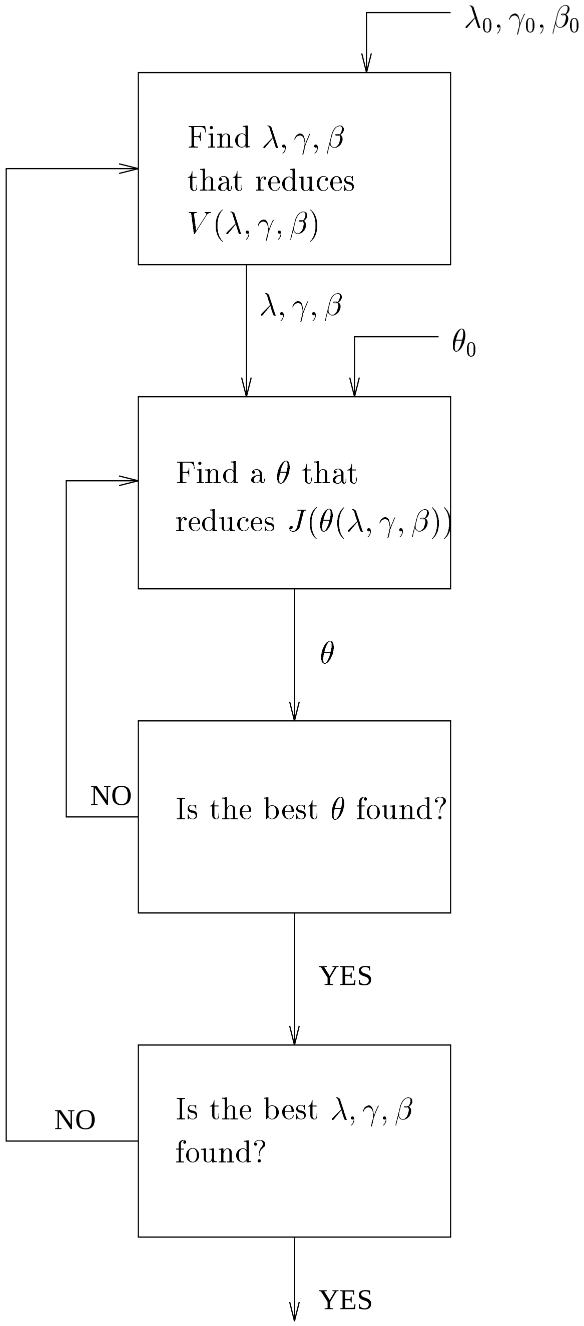

Figure 2: Hierarchical optimization approach. The inner loop can be viewed as parameter identification, while the outer loop can be viewed as a kind of structure identification as it estimates the optimal model complexity.

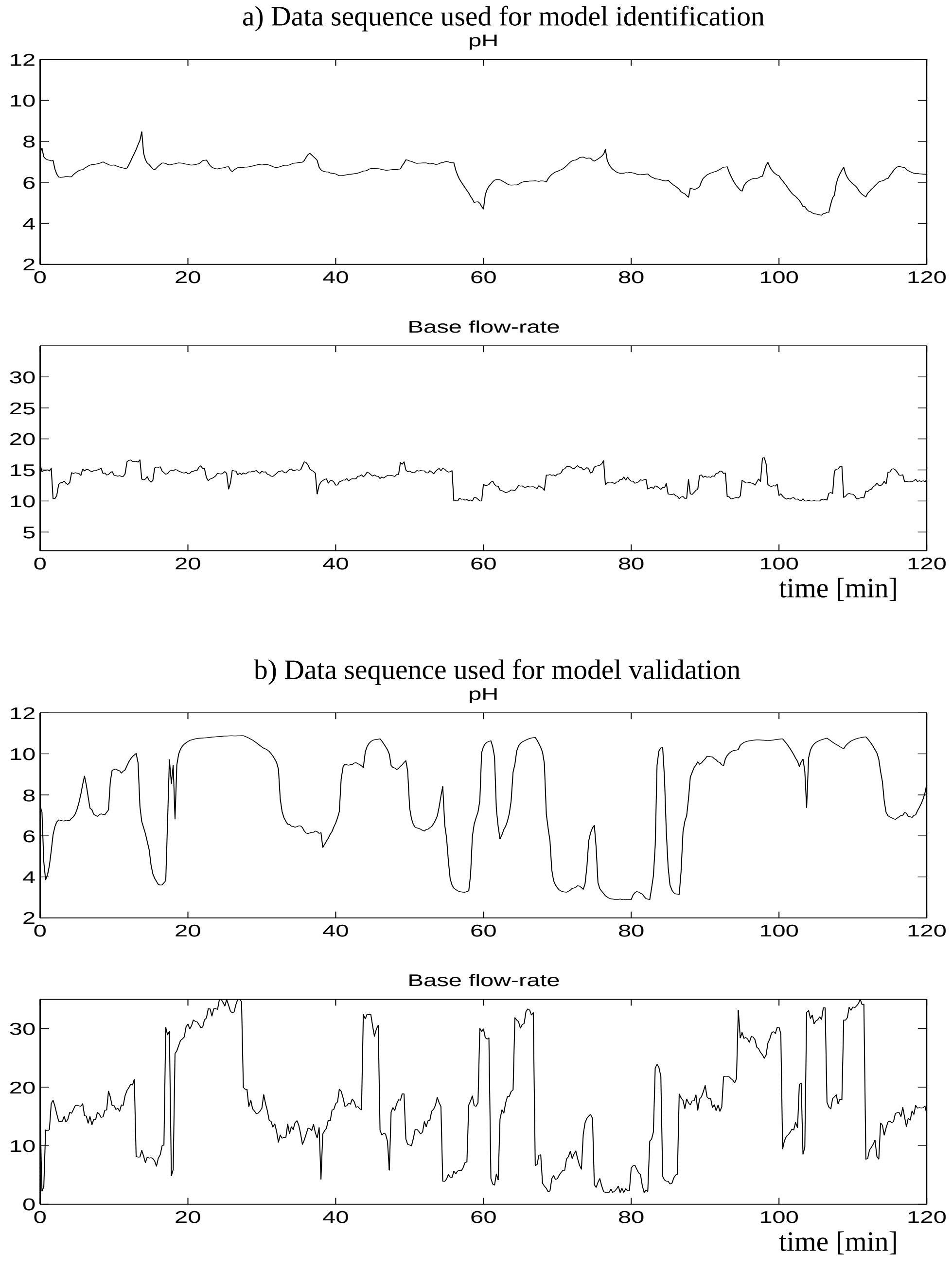

Figure 3: Simulated data sequences.

Tor A. Johansen was born in Steinkjer, Norway. He received his Dr.Ing. (PhD) and Siv.Ing. (M.Sc) degrees in electrical and computer engineering from the Norwegian Institute of Technology (NTH), Trondheim, in 1994 and 1989, respectively. He is current employed by SINTEF Automatic Control as a research scientist. From 1991 to 1994 he was a graduate student at the Norwegian Institute of Technology. During 1992 he was at the Department of Electrical Engineering – Systems at the University of Southern California as a research visitor. In 1990 he was at the Norwegian Defence Research Establishment (FFI) at Kjeller. His research interests include model representation, empirical and semi-mechanistic modeling techniques for industrial processes, system identification, and non-linear model-based and adaptive control.

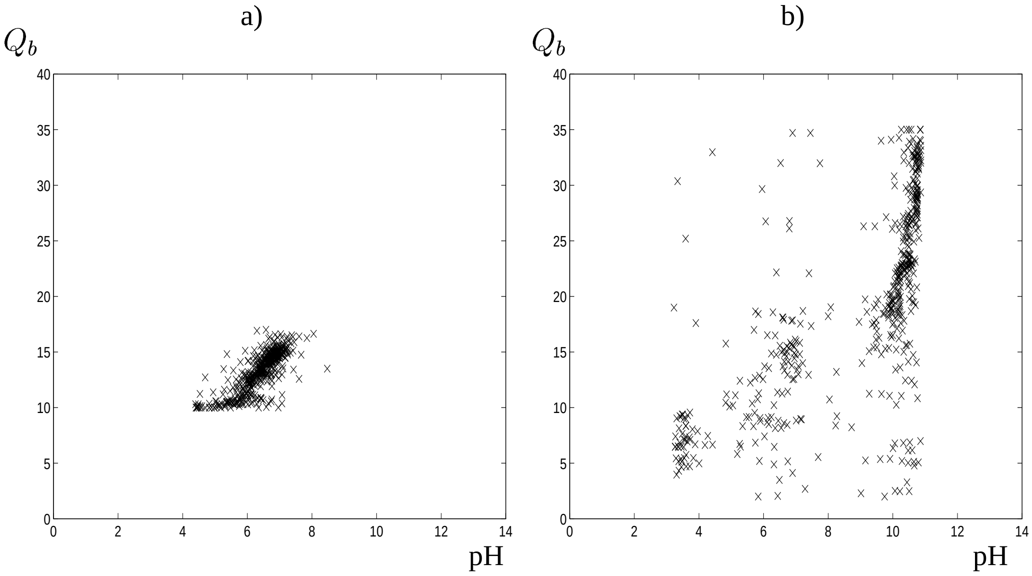

Figure 4: Empirical distribution of a) identification data, and b) validation data.

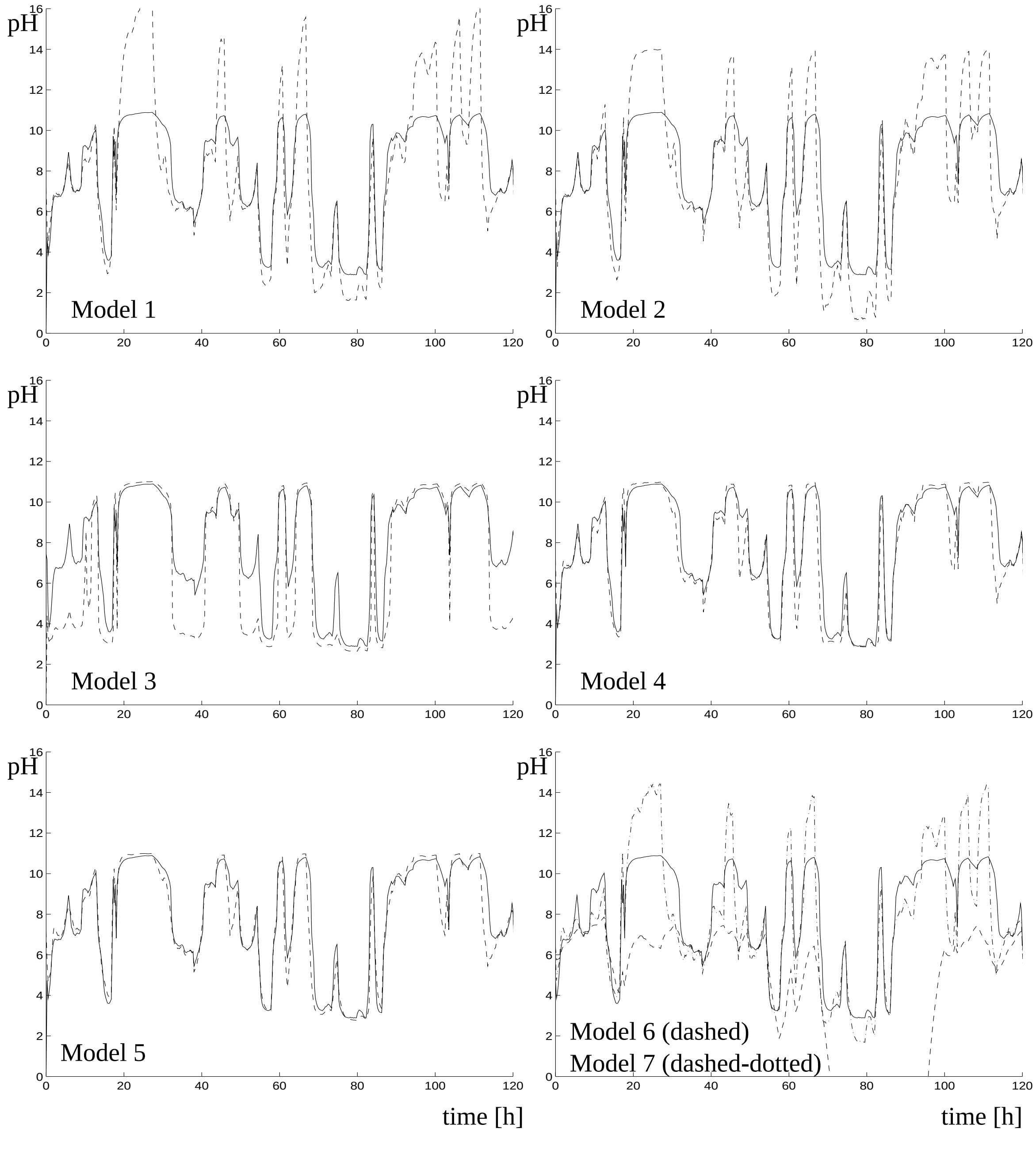

Figure 5: Simulation of the 7 models when the validation data input sequence is applied to the models. The solid curves are the validation data pH, while the dashed curves are simulated pH using the models. For models 6 and 7, the dashed and dashed-dotted curves are with different initial parameter estimates.

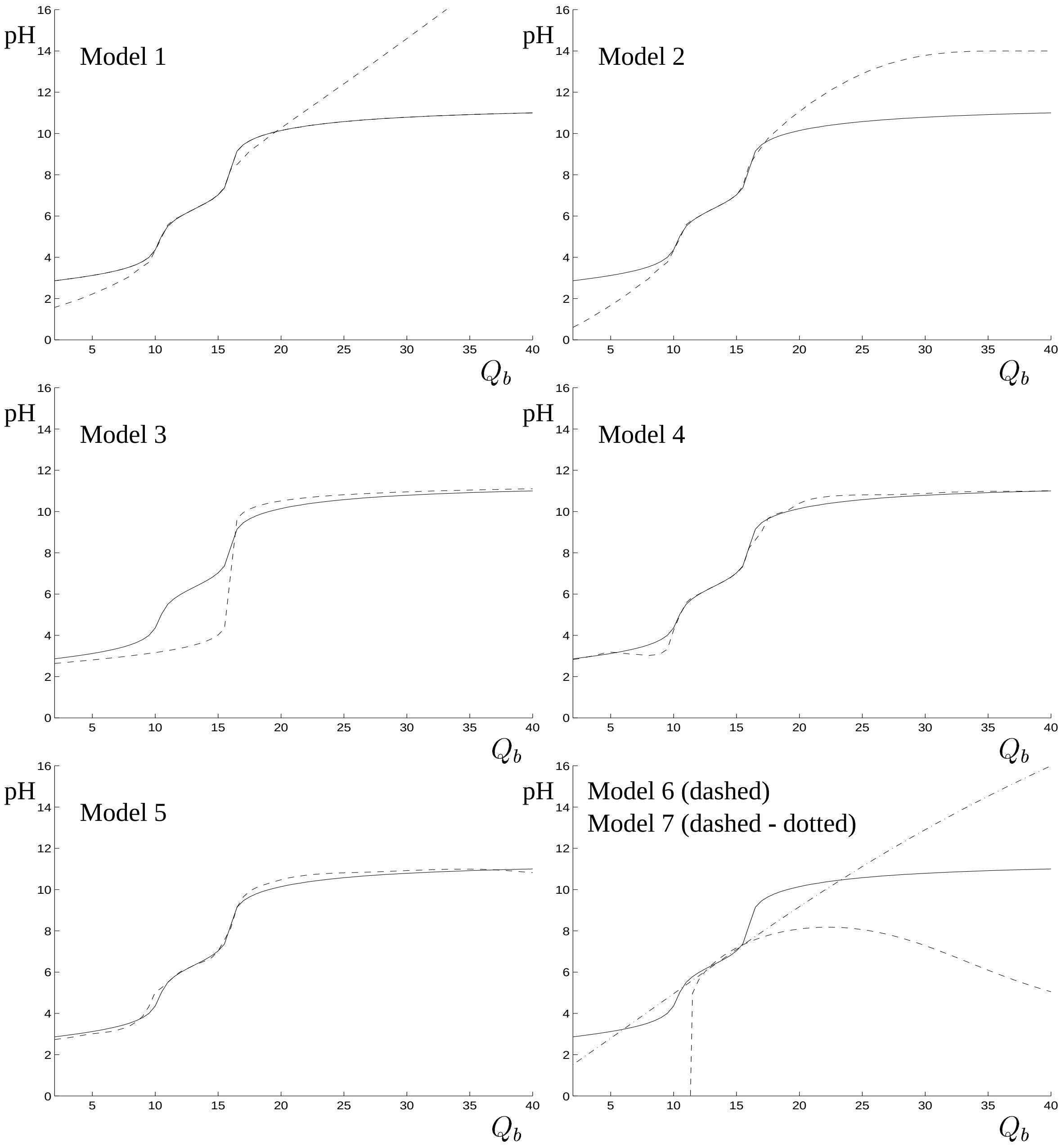

Figure 6: Titration curves (steady state response) of the 7 models. The solid curves are the system's true titration curve, and the dashed curves are the models. For models 6 and 7, the dashed and dashed-dotted curves are with different initial parameter estimates.