MULTI-OBJECTIVE IDENTIFICATION OF FIR MODELS

Tor A. Johansen [*]

[*] *Department of Engineering Cybernetics, Norwegian University of Science and Technology, N-7491 Trondheim, Norway. Email: Tor.Arne.Johansen@itk.ntnu.no*

Abstract: Identification of model parameters can be viewed as a problem with multiple objectives and constraints derived from empirical data (dynamic and steady-state), physical models and belief, empirical and qualitative belief, desired model properties etc. A fairly general approach to multi-objective system identification based on constrained optimization is suggested, and here we formalize the method for the identification of FIR models. Particular attention is paid to the analysis and selection of tradeoffs between conflicting objectives and constraints.

Keywords: Optimization, Parameter Estimation, Prior Knowledge, Linear Systems.

1. INTRODUCTION

The main objective of system identification is to identify a model with good prediction capabilities in the sense that it is able to accurately predict the system's response to a given class of excitations. Hence, a common identification objective is to minimize some penalty on mismatch between model prediction and observed data, which is the underlying principle of the least squares method, maximum likelihood and the prediction error methods, see e.g. (Söderström and Stoica 1988). However, in practical system identification it is often desirable to introduce additional objectives and constraints into the identification problem. There are several reasons for this:

(1) The data sequence used for identification may be incomplete or uncertain. Hence, it contains information only about certain aspects of the system, like a certain frequency interval or a certain region of the state space, while it may be desired to apply the model to a wider class of excitation signals or operating conditions.
(2) The model structure need not be identifiable, i.e. there may be several parameter vectors that yield the same model predictions for any excitations. Then there are excessive degrees of freedom available in the model that may be applied to meet different objectives.
(3) There is often additional information available. It makes sense to use this information to improve the accuracy and validity of the model. In general, any additional information will reduce the variance of the model due to uncertain data. Such information may be qualitative, such as stability, or it may be quantitative, such as steady-state data or explicit knowledge about the model response derived from experiments or other models.
(4) There may be some properties that we may want the model to have, for certain reasons. For example, regularity (or smoothness) of the model is usually desirable both for variance reduction reasons and because is provides advantages for model application.

The potential benefits with multi-objective system identification is better models, and reduced model development costs because less experimental data may be needed, or data from normal operation could replace data from designed experiments.

Here a quite general framework for multi-objective model identification based on multi-objective optimization is presented, and in particular its application to the identification of FIR models. The motivation for the special attention given to the FIR model representation is twofold; it is simple to treat analytically, and it is of great practical importance, for example as a basis for model predictive control (Qin and Badgwell 1996).

The multi-objective approach to system identification is certainly not new, but its potential has not been fully explored and exploited. Single constraints or penalties on the parameter space have been suggested, for example to ensure stability (Tulleken 1993, Johansen 1996), convexity of optimizing control criterion (Foss and Johansen 1997), fulfillment of balance equations and steady-state data (Thompson and Kramer 1994), and explicit belief about parameter values (Moons and De Moor 1995, Pages *et al.* 1996). Furthermore, regularization is sometimes applied to reduce the variance of the identified model (Tikhonov and Arsenin 1977, Larsen and Hansen 1994, Johansen 1997*b*, Dayal and MacGregor 1996). In (Eskinat 1995), it is also shown that constrained identification has the same variance-reducing effect. Moreover, penalties derived from explicit belief about parameters (Sjöberg *et al.* 1993) and other a prior known model (Thompson and Kramer 1994, Johansen 1996) have also been suggested. Prior knowledge in terms of constraints and penalties can be implemented directly in a prediction error method (PEM) framework, (Johansen 1996, Johansen 1997*a*), or the penalties can be reformulated into equivalent prior distributions in a Bayesian system identification framework (Peterka 1981, Tulleken 1993).

The FIR (finite impulse response) model is

$$\begin{pmatrix} y_1(t) \\ \vdots \\ y_m(t) \end{pmatrix} = \sum_{k=0}^{T} \begin{pmatrix} h_{11}^k & \cdots & h_{1r}^k \\ \vdots & & \vdots \\ h_{m1}^k & \cdots & h_{mr}^k \end{pmatrix} \begin{pmatrix} u_1(t-k) \\ \vdots \\ u_r(t-k) \end{pmatrix}$$

where it is assumed that the input vector $u(t) \in R^r$ and output vector $y(r) \in R^m$ both have zero mean. Furthermore, it is assumed that the inputs and outputs are properly scaled. The system identification problem is to determine the Markov parameters h_{ij}^k. Let θ be a vector containing all the model parameters (Markov parameters) that are to be chosen to meet the given objectives and constraints.

2. OBJECTIVES AND CONSTRAINTS

In this section we describe some model objectives and constraints, and suggest possible mathematical formulations. The list is by no means complete, and one can think of numerous variations.

2.1 *Time-series data*

The most common objective (and usually the only explicit one) in system identification is that the model should have good prediction capabilities, in some sense. This objective is typically implemented as a least squares penalty on the mismatch between the model prediction and the measured output, using a sequence of input and output data $(u(1), ..., u(N))$ and $(y(1), ..., y(N))$:

$$V_{\text{ts-data}}(\theta) = \frac{1}{Nm} \sum_{t=1}^{N} \sum_{i=1}^{m} \left(y_i(t) - \sum_{k=0}^{T} \sum_{j=1}^{r} h_{ij}^k u_j(t-k) \right)^2$$

This equation can be written in the form

$$V_{\text{ts-data}}(\theta) = \frac{1}{Nm} \left(Y - \Phi\theta\right)^T \left(Y - \Phi^T\theta\right)$$

where the Markov parameters are stacked in the $p \times 1$ vector θ, where $p = (T+1)mr$, the output data is stacked in the $N \times 1$ vector Y, and the input data is properly organized in the $N \times p$ matrix Φ.

2.2 *Steady-state gain*

The steady-state gain of the system is often of great importance, for example when the model is used in a controller that do not contain integral action. The steady-state gain is typically found by a step-response experiment or from a steady-state model/simulator. Another objective may therefore be that the model's steady-state gain matrix should match some given $m \times r$ matrix G, i.e. $y = Gu$ at steady state. This can be implemented as a set of constraints

$$\Delta \sum_{k=1}^{T} h_{ij}^k = g_{ij}$$

for $i = 1, .., m$ and $j = 1, ..., r$, or penalty

$$V_{\text{ss-gain}}(\theta) = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \left(g_{ij} - \Delta \sum_{k=0}^{T} h_{ij}^k \right)^2$$

where g_{ij} is the ith element of the matrix G, and Δ is the sampling interval. It is assumed that ΔT is much larger than the largest time-constant of the system. The constraints and penalty can be written in matrix form

$$G_{\text{ss-gain}}(\theta) = A_g\theta - b_g = 0$$

or

$$V_{\text{ss-gain}}(\theta) = (A_g\theta - b_g)^T (A_g\theta - b_g)$$

where A_g is a $mr \times p$ matrix with zeros and ones in appropriate places, and the $mr \times 1$ vector b_g containts the elements of G.

2.3 *Regularity*

Regularity properties (such as smoothness) of the impulse response are usually desirable, but should not be expected from the standard least squares estimation algorithm when there are many parameters (T is large). Another objective may therefore be that the impulse response function is a smooth funciton of time. This can be implemented as a penalty on an approximation to the squared second derivative of the impulse response function:

$$V_{\text{regl}}(\theta) = \frac{1}{mr(T-1)} \cdot \sum_{i=1}^{m} \sum_{j=1}^{r} \sum_{k=1}^{T-1} \left(\frac{h_{ij}^{k-1} - 2h_{ij}^{k} + h_{ij}^{k+1}}{\Delta^2} \right)^2 w_{ij}(k)$$

where Δ is the sampling interval, and w_{ij} is a weighting function. It is easy to see that this can be written in the matrix form

$$V_{\text{regl}}(\theta) = \theta^T S^T S \theta$$

where S is a positive definite $p \times p$ matrix.

Regularity penalties are often imposed the the reason of reducing the identified model's variance. It can be seen that regularization (or any penalty or active constraint) will in general reduce the model's effective number of degrees of freedom, which as a direct consequence reduces the model's variance (Johansen 1997*b*).

2.4 *Vanishing tail*

The impulse response of stable processes are characterized by a vanishing tail. This objective can be implemented as

$$V_{\text{tail}}(\theta) = \sum_{i=1}^{m} \sum_{j=1}^{r} \sum_{k=0}^{T} \left(h_{ij}^{k} \right)^2 \rho_{ij}(k)$$

where ρ_{ij} is a weighting function. In matrix form, this can be written as

$$V_{\text{tail}}(\theta) = \theta^T Q^T Q \theta$$

for some $p \times p$ matrix Q.

2.5 *Parameter bounds*

Some knowledge of the impulse response may be available, and an objective may be that the identified impulse reponse shold match this knowledge. This knowledge may originate from simulation of rigorous mathematical models, observed behavior of the system or the belief of engineers. Some quite general implementations of such knowledge are as upper and lower constraints on the impulse response

$$\underline{h}_{ij}^{k} \leq h_{ij}^{k} \leq \overline{h}_{ij}^{k}$$

for $i = 1, ..., m$, $j = 1, ..., r$, $k = 0, ..., T$, or as penalty on deviation from an a priori given impulse reponse function $\tilde{h}_{ij}^{k}$

$$V_{\text{prior}}(\theta) = \frac{1}{mr(T+1)} \sum_{i=1}^{m} \sum_{j=1}^{r} \sum_{k=0}^{T} \left(h_{ij}^{k} - \tilde{h}_{ij}^{k} \right)^2 \sigma_{ij}(k)$$

where σ_{ij} is a weighting function. These implementations can be written in matrix form

$$H_p(\theta) = A_p \theta - b_p \leq 0$$

for some $2p \times p$ matrix A_p and $2p \times 1$ vector b_p, or

$$V_{\text{prior}}(\theta) = (\theta - \tilde{\theta})^T W^T W (\theta - \tilde{\theta})$$

respectively. W is a $p \times p$ weighting matrix, and $\tilde{\theta}$ constain the parameters of the a priori impulse response function matrix $\tilde{h}$.

Quite often, model knowledge is available in the frequency domain as belief or partial knowledge about the system's frequency response matrix, see also (Eskinat 1995). The frequency response matrix $H(e^{j\omega})$ is related to the Markov parameters by the discrete Fourier transform

$$H_{ij}(e^{j\omega}) = \sum_{k=0}^{\infty} h_{ij}^{k} e^{-j\omega k}$$

Thus, if $H_{ij}(e^{-j\omega_l})$ is known for certain $\omega_1, ..., \omega_M$ we get the following pairs of constraints on the Markov parameters:

$$\Re \left(H_{ij}(e^{-j\omega_l}) \right) = \sum_{k=0}^{T} h_{ij}^{k} \cos(\omega_l k)$$

$$\Im \left(H_{ij}(e^{-j\omega_l}) \right) = -\sum_{k=0}^{T} h_{ij}^{k} \sin(\omega_l k)$$

for $l = 1, 2, ..., M$. Similar constraints can be constructed if upper or lower bounds on the frequency response matrix is known.

2.6 *Absolute noise bound*

If it is assumed that an upper bound on the uncertainty on each output variable is known, say $\overline{n}_i$, one may use this to impose the following constraints on the parameters:

$$y_i(t) - \overline{n}_i \leq \sum_{k=0}^{T} \sum_{j=1}^{r} h_{ij}^{k} u_j(t-k) \ \leq \ y_i(t) + \overline{n}_i$$

This is similar to set membership identification.

3. MULTI-OBJECTIVE OPTIMIZATION

Let a set of scalar objective functions $V_1, ..., V_n$, scalar inequality constraints $F_1, ..., F_l$ and equality constraints $H_1, ..., H_q$ be given. These can be any mixture of the objectives and criteria in section 2, or different ones. Notice that all the penalties in section 2 are quadratic and convex and all constraints are linear. In this case, the optimization problem is a convex quadratic program.

3.1 *Single Weighted Objective*

There are several approaches to multi-objective optimization in the literature. One simple and common approach is to formulate a single weighted objective function

$$V(\theta; \gamma_1, ..., \gamma_n) = \gamma_1 V_1(\theta) + ... + \gamma_n V_n(\theta)$$

on the basis of the multiple objective functions. The weighting parameters are required to satisfy $\gamma_1, ..., \gamma_n \geq 0$. This leads to the optimization problem

$$\min_{\theta} V(\theta; \gamma_1, ..., \gamma_n) \tag{1}$$

subject to

$$F_1(\theta) \leq 0, \ ..., \ F_l(\theta) \leq 0 \tag{2}$$

$$H_1(\theta) = 0, \ ..., \ H_q(\theta) = 0 \tag{3}$$

Hence, the relative weighting of the objectives must be defined a priori in terms of the weighting parameters. The multiple goals and constraints will typically be conflicting, to some degree. For example, the best fit to the time-series data may be in some conflict with the preference for smoothness of the impulse response. Also, prior belief about the impulse response (implemented as a penalty on deviation from the prior guess) may not always be consistent with the best fit to the time-series data (either due to unmodelled phenomena or incorrect prior knowledge). Some such conflicts may be acceptable, while others may indicate serious deficiencies in the data, models or prior belief/knowledge. Moreover, the ideal tradeoff between the various objectives and constraints is typically not known a priori. Multi-objective identification is therefore a highly interactive session where in each iteration the user will modify the weighting parameters and possibly also the objectives and constraints themseleves to get a better model.

To understand the tradeoffs the user faces in multi-objective system identification, consider the Karush-Kuhn-Tucker conditions (Luenberger 1984) that characterizes an optimal solution $\theta^\star$ (assuming it exists) to the problem (1)-(3) for a given $\gamma_1, ..., \gamma_n$:

$$\sum_{i=1}^{n} \gamma_i \frac{dV_i}{d\theta}(\theta^\star) + \sum_{i=1}^{l} \frac{dF_i}{d\theta}(\theta^\star)\mu_i + \sum_{i=1}^{q} \frac{dH_i}{d\theta}(\theta^\star)\lambda_i = 0 \tag{4}$$

$$\mu_i^T F(\theta^\star) = 0, \quad \mu_i \geq 0, \ i = 1, ..., l \tag{5}$$

The Lagrange multipliers $\lambda_1, ..., \lambda_q$ and $\mu_1, ..., \mu_l$ are uniquely defined by an optimal solution θ^*. Eq. (4) has some interesting interpretations. The different terms in (4) characterizes the tradeoff between the conflicting objectives and constraints. If a term is zero, it means that $\theta^\star$ is the optimal solution for the corresponding single objective in addition to being the optimal solution to the multi-objective problem. This means that this objective or constraint is in no conflict with the other objectives and constraints. On the other hand, a non-zero term indicates that improved fulfillment of that single objective can be achieved by sacrificing attainment of other objectives. A simple geometric interpretation is illustrated in Fig. 1. Each term in (4) is a vector pointing in the direction of largest increase for each single objective, or perpendicular to the constraint surface. The vectors sum to zero, since $\theta^\star$ is the optimum for the multiple objectives and constraints. The length of each vector can be interpreted as the degree of mismatch between the corresponding objective or constraint and the other objectives and constraints. This is useful information that can be used to detect inconsistencies in the process knowledge, data and desired properties, and also be used to tune the weighting parameters to get a sensible tradeoff between the objectives and constraints.

The direction of the vector also contain information about the tradeoff, but it is hard to interpret in higher dimensions. A compromise would be to consider some measure of directional match between any pair of vectors ξ_1 and ξ_2, such as

$$\phi = \cos^{-1}\left(\frac{\xi_1^T \xi_2}{||\xi_1||_2 \cdot ||\xi_2||_2}\right)$$

3.2 *Goal Programming*

Another common approach to multi-objective optimization is goal programming, where a quantitative goal ζ_i is associated with each objective function, and the following problem is solved

$$\min_{x,\theta} x$$

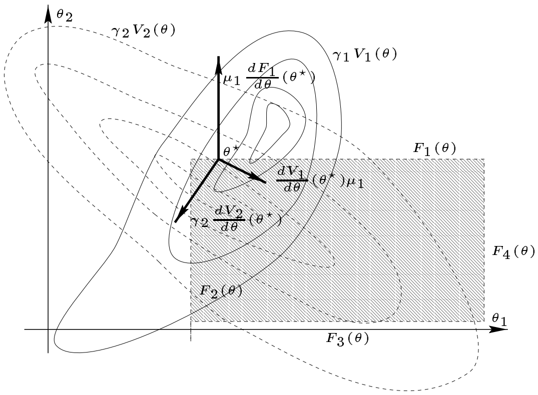

Fig. 1. Tradeoff between conflicting objectives and constraints. The solid curves are contour lines for the first objective, while the dashed curves are contour lines for the second objective. The shaded region represents an infeasible region of the parameter space.

subject to

$$V_1(\theta) - xs_1 \leq \zeta_1, \ \ldots, \ V_n(\theta) - xs_n \leq \zeta_n$$
$$F_1(\theta) \leq 0, \ \ldots, \ F_p(\theta) \leq 0$$
$$H_1(\theta) = 0, \ \ldots, \ H_q(\theta) = 0$$

The scalar parameters $s_1, \ldots, s_n \geq 0$ define the reward or penalty on over- and under-attainment of the goal. Minimizing the scalar dummy variable x leads to be best trade-off, in this sense.

4. ILLUSTRATIVE SIMULATION EXAMPLE

The following non-minimum phase linear system is simulated

$$h(s) = \frac{-2s + 1}{(s+3)(s+0.4)(s+0.2)}$$

A random input signal with 801 samples was generated by bandpass filtering white noise. The input was applied to this system, with a sampling frequency $f_s = 4$ Hz. The power spectra of the input and output signals are given in Figure 2. We observe that the experiment design is sub-optimal since there is very little low frequency excitation (the transfer function $h(s)$ has poles corresponding to frequencies well below 1 rad/s). Consequently, the identification results in Figure 3 with this data sequence show a FIR model with poor low-frequency accuracy using the pure least squares algorithm, as expected. It can also be seen that the results can be improved in various ways by utilizing additional prior knowledge to the data:

- Prior knowledge of the steady-state gain.
- Smoothness, regularization.
- Penalty on tail, stability.
- Upper and lower bounds on response.

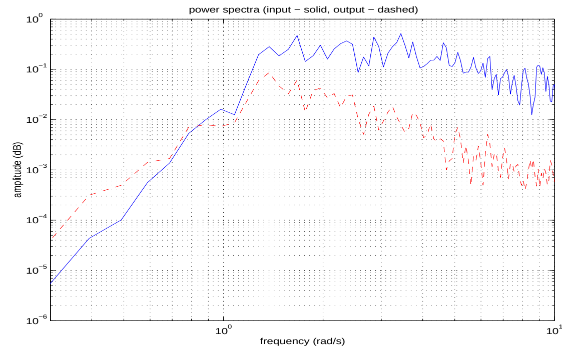

Fig. 2. Estimated power spectra of input (solid line) and output (dashed line) data.

The problem was solved by specifying weights on the different terms in the compostite obective function.

Consider the model identified on the mixture of all the objectives and the constraint. The contribution of the different terms in the Karush-Kuhn-Tucker equation (4) can be seen in Figure 4. We observe that only the time-series pentalty and the smoothness penalty have large contributions to (4). From this, we make the following conclusions:

- Since the terms corresponding to the steady-state gain and stability are very small, it is clear that the optimal model is consistent with these specifications. Furthermore, it can be concluded that there objectives are not in conflict with the evidence in the time series data or the smoothness objective.
- However, it is not possible to conclude that the assumed steady-state gain is correct, since another gain may still not be in conflict with the evidence in the data. In this case the assumed steady-state gain is actually correct, but using a different assumption here leads to a quite different impulse response model that is still consistent with the other pieces of data and belief. The reason for this is the lack of low-frequency information in the data, which means that the data are useless for the purpose of falsifying the low-frequency part of the model.
- The term correponding to the parameters bounds is nonzero, but fairly small. Since it is nonzero, the parameter bounds have a influence on the model. The term is small since the bounds constrains mainly the low-frequency part of the model, for which there is no evidence in the data, as discussed above.
- The time-series penalty and non-smoothness penalty balances each other and thus corresponds to conflicting objectives. Any further reduction of either of these objectives can only be achieved by accepting a higher degree of violation of one or more other objectives.

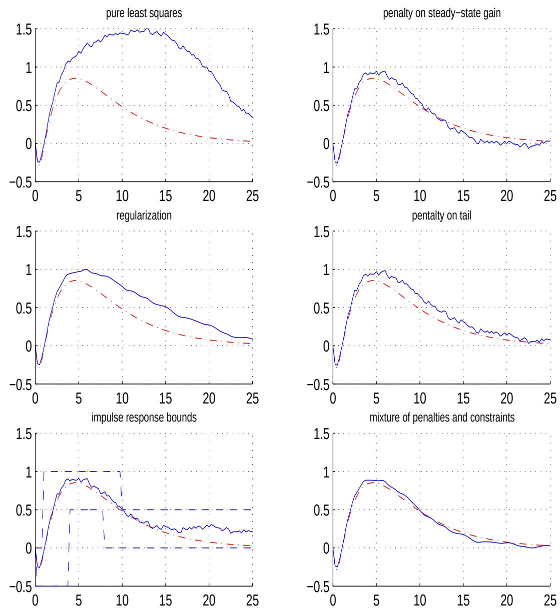

Fig. 3. Example: Identified FIR models. Dashed-dotted curves show the true system impulse response, while solid lines show the models' impulse responses.

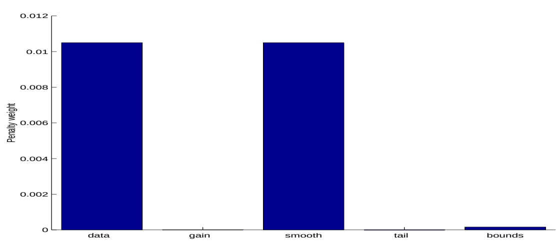

Fig. 4. The contribution of the different terms in the Karush-Kuhn-Tucker conditions (lengths of vector).

5. CONCLUDING REMARKS

An approach to identification of FIR models based on empirical data, prior knowledge and belief is suggested. The approach relies on multi-objective programming. The name of the game is thus to find an appropriate balance between our belief in the different pieces of data and knowledge in order to end up with a good model despite the fact that each on the pieces on data and belief may be incomplete or inaccurate. Rather than relying on objective statistical considerations (such as the bias/variance tradeoff (Johansen 1997*b*)), the method is subjective and driven by the engineer's choices. The different terms in the Karush-Kuhn-Tucker optimality conditions can be utilized to diagnose the model, i.e. to detect inconsistencies or consistency between different pieces of data and belief.

6. REFERENCES

Dayal, B. S. and J. F. MacGregor (1996). Identification of finite impulse response models: Methods and robustness issues. *Industrial and Engineering Chemistry Research* **35**, 4078–4090.

Eskinat, E. (1995). System identification using constrained estimation. In: *Proc. European Control Conference, Rome.* pp. 856–861.

Foss, B. A. and T. A. Johansen (1997). Identification and convexity in optimizing control. In: *Preprints IFAC Symposium on System Identification, Kitakyushu, Japan.* pp. 691–696.

Johansen, T. A. (1996). Identification of non-linear systems using empirical data and prior knowledge – An optimization approach. *Automatica* **32**, 337–356.

Johansen, T. A. (1997*a*). Constrained and regularized system identification. In: *Preprints IFAC Symposium on System Identification, Kitakyushu, Japan.* pp. 1467–1472.

Johansen, T. A. (1997*b*). On Tikhonov regularization, bias and variance in nonlinear system identification. *Automatica* **33**, 441–446.

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

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

Moons, C. and B. De Moor (1995). Parameter identification of induction motor drives. *Automatica* **31**, 1137–1147.

Pages, A., H. Pingaud, M. Meyer and X. Joulia (1996). From process simulation to general estimation. *Computers and Chemical Engineering* **20**, 1413–1424.

Peterka, V. (1981). Bayesian system identification. *Automatica* **17**, 41–53.

Qin, S. J. and T. A. Badgwell (1996). An overview of industrial model predictive control technology. Preprint CPC-V, Lake Tahoe.

Sjöberg, J., T. McKelvey and L. Ljung (1993). On the use of regularization in system identification. In: *Preprints 12th IFAC World Congress, Sydney.* Vol. 7. pp. 381–386.

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

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

Tikhonov, A. N. and V. Y. Arsenin (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.