Process Systems Engineering 2003
B. Chen and A.W. Westerberg (editors)

A Unified Framework for Systematic Model Improvement

Niels Rode Kristensen[a], Henrik Madsen[b] and Sten Bay Jørgensen[a]

[a]Department of Chemical Engineering, Technical University of Denmark, Building 229, DK-2800 Lyngby, Denmark

[b]Informatics and Mathematical Modelling, Technical University of Denmark, Building 321, DK-2800 Lyngby, Denmark

A unified framework for improving the quality of continuous time models of dynamic systems based on experimental data is presented. The framework is based on an interplay between stochastic differential equation (SDE) modelling, statistical tests and multivariate nonparametric regression. This combination provides systematic methods for pinpointing and repairing model deficiencies by uncovering their structural origin. The potential of the proposed framework in terms of modelling complex dynamic phenomena such as reaction kinetics is illustrated with a case study involving a model of a fed-batch bioreactor, where it is illustrated how an incorrectly modelled biomass growth rate can be pinpointed and an estimate provided of the functional relation needed to properly describe it.

1. INTRODUCTION

Dynamic process models are used in many areas of chemical engineering and for many different purposes. Dynamic model development is therefore an inherently purpose-driven act in the sense that the required accuracy of a model depends on its intended application, and finding a suitable model for a given purpose involves a trade-off between required model accuracy and affordable model complexity. Methodologies for model development that address this trade-off in an optimal manner are thus needed to enable fast business decision-making, especially in the biochemical, pharmaceutical and specialty chemicals industries, where time-to-market issues are of critical importance.

One such methodology is *grey-box modelling* [1], where the key idea is to find the simplest model for a given purpose, which is consistent with prior physical knowledge and not falsified by available experimental data. In the present paper a grey-box modelling framework is proposed, within which specific model deficiencies can be pinpointed and their structural origin uncovered in order to speed up the model development procedure. The key to uncovering the structural origin of model deficiencies is a possibility of obtaining estimates of unknown functional relations. An important tool for this purpose is nonparametric modelling, and the integration of nonparametric modelling with conventional grey-box modelling into a systematic framework for model improvement is the key contribution of the paper, the remainder of which is organized as follows: In Section 2 the details of the proposed framework are outlined; in Section 3 a case study illustrating its performance is presented and in Section 4 the conclusions of the paper are given.

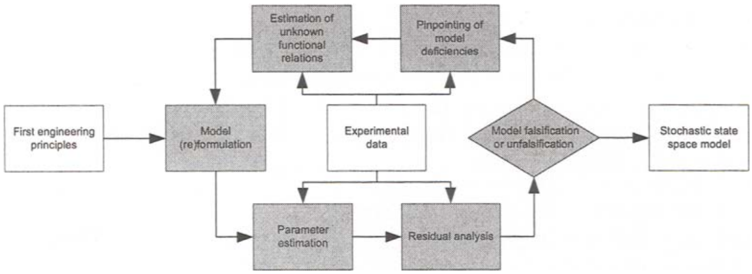


Figure 1. The proposed grey-box modelling cycle. The boxes in grey illustrate tasks and the boxes in white illustrate inputs to and outputs from the modelling cycle.

2. METHODOLOGY

The proposed grey-box modelling framework is shown in Figure 1 in the form of a modelling cycle, which shows the individual steps of the corresponding model development procedure. In the remainder of this section the individual steps are briefly described. A more elaborate discussion is given by Kristensen et al. [2].

2.1. Model (re)formulation

A key idea of grey-box modelling is to use all relevant prior physical knowledge, for which reason the first step within the modelling cycle is *model (re)formulation* based on first engineering principles, where the idea is to formulate an initial model structure in the form of a standard ODE model and subsequently translate this model into a grey-box model. Grey-box models are stochastic state space models consisting of a set of SDE's describing the dynamics of the system and a set of algebraic measurement equations, i.e.:

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})dt + \boldsymbol{\sigma}(\boldsymbol{u}_t, t, \boldsymbol{\theta})d\boldsymbol{\omega}_t \quad (1)$$

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta}) + \boldsymbol{e}_k \quad (2)$$

where $t \in \mathbb{R}$ is time, $\boldsymbol{x}_t \in \mathbb{R}^n$ is a state vector, $\boldsymbol{u}_t \in \mathbb{R}^m$ is an input vector, $\boldsymbol{y}_k \in \mathbb{R}^l$ is an output vector, $\boldsymbol{\theta} \in \mathbb{R}^p$ is a parameter vector, $\boldsymbol{f}(\cdot) \in \mathbb{R}^n$, $\boldsymbol{\sigma}(\cdot) \in \mathbb{R}^{n \times n}$ and $\boldsymbol{h}(\cdot) \in \mathbb{R}^l$ are nonlinear functions, $\{\boldsymbol{\omega}_t\}$ is an n-dimensional standard Wiener process and $\{\boldsymbol{e}_k\}$ is an l-dimensional white noise process with $\boldsymbol{e}_k \in N(\boldsymbol{0}, \boldsymbol{S}(\boldsymbol{u}_k, t_k, \boldsymbol{\theta}))$. The first term on the right-hand side of (1) is called the *drift* term and is equivalent to the term on the right-hand side of the standard ODE model. The second term is called the *diffusion* term and is included to accommodate random effects due to e.g. approximation errors.

2.2. Parameter estimation

The second step within the modelling cycle is *parameter estimation*, where the idea is to estimate the unknown parameters of the model in (1)-(2) from available experimental data, including the parameters of the diffusion term. The solution to (1) is a Markov process, and an estimation scheme based on probabilistic methods, e.g. *maximum likelihood* (ML) or *maximum a posteriori* (MAP), can therefore be applied. A detailed account of the estimation scheme used within the proposed framework is given by Kristensen et al. [3].

2.3. Residual analysis

The third step within the modelling cycle is *residual analysis*, where the idea is to evaluate the quality of the model by means of cross-validation, once the unknown parameters have been estimated. Depending on the intended application of the model this can be done in a one-step-ahead prediction setting as well as in a pure simulation setting. In either case a number of different methods can be applied [4].

2.4. Model falsification or unfalsification

The fourth step within the modelling cycle is the important step of *model falsification or unfalsification*, which deals with whether or not, based on the information obtained in the previous step, the model is sufficiently accurate to serve its intended purpose. If the model is unfalsified, the model development procedure can be terminated, but if the model is falsified, the modelling cycle must be repeated by re-formulating the model.

A key feature of the proposed framework is that, in the latter case, the properties of the model in (1)-(2) can be exploited to facilitate the task at hand.

2.5. Pinpointing of model deficiencies

The fifth step within the modelling cycle therefore deals with *pinpointing of model deficiencies*, and relies on the asymptotic Gaussianity of the ML estimator mentioned above [3], which allows t-tests to be performed to test whether the individual parameters are significant or not. Such tests are important because of the nature of the model in (1)-(2), where the diffusion term is included to account for random effects due to e.g. approximation errors, which means that the presence of *significant* parameters in this term is an indication that the corresponding drift term is incorrect. This in turn provides an uncertainty measure that allows model deficiencies to be detected. If a diagonal parameterization of the diffusion term is used, this even allows the deficiencies to be pinpointed in the sense that deficiencies in specific elements of the drift term can be detected, indicating that some of the inherent phenomena of this term may be incorrectly modelled.

If, by applying physical insights, a specific phenomenon can subsequently be selected for further investigation, the proposed framework also provides means to confirm if the suspicion that this phenomenon is incorrectly modelled is true or not. Typical such phenomena include reaction rates and similar complex dynamic phenomena, all of which are usually modelled using functions of the state and input variables, i.e. $r_t = \varphi(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta})$, where r_t is a phenomenon of interest and $\varphi(\cdot) \in \mathbb{R}$ is the nonlinear function used to model it. To confirm if the suspicion that $\varphi(\cdot)$ is incorrect is true, the parameter estimation step must be repeated with a re-formulated version of the model in (1)-(2) to give new statistical information. More specifically, if r_t is isolated by including it in the re-formulated model as an additional state variable, i.e.:

$$d\boldsymbol{x}_t^* = \boldsymbol{f}^*(\boldsymbol{x}_t^*, \boldsymbol{u}_t, t, \boldsymbol{\theta})dt + \boldsymbol{\sigma}^*(\boldsymbol{u}_t, t, \boldsymbol{\theta})d\boldsymbol{\omega}_t^* \tag{3}$$

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_k^*, \boldsymbol{u}_k, t_k, \boldsymbol{\theta}) + \boldsymbol{e}_k \tag{4}$$

where $\boldsymbol{x}_t^* = [\boldsymbol{x}_t^T \; r_t]^T$, $\boldsymbol{\sigma}^*(\cdot) \in \mathbb{R}^{(n+1)\times(n+1)}$ and $\{\boldsymbol{\omega}_t^*\}$ is an $(n+1)$-dimensional standard Wiener process and where

$$\boldsymbol{f}^*(\boldsymbol{x}_t^*, \boldsymbol{u}_t, t, \boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta}) \\ \frac{\partial \varphi(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta})}{\partial \boldsymbol{x}_t} \frac{d\boldsymbol{x}_t}{dt} + \frac{\partial \varphi(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta})}{\partial \boldsymbol{u}_t} \frac{d\boldsymbol{u}_t}{dt} \end{pmatrix} \tag{5}$$

then the presence of significant parameters in the corresponding diagonal element of the expanded diffusion term is a strong indication that $\varphi(\cdot)$ is in fact incorrect.

2.6. Estimation of unknown functional relations

The sixth step within the modelling cycle, which can only be used if specific model deficiencies have been pinpointed as described above, deals with *estimation of unknown functional relations*. The idea is to uncover the structural origin of these model deficiencies, and the procedure for doing this is based on a combination of the applicability of stochastic state space models for state estimation and the ability of nonparametric regression methods to provide visualizable estimates of unknown functional relations.

Using the re-formulated model in (3)-(4) and the corresponding parameter estimates, state estimates $\hat{\boldsymbol{x}}^*_{k|k}$, $k = 0, \ldots, N$, can be obtained for a given set of experimental data by applying the extended Kalman filter. In particular, since the incorrectly modelled phenomenon r_t is included as an additional state variable in this model, estimates $\hat{r}_{k|k}$, $k = 0, \ldots, N$, can be obtained, which in turn facilitates application of nonparametric regression to provide estimates of possible functional relations between r_t and the state and input variables. Several such techniques are available, but in the context of the proposed framework, *additive models* [5] are preferred. With additive models, the variation in r_t can be decomposed into the variation that can be attributed to each of the state and input variables in turn and the result can be visualized by means of partial dependence plots with associated bootstrap confidence intervals. In this manner, it may be possible to reveal the true structure of the function describing r_t, which in turn provides valuable information about how to re-formulate the model for the next modelling cycle iteration.

3. CASE STUDY: MODELLING A FED-BATCH BIOREACTOR

To illustrate the performance of the proposed framework in terms of improving the quality of an existing model, a simple simulation example is considered in the following.

The process considered is a fed-batch bioreactor, where the true model used to simulate the process is given as follows:

$$\frac{d}{dt}\begin{pmatrix} X \\ S \\ V \end{pmatrix} = \begin{pmatrix} \mu(S)X - \frac{FX}{V} \\ -\frac{\mu(S)X}{Y} + \frac{F(S_F - S)}{V} \\ F \end{pmatrix} \tag{6}$$

where X is the biomass concentration, S is the substrate concentration, V is the volume, F is the feed flow rate, Y is the yield coefficient of biomass, S_F is the feed concentration of substrate, and $\mu(S)$ is the biomass growth rate, i.e.:

$$\mu(S) = \mu_{\max}\frac{S}{K_2S^2 + S + K_1} \tag{7}$$

where $\mu_{\max}$, K_1 and K_2 are kinetic parameters. Simulated data sets from two batch runs are generated by perturbing the feed flow rate along a pre-determined trajectory and subsequently adding Gaussian measurement noise to the appropriate variables. Using these data sets it is now illustrated how the proposed modelling cycle can be used to improve an initial model, assuming that the intended purpose of the model is simulation.

3.1. First modelling cycle iteration

The first iteration through the modelling cycle starts with the model formulation step, where it is assumed that an initial model corresponding to (6) is available with the true structure of $\mu(S)$ in (7) unknown. This model is then translated into a grey-box model:

$$d\begin{pmatrix} X \\ S \\ V \end{pmatrix} = \begin{pmatrix} \mu X - \frac{FX}{V} \\ -\frac{\mu X}{Y} + \frac{F(S_F - S)}{V} \\ F \end{pmatrix} dt + \begin{bmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix} d\omega_t \tag{8}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}_k = \begin{pmatrix} X \\ S \\ V \end{pmatrix}_k + \boldsymbol{e}_k\ , \quad \boldsymbol{e}_k \in N(\boldsymbol{0}, \boldsymbol{S})\ , \quad \boldsymbol{S} = \begin{bmatrix} S_{11} & 0 & 0 \\ 0 & S_{22} & 0 \\ 0 & 0 & S_{33} \end{bmatrix} \tag{9}$$

where, because the true structure of $\mu(S)$ given in (7) is assumed to be unknown, a constant growth rate μ has been used instead. A diagonal parameterization of the diffusion term has been used to allow model deficiencies to be pinpointed.

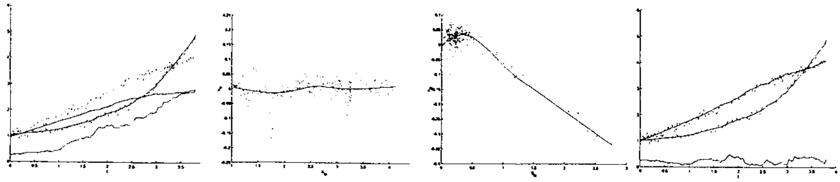
As the next step, the unknown parameters of the model in (8)-(9) are estimated with the ML method using the data from batch no. 1. Evaluating the quality of the resulting model by means of simple pure simulation comparison on cross-validation data from batch no. 2 gives the results shown in Figure 2a, which show that the model does a very poor job, particularly for y_1 and y_2. Moving to the model falsification or unfalsification step, the poor pure simulation capabilities falsify the model for its intended purpose, which means that the modelling cycle must be repeated by re-formulating the model.

To obtain information about how to do this in an intelligent way, model deficiencies should be pinpointed, if possible. t-tests for significance of the individual parameters show that, on a 5% level, the first two parameters of the diffusion term are both significant, which indicates that the the first two elements of the drift term may be incorrect. These elements both depend on μ, which is therefore an obvious suspect for being deficient. To confirm this suspicion, the model is re-formulated with μ as an additional state variable, and estimating the parameters of this model, using the same data set as before, gives t-test results that show that, of the parameters of the diffusion term, only the one corresponding to the equation for μ is now significant on a 5% level. This in turn indicates that there is substantial variation in μ and thus confirms the suspicion that μ is deficient.

Nonparametric modelling can now be applied to uncover the structural origin of the deficiency. Using the re-formulated model and the corresponding parameter estimates, state estimates $\hat{X}_{k|k}$, $\hat{S}_{k|k}$, $\hat{V}_{k|k}$, $\hat{\mu}_{k|k}$, $k = 0, \ldots, N$, are obtained and an additive model is fitted to reveal the true structure of the function describing μ by means of estimates of functional relations between μ and the state and input variables. It is reasonable to assume that μ does not depend on V and F, so only functional relations between $\hat{\mu}_{k|k}$ and $\hat{X}_{k|k}$ and $\hat{S}_{k|k}$ are estimated, giving the results shown in Figure 2b-c. These plots indicate that $\hat{\mu}_{k|k}$ does not depend on $\hat{X}_{k|k}$, but is highly dependent on $\hat{S}_{k|k}$, which in turn suggests to replace the assumption of constant μ with an assumption of μ being a function of S that complies with the revealed functional relation.

3.2. Second modelling cycle iteration

The functional relation revealed in the partial dependence plot between $\hat{\mu}_{k|k}$ and $\hat{S}_{k|k}$ in Figure 2b-c clearly indicates that the growth of biomass is governed by Monod kinetics and inhibited by substrate, which in the first step of the second iteration through the modelling cycle makes it possible to re-formulate the model in (8)-(9) accordingly by replacing μ with the true function $\mu(S)$ in (7). Estimation of the unknown parameters of the re-formulated model using the same data set as before gives much better results, which is evident from the cross-validation results shown in Figure 2d, which show that



(a) Initial comp. (b) $\hat{\mu}_{k|k}$ vs. $\hat{X}_{k|k}$ (c) $\hat{\mu}_{k|k}$ vs. $\hat{S}_{k|k}$ (d) Final comp.

Figure 2. Pure simulation comparison results using cross-validation data from batch no. 2 (sim. values in solid), and partial dependence plots of $\hat{\mu}_{k|k}$ vs. $\hat{X}_{k|k}$ and $\hat{S}_{k|k}$ (solid lines: Estimates; dotted lines: 95% bootstrap confidence intervals).

the pure simulation capabilities of the re-formulated model are very good. Moving to the model falsification or unfalsification step, the re-formulated model is thus unfalsified for its intended purpose, and the model development procedure can now be terminated.

4. CONCLUSION

A systematic framework for improving the quality of continuous time models of dynamic systems based on experimental data has been presented. The proposed grey-box modelling framework is based on an interplay between stochastic differential equation (SDE) modelling, statistical tests and nonparametric modelling and provides features that allow model deficiencies to be pinpointed and their structural origin to be uncovered to improve the model. A key result in this regard is that the proposed framework can be used to obtain nonparametric estimates of unknown functional relations, which allows unknown or incorrectly modelled dynamic phenomena to be uncovered and proper parametric expressions to be inferred from the estimated functional relations.

REFERENCES

[1] Bohlin, T. and Graebe, S. F. (1995). Issues in Nonlinear Stochastic Grey-Box Identification. *International Journal of Adaptive Control and Signal Processing*, **9**, 465–490.

[2] Kristensen, N. R.; Madsen, H. and Jørgensen, S. B. (2002b). A Method for Systematic Improvement of Stochastic Grey-Box Models. Submitted for publication.

[3] Kristensen, N. R.; Madsen, H. and Jørgensen, S. B. (2002a). Parameter Estimation in Stochastic Grey-Box Models. Submitted for publication.

[4] Holst, J.; Holst, U.; Madsen, H. and Melgaard, H. (1992). Validation of Grey Box Models. In L. Dugard; M. M'Saad and I. D. Landau, editors, *Selected Papers from the 4th IFAC Symposium on Adaptive Systems in Control and Signal Processing*, pages 407–414. Pergamon Press.

[5] Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall, London, England.