

Modeling and Prediction Using Stochastic Differential Equations

Rune Juhl, Jan Kloppenborg Møller, John Bagterp Jørgensen
and Henrik Madsen

Abstract Pharmacokinetic/pharmakodynamic (PK/PD) modeling for a single subject is most often performed using nonlinear models based on deterministic ordinary differential equations (ODEs), and the variation between subjects in a population of subjects is described using a population (mixed effects) setup that describes the variation between subjects. The ODE setup implies that the variation for a single subject is described by a single parameter (or vector), namely the variance (covariance) of the residuals. Furthermore the prediction of the states is given as the solution to the ODEs and hence assumed deterministic and can predict the future perfectly. A more realistic approach would be to allow for randomness in the model due to e.g., the model be too simple or errors in input. We describe a modeling and prediction setup which better reflects reality and suggests stochastic differential equations (SDEs) for modeling and forecasting. It is argued that this gives models and predictions which better reflect reality. The SDE approach also offers a more adequate framework for modeling and a number of efficient tools for model building. A software package (CTSM-R) for SDE-based modeling is briefly described.

1 Introduction

Pharmacokinetic/pharmakodynamic (PK/PD) modeling is often performed using nonlinear mixed effects models based on deterministic ordinary differential equations (ODEs), [24]. The ODE models the dynamics of the system as

$$\begin{aligned}\frac{dX}{dt} &= f(X(t), t) \\ y_k &= X(t_k) + e_k,\end{aligned}$$

R. Juhl (✉) · J.K. Møller · J.B. Jørgensen · H. Madsen
Department of Applied Mathematics and Computer Science, Technical University
of Denmark, Kongens Lyngby, Denmark
e-mail: ruju@dtu.dk

where $X(t)$ is the state of the system, $f(\cdot)$ the model, y_k the discrete observations, and e_k the measurement errors which are assumed independent and identically distributed (iid) Gaussian. Given an initial value, the solution to the ODE $X(t)$ is a perfect prediction of all future values. The ODE model is an input–output model, where the residuals are the difference between the solution to the ODE and the observations. In the population setup, this implies that the total variation in data for a population of individuals is split into inter- and intraindividual variation. However, due to the ODE framework, the interindividual variation can only come from the covariance of the iid residuals, i.e., there must be no autocorrelation in the residuals.

The ODE framework is built on the assumption that future values of the states $X(t)$ can be predicted exactly and that the residual error is independent of the prediction horizon. This is often too simplistic and implies that the uncertainty about future values of the states and observations is not adequately described. This again has consequences for the design of model-based controllers and proper planning of medical treatments in general.

The ODE-based model class has a restricted residual error structure, as it assumes serially uncorrelated prediction residuals. There are several reasons why this assumption is violated: (1) misspecification or approximations of the structural model due to the complexity of the biological system, (2) unrecognized inputs, and (3) unpredictable random behavior of the process due to measurement errors for the input variables (e.g., specification of meals or physical exercise; both factors are known to influence future values of the blood glucose). In addition to these issues, the intraindividual (residual) variability also accounts for various environmental errors such as those associated with assay, dosing, and sampling. Since most of these errors cannot be considered as uncorrelated measurement errors, the description of the total individual error should preferably be separated (see also [7, 13]). Furthermore, [8] describe three types of residual error models to population PK/PD data analysis to account for more complicated residual error structures.

Neglecting the correlated residuals in the model description not only leads to serious issues when the model is used for forecasting and control as mentioned above, but it also disables a possibility for using proper methods for statistical model validation, parameter testing, and model identification (see e.g., [16], pp. 46–47).

In this chapter, stochastic differential equations (SDEs) are introduced to address these issues. SDEs facilitate the ability to split the intraindividual error into two fundamentally different types: (1) serially uncorrelated measurement error typically caused by assay error and (2) system error caused by model and input misspecifications. The concept will first be studied for a single subject and later on in a mixed effects setup with a population of individuals.

The use of SDEs opens up for new tools for model development, as it quantifies the amount of system and measurement noise. Specifically the approach allows for tracking of unknown inputs and parameters overtime by modeling them as random walk processes. These principles lead to efficient methods for pinpointing model deficiencies, and subsequently for identifying model improvements. The SDE approach also provides methods for model validation. This modeling framework is often called gray box modeling [25].

In this study, we will use maximum likelihood techniques both for parameter estimation and for model identification, and both for a single subject and in the population setting. It is known that parameter estimation in nonlinear mixed effects models with SDEs is most effectively carried out by considering an approximation to the population likelihood function. The population likelihood function is then approximated using the first-order conditional estimation (FOCE) method, which is based on a second-order Taylor expansion of each individual likelihood function at its optimum—see [17]. Like in [12], the extended Kalman filter is used for evaluating the single subject likelihood function.

This algorithm introduces a two-level numerical optimization, since not only the population likelihood function has to be maximized, but also for each value of the population likelihood all the individual likelihood functions must be maximized. This makes estimation computationally demanding, but the algorithm facilitates parallelization at several places to reduce the estimation time. The method is implemented in the R-package CTSM-R (continuous time stochastic modeling in R) [2], which is used in the DIACON project [3] focusing on technologies for semi- and fully-automatic insulin administration for treatment of type 1 diabetes. This project takes advantage of the fact that the SDE approach provides probabilistic forecasts for future values of the system states, which is crucial for reliable semi- and fully-automatic (closed-loop) insulin administration using model predictive control.

Section 2 describes various scenarios for data (single subject, repeated experiments, and populations of subjects), and how the likelihood function is formulated for each of these scenarios. Section 3 describes the approach used for population data from an experiment conducted in DIACON. Some practical issues related to SDE-based modeling are discussed in Sect. 4, and finally Sect. 5 summarizes. Both simulated and real-life experimental data are used throughout the chapter for illustrating the modeling and prediction framework.

2 Data and Modeling

Experiments can be conducted in various ways and the appropriate modeling approach depends on this. The basics start with a single experiment (solid ellipse in Fig. 1) which results in a series of data points \mathcal{Y} sampled, possibly irregularly, at times $t_1 < t_2 < \dots < t_N$. This single time series and how it is modeled are described in Sect. 2.1. Repeating the same experiment multiple times (dashed ellipse in Fig. 1) may be modeled as independent data series assuming no random effects between the runs. This is described in Sect. 2.2. When an experiment is done using several subjects (dotted ellipse in Fig. 1), then it is normal to include random effects between them. This is the so called population extension which is described in Sect. 2.3. In addition to the structure of data prior information may be available or used as a modeling technique. This is described in Sect. 2.4.

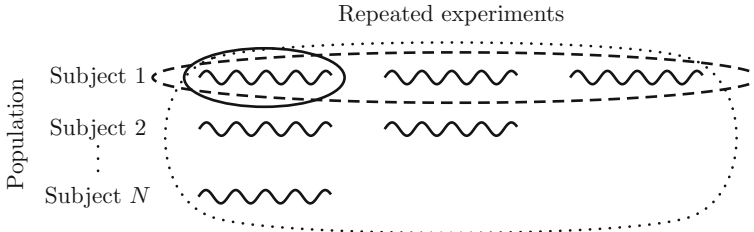


Fig. 1 A scenario of experiments in a study. The *solid ellipse* is a single time series trial. The *dashed ellipse* is a collection of three possibly independent repeated trials of subject 1. The *dotted ellipse* is a collection of subjects with random variation from a population

2.1 Single Data Series

This section begins by introducing a fundamental framework describing how to model physical phenomena. The aim is to provide a probabilistic model for a discrete time series $\mathcal{Y}_N = Y_1, Y_2, \dots, Y_N$. The formulation in this section is a general framework which is useful for all types of correlated time series data and not just physiological data.

The natural extension to the ODE framework is SDE's. We begin by introducing the stochastic process \mathbf{x}_t which satisfies an Itô SDE

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

where \mathbf{x}_t is the state, \mathbf{u}_t is an exogenous input, and $\boldsymbol{\theta}$ the parameters of the model. $\mathbf{f}()$ and $\boldsymbol{\sigma}()$ are possibly nonlinear functions called the drift and diffusion terms. $\boldsymbol{\omega}$ is the Wiener process driving the stochastic part of the process. (1) describes the dynamics and is called the system equation. Note that the ODE model is contained within the SDE when removing the diffusion term $\boldsymbol{\sigma}(\mathbf{x}_t, \mathbf{u}_t, t, \boldsymbol{\theta}) d\boldsymbol{\omega}_t$.

The solution to the SDE (1) is not in general known except for linear and a few other SDE's. Many methods for solving SDE's have been proposed, e.g., Hermite expansions, simulation-based methods and Kalman filtering, see [4]. This chapter focuses on the Kalman filter using CTSM-R. The Kalman filter restricts the diffusion to being independent of the states because the approximations required to integrate an SDE with state-dependent diffusion give undesirable results or performance. However, some SDEs with state-dependent diffusion can be transformed to an SDE with unit diffusion by the Lamperti transform (see Sect. 4.1).

The stochastic process is observed discretely and possibly partially with independent noise via the measurement equation

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

where $\mathbf{h}()$ is a possibly nonlinear function of the states and inputs. \mathbf{e}_k is an independent noise term attributed by the imperfect measurements. Due to the Kalman filter, the

measurement model is restricted to additive noise in CTSM-R

$$\mathbf{y}_k = \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k, t_k, \boldsymbol{\theta}) + \mathbf{e}_k, \tag{3}$$

where \mathbf{e}_k is Gaussian with $\mathcal{N}(0, \mathbf{S}(\mathbf{u}_k, t_k))$.

The combination (1) and (3) is the state space model formulation used in this paper to understand data. This is a gray box model as it bridges the gap between data driven black box models and pure physical white box models.

Example 1 As an example to illustrate the methods, we will use a simulation example (see Fig. 2). A linear 3 compartment transport model [15] similar to the real-data modeling example presented in Sect. 3 is used. We can think of the response (y) as venous glucose concentration in the blood of a patient, and the input (u) as exogenous glucagon.

The data are simulated according to the model

$$d\mathbf{x}_t = \left(\begin{bmatrix} u_t \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} -k_a & 0 & 0 \\ k_a & -k_a & 0 \\ 0 & k_a & -k_e \end{bmatrix} \mathbf{x}_t \right) dt + \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix} d\boldsymbol{\omega}_t \tag{4}$$

$$y_k = [0 \ 0 \ 1] \mathbf{x}_{t_k} + e_k, \tag{5}$$

where $\mathbf{x} \in \mathbb{R}^3$, $e_k \sim \mathcal{N}(0, s^2)$, $t_k = \{1, 11, 21, \dots\}$, and the specific parameters ($\boldsymbol{\theta}$) used for simulation are given in Table 1 (first column).

The structure of the model (4) will of course usually be hidden, and we will have to identify the structure based on the measurements as given in Fig. 2. As a general

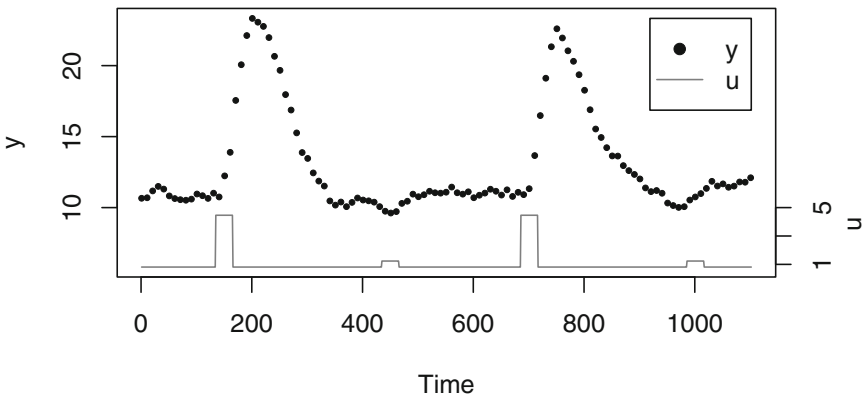


Fig. 2 Simulated data for the example (Eqs. (4), (5), and Table 1)

principle simple models are preferred over more complex models, and therefore a first hypothesis could be (Model 1)

$$dx_t = (u_t - k_e x_t) dt + \sigma_3 d\omega_t \quad (6)$$

$$y_k = x_{t_k} + e_k. \quad (7)$$

In this approach, the estimation is based on the likelihood function as defined in the following section.

2.1.1 Likelihood

Given a sequence of measurements

$$\mathcal{Y}_N = [\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_k, \dots, \mathbf{y}_N] \quad (8)$$

the likelihood of the unknown parameters θ given the model formulated as (1)–(3) is the joint probability density function (pdf)

$$L(\theta, \mathcal{Y}_N) = p(\mathcal{Y}_N | \theta), \quad (9)$$

where the likelihood L is the probability density function given θ . The joint probability density function is partitioned as the product of the one-step conditional probability functions

$$L(\theta, \mathcal{Y}_N) = \left(\prod_{k=1}^N p(\mathbf{y}_k | \mathcal{Y}_{k-1}, \theta) \right) p(\mathbf{y}_0 | \theta). \quad (10)$$

The solution to a linear SDE driven by a Brownian motion is a Gaussian process. Nonlinear SDEs do not result in a Gaussian process and thus the marginal probability is not Gaussian. By sampling, the nonlinearities fast enough in some sense then it is reasonable to assume that the conditional density is Gaussian.

The Gaussian density is fully described by the first and second-order moments

$$\hat{\mathbf{y}}_{k|k-1} = E[\mathbf{y}_k | \mathcal{Y}_{k-1}, \theta] \quad (11)$$

$$\Sigma_{k|k-1} = V[\mathbf{y}_k | \mathcal{Y}_{k-1}, \theta]. \quad (12)$$

Introducing the innovation error

$$\boldsymbol{\varepsilon}_k = \mathbf{y}_k - \hat{\mathbf{y}}_{k|k-1}, \quad (13)$$

the likelihood (10) becomes

$$L(\boldsymbol{\theta}, \mathcal{Y}_N) = \left(\prod_{k=1}^N \frac{\exp\left(-\frac{1}{2} \boldsymbol{\varepsilon}_k^T \boldsymbol{\Sigma}_{k|k-1}^{-1} \boldsymbol{\varepsilon}_k\right)}{\sqrt{|\boldsymbol{\Sigma}_{k|k-1}|} \sqrt{2\pi}^l} \right) p(\mathbf{y}_0 | \boldsymbol{\theta}). \quad (14)$$

The probability density of the initial observation $p(\mathbf{y}_0 | \boldsymbol{\theta})$ is parameterized through the probability density of the initial state $p(x_0 | \boldsymbol{\theta})$. The mean $\hat{\mathbf{y}}_{k|k-1}$ and covariance $\boldsymbol{\Sigma}_{k|k-1}$ are computed recursively using the extended Kalman filter, see Appendix A for a brief description, or [10] for a detailed description.

The unknown parameters are estimated by maximizing the likelihood function using an optimization algorithm. The likelihood (14) is a product of probability densities all less than 1, which causing numerical problems. Taking the logarithm of the likelihood (14) turns the product into a summation and cancels the exponentials thus stabilizing the calculation. The parameters are now found by maximizing the log-likelihood or by convention minimize the negative log-likelihood

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta} \in \Theta} (-\ln(L(\boldsymbol{\theta}, \mathcal{Y}_N))). \quad (15)$$

The uncertainty of the maximum likelihood parameter estimate $\hat{\boldsymbol{\theta}}$ is related to the curvature of the likelihood function. An estimate of the asymptotic covariance of $\hat{\boldsymbol{\theta}}$ is the inverse of the observed Fisher information matrix

$$V[\hat{\boldsymbol{\theta}}] = [\mathbf{I}(\hat{\boldsymbol{\theta}})]^{-1}, \quad (16)$$

where $\mathbf{I}(\hat{\boldsymbol{\theta}})$ is the observed Fisher information matrix, that is the negative Hessian matrix (curvature) of the likelihood function evaluated at the maximum likelihood estimate [17, 21].

Example 2 We continue with the simulated data from Example 1. As noted above, a first approach to model the data could be a first-state model (Eqs. (6)–(7)). The result of the estimation ($\hat{\boldsymbol{\theta}}_1$) is given in Table 1, the initial value of the state (x_{30}) and the time constant ($1/k_e$) are both captured quite well, while the uncertainty parameters are way off, the diffusion is too large and the observation variance is too small (with extremely large uncertainty).

The parameters in the model are all assumed to be greater than zero, and it is therefore advisable to estimate parameters in the log-domain, and then transform back to the original domain before presenting the estimates. The log-domain estimation is also the explanation for the nonsymmetric confidence intervals in Table 1, the confidence intervals are all based on the Hessian of the likelihood at the optimal parameter values, and confidence intervals are based on the Wald confidence interval in the transformed (log) domain [21]. Such intervals could be refined using profile likelihood-based confidence intervals [21] (see also Sect. 4.4).

Table 1 Parameter estimates from simulation example and confidence intervals for the individual parameters are given in parenthesis below the estimates

	θ	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
x_{10}	40.000	–		38.819 (29.172, 48.466)
x_{20}	35.000	–	107.960 (75.211, 140.710)	33.421 (29.778, 37.064)
x_{30}	11.000	10.657 (6.606, 14.708)	10.641 (10.392, 10.889)	10.604 (10.281, 10.927)
k_a	0.025	–	0.006 (0.0038, 0.0778)	0.026 (0.025, 0.027)
k_e	0.080	0.081 (0.071, 0.094)	0.056 (0.0418, 0.0743)	0.080 (0.078, 0.083)
σ_1	1.000	–	–	0.5500 (0.224, 1.353)
σ_2	0.200	–	3.616 (2.670, 4.898)	0.282 (0.113, 0.704)
σ_3	0.050	2.206 (1.848, 2.634)	0.001 ($2 \cdot 10^{-55}$, $3 \cdot 10^{48}$)	0.001 ($9 \cdot 10^{-56}$, $1 \cdot 10^{49}$)
s	0.025	0.0002 ($2 \cdot 10^{-33}$, $2.6 \cdot 10^{25}$)	0.016 (0.0065, 0.0388)	0.031 (0.020, 0.049)
$l(\hat{\theta}, \mathbf{y})$	–	–343.68	–67.85	–19.70
df	–	4	7	9

θ is the true values, $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$ are the estimated for the first-, second-, and third-state models, respectively. Last two rows present the log-likelihood and the number of degrees of freedom

In order to validate the model and suggest further development, we should inspect the innovation error. When the model is not time homogeneous, the standard error of the prediction will not be constant and the innovation error should be standardized

$$r_k = \frac{\varepsilon_k}{\sqrt{\Sigma_{k|k-1}}}, \tag{17}$$

where the innovation error (ε_k) is given in (13). All numbers needed to calculate the standardized residuals can be obtained directly from CTSM-R using the function `predict`. Both the autocorrelation and partial autocorrelation (Fig. 3) are significant in lag 1 and 2. This suggests a second-state model for the innovation error, and hence a third-state model should be used. Consequently we can go directly from the first-state model to the true structure (a third-state model).

Now we have assumed that a number of the parameters are actually zero, in a real-life situation, we might test these parameters using likelihood ratio tests, or indeed identify them through engineering principles. The parameter estimates are given in Table 1 ($\hat{\theta}_3$); in this case, the diffusion parameter (σ_3) has an extremely

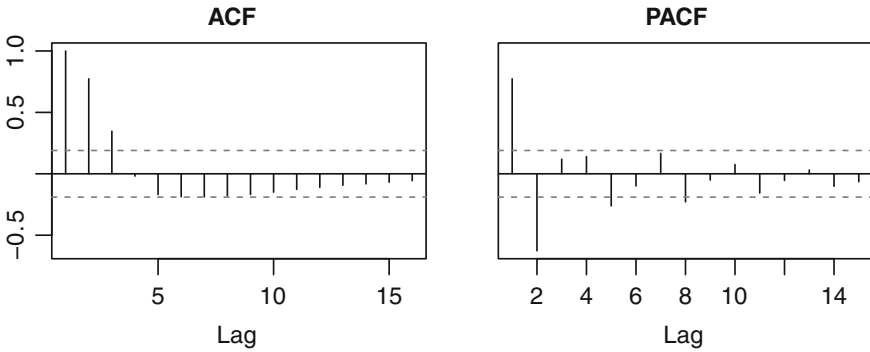


Fig. 3 Autocorrelation and partial autocorrelation from a simple (1 state) model

wide confidence interval, and it could be checked if these parameters should indeed be zero (again using likelihood ratio test), but for now we will proceed with the residual analysis which is an important part of model validation (see e.g., [16]). The autocorrelation and partial autocorrelation for the third-state model are shown in Fig. 4. We see that there are no values outside the 95 % confidence interval, and we can conclude that there is no evidence against the hypothesis of white noise residuals, i.e., the model sufficiently describes the data.

Autocorrelation and partial autocorrelations are based on short-term predictions (in this case 10 min) and hence we check the local behavior of the model. Depending on the application of the model, we might be interested in longer-term behavior of the model. Prediction can be made on any horizon using CTSM-R. In particular, we can compare deterministic simulation in CTSM-R (meaning conditioning only on the initial value of the states). Such a simulation plot is shown in Fig. 5, here we

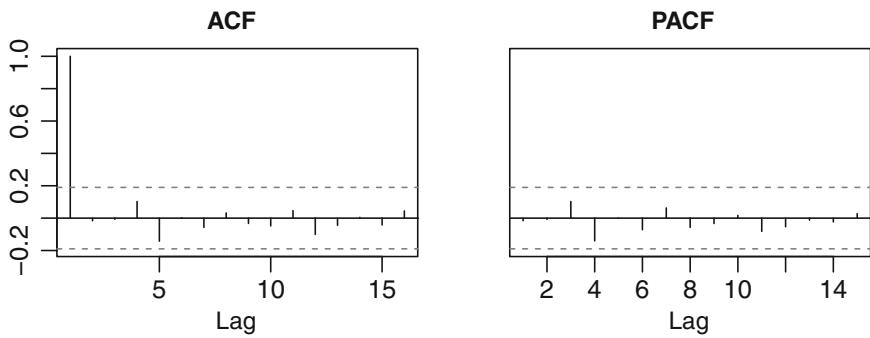


Fig. 4 Autocorrelation and partial autocorrelation from the third-state model (i.e., the correct model)

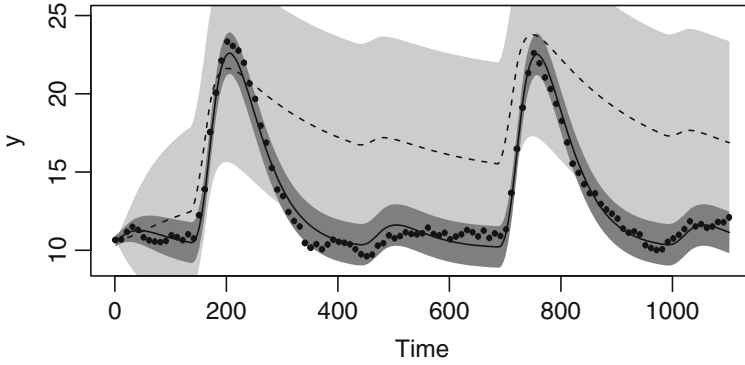


Fig. 5 Simulation with model 2 and 3, *dashed gray line* expectation of model 2, *black line* expectation of model 3, *light gray area* 95 % prediction interval for model 2, *dark gray area* 95 % prediction interval for model 3, and *black dots* are the observations

compare a second-state model (see Table 1) with the true third-state model. It is quite evident that model 2 is not suited for simulation, with the global structure being completely off, while “simulation” with a third-state model (with the true structure, but estimated parameters), gives narrow and reasonable simulation intervals. In the case of linear SDE-models with linear observation, this “simulation” is exact, but for nonlinear models it is recommended to use real simulations, e.g., using a Euler scheme.

The step from a second-state model (had we initialized our model development with a second-state model) to the third-state model is not at all trivial. However, Fig. 5 shows that simulation of model 2 does not contain the observations and thus model 2 will not be well suited for simulations. Also the likelihood ratio test (or AIC/BIC) supports that model 3 is far better than model 2, further it would be reasonable to fix σ_3 at zero (in practice a very small number).

2.2 Independent Data Series

An experiment may be repeated several times without expecting variation in the underlying parameters. Given S sequences of possibly varying length

$$\mathbf{Y} = [\mathcal{Y}_{N_1}^1, \mathcal{Y}_{N_2}^2, \dots, \mathcal{Y}_{N_i}^i, \dots, \mathcal{Y}_{N_S}^S], \quad (18)$$

the likelihood is the product of the likelihood (10) for each sequence

$$L(\boldsymbol{\theta}, \mathbf{Y}) = \prod_{i=1}^S \left(\left(\prod_{k=1}^N \frac{\exp\left(-\frac{1}{2} \boldsymbol{\varepsilon}_k^T \boldsymbol{\Sigma}_{k|k-1}^{-1} \boldsymbol{\varepsilon}_k\right)}{\sqrt{|\boldsymbol{\Sigma}_{k|k-1}|} \sqrt{2\pi^l}} \right) p(\mathbf{y}_{0,i} | \boldsymbol{\theta}) \right). \quad (19)$$

The unknown parameters are again estimated by minimizing the negative log-likelihood

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta} \in \Theta} (-\ln(L(\boldsymbol{\theta}, \mathbf{Y}))). \quad (20)$$

If the independence assumption is violated and the parameters vary between the time series, then the model performance would be lowered as the parameter estimates will be a compromise. The natural extension is to include a population effect.

2.3 Population Extension

The gray box model can be extended to include a hierarchical structure to model variation occurring between data series where each series has its own parameter set. This is useful for describing data from a number of individuals belonging to a population of individuals. The hierarchical modeling is also called mixed effects and population extension in pharmaceutical science. Nonlinear mixed effects modeling has long been used in pharmacokinetic/pharmacodynamic studies to account for variation from the natural grouping: multiple centers, multiple days, age and BMI of subjects, etc. Mixed effects modeling combines fixed and random effects [17]. The fixed effect is the average of that effect over the entire population while the random effect allows for variation around that average.

Consider N subjects in a clinical study. This is a single level grouping. The model for the i th subject is

$$d\mathbf{x}_{i,t} = \mathbf{f}(\mathbf{x}_{i,t}, \mathbf{u}_{i,t}, t, \boldsymbol{\theta}_i) dt + \boldsymbol{\sigma}(\mathbf{u}_{i,t}, t, \boldsymbol{\theta}_i) d\boldsymbol{\omega}_t \quad (21)$$

$$\mathbf{y}_{i,k} = \mathbf{h}(\mathbf{x}_{i,k}, \mathbf{u}_{i,k}, t_{i,k}, \boldsymbol{\theta}_i) + \mathbf{e}_{i,k}, \quad (22)$$

which is the general model extended with subscript i . The individual parameters $\boldsymbol{\theta}_i$ are

$$\boldsymbol{\theta}_i = z(\boldsymbol{\theta}_f, \mathbf{Z}_i, \boldsymbol{\eta}_i), \quad (23)$$

where z maps from subject covariates such (i.e., BMI and age) \mathbf{Z}_i , fixed effects parameters $\boldsymbol{\theta}_f$, and the random effects $\boldsymbol{\eta}_i \in \mathcal{R}^k \sim \mathcal{N}(0, \boldsymbol{\Omega})$ to subject parameters. The subject parameters are typically modeled as either normally or log-normally distributed by combining the fixed effect parameters and the random effects in either an additive $\boldsymbol{\theta}_i = \boldsymbol{\theta}_f + \boldsymbol{\eta}_i$ or an exponential transform $\boldsymbol{\theta}_i = \boldsymbol{\theta}_f e^{\boldsymbol{\eta}_i}$.

The likelihood of the fixed effects is the product of the marginal probability densities for each subject

$$L(\boldsymbol{\theta}_f, \Omega) = \prod_{i=1}^N p(\mathcal{Y}_i | \boldsymbol{\theta}_i, \Omega), \quad (24)$$

where the marginal density is found by integrating over the random effects $\boldsymbol{\eta}_i$

$$p(\mathcal{Y}_i | \boldsymbol{\theta}_i, \Omega) = \int p_1(\mathcal{Y}_i | \boldsymbol{\theta}_i, \boldsymbol{\eta}_i) p_2(\boldsymbol{\eta}_i | \Omega) d\boldsymbol{\eta}_i. \quad (25)$$

$p_1(\mathcal{Y}_i | \boldsymbol{\theta}, \boldsymbol{\eta})$ is the probability of the individual subject which given by (10). $p_2(\boldsymbol{\eta}_i | \Omega)$ is the probability of the second-stage model where the random effects describe the interindividual variation.

2.3.1 Approximation of the Marginal Density

The integral in (25) rarely has a closed-form solution and thus must be approximated in a computationally feasible way. This can be done in two ways: approximating (a) the integrand by Laplacian or (b) the entire integral by Gaussian quadrature.

Gaussian quadrature can approximate the integral by a weighted sum of the integrand evaluated at specific nodes. The accuracy of Gaussian quadrature increases as the order (number of nodes) increases. With adaptive Gaussian quadrature, the accuracy can be improved even further at higher cost. The computational complexity of Gaussian quadrature suffers from the curse of dimensionality and becomes infeasible even for few dimensions.

Now consider the Laplacian approximation which is widely used approximation to integrals [17]. Observe that the integrand in (25) is nonnegative such that

$$\begin{aligned} p_1(\mathcal{Y}_i | \boldsymbol{\theta}_i, \boldsymbol{\eta}_i) p_2(\boldsymbol{\eta}_i | \Omega) &= e^{\log(p_1(\mathcal{Y}_i | \boldsymbol{\theta}_i, \boldsymbol{\eta}_i) p_2(\boldsymbol{\eta}_i | \Omega))} \\ &= e^{g_i(\boldsymbol{\eta}_i)}, \end{aligned} \quad (26)$$

where $g_i(\boldsymbol{\eta}_i)$ is the log-posterior distribution for the i th subject. Now consider the second-order Taylor expansion of $g_i(\boldsymbol{\eta}_i)$ around its mode $\hat{\boldsymbol{\eta}}_i$

$$g_i(\boldsymbol{\eta}_i) \approx g_i(\hat{\boldsymbol{\eta}}_i) + \frac{1}{2} (\boldsymbol{\eta}_i - \hat{\boldsymbol{\eta}}_i)^T \Delta g_i(\hat{\boldsymbol{\eta}}_i) (\boldsymbol{\eta}_i - \hat{\boldsymbol{\eta}}_i), \quad (27)$$

since $\nabla g_i(\hat{\boldsymbol{\eta}}_i) = 0$ at the mode. By inserting (27) and (26) in (25), the Laplacian approximation of the marginal probability density is defined as

$$\begin{aligned} p(\mathcal{Y}_i | \boldsymbol{\theta}_f, \Omega) &\approx \int e^{g_i(\hat{\boldsymbol{\eta}}_i) + \frac{1}{2}(\boldsymbol{\eta}_i - \hat{\boldsymbol{\eta}}_i)^T \Delta g_i(\hat{\boldsymbol{\eta}}_i)(\boldsymbol{\eta}_i - \hat{\boldsymbol{\eta}}_i)} d\boldsymbol{\eta}_i \\ &= e^{g_i(\hat{\boldsymbol{\eta}}_i)} \int e^{\frac{1}{2}(\boldsymbol{\eta}_i - \hat{\boldsymbol{\eta}}_i)^T \Delta g_i(\hat{\boldsymbol{\eta}}_i)(\boldsymbol{\eta}_i - \hat{\boldsymbol{\eta}}_i)} d\boldsymbol{\eta}_i, \end{aligned} \quad (28)$$

where the integral is recognized as the scaled integral over a multivariate Gaussian¹ distribution with covariance $\Sigma = (-\Delta g(\boldsymbol{\eta}_i))^{-1}$. The marginal density becomes

$$p(\mathcal{Y}_i | \boldsymbol{\theta}_i, \Omega) \approx e^{g_i(\hat{\boldsymbol{\eta}}_i)} \sqrt{\frac{(2\pi)^k}{|-\Delta g(\hat{\boldsymbol{\eta}}_i)|}}. \quad (29)$$

Inserting (29) in (24) the likelihood becomes

$$L(\boldsymbol{\theta}_f, \Omega) \approx \prod_{i=1}^N e^{g_i(\hat{\boldsymbol{\eta}}_i)} \sqrt{\frac{(2\pi)^k}{|-\Delta g(\hat{\boldsymbol{\eta}}_i)|}}. \quad (30)$$

The Hessian $\Delta g(\hat{\boldsymbol{\eta}}_i)$ is found by analytically differentiating the expression for the log-posterior $g(\boldsymbol{\eta})$. After some derivation, the Hessian is

$$\begin{aligned} \Delta g(\boldsymbol{\eta}_i) &= \sum_{k=1}^N \left[\frac{\partial^2 \mathbf{y}^T}{\partial \boldsymbol{\eta}_i \partial \boldsymbol{\eta}_i} \Sigma_{k|k-1}^{-1} (\mathbf{y}_k - \hat{\mathbf{y}}_{k|k-1}) + 2 \frac{\partial \hat{\mathbf{y}}_{k|k-1}}{\partial \boldsymbol{\eta}_i} \frac{\partial \left[\Sigma_{k|k-1}^{-1} \right]}{\partial \boldsymbol{\eta}_i} (\mathbf{y} - \hat{\mathbf{y}}_{k|k-1}) \right. \\ &\quad - \frac{\partial \hat{\mathbf{y}}_{k|k-1}}{\partial \boldsymbol{\eta}_i} \Sigma_{k|k-1}^{-1} \frac{\partial \hat{\mathbf{y}}_{k|k-1}}{\partial \boldsymbol{\eta}_i} - \frac{1}{2} (\mathbf{y} - \hat{\mathbf{y}}_{k|k-1}) \frac{\partial^2 \left[\Sigma_{k|k-1}^{-1} \right]}{\partial \boldsymbol{\eta}_i \partial \boldsymbol{\eta}_i} (\mathbf{y} - \hat{\mathbf{y}}_{k|k-1}) \\ &\quad \left. + \text{tr} \left(\frac{\partial \left[\Sigma_{k|k-1}^{-1} \right]}{\partial \boldsymbol{\eta}_i} \frac{\partial \Sigma_{k|k-1}}{\partial \boldsymbol{\eta}_i} + \Sigma_{k|k-1}^{-1} \frac{\partial \Sigma_{k|k-1}}{\partial \boldsymbol{\eta}_i \partial \boldsymbol{\eta}_i} \right) \right] - \Omega^{-1}, \end{aligned} \quad (31)$$

where tr is the trace of a matrix. The second-derivative terms are generally complicated or inconvenient to compute. At the mode $\hat{\boldsymbol{\eta}}_i$, the contribution of the second-derivative terms is usually negligible and thus an approximation for the Hessian is

$$\Delta g(\hat{\boldsymbol{\eta}}_i) \approx - \sum_{k=1}^N \left(\frac{\partial \hat{\mathbf{y}}_{k|k-1}}{\partial \boldsymbol{\eta}_i} \Big|_{\boldsymbol{\eta}_i = \hat{\boldsymbol{\eta}}_i} \Sigma_{k|k-1}^{-1} \frac{\partial \hat{\mathbf{y}}_{k|k-1}}{\partial \boldsymbol{\eta}_i} \Big|_{\boldsymbol{\eta}_i = \hat{\boldsymbol{\eta}}_i} \right) - \Omega^{-1}. \quad (32)$$

¹The integral over the multivariate Gaussian density is $\frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \int e^{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})} d\mathbf{x} = 1$.

This approximation is similar to the Gauss–Newton and NONMEM’s first-order conditional estimation (FOCE) approximations of the Hessian where only first partial derivatives are included [9, 17].

The parameters are found by iteratively minimizing the first- and second-stage model. For a trial set of fixed effect parameters, an optimization of g must be done for all subjects. When all η_i have been found, the Laplacian and FOCE approximations can be computed to obtain the population likelihood. The population likelihood can then be optimized.

2.4 Prior Information

Bayesian analysis combines the likelihood of the data and already known information which is called a prior. When the prior probability density function is updated, it becomes the posterior probability density function. In true, Bayesian analysis the prior may be any distribution, although conjugated priors are used in practice to simplify the computations.

In the view of CTSM-R, priors are mainly used as (a) empirical prior or for (b) regularizing the estimation.

An empirical prior is a result from a previous estimation. Imagine an experiment has been analyzed and followed by rerunning the experiment. These two data series are stochastically independent sets and should be analyzed as in Sect. 2.2. However, using the results from the first analysis as a prior, only the new data series has to be analyzed. If the quadratic Wald approximation holds this prior is Gaussian.

Regularizing one or more parameters is sometimes required to achieve a feasible estimation of the parameters. State equations describe a physical phenomenon and as such the modeler often has knowledge (possibly partly subjective) about the parameters from, e.g., another study. The reported values are often a mean and a standard deviance. Thus a Gaussian prior is reasonable.

Updating the prior probability density function $p(\boldsymbol{\theta})$ forms the posterior probability density function through Bayes’ rule

$$p(\boldsymbol{\theta}|\mathcal{Z}_N) = \frac{p(\mathcal{Z}_N|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{Z}_N)} \propto p(\mathcal{Z}_N|\boldsymbol{\theta})p(\boldsymbol{\theta}), \quad (33)$$

where the probability density $p(\mathcal{Z}_N|\boldsymbol{\theta})$ is proportional to the likelihood of a single data series given in (10). No information is called a diffuse prior which is uniform over the entire domain. The posterior then reduces to the likelihood of the data.

Let the prior be described by a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)$ where

$$\boldsymbol{\mu}_\theta = E[\boldsymbol{\theta}] \quad (34)$$

$$\boldsymbol{\Sigma}_\theta = V[\boldsymbol{\theta}], \quad (35)$$

and let

$$\boldsymbol{\varepsilon}_\theta = \boldsymbol{\theta} - \boldsymbol{\mu}_\theta, \quad (36)$$

then the posterior probability density function is

$$p(\boldsymbol{\theta}|\mathcal{Y}_N) \propto \left(\prod_{k=1}^N \frac{\exp\left(-\frac{1}{2}\boldsymbol{\varepsilon}_k^T \boldsymbol{\Sigma}_{k|k-1}^{-1} \boldsymbol{\varepsilon}_k\right)}{\sqrt{|\boldsymbol{\Sigma}_{k|k-1}|} \sqrt{2\pi^l}} \right) p(\mathbf{y}_0|\boldsymbol{\theta}) \times \frac{\exp\left(-\frac{1}{2}\boldsymbol{\varepsilon}_\theta^T \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{\varepsilon}_\theta\right)}{\sqrt{|\boldsymbol{\Sigma}_\theta|} \sqrt{2\pi^p}}. \quad (37)$$

The parameters are estimated by maximizing the posterior density function (37), i.e., maximum a posteriori (MAP) estimation. The MAP parameter estimate is found by minimizing the negative logarithm of (37)

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta} \in \Theta} (-\ln(p(\boldsymbol{\theta}|\mathcal{Y}_N, \mathbf{y}_0))). \quad (38)$$

When there is no prior the MAP estimate reduces to the ML estimate.

3 Example: Modeling the Effect of Exercise on Insulin Pharmacokinetics in “Continuous Subcutaneous Insulin Infusion” Treated Type 1 Diabetes Patients

The artificial pancreas is believed to ease substantially the burden of constant management of type 1 diabetes for patients. An important aspect of the artificial pancreas development is the mathematical models used for control, prediction, and simulation. A major challenge to the realization of the artificial pancreas is the effect of exercise on the insulin and plasma glucose dynamics. This is the first step towards a population model of exercise effects in type 1 diabetes. The focus is on the effect on the insulin pharmacokinetics in continuous subcutaneous insulin infusion (CSII)-treated patients by modeling the absorption rate as a function of exercise. This example is described in detail in [5].

3.1 Data

The insulin data for this study originates from a clinical study on 12 subjects with type 1 diabetes treated with continuous subcutaneous insulin infusion (CSII). Each subject did two study days separated by at least three weeks. The insulin was observed by drawing blood nonequidistantly over the course of the trial. A detailed description of the data is found in [23].

Natural considerations toward the subjects limits how frequent the insulin can be sampled. This limits the amount of observations per time series and often care-

ful nonequidistant sampling becomes necessary. Both issues makes estimation of parameters more difficult. However, using all the subjects collectively increases the amount of data and improves estimation. The repeated trials per subject are considered independent trials, i.e., no random variation on the parameters. The subjects are assumed to have interindividual variation for several of the parameters.

3.2 The Gray Box Insulin Model

A linear three-compartment ODE model is used as basis to describe the pharmacokinetics of subcutaneous infused insulin in a single subject as suggested by [26]. The model is illustrated in Fig. 6.

The absorption is characterized by the rate parameter k_a between all three compartments. The two compartments I_{sc_2} and I_p are modeled with diffusion. Only the third-state I_p is being observed.

The compartment model is formulated as the following SDE

$$d \begin{bmatrix} I_{sc_1} \\ I_{sc_2} \\ I_p \end{bmatrix} = \left(\begin{bmatrix} -k_a & 0 & 0 \\ k_a & -k_a & 0 \\ 0 & \frac{k_a}{V_I} & -k_e \end{bmatrix} \begin{bmatrix} I_{sc_1} \\ I_{sc_2} \\ I_p \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} I_{pump} \right) dt + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \sigma_{I_{sc}} & 0 \\ 0 & 0 & \sigma_{I_p} \end{bmatrix} d\omega_t, \tag{39}$$

where I_{sc_1} [mU] and I_{sc_2} [mU] represent the subcutaneous layer and deeper tissues, respectively, and I_p [mU/L] represents plasma. I_{pump} is the input from the pump [mU/min]. k_a [min^{-1}] is the absorption rate and k_e [min^{-1}] is the clearance rate of insulin from plasma. V_I is the volume of distribution [L]. $\sigma_{I_{sc}}$ and σ_{I_p} are the standard deviation of the diffusion processes.

The observation equation is formulated through a transformation of the third-state I_p . The log transformation used here is a natural choice since I_p is a concentration which is a nonnegative number. Transformations are discussed in Sect. 4.1. The observation equation is

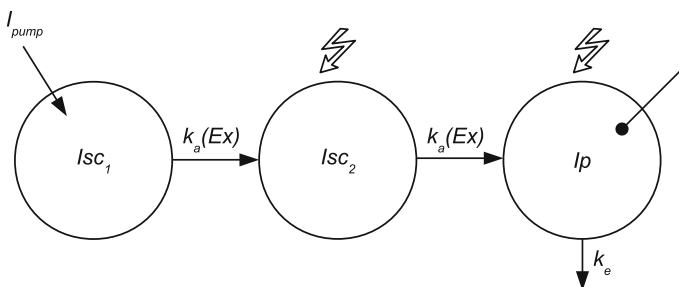


Fig. 6 Illustration of a three-compartment model describing the pharmacokinetics of insulin delivered continuously from an insulin pump. Lightning bolts indicate diffusion terms

$$\log(y_k) = \log(Ip_k) + e_k, \quad (40)$$

where y_k is the observed plasma insulin concentration and $e_k \sim N(0, \xi)$ is the measurement noise. The variance is further modeled such that $\xi = S_{\min} + S$, where S_{\min} is a known hardware specific measurement error variance of the equipment [5]. Note that the measurement error multiplicative in the natural domain of y_k . This works as an approximation of a proportional error model.

The full gray box model is the SDE system equation (39) and the observation equation (40).

Population Parameters

The individual parameters are modeled as a combination of fixed population effects and random individual effects

$$\theta_i = h(\theta_{pop}, Z_i) \cdot e^{\eta_i}, \quad (41)$$

where θ_i is the parameter value for individual i , $h(\cdot)$ is a possibly nonlinear function, θ_{pop} is the overall population parameter (fixed effect), Z_i are covariates (age, weight, gender etc.), and $\eta_i \sim N(0, \Omega)$ is the individual random effect.

For this model, four parameters were modeled with a random effect. The initial values of the two subcutaneous layer states are assumed to be affected by the same variation from the population mean

$$Isc_{10,i} = Isc_{10} \cdot e^{\eta_{i,1}} \quad Isc_{20,i} = Isc_{20} \cdot e^{\eta_{i,1}}.$$

The absorption rate k_a and the clearance rate k_e have separate random effects

$$k_{a,i} = k_a \cdot e^{\eta_{i,2}} \quad k_{e,i} = k_e \cdot e^{\eta_{i,3}}.$$

The volume of distribution V_I is scaled by the weight (kg) of the subject. The weight is a covariate

$$V_{I,i} = V_I \cdot \text{weight}_i.$$

The random effects are assumed Gaussian with

$$\eta_i = [\eta_{i1}, \eta_{i2}, \eta_{i3}] \sim \mathcal{N}(0, \text{diag}(\omega_{Isc}, \omega_{k_a}, \omega_{k_e})).$$

3.3 Exercise Effects

The model is further extended by making the absorption rate k_a dependent on exercising. Two extensions are investigated.

Model A

The first extension specifies k_a as

$$k_a = \bar{k}_a + \alpha \cdot \text{Ex}, \tag{42}$$

where \bar{k}_a is the basal rate and α is the effect of exercise. Ex is a binary input which is 1 when the subject is exercising and otherwise 0.

Model B

The subjects were exercising at two intensities and this extends (42) to

$$k_a = \bar{k}_a + \alpha_{\text{mild}} \cdot \text{Ex}_{\text{mild}} + \alpha_{\text{moderate}} \cdot \text{Ex}_{\text{moderate}}, \tag{43}$$

where \bar{k}_a is the basal rate, α_{mild} and α_{moderate} are the effects of mild and moderate exercise. Ex_{mild} and $\text{Ex}_{\text{moderate}}$ are binary inputs which is 1 during either mild or moderate exercising.

3.4 Model Comparison

The best model is selected by comparing the ML estimates with the likelihood ratio test, AIC, and BIC in Table 2. The base model is nested in both model A and B and model A is nested in B. The nested models can be compared with the likelihood ratio test. Both models A and B explain significantly more of the variability in the data than the base model. Model A is the preferred model based on the likelihood ratio test. The additional improvement in the likelihood with model B is not enough to justify the extra parameter. The difference in AIC and BIC between model A and B relatively small but indicate that model B is to be preferred. The relative likelihood between model A and B is $\exp(0.5 \cdot (1815 - 1817)) = 0.37$ and suggests that model A is 37% as probable as model B [1].

The parameter estimates for all three models are seen in Table 3. For model B, the moderate intensity exercise results in a larger absorption rate than mild exercise.

Table 2 Model comparison using likelihood ratio test, AIC and BIC

Model	df	$-\log(L)$	LRT (p)	AIC	BIC
Base	10	927	–	1878	1799
Model A versus Base	11	897	$<10^{-7}$	1817	1729
Model B versus A	12	895	0.16	1815	1720

Table 3 Parameter estimates from the three models: base, A and C

	Base	Model A	Model B
I_{sc1_0}	87.4	58.3	61.8
I_{sc2_0}	35.9	56.3	52.2
k_a	0.023	0.026	0.024
k_e	0.079	0.077	0.076
σ_{Isc}	2.94	2.61	2.48
σ_{Ip}	0.030	0.027	0.026
S	0.00028	0.00034	0.00075
ω_{Isc}	0.379	0.226	0.299
ω_{ka}	0.122	0.112	0.112
ω_{ke}	0.142	0.150	0.146
α		0.00762	
α_{mild}			0.00961
$\alpha_{moderate}$			0.00515

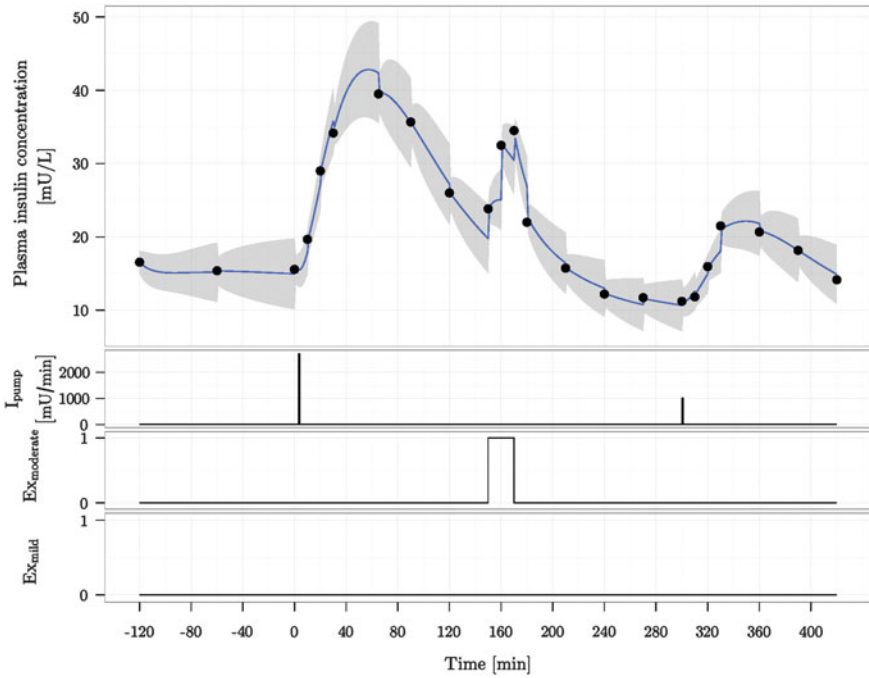


Fig. 7 Top One-step predictions from model A (Blue line). The observations are represented by dots. The gray area indicates 95% prediction interval. Middle and bottom Insulin and exercise inputs

3.5 Predictions

From the three models tried here, model A with a single absorption rate is the best to explain the data. One-step predictions using model A using a single trial of one subject are shown in Fig. 7. In general, the predictions are acceptable and the model does seem to capture the increase related to exercise. Especially, in Fig. 7, the compliance between the predictions and the observations is good. The width of the prediction interval is, however, large in this case. k-step predictions can also easily be calculated using CTSM-R and the `predict` function. A more detailed account of the exercise dependence analysis using population modeling is found in [5].

4 Other Topics

4.1 Transformations

In general, transformations should be applied whenever appropriate, and as all inference with CTSM-R assumes Gaussian random output, this should be ensured by transformations. Transformations can be applied in three different levels (1) state transformations, (2) transformation of observations, and (3) transformation of the parameters. We will briefly discuss each of these types of transformations and refer the interested reader to appropriate literature.

If there are natural restrictions of the state space, e.g., the natural state space is the positive real axis, or some interval, then these restrictions should be included in the SDE description. This implies a formulation of the form

$$dx_t = f(x_t, u_t)dt + \sigma(x_t)dw_t. \quad (44)$$

However, the Kalman filter requires the diffusion term to be independent of the state and therefore we should apply the Lamperti transform;

$$z_t = \int \frac{d\xi}{\sigma(\xi)} \Big|_{\xi=x_t} \quad (45)$$

and use Itô's Lemma to obtain a description where the SDE description is independent of the state (see [18], Paper D for a tutorial on the Lamperti transform, and [19] for a nontrivial application).

The usual comments on transformation of the observations also apply to the SDE models, i.e., the standardized residuals should have constant variance, this should be checked and if the residuals do not have constant variance the observations should be transformed (e.g., using log transformation).

As already discussed in the examples in this chapter, the parameters should be estimated on the real axis (implying e.g., log transformation of positive parameters).

4.2 Identification

We have already seen that the autocorrelation function and the partial autocorrelation functions can be used for identification. If data are not equivalently sampled, one might use linear SDE models on the residuals to identify model order (number of states).

For nonlinear models the usual autocorrelation function is also relevant. Nonlinear dependence in the residuals will almost always include a linear dependence which will appear in the autocorrelation function. It can be shown that some nonlinear functions does not have linear dependence and the autocorrelation functions will fail. Generalizations in the form of lag-dependent and partial lag-dependent functions might then be used instead [20].

Finally identification can be based on random walk identification, where one parameter is formulated as a random walk process and the reconstruction or smoothed parameter is compared with state estimates and/or input to identify possible model extensions (see also [18, 19], paper F, and [11]).

4.3 Simulation/Prediction Models

As we have already seen in the simulation example, misspecification of a model can lead to very poor performance in simulation (long-term prediction) performance of models. A way to ensure reasonable performance in long-term predictions is by forcing the diffusion parameters to be small. This is done by fixing diffusion parameters, see [14] for a discussion about simulation and multistep predictions in SDE-models.

4.4 Testing and Confidence Intervals

Often, in particular for data-rich situations, the standard Wald confidence intervals, as presented directly from CTSM-R, are good approximations of the “true” confidence intervals. These are, however, approximations, and conclusions regarding individual parameters should be based on likelihood ratio tests rather than confidence intervals. In cases where models are not nested, it is recommended to use likelihood-based information criteria (AIC or BIC) for model selection.

Still, confidence intervals provide useful information that should always be reported, also when parameters are significant. But as we saw in the simulation examples, the Wald confidence interval might fail completely (e.g., σ_3 in Models 2 and 3). The problem is that the Wald standard error uses the local curvature of the likelihood (the Hessian), to approximate the uncertainty, and e.g., if the curvature

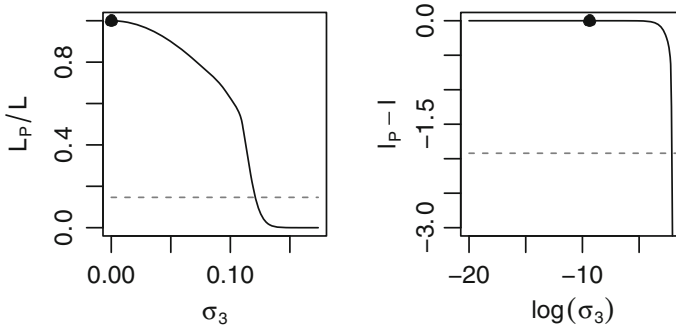


Fig. 8 Profile likelihood for σ_3 in the third-state simulation model of Examples 1–2

is close to zero (see Fig. 8), then the variance of the parameter estimates becomes infinite (as we saw in the examples).

As an alternative, we can calculate profile likelihood confidence intervals (see Fig. 8), we will not go into detail with the calculation of such intervals, but note that the profile likelihood confidence interval is based on the same statistical properties of the likelihood ratio as the likelihood ratio test. In the case of Model 3 of the simulation example, the profile likelihood confidence interval for σ_3 is $[0, 0.12]$, which seems much more reasonable than the values obtained by the Wald approximation. For further reading see [17, 21].

5 Summary

A general framework for modeling physical dynamical systems using stochastic differential equations has been demonstrated. CTSM-R is an efficient and parallelized implementation in the statistical language R. R facilitates easy data handling, visualization, and statistical tests essential for any modeling task. CTSM-R uses maximum likelihood and thus known techniques for model identification and selection can also be used for this framework as demonstrated.

This chapter has demonstrated the principles using linear models with transformations. CTSM-R has been used for a number of nonlinear problems see e.g., [18, 22].

CTSM-R has been extended to include hierarchical modeling. A study of exercise dependence in insulin absorption was modeled with a random effect between the subjects. This is an example of commonly used population modeling in PK/PD.

A detailed user guide and additional examples are available from <http://ctsm.info>.

Appendix A: Extended Kalman Filtering

For nonlinear models the innovation vectors \mathbf{e}_k (or \mathbf{e}_k^i) and their covariance matrices $\Sigma_{k|k-1}^{yy}$ (or $\Sigma_{k|k-1}^{yy,i}$) can be computed recursively by means of the extended Kalman filter (EKF) as outlined in the following.

Consider first the linear time-varying model

$$d\mathbf{X}_t = (\mathbf{A}(\mathbf{u}_t, t, \boldsymbol{\theta})\mathbf{X}_t + \mathbf{B}(\mathbf{u}_t, t, \boldsymbol{\theta})) dt + \boldsymbol{\sigma}(\mathbf{u}_t, t, \boldsymbol{\theta})d\boldsymbol{\omega}_t \quad (46)$$

$$\mathbf{Y}_k = \mathbf{C}(\mathbf{u}_k, t_k, \boldsymbol{\theta})\mathbf{X}_k + \mathbf{e}_k \quad (47)$$

in the following we will use $\mathbf{A}(t)$, $\mathbf{B}(t)$, and $\boldsymbol{\sigma}(t)$ as short-hand notation for $\mathbf{A}(\mathbf{u}_t, t, \boldsymbol{\theta})$, $\mathbf{B}(\mathbf{u}_t, t, \boldsymbol{\theta})$, and $\boldsymbol{\sigma}(\mathbf{u}_t, t, \boldsymbol{\theta})$.

We will restrict ourselves to the initial value problem; solve (46) for $t \in [t_k, t_{k+1}]$ given that the initial condition $X_{t_k} \sim N(\hat{\mathbf{x}}_{k|k}, \Sigma_{k|k}^{xx})$. This is the kind of solution we would get from the ordinary Kalman filter in the update step.

Now if we consider, the transformation

$$\mathbf{Z}_t = e^{-\int_{t_k}^t \mathbf{A}(s)ds} \mathbf{X}_t, \quad (48)$$

then by Itô's Lemma, it can be shown that the process \mathbf{Z}_t is governed by the Itô stochastic differential equation

$$d\mathbf{Z}_t = e^{-\int_{t_k}^t \mathbf{A}(s)ds} \mathbf{B}(t)dt + e^{-\int_{t_k}^t \mathbf{A}(s)ds} \boldsymbol{\sigma}(t)d\boldsymbol{\omega}_t \quad (49)$$

with initial conditions $\mathbf{Z}_{t_k} \sim N(\hat{\mathbf{x}}_{k|k}, \Sigma_{k|k}^{xx})$. The solution to (49) is given by the integral equation

$$\mathbf{Z}_t = \mathbf{Z}_{t_k} + \int_{t_k}^t e^{-\int_{t_k}^u \mathbf{A}(u)du} \mathbf{B}(s)ds + \int_{t_k}^t e^{-\int_{t_k}^s \mathbf{A}(u)du} \boldsymbol{\sigma}(s)d\boldsymbol{\omega}_s \quad (50)$$

Now inserting the inverse of the transformation (48) gives

$$\begin{aligned} \mathbf{X}_t &= e^{\int_{t_k}^t \mathbf{A}(s)ds} \mathbf{X}_0 + e^{\int_{t_k}^t \mathbf{A}(s)ds} \int_{t_k}^t e^{-\int_{t_k}^u \mathbf{A}(u)du} \mathbf{B}(s)ds \\ &\quad + e^{\int_{t_k}^t \mathbf{A}(s)ds} \int_{t_k}^t e^{-\int_{t_k}^s \mathbf{A}(u)du} \boldsymbol{\sigma}(s)d\boldsymbol{\omega}_s \end{aligned} \quad (51)$$

Taking the expectation and variance on both sides of (51) gives

$$E[\mathbf{X}_t] = e^{\int_{t_k}^t \mathbf{A}(s)ds} E[\mathbf{X}_{t_k}] + e^{-\int_{t_k}^t \mathbf{A}(s)ds} \int_{t_k}^t e^{-\int_{t_k}^u \mathbf{A}(u)du} \mathbf{B}(s)ds \quad (52)$$

$$\begin{aligned} V[\mathbf{X}_t] &= e^{\int_{t_k}^t \mathbf{A}(s)ds} V[\mathbf{X}_{t_k}] e^{\int_{t_k}^t \mathbf{A}(s)ds} + e^{\int_{t_k}^t \mathbf{A}(s)ds} V \left[\int_{t_k}^t e^{-\int_{t_k}^s \mathbf{A}(u)du} \boldsymbol{\sigma}(s) d\boldsymbol{\omega}_s \right] e^{\int_{t_k}^t \mathbf{A}(s)ds} \\ &= e^{\int_{t_k}^t \mathbf{A}(s)ds} V[\mathbf{X}_0] e^{\int_{t_k}^t \mathbf{A}(s)ds} \\ &\quad + e^{\int_{t_k}^t \mathbf{A}(s)ds} \int_{t_k}^t e^{-\int_{t_k}^s \mathbf{A}(u)du} \boldsymbol{\sigma}(s) \boldsymbol{\sigma}(s)^T e^{-\int_{t_k}^s \mathbf{A}(u)ds} ds e^{\int_{t_k}^t \mathbf{A}(s)ds}, \end{aligned} \quad (53)$$

where we have used Itô isometry in the second equation for the variance. Now differentiation the above expression w.r.t. time gives

$$\frac{dE[\mathbf{X}_t]}{dt} = \mathbf{A}(t)E[\mathbf{X}_t] + \mathbf{B}(t) \quad (54)$$

$$\frac{dV[\mathbf{X}_t]}{dt} = \mathbf{A}(t)V[\mathbf{X}_t] + V[\mathbf{X}_t]\mathbf{A}(t)^T + \boldsymbol{\sigma}(t)\boldsymbol{\sigma}(t)^T, \quad (55)$$

with initial conditions given by $E[\mathbf{X}_{t_k}] = \hat{\mathbf{x}}_{k|k}$ and $V[\mathbf{X}_{t_k}] = \Sigma_{k|k}^{xx}$.

For the nonlinear case

$$d\mathbf{X}_t = \mathbf{f}(\mathbf{X}_t, \mathbf{u}_t, t, \boldsymbol{\theta})dt + \boldsymbol{\sigma}(\mathbf{u}_t, t, \boldsymbol{\theta})d\boldsymbol{\omega}_t \quad (56)$$

$$\mathbf{Y}_k = \mathbf{h}(\mathbf{X}_k, \mathbf{u}_k, t_k, \boldsymbol{\theta}) + \mathbf{e}_k, \quad (57)$$

we introduce the Jacobian of \mathbf{f} around the expectation of \mathbf{X}_t ($\hat{\mathbf{x}}_t = E[\mathbf{X}_t]$), we will use the following short hand notation

$$\mathbf{A}(t) = \left. \frac{\partial \mathbf{f}(\mathbf{x}, \mathbf{u}_t, t, \boldsymbol{\theta})}{\partial \mathbf{x}} \right|_{\mathbf{x}=\hat{\mathbf{x}}_{t|k}}, \quad \mathbf{f}(t) = \mathbf{f}(\hat{\mathbf{x}}_{t|k}, \mathbf{u}_t, t, \boldsymbol{\theta}) \quad (58)$$

where $\hat{\mathbf{x}}_t$ is the expectation of \mathbf{X}_t at time t , this implies that we can write the first-order Taylor expansion of (56) as

$$d\mathbf{X}_t \approx [\mathbf{f}(t) + \mathbf{A}(t)(\mathbf{X}_t - \hat{\mathbf{x}}_{t|k})] dt + \boldsymbol{\sigma}(t)d\boldsymbol{\omega}_t. \quad (59)$$

Using the results from the linear time-varying system above, we get the following approximate solution to the (59)

$$\frac{dE[\mathbf{X}_t]}{dt} \approx \mathbf{f}(t) \quad (60)$$

$$\frac{dV[\mathbf{X}_t]}{dt} \approx \mathbf{A}(t)V[\mathbf{X}_t] + V[\mathbf{X}_t]\mathbf{A}^T(t) + \boldsymbol{\sigma}(t)\boldsymbol{\sigma}^T(t), \quad (61)$$

with initial conditions $E[\mathbf{X}_{t_k}] = \hat{\mathbf{x}}_{k|k}$ and $V[\mathbf{X}_{t_k}] = \Sigma_{k|k}^{xx}$. Equations (60) and (61) constitute the basis of the prediction step in the Extended Kalman Filter, which for completeness is given below

Theorem 1 (Continuous-discrete time extended Kalman filter) *With given initial conditions for the $\hat{\mathbf{x}}_{1|0} = \mathbf{x}_0$ and $\Sigma_{1|0}^{xx} = \Sigma_0^{xx}$ the extended Kalman filter approximations are given by; the output prediction equations:*

$$\hat{\mathbf{y}}_{k|k-1} = \mathbf{h}(\hat{\mathbf{x}}_{k|k-1}, \mathbf{u}_k, t_k, \boldsymbol{\theta}); \quad \Sigma_{k|k-1}^{yy} = \mathbf{C}_k \Sigma_{k|k-1}^{xx} \mathbf{C}_k^T + \mathbf{S}_k \quad (62)$$

the innovation and Kalman gain equation:

$$\boldsymbol{\varepsilon}_k = \mathbf{y}_k - \hat{\mathbf{y}}_{k|k-1}; \quad \mathbf{K}_k = \Sigma_{k|k-1}^{xx} \mathbf{C}_k^T \left(\Sigma_{k|k-1}^{yy} \right)^{-1} \quad (63)$$

the updating equations:

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k \boldsymbol{\varepsilon}_k; \quad \Sigma_{k|k}^{xx} = \Sigma_{k|k-1}^{xx} - \mathbf{K}_k \Sigma_{k|k-1}^{yy} \mathbf{K}_k^T \quad (64)$$

and the state prediction equations:

$$\frac{d\hat{\mathbf{x}}_{t|k}}{dt} = \mathbf{f}(\hat{\mathbf{x}}_{t|k}, \mathbf{u}_t, t, \boldsymbol{\theta}), \quad t \in [t_k, t_{k+1}[\quad (65)$$

$$\frac{d\Sigma_{t|k}^{xx}}{dt} = \mathbf{A}(t) \Sigma_{t|k}^{xx} + \Sigma_{t|k}^{xx} \mathbf{A}(t)^T + \boldsymbol{\sigma}(t) \boldsymbol{\sigma}(t)^T, \quad t \in [t_k, t_{k+1}[\quad (66)$$

where the following short-hand notation has been applied:

$$\mathbf{A}(t) = \left. \frac{\partial \mathbf{f}(\mathbf{x}, \mathbf{u}_t, t, \boldsymbol{\theta})}{\partial \mathbf{x}} \right|_{\mathbf{x}=\hat{\mathbf{x}}_{t|k-1}}, \quad \mathbf{C}_k = \left. \frac{\partial \mathbf{h}(\mathbf{x}, \mathbf{u}_k, t_k, \boldsymbol{\theta})}{\partial \mathbf{x}} \right|_{\mathbf{x}=\hat{\mathbf{x}}_{k|k-1}} \quad (67)$$

$$\boldsymbol{\sigma}(t) = \boldsymbol{\sigma}(\mathbf{u}_t, t, \boldsymbol{\theta}), \quad \mathbf{S}_k = \mathbf{S}(\mathbf{u}_k, t_k, \boldsymbol{\theta}) \quad (68)$$

The prediction step was covered above and the updating step can be derived from linearization of the observation equation and the projection theorem [6]. From the construction above, it is clear that the approximation is only likely to hold if the nonlinearities are not too strong. This implies that the sampling frequency is fast enough for the prediction equations to be a good approximation and that the accuracy in the observation equation is good enough for the Gaussian approximation to hold approximately. Even though “simulation” through the prediction equations is available in CTSM-R, it is recommended that simulation results are verified (or indeed performed), by real-stochastic simulations (e.g., by simple Euler simulations).

References

1. Burnham, K., Anderson, D.: *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer (2002)
2. CTSM-R (Continuous Time Stochastic Modelling in R). www.ctsm.info
3. DIACON Project. www.diacongroup.org. New Technologies for treatment of Type 1 diabetes
4. Donnet, S., Samson, A.: A review on estimation of stochastic differential equations for pharmacokinetic/pharmacodynamic models. In: *Advanced Drug Delivery Reviews* (2013). doi:10.1016/j.addr.2013.03.005. <http://www.sciencedirect.com/science/article/pii/S0169409X13000501>
5. Duun-Henriksen, A., Juhl, R., Schmidt, S., Nørgaard, K., Madsen, H.: Modelling the effect of exercise on insulin pharmacokinetics in “continuous subcutaneous insulin infusion” treated type 1 diabetes patients. Technical report DTU Compute-Technical Report-2013, Technical University of Denmark (2013)
6. Jazwinski, A.H.: *Stochastic processes and filtering theory*. Dover publications, Inc. (1970)
7. Jelliffe, R., Schumitzky, A., Van Guilder, M.: Population pharmacokinetics/pharmacodynamics modeling; parametric and nonparametric methods. *Ther. Drug Monit.* **22**, 354–365 (2000)
8. Karlsson, M., Beal, S., Sheiner, L.: Three new residual error models for population pk/pd analysis. *J. Pharmacokinet. Pharmacodyn.* **23**, 651–672 (1995)
9. Klim, S., Mortensen, S.B., Kristensen, N.R., Overgaard, R.V., Madsen, H.: Population stochastic modelling (PSM)-an R package for mixed-effects models based on stochastic differential equations. *Comput. Methods Progr. Biomed.* **94**(3), 279–289 (2009). doi:10.1016/j.cmpb.2009.02.001. <http://www.sciencedirect.com/science/article/pii/S0169260709000455>
10. Kristensen, N.R., Madsen, H.: Continuous time stochastic modelling—CTSM 2.3 Mathematics guide. Technical University of Denmark, DTU Informatics, Building 321 (2003). www.ctsm.info
11. Kristensen, N.R., Madsen, H., Jørgensen, S.B.: A method for systematic improvement of stochastic gray-box models. *Comput. Chem. Eng.* **116**, 1431–1449 (2004)
12. Kristensen, N.R., Madsen, H., Jørgensen, S.B.: Parameter estimation in stochastic grey-box models. *Automatica* **40**(2), 225–237 (2004)
13. Lindsey, J., Jones, B., Jarvis, P.: Some statistical issues in modelling pharmacokinetic data. *Stat. Med.* **20**, 2775–2783 (2001)
14. Löwe, R., Mikkelsen, P., Madsen, H.: Stochastic rainfall-runoff forecasting: parameter estimation, multi-step prediction, and evaluation of overflow risk. *Stoch. Environ. Res. Risk Assess.* **28**(3), 505–516 (2014). doi:10.1007/s00477-013-0768-0. (Offprint, no public access)
15. Lv, D., Breton, M.D., Farhy, L.S.: Pharmacokinetics modeling of exogenous glucagon in Type 1 diabetes mellitus patients. *Diabet. Technol. Ther.* **15**(11), 935–941 (2013). doi:10.1089/dia.2013.0150. <http://online.liebertpub.com/globalproxy.cvt.dk/doi/abs/10.1089/dia.2013.0150>
16. Madsen, H.: *Time Series Analysis*. Chapman and Hall (2008)
17. Madsen, H., Thyregod, P.: *Introduction to general and generalized linear models*. Chapman and Hall (2011)
18. Møller, J., Madsen, H.: Stochastic state space modelling of nonlinear systems—with application to marine ecosystems. In: IMM-PHD-2010-246. Technical University of Denmark, DTU Informatics, Building 321 (2010)
19. Møller, J., Phillipsen, K.R., Christensen, L.E., Madsen, H.: Development of a restricted state space stochastic differential equation model for bacterial growth in rich media. *J. Theor. Biol.* **305**, 78–87 (2012). doi:10.1016/j.jtbi.2012.04.015
20. Nielsen, H.A., Madsen, H.: A generalization of some classical time series tools. *Comput. Stat. Data Anal.* **37**, 13–31 (2001)
21. Pawitan, Y.: *In all likelihood: Statistical modelling and inference using likelihood*. Oxford Science Publications (2001)
22. Philipsen, K.R., Christiansen, L.E., Hasman, H., Madsen, H.: Modelling conjugation with stochastic differential equations. *J. Theor. Biol.* **263**(1), 134–142 (2010). doi:10.1016/j.jtbi.2009.11.011

23. Schmidt, S., Finan, D.A., Duun-Henriksen, A.K., Jørgensen, J.B., Madsen, H., Bengtsson, H., Holst, J.J., Madsbad, S., Nørgaard, K.: Effects of everyday life events on glucose, insulin, and glucagon dynamics in continuous subcutaneous insulin infusion treated type 1 diabetes: Collection of clinical data for glucose modeling. *Diabet. Technol. Ther.* **4**(3), 210–217 (2012). doi:[10.1089/dia.2011.0101](https://doi.org/10.1089/dia.2011.0101). <http://online.liebertpub.com/doi/abs/10.1089/dia.2011.0101>
24. Tornøe, C.W., Agersø, H., Jonsson, E.N., Madsen, H., Nielsen, H.A.: Non-linear mixed-effects pharmacokinetic/pharmacodynamic modelling in nlme using differential equations. *Computer Methods and Programs in Biomedicine. Comput. Methods Progr. Biomed.* **76**(1), 31–40 (2004). doi:[10.1016/j.cmpb.2004.01.001](https://doi.org/10.1016/j.cmpb.2004.01.001)
25. Tornøe, C.W., Jacobsen, J., Pedersen, O., Hansen, T., Madsen, H.: Grey-box modelling of pharmacokinetic/pharmacodynamic systems. *J. Pharmacokinet. Pharmacodyn.* **31**(5), 401–417 (2004)
26. Wilinska, M.E., Chassin, L.J., Acerini, C.L., Allen, J.M., Dunger, D.B., Hovorka, R.: Simulation environment to evaluate closed-loop insulin delivery systems in type 1 diabetes. *J. Diabet. Sci. Technol.* **4**(1), 132–144 (2010)