Parameter estimation in nonlinear mixed effect models based on ordinary differential equations: An optimal control approach

Quentin Clairon, Chloé Pasin, Irene Balelli, Rodolphe Thiébaut, Mélanie Prague

Parameter estimation in nonlinear mixed effect models based on ordinary differential equations: an optimal control approach

Quentin Clairon[1,2*], Chloé Pasin[3], Irene Balelli[4], Rodolphe Thiébaut[1,2] and Mélanie Prague[1,2]

[1*] University of Bordeaux, Inria Bordeaux Sud-Ouest Inserm, Bordeaux Population Health Research Center, SISTM Team, UMR1219, F-33000 Bordeaux, France.
[2*] Vaccine Research Institute, F-94000 Créteil, France.
[3]Institute of Medical Virology, University of Zurich Department of Infectious Diseases and Hospital Epidemiology University Hospital, Collegium Helveticum, Zurich, Switzerland.
[4] Centre Inria d'Université Côte d'Azur, EPIONE Research Project, Valbonne, France.

*Corresponding author(s). E-mail(s): quentin.clairon@u-bordeaux.fr;

Abstract

We present a method for parameter estimation for nonlinear mixed-effects models based on ordinary differential equations (NLME-ODEs). It aims to regularize the estimation problem in the presence of model misspecification and practical identifiability issues, while avoiding the need to know or estimate initial conditions as nuisance parameters. To this end, we define our estimator as a minimizer of a cost function that incorporates a possible gap between the assumed population-level model and the specific individual dynamics. The computation of the cost function leads to formulate and solve optimal control problems at the subject level. Compared to the maximum likelihood method, we show through simulation examples that our method improves the estimation accuracy in possibly partially observed systems with unknown initial conditions or poorly identifiable parameters with or without model

error. We conclude this work with a real-world application in which we model the antibody concentration after Ebola virus vaccination.

Keywords: Dynamic population models, Ordinary differential equations, Optimal control theory, Mechanistic models, Nonlinear mixed effects models, Clinical trial analysis

1 Introduction

Ordinary differential equation (ODE) models are standard in population dynamics, epidemiology, virology, pharmacokinetics, or genetic regulatory network analysis since they can describe the main mechanisms of interaction between different biological components of complex systems and their evolution over time and because they also provide reasonable approximations to stochastic dynamics (Perelson et al (1996); Engl et al (2009); Villain et al (2019)).

For experimental designs with a large number of subjects and a limited number of individual measurements, nonlinear mixed-effects models may be more relevant than single-subject models, since they allow to collect informations from the entire population while accounting for variability among individuals. For instance, clinical trials and pharmacokinetics/pharmacodynamics studies often fall into this category (M. Lavielle and Mentre (2011); Guedj et al (2007)). Formally, we are interested in a population where the dynamic of each subject $i \in [\![1, n]\!]$ is modeled by the d-dimensional ODE:

$$\begin{cases} \dot{x}_i(t) = f_{\theta,b_i}(t, x_i(t)) \\ x_i(0) = x_{i,0} \end{cases} \tag{1}$$

where f is a $d-$dimensional vector field, θ is a $p-$dimensional parameter, $b_i \sim N(0, \Psi)$ is a $q-$dimensional random effect where Ψ is a variance-covariance matrix, $x_{i,0}$ is the initial condition for subject i. We denote $X_{\theta,b_i,x_{i,0}}$ the solution of (1) for a given set $(\theta, b_i, x_{i,0})$. In (1), we can also incorporate covariate functions z_i which are omitted here for the purpose of clarity.

Our goal is to estimate the true population parameters (θ^*, Ψ^*) as well as the true subject specific realizations $\{b_i^*\}_{i\in[\![1,n]\!]}$ from partial and noisy observations coming from n subjects and described by the following observational model:

$$y_{ij} = CX_{\theta^*,b_i^*,x_{i,0}^*}(t_{ij}) + \epsilon_{ij}.$$

For the $i-$th subject, we denote t_{ij} its j-th measurement time-point on the observation interval $[0, T]$ and n_i its total number of available measurements. Here C is a $d^o \times d$ sized observation matrix emphasizing the potentially partially observed nature of the process and $\epsilon_{ij} \sim \sigma^* \times N(0, I_{d^o})$ is the measurement error. The vector $\mathbf{y_i} = \{y_{ij}\}_{j\in[\![1,n_i]\!]}$ corresponds to the set of observations available for the i-th subject and $\mathbf{y} = \{\mathbf{y_i}\}_{i\in[\![1,n]\!]}$ is the set of all observations in the

population. We also assume that only a subset $x_{i,0}^{k*}$ of $x_{i,0}^{*}$ is perfectly known, the other ones, denoted $x_{i,0}^{u*}$, being unknown and they are ordered as follows $x_{i,0} = \left(\left(x_{i,0}^{u}\right)^T, \left(x_{i,0}^{k}\right)^T\right)^T$ for the sake of clarity. Nonetheless, pre-existing information can be available for $x_{i,0}^{u*}$ under the form of a priori distribution with a possibly parameter dependent density $\mathbb{P}\left(x_{i,0}^{u} \mid \theta, \Psi, b_i\right)$. The same holds for (θ, Ψ) for which a priori distribution $\mathbb{P}(\theta, \Psi)$ can be available.

Our problem belongs to the class of parameter estimation problem in nonlinear mixed effect models based on ODEs (NLME-ODEs). In this context, frequentist methods based on likelihood maximization (via different numerical procedures: Laplace approximation Pinheiro and Bates (1994), Gaussian quadrature Guedj et al (2007) or SAEM Comets et al (2017); Lavielle and Mentré (2007)) and Bayesian ones aiming to reconstruct the a posteriori distribution or to derive the maximum a posteriori estimator (via MCMC algorithms Lunn et al (2000); Huang and Dagne (2011), importance sampling Raftery and Bao (2010), approximation of the asymptotic posterior distribution Prague et al (2013)) have been proposed. In particular, dedicated methods/softwares using the structure of ODE models have been implemented to increase numerical stability and speed up convergence rate (Tornoe et al (2004)), to reduce the computational time (Donnet and Samson (2006)) or to avoid the repeated model integration and estimation of initial conditions (Wang et al (2014)). However, all the preceding methods face similar pitfalls due to specific features of population models based on ODEs (with the exception of Wang et al (2014)):

1. They do not take into account the presence of model misspecification, a common feature of ODE models used in biology. Indeed, the ODE modeling process suffers from model inadequacy, understood as the discrepancy between the mean model response and the real process, as well as residual variability subject to specific stochastic perturbations or missing elements that disappear by averaging over the entire population (Kennedy and O'Hagan (2001)). As examples of the causes of model inadequacies, one can think of the ODE models used in epidemiology and virology, which are derived by approximations in which, for example, interactions are modeled by pairwise products, while higher order terms and/or the influence of unknown/unmeasured external factors are neglected. As for residual variability, recall that biological processes are often stochastic and the justification for deterministic modeling lies in the approximation of stochastic processes (Kurtz (1978); Kampen (1992)). Moreover, in the context of NLME-ODEs, new sources of model uncertainties emerge. Firstly, error measurement in covariates z_i can lead to use a proxy function $\widehat{z_i}$ instead of z_i (Huang and Dagne (2011)). Secondly, the sequential nature of most inference methods leads to estimate $\{b_i^*\}_{i \in [\![1,n]\!]}$ based on an approximation $\widehat{\theta}$ instead of θ^*. Thus, the structure of mixed-effect models spreads measurement uncertainty into the mechanistic model structure during the estimation. It turns classical statistical uncertainties into model error causes. Estimation of θ^*, Ψ^* and

$\{b_i^*\}_{i\in[\![1,n]\!]}$ must be performed in the presence of the model error, although it is known to dramatically affect the accuracy of methods that do not take it into account (Brynjarsdottir and O'Hagan (2014)).

2. They have to estimate or make assumptions on the true unknown initial conditions $x_{i,0}^{u*}$. In ODE models, the initial conditions are generally nuisance parameters in the sense that inferring their values does not bring answers to the scientific questions which motivate the model construction but is necessary for the estimation of the relevant parameters. For example, partially observed compartmental models used in pharmacokinetics/pharmacodynamics often involve unknown initial conditions which needs to be inferred to estimate the transmission rates between compartments, which are the true parameters of interest. Unknown initial conditions imply either assumptions on $x_{i,0}^{u*}$ values (M. Lavielle and Mentre (2011)), another potential cause of model misspecifications, or their estimation (Huang and Lu (2008)). This latter case requires a priori knowledge to derive $\mathbb{P}\left(x_{i,0}^{u} \mid \theta, \Psi, b_i\right)$ expression and simultaneous inference of $(b_i^*, x_{i,0}^{u*})$ as subject specific parameters. This increases the complexity of the related optimization problem and can degrade estimation accuracy.
3. They can face accuracy degradation when the inverse problem of parameter estimation is ill-posed (Engl et al (2009)) due to practical identifiability issues. Ill-posedness in ODE models is often due to the geometry induced by the mapping $(\theta, b_i, x_{i,0}) \longmapsto CX_{\theta,b_i,x_{i,0}}$, where there can be a small number of relevant directions of variation skewed from the original parameter axes (Gutenkunst et al (2007)). This problem, called sloppiness, often appears in ODE models used in biology (Gutenkunst et al (2007); Leary et al (2015)) and leads to an ill-conditioned Fisher Information Matrix. For maximum likelihood estimators this is a cause of high variance due to the Cramér-Rao bound. For Bayesian inference, it leads to a nearly singular asymptotic posterior distribution because of Bernstein–von Mises theorem (see Campbell (2007) for the computational induced problems). Although this problem is in part mitigated in NLME-ODEs which merge different subjects for estimating (θ^*, Ψ^*) and use distribution of $b_i \mid \Psi$ as prior at the subject level (Lavielle and Aarons (2015)), estimation accuracy can benefit from the use of regularization techniques.

These specific features of ODE-based population models limit the amount of information classic approaches can extract for estimation purposes from observations no matter their qualities or abundances. This advocates for the development of new estimation procedures. Approximate methods (Varah (1982); Ramsay et al (2007)) have already proven to be useful for ODE models to face such issues. They rely on an approximation of the solution of the original ODE (1) which is expected to have a smoother dependence with respect to the parameters and to relax the constraint imposed by the model during the estimation process. The interest of such approximations is twofold. Firstly, they produce estimators with a better conditioned variance matrix comparing

to classic likelihood based approaches. Secondly, they reduce the effect of model error on estimator accuracy. Also, some of these approximations bypass the need to estimate initial conditions (Ramsay et al (2007); Clairon (2020)). Still, these methods are currently limited to cases where observations are coming from one subject.

In this work, we develop a new estimation method adapted to NLME-ODEs integrating such approximations to mitigate the effect of model misspecification and poorly identifiable parameters on estimation accuracy, while avoiding the need to estimate $x_{i,0}^{u*}$ as additional subject specific parameters. We propose here a hierarchical profiling approach taking the form of a nested estimation procedure. The unknowns initial conditions $\left\{x_{i,0}^{u*}\right\}_{i\in[\![1,n]\!]}$ are seen as nuisance parameters for $\{b_i^*\}_{i\in[\![1,n]\!]}$ estimation, which are in turn considered as nuisance parameters for population parameter $(\theta^*, \Psi^*, \sigma^*)$ estimation. This lead to the construction of outer, middle and inner criteria for the estimation of $(\theta^*, \Psi^*, \sigma^*)$, $\{b_i^*\}_{i\in[\![1,n]\!]}$ and $\left\{x_{i,0}^{u*}\right\}_{i\in[\![1,n]\!]}$ respectively. The inner criteria is designed to incorporate $\mathbb{P}\left(x_{i,0}^{u} \mid \theta, \Psi, b_i\right)$ if an expression is proposed for it but can also be defined without if no prior information exists for $\left\{x_{i,0}^{u*}\right\}_{i\in[\![1,n]\!]}$. Also, this criterion accounts for model error presence by assuming that the actual dynamic of each subject is better described by a perturbed version of the ODE (1). This added perturbation aims to capture different sources of errors at the subject level (Brynjarsdottir and O'Hagan (2014); Tuo and Wu (2015)). We control the magnitude of the acceptable perturbations by defining the inner criteria through a cost function balancing the two contrary objectives of fidelity to the observations and to the original model: to this end, we introduce a model discrepancy penalization term. The practical computation of the chosen perturbations requires to solve optimal control problems (Clarke (2013)) known as tracking problems. This is done using a method inspired by Cimen and Banks (2004) which has the advantage to automatically provide an estimator for $x_{i,0}^{u*}$ with no additional computational costs. This is the key element to efficiently profile on initial conditions during b_i^* estimation.

In section 2, we present the inner, middle and outer criteria used to define our estimator. In section 3, we compare our approach with classic maximum likelihood in simulations. Then, we proceed to the real data analysis coming from clinical studies and a model of the antibody concentration dynamics following immunization with an Ebola vaccine in East African participants (Pasin et al (2019)). Section 5 concludes and discuss future extensions of the method.

2 Estimator construction: definition of the inner/middle/outer criteria

From now on, we use the Choleski decomposition $\sigma^2\Psi^{-1} = \triangle^T\triangle$ and the parametrization $\phi := (\theta, \Delta, \sigma)$ instead of (θ, Ψ, σ) to enforce positiveness

and symmetry of Ψ and denote in a summarized way the set of all population parameters. The norm $\|.\|_2$ denotes the classic Euclidean one defined by $\|b\|_2 = \sqrt{b^T b}$. Similarly as in the Expectation-Maximization (EM) algorithm, we estimate the population and individual parameters via a nested procedure:

- Estimator $\widehat{\phi}$ obtained by minimization of an **outer criterion** F based on an approximation of $\min_b \min_{x_0^u} (-\ln \mathbb{P}[\phi, b, x_0^u \mid \mathbf{y}])$, the log joint-distribution of (ϕ, b, x_0^u) sequentially profiled on $b := \{b_i\}_{i\in[\![1,n]\!]}$ and $x_0^u := \left\{x_{0,i}^u\right\}_{i\in[\![1,n]\!]}$, which are respectively the set of all random effects and unknown initial conditions among all subjects.
- Estimator $\widehat{b_i} := \widehat{b_i}(\phi)$ obtained for each subject i by minimization of a **middle criterion** G_i based on an approximation of $\min_{x_{0,i}^u} \left(-\ln \mathbb{P}(\mathbf{y_i}, b_i, x_{0,i}^u \mid \phi)\right)$, the log joint-distribution of the data, the random effects and unknown initial conditions profiled on the latter.
- Estimator $\widehat{x_{0,i}^u} := \widehat{x_{0,i}^u}(\phi, b_i)$ obtained for each subject i by minimization of an **inner criterion** H_i based on an approximation of $-\ln \mathbb{P}(\mathbf{y_i}, x_{0,i}^u \mid \phi, b_i)$, the log joint-distribution of the data and unknown initial conditions.

Our estimation procedure can be expressed in a pseudo-algorithmic way.

1/ Outer criteria minimization:

$$\begin{aligned}\widehat{\phi} &= \arg\min_\phi F(\phi) \\ &:= \arg\min_\phi -2\ln \widetilde{\mathbb{P}}(\phi, \widehat{b}, \widehat{x_0^u} \mid \mathbf{y}),\end{aligned}$$

for a given subject i and ϕ value:

2/ middle criteria minimization:

$$\begin{aligned}\widehat{b_i}(\phi) &= \arg\min_{b_i} G_i(b_i \mid \phi) \\ &:= \arg\min_{b_i} -2\ln \widetilde{\mathbb{P}}(\mathbf{y_i}, b_i, \widehat{x_{0,i}^u} \mid \phi),\end{aligned}$$

for a given b_i value:

3/ inner criteria minimization:

$$\begin{aligned}\widehat{x_{0,i}^u}(\phi, b_i) &= \arg\min_{x_{0,i}^u} H_i(x_{0,i} \mid \phi, b_i) \\ &:= \arg\min_{x_{0,i}^u} -2\ln \widetilde{\mathbb{P}}(\mathbf{y_i}, x_{0,i}^u \mid \phi, b_i).\end{aligned}$$

In the following sections, we derive the expressions of F, G_i and H_i starting with H_i since each criterion construction rely on lower level ones. Finally, despite that the following formal presentation of criteria are made for any $\mathbb{P}\left(x_{i,0}^u \mid \phi, b_i\right)$ expressions, we have to restrict ourselves to uniform, normal and log-normal densities in practice to use our numerical procedures.

2.1 Inner criterion

In this section, we construct the criteria H_i used to estimate $x_{0,i}^{u*}$ for a given (ϕ, b_i) value. A classic procedure would lead to jointly estimate $\left(b_i^*, x_{0,i}^{u*}\right)$ by maximization of the log joint-likelihood function of the data and $(b_i, x_{0,i}^u)$. However for each subject, we want to:

1. profile on $x_{0,i}^{u*}$ during random effects estimation to limit b_i^* estimation degradation due to presence of nuisance parameters,
2. use prior knowledge given by $\mathbb{P}\left(x_{i,0}^u \mid \phi, b_i\right)$ if available,
3. allow an acceptable deviation from the assumed model at the population level to take into account possible model misspecifications.

To solve the first and second point, we define our estimator:

1. as the maximizer of the joint conditional likelihood $\mathbb{P}(\mathbf{y_i}, x_{0,i}^u \mid \phi, b_i)$ if $\mathbb{P}\left(x_{i,0}^u \mid \phi, b_i\right)$ is available,
2. otherwise as the maximizer of $\mathbb{P}(\mathbf{y_i} \mid \phi, b_i, x_{0,i}^u)$.

Since $\mathbb{P}(\mathbf{y_i}, x_{0,i}^u \mid \phi, b_i) = \mathbb{P}(\mathbf{y_i} \mid \phi, b_i, x_{0,i}^u)\mathbb{P}\left(x_{i,0}^u \mid \phi, b_i\right)$, we have $\arg\max_{x_{0,i}^u} \mathbb{P}(\mathbf{y_i} \mid \phi, b_i, x_{0,i}^u) = \arg\max_{x_{0,i}^u} \mathbb{P}(\mathbf{y_i}, x_{0,i}^u \mid \phi, b_i)$ if $\mathbb{P}\left(x_{i,0}^u \mid \phi, b_i\right)$ is constant. So, the estimation criteria in absence of prior information is equivalent to choosing a uniform prior over $x_{0,i}^u$ space and constitute only a particular case. We will thus focus on $\mathbb{P}(\mathbf{y_i}, x_{0,i}^u \mid \phi, b_i)$ from now one. We have:

$$\begin{aligned}\mathbb{P}(\mathbf{y_i}, x_{0,i}^u \mid \phi, b_i) &= \textstyle\prod_j \mathbb{P}(y_{ij} \mid \phi, b_i, x_{0,i}^u)\mathbb{P}\left(x_{i,0}^u \mid \phi, b_i\right) \\ &= \textstyle\prod_j (2\pi)^{-d^o/2}\, \sigma^{-d^o} e^{-0.5\left\|CX_{\theta,b_i,x_{0,i}}(t_{ij}) - y_{ij}\right\|_2^2/\sigma^2} \mathbb{P}\left(x_{i,0}^u \mid \phi, b_i\right),\end{aligned}$$

from which we derive the joint likelihood estimator:

$$\begin{aligned}\widehat{x_{0,i}^u}(\phi, b_i) &= \arg\min_{x_{0,i}^u} -2\ln \mathbb{P}(\mathbf{y_i}, x_{0,i}^u \mid \phi, b_i, x_{0,i}^u) \\ &= \arg\min_{x_{0,i}^u} \left\{ \tfrac{1}{\sigma^2} \textstyle\sum_j \left\|CX_{\theta,b_i,x_{0,i}}(t_{ij}) - y_{ij}\right\|_2^2 - 2\ln \mathbb{P}\left(x_{i,0}^u \mid \phi, b_i\right) \right\}.\end{aligned}$$

We also want to allow the presence of perturbations at the subject scale comparing to the original model defined at the population level. For this, we assume the regression function is no longer $X_{\theta,b_i,x_{0,i}}$, but rather $X_{\theta,b_i,x_{i,0},u_i}$, the solution of:

$$\begin{cases} \dot{x}_i(t) = f_{\theta,b_i}(t, x_i(t)) + Bu_i(t) \\ x_i(0) = x_{i,0}. \end{cases} \tag{2}$$

This perturbed ODE has been obtained by the addition of the forcing term $t \mapsto Bu_i(t)$ to ODE (1) with B a $d \times d_u$ matrix and u_i a function in $L^2\left([0,T], \mathbb{R}^{d_u}\right)$ representing the perturbation. However, to ensure the possible perturbations remain small, we replace the data fitting criterion $\sum_j \left\|CX_{\theta,b_i,x_{0,i}}(t_{ij}) - y_{ij}\right\|_2^2$ by $\min_{u_i} \mathcal{C}_i(x_{i,0}^u, u_i \mid \theta, b_i, U)$, where

$$\mathcal{C}_i(x_{i,0}^u, u_i \mid \theta, b_i, U) = \sum_j \left\|CX_{\theta,b_i,x_{0,i},u_i}(t_{ij}) - y_{ij}\right\|_2^2 + \|u_i\|_{U,L^2}^2,$$

and $\|u_i\|^2_{U,L^2} = \int_0^T u_i(t)^T U u_i(t) dt$ is the weighted Euclidean norm. Here, the magnitude of the allowed perturbations is controlled by a positive definite and symmetric weighting matrix U. Finally, we obtain:

$$\widehat{x^u_{0,i}}(\phi, b_i) := \arg\min_{x^u_{0,i}} H_i(x^u_{0,i} \mid \phi, b_i) \tag{3}$$

where

$$H_i(x^u_{0,i} \mid \phi, b_i) = \min_{x^u_{0,i}} \left\{ \frac{1}{\sigma^2} \min_{u_i} \mathcal{C}_i(x^u_{i,0}, u_i \mid \theta, b_i, U) - 2 \ln \mathbb{P}\left(x^u_{i,0} \mid \phi, b_i\right) \right\}.$$

Computing $H_i(x^u_{0,i} \mid \phi, b_i)$ requires to solve the infinite dimensional optimization problem $\min_{u_i} \mathcal{C}_i(x^u_{i,0}, u_i \mid \theta, b_i, U)$ in $L^2\left([0,T], \mathbb{R}^{d_u}\right)$. This problem belongs to the field of optimal control theory for which dedicated approaches have been developed (Sontag (1998); Aliyu (2011); Clarke (2013)). Here we use the same method as in Clairon (2020) which is detailed in Appendix A. All it requires from the user is to specify a pseudo-linear representation of ODE (1), i.e., a possibly state-dependent matrix $A_{\theta,b_i}(t, x_i(t))$ and state-independent vector $r_{\theta,b_i}(t)$ such that:

$$f_{\theta,b_i}(t, x_i(t)) = A_{\theta,b_i}(t, x_i(t))\, x_i(t) + r_{\theta,b_i}(t). \tag{4}$$

This formulation is crucial for solving the optimal control problem in a computationally efficient way. Linear models already fit in this formalism with $A_{\theta,b_i}(t) := A_{\theta,b_i}(t, x_i(t))$. For nonlinear models, the pseudo-linear representation is not unique but always exists (in order to exploit this non-uniqueness as an additional degree of freedom, see Cimen (2008) section 6). This method presents the advantage of formulating $\min_{u_i} \mathcal{C}_i(x^u_{i,0}, u_i \mid \theta, b_i, U)$ as a quadratic form (or a sequence of quadratic forms) with respect to $x^u_{0,i}$. Thus, if we choose a uniform, normal or log-normal law for $\mathbb{P}\left(x^u_{i,0} \mid \phi, b_i\right)$, $\arg\min_{x^u_{0,i}} H_i(x^u_{0,i} \mid \phi, b_i)$ has a closed form expression (approximated for log-normal), and obtaining $\widehat{x^u_{0,i}}(\phi, b_i)$ does not add any computational complexity comparing to $\min_{u_i} \mathcal{C}_i(x^u_{i,0}, u_i \mid \theta, b_i, U)$.

For a given $x^u_{0,i}$, the perturbation u_i corresponding to the solution of $\min_{u_i} \mathcal{C}_i(x^u_{i,0}, u_i \mid \theta, b_i, U)$ is named optimal control and denoted $\overline{u}_{i,\phi,b_i,x^u_{0,i}}$. In particular, we denote $\overline{u}_{i,\theta,b_i} := \overline{u}_{i,\theta,b_i,\widehat{x_{0,i}}}$ the optimal control corresponding to the initial condition estimator $\widehat{x_{0,i}} = \left(\widehat{x^u_{0,i}}(\phi, b_i)^T, \left(x^k_{i,0}\right)^T\right)^T$. The solution of (2) corresponding to the optimal control $u_i := \overline{u}_{i,\theta,b_i}$ is denoted $\overline{X}_{\theta,b_i}$ and named optimal trajectory: this will be considered as the regression function for the i-th subject. $\overline{X}_{\theta,b_i}$ is thus defined as solution of ODE (2) which needs the smallest perturbation in order to get close to the observations. In particular, $\overline{X}_{\theta,b_i}$ and $\overline{u}_{i,\theta,b_i}$ are respectively the subject specific state variable and

perturbation such that:

$$\begin{aligned} H_i(\widehat{x^u_{0,i}}(\phi, b_i) \mid \phi, b_i) \;=\; & \tfrac{1}{\sigma^2}\left\{\textstyle\sum_j \left\|C\overline{X}_{\theta,b_i}(t_{ij}) - y_{ij}\right\|_2^2 + \|\overline{u}_{i,\theta,b_i}\|^2_{U,L^2}\right\} \\ & - 2\ln \mathbb{P}(\widehat{x^u_{0,i}}(\phi, b_i) \mid \phi, b_i). \end{aligned} \tag{5}$$

Again, formal expressions can be derived for both $\overline{u}_{i,\theta,b_i}$ and $\widehat{x^u_{0,i}}(\phi, b_i)$, but they present no interest for the sake of explanation and are left in Appendix A.

2.2 Middle criterion

To construct an estimator $\widehat{b_i}$ of the random effects, we rely on an approximation of $\ln \mathbb{P}(\mathbf{y_i}, b_i, x^u_{0,i} \mid \phi)$ profiled on the unknown initial conditions. Since

$$\begin{aligned} \mathbb{P}(\mathbf{y_i}, b_i, x^u_{0,i} \mid \phi) \;&=\; \mathbb{P}(\mathbf{y_i} \mid \phi, b_i, x^u_{0,i})\mathbb{P}(b_i, x^u_{0,i} \mid \phi) \\ &=\; \mathbb{P}(\mathbf{y_i} \mid \phi, b_i, x^u_{0,i})\mathbb{P}(x^u_{0,i} \mid \phi, b_i)\mathbb{P}(b_i \mid \phi), \end{aligned}$$

with $\mathbb{P}(b_i \mid \phi)$ the density of $b_i \sim N(0, \sigma^2 \left(\triangle^T \triangle\right)^{-1})$, we can define as estimator:

$$\begin{aligned} \widehat{b_i}(\phi) \;&=\; \arg\min_{b_i} \min_{x^u_{0,i}} -2\ln \mathbb{P}(\mathbf{y_i}, b_i, x^u_{0,i} \mid \phi) \\ &=\; \arg\min_{b_i} \left\{ \min_{x^u_{0,i}} \left\{ \tfrac{1}{\sigma^2} \textstyle\sum_j \left\|CX_{\theta,b_i,x_{0,i}}(t_{ij}) - y_{ij}\right\|_2^2 - 2\ln \mathbb{P}\left(x^u_{i,0} \mid \phi, b_i\right) \right\} \right. \\ &+\; \left. \tfrac{\|\Delta b_i\|_2^2}{\sigma^2} \right\}. \end{aligned}$$

Still, we use the same relaxation & penalization scheme as in the previous section to account for model error presence for b_i^* estimation. We replace again the term $\sum_j \left\|CX_{\theta,b_i,x_{0,i}}(t_{ij}) - y_{ij}\right\|_2^2$ by $\min_{u_i} \mathcal{C}_i(x^u_{i,0}, u_i \mid \theta, b_i, U)$ in the previous criteria and we end up with the following estimator:

$$\widehat{b_i}(\phi) \;:=\; \arg\min_{b_i} G_i(b_i \mid \phi) \tag{6}$$

where:

$$G_i(b_i \mid \phi) = H_i(\widehat{x^u_{0,i}}(\phi, b_i) \mid \phi, b_i) + \frac{\|\Delta b_i\|_2^2}{\sigma^2}. \tag{7}$$

2.3 Outer criterion

2.3.1 *F* general expression

We focus in this section on population parameter estimation. Classic maximum likelihood based approaches generally consider as estimator: $\widehat{\phi} := \mathbb{E}_{b_1}\left[\mathbb{P}(\phi, b_1 \mid \mathbf{y})\right]$. That is, they get rid of the unknown subject specific parameters by taking the mean value of $\mathbb{P}(\phi, b_1 \mid \mathbf{y})$ where $b_1 \sim N(0, \sigma^2 \left(\triangle^T \triangle\right)^{-1})$. This generally requires the numerical approximation of integrals of possibly high dimensions (the same as b_1), a source of approximation and computational issues (Pinheiro and Bates (1994)). To avoid this, we consider the random

effects as nuisance parameters and rely on a classic profiling approach for ϕ^* estimation (Murphy and der Vaart (2000)). Instead of taking the mean, we rely on the profiled joint distribution sequentially with respect to $b := \{b_i\}_{i\in[\![1,n]\!]}$ and $x_0^u = \left\{x_{0,i}^u\right\}_{i\in[\![1,n]\!]}$, or equivalently $\min_b \min_{x_0^u}(-2\ln\mathbb{P}(\phi, b, x_0^u \mid \mathbf{y}))$. Bayes formula gives us $\mathbb{P}(\phi, b, x_0^u \mid \mathbf{y}) \propto \mathbb{P}(\mathbf{y}, b, x_0^u \mid \phi)\mathbb{P}(\phi)$ and we get $\mathbb{P}(\phi, b, x_0^u \mid \mathbf{y}) \propto \left(\prod_i \mathbb{P}(\mathbf{y_i}, b_i, x_{0,i}^u \mid \phi)\right)\mathbb{P}(\phi)$ by conditional independence of subject by subject observations and subject specific parameters. It follows that

$$\min_b \min_{x_0^u}(-2\ln\mathbb{P}\left[\phi, b, x_0^u \mid \mathbf{y}\right]) \propto \sum_i \min_{b_i} \min_{x_{0,i}^u}\left\{-2\ln\mathbb{P}(\mathbf{y_i}, b_i, x_{0,i}^u \mid \phi)\right\} - 2\ln\mathbb{P}(\phi),$$

from which we derive the estimator

$$\begin{aligned}\widehat{\phi} = \arg\min_\phi \Bigg\{ & \sum_i \min_{b_i} \Bigg\{ \min_{x_{0,i}^u} \Bigg\{ \frac{1}{\sigma^2} \sum_j \left\| CX_{\theta,b_i,x_0^i}(t_{ij}) - y_{ij} \right\|_2^2 - 2\ln\mathbb{P}(x_{0,i}^u \mid \phi, b_i) \Bigg\} \\ & + \frac{\|\Delta b_i\|_2^2}{\sigma^2} \Bigg\} \\ & + (d^o \textstyle\sum_i n_i + nq)\ln\sigma^2 - 2n\ln|\triangle| - 2\ln\mathbb{P}(\phi) \Bigg\}\end{aligned}$$

by using the exact expression of $\ln\mathbb{P}(\mathbf{y_i}, b_i, x_{0,i}^u \mid \phi)$ (computational details are recalled in Appendix B). In order to account for the presence of model error and limit its effect on estimation, we replace in the last expression the classic profiled likelihood estimator for b_i^* and $x_{0,i}^{u*}$ by $\widehat{b_i}(\phi)$ and $\widehat{x_{0,i}^u}(\phi, b_i)$ respectively and $X_{\theta,b_i,x_{0,i}}$ by $\overline{X}_{\theta,b_i}$. This leads us to the following population parameter estimator:

$$\widehat{\phi} := \arg\min_\phi F(\phi)\,, \tag{8}$$

where:

$$\begin{aligned}F(\phi) = & \frac{1}{\sigma^2}\sum_i \left(\sum_j \left\| C\overline{X}_{\theta,\widehat{b_i}(\phi)}(t_{ij}) - y_{ij} \right\|_2^2 + \left\| \Delta\widehat{b_i}(\phi) \right\|_2^2 \right) \\ & - 2\ln\mathbb{P}(\widehat{x_{0,i}^u}(\phi, \widehat{b_i}(\phi)) \mid \phi, \widehat{b_i}(\phi)) \\ & + (d^o \textstyle\sum_i n_i + nq)\ln\sigma^2 - 2n\ln|\triangle| - 2\ln\mathbb{P}(\phi)\end{aligned}\,. \tag{9}$$

2.3.2 F profiling on σ for uniform $x_{0,i}^u$ distribution

If $\mathbb{P}(x_{0,i}^u \mid \phi, b_i)$ is constant then $\widehat{x_{0,i}^u}(\phi, b_i)$ and $\widehat{b_i}(\phi)$ do not depend on σ i.e. $\widehat{x_{0,i}^u}(\phi, b_i) = \widehat{x_{0,i}^u}(\theta, b_i)$ and $\widehat{b_i}(\phi) = \widehat{b_i}(\theta, \Delta)$ and consequentially neither does $\overline{X}_{\theta,\widehat{b_i}(\theta,\Delta)}$. So, for each (θ, Δ), the maximizer in σ^2 of F has a closed form expression:

$$\sigma^2(\theta, \Delta) = \frac{1}{(d^o \sum_i n_i + qn)} \sum_i \left(\sum_j \left\| C\overline{X}_{\theta,\widehat{b_i}(\theta,\Delta)}(t_{ij}) - y_{ij} \right\|_2^2 + \left\| \Delta\widehat{b_i}(\theta, \Delta) \right\|_2^2 \right). \tag{10}$$

By using $\sigma^2(\theta, \Delta)$ expression, we get $\min_{\sigma^2} F(\theta, \Delta, \sigma \mid \mathbf{y}) = \overline{F}[\theta, \Delta \mid \mathbf{y}]$ where:

$$\overline{F}[\theta, \Delta \mid \mathbf{y}] = \left(d^o \sum_i n_i + qn \right) \ln\left(\sigma^2(\theta, \Delta)\right) - 2n \ln|\triangle| - 2 \ln \mathbb{P}(\phi).$$

Thus, we can profile F on σ^2 and define our estimator as:

$$\left(\widehat{\theta}, \widehat{\Delta}\right) = \arg\min_{(\theta,\Delta)} \overline{F}[\theta, \Delta \mid \mathbf{y}]. \tag{11}$$

An estimator of σ^* is obtained from there by computing $\sigma^2\left(\widehat{\theta}, \widehat{\Delta}\right)$, given by equation (10). The details of $\overline{F}$ derivation are left in appendix B.

2.4 Asymptotic Variance-Covariance matrix estimator for $\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Delta}}\right)$

We derive an estimator of the asymptotic variance of $\left(\widehat{\theta}, \widehat{\Delta}\right)$. Here we restrict to the case described in section 2.3.2 when a uniform distribution is chosen for $x^u_{i,0}$ and the outer criterion is profiled on σ. The general case can be considered similarly, but we withdraw it for the sake of clarity since it is not used in following simulation works. We highlight that in practice the matrix Δ is parametrized by a vector δ of dimension q', i.e $\triangle := \triangle(\delta)$ and we give here a variance estimator of $\left(\widehat{\theta}, \widehat{\delta}\right)$. From this, the variance of $\widehat{\Delta}$ can be obtained using classic delta-methods (see van der Vaart (1998) chapter 3). Conditions on model structural identifiability and regularity are required to derive the existence of this asymptotic variance, we precise such sufficient conditions in appendix D.

Theorem 1 *There is a model dependent lower bound λ such that if $\|U\|_2 > \lambda$ then the estimator $\left(\widehat{\theta}, \widehat{\delta}\right)$ converges almost surely to a constant value $\left(\overline{\theta}, \overline{\delta}\right)$ such that:*

$$\sqrt{n}(\widehat{\theta} - \overline{\theta}, \widehat{\delta} - \overline{\delta}) \rightsquigarrow N\left(0, A(\overline{\theta}, \overline{\delta})^{-1} B(\overline{\theta}, \overline{\delta}) \left(A(\overline{\theta}, \overline{\delta})^{-1}\right)^T\right),$$

where $A(\overline{\theta}, \overline{\delta}) = \lim_n \frac{1}{n} \sum_{i=1}^n \left[\frac{\partial \widetilde{J}(\overline{\theta}, \overline{\delta}, \mathbf{y_i})}{\partial(\theta, \delta)}\right]$, $B(\overline{\theta}, \overline{\delta}) = \lim_n \frac{1}{n}\left[\sum_i \widetilde{J}(\overline{\theta}, \overline{\delta}, \mathbf{y_i}) \widetilde{J}(\overline{\theta}, \overline{\delta}, \mathbf{y_i})^T\right]$ *and the vector valued function* $\widetilde{J}(\theta, \delta, \mathbf{y_i}) = \begin{pmatrix} \widetilde{J}_\theta(\theta, \delta, \mathbf{y_i}) \\ \widetilde{J}_\delta(\theta, \delta, \mathbf{y_i}) \end{pmatrix}$ *is given by:*

$$\begin{aligned} \widetilde{J}_\theta(\theta, \delta, y_i) &= \tfrac{d}{d\theta} h(\widehat{b}(\theta, \Delta(\delta)), \theta, \Delta(\delta), y_i) \\ \widetilde{J}_\delta(\theta, \delta, y_i) &= \tfrac{d}{d\delta} h(\widehat{b}_i(\theta, \Delta(\delta)), \theta, \Delta(\delta), y_i) \\ &\quad - \tfrac{2}{d^o \mathbb{E}[n_1] + q} Tr\left(\triangle(\delta)^{-1} \tfrac{\partial \triangle(\delta)}{\partial \delta_k}\right) h(\widehat{b}_i(\theta, \Delta(\delta)), \theta, \Delta(\delta), y_i) \end{aligned},$$

where $h(b_i, \theta, \Delta, y_i) = \|\Delta b_i\|_2^2 + \sum_j \left\|C\overline{X}_{\theta, b_i}(t_{ij}) - y_{ij}\right\|_2^2$.

The proof is left in appendix D. The practical interest of this theorem is to give an estimator of Variance-Covariance $V(\widehat{\theta},\widehat{\delta}) \simeq \widehat{A}(\widehat{\theta},\widehat{\delta})^{-1}\widehat{B}(\widehat{\theta},\widehat{\delta})\left(\widehat{A}(\widehat{\theta},\widehat{\delta})^{-1}\right)^T/n$ with the matrices $\widehat{A}(\widehat{\theta},\widehat{\delta}) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial J(\widehat{\theta},\widehat{\delta},\mathbf{y_i})}{\partial(\theta,\delta)}$ and $\widehat{B} = \frac{1}{n}\sum_{i=1}^{n} J(\widehat{\theta},\widehat{\delta},\mathbf{y_i})J(\widehat{\theta},\widehat{\delta},\mathbf{y_i})^T$. The $(p+q)$ components of the vector valued function J for $1 \leq k \leq p$ are given by :

$$J_k(\theta,\delta,\mathbf{y_i}) = \frac{d}{d\theta_k}h(\widehat{b}(\theta,\Delta(\delta)),\theta,\Delta(\delta),\mathbf{y_i}),$$

and for $p+1 \leq k \leq p+q$ by

$$\begin{aligned} J_k(\theta,\delta,\mathbf{y_i}) &= \tfrac{d}{d\delta_k}h(\widehat{b}_i(\theta,\Delta(\delta)),\theta,\Delta(\delta),\mathbf{y_i}) \\ &- \tfrac{2n}{d^o\sum_i n_i+qn}Tr\left(\triangle(\delta)^{-1}\tfrac{\partial\triangle(\delta)}{\partial\delta_k}\right)h(\widehat{b}_i(\theta,\Delta(\delta)),\theta,\Delta(\delta),\mathbf{y_i}). \end{aligned}$$

Now that we have proven the existence of the variance-covariance matrix $V(\overline{\theta},\overline{\delta})$ such that $\widehat{\delta}-\overline{\delta} \rightsquigarrow N\left(0,V(\overline{\theta},\overline{\delta})\right)$, we can use the Delta method to derive the asymptotic normality of the original matrix $\Psi\left(\widehat{\delta}\right) = \sigma^2\left(\Delta(\widehat{\delta})^T\Delta(\widehat{\delta})\right)^{-1}$ as well as an estimator of its asymptotic variance. In the case of a diagonal matrix Ψ, composed of the elements $\left(\Psi_1^2,\dots\Psi_q^2\right)$ and of the parametrization $\triangle(\delta) = Diag(\left\{e^{\delta_l}\right\}_{l\in[\![1,q]\!]})$ used in section 3, we derive:

$$\begin{pmatrix}\Psi_1(\widehat{\delta})\\ \vdots \\ \Psi_q(\widehat{\delta})\end{pmatrix} - \begin{pmatrix}\Psi_1(\delta^*)\\ \vdots \\ \Psi_q(\delta^*)\end{pmatrix} \rightsquigarrow N\left(0,\sigma^2\begin{pmatrix}e^{-\delta_1^*} & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & e^{-\delta_q^*}\end{pmatrix}V(\theta^*,\delta^*)\begin{pmatrix}e^{-\delta_1^*} & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & e^{-\delta_q^*}\end{pmatrix}\right).$$

Remark 1 *The previous theorem 1 states that we retrieve a parametric convergence rate. Thus, we avoid the pitfall described in Sartori (2003) for profiled methods in presence of a number of nuisance parameters increasing with the number of subjects (or strata to resume Sartori (2003) terminology) potentially leading to bias accumulation for score functions among subjects. The i.i.d structure of random effects allows us to rely on central limit theorem to avoid this accumulation phenomenon.*

3 Results on simulated data

We compare the accuracy of our approach with maximum likelihood (ML) in different models and experimental designs reflecting the problems exposed in the introduction, that is estimation in 1/presence of model error, 2/partially observed framework with unknown initial conditions and 3/presence of poorly identifiable parameters. We proceed to Monte-Carlo simulations based on $N_{MC} = 100$ runs. At each run, we generate n_i observations coming from n subjects on an observation interval $[0,T]$ with Gaussian measurement noise of standard deviation σ^*. From these data, we estimate θ^*, Ψ^*

and b_i^* with both estimation methods. We quantify the accuracy of each estimator $\widehat{\psi}_p$ of the population parameters estimate $\widehat{\psi} = \left(\widehat{\theta}, \widehat{\Psi}\right)$ via Monte-Carlo computation of the bias $Bias(\widehat{\psi}_p) = \mathbb{E}\left[\widehat{\psi}_p\right] - \psi_p^*$, the empirical variance $V^e(\widehat{\psi}_p) = \mathbb{E}\left[\left(\mathbb{E}\left[\widehat{\psi}_p\right] - \psi_p^*\right)^2\right]$, the mean squared error $MSE(\widehat{\psi}_p) = Bias(\widehat{\psi}_p)^2 + V^e(\widehat{\psi}_p)$, the estimated variance $\widehat{V}\left(\widehat{\psi}_p\right)$, as well as the coverage rate of the 95%-confidence interval derived from it. This coverage rate, denoted CR in the following results, corresponds to the frequency at which the interval $\left[\widehat{\psi}_p \pm z_{0.975}\sqrt{\widehat{V}\left(\widehat{\psi}_p\right)}\right]$ contains ψ_p^* with $z_{0.975}$ the $0.975-$quantile of the centered Gaussian law. We compute the previous quantities for the normalized values $\widehat{\psi}_p^{norm} := \frac{\widehat{\psi}_p}{\psi_p^*}$ to make relevant comparisons among parameters with different order of magnitude. For b_i^*, we estimate the mean squared error $MSE(\widehat{b}_i) = \mathbb{E}\left[\left\|b_i^* - \widehat{b}_i\right\|_2^2\right]$. For each subsequent examples, we give the results for $n = 50$ and present in appendix C the case $n = 20$ to analyze the evolution of each estimator accuracy with respect to data sparsity.

In the following, we use the superscript ML to denote the ML estimator. For the fairness of comparison with ML, we choose a non-informative prior i.e. $\ln \mathbb{P}[\theta, \Delta] = 0$ for our method throughout this section. Also, we do not use a distribution for $x_{i,0}^u$ for our approach. For ML which requires it, we will use the right parametric form for $\mathbb{P}\left(x_{i,0}^u \mid \phi, b_i\right)$. If the ODE (1) has an analytical solution, the ML estimator is computed via SAEM algorithm (SAEMIX package Comets et al (2017)). Otherwise, it is done via a restricted likelihood method dedicated to ODE models implemented in the nlmeODE package (Tornoe et al (2004)). For our method, we need to select U balancing model and data fidelity in the inner and middle criteria (5)-(7). We use the method presented in G. Hooker and Earn (2011) to compute $\mathrm{EP}_i(U)$, the prediction error for the subject i corresponding to the estimators $\widehat{\theta_U}$, $\left\{\widehat{b_{i,U}}\right\}_{i\in[\![1,n]\!]}$ obtained for a given matrix U. From this, we compute $\mathrm{EP}(U) = \sum_i \mathrm{EP}_i(U)$ the global prediction error for the whole population. We test a trial of weighing matrices $\{U_l\}_{l\in[\![1,L]\!]}$ and retain the one minimizing EP and denote $\widehat{\theta}, \widehat{\Psi}$, $\left\{\widehat{b_i}\right\}_{i\in[\![1,n]\!]}$ the corresponding estimator. For solving the optimization problems required for computing our criteria, we use the Nelder-Mead algorithm implemented in the optimr package (Nash (2016)). All optimization algorithms used here require a starting guess value. We start from the true parameter value for each of them. By doing so, we aim to keep distinct two problems: 1)the numerical stability of the estimation procedures, 2)the intrinsic accuracy of the different estimators. These two problems are correlated, but we aim to adress only the latter which corresponds to the issues raised in introduction. Still, we check on preliminary analysis that local minima presence was not an issue in the neighborhood of $(\theta^*, \triangle^*)$ by testing different starting points for all methods.

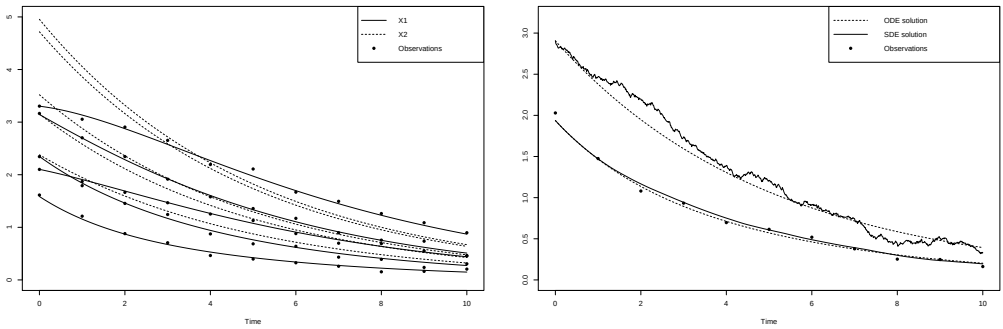

Fig. 1 Left: Examples of solutions of (12) and corresponding observations. Right: Solution of (12) and a realization of (13) for the same parameter values.

No problem appears for our method and SAEMIX. A negligible number of non convergence cases appear for nlmeODE which have been discarded thanks to the convergence criteria embedded in the package.

3.1 Application 1 - Partially observed linear model

We consider the population model where each subject i follows the ODE:

$$\begin{cases} \dot{X}_{1,i} = \phi_{2,i} X_{2,i} - \phi_{1,i} X_{1,i} \\ \dot{X}_{2,i} = -\phi_{2,i} X_{2,i} \\ (X_{1,i}(0), X_{2,i}(0)) = (x_{1,0}, x_{2,0,i}) \end{cases} \tag{12}$$

with the following parametrization: $\log(\phi_{1,i}) = \theta_1 + b_i$ and $\log(\phi_{2,i}) = \theta_2$ where $b_i \sim N(0, \Psi)$. The true population parameter values are $\theta^* = (\theta_1^*, \theta_2^*) = (\log(0.5), \log(2))$ and $\Psi^* = 0.5^2$ and we are in a partially observed framework where only $X_{1,i}$ is accessible. The true initial conditions are distributed with $x_{1,0,i}^* \sim N(2, 0.5)$ and $x_{2,0,i}^* \sim N(3, 1)$. An analytic solution exists for ODE (12). In particular the first component is given by $X_{1,i}(t) = e^{-\phi_{1,i}t}(x_{1,0} + \frac{x_{2,0}\phi_{2,i}}{\phi_{1,i}-\phi_{2,i}}(e^{(\phi_{1,i}-\phi_{2,i})t} - 1))$ and will be used for estimation with the SAEMIX package. We generate $n_i = 11$ longitudinal observations per subject on $[0, T] = [0, 10]$ with measurement noise of standard deviation $\sigma = 0.05$. An example of sampled observations and corresponding solutions are plotted in figure 1. We want to investigate the impact of initial condition, especially the unobserved one $x_{2,0,i}^*$, on the ML estimator accuracy. Indeed, our method does not need to estimate $x_{2,0,i}^*$ and thus no additional difficulties appear in this partially observed framework. For the ML, however, it is a nuisance subject-specific parameter that should be estimated and for which no observations are available. For this, we compute $\widehat{\theta}_{x_0}^{ML}$, $\widehat{\theta}_{x_{0,2}}^{ML}$ and $\widehat{\theta}^{ML}$ the ML estimator respectively when: 1) both initial conditions are perfectly known, 2) $x_{1,0,i}^*$ is replaced by the measured value, 3) in addition, $x_{2,0,i}^*$ has to be estimated.

		Well-specified						Misspecified					
		MSE	Bias	V^e	$\widehat{V}$	CR	MSE b_i	MSE	Bias	V^e	$\widehat{V}$	CR	MSE b_i
θ_1	$\widehat{\theta}^{ML}_{x_0}$	0.01	0.01	0.01	0.01	0.95		0.01	4e-4	0.01	0.01	0.91	
	$\widehat{\theta}^{ML}_{x_{0,2}}$	0.01	0.01	0.01	0.01	0.94		0.01	-3e-4	0.01	1e-4	0.89	
	$\widehat{\theta}^{ML}$	0.04	-0.04	0.04	0.01	0.86		0.05	0.02	0.05	0.01	0.81	
	$\widehat{\theta}$	**5e-3**	**8e-3**	**8e-3**	**1e-2**	**0.97**		**0.01**	**-8e-3**	**7e-3**	**0.05**	**0.97**	
θ_2	$\widehat{\theta}^{ML}_{x_0}$	4e-5	1e-3	4e-5	4e-5	0.95		1e-4	-1e-3	1e-4	1e-4	0.83	
	$\widehat{\theta}^{ML}_{x_{0,2}}$	6e-5	1e-3	6e-5	8e-5	0.94		1e-4	-1e-3	2e-4	0.01	0.82	
	$\widehat{\theta}^{ML}$	4e-3	-0.01	3e-3	1e-4	0.80		4e-3	-2e-3	4e-3	2e-4	0.63	
	$\widehat{\theta}$	**5e-5**	**2e-3**	**4e-5**	**4e-5**	**0.93**		**1e-4**	**2e-5**	**1e-4**	**1e-4**	**0.92**	
Ψ	$\widehat{\theta}^{ML}_{x_0}$	0.01	-0.03	0.01	7e-3	1	0.01	0.01	-0.003	0.01	0.01	1	0.01
	$\widehat{\theta}^{ML}_{x_{0,2}}$	0.02	-0.03	0.01	7e-3	1	0.01	0.01	-0.005	0.01	0.01	1	0.01
	$\widehat{\theta}^{ML}$	0.05	0.17	0.02	0.02	1	0.10	0.09	0.21	0.04	0.03	1	0.12
	$\widehat{\theta}$	**0.01**	**-0.01**	**0.01**	**0.01**	**0.92**	**0.01**	**0.02**	**-0.02**	**0.02**	**0.01**	**0.90**	**0.01**

Table 1 Results of estimation for model (12). The different subscripts stand for the following estimation scenarios: 1) x_0 when both initial conditions are set to $\left(x^*_{0,1}, x^*_{0,2}\right)$, 2) $x_{0,2}$ when $x_{0,i}$ is set to $y_{i,0}$ and $x_{0,2}$ to $x^*_{0,2}$, 3) absence of subscript when $x_{0,i}$ is set to $y_{i,0}$ and $x_{0,2}$ is estimated. Results from our method are in bold.

3.1.1 Well-specified case

We used the exact model described in Section 3.1 for the estimation procedure. Thus we are in a completely well-specified setting, with all mechanisms modeled. We present the estimation results in table 1 - left side. For ML, the results are goods in terms of accuracy and consistent in terms of asymptotic confidence interval coverage rate when both initial conditions are known: 95% for θ_1 and θ_2, which is consistent with theoretical results. However, there is a significant drop in accuracy when $x^*_{2,0,i}$ has to be estimated. In particular, the coverage rate drops to 86% and 80% for θ_1 and θ_2 respectively. Interestingly, ML inaccuracy is driven by bias and under-estimated variance when initial conditions are not known (as shown by a greater V^e than $\widehat{V}$). In this case our method provides a relevant alternative: it gives accurate estimations with a good coverage rate for all parameters while avoiding the estimation of $x^*_{2,0,i}$. Variances are properly estimated compared to empirical variances. Estimation of individual random effects is also more accurate with our method, with a MSE for b_i 10 times smaller compared to ML with unknown initial conditions.

3.1.2 Misspecified case in presence of model error at the subject level

To mimic the presence of misspecification, we now generate the observations from the hypoelliptic stochastic model:

$$\begin{cases} dX_{1,i} = \phi_{2,i} X_{2,i} dt - \phi_{1,i} X_{1,i} dt \\ dX_{2,i} = -\phi_{2,i} X_{2,i} dt + \alpha dB_t \\ (X_{1,i}(0), X_{2,i}(0)) = (x_{1,0}, x_{2,0,i}) \end{cases} \tag{13}$$

with B_t a Wiener process and $\alpha = 0.1$ the diffusion coefficient. For the sake of comparison, a solution of (12) and a realization of its perturbed counterpart given by (13) are plotted in figure 1. This framework where stochasticity only

affects the unmeasured compartment is known to be problematic for parameter estimation and inference procedures are yet to be developed for sparse sampling case. From figure 1 it is easy to see that the diffusion α will be hard to estimate when we only have observations for $X_{1,i}$. Thus, we still estimate the parameters from the model (12) which is now seen as a deterministic approximation of the true stochastic process. Still, it is expected that our method will mitigate the effect of stochasticity on the estimation accuracy by taking into account model misspecification. Results are presented in table 1 - right side. The differences between the two methods are similar to the previous well-specified case with an additional loss of accuracy coming from model error for both estimators. However, the misspecification effect for ML is more pronounced comparing to our method which manages to limit the damages done. This illustrates the benefits of taking into account model uncertainty for estimation, in particular here when model error occurs in the unobserved compartment, a situation in which classic statistical criteria for model assessment based on a data fitting criterion are difficult to use.

3.2 Application 2 - Partially observed nonlinear model

We consider the model presented in De Gaetano and Arino (2000) for the analysis of glucose and insulin regulation:

$$\begin{cases} \dot{G}_i = S_G(G_B - G_i) - X_i G_i \\ \dot{I}_i = \gamma t(G_i - h) - m_i(I_i - I_B) \\ \dot{X}_i = -p_2(X_i + S_I(I_i - I_B)). \end{cases} \tag{14}$$

The ODE system (14) rules the behavior of circulating glucose G_i and insulin I_i in blood as well as insulin X_i present in interstitial fluid. We are in a partially observed case where only G_i and I_i are measured. The values of parameters $(p_2, \gamma, h, G_B, I_B)$ are set to $(-4.93, -6.85, 4.14, 100, 100)$ and we aim to estimate $\theta = (\theta_{S_G}, \theta_{S_I}, \theta_m)$, linked to the original model via the parametrization: $\log(S_G) = \theta_{S_G}$, $\log(S_I) = \theta_{S_I}$ and $\log(m_i) = \theta_m + b_i$ where $b_i \sim N(0, \Psi)$. The true population parameter values are $\theta^* = (-3.89, -7.09, -1.81)$ and $\Psi^* = 0.26^2$. The true subject-specific initial conditions $x^*_{i,0} = \left(G^*_{0,i}, I^*_{0,i}, X^*_{0,i}\right)$ are distributed according to $\ln(x^*_{i,0}) \sim N(l_{x^*_0}, \Psi_{l_{x^*_0}})$ with $l_{x^*_0} = (5.52, 4.88, -7)$ and $\Psi_{l_{x^*_0}} = \left(0.17^2, 0.1^2, 10^{-4}\right)$. We generate $m_i = 5$ observations on $[0, T] = [0, 180]$ with Gaussian measurement noise of standard deviation $\sigma^* = 3$. As in the previous example, we investigate the impact of unknown initial conditions on the estimators accuracy. We are particularly interested in the joint estimation of θ_{S_I}, which appears only in the equation ruling the unobserved state variable X_i, and $x^*_{0,i}$ required for each subject by ML. For this, we distinguish two cases, 1) when θ_{S_I} is known, 2) when θ_{S_I} has to be estimated and we denote respectively $\widehat{\theta_{S_i}}$ and $\widehat{\theta}$ the corresponding estimators. Finally, since the model is nonlinear we have to specify a pseudo-linear representation of the

		Well-specified						Misspecified					
		MSE	Bias	V^e	$\widehat{V}$	CR	MSE b_i	MSE	Bias	V^e	$\widehat{V}$	CR	MSE b_i
θ_{S_G}	$\widehat{\theta}^{ML}_{S_i}$	5e-5	2e-3	4e-5	9e-6	0.95		6e-5	3e-3	6e-5	2e-5	0.85	
	$\widehat{\theta}^{ML}$	2e-3	0.03	1e-3	8e-5	0.85		2e-3	3e-3	1e-3	2e-4	0.54	
	$\widehat{\theta}_{S_i}$	**1e-5**	**4e-4**	**1e-5**	**8e-6**	**0.95**		**2e-5**	**-2e-5**	**2e-5**	**2e-5**	**0.93**	
	$\widehat{\theta}$	**2e-4**	**-6e-4**	**2e-4**	**2e-4**	**0.96**		**3e-4**	**-1e-3**	**3e-4**	**4e-4**	**0.93**	
θ_{S_I}	$\widehat{\theta}^{ML}_{S_i}$	known						known					
	$\widehat{\theta}^{ML}$	2e-3	0.03	1e-3	6e-5	0.90		0.01	0.04	0.01	1e-3	0.55	
	$\widehat{\theta}_{S_i}$	**known**						**known**					
	$\widehat{\theta}$	**1e-4**	**-7e-4**	**1e-4**	**1e-4**	**0.96**		**3e-4**	**-1e-3**	**3e-4**	**3e-4**	**0.92**	
θ_m	$\widehat{\theta}^{ML}_{S_i}$	7e-4	3e-3	6e-4	5e-4	0.94		8e-4	-3e-3	8e-4	5e-4	0.89	
	$\widehat{\theta}^{ML}$	9e-4	8e-3	8e-4	5e-4	0.86		5e-3	-5e-3	5e-3	5e-4	0.88	
	$\widehat{\theta}_{S_i}$	**5e-4**	**6e-3**	**5e-4**	**5e-4**	**0.95**		**4-4**	**7e-4**	**4e-4**	**5e-4**	**0.95**	
	$\widehat{\theta}$	**6e-4**	**6e-3**	**5e-4**	**5e-4**	**0.95**		**4e-4**	**6e-4**	**4e-4**	**5e-4**	**0.96**	
Ψ	$\widehat{\theta}^{ML}_{S_i}$	0.02	7e-4	0.02	0.02	0.95	0.02	0.03	-3e-3	0.03	0.02	0.93	0.03
	$\widehat{\theta}^{ML}$	0.04	-0.09	0.03	0.02	0.88	0.02	0.03	-8e-3	0.02	0.02	0.87	0.03
	$\widehat{\theta}_{S_i}$	**0.01**	**-2e-3**	**0.01**	**0.01**	**0.95**	**0.01**	**0.01**	**-4e-3**	**0.01**	**0.02**	**0.94**	**0.01**
	$\widehat{\theta}$	**0.01**	**3e-3**	**0.01**	**0.01**	**0.94**	**0.01**	**0.02**	**-7e-3**	**0.02**	**0.02**	**0.94**	**0.02**

Table 2 Results of estimation for model (14). The different subscripts stand for the following estimation scenarios: 1) S_i when S_i is set to S_i^*, 2) absence of subscript when S_i is estimated. Results from our method are in bold.

vector field as in (4):

$$A_{\theta,b_i}(t, G_i, I_i, X_i) = \begin{pmatrix} -S_G & 0 & -G_i \\ \gamma t & -m_i & 0 \\ 0 & -p_2 S_I & -p_2 \end{pmatrix}, \; r_{\theta,b_i}(t) = \begin{pmatrix} S_G G_B \\ -\gamma t h + m_i I_B \\ p_2 S_I I_B \end{pmatrix}.$$

3.2.1 Well-specified case

We present the estimation results in table 2 - left side. Our method has small bias and achieve good coverages in all cases. We obtain smaller MSE than ML and avoid the drop in coverage rate of the confidence interval in the case of $\theta^*_{S_I}$ estimation, which is often needed in practice. The difference between the two estimators behaviors is explained by the fact that they are defined through the construction of two different optimization problems. At the population level, our approach leads to minimize a cost function depending on a 4-dimensional parameter whereas ML, due to its need to estimate $x^*_{i,0}$, considers a 10-dimensional one. Thus, the parameter spaces explored by each method to look for the minimum are very different.

3.2.2 Misspecified case in presence of model error at the subject level

To mimic misspecification presence, we generate the observations from the stochastic model:

$$\begin{cases} dG_i = (S_G(G_B - G_i) - X_i G_i)\, dt + \alpha_1 dB_{1,t} \\ dI_i = (\gamma t(G_i - h) - m_i(I_i - I_B))dt + \alpha_2 dB_{2,t} \\ dX_i = (-p_2(X_i + S_I(I_i - I_B)))\, dt + \alpha_3 dB_{3,t} \end{cases}, \tag{15}$$

where the $B_{i,t}$ are Wiener processes and $(\alpha_1, \alpha_2, \alpha_3) = \left(2, 2, 2 \times 10^{-4}\right)$ their diffusion coefficients. We present the estimation results in table 2 - right side.

Parameters		Biological interpretation	Values
δ_L		long-lived B-cells declining rate	$\log(2)/(364 \times 6)$
θ^*	$\theta^*_{\delta_S}$	Mean log-value for δ_S, the short-lived cells declining rate	$\log(\log(2)/1.2) \simeq -0.54$
	$\theta^*_{\phi_S}$	Mean log-value for ϕ_S, the antibodies influx from short-lived cells	$\log(2755) \simeq 7.92$
	$\theta^*_{\phi_L}$	Mean log-value for ϕ_L, the antibodies influx from long-lived cells	$\log(16) \simeq 2.78$
	$\theta^*_{\delta_{Ab}}$	Mean log-value for δ_{Ab}, the antibodies declining rate	$\log(\log(2)/24) \simeq -3.54$
Ψ^*	$\Psi^*_{\phi_S}$	Inter individual variance for $\log(\phi_{S,i})$	0.92^2
	$\Psi^*_{\phi_L}$	Inter individual variance for $\log(\phi_{L,i})$	0.85^2
	$\Psi^*_{\delta_{Ab}}$	Inter individual variance for $\log(\delta_{Ab,i})$	0.3^2

Table 3 Biological interpretation and parameter values

For ML, the drop in coverage rate for $\theta^*_{S_G}$ and $\theta^*_{S_I}$ is even more striking when $\theta^*_{S_I}$ needs to be estimated. This is explained by the effect of model misspecification which increases bias and the fact that ML does not take into account this new source of uncertainty which leads to under-estimation of variance and too narrow confidence intervals. Our method achieved small biais, nominal coverages and small MSE for random effects.

3.3 Application 3 - Antibody concentration evolution model

We consider the model presented in Pasin et al (2019) to analyze the antibody concentration, denoted A_i, generated by two populations of antibody secreting cells: the short lived, denoted S_i, and the long-lived, denoted L_i:

$$\begin{cases} \dot{S}_i = -\delta_S S_i \\ \dot{L}_i = -\delta_L L_i \\ \dot{A}_i = \vartheta_{S,i} S_i + \vartheta_{L,i} L_i - \delta_{Ab,i} A_i \\ (S_i(0), L_i(0), A_i(0)) = (S_{0,i}, L_{0,i}, A_{0,i})\,. \end{cases} \tag{16}$$

This model is used to quantify the humoral response on different populations after an Ebola vaccine injection with a 2 doses regimen seven days after the second injection when the antibody secreting cells enter in a decreasing phase. These cells being unobserved, the preceding equation can be simplified to focus on antibody concentration evolution:

$$\dot{A}_i = \phi_{S,i} e^{-\delta_S t} + \phi_{L,i} e^{-\delta_L t} - \delta_{Ab,i} A_i \tag{17}$$

with $\phi_{S,i} := \vartheta_{S,i} S_{0,i}$ and $\phi_{L,i} := \vartheta_{L,i} L_{0,i}$. This equation has an analytic solution provided in 17 which will be used for ML. We consider the following parametrization: $\log(\delta_S) = \theta_{\delta_S}$, $\log(\phi_{S,i}) = \theta_{\phi_S} + b_{\phi_S,i}$, $\log(\phi_{L,i}) = \theta_{\phi_L} + b_{\phi_L,i}$ and $\log(\delta_{Ab,i}) = \theta_{\delta_{Ab}} + b_{\delta_{Ab},i}$. The true parameter values are presented in table 3. According to Pasin et al (2019), δ_L was non-identifiable based on the available data and only a lower bound has been derived for it via profiled likelihood. So, to make fair comparisons between our approach and maximum likelihood, we do not estimate it. Regarding population parameters, we are particularly interested in the behavior of estimation methods for θ_{δ_S} and θ_{ϕ_S}. Indeed, a parameter sensitivity analysis shows the symmetric role of θ_{δ_S} and θ_{ϕ_S} on the

		Well-specified						Misspecified					
		MSE	Bias	V^e	$\widehat{V}$	CR	MSE b_i	MSE	Bias	V^e	$\widehat{V}$	CR	MSE b_i
θ_{δ_S}	$\widehat{\theta}_{\delta_S}^{ML}$	known						known					
	$\widehat{\theta}_{ML}$	2.13	0.78	1.51	70.64	0.92		3.88	1.48	1.68	4.10	0.80	
	$\widehat{\theta}_{\delta_S}$	**known**						**known**					
	$\widehat{\theta}$	**0.62**	**-0.34**	**0.50**	**0.66**	**0.92**		**0.93**	**-0.40**	**0.77**	**0.62**	**0.90**	
θ_{ϕ_S}	$\widehat{\theta}_{\delta_S}^{ML}$	4e-4	0.01	3e-4	3e-4	0.94		1e-3	0.02	1e-3	5e-4	0.91	
	$\widehat{\theta}_{ML}$	0.01	-0.05	7e-3	0.40	0.92		0.02	-0.10	0.01	0.02	0.88	
	$\widehat{\theta}_{\delta_S}$	**2e-3**	**-0.05**	**2e-4**	**1e-3**	**0.94**		**7e-4**	**-0.02**	**3e-4**	**1e-3**	**0.92**	
	$\widehat{\theta}$	**2e-3**	**1e-3**	**2e-3**	**2e-3**	**0.93**		**4e-3**	**-6e-3**	**3e-3**	**0.01**	**0.90**	
θ_{ϕ_L}	$\widehat{\theta}_{\delta_S}^{ML}$	3e-3	0.02	3e-3	2e-3	0.95		5e-3	0.03	4e-3	3e-3	0.93	
	$\widehat{\theta}_{ML}$	4e-3	0.03	4e-3	3e-3	0.90		9e-3	0.05	7e-3	4e-3	0.90	
	$\widehat{\theta}_{\delta_S}$	**7e-4**	**-0.01**	**5e-4**	**3e-3**	**0.95**		**2e-3**	**-0.02**	**3e-3**	**2e-3**	**0.97**	
	$\widehat{\theta}$	**3e-3**	**-3e-3**	**3e-3**	**2e-3**	**0.91**		**6e-3**	**-8e-3**	**6e-3**	**7e-3**	**0.90**	
$\theta_{\delta_{Ab}}$	$\widehat{\theta}_{\delta_S}^{ML}$	7e-4	-0.02	5e-4	3e-4	0.93		2e-3	-0.03	1e-3	1e-3	0.92	
	$\widehat{\theta}_{ML}$	2e-3	-0.02	1e-3	4e-4	0.88		4e-3	-0.04	3e-3	7e-4	0.88	
	$\widehat{\theta}_{\delta_S}$	**2e-4**	**0.01**	**1e-4**	**3e-4**	**0.95**		**3e-4**	**2e-3**	**3e-4**	**3e-4**	**0.96**	
	$\widehat{\theta}$	**4e-4**	**0.01**	**3e-4**	**2e-4**	**0.90**		**3e-4**	**8e-3**	**3e-4**	**2e-3**	**0.89**	
Ψ_{ϕ_S}	$\widehat{\theta}_{\delta_S}^{ML}$	0.04	-1e-3	0.04	0.07	1	0.15	0.05	0.03	0.05	0.08	1	0.17
	$\widehat{\theta}_{ML}$	0.11	0.01	0.11	0.05	1	0.17	0.13	0.01	0.13	0.25	1	0.21
	$\widehat{\theta}_{\delta_S}$	**0.02**	**8e-3**	**0.02**	**0.01**	**0.94**	**0.06**	**0.02**	**2e-3**	**0.02**	**0.02**	**0.94**	**0.11**
	$\widehat{\theta}$	**0.02**	**-0.03**	**0.02**	**0.02**	**0.94**	**0.07**	**0.02**	**-0.05**	**0.02**	**0.03**	**0.92**	**0.08**
Ψ_{ϕ_L}	$\widehat{\theta}_{\delta_S}^{ML}$	0.03	0.04	0.02	0.04	1	0.30	0.05	0.03	0.05	0.06	1	0.73
	$\widehat{\theta}_{ML}$	0.03	0.05	0.02	0.04	1	0.60	0.03	0.05	0.02	0.07	1	0.74
	$\widehat{\theta}_{\delta_S}$	**0.02**	**-0.1**	**5e-3**	**8e-3**	**0.93**	**0.07**	**0.02**	**-0.10**	**0.01**	**0.02**	**0.91**	**0.10**
	$\widehat{\theta}$	**0.03**	**-0.06**	**0.02**	**0.01**	**0.92**	**0.08**	**0.03**	**-0.06**	**0.02**	**0.03**	**0.87**	**0.12**
$\Psi_{\delta_{Ab}}$	$\widehat{\theta}_{\delta_S}^{ML}$	0.11	0.18	0.08	0.02	1	0.10	0.33	0.41	0.17	0.05	1	0.56
	$\widehat{\theta}_{ML}$	0.20	0.29	0.11	0.02	1	0.50	0.30	0.34	0.19	0.05	1	0.69
	$\widehat{\theta}_{\delta_S}$	**0.10**	**-0.30**	**0.01**	**0.01**	**0.95**	**0.03**	**0.10**	**-0.16**	**0.08**	**0.06**	**0.91**	**0.04**
	$\widehat{\theta}$	**0.11**	**-0.27**	**0.04**	**0.04**	**0.95**	**0.04**	**0.15**	**-0.29**	**0.06**	**0.10**	**0.88**	**0.06**

Table 4 Results of estimation for model (17). The different subscripts stand for the following estimation scenarios: 1) δ_S when θ_{δ_S} is set to $\theta^*_{\delta_S}$, 2) absence of subscript when θ_{δ_S} is estimated. Results from our method are in bold.

ODE solution (see Balelli et al (2020)). Thus, they are likely to face practical identifiability problems. To investigate this effect, we estimate the parameters when $\theta^*_{\delta_S}$ 1) is known (the corresponding estimators will be denoted with the subscript δ_S), or 2) has to be estimated as well.

3.3.1 Well-specified case

We generate $n_i = 11$ longitudinal observations on the interval $[0, T] = [0, 364]$ with measurement noise of standard deviation $\sigma^* = 100$. For each subject i, the initial condition has been generated according to $A^*_{0,i} \sim N(\overline{A_0}, \sigma^2_{\overline{A_0}})$ with $\overline{A_0} = 500$ and $\sigma_{\overline{A_0}} = 260$ to reflect the dispersion observed in data presented in Pasin et al (2019). We present the estimation results in table 4 - left side.

Our method gives an improved estimation with a dramatically reduced variance for $\theta^*_{\delta_S}$ comparing to ML, as well as an improved estimate for the $\{b^*_i\}_{i\in[\![1,n]\!]}$ in all cases. We assume that is due to the committed estimation error for θ^* which causes model error during $\{b^*_i\}_{i\in[\![1,n]\!]}$ estimation, not accounted for by ML. This in turn explains why variance Ψ^* is better estimated with our approach. In this mixed-effect context, this cause of model error is systematically present and claims for the use of estimation methods taking it into account when subject specific parameters are critical for the practitioner.

3.3.2 Misspecified case in presence of model error at the subject level

The data are generated with a stochastic perturbed version of ODE (17):

$$dA_i = \left(\phi_{S,i}e^{-\delta_S t} + \phi_{L,i}e^{-\delta_L t} - \delta_{Ab,i}A_i\right)dt + \alpha dB_t \tag{18}$$

where B_t is a Wiener process and $\alpha = 10$ its diffusion coefficient. The value for α has been chosen big enough to produce significantly perturbed trajectories but small enough to ensure that ODE (17) is still a relevant approximation for estimation purpose. The results are presented in table 4 - right side. Our method outperforms the ML for $\theta^*_{\delta_S}$ as well as for $\{b_i^*\}_{i\in[1,\,n]}$ estimation and their variances. However, we acknowledge that this last simulation setting is challenging even for our approach with confidence coverage around 90% for most of parameters, below the theoretical rate of 95%.

4 Real data analysis

We use the presented estimation approach to address the same problem as Pasin et al (2019). This real data example is similar to the synthetic scenario performed in Section 3.3. In brief, we use data from a phase I trial in East Africa evaluating the effect of an heterologous anti-Ebola vaccine strategy in which Ad26.ZEBOV was injected first and then MVA-BN-Filo with a delay of 28 days between the two doses. We consider a population of n=28 individuals, with in average 5 measurements per subject. In order to ensure a fair comparison, we adopt a Bayesian framework for $\theta = (\theta_{\delta_S}, \theta_{\phi_S}, \theta_{\phi_L}, \theta_{\delta_{Ab}})$ and used the same prior distribution as in the original paper:

$$\pi(\theta) \sim N\left(\begin{pmatrix} -1 \\ 0 \\ 0 \\ -4.1 \end{pmatrix}, \begin{pmatrix} 25 & 0 & 0 & 0 \\ 0 & 100 & 0 & 0 \\ 0 & 0 & 100 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}\right).$$

We set our mesh-size to get 200 discretization points for each subject and we use $U = 10$ i.e., a value lower than in the simulated data case because of the presence of model error. We also proceed to the log-transformation of the data to stabilize the measurement noise variance. Using the transformation $\widetilde{A_i}(t) := \log_{10} A_i(t)$ in Equation (17), this drives us to use the following nonlinear model:

$$\dot{\widetilde{A_i}}(t) = \frac{1}{\ln(10)}\left(\phi_{S,i}e^{-\delta_S t} + \phi_{L,i}e^{-\delta_L t}\right)10^{-\widetilde{A_i}(t)} - \frac{\delta_{Ab,i}}{\ln(10)}. \tag{19}$$

We choose $A_{\theta,b_i}(t, x, z_i(t)) = \frac{1}{\ln(10)}\left(\phi_{S,i}e^{-\delta_S t} + \phi_{L,i}e^{-\delta_L t}\right)\frac{10^{-x}}{x}$ and $r_{\theta,b_i}(t, z_i(t)) = -\frac{\delta_{Ab,i}}{\ln(10)}$ for the pseudo-linear formulation of the model. In Table 5, we compare our estimations with those presented in Pasin et al (2019) obtained using the NIMROD software (Prague et al (2013)). Both methods

Parameter	Estimations from Pasin et al (2019) Mean IC95%	Optimal Control appraoch Mean IC95%
θ_{δ_S}	-0.57 [-1.02, -0.02]	-0.18 [-0.58, 0.22]
θ_{ϕ_S}	7.92 [7.52, 8.30]	7.45 [6.85, 7.96]
θ_{ϕ_L}	2.78 [2.62, 3.01]	2.58 [2.15, 3.01]
$\theta_{\delta_{Ab}}$	-3.54 [-3.62, -3.45]	-3.48 [-3.95, -3.01]
Ψ_{ϕ_S}	0.92 [0.83, 1.01]	0.64 [0.60, 0.70]
Ψ_{ϕ_L}	0.85 [0.78, 0.92]	0.70 [0.55, 0.90]
$\Psi_{\delta_{Ab}}$	0.30 [0.24, 0.36]	0.25 [0.19, 0.31]

Table 5 Estimation presented in Pasin et al (2019) and via our approach.

produce estimations with overlapping confidence intervals for θ supporting the previous published results in term of antibodies concentrations dynamics over time. Still, significant differences appear for $(\Psi_{\phi_S}, \Psi_{\phi_L}, \Psi_{\delta_{Ab}})$ estimation with lower dispersion of random effects in the optimal control approach. This is explained by the fact that a part of the variability is now carried out by subject-specific perturbations $\left\{\overline{u}_{i,\widehat{\theta},b_i(\widehat{\theta})}\right\}_{i\in[\![1,\,n]\!]}$. Figure 2 allows to visually check that the individual fits are comparable between the two approaches.

Finally, our method can be used to assess the model adequacy via the temporal evolution analysis of $\left\{\overline{u}_{i,\widehat{\theta},b_i(\widehat{\theta})}\right\}_{i\in[\![1,\,n]\!]}$ estimated as byproducts of our method. In Section 2.1, we have also indicated that perturbations $\overline{u}_{i,\theta,b_i,x_{0,i}}$ can be computed for an arbitrary set $(\theta, b_i, x_{0,i})$. In particular, we estimate $\left\{\overline{u}_{i,\widehat{\theta}^P,\widehat{b_i}^P,y_{0,i}}\right\}_{i\in[\![1,\,n]\!]}$, the committed error corresponding to $(\widehat{\theta}^P, \widehat{b_i}^P)$, the population and subject specific estimators obtained in Pasin et al (2019). In Figure 3, we plot both perturbation sets. Our method leads to residual perturbations of smaller magnitudes and narrower confidence intervals. This means that our approach produces an estimation which minimizes the committed model error for each subject comparing to a method based only on a data fitting criterion as in Pasin et al (2019). Moreover, by reducing the size of the confidence interval for estimated perturbations, we conclude to a mean perturbation among the population which is statistically different from zero at the beginning of observation interval. This may indicates presence of model misspecification.

5 Conclusion

In this paper, we propose an estimation method that addresses problems encountered by classical approaches in NLME-ODE models. We identify the following shortcomings for exact methods such as likelihood-based inference: their difficulty in the presence of model misspecification, their need to estimate initial conditions as regular random effects, and their dramatic performance degradation in the presence of poorly identifiable parameters. We propose here a method based on control theory that accounts for the presence of potential model uncertainty at the subject level and that can be easily profiled on initial conditions. Simulations with and without model error illustrate the advantages

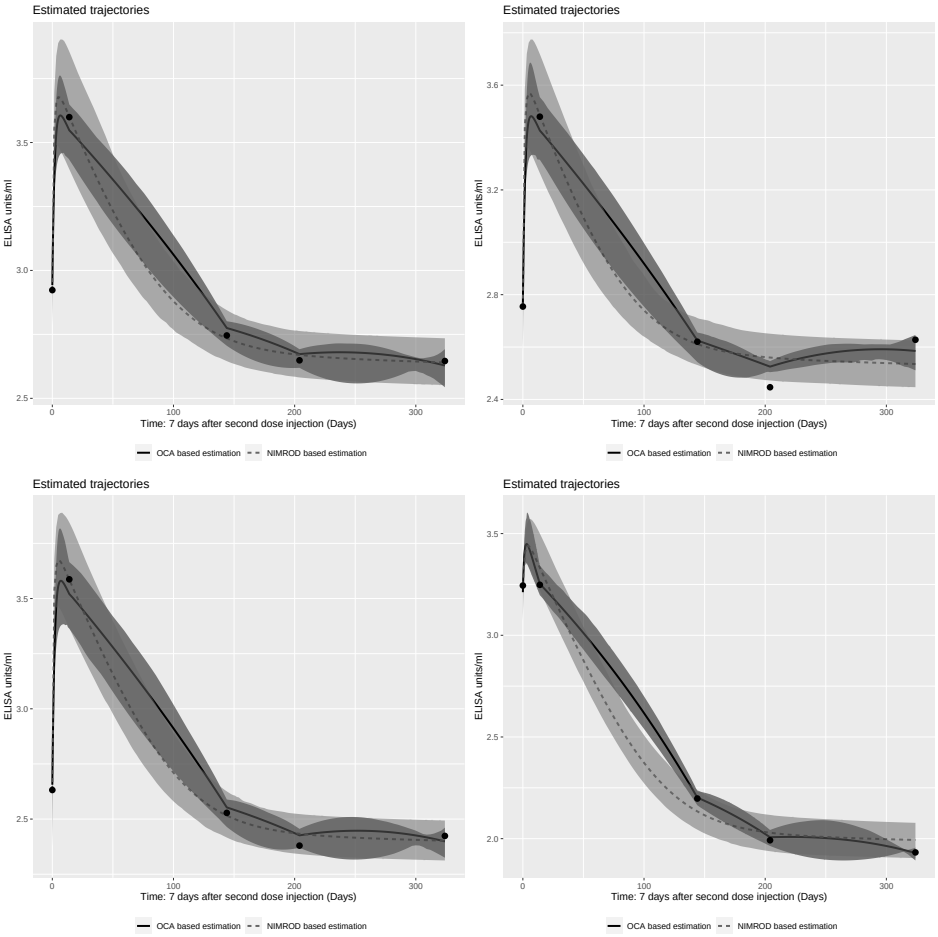

Fig. 2 Examples of fitted trajectories for both methods for four different random subjects. Dashed lines: fitted ODE solutions from Pasin et al (2019). Solid line: optimal trajectories $\overline{X}_{\widehat{\theta},\widehat{b_i}}$ obtained with optimal control approach. Shaded area are the 95% confidence intervals.

of regularization techniques for estimating poorly identifiable parameters, subject-specific parameters, and their variances in NLME-ODEs. In addition, bypassing the estimation of initial conditions represents a clear advantage for partially observed systems comparing to likelihood based approaches, as emphasized in the simulations.

Still, this benefit in term of estimation accuracy comes with a computational price. On a server (see https://plafrim-users.gitlabpages.inria.fr/doc/ for more server details) with the parallelization package Snow in R language, it takes approximately 10-15 minutes to obtain an estimation for the two-dimensional linear model, 30 minutes for the insulin model and 3-4 hour for the antibody concentration evolution one, whereas it was a matter of minutes for the other approaches. Nevertheless, the use of compiled languages and proper

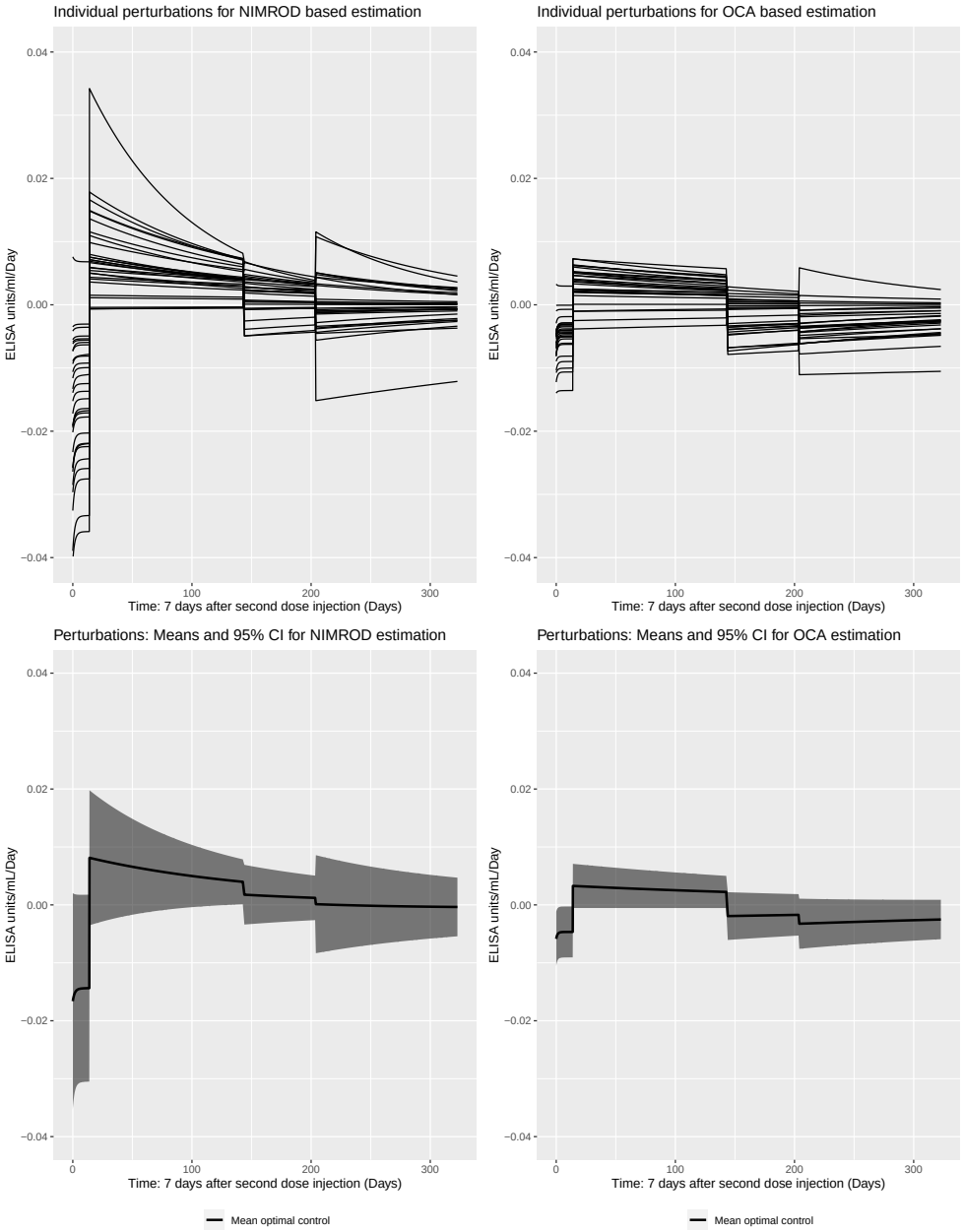

Fig. 3 1) Up: Estimated residual controls for each subject, 2) bottom: mean optimal control and 95% confidence interval for the optimal controls a) left: $\overline{u}_{i,\widehat{\theta}^P,\widehat{b_i}^P,y_{0,i}}$ obtained from parameter estimation in Pasin et al (2019), b) right: $\overline{u}_{i,\widehat{\theta},b_i(\widehat{\theta})}$ obtained from our estimation.

parallelization could reduce the computation time. Moreover, we have willingly separated the formal definition of the optimal control problem required by our method and the numerical procedure used to solve it, in case it may exists better suited approaches for this specific control problem. Right now, our current strategy allows us to profile on initial conditions, therefore looking for another numerical procedure is beyond the scope of this paper.

Finally, the qualitative assessment of model misspecification exposed in section 4 can be made more rigorous. In a one subject setting, the estimation of a perturbation term at the derivative level via non-parametric procedures to test model error presence has been already explored (Hooker et al (2015); Engelhardt et al (2017)). Comparing to statistical methods solely based on data fitting criteria, they generally produce more sensitive statistical tests and can explore misspecification presence even for unobserved state-variables. Our control based approach can extend such tests to a population framework, while avoiding issues due to hyperparameter selection required for non-parametric statistical methods which can appear for a growing number of subjects. For example, to stay in a Bayesian setting, we can specify a prior distribution for the controls and then compare it with the obtained posterior once the inference is made. This would lead to a semi-parametric inference problem for which an optimal control based approach has already been proven useful (see Clairon (2020)). This is a subject for further work.

Supplementary information. A supplementary file containing the appendixes and proof of theorem 1 is available alongside this article.

Acknowledgments. Experiments presented in this paper were carried out using the PlaFRIM experimental testbed, supported by Inria, CNRS (LABRI and IMB), Université de Bordeaux, Bordeaux INP and Conseil Régional d'Aquitaine (see https://www.plafrim.fr/). This manuscript was developed under WP4 of EBOVAC3.

Funding: This work has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under projects EBOVAC1 and EBOVAC3 (respectively grant agreement No 115854 and No 800176). The IMI2 Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and the European Federation of Pharmaceutical Industries and Association.

Declarations

- Funding: See Acknowledgments.
- Conflict of interest/Competing interests: The authors have no conflicts of interest to declare that are relevant to the content of this article.
- Ethics approval: Not applicable.
- Consent to participate: Not applicable.
- Consent for publication: Not applicable.
- Availability of data and materials: Not applicable.

- Code availability: Our estimation method is implemented in R and a code reproducing the examples of Section 3 is available on a GitHub repository located here.

References

Aliyu M (2011) Nonlinear H-Infinity Control, Hamiltonian Systems and Hamilton-Jacobi Equations. CRC Press

Balelli I, Pasin C, Prague M, et al (2020) A model for establishment, maintenance and reactivation of the immune response after vaccination against ebola virus. Journal of Theoretical Biology p 110254

Brynjarsdottir J, O'Hagan A (2014) Learning about physical parameters: The importance of model discrepancy. Inverse Problems 30:24

Campbell D (2007) Bayesian collocation tempering and generalized profiling for estimation of parameters from differential equation models. PhD thesis, McGill University Montreal,Quebec

Cimen T (2008) State-dependent riccati equation (sdre) control: A survey. IFAC Proceedings 41:3761–3775

Cimen T, Banks S (2004) Global optimal feedback control for general nonlinear systems with nonquadratic performance criteria. Systems and Control Letters 53:327–346

Clairon Q (2020) A regularization method for the parameter estimation problem in ordinary differential equations via discrete optimal control theory. Journal of Statistical Planning and Inference

Clarke F (2013) Functional Analysis, Calculus of Variations and Optimal Control. Graduate Texts in Mathematics, Springer-Verlag London

Comets E, Lavenu A, Lavielle M (2017) Parameter estimation in nonlinear mixed effect models using saemix, an r implementation of the saem algorithm. Journal of Statistical Software 80:1–42

De Gaetano A, Arino O (2000) Mathematical modelling of the intravenous glucose tolerance test. Journal of mathematical biology 40(2):136–168

Donnet S, Samson A (2006) Estimation of parameters in incomplete data models defined by dynamical systems. Journal of Statistical Planning and Inference 137(9):2815–2831

Engelhardt B, Kschischo M, Fröhlich H (2017) A bayesian approach to estimating hidden variables as well as missing and wrong molecular interactions

in ordinary differential equation-based mathematical models. Journal of The Royal Society Interface 14(131):20170,332

Engl H, Flamm C, Kügler P, et al (2009) Inverse problems in systems biology. Inverse Problems 25(12)

G. Hooker LDVRS. P. Ellner, Earn DJD (2011) Parameterizing state-space models for infectious disease dynamics by generalized profiling: measles in ontario. Journal of the Royal Society 8:961–974

Guedj J, Thiebaut R, Commenges D (2007) Maximum likelihood estimation in dynamical models of hiv. Biometrics 63:1198–206

Gutenkunst RN, Waterfall J, Casey F, et al (2007) Universally sloppy parameter sensitivities in systems biology models. Public Library of Science Computational Biology 3:e189

Hooker G, Ellner SP, et al (2015) Goodness of fit in nonlinear dynamics: misspecified rates or misspecified states? The Annals of Applied Statistics 9(2):754–776

Huang Y, Dagne G (2011) A bayesian approach to joint mixed-effects models with a skew normal distribution and measurement errors in covariates. Biometrics 67:260–269

Huang Y, Lu T (2008) Modeling long-term longitudinal hiv dynamics with application to an aids clinical study. Annal of Applied Statistics 2:1348–1408

Kampen NV (1992) Stochastic Process in Physics and Chemistry. Elsevier

Kennedy MC, O'Hagan A (2001) Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63(3):425–464

Kurtz T (1978) Strong approximation theorems for density dependent markov chains. Stochastic Processes and their Applications 6:223–240

Lavielle M, Aarons L (2015) What do we mean by identifiability in mixed effects models? Journal of pharmacokinetics and pharmacodynamics

Lavielle M, Mentré F (2007) Estimation of population pharmacokinetic parameters of saquinavir in hiv patients with the monolix software. Journal of Pharmacokinetics and Pharmacodynamics 34

Leary TO, Sutton A, Marder E (2015) Computational models in the age of large datasets. Current Opinion in Neurobiology 32:87–94

Lunn D, A.Thomas, Best N, et al (2000) Winbugs - a bayesian modelling framework: Concepts, structure and extensibility. Statistics and Computing 10:325–337

M. Lavielle AFA. Samson, Mentre F (2011) Maximum likelihood estimation of long terms hiv dynamic models and antiviral response. Biometrics 67:250–259

Murphy S, der Vaart AV (2000) On profile likelihood. Journal of American Statistical Association 95:449–465

Nash JC (2016) Using and extending the optimr package

Pasin C, Balelli I, Van Effelterre T, et al (2019) Dynamics of the humoral immune response to a prime-boost ebola vaccine: quantification and sources of variation. Journal of virology 93(18):e00,579–19

Perelson A, Neumann A, Markowitz M, et al (1996) Hiv-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science 271:1582–1586

Pinheiro J, Bates DM (1994) Approximations to the loglikelihood function in the nonlinear mixed effects model. Journal of the Computational and Graphical Statistics 4:12–35

Prague M, Commengues D, Guedj J, et al (2013) Nimrod: A program for inference via a normal approximation of the posterior in models with random effects based on ordinary differential equations. Computer Methods and Programs in Biomedicine 111:447–458

Raftery A, Bao L (2010) Estimating and projecting trends in hiv/aids generalized epidemics using incremental mixture importance sampling. Biometrics 66:1162–1173

Ramsay J, Hooker G, Cao J, et al (2007) Parameter estimation for differential equations: A generalized smoothing approach. Journal of the Royal Statistical Society (B) 69:741–796

Sartori N (2003) Modified profile likelihood in models with stratum nuisance parameters. Biometrika 90:553–549

Sontag E (1998) Mathematical Control Theory: Deterministic finite-dimensional systems. Springer-Verlag (New-York)

Tornoe C, Agerso H, Jonsson EN, et al (2004) Non-linear mixed-effects pharmacokinetic/pharmacodynamic modelling in nlme using differential equations. Computer Methods and Programs in Biomedicine 76:31–41

Tuo R, Wu C (2015) Efficient calibration for imperfect computer models. Annals of Statistics

van der Vaart A (1998) Asymptotic Statistics. Cambridge Series in Statistical and Probabilities Mathematics, Cambridge University Press

Varah JM (1982) A spline least squares method for numerical parameter estimation in differential equations. SIAM Jsci Stat Comput 3(1):28–46

Villain L, Commenges D, Pasin C, et al (2019) Adaptive protocols based on predictions from a mechanistic model of the effect of il7 on cd4 counts. Statistics in medicine 38(2):221–235

Wang L, Cao J, Ramsay J, et al (2014) Estimating mixed-effects differential equation models. Statistics and Computing 24:111–121