European Symposium on Computer Aided Process Engineering – 12
J. Grievink and J. van Schijndel (Editors)

Using Continuous Time Stochastic Modelling and Nonparametric Statistics to Improve the Quality of First Principles Models

Niels Rode Kristensen[a], Henrik Madsen[b] and Sten Bay Jørgensen[a]
[a)] Computer Aided Process Engineering Center, Department of Chemical Engineering
[b)] Mathematical Statistics Section, Informatics and Mathematical Modelling
Technical University of Denmark, DTU, DK-2800 Lyngby, Denmark

Abstract

A methodology is presented that combines modelling based on first principles and data based modelling into a modelling cycle that facilitates fast decision-making based on statistical methods. A strong feature of this methodology is that given a first principles model along with process data, the corresponding modelling cycle can be used to easily, rapidly and in a statistically sound way produce a more reliable model of the given system for a given purpose. A computer-aided tool, which integrates the elements of the modelling cycle, is also presented, and an example is given of modelling a fed-batch bioreactor.

1. Introduction

The increasing use of computer simulations in analysis and design of process systems and recent advances in model based process control and process optimisation have made the development of rigorous dynamic process models increasingly important over the past couple of decades. Particularly in view of the increasing focus on batch and fed-batch operation in many areas of the process industry, the ability of such process models to describe nonlinear and time-varying behaviour has also become more important.

Altogether, these developments have necessitated faster development of new and improvement of existing first principles models, i.e. models based on physical insights and conservation balances. The purpose of this paper is to show how *continuous time stochastic modelling* and time series analysis tools based on *nonparametric statistics* can be used to facilitate this. Continuous time stochastic modelling is a grey-box approach to process modelling that combines deterministic and stochastic modelling through the use of stochastic differential equations (SDE's) and has previously been described in Kristensen et al. (2001a). Other previous contributions in the area of grey-box modelling include the work of Madsen and Melgaard (1991) and Bohlin and Graebe (1995) and references therein.

The outline of the paper is as follows: In Section 2 the overall methodology is described in terms of a modelling cycle, some details of the individual elements of this cycle are given, and a computer aided tool that facilitates the use of the overall methodology is briefly described. In Section 3 a case study is presented that shows how the

methodology can be used to improve the quality of a first principles model of a simple fed-batch bioreactor. Conclusions are given in Section 4.

2. Methodology

The overall methodology can be described in terms of Figure 1, which shows the proposed continuous time stochastic modelling cycle described in the following.

2.1 Model construction

The first step in the modelling cycle deals with construction of the basic model, which is a *continuous-discrete stochastic state space model* consisting of a set of SDE's describing the dynamics of the system in continuous time and a set of algebraic equations describing measurements at discrete time instants, i.e.

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

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

where t is time, $\mathbf{x}_t$ is a vector of state variables, $\mathbf{u}_t$ is a vector of input variables, $\mathbf{y}_k$ is a vector of measured output variables, $\boldsymbol{\theta}$ is a vector of unknown parameters, $\mathbf{f}(\cdot)$, $\boldsymbol{\sigma}(\cdot)$ and $\mathbf{h}(\cdot)$ are nonlinear functions, $\boldsymbol{\omega}_t$ is a Wiener process and $\mathbf{e}_k$ is a $N(\mathbf{0},\mathbf{S}(\mathbf{u}_k,t_k,\boldsymbol{\theta}))$ process. A detailed account of the advantages of using SDE's is given in Kristensen et al. (2001a).

2.2 Parameter estimation

The second step in the modelling cycle deals with estimation of the unknown parameters $\boldsymbol{\theta}$ in (1) and (2) using data sets from one or more experiments. The properties of the basic model allow statistical methods to be applied for this purpose, e.g. *maximum likelihood* (ML) estimation, or *maximum a posteriori* (MAP) estimation if prior information about the parameters is available. More specifically, the unknown parameters can be determined by solving a variant of the optimisation problem

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \left\{ -\ln\left(\left(\prod_{i=1}^{S} \left(\prod_{k=1}^{N_i} \frac{\exp\left(-\frac{1}{2} {\boldsymbol{\varepsilon}^i_k}^T {\mathbf{R}^i_{k|k-1}}^{-1} \boldsymbol{\varepsilon}^i_k \right)}{\sqrt{\det \mathbf{R}^i_{k|k-1}}\sqrt{2\pi}^l} \right) p\left(\mathbf{y}^i_0 \middle| \boldsymbol{\theta}\right) \right) \frac{\exp\left(-\frac{1}{2}\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\varepsilon}_{\boldsymbol{\theta}}\right)}{\sqrt{\det \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}\sqrt{2\pi}^p} \right) \right\} \tag{3}$$

by further conditioning on the initial conditions $\mathbf{y}_0^i$ in the individual experiments. $\boldsymbol{\varepsilon}^i_k$ and $\mathbf{R}^i_{k|k-1}$ are the mean and covariance of the innovations from an extended Kalman filter at the k'th sample in the i'th experiment, and $\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ are the deviation from, and the covariance of a prior estimate of the parameters. A more detailed account of this formulation is given in Kristensen et al. (2001a), and details about the algorithms behind the corresponding estimation methods can be found in Kristensen et al. (2001b).

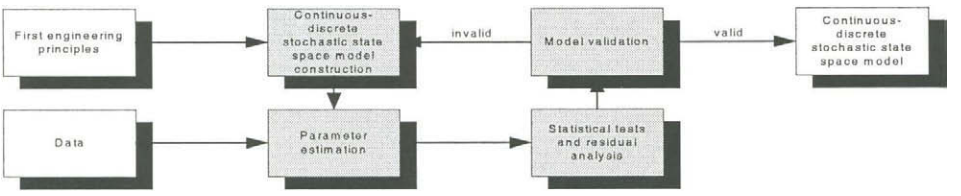

Figure 1. The continuous time stochastic modelling cycle.

2.3 Statistical tests and residual analysis

The third step in the modelling cycle deals with assessing the quality of the model once the unknown parameters have been estimated. The estimators described above are all approximately Gaussian, meaning that t-tests can be performed to test the hypothesis that a parameter is marginally insignificant. The test quantity is the value of the estimate of the parameter divided by the standard deviation of the estimate and is approximately t-distributed with a number of degrees of freedom that equals the number of data points minus the number of estimated parameters. To test the hypothesis that some parameters are simultaneously insignificant, several tests can be applied, e.g. a likelihood ratio test, a Lagrange multiplier test or a test based on Wald's W-statistic. These test quantities all have the same asymptotic χ^2-distribution with a number of degrees of freedom that equals the number of parameters to be tested for insignificance, but in the context of the proposed modelling cycle Wald's test has the advantage that no re-estimation is required. Details about the derivation of this statistic are given in Holst et al. (1992).

Another important aspect in assessing the quality of the model is to investigate its predictive capabilities by performing cross-validation and examining the corresponding residuals. Depending on the intended application of the model this can be done in a one-step-ahead prediction setting or in a pure simulation setting, and one of the most powerful methods is to compute and inspect the sample autocorrelation function (SACF) and the sample partial autocorrelation function (SPACF) of the residuals to detect if there are any significant lag dependencies, as this indicates that the model is incorrect. Nielsen and Madsen (2001) recently presented extensions of these linear tools to nonlinear systems, the lag-dependence function (LDF) and the partial lag-dependence function (PLDF), which are based on the close relation between correlation coefficients and values of the coefficients of determination for regression models and which extend to nonlinear systems by incorporating nonparametric regression in the form of additive models. In the context of the proposed modelling cycle the ability of the LDF and the PLDF to detect nonlinear lag-dependencies is particularly important.

2.4 Model validation

The last step in the modelling cycle deals with model validation or invalidation, or, more specifically, with whether, based on the information gathered in the previous step, the model is invalidated with respect to its intended application or not. If the model is invalidated, the modelling cycle is repeated by first changing the structure of the model in accordance with the information gathered in all steps of the previous cycle.

2.5 A computer aided tool for continuous time stochastic modelling

To facilitate the use of the proposed modelling cycle, a GUI-based computer-aided tool, called CTSM, has been developed, cf. Kristensen et al. (2001b). Within CTSM models of the kind (1)-(2) can be set up, unknown parameters can be estimated using a variant of (3), and statistical tests and residual analysis can be performed. CTSM is very flexible with respect to the data sets that can be used for estimation, as features for dealing with occasional outliers, irregular sample intervals and missing observations have been implemented. CTSM runs on Win32, Solaris and Linux platforms, and on Solaris platforms the program supports shared memory parallellization using OpenMP for improved performance.

3. Case study: Modelling a fed-batch bioreactor

To illustrate how the proposed modelling cycle can be used to improve the quality of a first principles model, a simple example is given. The process considered is a fed-batch bioreactor described by a simple unstructured model of biomass growth, i.e.

$$d\begin{pmatrix} X \\ S \\ V \end{pmatrix} = \begin{pmatrix} \mu(S)X - \dfrac{XF}{V} \\ -\dfrac{\mu(S)X}{Y} + \dfrac{(S_F - S)F}{V} \\ F \end{pmatrix} dt + \begin{bmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix} d\omega_t \tag{4}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}_k = \begin{pmatrix} X \\ S \\ V \end{pmatrix}_k + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}_k \quad , \quad \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}_k \in \begin{pmatrix} N(0, S_{11}) \\ N(0, S_{22}) \\ N(0, S_{33}) \end{pmatrix} \tag{5}$$

where X is the biomass concentration, S is the substrate concentration, V is the volume of the fermenter, F is the feed flow rate, S_F (=10) is the feed concentration of substrate, Y (=0.5) is the yield coefficient of biomass and $\mu(S)$ is the growth rate. σ_{11}, σ_{22}, σ_{33}, S_{11}, S_{22} and S_{33} are stochastic parameters. Two different cases are considered for $\mu(s)$, corresponding to Monod kinetics with and without substrate inhibition, i.e.

$$\mu(S) = \mu_{\max} \frac{S}{K_2 S^2 + S + K_1} \tag{6}$$

$$\mu(S) = \mu'_{\max} \frac{S}{S + K'_1} \tag{7}$$

In the following the model consisting of equations (4), (5) and (6), with K_2=0.5, is regarded as the true process to be modelled, and using the true parameter values in Table 2 two sets of data are generated by stochastic simulation. One data set is used for estimation and the other is used for validation. The model consisting of equations (4), (5) and (7) is regarded as an original first principles model, which in the context of the modelling cycle is the basic model. Using the estimation data set, the unknown parameters of the model are estimated with CTSM, giving the results in Table 1.

Table 1. Estimation results using the incorrect model structure.

Parameter	X_0	S_0	V_0	μ'_{max}	K'_1	σ_{11}	σ_{22}	σ_{33}	S_{11}	S_{22}	S_{33}
True value	1	0.245	1	-	-	0	0	0	0.01	0.001	0.01
Estimate	1.042	0.250	0.993	0.737	0.003	0.104	0.182	0.000	0.008	0.000	0.011
Std. Dev.	0.014	0.010	0.001	0.008	0.001	0.018	0.010	0.000	0.001	0.000	0.003
t-score	72.93	24.94	689.3	96.02	2.396	5.867	18.26	1.632	6.453	3.467	3.801
Significant	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes

Results of marginal t-tests show that the only insignificant stochastic parameter is σ_{33}, whereas σ_{11} and σ_{22} are significant. This in turn indicates that the deterministic parts of the equations for X and S in (4) are incorrect in terms of describing the variations in the estimation data set. To investigate this further, residual analysis is performed. One-step-ahead prediction results on the validation data set are shown in Figure 2 and Figure 3

shows the SACF, SPACF, LDF and PLDF for the corresponding residuals. There are no significant lag dependencies in the residuals for y_1 and y_3, whereas in the residuals for y_2 there is a significant lag dependence at lag 1. This is an additional indication that the equation for S in (4) is incorrect. A final piece of evidence that something is wrong is gathered from the pure simulation results in Figure 2. The information now available clearly invalidates the model, particularly if its intended purpose is simulation, and the modelling cycle is repeated by modifying the structure of the model.

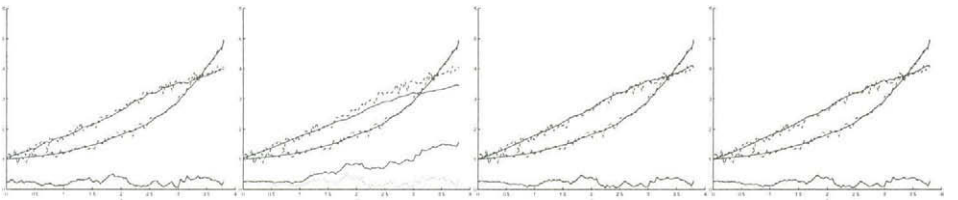

Figure 2. Cross-validation results. From left to right: One-step-ahead prediction and pure simulation using the incorrect model structure and one-step-ahead prediction and pure simulation using the correct model structure. (Solid: Predicted values, dashed: true y_1, dotted: true y_2, dash-dotted: true y_3).

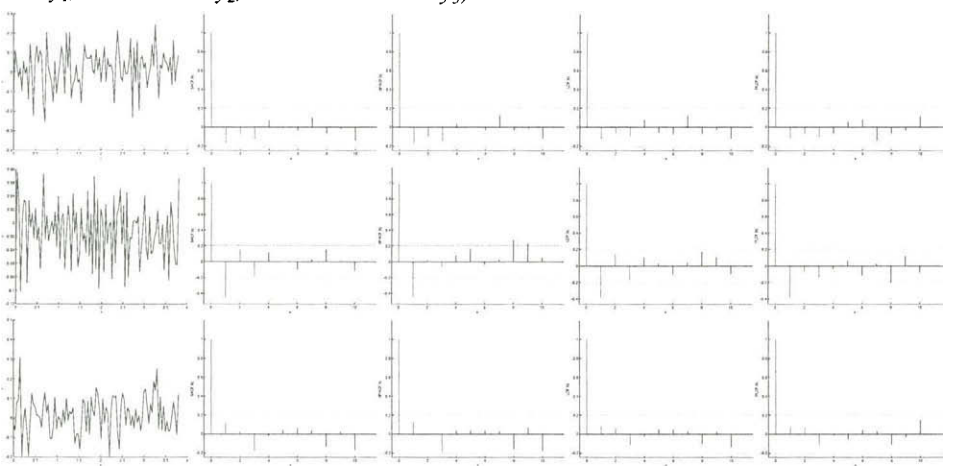

Figure 3. One-step-ahead prediction cross-validation residuals and corresponding SACF, SPACF, LDF and PLDF using the incorrect model structure. (Top: y_1, middle: y_2, bottom: y_3).

Table 2. Estimation results using the correct model structure.

Parameter	X_0	S_0	V_0	μ_{max}	K_1	σ_{11}	σ_{22}	σ_{33}	S_{11}	S_{22}	S_{33}
True value	1	0.245	1	1	0.03	0	0	0	0.01	0.001	0.01
Estimate	1.004	0.262	1.003	0.999	0.030	0.000	0.000	0.000	0.009	0.001	0.011
Std. Dev.	0.010	0.008	0.007	0.009	0.007	0.000	0.000	0.000	0.001	0.000	0.001
t-score	101.0	32.75	143.3	109.4	4.240	0.003	0.005	0.003	7.142	7.391	7.193
Significant	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes

The information available suggests that the deterministic parts of the equations for X and S in (4) are incorrect, i.e. those parts of the model that depend on $\mu(S)$. Replacing (7) with the correct structure in (6) and re-estimating the unknown parameters with

CTSM, the results shown in Table 2 are obtained. Marginal t-tests indicate that all three stochastic parameters, σ_{11}, σ_{22} and σ_{33}, are now insignificant, and the hypothesis of simultaneous insignificance cannot be rejected when performing a test based on Wald's W-statistic. Additional evidence that the modified model is correct is gathered by performing residual analysis. One-step-ahead prediction results on the validation data set are shown in Figure 2, and the SACF, SPACF, LDF and PLDF (not shown) for the corresponding residuals show no significant lag dependencies. A final piece of evidence of the validity of the modified model is gathered from the pure simulation results in Figure 2. In summary, if the intended purpose of the original model was simulation or infinite-horizon prediction, e.g. for use in an MPC controller, it has been now been invalidated and a more reliable model has been developed. However, if the intended purpose of the original model was one-step-ahead prediction, it might still be suitable.

4. Conclusion

A methodology has been presented that combines modelling based on first principles and data based modelling through the use of stochastic differential equations and statistical methods for parameter estimation and model validation. The methodology features a modelling cycle that can be used to easily, rapidly and in a statistically sound way develop a reliable model of a given system. A computer-aided tool, called CTSM, which integrates the elements of the modelling cycle, has also been presented.

5. References

Bohlin, Torsten and Stefan F. Graebe (1995). Issues in Nonlinear Stochastic Grey-Box Identification. *International Journal of Adaptive Control and Signal Processing* **9**, pp. 465-490.

Holst, Jan, Ulla Holst, Henrik Madsen and Henrik Melgaard (1992). Validation of Grey-Box Models. In: *Preprints of the IFAC Symposium on Adaptive Systems in Control and Signal Processing*, Grenoble, France, pp. 407-414.

Kristensen, Niels Rode, Henrik Madsen and Sten Bay Jørgensen (2001a). Computer Aided Continuous Time Stochastic Process Modelling. In: *European Symposium on Computer Aided Process Engineering – 11* (Rafiqul Gani and Sten Bay Jørgensen, Eds.), pp. 189-194.

Kristensen, Niels Rode, Henrik Melgaard and Henrik Madsen (2001b). *CTSM 2.0 – User's Guide*. DTU, Lyngby, Denmark.

Madsen, Henrik and Henrik Melgaard (1991). The Mathematical and Numerical Methods used in CTLSM. Technical Report 7/1991. IMM, DTU, Lyngby, Denmark.

Nielsen, Henrik Aalborg and Henrik Madsen (2001). A Generalization of Some Classical Time Series Tools. *Computational Statistics and Data Analysis* **37**(1), pp. 13-31.