
Research Article

Exact Gradients Improve Parameter Estimation in Nonlinear Mixed Effects Models with Stochastic Dynamics

Helga Kristin Olafsdottir,^{1,2,5} Jacob Leander,^{1,2,3} Joachim Almquist,^{1,3,4} and Mats Jirstrand¹

Received 4 March 2018; accepted 14 May 2018; published online 1 August 2018

Abstract. Nonlinear mixed effects (NLME) modeling based on stochastic differential equations (SDEs) have evolved into a promising approach for analysis of PK/PD data. SDE-NLME models go beyond the realm of standard population modeling as they consider stochastic dynamics, thereby introducing a probabilistic perspective on the state variables. This article presents a summary of the main contributions to SDE-NLME models found in the literature. The aims of this work were to develop an exact gradient version of the first-order conditional estimation (FOCE) method for SDE-NLME models and to investigate whether it enabled faster estimation and better gradient precision/accuracy compared to the use of gradients approximated by finite differences. A simulation-estimation study was set up whereby finite difference approximations of the gradients of each level were interchanged with the exact gradients. Following previous work, the uncertainty of the state variables was accounted for using the extended Kalman filter (EKF). The exact gradient FOCE method was implemented in Mathematica 11 and evaluated on SDE versions of three common PK/PD models. When finite difference gradients were replaced by exact gradients at both FOCE levels, relative runtimes improved between 6- and 32-fold, depending on model complexity. Additionally, gradient precision/accuracy was significantly better in the exact gradient case. We conclude that parameter estimation using FOCE with exact gradients can successfully be applied to SDE-NLME models.

KEY WORDS: extended Kalman filter; first-order conditional estimation (FOCE); nonlinear mixed effects modeling; sensitivity equations; stochastic differential equations.

INTRODUCTION

Nonlinear mixed effects (NLME) models are used to describe populations of individuals that behave qualitatively similar, but where every individual has its own quantitative characteristics. These models have been highly applicable in pharmacokinetics (PK) and pharmacodynamics (PD) (1–4). Physiological systems are often modeled with a continuous-time deterministic model and noisy observations that are discrete in time. However, since the mechanisms of the physiological systems of interest are often not completely

understood, a natural extension is to account for the uncertainty in the dynamics. This is the motivation behind extending the NLME models with yet a stochastic part (5).

Stochastic differential equations (SDEs) greatly extend the descriptive power of ordinary differential equations (ODEs) that are usually used to encode dynamical system models in pharmacometrics. An ODE model can be turned into an SDE model by the addition of possibly non-linearly scaled noise terms. These additional terms can be thought of as a type of slack-variable introduced to pick-up potential discrepancies between the mechanistically modeled dynamics and unknown but indirectly observed dynamical effects. In other words, these noise terms in the dynamical system equations represent the lumped effect of all not explicitly mechanistically modeled effects in the system. This includes turning model parameters into state variables with random dynamics, which for instance could be used to model inter occasion variability (6) or time-dependent changes in parameters where the trend is unknown *a priori*, and for the estimation of an unknown input signal (7).

The application of SDEs to model uncertainty in dynamics has long been used in other fields, such as finance and control theory, and various parameter estimation

¹ Fraunhofer-Chalmers Centre, Chalmers Science Park, Gothenburg, Sweden.

² Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden.

³ Quantitative Clinical Pharmacology, Early Clinical Development, IMED Biotech Unit, AstraZeneca, Gothenburg, Sweden.

⁴ Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden.

⁵ To whom correspondence should be addressed. (e-mail: khelga@chalmers.se)

methods have been developed (8). However, the sparsity and irregular sampling of PK/PD data has made it difficult to directly apply these parameter estimation methods. Nevertheless, some attempts to develop methods for parameter estimation in NLME models with stochastic dynamics have been successful, using for example (i) Bayesian inference (9,10), (ii) expectation maximization (EM) methods (11–13), and (iii) by expanding the traditional gradient-based estimation methods using Kalman filters (14,15). These methods have been used for several PK/PD applications (16–19). This paper focuses on gradient-based methods. A general overview of some of the related method and application papers is shown in Table I.

ODE models are a subset of SDE models because SDEs reduce to ODEs when the system noise (the Gaussian process covariance) is decreased towards zero. Ideally, the dynamical system under study is well characterized and the mechanistic terms present in the dynamical system model equations give an accurate description of the model’s dynamical behavior. Then ODEs are a suitable model description. To account for incorrect model definitions, Kristensen *et al.* (20) introduced a method for evaluating the choice of a deterministic population model described

with ODEs and iteratively improving it by using an SDE model to pinpoint where in the dynamical model equations misspecification may be present. A similar idea had previously been used for gray box models (21).

Instead of opting to find a deterministic model, Tornøe *et al.* (14) and Overgaard *et al.* (15) captured the additional uncertainty due to model misspecification, oversimplifications, and approximations by using SDEs in NLME models. The well-known parameter estimation method first-order conditional estimation (FOCE) (22) was expanded to account for the new mathematical framework. This involved state variable estimation using extended Kalman filters (EKF) (23). The algorithm was initially implemented in NONMEM (14) and later implemented in both MATLAB (24) and R (25).

Optimizing the FOCE approximation of the log likelihood using gradient-based methods, such as the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (26), requires sensitivities of the state variables of the underlying model. However, NLME models defined by differential equations often lack an analytic solution. Hence, the traditional way to compute the gradient is to use a finite difference approximation. An alternative approach is to obtain the gradients

Table I. Timeline showing the main contributions to nonlinear mixed effects models with stochastic differential equations and to the S-FOCE method

	Method	Application	
2004	Grey-box modelling		Tornøe et al. [21]
2005	NLME models with SDEs, estimation algorithm		Overgaard et al. [15]
	SDEs for PK/PD model development		Kristensen et al. [20]
2007	NONMEM implementation		Tornøe et al. [14]
	Matlab framework		Mortensen et al. [24]
	Interleukin-21 effects on thermoregulation in monkeys	Overgaard et al. [16]
2009	R package		Klim et al. [25]
2011	Compartmental model for leucine kinetics	Berglund et al. [17]
2013	Measurement time uncertainty		Krengel et al. [18]
2014	SDEs for single individual		Leander et al. [28]
2015	NLME modeling using SDEs		Leander et al. [29]
	S-FOCE for NLME models using ODEs		Almquist et al. [27]
2016	A metformin case study	Matzuka et al. [19]

utilizing the sensitivity equations for the original set of differential equations. The algorithm originally proposed by Almquist *et al.* (27) is an extension of the FOCE algorithm that uses these sensitivity-based gradients for population NLME models based on ODEs. This algorithm is here referred to as S-FOCE. The sensitivity approach was also applied by Leander *et al.* (28) for single individual models using SDEs. For the single individual, the likelihood was approximated using the EKF, which in turn requires a first-order differentiation of the EKF. These ideas have been combined for parameter estimation of NLME models using SDEs (29). In that study, the exact gradient approach was used instead of the traditional finite difference approach when using the FOCE-EKF approach by using the mixed symbolic and numeric algebra capabilities in Mathematica (30). However, the details of the algorithm have until now not been presented.

This paper extends the S-FOCE algorithm to a general NLME model based on SDEs using the EKF approach. This involves both the first- and the second-order sensitivities of the EKF. Earlier stage of this work has been presented by (31). The method is evaluated on synthetically generated datasets from common PK/PD models with uncertainty in the dynamics, in addition to the traditional observation uncertainty and interindividual variability. Relative computational time of the sensitivity-based gradients is compared to the finite difference approach. Moreover, the precision and the accuracy of both gradient approaches are analyzed.

The paper is structured as follows. First, the necessary theory for the proposed extension is introduced, including the approximate population likelihood and the EKF, followed by the proposed algorithm. Second, the methods for evaluating and comparing of the algorithm are described. Third, the results of the comparison are presented, showing speedup in estimation time as well as increased precision and accuracy of gradients for the proposed algorithm. Finally, the results are discussed along with an outlook towards implementation of the algorithm in an existing parameter estimation framework.

THEORETICAL BACKGROUND

This section introduces the theory needed for extending the S-FOCE method to NLME models with stochastic dynamics (SDE-NLME).

Nonlinear Mixed Effects Models Based on Stochastic Differential Equations

NLME models can be defined using SDEs to describe the dynamics of the continuous-time state variables \mathbf{x} . The continuous-time, stochastic dynamics are modeled as

$$\begin{aligned} dx_i &= f(\mathbf{x}_i, \mathbf{u}_i, t, \boldsymbol{\theta}, \boldsymbol{\eta}_i)dt + \mathbf{G}(\mathbf{x}_i, \mathbf{u}_i, t, \boldsymbol{\theta}, \boldsymbol{\eta}_i)d\mathbf{W}_i \\ \mathbf{x}_i(0) &= \mathbf{x}_0(\boldsymbol{\theta}, \boldsymbol{\eta}_i) \end{aligned} \tag{1}$$

with discrete observations modeled using an observation function \mathbf{h} and Gaussian noise $\mathbf{e}_{ij} \sim N(\mathbf{0}, \boldsymbol{\Sigma}(\mathbf{x}_{ij}, \mathbf{u}_i, t_{ij}, \boldsymbol{\theta}, \boldsymbol{\eta}_i))$,

$$\mathbf{y}_{ij} = \mathbf{h}(\mathbf{x}_{ij}, \mathbf{u}_i, t_{ij}, \boldsymbol{\theta}, \boldsymbol{\eta}_i) + \mathbf{e}_{ij}. \tag{2}$$

The function $f(\mathbf{x}_i, \mathbf{u}_i, t, \boldsymbol{\theta}, \boldsymbol{\eta}_i)$, describing the deterministic part of the dynamics, is called the drift function and the part introducing randomness to the state variables, $\mathbf{G}(\mathbf{x}_i, \mathbf{u}_i, t, \boldsymbol{\theta}, \boldsymbol{\eta}_i)d\mathbf{W}_i$, is called the system noise. Here, \mathbf{W}_i is defined as a Wiener process, with $d\mathbf{W}_i \sim N(\mathbf{0}, dt\mathbf{I})$. The indices i and j correspond to individual i and its j -th observation at time t_{ij} for each individual. Here, $\boldsymbol{\theta}$, $\boldsymbol{\eta}_i$, and \mathbf{u}_i are, respectively, the fixed effects, random effects, and known inputs, such as covariates including dosage. The random effects $\boldsymbol{\eta}_i$ are chosen to be normally distributed with mean zero and covariance $\boldsymbol{\Omega}(\boldsymbol{\theta})$. This way of formulating the NLME model gives three sources of variability in the response (29), namely (i) observation noise, \mathbf{e}_{ij} ; (ii) system noise, $\mathbf{G}(\mathbf{x}_i, \mathbf{u}_i, t, \boldsymbol{\theta}, \boldsymbol{\eta}_i)d\mathbf{W}_i$; and (iii) parameter variability, $\boldsymbol{\eta}_i$.

The Approximate Population Likelihood

Since \mathbf{x}_i are stochastic processes, the distribution of \mathbf{y}_{ij} changes if conditioned on previous observations. Let $\mathcal{Y}_{i(j-1)} = \{\mathbf{v}_{i1}, \mathbf{v}_{i2}, \dots, \mathbf{v}_{i(j-1)}\}$ denote those observations. The residuals $\boldsymbol{\varepsilon}_{ij}$ are defined as

$$\boldsymbol{\varepsilon}_{ij} = \mathbf{v}_{ij} - \hat{\mathbf{y}}_{ij} \tag{3}$$

with the expected observation value $\hat{\mathbf{y}}_{ij}$ and covariance \mathbf{R}_{ij} conditioned on $\mathcal{Y}_{i(j-1)}$ and $\boldsymbol{\theta}$ defined as

$$\begin{aligned} \hat{\mathbf{y}}_{ij} &= E[\mathbf{y}_{ij} | \mathcal{Y}_{i(j-1)}, \boldsymbol{\theta}] \\ \mathbf{R}_{ij} &= \text{Cov}[\mathbf{y}_{ij} | \mathcal{Y}_{i(j-1)}, \boldsymbol{\theta}]. \end{aligned} \tag{4}$$

As shown in, e.g., (14,29), the combined likelihood L for all individuals simplifies to

$$L(\boldsymbol{\theta}) = \prod_{i=1}^N \int \exp(l_i(\boldsymbol{\theta}, \boldsymbol{\eta}_i)) d\boldsymbol{\eta}_i, \tag{5}$$

where

$$\begin{aligned} -2l_i &= \sum_{j=1}^{n_i} \left(\boldsymbol{\varepsilon}_{ij}^T \mathbf{R}_{ij}^{-1} \boldsymbol{\varepsilon}_{ij} + \log \det(2\pi \mathbf{R}_{ij}) \right) \\ &\quad + \boldsymbol{\eta}_i^T \boldsymbol{\Omega}^{-1} \boldsymbol{\eta}_i + \log \det(2\pi \boldsymbol{\Omega}), \end{aligned} \tag{6}$$

and where the indirect dependence of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are suppressed for simplicity.

The FOCE approximation of the population likelihood becomes

$$\log L_F(\boldsymbol{\theta}) = \sum_{i=1}^N \left(l_i(\boldsymbol{\theta}, \boldsymbol{\eta}_i^*) - \frac{1}{2} \log \det \left(\frac{-\mathbf{H}_i(\boldsymbol{\theta}, \boldsymbol{\eta}_i^*)}{2\pi} \right) \right). \tag{7}$$

Here, $\boldsymbol{\eta}_i^*$ maximizes the individual likelihood l_i for a given $\boldsymbol{\theta}$, and \mathbf{H}_i is a first-order approximation of the Hessian Δl_i . Further details on this approximation can be found in, e.g., (27).

The Extended Kalman Filter

The SDEs introduce uncertainty to the time evolution of the state variables of the system. The EKF can be used to estimate the state variables, as suggested in (14,15,29). The continuous-discrete EKF is a state-space estimator for the continuous-discrete state-space models of the form introduced in Eq. (1), with the exception that the state variables should be independent from both the observation noise and the system noise (32). The EKF estimates the conditional expectation of the state $\hat{\mathbf{x}}_{i(j|j)} = E[\mathbf{x}_{it_j} | \mathbf{Y}_j, \boldsymbol{\phi}_i]$ and its covariance $\mathbf{P}_{i(j|j)} = \text{Cov}[\mathbf{x}_{it_j} | \mathbf{Y}_j, \boldsymbol{\phi}_i]$. Here, $\boldsymbol{\phi}_i = (\boldsymbol{\theta}, \boldsymbol{\eta}_i, \mathbf{u}_i)$ is a vector containing all parameters corresponding to individual i .

In the rest of this section, the notation is simplified by omitting the individual index i . The drift function \mathbf{f} and the observation function \mathbf{h} from Eqs. (2) and (1) are linearized by introducing

$$\begin{aligned} \mathbf{A}_t &= \left. \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right|_{\mathbf{x}=\hat{\mathbf{x}}_{t|j-1}}, t \in [t_{j-1}, t_j] \\ \mathbf{C}_j &= \left. \frac{\partial \mathbf{h}}{\partial \mathbf{x}} \right|_{\mathbf{x}=\hat{\mathbf{x}}_{j|j-1}} \end{aligned} \tag{8}$$

The EKF has two main steps, a time-update step and a measurement-update step (33). In the time-update step, the state variables and covariance are predicted using all previous observations. This is done by solving the differential equations

$$\begin{aligned} \frac{d\hat{\mathbf{x}}_{t|j-1}}{dt} &= \mathbf{f}(\hat{\mathbf{x}}_{t|j-1}, \mathbf{u}_t, t, \boldsymbol{\phi}), \quad t \in [t_{j-1}, t_j] \\ \frac{d\mathbf{P}_{t|j-1}}{dt} &= \mathbf{A}_t \mathbf{P}_{t|j-1} + \mathbf{P}_{t|j-1} \mathbf{A}_t^T + \mathbf{G}_t \mathbf{G}_t^T, \quad t \in [t_{j-1}, t_j]. \end{aligned} \tag{9}$$

The above prediction gives

$$\begin{aligned} \hat{\mathbf{y}}_j &= \mathbf{h}(\hat{\mathbf{x}}_{j|j-1}, \mathbf{u}_j, t_j, \boldsymbol{\phi}) \\ \mathbf{R}_j &= \mathbf{C}_j \mathbf{P}_{j|j-1} \mathbf{C}_j^T + \boldsymbol{\Sigma}_j. \end{aligned} \tag{10}$$

In the measurement-update step, the prediction is used to compute the *Kalman gain*

$$\mathbf{K}_j = \mathbf{P}_{j|j-1} \mathbf{C}_j^T \mathbf{R}_j^{-1} \tag{11}$$

which is used to update the state and its covariance from the current observation

$$\begin{aligned} \hat{\mathbf{x}}_{j|j} &= \hat{\mathbf{x}}_{j|j-1} + \mathbf{K}_j \boldsymbol{\varepsilon}_j \\ \mathbf{P}_{j|j} &= \mathbf{P}_{j|j-1} - \mathbf{K}_j \mathbf{R}_j \mathbf{K}_j^T. \end{aligned} \tag{12}$$

The filter is initialized with

$$\begin{aligned} \hat{\mathbf{x}}_{1|0} &= \mathbf{x}_0 \\ \mathbf{P}_{1|0} &= \mathbf{P}_0. \end{aligned} \tag{13}$$

PROPOSED ALGORITHM

This section presents the extension of the S-FOCE method, originally proposed by Almquist *et al.* (27) for parameter estimation of general NLME models based on ODEs, to the use of SDEs. The method utilizes sensitivity equations to calculate exact gradients for a gradient-based optimization of the model parameters. The sensitivities are the derivatives of a function with regard to its parameters. The sensitivities needed for the gradient-based optimization in addition to the sensitivities in the existing S-FOCE algorithm are presented. A detailed derivation of terms needed for the sensitivity equations can be found in the [Appendix](#).

The fixed effects parameters $\boldsymbol{\theta}$ can be estimated by maximizing the approximate population likelihood (APL) in Eq. (7) using gradient-based methods. This means a need for computing (27)

$$\frac{d \log L_f(\boldsymbol{\theta})}{d \boldsymbol{\theta}_m} = \sum_{i=1}^N \left(\frac{d l_i(\boldsymbol{\theta}, \boldsymbol{\eta}_i^*)}{d \boldsymbol{\theta}_m} - \frac{1}{2} \text{tr} \left[\mathbf{H}_i^{-1}(\boldsymbol{\theta}, \boldsymbol{\eta}_i^*) \frac{d \mathbf{H}_i(\boldsymbol{\theta}, \boldsymbol{\eta}_i^*)}{d \boldsymbol{\theta}_m} \right] \right). \tag{14}$$

Unlike the ODE case, the APL is formulated in terms of conditional expected values of state variables and their covariances given by the EKF. Hence, sensitivities of entities computed by the EKF equations should be computed. These are the first-order sensitivities

$$\begin{aligned} \frac{d \boldsymbol{\varepsilon}_j}{d \boldsymbol{\eta}_{ik}}, \quad \frac{d \boldsymbol{\varepsilon}_j}{d \boldsymbol{\theta}_n}, \quad \frac{d \mathbf{R}_{ij}}{d \boldsymbol{\eta}_{ik}}, \quad \frac{d \mathbf{R}_{ij}}{d \boldsymbol{\theta}_n}, \\ \frac{d \hat{\mathbf{x}}_{i(j|j-1)}}{d \boldsymbol{\eta}_{ik}}, \quad \frac{d \hat{\mathbf{x}}_{i(j|j-1)}}{d \boldsymbol{\theta}_n} \end{aligned} \tag{15}$$

and the second-order sensitivities

$$\begin{aligned} \frac{d^2 \boldsymbol{\varepsilon}_j}{d \boldsymbol{\eta}_{ik} d \boldsymbol{\theta}_n}, \quad \frac{d^2 \boldsymbol{\varepsilon}_j}{d \boldsymbol{\eta}_{ik} d \boldsymbol{\eta}_{il}}, \quad \frac{d^2 \mathbf{R}_{ij}}{d \boldsymbol{\eta}_{ik} d \boldsymbol{\theta}_n}, \quad \frac{d^2 \mathbf{R}_{ij}}{d \boldsymbol{\eta}_{ik} d \boldsymbol{\eta}_{il}}, \\ \frac{d^2 \hat{\mathbf{x}}_{i(j|j-1)}}{d \boldsymbol{\eta}_{ik} d \boldsymbol{\theta}_n}, \quad \frac{d^2 \hat{\mathbf{x}}_{i(j|j-1)}}{d \boldsymbol{\eta}_{ik} d \boldsymbol{\eta}_{il}}. \end{aligned} \tag{16}$$

In addition, the sensitivities of the state variable covariance \mathbf{P} are needed. The derivations of the results can be found in the [Appendix](#). The resulting SDE-EKF extension of S-FOCE is outlined in Algorithm 1.

METHODS

Initializing the EKF

The use of the EKF requires initial values of the state variable \mathbf{x}_0 and state variable covariance \mathbf{P}_0 . The state variable is set to the initial value of the system equations. To estimate the covariance of the state, different methods can be used. Setting $\mathbf{P}_0 = \mathbf{I}$ is a common but arbitrary way of initializing the EKF. Tornøe *et al.* (14) set \mathbf{P}_0 to the integral of the Wiener processes and system dynamics between the first

Algorithm 1 Extended S-FOCE parameter estimation algorithm for NLME models based on SDEs

```

1:  $s := 0, \theta_s := \theta_{start}$  ▷ Initialize the algorithm
2: for all individuals do
3:    $u := 0, \eta_s^u := 0$ 
4: end for
5: repeat ▷ Solve the outer problem
6:   for all individuals do
7:      $u := 0$ 
8:     repeat ▷ Solve the inner problem
9:       Solve for  $x$  and the sensitivities  $d\hat{x}/d\eta, \partial P/\partial\eta, \partial P/\partial\hat{x}$ 
10:      Compute  $l$  and  $dl/d\eta$ 
11:      Update  $\eta_s^{u+1}$  according to BFGS
12:       $u := u + 1$ 
13:     until  $\eta_s^*$  is obtained
14:   end for
15:   for all individuals do
16:     Set  $\eta := \eta_s^*$ 
17:     Solve for  $x$  and the sensitivities  $d\hat{x}/d\eta, d\hat{x}/d\theta, d^2\hat{x}/d\eta^2, d^2\hat{x}/d\eta d\theta, \partial P/\partial\eta, \partial P/\partial\theta,$ 
        $\partial^2 P/\partial\eta^2, \partial^2 P/\partial\eta\partial\theta, \partial^2 P/\partial\hat{x}\partial\eta, \partial^2 P/\partial\hat{x}\partial\theta, \partial^2 P/\partial\hat{x}^2$ 
18:   end for
19:   Compute  $\log L_F$  and  $d\log L_f/d\theta$ 
20:   Update  $\theta_{s+1}$  according to BFGS
21:   for all individuals do ▷ Set the starting values for inner problem
22:      $\eta_{s+1}^0 := \eta_s^*$ 
23:   end for  $s := s + 1$ 
24: until convergence of  $\theta$ 

```

two observations to get a representative value. However, in the case of PK/PD models, one often starts in a steady state and can solve the initial covariance analytically from Eq. (9).

Models and Data

To evaluate the proposed method, a simulation-estimation study was set up. Data was simulated from three common PK and PK/PD models: a two-compartmental PK model (M1), a PK/PD model consisting of a two-compartmental PK together with a direct PD response model (M2), and a PK/PD model consisting of a two-compartmental PK with an indirect PD response model (M3). All model dynamics are described using SDEs, with a one-dimensional Wiener process describing uncertainty of the absorption reaction. The number of state variables and parameters is shown in Table II. Each experimental setup included 50 individuals divided into five dose groups. Representative values were chosen for the simulation in compatible units. The sampling times are chosen to take the characteristics of the PK and PD responses into account. A full description of

the model and experimental design can be found in the [Appendix](#).

Comparison

Algorithm 1 consists of two levels of optimization, an inner level to find the empirical Bayes estimates η_i^* given θ ,

Table II. Overview of the benchmark models

Model	M1	M2	M3
Fixed effects parameters (θ)	10	14	16
Random effects parameters (η)	2	3	3
State variables (x)	3	3	4
Outputs (y)	1	2	2
Wiener processes	1	1	1
ODEs per individual, inner problem	27	36	56
ODEs per individual, outer problem	333	621	1078
Individuals	50	50	50

and an outer level to find the optimal fixed effects θ^* . Both levels of optimization use gradient-based methods. Three different ways of computing gradients are considered. One uses sensitivities in both levels of optimization ($S-S$), another uses sensitivities in the inner level and finite differences in the outer level ($S-F$), and the last uses finite differences in both levels of optimization ($F-F$). Furthermore, the finite differences are either computed according to forward differences

$$\frac{f(\theta_k(1+10^{-q})) - f(\theta_k)}{\theta_k 10^{-q}}, \quad (17)$$

or central differences

$$\frac{f(\theta_k(1+10^{-q})) - f(\theta_k(1-10^{-q}))}{2\theta_k 10^{-q}}, \quad (18)$$

where q controls the relative step length and f denotes the function to be differentiated. The proposed algorithm is evaluated in terms of runtime and in terms of precision and accuracy of the gradient. The initial values of the optimization are chosen within 15% from the true value, drawn from a uniform distribution. The algorithm is implemented and run using Mathematica 11.0 (30). Calculations are performed on a workstation with a 4.00-GHz Intel Core i7-6700K CPU and 32.0 GB of RAM. A running version is available upon request.

Timing Comparison

The timing comparison is done in a similar way as the previously performed time comparison for S-FOCE with ODEs (27). For a given model and a relevant simulated dataset, the total evaluation time is calculated using each of the above approaches to obtain the gradient. To isolate the effect of computing the different gradients, they were calculated for the same set of points. The points were chosen as the points from the gradient-based optimization using the $S-S$ gradient approach. This ensures that equally many evaluations were performed for each gradient. For the finite differences, the step length is kept constant, by letting $q=4$, corresponding to relative step length of 0.0001.

Precision Comparison

Precision is compared only for finite differences in the outer level of optimization. For the inner level, both methods use sensitivity-based gradients. The gradients are computed at the found optimum, θ^* , using the $S-F$ and $S-S$ gradients. The step length varies, with q ranging from 3.2 to 6.4. To capture the effect of new realizations on the numerical precision in $\log L_F(\theta)$, the gradient is computed 300 times for each q using different randomized starting values of η for the inner level of optimization. The starting values were drawn from a uniform distribution between -0.3 and 0.3 .

RESULTS

Faster Estimation

The proposed algorithm was used to estimate the parameter values for models M1–M3. The resulting parameter estimates using the $S-S$ gradients along with the relative standard errors can be found in Table III. For each of the finite difference methods and models (M1–M3), the runtime of the parameter estimation was measured for all three previously mentioned optimization schemes. The speedup in each case is the relative gain when going from an $F-F$ scheme to either of the $S-F$ or $S-S$ schemes. This is shown in Fig. 1. As the models become more complex in terms of number of fixed effects and number of state variables, the relative benefit of using sensitivities instead of finite differences increases. This holds both for forward differences and central differences. However, for the central differences, the advantage is more significant, since the central difference gradient requires approximately double the amount of calls to the likelihood function, compared to forward differences.

Improved Precision and Accuracy

The gradient error was computed as the difference between the obtained value and the average value of the values obtained using the sensitivity-based gradient method. The exact method does not utilize a finite difference and is independent of q .

As shown in Fig. 2, the sensitivity-based gradient is more precise than either of the other two methods. For large step size (small q), the gradients computed using either forward or central differences are biased but relatively precise. However, as shown in Fig. 3, the precision of the finite difference methods is not better than the precision of the sensitivity-based method. For small step size (large q), the accuracy increases but both methods suffer from a significant loss of precision on a relative scale compared to the sensitivity-based method.

Figure 3 shows the standard deviation of the gradient error as a function of the step size. The linear dependency of the standard deviation on q can be reasoned directly from Eqs. (17) and (18), by assuming a numerical error term in relation to each calculation of the function value that is independent of the step size. The forward finite difference gradient $\Delta f/\Delta\theta_k$ can be written as

$$\frac{\Delta f}{\Delta\theta_k} = \frac{(f(\theta_k(1+10^{-q})) + \epsilon_1) - (f(\theta_k) + \epsilon_2)}{\theta_k 10^{-q}}, \quad (19)$$

where ϵ_1, ϵ_2 are independent, identically distributed, numerical error terms with mean zero and variance σ^2 . The total variance of the forward finite difference gradient due to measurement error is therefore

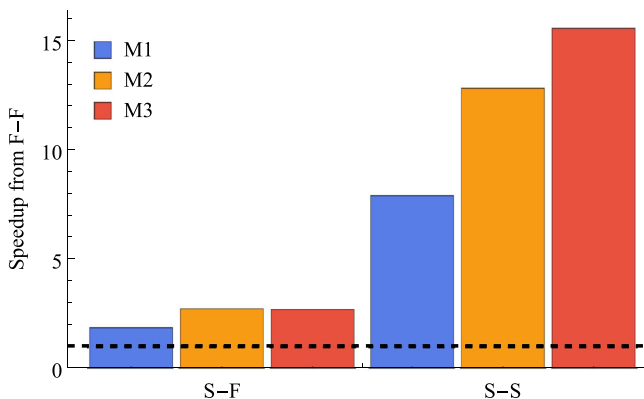
$$\text{Var}\left(\frac{\Delta f}{\Delta\theta_k}\right) = \frac{2\sigma^2}{\theta_k^2 10^{-2q}}. \quad (20)$$

Table III. Parameter values used for simulating data and parameter estimates for the benchmark models. relative standard errors in percent are shown in parenthesis. the parameters that are not a part of the model are denoted with a dash (-)

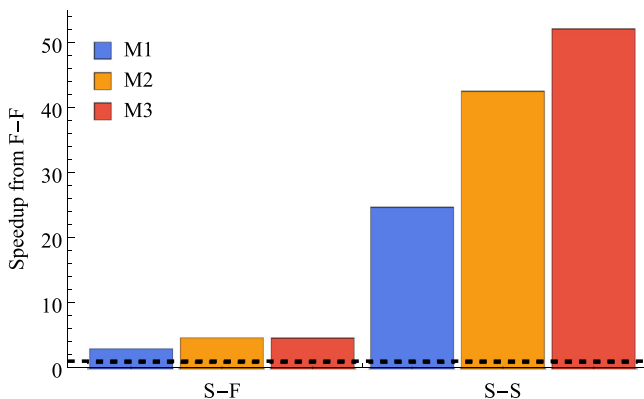
Parameter	Description	True value	M1	M2	M3
k_a	First-order absorption	1.5	1.6 (2.5)	1.5 (2.2)	1.5 (2.3)
CL	Clearance from central compartment	10	8.6 (4.4)	9.6 (2.8)	9.6 (2.8)
Q	Inter-compartmental clearance	5	5.3 (2.7)	5.0 (1.5)	5.0 (1.5)
V_2	Volume of central compartment	50	51 (1.2)	50 (1.1)	50 (1.1)
V_3	Volume of peripheral compartment	80	99 (11)	83 (2.6)	85 (2.9)
I_{max}	Maximum inhibition	0.8	-	0.8 (0.56)	0.83 (1.2)
IC_{50}	Concentration at half maximal inhibition	50	-	50 (3.5)	55 (4.4)
k_{in}	Zero-order production rate constant	50	-	-	48 (2.3)
k_{out}	First-order elimination rate constant	0.5	-	-	0.48 (2.3)
g_1	Relative system noise standard deviation	0.1	0.1 (24)	0.1 (15)	0.1 (2.6)
σ_1	Additive error model parameter PK	1	1.0 (12)	1.0 (11)	1.0 (11)
σ_2	Proportional error model parameter PK	0.1	0.1 (2.7)	0.1 (2.6)	0.075 (4.1)
σ_3	Additive error model parameter PD	5	-	5.2 (2.0)	5.16 (2.0)
ω_{11}^2	Interindividual variance k_a	0.01	0.008 (49)	0.007 (47)	0.0056 (55)
ω_{22}^2	Interindividual variance CL	0.09	0.13 (15)	0.08 (14)	0.077 (14)
ω_{33}^2	Interindividual variance IC_{50}	0.04	-	0.06 (21)	0.056 (23)

Taking the logarithm of this gradient’s standard deviation yields

$$\log_{10} \left(\sqrt{\text{Var} \left(\frac{\Delta f}{\Delta \theta_k} \right)} \right) = q + \log_{10}[\sigma] + \frac{1}{2} \log_{10}[2] - \log_{10}[\theta_k], \quad (21)$$



a Forward finite differences



b Central finite differences

Fig. 1. Comparison of speedup from one estimation scheme (F-F) to another (S-F and S-S). The dotted line represents the speedup baseline. The gain increases with the increased complexity of the model

which is linear in q with unit slope. Similar results can be derived for central difference gradients.

DISCUSSION

Misspecification of the dynamics in population modeling can be identified and quantified through an SDE-NLME approach. This is accomplished by estimating the magnitude of the system noise terms of the SDEs. These terms serve as a source of uncertainty that affects the time evolution of the model state variables, and they are used in parallel with the classical stochastic model components for the residual error and the between-subject variability.

In this work, a novel method for parameter estimation in SDE-NLME models based on exact gradients is presented. It is evaluated by performing parameter estimation for PK/PD models of different complexity, and comparing the computational time, precision, and accuracy to a standard approach based on a finite difference gradient. The method performs

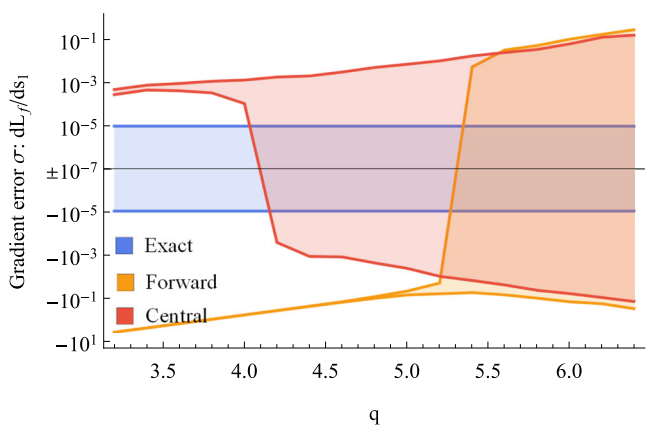


Fig. 2. The 10th and 90th percentiles of the sample-based gradient errors of the exact method (blue), or using forward (red) or central (yellow) finite differences plotted as a function of the finite difference step size, quantified as 10^{-q} . Errors of both signs are displayed using a double logarithmic y-axis originating from $\pm 10^{-7}$

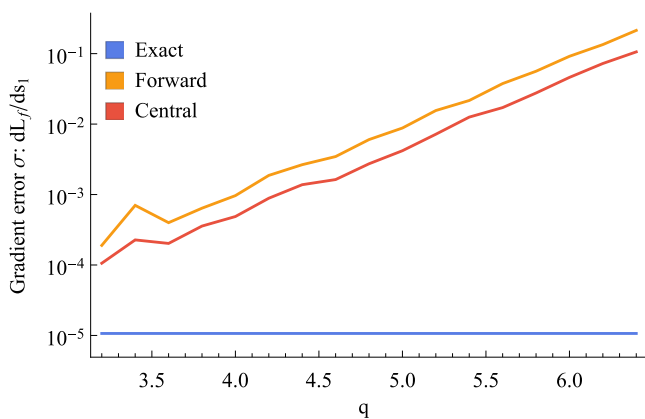


Fig. 3. Sample standard deviation of the gradient using the exact method (blue), forward differences (red), or central differences (yellow), plotted as a function of the finite difference step size, quantified as 10^{-q}

well and shows potential for fast and robust analysis of PK/PD data.

Parameter estimation for NLME PK/PD models can be a time-consuming task, especially when the model of the physiological system is defined by ODEs. The computational burden increases even more if the mathematical framework is expanded to an SDE-NLME setting. By computing the exact gradient of the FOCE likelihood, the need for repeatedly having to solve the inner optimization problem for each perturbation of the fixed effect parameters is avoided, significantly decreasing computational time complexity. As seen in Fig. 1, larger and more complex models of this study benefit more from exact gradients. This suggests that the speedup provided by the exact gradient computations in general will be largest for complex models where this time-saving matters the most. Similar to the S-FOCE method proposed by Almquist *et al.* (27), the algorithm presented in this paper can easily be parallelized over individuals, further increasing its feasibility.

The exact gradients are computed from numerical solutions of ODEs and may therefore contain numerical errors. However, they are still considered exact since the numerical integration does not introduce any bias, and since the numerical error can be made arbitrarily small by controlling the tolerance of the integrator. The quality of the gradients computed with finite differences was evaluated by comparison to the mean of the exactly computed gradient (Fig. 2). For large finite difference step sizes, the relative precision is high but the gradient is biased. Although the accuracy of the finite difference gradient increased with smaller step size, the sample standard deviation clearly shows a simultaneous loss of precision. This demonstrates a fundamental problem with the finite difference method: it is hard to obtain both good precision and accuracy at the same time. Central differences may improve issues with bias, but still suffer from precision issues to a similar extent as forward differences. Exact gradients, on the other hand, elegantly circumvent the step length conundrum inherent to the finite difference approximation. They are unbiased per construction and the absolute precision is always comparatively high. Although exact gradients outperform the finite difference gradients when compared as such, it is unknown whether

exact gradients improve robustness in the parameter estimation step, or in the covariance step, which ultimately is what matters. It may appear plausible that this indeed is the case, but future research is required to prove it.

The EKF has been the standard way of handling stochastic dynamics in NLME models. To fulfill the prerequisites of the EKF, the model state variables must be independent of both the observation noise and the system noise (32). However, in PK/PD models, it is often more realistic to assume proportional noise (1,34). A pragmatic way around this issue is to replace the state variables with their conditional expectation in the model definition. For some models, one could also consider the Lamperti transformation (35) for eliminating the state variable dependency of the system noise. Another filter that might also be applicable is the unscented Kalman filter (UKF), introduced by (36), which instead of linearization of the nonlinear system, uses a so-called unscented transformation. The UKF has been shown to perform very well for a variety of nonlinear models (37), and should be considered as an alternative for state variable estimation in SDE-NLME models.

The EKF was initialized by assuming steady state and solving Eq. (9). In the models chosen for the simulation-estimation study, the steady-state approach is trivial, with $P_0 = \mathbf{0}$. This way of initializing the EKF uses information about the particular model and is superior to using simulations of the initial behavior (14) or simply choosing an arbitrary initialization.

To gain impact in the pharmacometric community, two important steps remain. First, additional applications of SDE-NLME models are needed to further demonstrate the benefits of PK/PD models with uncertain dynamics. Some applications have already been mentioned in this work (6,16,17,19). Second, for novel methods to be widely used, they must be implemented in software with large user-bases, such as NONMEM, Phoenix, and Monolix (22,38,39), or open-source initiatives such as nlmixr (40). Tornøe *et al.* (14) have previously reported on the implementation of SDE-NLME in the industry-standard software NONMEM. The prediction differential equations Eq. (9) were then introduced as a system of ODEs and event tags were introduced to account for the time-update and measurement-update steps. In version 7.4, NONMEM introduced a FAST option to perform the optimization using the exact gradient approach introduced for ODE models by Almquist *et al.* (27). To our understanding, the combination of the recent implementation of the FAST method in NONMEM and the filter implementation of Tornøe *et al.* (14) should be enough for provide this functionality in NONMEM. However, it requires the user to explicitly state the update equations of the Kalman equations. It is only through such wide availability that SDE-NLME methods will get acceptance and popularity among PK/PD modelers.

CONCLUSION

A feasible method for parameter estimation of SDE-NLME models that extends S-FOCE has been proposed. It provides shorter computational time compared to previously used finite difference gradient-based methods. The gradients computed during optimization of the APL are both more precise and more accurate than those computed using finite differences.

ACKNOWLEDGEMENTS

This project has been supported by the Swedish Foundation for Strategic Research, which is gratefully acknowledged.

APPENDIX

Derivation of EKF Sensitivities

This section derives the equations needed for performing both the first- and second-order sensitivities calculation and thereby obtaining the gradients needed for extension of the S-FOCE method to the case of SDEs.

First-Order Sensitivities

The sensitivities $d\boldsymbol{\epsilon}_{ij}/d\boldsymbol{\eta}_{ik}$ and $d\mathbf{R}_{ij}/d\boldsymbol{\eta}_{ik}$ need to be determined. Using the chain rule gives

$$\begin{aligned} \frac{d\boldsymbol{\epsilon}_{ij}}{d\boldsymbol{\eta}_{ik}} &= \frac{d(\mathbf{v}_{ij} - \hat{\mathbf{y}}_{ij})}{d\boldsymbol{\eta}_{ik}} \\ &= -\frac{d\hat{\mathbf{y}}_{ij}}{d\boldsymbol{\eta}_{ik}} \\ &= -\left(\frac{\partial \mathbf{h}}{\partial \boldsymbol{\eta}_{ik}} + \frac{\partial \mathbf{h}}{\partial \hat{\mathbf{x}}_{i(j|j-1)}} \frac{d\hat{\mathbf{x}}_{i(j|j-1)}}{d\boldsymbol{\eta}_{ik}} \right) \end{aligned}$$

and

$$\frac{d\mathbf{R}_{ij}}{d\boldsymbol{\eta}_{ik}} = \frac{\partial \mathbf{R}_{ij}}{\partial \boldsymbol{\eta}_{ik}} + \frac{\partial \mathbf{R}_{ij}}{\partial \hat{\mathbf{x}}_{i(j|j-1)}} \frac{d\hat{\mathbf{x}}_{i(j|j-1)}}{d\boldsymbol{\eta}_{ik}}. \quad (23)$$

The derivatives

$$\frac{\partial \mathbf{h}}{\partial \boldsymbol{\eta}_{ik}}, \text{ and } \frac{\partial \mathbf{h}}{\partial \hat{\mathbf{x}}_{i(j|j-1)}} \quad (24)$$

can be found directly from the definition of \mathbf{h} . However, the derivatives

$$\frac{d\hat{\mathbf{x}}_{i(j|j-1)}}{d\boldsymbol{\eta}_{ik}}, \frac{\partial \mathbf{R}_{ij}}{\partial \boldsymbol{\eta}_{ik}}, \text{ and } \frac{\partial \mathbf{R}_{ij}}{\partial \hat{\mathbf{x}}_{i(j|j-1)}}. \quad (25)$$

remain to be found. In further calculations, the individual index i will be suppressed.

Differentiating Predicted Expected State Variables. Start by considering $\frac{d\hat{\mathbf{x}}_{i(j|j-1)}}{d\boldsymbol{\eta}_k}$. This derivative can be obtained by solving the sensitivity equation

$$\frac{d}{dt} \left(\frac{d\hat{\mathbf{x}}_{t|j}}{d\boldsymbol{\eta}_k} \right) = \frac{\partial \mathbf{f}}{\partial \boldsymbol{\eta}_k} + \frac{\partial \mathbf{f}}{\partial \hat{\mathbf{x}}_{t|j}} \frac{d\hat{\mathbf{x}}_{t|j}}{d\boldsymbol{\eta}_k}, \quad t \in [t_j, t_{j+1}] \quad (26)$$

$$\frac{d\hat{\mathbf{x}}_{t|j}}{d\boldsymbol{\eta}_k}(t_j) = \frac{d\hat{\mathbf{x}}_{j|j}}{d\boldsymbol{\eta}_k}$$

where

$$\hat{\mathbf{x}}_{j|j} = \hat{\mathbf{x}}_{j|j-1} + \mathbf{K}_j \boldsymbol{\epsilon}_j \quad (27)$$

and thus,

$$\frac{d\hat{\mathbf{x}}_{j|j}}{d\boldsymbol{\eta}_k} = \frac{d\hat{\mathbf{x}}_{j|j-1}}{d\boldsymbol{\eta}_k} + \frac{d\mathbf{K}_j}{d\boldsymbol{\eta}_k} \boldsymbol{\epsilon}_j + \mathbf{K}_j \frac{d\boldsymbol{\epsilon}_j}{d\boldsymbol{\eta}_k} \quad (28)$$

Differentiating Predicted State Variable Covariance. Now consider the other two remaining derivatives, $\frac{\partial \mathbf{R}_j}{\partial \boldsymbol{\eta}_k}$ and $\frac{\partial \mathbf{R}_j}{\partial \hat{\mathbf{x}}_{j|j-1}}$.

The equation

$$\mathbf{R}_j = \mathbf{C}_j \mathbf{P}_{j|j-1} \mathbf{C}_j^T + \boldsymbol{\Sigma}_j \quad (29)$$

gives

$$\frac{\partial \mathbf{R}_j}{\partial \boldsymbol{\eta}_k} = \frac{\partial \mathbf{C}_j}{\partial \boldsymbol{\eta}_k} \mathbf{P}_{j|j-1} \mathbf{C}_j^T + \mathbf{C}_j \frac{\partial \mathbf{P}_{j|j-1}}{\partial \boldsymbol{\eta}_k} \mathbf{C}_j^T + \mathbf{C}_j \mathbf{P}_{j|j-1} \frac{\partial \mathbf{C}_j^T}{\partial \boldsymbol{\eta}_k} + \frac{\partial \boldsymbol{\Sigma}_j}{\partial \boldsymbol{\eta}_k}. \quad (30)$$

From the definition of $\boldsymbol{\Sigma}_j$ and \mathbf{h} , $\partial \boldsymbol{\Sigma}_j / \partial \boldsymbol{\eta}_k$ and $\partial \mathbf{C}_j / \partial \boldsymbol{\eta}_k$ can be directly obtained. The remaining derivative is $\partial \mathbf{P}_{j|j-1} / \partial \boldsymbol{\eta}_k$. By definition,

$$\mathbf{P}_{1|0} = \mathbf{P}_0 \quad (31)$$

so

$$\frac{\partial \mathbf{P}_{1|0}}{\partial \boldsymbol{\eta}_k} = \frac{\partial \mathbf{P}_0}{\partial \boldsymbol{\eta}_k} \quad (32)$$

directly follows. For positive integers j , the following sensitivity equation for $\mathbf{P}_{j+1|j}$

$$\begin{aligned} \frac{d}{dt} \left(\frac{\partial \mathbf{P}_{t|j}}{\partial \boldsymbol{\eta}_k} \right) &= \frac{\partial \mathbf{A}_t}{\partial \boldsymbol{\eta}_k} \mathbf{P}_{t|j} + \mathbf{A}_t \frac{\partial \mathbf{P}_{t|j}}{\partial \boldsymbol{\eta}_k} + \mathbf{P}_{t|j} \frac{\partial \mathbf{A}_t^T}{\partial \boldsymbol{\eta}_k} + \frac{\partial \mathbf{P}_{t|j}}{\partial \boldsymbol{\eta}_k} \mathbf{A}_t^T \\ &\quad + \frac{\partial \mathbf{G}_t}{\partial \boldsymbol{\eta}_k} \mathbf{G}_t^T + \mathbf{G}_t \frac{\partial \mathbf{G}_t^T}{\partial \boldsymbol{\eta}_k}, \quad t \in [t_j, t_{j+1}] \quad (33) \\ \frac{\partial \mathbf{P}_{t|j}}{\partial \boldsymbol{\eta}_k}(t_j) &= \frac{\partial \mathbf{P}_{j|j}}{\partial \boldsymbol{\eta}_k} \end{aligned}$$

where

$$\frac{\partial \mathbf{P}_{j|j}}{\partial \boldsymbol{\eta}_k} = \frac{\partial \mathbf{P}_{j|j-1}}{\partial \boldsymbol{\eta}_k} - \left(\frac{\partial \mathbf{K}_j}{\partial \boldsymbol{\eta}_k} \mathbf{R}_j \mathbf{K}_j^T + \mathbf{K}_j \frac{\partial \mathbf{R}_j}{\partial \boldsymbol{\eta}_k} \mathbf{K}_j^T + \mathbf{K}_j \mathbf{R}_j \frac{\partial \mathbf{K}_j^T}{\partial \boldsymbol{\eta}_k} \right) \quad (34)$$

In the same way as for $\frac{\partial \mathbf{R}_j}{\partial \boldsymbol{\eta}_k}$, the equations for $\frac{\partial \mathbf{R}_j}{\partial \hat{\mathbf{x}}_{j|j-1}}$ are obtained by replacing $\boldsymbol{\eta}_k$ with $\hat{\mathbf{x}}_{j|j-1}$.

Differentiating the Kalman Gain. In the above cases, the derivative and partial derivative of the Kalman gain \mathbf{K}_j are required. This can be done by first noting that

$$\frac{d\mathbf{K}_j}{d\boldsymbol{\eta}_k} = \frac{\partial\mathbf{K}_j}{\partial\boldsymbol{\eta}_k} + \frac{\partial\mathbf{K}_j}{\partial\hat{\mathbf{x}}_{j|j-1}} \frac{d\hat{\mathbf{x}}_{j|j-1}}{d\boldsymbol{\eta}_k} \quad (35)$$

and then calculating the partial derivative of \mathbf{K}_j with respect to $\boldsymbol{\eta}_k$ as follows:

$$\begin{aligned} \frac{\partial\mathbf{K}_j}{\partial\boldsymbol{\eta}_k} &= \frac{\partial}{\partial\boldsymbol{\eta}_k} \left(\mathbf{P}_{j|j-1} \mathbf{C}_j^T \mathbf{R}_j^{-1} \right) \\ &= \frac{\partial\mathbf{P}_{j|j-1}}{\partial\boldsymbol{\eta}_k} \mathbf{C}_j^T \mathbf{R}_j^{-1} + \mathbf{P}_{j|j-1} \frac{\partial\mathbf{C}_j^T}{\partial\boldsymbol{\eta}_k} \mathbf{R}_j^{-1} + \mathbf{P}_{j|j-1} \mathbf{C}_j^T \frac{\partial\mathbf{R}_j^{-1}}{\partial\boldsymbol{\eta}_k} \\ &= \frac{\partial\mathbf{P}_{j|j-1}}{\partial\boldsymbol{\eta}_k} \mathbf{C}_j^T \mathbf{R}_j^{-1} + \mathbf{P}_{j|j-1} \frac{\partial\mathbf{C}_j^T}{\partial\boldsymbol{\eta}_k} \mathbf{R}_j^{-1} \\ &\quad - \mathbf{P}_{j|j-1} \mathbf{C}_j^T \mathbf{R}_j^{-1} \frac{\partial\mathbf{R}_j}{\partial\boldsymbol{\eta}_k} \mathbf{R}_j^{-1} \\ &= \frac{\partial\mathbf{P}_{j|j-1}}{\partial\boldsymbol{\eta}_k} \mathbf{C}_j^T \mathbf{R}_j^{-1} + \mathbf{P}_{j|j-1} \frac{\partial\mathbf{C}_j^T}{\partial\boldsymbol{\eta}_k} \mathbf{R}_j^{-1} - \mathbf{K}_j \frac{\partial\mathbf{R}_j}{\partial\boldsymbol{\eta}_k} \mathbf{R}_j^{-1} \\ &= \left(\frac{\partial\mathbf{P}_{j|j-1}}{\partial\boldsymbol{\eta}_k} \mathbf{C}_j^T + \mathbf{P}_{j|j-1} \frac{\partial\mathbf{C}_j^T}{\partial\boldsymbol{\eta}_k} - \mathbf{K}_j \frac{\partial\mathbf{R}_j}{\partial\boldsymbol{\eta}_k} \right) \mathbf{R}_j^{-1} \end{aligned} \quad (36)$$

In the same way, it follows that

$$\frac{\partial\mathbf{K}_j}{\partial\hat{\mathbf{x}}_{j|j-1}} = \left(\frac{\partial\mathbf{P}_{j|j-1}}{\partial\hat{\mathbf{x}}_{j|j-1}} \mathbf{C}_j^T + \mathbf{P}_{j|j-1} \frac{\partial\mathbf{C}_j^T}{\partial\hat{\mathbf{x}}_{j|j-1}} - \mathbf{K}_j \frac{\partial\mathbf{R}_j}{\partial\hat{\mathbf{x}}_{j|j-1}} \right) \mathbf{R}_j^{-1} \quad (37)$$

Moreover, the derivative of \mathbf{K}_j with respect to $\boldsymbol{\theta}$ is calculated in the same way as the derivative with respect to $\boldsymbol{\eta}$.

Second-Order Sensitivities

This section derives the equations needed to compute the derivatives

$$\frac{d^2\boldsymbol{\varepsilon}_j}{d\boldsymbol{\eta}_k d\boldsymbol{\theta}_n} \text{ and } \frac{d^2\mathbf{R}_j}{d\boldsymbol{\eta}_k d\boldsymbol{\theta}_n}. \quad (38)$$

The chain rule yields

$$\begin{aligned} \frac{d^2\boldsymbol{\varepsilon}_j}{d\boldsymbol{\eta}_k d\boldsymbol{\theta}_n} &= -\frac{d}{d\boldsymbol{\theta}_n} \left(\frac{\partial\mathbf{h}}{\partial\boldsymbol{\eta}_k} + \frac{\partial\mathbf{h}}{\partial\hat{\mathbf{x}}_{j|j-1}} \frac{d\hat{\mathbf{x}}_{j|j-1}}{d\boldsymbol{\eta}_k} \right) \\ &= -\left(\frac{\partial^2\mathbf{h}}{\partial\boldsymbol{\eta}_k \partial\boldsymbol{\theta}_n} + \frac{\partial^2\mathbf{h}}{\partial\boldsymbol{\eta}_k \partial\hat{\mathbf{x}}_{j|j-1}} \frac{d\hat{\mathbf{x}}_{j|j-1}}{d\boldsymbol{\theta}_n} \right. \\ &\quad \left. + \left(\frac{\partial^2\mathbf{h}}{\partial\hat{\mathbf{x}}_{j|j-1} \partial\boldsymbol{\theta}_n} + \frac{\partial^2\mathbf{h}}{\partial\hat{\mathbf{x}}_{j|j-1}^2} \frac{d\hat{\mathbf{x}}_{j|j-1}}{\boldsymbol{\theta}_n} \right) \frac{d\hat{\mathbf{x}}_{j|j-1}}{d\boldsymbol{\eta}_k} \right. \\ &\quad \left. + \frac{\partial\mathbf{h}}{\partial\hat{\mathbf{x}}_{j|j-1}} \frac{d^2\hat{\mathbf{x}}_{j|j-1}}{d\boldsymbol{\eta}_k d\boldsymbol{\theta}_n} \right) \end{aligned} \quad (39)$$

and

$$\begin{aligned} \frac{d^2\mathbf{R}_j}{d\boldsymbol{\eta}_k d\boldsymbol{\theta}_n} &= \frac{d}{d\boldsymbol{\theta}_n} \left(\frac{\partial\mathbf{R}_j}{\partial\boldsymbol{\eta}_k} + \frac{\partial\mathbf{R}_j}{\partial\hat{\mathbf{x}}_{j|j-1}} \frac{d\hat{\mathbf{x}}_{j|j-1}}{d\boldsymbol{\eta}_k} \right) \\ &= \frac{\partial^2\mathbf{R}_j}{\partial\boldsymbol{\eta}_k \partial\boldsymbol{\theta}_n} + \frac{\partial^2\mathbf{R}_j}{\partial\boldsymbol{\eta}_k \partial\hat{\mathbf{x}}_{j|j-1}} \frac{d\hat{\mathbf{x}}_{j|j-1}}{d\boldsymbol{\theta}_n} \\ &\quad + \left(\frac{\partial^2\mathbf{R}_j}{\partial\hat{\mathbf{x}}_{j|j-1} \partial\boldsymbol{\theta}_n} + \frac{\partial^2\mathbf{R}_j}{\partial\hat{\mathbf{x}}_{j|j-1}^2} \frac{d\hat{\mathbf{x}}_{j|j-1}}{d\boldsymbol{\theta}_n} \right) \frac{d\hat{\mathbf{x}}_{j|j-1}}{d\boldsymbol{\eta}_k} \\ &\quad + \frac{\partial\mathbf{R}_j}{\partial\hat{\mathbf{x}}_{j|j-1}} \frac{d^2\hat{\mathbf{x}}_{j|j-1}}{d\boldsymbol{\eta}_k d\boldsymbol{\theta}_n} \end{aligned} \quad (40)$$

Note that in the above calculations, $\boldsymbol{\theta}_n$ can be replaced by $\boldsymbol{\eta}_l$ in order to obtain the derivatives

$$\frac{d^2\boldsymbol{\varepsilon}_j}{d\boldsymbol{\eta}_k d\boldsymbol{\eta}_l} \text{ and } \frac{d^2\mathbf{R}_j}{d\boldsymbol{\eta}_k d\boldsymbol{\eta}_l}. \quad (41)$$

Second Derivative of Predicted Expected State.

$$\begin{aligned} \frac{d}{dt} \left(\frac{d^2\hat{\mathbf{x}}_{t|j}}{d\boldsymbol{\eta}_k d\boldsymbol{\theta}_n} \right) &= \frac{\partial^2\mathbf{f}}{\partial\boldsymbol{\eta}_k \partial\boldsymbol{\theta}_n} + \frac{\partial^2\mathbf{f}}{\partial\boldsymbol{\eta}_j \partial\hat{\mathbf{x}}_{t|j}} \frac{d\hat{\mathbf{x}}_{t|j-1}}{d\boldsymbol{\theta}_n} \\ &\quad + \left(\frac{\partial^2\mathbf{f}}{\partial\hat{\mathbf{x}}_{t|j} \partial\boldsymbol{\theta}_n} + \frac{\partial^2\mathbf{f}}{\partial\hat{\mathbf{x}}_{t|j}^2} \frac{d\hat{\mathbf{x}}_{t|j}}{d\boldsymbol{\theta}_n} \right) \frac{d\hat{\mathbf{x}}_{t|j}}{d\boldsymbol{\eta}_k} \\ &\quad + \frac{\partial\mathbf{f}}{\partial\hat{\mathbf{x}}_{t|j}} \frac{d^2\hat{\mathbf{x}}_{t|j}}{d\boldsymbol{\eta}_k \partial\boldsymbol{\theta}_n}, \quad t \in [t_j, t_{j+1}] \\ \frac{d^2\hat{\mathbf{x}}_{t|j}}{d\boldsymbol{\eta}_k \partial\boldsymbol{\theta}_n}(t_j) &= \frac{d^2\hat{\mathbf{x}}_{j|j}}{d\boldsymbol{\eta}_k \partial\boldsymbol{\theta}_n} \end{aligned} \quad (42)$$

where

$$\frac{d^2\hat{\mathbf{x}}_{j|j}}{d\boldsymbol{\eta}_k \partial\boldsymbol{\theta}_n} = \frac{d^2\hat{\mathbf{x}}_{j|j-1}}{d\boldsymbol{\eta}_k \partial\boldsymbol{\theta}_n} + \frac{d^2\mathbf{K}_j}{d\boldsymbol{\eta}_k d\boldsymbol{\theta}_n} \boldsymbol{\varepsilon}_j + \frac{d\mathbf{K}_j}{d\boldsymbol{\eta}_k} \frac{\boldsymbol{\varepsilon}_j}{d\boldsymbol{\theta}_n} + \frac{d\mathbf{K}_j}{d\boldsymbol{\theta}_n} \frac{\boldsymbol{\varepsilon}_j}{d\boldsymbol{\eta}_k} + \mathbf{K}_j \frac{d\boldsymbol{\varepsilon}_j}{d\boldsymbol{\eta}_k d\boldsymbol{\theta}_n} \quad (43)$$

Second Derivative of Predicted State Covariance.

$$\begin{aligned} \frac{\partial^2\mathbf{R}_j}{\partial\boldsymbol{\eta}_k \partial\boldsymbol{\theta}_n} &= \frac{\partial}{\partial\boldsymbol{\theta}_n} \left(\frac{\partial\mathbf{C}_j}{\partial\boldsymbol{\eta}_k} \mathbf{P}_{j|j-1} \mathbf{C}_j^T + \mathbf{C}_j \frac{\partial\mathbf{P}_{j|j-1}}{\partial\boldsymbol{\eta}_k} \mathbf{C}_j^T + \mathbf{C}_j \mathbf{P}_{j|j-1} \frac{\partial\mathbf{C}_j^T}{\partial\boldsymbol{\eta}_k} + \frac{\partial\boldsymbol{\Sigma}_j}{\partial\boldsymbol{\eta}_k} \right) \\ &= \frac{\partial^2\mathbf{C}_j}{\partial\boldsymbol{\eta}_k \partial\boldsymbol{\theta}_n} \mathbf{P}_{j|j-1} \mathbf{C}_j^T + \frac{\partial\mathbf{C}_j}{\partial\boldsymbol{\eta}_k} \frac{\partial\mathbf{P}_{j|j-1}}{\partial\boldsymbol{\theta}_n} \mathbf{C}_j^T + \frac{\partial\mathbf{C}_j}{\partial\boldsymbol{\eta}_k} \mathbf{P}_{j|j-1} \frac{\partial\mathbf{C}_j^T}{\partial\boldsymbol{\theta}_n} \\ &\quad + \frac{\partial\mathbf{C}_j}{\partial\boldsymbol{\theta}_n} \frac{\partial\mathbf{P}_{j|j-1}}{\partial\boldsymbol{\eta}_k} \mathbf{C}_j^T + \mathbf{C}_j \frac{\partial^2\mathbf{P}_{j|j-1}}{\partial\boldsymbol{\eta}_k \partial\boldsymbol{\theta}_n} \mathbf{C}_j^T + \mathbf{C}_j \frac{\partial\mathbf{P}_{j|j-1}}{\partial\boldsymbol{\eta}_k} \frac{\partial\mathbf{C}_j^T}{\partial\boldsymbol{\theta}_n} \\ &\quad + \frac{\partial\mathbf{C}_j}{\partial\boldsymbol{\theta}_n} \mathbf{P}_{j|j-1} \frac{\partial\mathbf{C}_j^T}{\partial\boldsymbol{\eta}_k} + \mathbf{C}_j \frac{\partial\mathbf{P}_{j|j-1}}{\partial\boldsymbol{\theta}_n} \frac{\partial\mathbf{C}_j^T}{\partial\boldsymbol{\eta}_k} + \mathbf{C}_j \mathbf{P}_{j|j-1} \frac{\partial^2\mathbf{C}_j^T}{\partial\boldsymbol{\eta}_k \partial\boldsymbol{\theta}_n} \\ &\quad + \frac{\partial^2\boldsymbol{\Sigma}_j}{\partial\boldsymbol{\eta}_k \partial\boldsymbol{\theta}_n} \end{aligned} \quad (44)$$

with

$$\begin{aligned} \frac{d}{dt} \left(\frac{\partial^2 \mathbf{P}_{t|j}}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n} \right) &= \frac{\partial^2 \mathbf{A}_t}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n} \mathbf{P}_{t|j} + \frac{\partial \mathbf{A}_t}{\partial \boldsymbol{\eta}_k} \frac{\partial \mathbf{P}_{t|j}}{\partial \boldsymbol{\theta}_n} + \frac{\partial \mathbf{A}_t}{\partial \boldsymbol{\theta}_n} \frac{\partial \mathbf{P}_{t|j}}{\partial \boldsymbol{\eta}_k} + \mathbf{A}_t \frac{\partial^2 \mathbf{P}_{t|j}}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n} \\ &+ \frac{\partial \mathbf{P}_{t|j}}{\partial \boldsymbol{\theta}_n} \frac{\partial \mathbf{A}_t^T}{\partial \boldsymbol{\eta}_k} + \mathbf{P}_{t|j} \frac{\partial^2 \mathbf{A}_t^T}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n} + \frac{\partial^2 \mathbf{P}_{t|j}}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n} \mathbf{A}_t^T + \frac{\partial \mathbf{P}_{t|j}}{\partial \boldsymbol{\eta}_k} \frac{\partial \mathbf{A}_t^T}{\partial \boldsymbol{\theta}_n} \\ &+ \frac{\partial^2 \mathbf{G}_t}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n} \mathbf{G}_t^T + \frac{\partial \mathbf{G}_t}{\partial \boldsymbol{\eta}_k} \frac{\partial \mathbf{G}_t^T}{\partial \boldsymbol{\theta}_n} + \frac{\partial \mathbf{G}_t}{\partial \boldsymbol{\theta}_n} \frac{\partial \mathbf{G}_t^T}{\partial \boldsymbol{\eta}_k} + \mathbf{G}_t \frac{\partial^2 \mathbf{G}_t^T}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n}, \quad t \in [t_j, t_{j+1}] \\ \frac{\partial^2 \mathbf{P}_{t|j}}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n} (t_j) &= \frac{\partial^2 \mathbf{P}_{j|j}}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n} \end{aligned} \quad (45)$$

where

$$\begin{aligned} \frac{\partial^2 \mathbf{P}_{j|j}}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n} &= \frac{\partial^2 \mathbf{P}_{j|j-1}}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n} \\ &- \left(\frac{\partial^2 \mathbf{K}_j}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n} \mathbf{R}_j \mathbf{K}_j^T + \frac{\partial \mathbf{K}_j}{\partial \boldsymbol{\eta}_k} \frac{\partial \mathbf{R}_j}{\partial \boldsymbol{\theta}_n} \mathbf{K}_j^T + \frac{\partial \mathbf{K}_j}{\partial \boldsymbol{\theta}_n} \mathbf{R}_j \frac{\partial \mathbf{K}_j^T}{\partial \boldsymbol{\eta}_k} \right. \\ &+ \frac{\partial \mathbf{K}_j}{\partial \boldsymbol{\theta}_n} \frac{\partial \mathbf{R}_j}{\partial \boldsymbol{\eta}_k} \mathbf{K}_j^T + \mathbf{K}_j \frac{\partial^2 \mathbf{R}_j}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n} \mathbf{K}_j^T + \mathbf{K}_j \frac{\partial \mathbf{R}_j}{\partial \boldsymbol{\eta}_k} \frac{\partial \mathbf{K}_j^T}{\partial \boldsymbol{\theta}_n} \\ &\left. + \frac{\partial \mathbf{K}_j}{\partial \boldsymbol{\theta}_n} \mathbf{R}_j \frac{\partial \mathbf{K}_j^T}{\partial \boldsymbol{\eta}_k} + \mathbf{K}_j \frac{\partial \mathbf{R}_j}{\partial \boldsymbol{\theta}_n} \frac{\partial \mathbf{K}_j^T}{\partial \boldsymbol{\eta}_k} + \mathbf{K}_j \mathbf{R}_j \frac{\partial^2 \mathbf{K}_j^T}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n} \right) \end{aligned} \quad (46)$$

The derivatives $\frac{\partial^2 \mathbf{R}_j}{\partial \boldsymbol{\eta}_k \partial \hat{\mathbf{x}}_{j|j-1}}$, $\frac{\partial^2 \mathbf{R}_j}{\partial \hat{\mathbf{x}}_{j|j-1} \partial \boldsymbol{\theta}_n}$, and $\frac{\partial^2 \mathbf{R}_j}{\partial \hat{\mathbf{x}}_{j|j-1}^2}$ are calculated in the same way. In the special case of $\frac{\partial^2 \mathbf{R}_j}{\partial \hat{\mathbf{x}}_{j|j-1}^2}$, the equations can be simplified to

$$\begin{aligned} \frac{\partial^2 \mathbf{R}_j}{\partial \hat{\mathbf{x}}_{j|j-1}^2} &= \frac{\partial^2 \mathbf{C}_j}{\partial \hat{\mathbf{x}}_{j|j-1}^2} \mathbf{P}_{j|j-1} \mathbf{C}_j^T + \mathbf{C}_j \frac{\partial^2 \mathbf{P}_{j|j-1}}{\partial \hat{\mathbf{x}}_{j|j-1}^2} \mathbf{C}_j^T + \mathbf{C}_j \mathbf{P}_{j|j-1} \frac{\partial^2 \mathbf{C}_j^T}{\partial \hat{\mathbf{x}}_{j|j-1}^2} \\ &+ 2 \left(\frac{\partial \mathbf{C}_j}{\partial \hat{\mathbf{x}}_{j|j-1}} \frac{\partial \mathbf{P}_{j|j-1}}{\partial \hat{\mathbf{x}}_{j|j-1}} \mathbf{C}_j^T + \frac{\partial \mathbf{C}_j}{\partial \hat{\mathbf{x}}_{j|j-1}} \mathbf{P}_{j|j-1} \frac{\partial \mathbf{C}_j^T}{\partial \hat{\mathbf{x}}_{j|j-1}} \right. \\ &\left. + \mathbf{C}_j \frac{\partial \mathbf{P}_{j|j-1}}{\partial \hat{\mathbf{x}}_{j|j-1}} \frac{\partial \mathbf{C}_j^T}{\partial \hat{\mathbf{x}}_{j|j-1}} \right) + \frac{\partial^2 \boldsymbol{\Sigma}_j}{\partial \hat{\mathbf{x}}_{j|j-1}^2} \end{aligned} \quad (47)$$

with

$$\begin{aligned} \frac{d}{dt} \left(\frac{\partial^2 \mathbf{P}_{t|j}}{\partial \hat{\mathbf{x}}_{j|j-1}^2} \right) &= \frac{\partial^2 \mathbf{A}_t}{\partial \hat{\mathbf{x}}_{j|j-1}^2} \mathbf{P}_{t|j} + \mathbf{A}_t \frac{\partial^2 \mathbf{P}_{t|j}}{\partial \hat{\mathbf{x}}_{j|j-1}^2} + \mathbf{P}_{t|j} \frac{\partial^2 \mathbf{A}_t^T}{\partial \hat{\mathbf{x}}_{j|j-1}^2} + \frac{\partial^2 \mathbf{P}_{t|j}}{\partial \hat{\mathbf{x}}_{j|j-1}^2} \mathbf{A}_t^T \\ &+ 2 \left(\frac{\partial \mathbf{A}_t}{\partial \hat{\mathbf{x}}_{j|j-1}} \frac{\partial \mathbf{P}_{t|j}}{\partial \hat{\mathbf{x}}_{j|j-1}} + \frac{\partial \mathbf{P}_{t|j}}{\partial \hat{\mathbf{x}}_{j|j-1}} \frac{\partial \mathbf{A}_t^T}{\partial \hat{\mathbf{x}}_{j|j-1}} \right) \\ &+ \frac{\partial^2 \mathbf{G}_t}{\partial \hat{\mathbf{x}}_{j|j-1}^2} \mathbf{G}_t^T + 2 \frac{\partial \mathbf{G}_t}{\partial \hat{\mathbf{x}}_{j|j-1}} \frac{\partial \mathbf{G}_t^T}{\partial \hat{\mathbf{x}}_{j|j-1}} + \mathbf{G}_t \frac{\partial^2 \mathbf{G}_t^T}{\partial \hat{\mathbf{x}}_{j|j-1}^2}, \quad t \in [t_j, t_{j+1}] \\ \frac{\partial^2 \mathbf{P}_{t|j}}{\partial \hat{\mathbf{x}}_{j|j-1}^2} (t_j) &= \frac{\partial^2 \mathbf{P}_{j|j}}{\partial \hat{\mathbf{x}}_{j|j-1}^2} \end{aligned} \quad (48)$$

where

$$\begin{aligned} \frac{\partial^2 \mathbf{P}_{j|j}}{\partial \hat{\mathbf{x}}_{j|j-1}^2} &= \frac{\partial^2 \mathbf{P}_{j|j-1}}{\partial \hat{\mathbf{x}}_{j|j-1}^2} \\ &- \left(\frac{\partial^2 \mathbf{K}_j}{\partial \hat{\mathbf{x}}_{j|j-1}^2} \mathbf{R}_j \mathbf{K}_j^T + \mathbf{K}_j \frac{\partial^2 \mathbf{R}_j}{\partial \hat{\mathbf{x}}_{j|j-1}^2} \mathbf{K}_j^T + \mathbf{K}_j \mathbf{R}_j \frac{\partial^2 \mathbf{K}_j^T}{\partial \hat{\mathbf{x}}_{j|j-1}^2} \right. \\ &+ 2 \left(\frac{\partial \mathbf{K}_j}{\partial \hat{\mathbf{x}}_{j|j-1}} \frac{\partial \mathbf{R}_j}{\partial \hat{\mathbf{x}}_{j|j-1}} \mathbf{K}_j^T + \frac{\partial \mathbf{K}_j}{\partial \hat{\mathbf{x}}_{j|j-1}} \mathbf{R}_j \frac{\partial \mathbf{K}_j^T}{\partial \hat{\mathbf{x}}_{j|j-1}} \right. \\ &\left. \left. + \mathbf{K}_j \frac{\partial \mathbf{R}_j}{\partial \hat{\mathbf{x}}_{j|j-1}} \frac{\partial \mathbf{K}_j^T}{\partial \hat{\mathbf{x}}_{j|j-1}} \right) \right) \end{aligned} \quad (49)$$

Second Derivative of the Kalman Gain. In both cases above, the second partial derivative of the Kalman gain \mathbf{K}_j is required. This can be done as follows:

$$\begin{aligned} \frac{\partial^2 \mathbf{K}_j}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n} &= \frac{\partial^2 \mathbf{P}_{j|j-1}}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n} \mathbf{C}_j^T \mathbf{R}_j^{-1} + \frac{\partial \mathbf{P}_{j|j-1}}{\partial \boldsymbol{\eta}_k} \frac{\partial \mathbf{C}_j^T}{\partial \boldsymbol{\theta}_n} \mathbf{R}_j^{-1} + \frac{\partial \mathbf{P}_{j|j-1}}{\partial \boldsymbol{\theta}_n} \mathbf{C}_j^T \frac{\partial \mathbf{R}_j^{-1}}{\partial \boldsymbol{\eta}_k} \\ &+ \frac{\partial \mathbf{P}_{j|j-1}}{\partial \boldsymbol{\theta}_n} \frac{\partial \mathbf{C}_j^T}{\partial \boldsymbol{\eta}_k} \mathbf{R}_j^{-1} + \mathbf{P}_{j|j-1} \frac{\partial^2 \mathbf{C}_j^T}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n} \mathbf{R}_j^{-1} + \mathbf{P}_{j|j-1} \frac{\partial \mathbf{C}_j^T}{\partial \boldsymbol{\eta}_k} \frac{\partial \mathbf{R}_j^{-1}}{\partial \boldsymbol{\theta}_n} \\ &- \frac{\partial \mathbf{K}_j}{\partial \boldsymbol{\theta}_n} \frac{\partial \mathbf{R}_j}{\partial \boldsymbol{\eta}_k} \mathbf{R}_j^{-1} - \mathbf{K}_j \frac{\partial^2 \mathbf{R}_j}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n} \mathbf{R}_j^{-1} - \mathbf{K}_j \frac{\partial \mathbf{R}_j}{\partial \boldsymbol{\eta}_k} \frac{\partial \mathbf{R}_j^{-1}}{\partial \boldsymbol{\theta}_n} \\ &= \left(\frac{\partial^2 \mathbf{P}_{j|j-1}}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n} \mathbf{C}_j^T + \mathbf{P}_{j|j-1} \frac{\partial^2 \mathbf{C}_j^T}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n} \mathbf{R}_j^{-1} - \mathbf{K}_j \frac{\partial^2 \mathbf{R}_j}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\theta}_n} \mathbf{R}_j^{-1} \right. \\ &\left. + \frac{\partial \mathbf{P}_{j|j-1}}{\partial \boldsymbol{\eta}_k} \frac{\partial \mathbf{C}_j^T}{\partial \boldsymbol{\theta}_n} + \frac{\partial \mathbf{P}_{j|j-1}}{\partial \boldsymbol{\theta}_n} \frac{\partial \mathbf{C}_j^T}{\partial \boldsymbol{\eta}_k} - \frac{\partial \mathbf{K}_j}{\partial \boldsymbol{\theta}_n} \frac{\partial \mathbf{R}_j}{\partial \boldsymbol{\eta}_k} \right) \mathbf{R}_j^{-1} \\ &- \left(\frac{\partial \mathbf{P}_{j|j-1}}{\partial \boldsymbol{\eta}_k} \mathbf{C}_j^T + \mathbf{P}_{j|j-1} \frac{\partial \mathbf{C}_j^T}{\partial \boldsymbol{\eta}_k} - \mathbf{K}_j \frac{\partial \mathbf{R}_j}{\partial \boldsymbol{\eta}_k} \right) \mathbf{R}_j^{-1} \frac{\partial \mathbf{R}_j}{\partial \boldsymbol{\theta}_n} \mathbf{R}_j^{-1} \end{aligned} \quad (50)$$

The derivatives $\frac{\partial^2 \mathbf{K}_j}{\partial \boldsymbol{\eta}_k \partial \hat{\mathbf{x}}_{j|j-1}}$, $\frac{\partial^2 \mathbf{K}_j}{\partial \hat{\mathbf{x}}_{j|j-1} \partial \boldsymbol{\theta}_n}$, and $\frac{\partial^2 \mathbf{K}_j}{\partial \hat{\mathbf{x}}_{j|j-1}^2}$ are calculated in the same way. In the special case of $\frac{\partial^2 \mathbf{K}_j}{\partial \hat{\mathbf{x}}_{j|j-1}^2}$, the derivative can be simplified to

$$\begin{aligned} \frac{\partial^2 \mathbf{K}_j}{\partial \hat{\mathbf{x}}_{j|j-1}^2} &= \left(\frac{\partial^2 \mathbf{P}_{j|j-1}}{\hat{\mathbf{x}}_{j|j-1}^2} \mathbf{C}_j^T + \mathbf{P}_{j|j-1} \frac{\partial^2 \mathbf{C}_j^T}{\hat{\mathbf{x}}_{j|j-1}^2} \mathbf{R}_j^{-1} - \mathbf{K}_j \frac{\partial^2 \mathbf{R}_j}{\hat{\mathbf{x}}_{j|j-1}^2} \mathbf{R}_j^{-1} \right. \\ &\left. + \frac{\partial \mathbf{P}_{j|j-1}}{\partial \hat{\mathbf{x}}_{j|j-1}} \frac{\partial \mathbf{C}_j^T}{\partial \hat{\mathbf{x}}_{j|j-1}} - \frac{\partial \mathbf{K}_j}{\partial \hat{\mathbf{x}}_{j|j-1}} \frac{\partial \mathbf{R}_j}{\partial \hat{\mathbf{x}}_{j|j-1}} \right) \mathbf{R}_j^{-1} \\ &- \left(\frac{\partial \mathbf{P}_{j|j-1}}{\partial \hat{\mathbf{x}}_{j|j-1}} \mathbf{C}_j^T + \mathbf{P}_{j|j-1} \frac{\partial \mathbf{C}_j^T}{\partial \hat{\mathbf{x}}_{j|j-1}} - \mathbf{K}_j \frac{\partial \mathbf{R}_j}{\partial \hat{\mathbf{x}}_{j|j-1}} \right) \mathbf{R}_j^{-1} \frac{\partial \mathbf{R}_j}{\partial \hat{\mathbf{x}}_{j|j-1}} \mathbf{R}_j^{-1} \end{aligned} \quad (51)$$

MODEL DEFINITIONS

In all following models, A_1 , A_2 , and A_3 represent the absorption, central, and peripheral compartments of the PK part of the model and R represents the PD response. The measured concentration in the central compartment is denoted by

$$C(t) = \frac{A_2(t)}{V_2}. \quad (52)$$

The two-compartment PK model is described with the equations

$$\begin{pmatrix} dA_1 \\ dA_2 \\ dA_3 \end{pmatrix} = \begin{pmatrix} -k_a A_1 & & & + \text{Input}(t) \\ k_a A_1 & -\frac{CL}{V_2} A_2 & -\frac{Q}{V_2} A_2 + \frac{Q}{V_3} A_3 & \\ & \frac{Q}{V_2} A_2 & -\frac{Q}{V_3} A_3 & \end{pmatrix} dt + \begin{pmatrix} -g_1 A_1 \\ g_1 A_1 \\ 0 \end{pmatrix} dW, \quad (53)$$

In model M1, the initial values are

$$(A_1(0), A_2(0), A_3(0)) = (\text{Dose}, 0, 0), \quad (54)$$

$\text{Input}(t) = 0$, and the output is considered to be

$$\mathbf{y}_t = C(t) + \mathbf{e}_t, \quad \mathbf{e}_t \sim N(0, \boldsymbol{\Sigma}_t) \quad (55)$$

with

$$\Sigma_t = \begin{pmatrix} \sigma_1^2 + \sigma_2^2 C(t)^2 & \\ & \sigma_3^2 \end{pmatrix}. \quad (56)$$

The parameters k_a and CL are considered to be log-normally distributed. The covariance matrix for the random effects is

$$\Omega = \begin{pmatrix} \omega_{11}^2 & 0 \\ 0 & \omega_{22}^2 \end{pmatrix} \quad (57)$$

For model M2, the dynamical equations displayed in Eq. (53) are kept the same, but the initial values are

$$(A_1(0), A_2(0), A_3(0)) = (0, 0, 0), \quad (58)$$

Input(t) represents the dose given at time t , and the output is

$$y_t = \begin{pmatrix} C(t) \\ R(t) \end{pmatrix} + e_t, \quad e_t \sim N(0, \Sigma_t) \quad (59)$$

where

$$R(t) = 100 \left(1 - \frac{I_{\max} C(t)}{IC_{50} + C(t)} \right) \quad (60)$$

and

$$\Sigma_t = \begin{pmatrix} \sigma_1^2 + \sigma_2^2 C(t)^2 & 0 \\ 0 & \sigma_3^2 \end{pmatrix}. \quad (61)$$

The parameters k_a , CL, and IC_{50} are considered to be log-normally distributed. The covariance matrix for the random effects is

$$\Omega = \begin{pmatrix} \omega_{11}^2 & 0 & 0 \\ 0 & \omega_{22}^2 & 0 \\ 0 & 0 & \omega_{33}^2 \end{pmatrix} \quad (62)$$

For model M3, the dynamical system of equation is

$$\begin{pmatrix} dA_1 \\ dA_2 \\ dA_3 \\ dR \end{pmatrix} = \begin{pmatrix} -k_a A_1 & & & & + \text{Input}(t) \\ k_a A_1 & -\frac{CL}{V_2} A_2 & -\frac{Q}{V_2} A_2 + \frac{Q}{V_3} A_3 & & \\ & \frac{Q}{V_2} A_2 - \frac{Q}{V_3} A_3 & & & \\ & & k_{in} \left(1 - I_{\max} \frac{A_2}{V_2} / \left(IC_{50} + \frac{A_2}{V_2} \right) \right) - k_{out} R & & \end{pmatrix} dt + \begin{pmatrix} -g_1 A_1 \\ g_1 A_1 \\ 0 \\ 0 \end{pmatrix} dW_t, \quad (63)$$

the initial values are

$$(A_1(0), A_2(0), A_3(0), R(0)) = \left(0, 0, 0, \frac{k_{in}}{k_{out}} \right), \quad (64)$$

Input(t) represents the dose given at time t and the output is

$$y_t = \begin{pmatrix} C(t) \\ R(t) \end{pmatrix} + e_t, \quad e_t \sim N(0, \Sigma_t) \quad (65)$$

with Σ_t as given by Eq. (61). The parameters k_a , CL, and IC_{50} are considered to be log-normally distributed. The covariance matrix for the random effects is as given by Eq. (62).

The sampling scheme was the same for all three models, relative to dosing. The PK part was sampled at hours 0.25, 0.5, 0.75, 1, 1.5, 2, 4, 6, 8, 12, 16, and 24 and the PD part was sampled at hours -24, -18, -12, 0, 1, 2, 4, 12, 24, 48, 72, 96, and 124. The doses of the hypothetical drug were 5, 10, 50, 100, and 200 dose units.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

REFERENCES

- Andersson R, Kroon T, Almquist J, Jirstrand M, Oakes ND, Evans ND, *et al.* Modeling of free fatty acid dynamics: insulin and nicotinic acid resistance under acute and chronic treatments. *J Pharmacokinet Pharmacodyn.* 2017;44(3):203–22.
- Jiang D, *et al.* Effects of CYP2C19 and CYP2C9 genotypes on pharmacokinetic variability of valproic acid in Chinese epileptic patients: nonlinear mixed-effect modeling. *Eur J Clin Pharmacol.* 2009;65(12):1187.
- Silber HE, Jauslin PM, Frey N, Gieschke R, Simonsson US, Karlsson MO. An integrated model for glucose and insulin regulation in healthy volunteers and type 2 diabetic patients following intravenous glucose provocations. *J Clin Pharmacol.* 2007;47(9):1159–71.
- Mould D, Upton RN. Basic concepts in population modeling, simulation, and model-based drug development—part 2: introduction to pharmacokinetic modeling methods. *CPT: Pharmacometrics Systems & Pharmacology.* 2013;2(4):1–14.
- Donnet S, Samson A. A review on estimation of stochastic differential equations for pharmacokinetic/pharmacodynamic models. *Adv Drug Deliv Rev.* 2013;65(7):929–39.
- Deng C, Plan EL, Karlsson MO. Approaches for modeling within subject variability in pharmacometric count data analysis: dynamic inter-occasion variability and stochastic differential equations. *J Pharmacokinet Pharmacodyn.* 2016;43(3):305–14.
- Trägårdh M, Chappell MJ, Ahnmark A, Lindén D, Evans ND, Gennemark P. Input estimation for drug discovery using optimal control and Markov chain Monte Carlo approaches. *J Pharmacokinet Pharmacodyn.* 2016;43(2):207–21.
- Sørensen H. Parametric inference for diffusion processes observed at discrete points in time: a survey. *Int Stat Rev.* 2004;72(3):337–54.
- Donnet S, Foulley JL, Samson A. Bayesian analysis of growth curves using mixed models defined by stochastic differential equations. *Biometrics.* 2010;66(3):733–41.
- Yan FR *et al.* Parameter estimation of population pharmacokinetic models with stochastic differential equations: implementation of an estimation algorithm. *Journal of Probability and Statistics.* 2014; 2014.

11. Donnet S, Samson A. Parametric inference for mixed models defined by stochastic differential equations. *ESAIM-Probab Stat.* 2008;12:196–218.
12. Donnet S and Samson A. EM algorithm coupled with particle filter for maximum likelihood parameter estimation of stochastic differential mixed-effects models. 2011.
13. Delattre M, Lavielle M. Coupling the SAEM algorithm and the extended Kalman filter for maximum likelihood estimation in mixed-effects diffusion models. *Stat Interface.* 2013;6(4):519–32.
14. Tornøe CW, Overgaard RV, Agersø H, Nielsen HA, Madsen H, Jonsson EN. Stochastic differential equations in NONMEM: implementation, application, and comparison with ordinary differential equations. *Pharm Res.* 2005;22(8):1247–58.
15. Overgaard RV, Jonsson EN, Tornøe CW, Madsen H. Non-linear mixed-effects models with stochastic differential equations: implementation of an estimation algorithm. *J Pharmacokinet Pharmacodyn.* 2005;32(1):85–107.
16. Overgaard RV, Holford N, Rytved KA, Madsen H. PKPD model of interleukin-21 effects on thermoregulation in monkeys—application and evaluation of stochastic differential equations. *Pharm Res.* 2007;24(2):298–309.
17. Berglund M, Sunnåker M, Adiels M, Jirstrand M, Wennberg B. Investigations of a compartmental model for leucine kinetics using non-linear mixed effects models with ordinary and stochastic differential equations. *Math Med Biol.* 2011;29(4):361–84.
18. Kregel A, Hauth J, Taskinen MR, Adiels M, Jirstrand M. A continuous-time adaptive particle filter for estimations under measurement time uncertainties with an application to a plasma-leucine mixed effects model. *BMC Syst Biol.* 2013;7(1):8.
19. Matzuka B, Chittenden J, Monteleone J, Tran H. Stochastic nonlinear mixed effects: a metformin case study. *J Pharmacokinet Pharmacodyn.* 2016;43(1):85–98.
20. Kristensen NR, Madsen H, Ingwersen SH. Using stochastic differential equations for PK/PD model development. *J Pharmacokinet Pharmacodyn.* 2005;32(1):109–41.
21. Tornøe CW, Jacobsen JL, Pedersen O, Hansen T, Madsen H. Grey-box modelling of pharmacokinetic/pharmacodynamic systems. *J Pharmacokinet Pharmacodyn.* 2004;31(5):401–17.
22. Beal SL, Sheiner LB, Boeckmann AJ, and Bauer RJ. NONMEM 7.4 users guides.
23. Jazwinski AH. Stochastic processes and filtering theory. Chelmsford: Courier Corporation; 2007.
24. Mortensen SB, Klim S, Dammann B, Kristensen NR, Madsen H, Overgaard RV. A matlab framework for estimation of NLME models using stochastic differential equations. *J Pharmacokinet Pharmacodyn.* 2007;34(5):623–42.
25. Klim S, Mortensen SB, Kristensen NR, Overgaard RV, Madsen H. Population stochastic modelling (PSM)—an R package for mixed-effects models based on stochastic differential equations. *Comput Methods Prog Biomed.* 2009;94(3):279–89.
26. Fletcher R. Practical methods of optimization. Hoboken: John Wiley & Sons; 2013.
27. Almquist J, Leander J, Jirstrand M. Using sensitivity equations for computing gradients of the FOCE and FOCEI approximations to the population likelihood. *J Pharmacokinet Pharmacodyn.* 2015;42:191–209.
28. Leander J, Lundh T, Jirstrand M. Stochastic differential equations as a tool to regularize the parameter estimation problem for continuous time dynamical systems given discrete time measurements. *Math Biosci.* 2014;251:54–62.
29. Leander J, Almquist J, Ahlström C, Gabrielsson J, Jirstrand M. Mixed effects modeling using stochastic differential equations: illustrated by pharmacokinetic data of nicotin acid in obese Zucker rats. *AAPS J.* 2015;17:586–96.
30. Wolfram Research, Inc., Mathematica, version 11.0. 2016.
31. Olafsdottir HK. Sensitivity-based gradients for parameter estimation in nonlinear mixed effects models with deterministic and stochastic dynamics, Master's thesis, Institutionen för signaler och system, Chalmers tekniska högskola, 2016.
32. Spinello D, Stilwell DJ. Nonlinear estimation with state-dependent Gaussian observation noise. *IEEE Trans Autom Control.* 2010;55(6):1358–66.
33. Anderson BD, Moore JB. Optimal filtering. Englewood Cliffs. 1979;21:22–95.
34. Nguyen T, *et al.* Model evaluation of continuous data pharmacometric models: metrics and graphics. *CPT: Pharmacometrics Systems & Pharmacology.* 2017;6(2):87–109.
35. Iacus SM. Simulation and inference for stochastic differential equations: with R examples. Berlin: Springer Science & Business Media; 2009.
36. Julier SJ, Uhlmann JK. New extension of the Kalman filter to nonlinear systems. In: Kadar I, editor. Signal processing, sensor fusion, and target recognition VI, vol. 3068. International Society for Optics and Photonics; 1997. pp. 182–194. <https://doi.org/10.1117/12.280797>.
37. Wan EA and Van Der Merwe R. The unscented Kalman filter for nonlinear estimation. In Adaptive systems for signal processing, communications, and control symposium 2000. AS-SPCC The IEEE. 2000; 2000:153–158.
38. Cetara, Phoenix 8.0..
39. Lixoft, Monolix.
40. Fidler M, Xiong Y, Schoemaker R, Wilkins J, Trame M, and Wang W. *Nlmixr*: nonlinear mixed effects models in population pharmacokinetics and pharmacodynamics. 2017.