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

Abstract

We present a parameter estimation method for nonlinear mixed effect models based on ordinary differential equations (NLME-ODEs). The method presented here aims at regularizing the estimation problem in presence of model misspecifications, practical identifiability issues and unknown initial conditions. For doing so, we define our estimator as the minimizer of a cost function which incorporates a possible gap between the assumed model at the population level and the specific individual dynamic. The cost function computation leads to formulate and solve optimal control problems at the subject level. This control theory approach allows to bypass the need to know or estimate initial conditions for each subject and it regularizes the estimation problem in presence of poorly identifiable parameters. Comparing to maximum likelihood, we show on simulation examples that our method improves 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 application on antibody concentration data after vaccination against Ebola virus coming from phase 1 trials. We use the estimated model discrepancy at the subject level to analyze the presence of model misspecification.

1 Introduction

ODE models are standard in population dynamics, epidemiology, virology, pharmacokinetics, or genetic regulation networks analysis due to their ability to describe the main mechanisms of interaction between different biological components of complex systems, their evolution in time and to provide reasonable approximations of stochastic dynamics [47, 37, 64, 4, 46, 42, 38, 18, 66]. Evidence of the relevance of ODEs resides for example in their joint use with control theory methods for the purpose of optimal treatment design [22, 1, 67, 45, 63]. In cases of experimental designs involving a large number of subjects and limited number of individual measurements, non-linear mixed-effect models may be more relevant than subject-by-subject model to gather information from the whole population while allowing between-individual variability. For example, clinical trials and pharmacokinetics/pharmacodynamics studies often fall into this category [42, 21, 27, 65, 49]. Formally, we are interested in a population where the dynamics of the compartments of each subject $i \in [\![1,\, n]\!]$ is modeled by the d-dimensional ODE:

$$\begin{cases} \dot{x}_i(t) = f_{\theta,b_i}(t, x_i(t), z_i(t)) \\ x_i(0) = x_{i,0} \end{cases} \tag{1.1}$$

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 belonging to $\mathbb{R}^d$ and z_i is a covariate function. We denote $X_{\theta,b_i,x_{i,0}}$ the solution of (1.1) for a given set $(\theta, b_i, x_{i,0})$.

Our goal is to estimate the true population parameters $(\theta^*,\, \Psi^*)$ 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:

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

where t_{ij} is the j-th measurement time-point for the $i-$th subject on the observation interval $[0,\, T]$. 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. We also assume only a subset of the true initial condition $x_{i,0}^*$, denoted $x_{i,0}^{k*}$, is known, the other ones, denoted $x_{i,0}^{u*}$, being unknown. For the sake of clarity, we order the state variables as follows: $x_{i,0} = \left(\left(x_{i,0}^u\right)^T, \left(x_{i,0}^k\right)^T\right)^T$. We denote n_i the number of observations for the i-th subject, $\mathbf{y_i} = \{y_{ij}\}_{j\in[\![1,\, n_i]\!]}$ its corresponding set of observations and $\mathbf{y} = \{\mathbf{y_i}\}_{i\in[\![1,\, n]\!]}$ the set of all observations in the population.

Our problem belongs to the class parameter estimation problem in nonlinear mixed effect models. In this context, frequentist methods based on likelihood maximization (via different numerical procedures: Laplace approximation [48], Gaussian quadrature [48, 40, 21] or SAEM [34, 37, 15]) and Bayesian ones aiming to reconstruct the a posteriori distribution or to derive the maximum a posteriori estimator (via MCMC algorithms [41, 25, 28], importance sampling [50], approximation of the asymptotic posterior distribution [49]) 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 [57], to reduce the computational time [17] or to avoid the repeated model integration and estimation of initial conditions [65]. However, all the preceding methods face similar pitfalls due to specific features of population models based on ODEs (with the exception of [65]):

1. They do not account for model misspecification presence, a common feature in ODE models used in biology. Indeed, the ODE modeling process suffers from model inadequacy, understood as the discrepancy between the mean model response and real world process, and residual variability issues, that is subject specific stochastic perturbations or missed elements which disappear by averaging over the whole population [30]. As examples of model inadequacy causes, one can think of ODE models used in epidemiology and virology which are derived by approximations where for instance, interactions are modeled by pairwise products while higher order terms and/or the influence of unknown/unmeasured external factors are neglected [54]. Regarding residual variability, let us remind that biological processes are often stochastic [6, 33] and the justification of deterministic modeling comes from the approximation of stochastic processes [35, 20, 29]. Moreover, in the context of population models, new sources of model uncertainties emerge. Firstly, error measurement in covariates z_i which is not often considered leads to use a proxy function $\widehat{z_i}$ instead of z_i [25]. 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 the true population parameter value θ^*. Thus, the structure of mixed-effect models spread measurement uncertainty into the mechanistic model structure during the estimation. It turns classical statistical uncertainties into model error causes. Estimation of θ^*, Ψ^* and $\{b_i^*\}_{i\in[\![1,\,n]\!]}$ has to be done with model misspecification presence although it is known to dramatically impair the accuracy of methods which do not take into account potential modeling error [7, 32].

2. They have to estimate or make assumptions on $x_{i,0}^{u*}$ values. In ODE models, the initial conditions are generally nuisance parameters in the sense that knowing their values does not bring answers to the scientific questions which motivate the model construction but the estimation of the relevant parameters requires $x_{i,0}^*$ inference as well. 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 their values [42, 21, 56], another potential cause of model misspecifications, or the need to estimate them [26, 27] which increases the optimization problem dimension and degrades estimation accuracy due to covariance effect between (θ^*, Ψ^*) and $x_{i,0}^{u*}$ estimate.

3. They can face accuracy degradation when the inverse problem of parameter estimation is

ill-posed [18, 55] 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 [23, 59, 58, 39]. This problem, called sloppiness, often appears in ODE models used in biology [23, 39] and leads to an ill-conditioned Fisher Information Matrix. For maximum likelihood estimators this is a cause for high variance due to the Cramér-Rao bound. For Bayesian inference, it leads to a nearly singular asymptotic a posteriori distribution because of Bernstein–von Mises theorem (see [8] for the computational induced problems). Despite this problem is in part mitigated by the population approach which merges different subjects for estimating (θ^*, Ψ^*) and uses distribution of $b_i \mid \Psi$ as prior at the subject level [36], estimation accuracy can benefit from the use of regularization techniques for the inverse problem.

These specific features of ODE-based population models limit the amount of information classic approaches can extract for estimation purposes from observations not matter their qualities or abundances. This advocates for the development of new estimation procedures. Approximate methods [62, 51, 12, 13] have already proven to be useful for ODE models facing these issues with observations coming from one subject. These approaches rely on an approximation of the solution of the original ODE (1.1) 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 and they reduce the effect of model error on estimator accuracy. Secondly, some of these approximations bypass the need to estimate initial conditions [51, 13]. In this work, we generalize one of these approaches to population models by developing a new estimation method specific to NLME-ODEs aiming to integrate such approximations to mitigate the effect of model misspecification and poorly identifiable parameter on estimation accuracy, while avoiding the need to estimate $x_{i,0}^{u*}$. We propose here a nested estimation procedure where population parameters $(\theta^*, \Psi^*, \sigma^*)$ are estimated through the maximization of an outer criterion. This requires in turn an estimator for the $\{b_i^*\}_{i\in[\![1,\,n]\!]}$ obtained through the repeated optimization of inner criteria. We consider that the actual dynamic for each subject is described by a perturbed version of the ODE (1.1) where the added perturbation captures different sources of errors at the subject level [7, 60]. 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 $\{b_i^*\}_{i\in[\![1,\,n]\!]}$ estimators require to solve optimal control problems [14, 31, 53] known as tracking problems. This is done using a method inspired by [11, 10] based on pseudo-linear representation and Linear-Quadratic theory. In addition, our method does not need to estimate $x_{i,0}^{u*}$. Nevertheless, it can provide an estimator of $x_{i,0}^{u*}$ as a direct byproduct of structural parameters estimation with no additional computational costs.

In section 2, we present the estimation method and derive the inner and outer criteria. In section 3, we analyse the asymptotic behavior of (θ^*, Ψ^*) estimator and derive an approximation of its asymptotic Variance-Covariance matrix from it. In section 4, we compare our approach with classic maximum likelihood in simulations. We then 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 [46]. Section 6 concludes and discuss further applications and extensions of the method.

2 Construction of the estimator: definition of the inner and outer criteria

From now on, we use the following Choleski decomposition $\sigma^2\Psi^{-1} = \triangle^T\triangle$ (or equivalently $\Psi = \sigma^2 \left(\triangle^T\triangle\right)^{-1}$) and the parametrization (θ, Δ, σ) instead of (θ, Ψ, σ). This parametrization will allows us to enforce positiveness and symmetry of Ψ and to derive an explicit estimator of σ given a value for (θ, Δ). The norm $\|.\|_2$ will denote 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:

- Estimation of $\widehat{b_i} := \widehat{b_i}(\theta, \Delta)$ for each subject i by minimization of the **inner criterion** g_i, a modified version of the log joint-likelihood function of the data and the random effects.
- Estimation of (θ, Δ, σ) via the maximization of an **outer criterion** defined as an approximation of the profiled joint distribution of $(\theta, \Delta, \sigma, b)$ with respect to b and denoted $G(\theta, \Delta, \sigma)$.

2.1 Inner criteria

In this section, we describe the procedure used to estimate the $q-$dimensional random effects $\{b_i^*\}_{i\in[\![1,\, n]\!]}$ for a given (θ, Δ, σ) value. A straightforward approach would be to look for the minimum of the log joint-likelihood function of the data and $\left\{b_i, x_{0,i}^u\right\}$. However, we want to:

1. avoid estimation of unknown initial conditions,
2. allow for each subject an acceptable departure from the assumed model at the population level to take into account possible model misspecifications.

To solve the first point, we define our estimator as the maximizer of the joint conditional likelihood $\mathbb{P}(\mathbf{y_i}, b_i \mid x_{0,i}^u, \theta, \Delta, \sigma)$ profiled on the unknown initial condition. Since

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

by using $\mathbb{P}(\mathbf{y_i} \mid b_i, \theta, \Delta, \sigma) = \prod_j \mathbb{P}(y_{ij} \mid b_i, \theta, \Delta, \sigma) = \prod_j (2\pi)^{-d^o/2}\, \sigma^{-d^o}\, e^{-0.5\left\|CX_{\theta,b_i,x_{0,i}}(t_{ij})-y_{ij}\right\|_2^2/\sigma^2}$, $\mathbb{P}(b_i \mid \theta, \Delta, \sigma) = (2\pi)^{-q/2}\, |\Psi|^{-1/2}\, e^{-0.5 b_i^T \Psi^{-1} b_i}$ and $\sigma^{2q}\, |\Psi|^{-1} = |\triangle|^2$, a straightforward mixed-effect estimator would be $\widehat{b_i} = \arg\min_{b_i} \min_{x_{0,i}^u} \left\{\sum_j \left\|CX_{\theta,b_i,x_{0,i}}(t_{ij}) - y_{ij}\right\|_2^2 + \|\Delta b_i\|_2^2\right\}$ that is, the classic maximum likelihood criteria profiled on $x_{0,i}^u$. Concerning the second point, we allow perturbations comparing to the original model, by assuming that the dynamic of each subject i follows a perturbed version of ODE (1.1):

$$\begin{cases} \dot{x}_i(t) = f_{\theta,b_i}(t, x_i(t), z_i(t)) + Bu_i(t) \\ x_i(0) = x_{i,0} \end{cases} \tag{2.1}$$

with the addition of the forcing term $t \mapsto Bu_i(t)$ with B a $d \times d_u$ matrix and u_i a function in $L^2\left([0,T], \mathbb{R}^{d_u}\right)$. We denote $X_{\theta,b_i,x_{i,0},u_i}$ the solution of this new ODE (2.1). However, to ensure the possible perturbation remains 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 $\min_{u_i} \mathcal{C}_i(b_i, x_{i,0}, u_i \mid \theta, U)$ where $\mathcal{C}_i(b_i, x_{i,0}, u_i \mid \theta, U) = \sum_j \left\|CX_{\theta,b_i,x_{0,i},u_i}(t_{ij}) - y_{ij}\right\|_2^2 + \|u_i\|_{U,L^2}^2$ and $\|u_i\|_{U,L^2}^2 = \int_0^T u_i(t)^T U u_i(t) dt$ is the weighted Euclidean norm. Therefore the magnitude of the allowed perturbations is controlled by a positive definite and symmetric weighting matrix U. Finally, we obtain:

$$\widehat{b_i}\,(\theta, \Delta) \quad := \quad \arg\min_{b_i} g_i(b_i \mid \theta, \Delta, U) \tag{2.2}$$

where:

$$g_i(b_i \mid \theta, \Delta, U) = \min_{x_{0,i}^u} \left\{ \min_{u_i} \mathcal{C}_i(b_i, x_{i,0}, u_i \mid \theta, U) + \|\Delta b_i\|_2^2 \right\}.$$

This requires to solve the infinite dimensional optimization problem $\min_{u_i} \mathcal{C}_i(b_i, x_{i,0}, u_i \mid \theta, U)$ 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 to solve them [53, 2, 14]. Here we use the same method as in [13], which ensures the existence and uniqueness of the solution and provides a computationally efficient way to find it for linear ODEs. This method can be extended to non-linear ODEs through an iterative procedure where the original problem is replaced by a sequence of problems involving only linear ODEs. In addition, the methods from [13] presents the advantage of formulating $\min_{u_i} \mathcal{C}_i(b_i, x_{i,0}, u_i \mid \theta, U)$ as a quadratic form (or a sequence of quadratic forms) with respect to

$x^u_{0,i}$. Thus, the computation of $\min_{x^u_{0,i}} \{\min_{u_i} \mathcal{C}_i(b_i, x_{i,0}, u_i \mid \theta, U)\}$ does not add any computational complexity comparing to $\min_{u_i} \mathcal{C}_i(b_i, x_{i,0}, u_i \mid \theta, U)$.

The control corresponding to the solution of $\min_{x^u_{0,i}} \{\min_{u_i} \mathcal{C}_i(b_i, x_{i,0}, u_i \mid \theta, U)\}$ is named optimal control and denoted $\overline{u}_{i,\theta,b_i}$. The corresponding solution of (2.1) is denoted $\overline{X}_{\theta,b_i}$ and named optimal trajectory. 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:

$$g_i(b_i \mid \theta, \Delta, U) = \sum_j \left\|C\overline{X}_{\theta,b_i}(t_{ij}) - y_{ij}\right\|_2^2 + \left\|\overline{u}_{i,\theta,b_i}\right\|_{U,L^2}^2 + \left\|\Delta b_i\right\|_2^2. \tag{2.3}$$

To incorporate possible model errors in the estimation process, e.g. due to subject specific exogenous perturbations, $\overline{X}_{\theta,b_i}$ is now assumed to be the subject specific regression function, defined as the state-variable which needs the smallest perturbation in order to get close to the observations. The numerical procedure to derive $\overline{X}_{\theta,b_i}$ and g_i is presented in appendix B and is based on the method proposed by [13]. All it requires for the user is to specify a pseudo-linear representation of ODE (1.1), i.e a possibly state-dependent matrix $A_{\theta,b_i}(t, x_i(t), z_i(t))$ and state-independent vector $r_{\theta,b_i}(t, z_i(t))$ such that:

$$f_{\theta,b_i}(t, x_i(t), z_i(t)) = A_{\theta,b_i}(t, x_i(t), z_i(t))\, x_i(t) + r_{\theta,b_i}(t, z_i(t)). \tag{2.4}$$

This formulation is crucial for solving the optimal control problem in a computationally efficient way. Linear models already fit in this formalism with $A_{\theta,b_i}(t, z_i(t)) := A_{\theta,b_i}(t, x_i(t), z_i(t))$. For nonlinear models, the pseudo-linear representation is not unique but always exists [11] (in order to exploit this non-uniqueness as an additional degree of freedom, see [9] section 6).

Remark 2.1. The definition of the optimal control $\overline{u}_{i,\theta,b_i}$ has an interpretation in terms of Bayesian inference in an infinite dimensional space. According to [16] (theorem 3.5 and Corollary 3.10), $\overline{u}_{i,\theta,b_i}$ is a maximum a posteriori estimator where the chosen prior measure is a centered Gaussian random field with the covariance operator determined by U. This link can be fruitful to import tools coming from deterministic control theory to solve statistical problem formalized in functional spaces.

2.2 Outer criteria definition

We focus in this section on population parameter estimation. Classic approaches rely on maximum a posteriori distribution or the likelihood of the observations in which they get rid of the unknown subject specific parameters by taking the mean value of $\mathbb{P}[\theta, \Delta, \sigma, b \mid \mathbf{y}]$ or $\mathbb{P}[\mathbf{y} \mid \theta, \Delta, \sigma, b]$,

$\mathbb{E}_b\left[\mathbb{P}\left[\theta,\Delta,\sigma,b \mid \mathbf{y}\right]\right]$ or $\mathbb{E}_b\left[\mathbb{P}\left[\mathbf{y} \mid \theta,\Delta,\sigma,b\right]\right]$ respectively, as outer criteria. This generally requires the numerical approximation of integrals of possibly high dimensions (the same as b), a source of approximation and computational issues [48]. To avoid this, we consider the random effects as nuisance parameters and rely on a classic profiling approach for $(\theta^*,\triangle^*)$ estimation [43]. Instead of taking the mean, we rely on the maximal value of the joint distribution with respect to b. We consider the cost function $\max_b \mathbb{P}\left[\theta,\Delta,\sigma,b \mid \mathbf{y}\right]$ (or equivalently $\max_b \ln \mathbb{P}\left[\theta,\Delta,\sigma,b \mid \mathbf{y}\right]$). Bayes formula gives us $\mathbb{P}\left[\theta,\Delta,\sigma,b \mid \mathbf{y}\right] \propto \mathbb{P}\left[\mathbf{y} \mid \theta,\Delta,\sigma,b\right]\mathbb{P}\left[\theta,\Delta,\sigma,b\right]$. Since $\mathbb{P}\left[\theta,\Delta,\sigma,b\right] = \mathbb{P}\left[b \mid \theta,\Delta,\sigma\right]\mathbb{P}\left[\theta,\Delta\right]$, we get $\mathbb{P}\left[\theta,\Delta,\sigma,b \mid \mathbf{y}\right] \propto \left(\prod_i \mathbb{P}\left[\mathbf{y_i} \mid \theta,\Delta,\sigma,b_i\right]\mathbb{P}\left[b_i \mid \theta,\Delta,\sigma\right]\right)\mathbb{P}\left[\theta,\Delta\right]$ by conditional independence of subject by subject observations and subject specific parameters. It follows that $\max_b \ln \mathbb{P}\left[\theta,\Delta,\sigma,b \mid \mathbf{y}\right] \propto \sum_i \max_{b_i}\left(\ln \mathbb{P}\left[\mathbf{y_i} \mid \theta,\Delta,\sigma,b_i\right] + \ln \mathbb{P}\left[b_i \mid \theta,\Delta,\sigma\right]\right) + \ln \mathbb{P}\left[\theta,\Delta\right]$. From now on we will use the estimate (2.2) of the previous section to construct a suitable approximation of

$$\overline{G}^{(1)}(\theta,\Delta,\sigma \mid \mathbf{y}) = \sum_i \max_{b_i}\left(\ln \mathbb{P}\left[\mathbf{y_i} \mid \theta,\Delta,\sigma,b_i\right] + \ln \mathbb{P}\left[b_i \mid \theta,\Delta,\sigma\right]\right) + \ln \mathbb{P}\left[\theta,\Delta\right]$$

as our criteria to estimate population parameters. As said in the previous section, we define the optimal trajectory $\overline{X}_{\theta,b_i}$ as the regression function for each subject. Therefore, we approximate $\mathbb{P}\left[\mathbf{y_i} \mid \theta,\Delta,\sigma,b_i\right]$ by $\widetilde{\mathbb{P}}\left[\mathbf{y_i} \mid \theta,\Delta,\sigma,b_i\right] \simeq \prod_j (2\pi)^{-d^o/2}\sigma^{-d^o}e^{-0.5\left\|C\overline{X}_{\theta,b_i}(t_{ij})-y_{ij}\right\|_2^2/\sigma^2}$. By similar computations as in the previous section, we derive $\arg\max_{b_i}\left(\ln \widetilde{\mathbb{P}}\left[\mathbf{y_i} \mid \theta,\Delta,\sigma,b_i\right] + \ln \mathbb{P}\left[b_i \mid \theta,\Delta,\sigma\right]\right) = \arg\max_{b_i}\left(\sum_j \left\|C\overline{X}_{\theta,b_i}(t_{ij}) - y_{ij}\right\|_2^2 + \left\|\Delta b_i\right\|_2^2\right)$. We regularize this estimation problem by approximating it via the addition of the Tikhonov penalization term on perturbation magnitude $\left\|\overline{u}_{i,\theta,b_i}\right\|_{U,L^2}^2$, thus $\arg\max_{b_i}\left(\ln \widetilde{\mathbb{P}}\left[\mathbf{y_i} \mid \theta,\Delta,\sigma,b_i\right] + \ln \mathbb{P}\left[b_i \mid \theta,\Delta,\sigma\right]\right) \simeq \arg\max_{b_i} g_i(b_i \mid \theta,\Delta,U) = \widehat{b_i}\left(\theta,\Delta\right)$ by using definition (2.3). Also, we use

$$\overline{G}^{(2)}\left[\theta,\Delta,\sigma \mid \mathbf{y}\right] = \sum_i \left(\ln \widetilde{\mathbb{P}}\left[\mathbf{y_i} \mid \theta,\Delta,\sigma,\widehat{b_i}\left(\theta,\Delta\right)\right] + \ln \mathbb{P}\left[\widehat{b_i}\left(\theta,\Delta\right) \mid \theta,\Delta,\sigma\right]\right) + \ln \mathbb{P}\left[\theta,\Delta\right]$$

as an approximation of $\overline{G}^{(1)}$. By replacing $\widetilde{\mathbb{P}}\left[\mathbf{y_i} \mid \theta,\Delta,\sigma,b_i\right]$ and $\mathbb{P}\left[b_i \mid \theta,\Delta,\sigma\right]$ by their values, we notice that $\arg\max_{(\theta,\Delta)}\left\{\overline{G}^{(2)}\left[\theta,\Delta,\sigma \mid \mathbf{y}\right]\right\} = \arg\max_{(\theta,\Delta)}\left\{\overline{G}^{(3)}\left[\theta,\Delta,\sigma \mid \mathbf{y}\right]\right\}$ for every $\sigma > 0$ where

$$\begin{aligned}\overline{G}^{(3)}\left[\theta,\Delta,\sigma \mid \mathbf{y}\right] &= -\tfrac{1}{2\sigma^2}\textstyle\sum_i\left(\sum_j\left\|C\overline{X}_{\theta,\widehat{b_i}(\theta,\Delta)}(t_{ij}) - y_{ij}\right\|_2^2 + \left\|\triangle\widehat{b_i}\left(\theta,\Delta\right)\right\|_2^2\right)\\ &- 0.5\left(d^o\textstyle\sum_i n_i + qn\right)\ln\left(\sigma^2\right) + 0.5n\ln\left(\left|\triangle^T\triangle\right|\right) + \ln \mathbb{P}\left[\theta,\Delta\right].\end{aligned}$$

Moreover, for each (θ, Δ), the maximizer in σ^2 of $\overline{G}^{(3)}$ has a closed form expression given by:

$$\sigma^2\left(\theta,\Delta\right) = \frac{1}{\left(d^o \sum_i n_i + qn\right)} \sum_i \left(\sum_j \left\| C\overline{X}_{\theta,\widehat{b_i}(\theta,\Delta)}(t_{ij}) - y_{ij} \right\|_2^2 + \left\| \triangle \widehat{b_i}\left(\theta,\Delta\right) \right\|_2^2 \right). \tag{2.5}$$

By using the expression of $\sigma^2\left(\theta,\Delta\right)$ given by equation (2.5), we get that $\arg\max_{(\theta,\Delta)} \max_{\sigma^2} \overline{G}^{(3)}(\theta,\Delta,\sigma \mid \mathbf{y}) = \arg\max_{(\theta,\Delta)} \left\{G\left[\theta,\Delta \mid \mathbf{y}\right]\right\}$ where:

$$G\left[\theta,\Delta \mid \mathbf{y}\right] = -0.5\left(d^o \sum_i n_i + qn\right) \ln\left(\sigma^2\left(\theta,\Delta\right)\right) + n \ln\left|\triangle\right| + \ln \mathbb{P}\left[\theta,\Delta\right].$$

Thus we can profile $\overline{G}^{(3)}$ on sigma σ^2 and define our estimator as:

$$\left(\widehat{\theta},\widehat{\Delta}\right) \quad = \quad \arg\max_{(\theta,\Delta)} \left\{G\left[\theta,\Delta \mid \mathbf{y}\right]\right\} \tag{2.6}$$

to reduce the optimization problem dimension and focus on the structural parameters. An estimator of σ^* is obtained from there by computing $\sigma^2\left(\widehat{\theta},\widehat{\Delta}\right)$ given by equation (2.5). The details of the outer criteria derivation are left in appendix A.

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

In this section, we derive an estimator of the asymptotic variance of $\left(\widehat{\theta},\widehat{\Delta}\right)$. We highlight that in practice the matrix Δ is parametrized by a vector δ of dimension q', i.e $\triangle := \triangle(\delta)$. We give here a variance estimator of $\left(\widehat{\theta},\widehat{\delta}\right)$. The variance of $\widehat{\Delta}$ can be obtained using classic delta-methods (see [61] chapter 3). First of all, we drop the vector field dependence in z and we introduce the function:

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

in order to present sufficient conditions ensuring our estimator is asymptotically normal:

1. the function $\widetilde{G}\left[\theta,\Delta(\delta)\right] = -0.5\left(d^o\mathbb{E}\left[n_1\right]+q\right)\ln\left(\frac{\lim_n \frac{1}{n}\sum_i^n \mathbb{E}\left[h(\widehat{b}(\theta,\Delta(\delta)),\theta,\Delta(\delta),y_i)\right]}{d^o\mathbb{E}[n_1]+q}\right) + \ln\left|\Delta(\delta)\right|$ has a well separated minimum $\left(\overline{\theta},\overline{\delta}\right)$ belonging to the interior of a compact $\Theta \times \Omega \subset \mathbb{R}^{d\times q'}$

2. the true initial conditions $\left\{x_{0,i}^*\right\}_{i\in[\![1,n]\!]} \in [\![1,\, n]\!]$ have finite variance and either

   (a) they are i.i.d,

   (b) for $\nu = 0$ and $\nu = 1$:

$$\lim_{n\longrightarrow\infty} \frac{1}{\left(V^{(\nu)}\right)^2}\mathbb{E}\left[\sum_{i=1}^n \left(\overline{h}^{(\nu)}(\mathbf{y_i}) - \mathbb{E}\left[\overline{h}^{(\nu)}(\mathbf{y_i})\right]\right)^2 1_{\left\{\overline{h}(\mathbf{y_i})-\mathbb{E}\left[\overline{h}(\mathbf{y_i})\right]>\varepsilon\sqrt{V^{(\nu)}}\right\}}\right]$$

where $\overline{h}^{(\nu)}(\mathbf{y_i}) = \frac{d^{(\nu)}h}{d^{(\nu)}(\theta,\delta)}(\widehat{b}_i(\overline{\theta}, \Delta(\overline{\delta})), \overline{\theta}, \Delta(\overline{\delta}), \mathbf{y_i})$ and $V^{(\nu)} = \sqrt{\sum_i Var(\overline{h}^{(\nu)}(\mathbf{y_i}))^2}$,

3. the subject specific number of observations $\{n_i\}_{i\in[\![1,n]\!]}$ are i.i.d and uniformly bounded,

4. for all possibles values (θ, b_i), the solution $X_{\theta,b_i,x^*_{0,i}}$ belongs to a compact χ of $\mathbb{R}^d$, and for all (t, θ, x), the mapping $b_i \longmapsto f_{\theta,b_i}(t,x)$ has a compact support Θ_b,

5. $(\theta, b_i, t, x) \longmapsto f_{\theta,b_i}(t,x)$ belongs to $C^1(\Theta \times \Theta_b \times [0,\,T] \times \chi, \mathbb{R}^d)$,

6. the matrices $\frac{\partial^2}{\partial^2 b_i} g_i(\widehat{b_i}\left(\overline{\theta}, \Delta(\overline{\delta})\right) \mid \overline{\theta}, \Delta(\overline{\delta}), U)$ and $\frac{\partial^2 \mathcal{C}_i}{\partial^2 x_{0,i}}(\widehat{b_i}\left(\overline{\theta}, \Delta(\overline{\delta})\right), \overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\theta},\Delta(\overline{\delta}))}(0), \overline{u}_{\overline{\theta},\widehat{b}_i(\overline{\theta},\Delta(\overline{\delta}))} \mid \overline{\theta}, U)$ are of full rank almost surely for every sequence $\mathbf{y_i}$,

7. there is a neighborhood $\Theta_{\overline{\theta}}$ of $\overline{\theta}$ such that $(\theta, b_i, t, x) \longmapsto f_{\theta,b_i}(t,x) \in C^5(\Theta_{\overline{\theta}} \times \Theta_b \times [0,\,T] \times \chi, \mathbb{R}^d)$.

Conditions 1-4 are used to derive the consistency of our estimator toward $\left(\overline{\theta}, \overline{\delta}\right)$ by following classic steps for M-estimator by proving 1/the uniform convergence of our stochastic cost function to a deterministic one, 2/the existence of a well-separated minimum for this deterministic function ([61] chapter 5). Conditions 6-7 ensures that our cost function is asymptotically smooth enough in the vicinity of $\left(\overline{\theta}, \overline{\delta}\right)$ to proceed to a Taylor expansion and transfer the regularity of the cost function to the asymptotic behavior of $\sqrt{n}(\widehat{\theta} - \overline{\theta}, \widehat{\delta} - \overline{\delta})$. Less restrictive conditions can be established under which our estimator is still asymptotically normal, in particular regarding f_{θ,b_i} regularity with respect to t. Also, we emphasize that the second assumption does not require to know the distribution of the $x^*_{0,i}$.

Theorem 3.1. *Under conditions 1-7, there is a model dependent lower bound λ such that if $\|U\|_2 > \lambda$ then the estimator $\left(\widehat{\theta}, \widehat{\delta}\right)$ is asymptotically normal and:*

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

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:*

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

The proof is left in appendix D. The practical interest of this theorem is to give an estimator of Variance-Covariance:

$$V(\widehat{\theta},\widehat{\delta}) \simeq \widehat{A}(\widehat{\theta},\widehat{\delta})^{-1}\widehat{B}(\widehat{\theta},\widehat{\delta})\left(\widehat{A}(\widehat{\theta},\widehat{\delta})^{-1}\right)^T/n.$$

In the last equation the matrices $\widehat{A}$ and $\widehat{B}$ are defined by:

$$\begin{cases} \widehat{A}(\widehat{\theta},\widehat{\delta}) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial J(\widehat{\theta},\widehat{\delta},\mathbf{y_i})}{\partial(\theta,\delta)} \\ \widehat{B}(\widehat{\theta},\widehat{\delta}) = \frac{1}{n}\sum_{i=1}^{n} J(\widehat{\theta},\widehat{\delta},\mathbf{y_i})J(\widehat{\theta},\widehat{\delta},\mathbf{y_i})^T \end{cases}$$

where 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

$$J_k(\theta,\delta,\mathbf{y_i}) = \frac{d}{d\delta_k}h(\widehat{b_i}(\theta,\Delta(\delta)),\theta,\Delta(\delta),\mathbf{y_i}) - \frac{2n}{d^o\sum_i n_i + qn}Tr\left(\triangle(\delta)^{-1}\frac{\partial\triangle(\delta)}{\partial\delta_k}\right)h(\widehat{b_i}(\theta,\Delta(\delta)),\theta,\Delta(\delta),\mathbf{y_i}).$$

Now that we have proven the existence of the variance matrix $V(\theta^*,\delta^*)$ such that $\widehat{\delta}-\delta^* \rightsquigarrow N\left(0,V(\theta^*,\delta^*)\right)$, 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 $\triangle(\delta) = \begin{pmatrix} e^{\delta_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & e^{\delta_q}\end{pmatrix}$ used in section 4, 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 3.2. The previous theorem 3.1 states that we retrieve a parametric convergence rate despite a number of nuisance parameter increasing with the number of subjects. We avoid the pitfall described in [52] for profiled methods, thanks to the i.i.d structure of the nuisance parameters. This allows us to prevent bias accumulation for score functions among subjects by using the central limit theorem. Our estimator shares similarities with conditional maximum likelihood ones and our proof for asymptotic normality follows similar steps as in [3] since the $\{b_i\}_{i\in[\![1,n]\!]}$ are i.i.d.

4 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 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. For the fairness of comparison with ML where no prior is specified, we choose a non-informative one i.e. $\ln \mathbb{P}[\theta, \Delta] = 0$ for our method throughout this section. If the differential equation (1.1) has an analytical solution, the ML estimator is computed via SAEM algorithm (SAEMIX package [15]). Otherwise it is done via a restricted likelihood method dedicated to ODE models implemented in the nlmeODE package [57]. For both our method and the ML, 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 the true population parameters θ^*, Ψ^* as well as the subject parameter realizations $b_i^* \sim N(0, \Psi^*)$ with both estimation methods. We quantify the accuracy of each entry $\widehat{\psi}_p$ 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 square error $MSE(\widehat{\psi}_p) = Bias(\widehat{\psi}_p)^2 + V_{emp}(\widehat{\psi}_p)$, 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 the subject specific parameter, we estimate the mean square 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 the sparsity of the available observations.

For our method, we need to select U the matrix appearing in the inner criteria definition (2.3) balancing model and data fidelity. We use for this the forward cross-validation method presented in [19]. Let us denote $\widehat{\theta_U}$, $\left\{\widehat{b_{i,U}}\right\}_{i\in[\![1,\,n]\!]}$ the estimators obtained for a given matrix U. For each subject i, we split $[0, T]$ into $H = 2$ sub-intervals $[t_h,\, t_{h+1}]$, such that $t_1 = 0$ and $t_H = T$. We denote $X_{\theta,b_i}(.,t_h,x_h)$ the solution of $\dot{x}(t) = f_{\theta,b_i}\left(t, x_i(t), z_i(t)\right)$ defined on the interval $[t_h,\, t_{h+1}]$ with initial condition $X_{\theta,b_i}(t_h,t_h,x_h) = x_h$.The forward cross-validation uses the causal relation imposed to the data by the ODE to quantify the prediction error:

$$\mathrm{EP}(i,U)=\sum_{h=1}^{H}\sum_{\{t_{ij}\in[t_h,\,t_{h+1}]\}}\left\|y_{ij}-CX_{\widehat{\theta_U},\widehat{b_{i,U}}}(t_{ij},t_h,\overline{X}_{\widehat{\theta_U},\widehat{b_{i,U}}}(t_h))\right\|_2^2.$$

The rationale of this selection method is the following: if U is too small, $C\overline{X}_{\widehat{\theta_U},\widehat{b_{i,U}}}(t_h)$ will be close to y_h but not to the actual ODE solution, and $t\longmapsto CX_{\widehat{\theta_U},\widehat{b_{i,U}}}(t,t_h,\overline{X}_{\widehat{\theta_U},\widehat{b_{i,U}}}(t_h))$ will diverge from the observations on $[t_h,\,t_{h+1}]$. If U is too large, $\overline{X}_{\widehat{\theta_U},\widehat{b_{i,U}}}(t_h)$ will be close to the ODE solution but far from y_h and it will lead to a large value for $\mathrm{EP}(i,U)$. Thus, a proper value for U which minimizes $\mathrm{EP}(i,U)$ will be chosen between these two extreme cases. The global prediction error for the whole population is computed with $\mathrm{EP}(U)=\sum_i \mathrm{EP}(i,U)$. We retain the matrix U which minimizes EP among a trial of tested values and we denote $\widehat{\theta},\widehat{\Psi},\left\{\widehat{b_{i,}}\right\}_{i\in[\![1,\,n]\!]}$ the corresponding estimator. In the following, we use the superscript ML to denote the ML estimator.

For solving the optimization problems required for computing our inner and outer criteria, we use the Nelder-Mead algorithm implemented in the optimr package [44]. All optimization algorithms used by the estimation methods require a starting guess value. We start from the true parameter value for each of them. By doing so, we aim to do not mix two distinct problems: 1)the numerical stability of the estimation procedures, 2)the intrinsic accuracy of the different estimators. Obviously these two problems are correlated, butwe 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 vicinity of $(\theta^*,\triangle^*)$ by testing different starting points for all methods. No problem appears for our method and SAEMIX. A negligible number of non convergence cases appear for nlmODE which have been discarded thanks to the convergence criteria embedded in the package.

4.1 Partially observed linear model

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

$$\begin{cases}\dot{X}_{1,i}=\phi_{2,i}X_{2,i}-\phi_{1,i}X_{1,i}\\ \dot{X}_{2,i}=-\phi_{2,i}X_{2,i}\\ (X_{1,i}(0),X_{2,i}(0))=(x_{1,0},x_{2,0,i})\end{cases}\tag{4.1}$$

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. Regarding the true initial

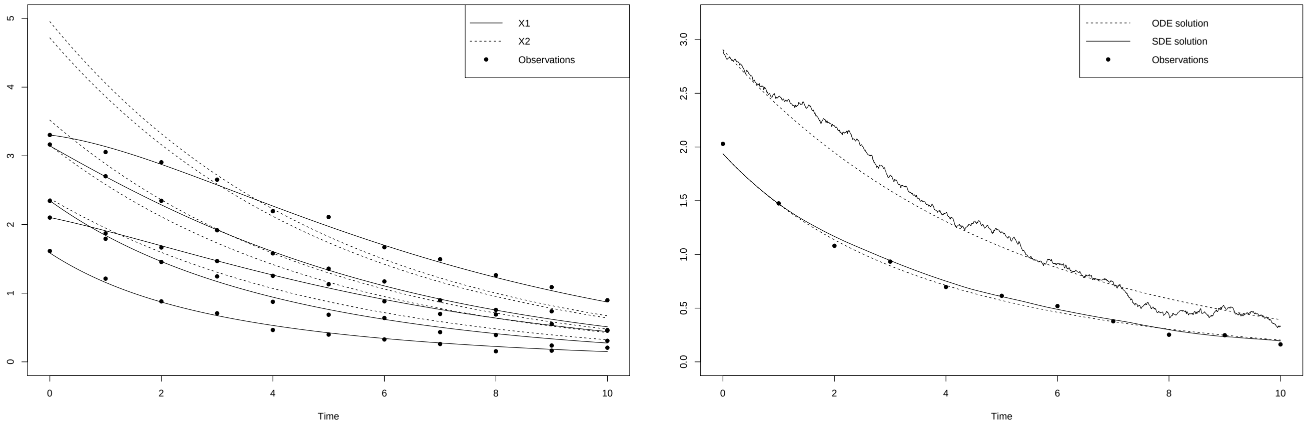

Fig. 4.1: Left: Examples of (4.1) solutions and corresponding observations (4.1). Right: Solution of (4.1) and a realization of (4.2) for the same parameter values.

conditions, they are subject specifics and normally distributed with $x^*_{1,0,i} \sim N(2,\, 0.5)$ and $x^*_{2,0,i} \sim N(3,\, 1)$. ODE (4.1) has an analytic solution given by $X_{1,i}(t) = e^{-\phi_{1,i}t}(x_{1,0} + \frac{x_{2,0}\phi_{2,i}}{\phi_{1,i}-\phi_{2,i}}(e^{(\phi_{1,i}-\phi_{2,i})t} - 1))$ for its first component which will be used for parameter estimation with the SAEMIX package. We generate $n_i = 11$ observations per subject on $[0,\, T] = [0,\, 10]$ with Gaussian measurement noise of standard deviation $\sigma = 0.05$. An example of observations and corresponding solution is plotted in figure 4.1.

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 nuisance subject-specific parameter that should be estimated and for which no observations are available. For this, we compute $\widehat{\theta}^{ML}_{x_0}$, $\widehat{\theta}^{ML}_{x_{0,2}}$ 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.

4.1.1 Correct model case

We present the estimation results in table 4.1. 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 in accordance with theoretical results. However, there is a significant drop in accuracy when $x^*_{2,0,i}$ has to be estimated, especially for θ_2. 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. 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 the unobserved initial conditions. Estimation of individual random effects is also more accurate with our method, with a decrease of more than 90% of MSE for b_i

		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	5e-3	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	5e-3	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**

Tab. 4.1: Results of estimation for model (4.1). 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.

comparing to ML.

4.1.2 Estimation in presence of model error at the subject level

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

$$\begin{cases} dX_{1,i} = \phi_{2,i} X_{2,i} dt - \phi_{1,i} X_{1,i} dt \\ dX_{2,i} = -\phi_{2,i} X_{2,i} dt + \alpha dB_t \\ (X_{1,i}(0), X_{2,i}(0)) = (x_{1,0}, x_{2,0,i}) \end{cases} \tag{4.2}$$

with B_t a Wiener process and $\alpha = 0.1$ the diffusion coefficient. For the sake of comparison, a solution of (4.1) and a realization of its perturbed counterpart given by (4.2) are plotted in figure 4.1. 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 4.1 it is easy to see 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 (4.1) 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 error presence. Results are presented in table 4.1. 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 SAEM is more pronounced than for our method which manages to limit the damages done. This confirms the benefits of taking into account model uncertainty for the regularization of the inverse problem, in particular 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.

4.2 Partially observed nonlinear model

We consider a simplified version of the model used in [57] for the analysis of glucose and insulin regulation:

$$\begin{cases} \dot{G}_i = S_G(G_B - G_i) - X_i G_i \\ \dot{I}_i = \gamma t(G_i - h) - n_i(I_i - I_B) \\ \dot{X}_i = -p_2(X_i + S_I(I_i - I_B)). \end{cases} \tag{4.3}$$

We are in a partially observed framework where only the glucose (G_i) and insulin (I_i) concentration are measured. The values of parameters $(p_2, \gamma, h, G_B, I_B)$ are fixed to $(-4.93, -6.85, 4.14, 100, 100)$ and we aim to estimate $\theta = (\theta_{S_G}, \theta_{S_I}, \theta_n)$, linked to the original model via the parametrization: $\log(S_G) = \theta_{S_G}$, $\log(S_I) = \theta_{S_I}$ and $\log(n_i) = \theta_n + b_i$ where $b_i \sim N(0, \Psi)$. The true population parameter values are $\theta^* = (-3.89,\ -7.09,\ -1.81)$ and $\Psi^* = 0.26^2$. Regarding the true initial conditions $x^*_{i,0} = \left(G^*_{0,i}, I^*_{0,i}, X^*_{0,i}\right)$ their values are all subject specific and 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 $n_i = 5$ 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 estimators accuracy. We are particularly interested by 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 the maximum likelihood based method. For this, we distinguish two cases, 1)when θ_{S_I} is known, 2)when θ_{S_I} has to be estimated and we respectively denote $\widehat{\theta_{S_i}}$ and $\widehat{\theta}$ the corresponding estimators. Finally, since the model is nonlinear we have to specify a pseudo-linear representation of the vector field as in (2.4), we take:

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

4.2.1 Correct model case

We present the estimation results in table 4.2. Our method obtains smaller MSE than ML and escapes the drop in coverage rate of the confidence interval in the case of $\theta^*_{S_I}$ estimation. The difference between the two estimators behavior 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 7-dimensional one. Thus, the topology of the parameter

		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}	$\widehat{\theta}^{MS}_{S_i}$	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}	$\widehat{\theta}^{MS}_{S_i}$	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_n	$\widehat{\theta}^{MS}_{S_i}$	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	$\widehat{\theta}^{MS}_{S_i}$	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**

Tab. 4.2: Results of estimation for model (4.3). 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.

spaces explored by each method to look for the minimum are very different.

4.2.2 Estimation in presence of model error at the subject level

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

$$\begin{cases} dG_i = \left(S_G(G_B - G_i) - X_i G_i\right) dt + \alpha_1 dB_{1,t} \\ dI_i = (\gamma t(G_i - h) - n_i(I_i - I_B))dt + \alpha_2 dB_{2,t} \\ dX_i = \left(-p_2(X_i + S_I(I_i - I_B))\right) dt + \alpha_3 dB_{3,t} \end{cases} \tag{4.4}$$

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 4.2. 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 leading to under-estimation of variance and too narrow confidence intervals.

4.3 Antibody concentration evolution model

We consider the model presented in [46] to analyze the antibody concentration, denoted Ab_i, generated by two populations of antibody secreting cells: the short lived, denoted S_i, and the

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

Tab. 4.3: Biological interpretation and parameter values

long-lived, denoted L_i:

$$\begin{cases} \dot{S}_i = -\delta_S S_i \\ \dot{L}_i = -\delta_L L_i \\ \dot{Ab}_i = \vartheta_{S,i} S_i + \vartheta_{L,i} L_i - \delta_{Ab,i} Ab_i \\ (S_i(0), L_i(0), Ab_i(0)) = (S_{0,i}, L_{0,i}, Ab_{0,i})\,. \end{cases} \tag{4.5}$$

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:

$$\dot{Ab}_i = \phi_{S,i} e^{-\delta_S t} + \phi_{L,i} e^{-\delta_L t} - \delta_{Ab,i} Ab_i \tag{4.6}$$

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 which will be used for maximum likelihood estimation with SAEMIX. We consider the following parametrization:

$$\begin{cases} \log(\delta_S) = \theta_{\delta_S} \\ \log(\phi_{S,i}) = \theta_{\phi_S} + b_{\phi_S,i} \\ \log(\phi_{L,i}) = \theta_{\phi_L} + b_{\phi_L,i} \\ \log(\delta_{Ab,i}) = \theta_{\delta_{Ab}} + b_{\delta_{Ab},i}. \end{cases}$$

The true parameter values are presented in table 4.3. According to [46], the parameter δ_L was non-identifiable 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 by 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 ODE solution (see [5]). Thus, they are likely to face practical identifiability problems.

		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}	$\widehat{\theta}^{ML}_{\delta_S}$	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}	$\widehat{\theta}^{ML}_{\delta_S}$	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}	$\widehat{\theta}^{ML}_{\delta_S}$	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**	
$\theta_{\delta_{Ab}}$	$\widehat{\theta}^{ML}_{\delta_S}$	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}	$\widehat{\theta}^{ML}_{\delta_S}$	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}	$\widehat{\theta}^{ML}_{\delta_S}$	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**
$\Psi_{\delta_{Ab}}$	$\widehat{\theta}^{ML}_{\delta_S}$	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**

Tab. 4.4: Results of estimation for model (4.6). 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.

To investigate this effect, we estimate the parameters when 1) $\theta^*_{\delta_S}$ is known (the corresponding estimators will be denoted with the subscript θ_{δ_S}), 2) it has to be estimated as well.

4.3.1 Correct model case

We generate $n_i = 11$ observations on the interval $[0, T] = [0, 364]$ with Gaussian measurement noise of standard deviation $\sigma^* = 100$. For each subject i, the initial condition has been generated according to $Ab^*_{0,i} \sim N(\overline{Ab_0}, \sigma^2_{\overline{Ab_0}})$ with $\overline{Ab_0} = 500$ and $\sigma_{\overline{Ab_0}} = 260$ to reflect the dispersion observed in data presented in [46]. We present the estimation results in table 4.4. Our method improves the estimation of $\theta^*_{\delta_S}$ facing practical identifiability problems comparing to the ML. In particular our method reduces the variance. As advocated in the introduction, our approach provides an improved estimate for the $\{b^*_i\}_{i\in[\![1,\,n]\!]}$. We assume that is due to the committed estimation error for θ^*, which causes model error for $\{b^*_i\}_{i\in[\![1,\,n]\!]}$ estimation, which is not taken into account by exact methods. This in turn explains why their 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 into account modeling uncertainties when subject specific parameters are critical for the practitioner.

4.3.2 Estimation in presence of model error at the subject level

The data are now generated with a stochastic perturbed version of the original model:

$$dAb_i = \left(\phi_{S,i}e^{-\delta_S t} + \phi_{L,i}e^{-\delta_L t} - \delta_{Ab,i}Ab_i\right) dt + \alpha dB_t \tag{4.7}$$

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 (4.6) is still a relevant approximation for estimation purpose. We keep the same parameter values and measurement noise level as in the previous section. The results are presented in table 4.4. Our method still outperforms the maximum likelihood for $\theta^*_{\delta_S}$ as well as the $\{b^*_i\}_{i\in[1,\, n]}$ estimation and their variances. In addition, we mitigate the effect of model error on estimation accuracy.

5 Real data analysis

We now proceed to the estimation starting from real data presented in [46] from which the parameter values given in table 4.3 come from. In [46] , the estimation is made from cohorts coming from three phase I trials performed in African and European countries. Each subject was vaccinated with two doses, Ad26.ZEBOV (Janssen Vaccines and Prevention) and MVA-BN-Filo (Bavarian Nordic). In these cohorts, both the effect of injection order, either Ad26.ZEBOV first and MVA-BN-Filo second, or MVA-BN-Filo first and Ad26.ZEBOV second, and the delay between, 28 or 56 days, were evaluated. In this study, we focus on an east African subpopulation where Ad26.ZEBOV was injected first and then MVA-BN-Filo with a delay of 28 days between the two doses. As in [46] and the simulation section, to stay in the temporal domain of validity of the model we use measurements made seven days after the second dose injection. It leaves us with 5 measurements of antibody concentration between days 7 up to days 330 per subject. The estimation in the original work has been done using the NIMROD software [49] and log-transformed antibody concentration measurement. We now estimate the parameters with our method with the aim to compare our results with the ex-

	Pasin et al.	CI (95%)	OCA	CI (95%)
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]
$\Psi_{\delta_{Ab}}$	0.3	[0.24, 0.36]	0.25	[0.19, 0.31]

Tab. 5.1: Estimation presented in [46] (left) and via our approach (right)

isting one. We used the same prior distribution $\pi(\theta) \sim N\left(\begin{pmatrix} -1 \\ 0 \\ 0 \\ -4.1 \end{pmatrix}, \begin{pmatrix} 25 & 0 & 0 & 0 \\ 0 & 100 & 0 & 0 \\ 0 & 0 & 100 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}\right)$ for $\theta = (\theta_{\delta_S}, \theta_{\phi_S}, \theta_{\phi_L}, \theta_{\delta_{Ab}})$ as the one defined in the NIMROD software. We choose our mesh-size such that we get 200 discretization points for each subject on the observation interval and we use $U = 10$ i.e. a value lower than in the simulated data case because of the model error presence. We also proceed to the log-transformation of the data to stabilize the measurement noise variance. This drives us to use the nonlinear model:

$$\dot{\widetilde{Ab_i}}(t) = \frac{1}{\ln(10)}\left(\phi_{S,i}e^{-\delta_S t} + \phi_{L,i}e^{-\delta_L t}\right)10^{-\widetilde{Ab_i}(t)} - \frac{\delta_{Ab,i}}{\ln(10)} \tag{5.1}$$

describing the dynamic of $\widetilde{Ab_i}(t) := \log_{10} Ab_i(t)$ for parameter estimation purpose. We use $A_{\theta,b_i}(t,x,z_i(t)) = \frac{1}{\ln(10)}\left(\phi_{S,i}e^{-\delta_S t} + \phi_{L,i}e^{-\delta_L t}\right)\frac{10^{-x}}{x}$ and $r_{\theta,b_i}(t,z_i(t)) = -\frac{\delta_{Ab,i}}{\ln(10)}$ for the pseudo-linear formulation of the model. Our estimations and the ones from the original paper [46] are presented in Table 5.1 for the sake of comparison. In the following, we denote $\left(\widehat{\theta}^P, \widehat{b_i}^P\right)$ (respectively $\left(\widehat{\theta}, \widehat{b_i}\right)$) the estimation obtained by [46] (respectively our approach) for the mean population parameter and subject specific ones. Both methods produce estimations with overlapping confidence intervals for θ. Still, significant differences appear for $(\Psi_{\phi_S}, \Psi_{\phi_L}, \Psi_{\delta_{Ab}})$ estimation which quantifies the dispersion of random effects. This is explained by the fact that we only consider a subset of the subjects used in [46] for estimation. This has an effect on the observed diversity within the cohort of patients and thus on $(\Psi_{\phi_S}, \Psi_{\phi_L}, \Psi_{\delta_{Ab}})$ estimation. Regarding the predictions, we present in figure 5.1 examples of estimated trajectories. The confidence intervals are computed via Monte-Carlo sampling from the approximated normal laws $\mathcal{N}(\widehat{\theta}, V(\widehat{\theta}))$ and $\mathcal{N}(\widehat{\theta}^P, V(\widehat{\theta}^P))$ to quantify the effect of estimation uncertainy on θ on the predicted trajectories. For NIMROD estimation, for a given sampled value $\widetilde{\theta}^P \sim \mathcal{N}(\widehat{\theta}^P, V(\widehat{\theta}^P))$ and subject i, the sampled regression function $X_{\widetilde{\theta}^P, \widehat{b_i}^P, y_{0,i}}$ is obtained by solving ODE (5.1) for parameter values $(\theta, b_i, x_{0,i}) = \left(\widetilde{\theta}^P, \widehat{b_i}^P, y_{0,i}\right)$.

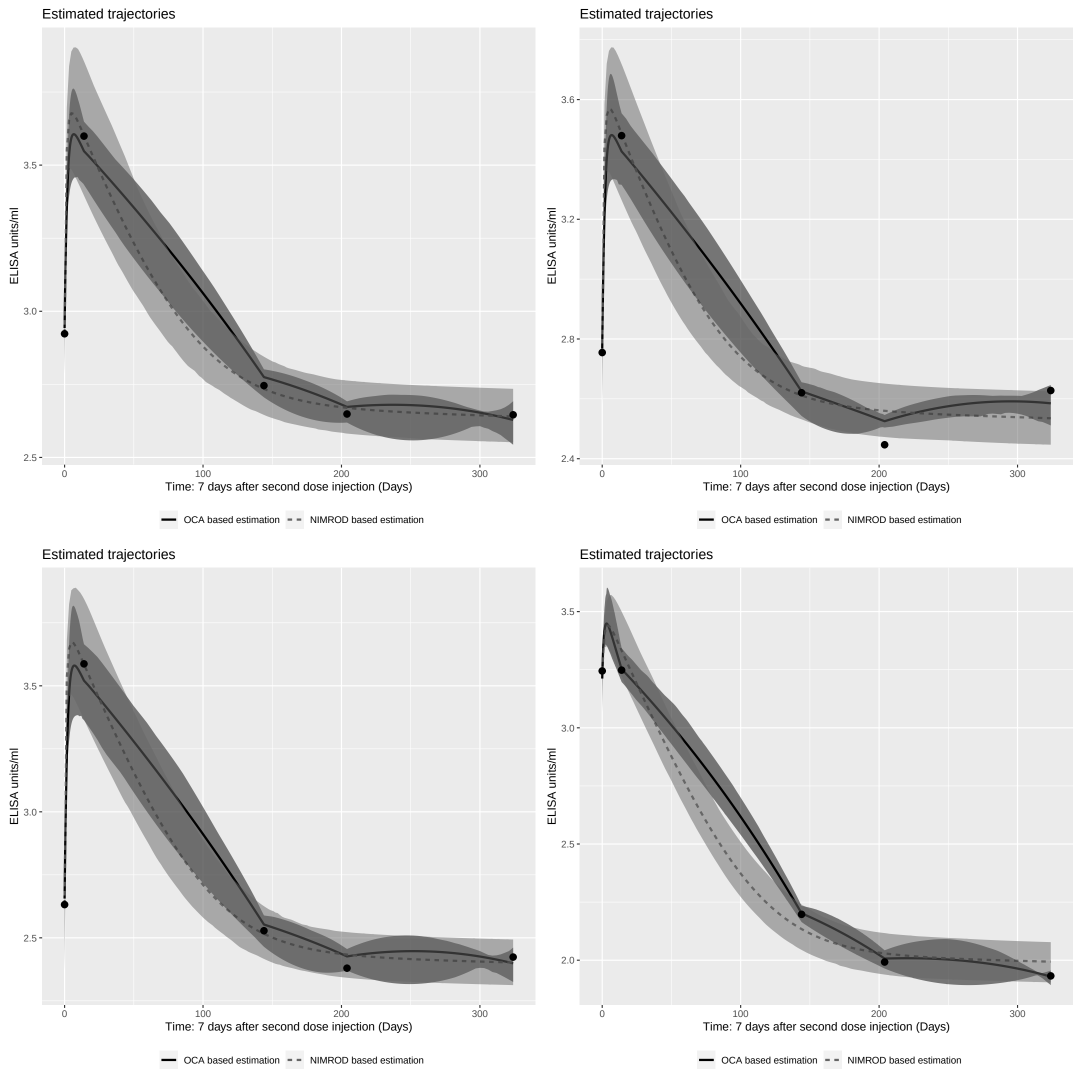

Fig. 5.1: Examples of fitted trajectories for both methods for different subjects. Here Time=0 is the 7th day post-second dose. Dashed lines: fitted ODE solutions (5.1) with $\left(\widehat{\theta}^P, \widehat{b_i}^P\right)$. Solid line: optimal trajectories $\overline{X}_{\widehat{\theta},\widehat{b_i}}$. Shaded area are the 95% confidence intervals.

Regarding our approach we recall the regression functions are now defined as optimal trajectories. So, for $\widetilde{\theta} \sim \mathcal{N}(\widehat{\theta}, V(\widehat{\theta}))$ the sampled regression function for subject i is the optimal trajectory $\overline{X}_{\widetilde{\theta},\widehat{b_i}}$ obtained via the minimization of the cost function $\mathcal{C}_i(\widehat{b_i}, x_{i,0}, u_i \mid \widetilde{\theta}, U)$. This explain the differences between the two confidence intervals in terms of shape and width. Our method gives narrower intervals because for each sampled value an optimal control problem is solved to obtain the related optimal trajectory. This imposes a common goal of data fidelity to each sampled $\overline{X}_{\widetilde{\theta},\widehat{b_i}}$ which limits their inter-variability. Still, despite these differences in shapes, both prediction intervals cover the same points. Morever, on the long-term our intervals are nearly always contained in the ones given by NIMROD.

Our estimation of θ supports the parameter inference obtained in [46] via another method and the subsequent analysis made on the antibody concentration dynamics. In addition to this parametric comparison, we want to assess the model adequacy via the temporal evolution analysis of the optimal controls $\overline{u}_{i,\widehat{\theta},b_i(\widehat{\theta})}$ estimated as byproducts of our method. Indeed, they quantify the exogenous perturbations u_i we need to add to model (5.1) so that the solution of its perturbed counterpart,

$$\dot{\widetilde{Ab_{i,u}}}(t) = \frac{1}{\ln(10)} \left(\phi_{S,i} e^{-\delta_S t} + \phi_{L,i} e^{-\delta_L t}\right) 10^{-\widetilde{Ab_{i,u}}(t)} - \frac{\delta_{Ab,i}}{\ln(10)} + u_i \tag{5.2}$$

reproduce the observations. This approach is similar to the one developed in [24] where control theory replaces non-parametric procedures to estimate u_i. Still, their approach relies on a finite basis approximation of $\widetilde{Ab_{i,u}}$ which requires to specify a basis function family, its dimension as well as a penalization parameter similar to U. At the contrary, our method avoids this complex step of hyper-parameter selection and only needs U. For comparison, we also quantify the committed model error for $\left(\widehat{\theta}^P, \widehat{b_i}^P\right)$. To do so we compute $\overline{u}_i^P$, the solution of the optimal control problem: $\overline{u}_i^P = \arg\min_{u_i} \left\{\sum_j \left\|\widetilde{Ab}_{i,\widehat{\theta}^P,\widehat{b_i}^P,y_{i0},u_i}(t_{ij}) - y_{ij}\right\|_2^2 + \|u_i\|_{U,L^2}^2\right\}$. In the last expression $\widetilde{Ab}_{i,\widehat{\theta}^P,\widehat{b_i}^P,y_{i0},u_i}$ is the solution of the perturbed ODE (5.2) for $(\theta, b_i) = \left(\widehat{\theta}^P, \widehat{b_i}^P\right)$ and y_{i0} is the measured concentration at $t = 0$ used a surrogate value for the initial condition (as they did in [46]). We still use $U = 10$ for this optimal control problem to allow for the same level of perturbation magnitude for both methods. In figure (5.2), we plot $\overline{u}_{i,\widehat{\theta},b_i(\widehat{\theta})}$ and $\overline{u}_i^P$ as well as their mean values and confidence intervals. Our method leads to residual perturbations of smaller magnitudes and narrower confidence intervals. This means our approach produces an estimation which minimizes the committed model error for each subject comparing to a method based only on a data fitting criteria. This is particularly clear at the beginning of the observation interval when the influence of the initial conditions is the highest. In this case our narrower confidence interval clearly excludes

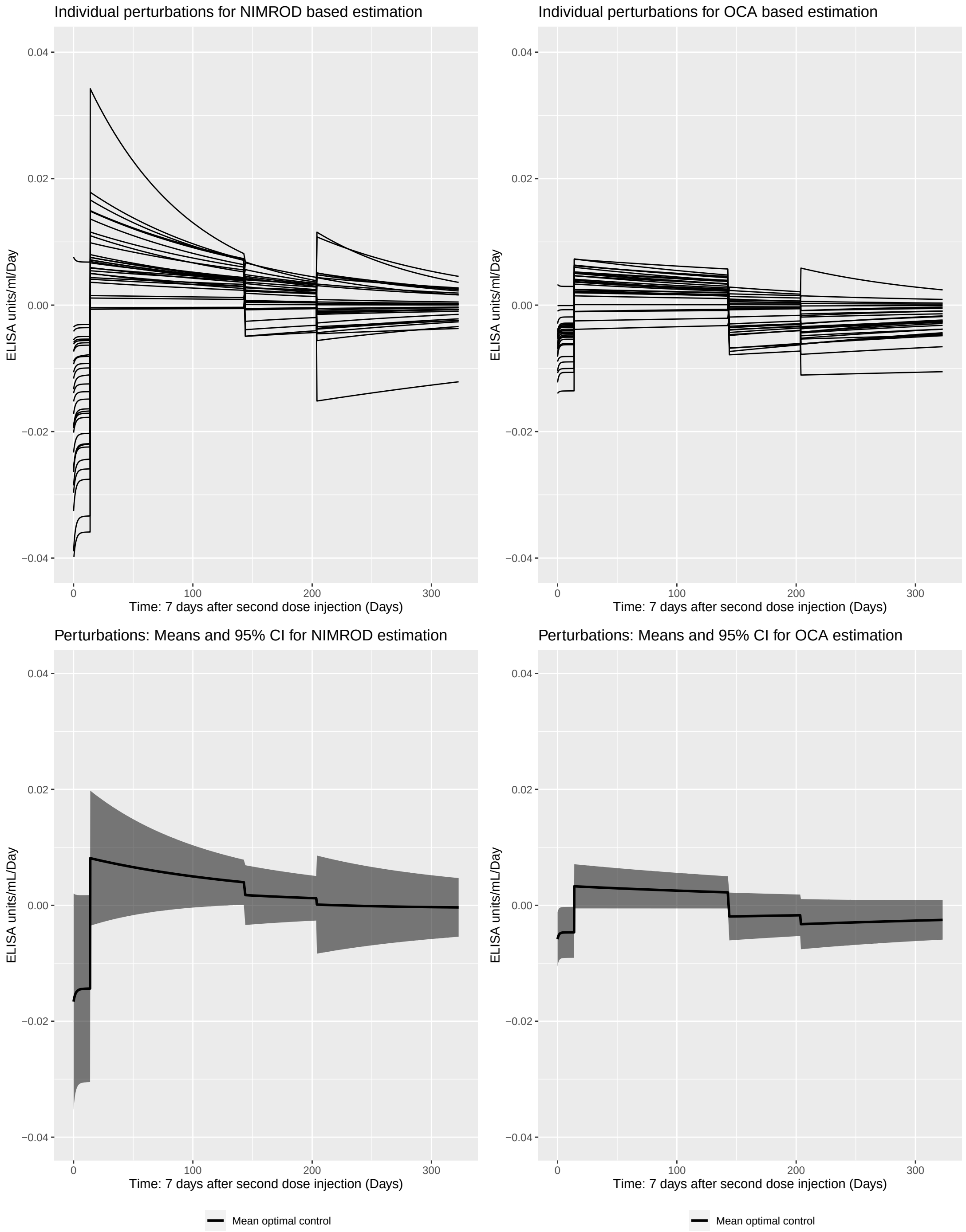

Fig. 5.2: 1) Up: Estimated residual controls for each subject, 2) bottom: mean optimal control and 95% confidence interval for the optimal controls a) left: $\overline{u}_i^P$ obtained from parameter estimation in [46], b) right: $\overline{u}_{i,\widehat{\theta},b_i(\widehat{\theta})}$ obtained from our estimation.

a null perturbation and advocates for an over-estimation of the predicted antibody concentration by the model. This makes sense because model (4.5) assumes that both populations of antibody secreting cells decrease with time, and that is probably not completely true at the beginning of the dynamic. Thus, despite similar results regarding parameter values between our estimation and [46], the insight given by our method at the dynamic scale leads us to the additional conclusion of model misspecification presence at the beginning of the observation interval.

6 Conclusion

In this work, we propose an estimation method addressing issues encountered by classic approaches for the problem of parameter estimation in NLME-ODEs. We identify three potential sources of problems for exact methods such as likelihood based inference: their difficulties in presence of model error, their need to estimate initial conditions and their dramatic performance degradation when facing poorly identifiable parameters. We propose here a method based on control theory accounting for the presence of potential model uncertainty at the subject level and which can be easily profiled on the initial conditions. Simulations with both presence and absence of model errors illustrate the benefits of regularization techniques for estimating poorly identifiable parameters, subject specific parameters as well as their variances in NLME-ODEs. In addition, bypassing estimation of initial conditions represents a clear advantage for partially observed systems comparing to likelihood based approaches, as emphasized in simulations.

Still, this benefit in term of estimation accuracy comes with a computational price. On a server 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 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, so looking for another numerical procedure is beyond the scope of this paper.

An under-exploited feature of the method so far is the obtained optimal controls. The qualitative based analysis exposed in section 5 can be made more rigorous. 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 [12, 13]). This is a subject for further work.

Software

Our estimation method is implemented in R and a code reproducing the examples of Section 4 is available on a GitHub repository located here.

Acknowledgement

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. 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.

References

[1] F.B. Agusto and A.I Adekunle. Optimal control of a two-strain tuberculosis-hiv/aidsco-infection model. *BioSystems*, 119:20–44, 2014.

[2] M.D.S. Aliyu. *Nonlinear H-Infinity Control, Hamiltonian Systems and Hamilton-Jacobi Equations*. CRC Press, 2011.

[3] E.B. Andersen. Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society*, 32:283–301, 1970.

[4] M. Andraud, O. Lejeune, J.Z. Musoro, B. Ogunjimi, P. Beutels, and N. Hens. Living on three time scales: the dynamics of plasma cell and antibody populations illustrated for hepatitis a virus. *Plos Computational Biology*, 63, 2012.

[5] Irene Balelli, Chloé Pasin, Mélanie Prague, Fabien Crauste, Thierry Van Effelterre, Viki Bockstal, Laura Solforosi, and Rodolphe Thiébaut. A model for establishment, maintenance and reactivation of the immune response after vaccination against ebola virus. *Journal of Theoretical Biology*, page 110254, 2020.

[6] C. G. Bowsher and P.S. Swain. Identifying source of variation and the flow of information in biochemical networks. *PNAS*, 109:1320–1328, 2012.

[7] J. Brynjarsdottir and A. O'Hagan. Learning about physical parameters: The importance of model discrepancy. *Inverse Problems*, 30:24, 2014.

[8] D.A Campbell. *Bayesian Collocation Tempering and Generalized Profiling for Estimation of Parameters from Differential Equation Models.* PhD thesis, McGill University Montreal,Quebec, 2007.

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

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

[11] T. Cimen and S.P. Banks. Nonlinear optimal tracking control with application to super-tankers for autopilot design. *Automatica*, 40:1845–1863, 2004.

[12] Q. Clairon and N. J-B. Brunel. Optimal control and additive perturbations help in estimating ill-posed and uncertain dynamical systems. *Journal of the American Statistical Association*, 113:1195–1209, 2018.

[13] Quentin Clairon. A regularization method for the parameter estimation problem in ordinary differential equations via discrete optimal control theory. *Journal of Statistical Planning and Inference*, 2020.

[14] Francis Clarke. *Functional Analysis, Calculus of Variations and Optimal Control.* Graduate Texts in Mathematics. Springer-Verlag London, 2013.

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

[16] M. Dashti, K. J H. Law, A.M Stuart, and J. Voss. Map estimators and their consistency in bayesian nonparametric inverse problems. *Inverse Problems*, 29, 2013.

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

[18] H.W. Engl, C. Flamm, P. Kügler, J. Lu, S. Müller, and P. Schuster. Inverse problems in systems biology. *Inverse Problems*, 25(12), 2009.

[19] L. De Vargas Roditi G. Hooker, S. P. Ellner and D. J. D. Earn. Parameterizing state-space models for infectious disease dynamics by generalized profiling: measles in ontario. *Journal of the Royal Society*, 8:961–974, 2011.

[20] D.T. Gillespie. The chemical langevin equation. *Journal of Chemical Physics*, 113(1):297–306, 2000.

[21] J. Guedj, R. Thiebaut, and D. Commenges. Maximum likelihood estimation in dynamical models of hiv. *Biometrics*, 63:1198–206, 2007.

[22] B.Z. Guo and B. Sun. Dynamic programming approach to the numerical solution of optimal control with paradigm by a mathematical model for drug therapies. *Optimization and Engineering*, pages 1–18, 2012.

[23] R. N. Gutenkunst, J.J Waterfall, F.P. Casey, K.S. Brown, C.R. Myers, and J.P. Sethna. Universally sloppy parameter sensitivities in systems biology models. *Public Library of Science Computational Biology*, 3:e189, 2007.

[24] Giles Hooker, Stephen P Ellner, et al. Goodness of fit in nonlinear dynamics: misspecified rates or misspecified states? *The Annals of Applied Statistics*, 9(2):754–776, 2015.

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

[26] Y. Huang, D. Liu, and H. Wu. Hierachical bayesian methods for estimation of parameters in a longitudinal hiv dynamic system. *Biometrics*, 62:413–423, 2006.

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

[28] Y. Huang, H. Wu, and E. P. Acosta. Hierarchical bayesian inference for hiv dynamic differential equation models incorporating multiple treatment factors. *Biom J*, 52:470–486, 2010.

[29] N.G. Van Kampen. *Stochastic Process in Physics and Chemistry*. Elsevier, 1992.

[30] Marc C Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

[31] Donald E. Kirk. *Optimal Control Theory: An Introduction*. Dover Publication, 1998.

[32] P. Kirk, D. Silk, and M. Michael. Reverse engineering under uncertainty. In *Uncertainty in Biology*, pages 15–32. Springer, 2016.

[33] M. Komorowski, J. Miekisz, and M.P.H. Stumpf. Decomposing noise in biochemical signaling systems highlights the role of protein degradation. *Biophysical journal*, 10:1783–1793, 2013.

[34] E. Kuhn and M. Lavielle. Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics and Data Analysis*, 49:1020–1038, 2005.

[35] T.G Kurtz. Strong approximation theorems for density dependent markov chains. *Stochastic Processes and their Applications*, 6:223–240, 1978.

[36] M. Lavielle and L. Aarons. What do we mean by identifiability in mixed effects models? *Journal of pharmacokinetics and pharmacodynamics*, 2015.

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

[38] D. Le, J.D. Miller, and V.V Ganusov. Mathematical modeling provides kinetic details of the human immune response to vaccination. *Frontiers in Cellular and Infection Microbiology*, 4:177, 2015.

[39] T. O Leary, A.C. Sutton, and E. Marder. Computational models in the age of large datasets. *Current Opinion in Neurobiology*, 32:87–94, 2015.

[40] M. J. Lindstrom and D. M. Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46:673–687, 1990.

[41] D.J. Lunn, A.Thomas, N. Best, and D. Spiegelhalter. Winbugs - a bayesian modelling framework: Concepts, structure and extensibility. *Statistics and Computing*, 10:325–337, 2000.

[42] A.K Fermin M. Lavielle, A. Samson and F. Mentre. Maximum likelihood estimation of long terms hiv dynamic models and antiviral response. *Biometrics*, 67:250–259, 2011.

[43] S.A. Murphy and A.W. Van der Vaart. On profile likelihood. *Journal of American Statistical Association*, 95:449–465, 2000.

[44] John C Nash. Using and extending the optimr package. 2016.

[45] C. Pasin, F. Dufour, L. Villain, H. Zhang, and R. Thiebaut. Controlling il-7 injections in hiv-infected patients. *Bulletin of Mathematical Biology*, 80:2349–2377, 2018.

[46] Chloé Pasin, Irene Balelli, Thierry Van Effelterre, Viki Bockstal, Laura Solforosi, Mélanie Prague, Macaya Douoguih, and Rodolphe Thiébaut. Dynamics of the humoral immune response to a prime-boost ebola vaccine: quantification and sources of variation. *Journal of virology*, 93(18):e00579–19, 2019.

[47] A. Perelson, A. Neumann, M. Markowitz, J. Leonard, and D. Ho. Hiv-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science*, 271:1582–1586, 1996.

[48] J.C. Pinheiro and D. M. Bates. Approximations to the loglikelihood function in the nonlinear mixed effects model. *Journal of the Computational and Graphical Statistics*, 4:12–35, 1994.

[49] M. Prague, D. Commengues, J. Guedj, J. Drylewicz, and R. Thiébaut. 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, 2013.

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

[51] J.O. Ramsay, G. Hooker, J. Cao, and D. Campbell. Parameter estimation for differential equations: A generalized smoothing approach. *Journal of the Royal Statistical Society (B)*, 69:741–796, 2007.

[52] N. Sartori. Modified profile likelihood in models with stratum nuisance parameters. *Biometrika*, 90:553–549, 2003.

[53] E. Sontag. *Mathematical Control Theory: Deterministic finite-dimensional systems.* Springer-Verlag (New-York), 1998.

[54] R.R. Stein, V. Bucci, N.C. Toussaint, C.G Buffie, G. Ratsch, E.G Pamer, C. Sander, and J.B. Xavier. Ecological modeling from time-series inference: Insight into dynamics and stability of intestinal microbiota. *Public Library of Science Computational Biology*, 9:12, 2013.

[55] A.M. Stuart. Inverse problems: A bayesian perspective. *Acta Numerica*, pages 451–559, 2010.

[56] R. Thiebaut, J. Drylewicz, M. Prague, C. Lacabaratz, and S. Beq et al. Quantifying and predicting the effect of exogenous interleukin on cd4+t cells in hiv-1 infection. *Plos Computational Biology*, 10 (5), 2014.

[57] C.W. Tornoe, H. Agerso, E. N. Jonsson, H. Madsen, and H. A. Nielsen. Non-linear mixed-effects pharmacokinetic/pharmacodynamic modelling in nlme using differential equations. *Computer Methods and Programs in Biomedicine*, 76:31–41, 2004.

[58] Mark K Transtrum, Benjamin B Machta, Kevin S Brown, Bryan C Daniels, Christopher R Myers, and James P Sethna. Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *The Journal of chemical physics*, 143(1):07B201_1, 2015.

[59] M.K. Transtrum, B.B. Machta, and J.P. Sethna. Geometry of nonlinear least squares with applications to sloppy models and optimization. *Physical Review*, 83:35, 2011.

[60] R. Tuo and C.F.J. Wu. Efficient calibration for imperfect computer models. *Annals of Statistics*, 2015.

[61] A.W. van der Vaart. *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilities Mathematics. Cambridge University Press, 1998.

[62] J. M. Varah. A spline least squares method for numerical parameter estimation in differential equations. *SIAM J.sci. Stat. Comput.*, 3(1):28–46, 1982.

[63] Laura Villain, Daniel Commenges, Chloé Pasin, Mélanie Prague, and Rodolphe Thiébaut. Adaptive protocols based on predictions from a mechanistic model of the effect of il7 on cd4 counts. *Statistics in medicine*, 38(2):221–235, 2019.

[64] J. Wakefield and A. Racine-Poon. An application of bayesian population pharmacokinetic/pharmacodynamic models to dose recommendation. *Statistics in Medicine*, 14:971–986, 1995.

[65] L. Wang, J. Cao, J.O. Ramsay, D.M Burger, C.J.L Laporte, and J.K Rockstroh. Estimating mixed-effects differential equation models. *Statistics and Computing*, 24:111–121, 2014.

[66] H. Wu, T. Lu, H. Xue, and H. Liang. Sparse additive odes for dynamic gene regulatory network modeling. *Journal of the American Statistical Association*, 109:700–716, 2014.

[67] S. Zhang and X. Xu. Dynamic analysis and optimal control for a model of hepatitis c with treatment. *Communications in Nonlinear Science and Numerical Simulation*, 46:14–25, 2016.

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

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

Appendix A: Outer criteria derivation

We recall the definition of $\overline{G}^{(2)}$:

$$\overline{G}^{(2)}\left[\theta,\Delta,\sigma \mid \mathbf{y}\right] = \sum_i \left(\ln \widetilde{\mathbb{P}}\left[\mathbf{y_i} \mid \theta,\Delta,\sigma,\widehat{b_i}\left(\theta,\Delta\right)\right] + \ln \mathbb{P}\left[\widehat{b_i}\left(\theta,\Delta\right) \mid \theta,\Delta,\sigma\right]\right) + \ln \mathbb{P}\left[\theta,\Delta\right].$$

By using $\widetilde{\mathbb{P}}\left[\mathbf{y_i} \mid \theta,\Delta,\sigma,\widehat{b_i}\left(\theta,\Delta\right)\right] = \prod_j p(y_{ij} \mid \widehat{b_i}\left(\theta,\Delta\right),\theta,\Delta,\sigma) = \prod_j \left(2\pi\right)^{-d^o/2} \sigma^{-d^o} e^{-0.5\left\|C\overline{X}_{\theta,\widehat{b_i}(\theta,\Delta)}(t_{ij})-y_{ij}\right\|_2^2/\sigma^2}$, $\mathbb{P}\left[\widehat{b_i}\left(\theta,\Delta\right) \mid \theta,\Delta,\sigma\right] = \left(2\pi\right)^{-q/2}\left|\Psi\right|^{-1/2} e^{-0.5\widehat{b_i}(\theta,\Delta)\Psi^{-1}\widehat{b_i}(\theta,\Delta)}$, we get

$$\begin{aligned}
\ln \widetilde{\mathbb{P}}\left[\mathbf{y_i} \mid \theta,\Delta,\sigma,\widehat{b_i}\left(\theta,\Delta\right)\right] &= \textstyle\sum_j \left(-d^o \ln\sigma - d^o \ln\left(2\pi\right)/2 - 0.5/\sigma^2 \left\|C\overline{X}_{\theta,\widehat{b_i}(\theta,\Delta)}(t_{ij}) - y_{ij}\right\|_2^2\right) \\
&= -d^o n_i \left(\ln\sigma + \ln\left(2\pi\right)/2\right) - 0.5/\sigma^2 \textstyle\sum_j \left\|C\overline{X}_{\theta,\widehat{b_i}(\theta,\Delta)}(t_{ij}) - y_{ij}\right\|_2^2 \\
\ln \mathbb{P}\left[\widehat{b_i}\left(\theta,\Delta\right) \mid \theta,\Delta,\sigma\right] &= -q/2\ln\left(2\pi\right) - 0.5\ln\left|\Psi\right| - 0.5\widehat{b_i}\left(\theta,\Delta\right)^T \Psi^{-1}\widehat{b_i}\left(\theta,\Delta\right)
\end{aligned}$$

so by reinjecting in $\overline{G}^{(2)}$ we obtain:

$$\begin{aligned}
&\overline{G}^{(2)}\left[\theta,\Delta,\sigma \mid \mathbf{y}\right] \\
&= -\tfrac{0.5}{\sigma^2}\textstyle\sum_i\sum_j \left\|C\overline{X}_{\theta,\widehat{b_i}(\theta,\Delta)}(t_{ij}) - y_{ij}\right\|_2^2 \\
&-\textstyle\sum_i \left(0.5\widehat{b_i}\left(\theta,\Delta\right)^T \Psi^{-1}\widehat{b_i}\left(\theta,\Delta\right) + d^o n_i\left(\ln\sigma + \ln\left(2\pi\right)/2\right) + q/2\ln\left(2\pi\right) + 0.5\ln\left|\Psi\right|\right) + \ln\mathbb{P}\left[\theta,\Delta\right].
\end{aligned}$$

From this, we have $\arg\max_{(\theta,\Delta,\sigma)}\left\{\overline{G}^{(2)}\left[\theta,\Delta,\sigma \mid \mathbf{y}\right]\right\} = \arg\max_{(\theta,\Delta,\sigma)}\left\{\overline{G}^{(3)}\left[\theta,\Delta,\sigma \mid \mathbf{y}\right]\right\}$ with

$$\begin{aligned}
\overline{G}^{(3)}\left[\theta,\Delta,\sigma \mid \mathbf{y}\right] &= -0.5\textstyle\sum_i\left(\sum_j \left\|C\overline{X}_{\theta,\widehat{b_i}(\theta,\Delta)}(t_{ij}) - y_{ij}\right\|_2^2/\sigma^2 + \widehat{b_i}\left(\theta,\Delta\right)^T\Psi^{-1}\widehat{b_i}\left(\theta,\Delta\right)\right) \\
&- \left(d^o\textstyle\sum_i n_i\right)\ln\sigma - 0.5n\ln\left|\Psi\right| + \ln\mathbb{P}\left[\theta,\Delta\right].
\end{aligned}$$

Now let us use the relationship $\sigma^2\Psi^{-1} = \triangle^T\triangle$ (or equivalently $\Psi = \sigma^2\left(\triangle^T\triangle\right)^{-1}$), we get:

$$\begin{aligned}
&\overline{G}^{(3)}\left[\theta, \Delta, \sigma \mid \mathbf{y}\right] \\
&= -0.5\textstyle\sum_i \left(\sum_j \left\|C\overline{X}_{\theta,\widehat{b_i}(\theta,\Delta)}(t_{ij}) - y_{ij}\right\|_2^2/\sigma^2 + \left\|\triangle\widehat{b_i}\left(\theta,\Delta\right)\right\|_2^2/\sigma^2\right) \\
&- \left(d^o\textstyle\sum_i n_i\right)\ln\sigma - 0.5n\ln\left(\left|\sigma^2\left(\triangle^T\triangle\right)^{-1}\right|\right) + \ln\mathbb{P}\left[\theta,\Delta\right] \\
&= -\tfrac{0.5}{\sigma^2}\textstyle\sum_i \left(\sum_j \left\|C\overline{X}_{\theta,\widehat{b_i}(\theta,\Delta)}(t_{ij}) - y_{ij}\right\|_2^2 + \left\|\triangle\widehat{b_i}\left(\theta,\Delta\right)\right\|_2^2\right) - \left(d^o\sum_i n_i\right)\ln\sigma \\
&-0.5n\left(\ln\left(\left|\left(\triangle^T\triangle\right)^{-1}\right|\right) + \ln\left(\sigma^{2q}\right)\right) + \ln\mathbb{P}\left[\theta,\Delta\right] \\
&= -\tfrac{0.5}{\sigma^2}\textstyle\sum_i \left(\sum_j \left\|C\overline{X}_{\theta,\widehat{b_i}(\theta,\Delta)}(t_{ij}) - y_{ij}\right\|_2^2 + \left\|\triangle\widehat{b_i}\left(\theta,\Delta\right)\right\|_2^2\right) - 0.5\left(d^o\sum_i n_i + qn\right)\ln\left(\sigma^2\right) \\
&-0.5n\ln\left(\left|\left(\triangle^T\triangle\right)^{-1}\right|\right) + \ln\mathbb{P}\left[\theta,\Delta\right]
\end{aligned}$$

From this, we derive:

$$\nabla_{\sigma^2}\overline{G}^{(3)}\left[\theta,\Delta,\sigma \mid \mathbf{y}\right] = \frac{0.5}{\sigma^4}\sum_i\left(\sum_j\left\|C\overline{X}_{\theta,\widehat{b_i}(\theta,\Delta)}(t_{ij}) - y_{ij}\right\|_2^2 + \left\|\triangle\widehat{b_i}\left(\theta,\Delta\right)\right\|_2^2\right) - \frac{0.5}{\sigma^2}\left(d^o\sum_i n_i + qn\right)$$

and we have $\nabla_{\sigma^2}\overline{G}^{(3)}\left[\theta,\Delta,\sigma \mid \mathbf{y}\right] = 0$ when:

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

By re-injecting, we get:

$$\begin{aligned}
&\overline{G}^{(3)}\left[\theta,\Delta,\sigma^2\left(\theta,\Delta\right) \mid \mathbf{y}\right] \\
&= -0.5/\left(\textstyle\sum_i\left(\sum_j\left\|C\overline{X}_{\theta,\widehat{b_i}(\theta,\Delta)}(t_{ij}) - y_{ij}\right\|_2^2 + \left\|\triangle\widehat{b_i}\left(\theta,\Delta\right)\right\|_2^2\right) / \left(d^o\sum_i n_i + qn\right)\right) \\
&\times\textstyle\sum_i\left(\sum_j\left\|C\overline{X}_{\theta,\widehat{b_i}(\theta,\Delta)}(t_{ij}) - y_{ij}\right\|_2^2 + \left\|\triangle\widehat{b_i}\left(\theta,\Delta\right)\right\|_2^2\right) \\
&-0.5\left(d^o\textstyle\sum_i n_i + qn\right)\ln\left(\sigma^2\left(\theta,\Delta\right)\right) - 0.5n\ln\left(\left|\left(\triangle^T\triangle\right)^{-1}\right|\right) + \ln\mathbb{P}\left[\theta,\Delta\right] \\
&= -0.5\left(d^o\textstyle\sum_i n_i + qn\right) - 0.5\left(d^o\sum_i n_i + qn\right)\ln\left(\sigma^2\left(\theta,\Delta\right)\right) + n\ln\left|\triangle\right| + \ln\mathbb{P}\left[\theta,\Delta\right]
\end{aligned}$$

Thus $\arg\max_{(\theta,\Delta)}\left\{\overline{G}^{(3)}\left[\theta,\Delta,\sigma^2\left(\theta,\Delta\right) \mid \mathbf{y}\right]\right\} = \arg\max_{(\theta,\Delta)}\left\{G\left[\theta,\Delta \mid \mathbf{y}\right]\right\}$ with:

$$G\left[\theta,\Delta \mid \mathbf{y}\right] = -0.5\left(d^o\sum_i n_i + qn\right)\ln\left(\sigma^2\left(\theta,\Delta\right)\right) + n\ln\left|\triangle\right| + \ln\mathbb{P}\left[\theta,\Delta\right].$$

Appendix B: Numerical procedure for $\overline{u}_{i,\theta,b_i}$, $\overline{X}_{\theta,b_i}$ and g_i computation

In this section we explain how to get numerical approximations for $\min_{x_{0,i}^u}\{\min_{u_i}\mathcal{C}_i(b_i,x_{i,0},u_i\mid\theta,U)\}$ and $\overline{u}_{i,\theta,b_i}$ linked to the perturbed ODE

$$\begin{cases} \dot{x}_i(t) = f_{\theta,b_i}(t,x_i(t),z_i(t)) + Bu_i(t) \\ x_i(0) = x_{i,0} \end{cases} \tag{0.1}$$

which are then used to evaluate $\overline{X}_{\theta,b_i}$ and g_i. Firstly we approximate g_i with a special type of discrete time optimal control problem, known as 'tracking problem'. Secondly we adapt the method proposed by [2, 3] to solve it.

g_i expression as an optimal control problem

We introduce a pseudo-linear version of model (0.1):

$$\begin{cases} \dot{x}_i(t) = A_{\theta,b_i}\left(t,x_i(t),z_i(t)\right)x_i(t) + r_{\theta,b_i}(t,z_i(t)) + Bu_i(t) \\ x_i(0) = x_{i,0} \end{cases} \tag{0.2}$$

where A_{θ,b_i} (resp. r_{θ,b_i}) is a $d\times d$ sized matrix (resp. d dimensional vector) valued function, linked to the original model by the relation $A_{\theta,b_i}\left(t,x_i(t),z_i(t)\right)x_i(t) + r_{\theta,b_i}(t,z_i(t)) = f_{\theta,b_i}(t,x_i(t),z_i(t))$. We consider a discretized version of the perturbed ODE (0.1) to proceed to parametric estimation:

$$\begin{cases} x_i(t_{k+1}^d) = \left(I_d + \Delta_k A_{\theta,b_i}(t_k^d,x_i(t_k^d),z_i(t_k^d))\right)x_i(t_k^d) + \Delta_k r_{\theta,b_i}(t_k^d,z_i(t_k^d)) + B\Delta_k u_i(t_k^d) \\ x_i(0) = x_{i,0} \end{cases} \tag{0.3}$$

where the discretization is made at K_i+1 time points $\{t_k^d\}_{0\leq k\leq K_i}$ with $t_0^d = 0$ and $t_{K_i}^d = t_{in_i}$. This set contains the observations time points i.e. $\{t_{ij}\}_{0\leq j\leq n_i}\subset\{t_k^d\}_{0\leq k\leq K_i}$, but can be bigger and patient specific, allowing to accurately approximate $X_{\theta,b_i,x_{i,0}}$ even when the observations are sparse on $[0,\,T]$. We define:

- $\Delta_k = t_{k+1}^d - t_k^d$, the mesh size between two discretization time-points,
- u_i^d the set of discrete values taken by the control at each time step i.e $u_i^d = \left(u(t_k^d),\ldots,u(t_{K_i-1}^d)\right)$,
- $w_k = 1_{\left\{\exists t_{ij}\mid t_{ij}=t_k^d\right\}}/(t_{k+1}^d - t_k^d)$ i.e. w_k is equal to $1/(t_{k+1}^d - t_k^d)$ if t_k^d corresponds to an observation time t_{ij}, otherwise $w_k = 0$,
- y_k^d=y_{ij} if $t_k^d = t_{ij}$, 0 otherwise,

- $X^d_{\theta,b_i,x_{i,0},u^d_i}$ the solution of (0.3).

The weights w_k and the set of extended data $\{y^d_k\}$ are introduced to have a vector of observations with the same length as $\{t^d_k\}_{0\leq k\leq K_i}$. We now introduce the discretized version of the cost $\mathcal{C}_i$ to be minimized:

$$\begin{aligned}\mathcal{C}^d_i(b_i,x_{i,0},u^d_i \mid \theta,U) &= \textstyle\sum_{j=0}^{n_i}\left\|CX^d_{\theta,b_i,x_{i,0},u^d_i}(t_{ij})-y_{ij}\right\|_2^2+\sum_{k=0}^{K_i-1}\triangle_k u_i(t_k)^T U u_i(t_k)\\ &= \left\|CX^d_{\theta,b_i,x_{i,0},u^d_i}(t_{in_i})-y_{in_i}\right\|_2^2 \\ &+ \textstyle\sum_{k=0}^{K_i-1}\triangle_k\left(\left\|CX^d_{\theta,b_i,x_{i,0},u^d_i}(t^d_k)-y^d_k\right\|_2^2 w_k+u_i(t_k)^T U u_i(t_k)\right).\end{aligned} \tag{0.4}$$

such that our inner criteria g_i can be approximated by:

$$g_i(b_i \mid \theta,\Delta,U)\simeq \min_{x^u_{0,i}}\min_{u^d_i}\mathcal{C}^d_i(b_i,x_{i,0},u^d_i \mid \theta,U)+\|\Delta b_i\|_2^2 .$$

The solution of this discrete control problem will be denoted $\overline{u}^d_{i,\theta,b_i}$, and the related optimal trajectory $\overline{X}^d_{\theta,b_i}$: they will be used as numerical approximations of $\overline{u}_{i,\theta,b_i}$ and $\overline{X}_{\theta,b_i}$ respectively.

Numerical methods for solving the tracking problem

We present how to numerically obtain $\min_{x^u_{0,i}}\min_{u^d_i}\mathcal{C}^d_i(b_i,x_{i,0},u^d_i \mid \theta,U)$ as well as the corresponding minimizer $\overline{u}^d_{i,\theta,b_i}$. We start with linear ODE models, then we consider nonlinear models following the steps detailed in [3]. In the rest of the section, we drop the dependence in the covariate $z_i(t^d_k)$ in A_{θ,b_i} and r_{θ,b_i} for the sake of notation clarity.

Linear models

Here, we suppose $A_{\theta,b_i}(t) := A_{\theta,b_i}(t,x)$ in the pseudo-linear model formulation. For a given set $(\theta,b_i,x_{i,0})$, Linear-Quadratic theory ensures the existence and uniqueness of the optimal control $\overline{u}^d_{i,\theta,b_i}$ and that $\min_{x^u_{0,i}}\min_{u^d_i}\mathcal{C}^d_i(b_i,x_{i,0},u^d_i \mid \theta,U)$ can be computed by solving a discrete final value problem, called the Riccati equation (e.g. [6, 1]).

Proposition 0.1. *Let us introduce $(R_{\theta,b_i,k}, h_{\theta,b_i,k})$ for $1\leq k\leq K_i$, the solution of the discrete*

Riccati equation:

$$\left\{\begin{array}{lcl} R_{\theta,b_i,k} & = & R_{\theta,b_i,k+1} + \triangle_k w_k C^T C + \Delta_k \left(R_{\theta,b_i,k+1} A_{\theta,b_i}(t_k^d) + A_{\theta,b_i}(t_k^d)^T R_{\theta,b_i,k+1}\right) \\ & + & \triangle_k^2 A_{\theta,b_i}(t_k^d)^T R_{\theta,b_i,k+1} A_{\theta,b_i}(t_k^d) \\ & - & \triangle_k (I_d + \triangle_k A_{\theta,b_i}(t_k^d)^T) R_{\theta,b_i,k+1} BG(R_{\theta,b_i,k+1}) B^T R_{\theta,b_i,k+1} (I_d + \triangle_k A_{\theta,b_i}(t_k^d)) \\ h_{\theta,b_i,k} & = & h_{\theta,b_i,k+1} - \triangle_k w_k C^T y_k^d + \triangle_k A_{\theta,b_i}(t_k^d)^T h_{\theta,b_i,k+1} \\ & + & \Delta_k \left(I_d + \Delta_k A_{\theta,b_i}(t_k^d)\right)^T R_{\theta,b_i,k+1} r_{\theta,b_i}(t_k^d) \\ & - & \Delta_k (I_d + \Delta_k A_{\theta,b_i}(t_k^d))^T R_{\theta,b_i,k+1} BG(R_{\theta,b_i,k+1}) B^T \left(h_{\theta,b_i,k+1} + \Delta_k R_{\theta,b_i,k+1} r_{\theta,b_i}(t_k^d)\right) \end{array}\right. \tag{0.5}$$

with final condition $(R_{\theta,b_i,K_i},\ h_{\theta,b_i,K_i}) = (C^T C,\ -C^T y_{in_i})$ *and* $G(R_{\theta,b_i,k+1})\ :=\ \left[U + \triangle_k B^T R_{\theta,b_i,k+1} B\right]^{-1}$.

Hence we get:

$$\begin{aligned} & g_i(b_i \mid \theta, \Delta, U) = \|\Delta b_i\|_2^2 + y_{in_i}^T y_{in_i} \\ & - \left(R_{\theta,b_i,0}^{uk} x_{0,i}^k + h_{\theta,b_i,0}^u\right)^T \left(R_{\theta,b_i,0}^u\right)^{-1} \left(R_{\theta,b_i,0}^{uk} x_{0,i}^k + h_{\theta,b_i,0}^u\right) + \left(x_{0,i}^k\right)^T R_{\theta,b_i,0}^k x_{0,i}^k + 2\left(h_{\theta,b_i,0}^k\right)^T x_{0,i}^k \\ & + \textstyle\sum_{k=0}^{K_m-1} \triangle_k \left(w_k \left(y_k^d\right)^T y_k^d + \left(2\left(h_{\theta,b_i,k+1}\right)^T + \Delta_k r_{\theta,b_i}(t_k^d)^T R_{\theta,b_i,k+1}\right) r_{\theta,b_i}(t_k^d)\right) \\ & - \textstyle\sum_{k=0}^{K_m-1} \triangle_k \left(h_{\theta,b_i,k+1} + \Delta_k R_{\theta,b_i,k+1} r_{\theta,b_i}(t_k^d)\right)^T BG(R_{\theta,b_i,k+1}) B^T \left(h_{\theta,b_i,k+1} + \Delta_k R_{\theta,b_i,k+1} r_{\theta,b_i}(t_k^d)\right) \end{aligned} \tag{0.6}$$

where $R_{\theta,b_i,0}^u, R_{\theta,b_i,0}^{uk}, R_{\theta,b_i,0}^k, h_{\theta,b_i,0}^u$ *and* $h_{\theta,b_i,0}^k$ *are given by the following decomposition* $R_{\theta,b_i,0} := \begin{pmatrix} R_{\theta,b_i,0}^u & R_{\theta,b_i,0}^{uk} \\ \left(R_{\theta,b_i,0}^{uk}\right)^T & R_{\theta,b_i,0}^k \end{pmatrix}$ *and* $h_{\theta,b_i,0} := \begin{pmatrix} h_{\theta,b_i,0}^u & h_{\theta,b_i,0}^k \end{pmatrix}$. *Moreover, the control* $\overline{u}_{i,\theta,b_i}^d$ *which minimizes the cost (0.4) is unique and equal to:*

$$\overline{u}_{i,\theta,b_i}^d(t_k^d) = -G(R_{\theta,b_i,k+1}) B^T \left(R_{\theta,b_i,k+1} \left(\left(I_d + \triangle_k A_{\theta,b_i}(t_k^d)\right) \overline{X}_{\theta,b_i}^d(t_k^d) + \Delta_k r_{\theta,b_i}(t_k^d)\right) + h_{\theta,b_i,k+1}\right) \tag{0.7}$$

where $\overline{X}_{\theta,b_i}^d$ *is the optimal trajectory, i.e. the solution of the initial value problem:*

$$\left\{\begin{array}{lcl} \overline{X}_{\theta,b_i}^d(t_{k+1}^d) & = & \left(I_d + \triangle_k A_{\theta,b_i}(t_k^d)\right) \overline{X}_{\theta,b_i}^d(t_k^d) + \Delta_k r_{\theta,b_i}(t_k^d) \\ & - & \triangle_k BG(R_{\theta,b_i,k+1}) B^T R_{\theta,b_i,k+1} \left(\left(I_d + \triangle_k A_{\theta,b_i}(t_k^d)\right) \overline{X}_{\theta,b_i}^d(t_k) + \triangle_k r_{\theta,b_i}(t_k^d)\right) \\ & - & \triangle_k BG(R_{\theta,b_i,k+1}) B^T h_{\theta,b_i,k+1} \end{array}\right. \tag{0.8}$$

with estimator $\widehat{x_{i,0}^u} = -\left(R_{\theta,b_i,0}^u\right)^{-1} \left(R_{\theta,b_i,0}^{uk} x_0^k + h_{\theta,b_i,0}^u\right)$ *for* $x_{i,0}^u$.

Remark 0.2. The theoretical basis for replacing $\mathcal{C}_i^d$ and the perturbed ODE (0.1) by their discretized counterparts can be found in [4] where, under mild regularity conditions on A_{θ,b_i} and r_{θ,b_i}, $\overline{X}_{\theta,b_i}^d$ and $\overline{u}_{i,\theta,b_i}^d$ converge to the solution of the continuous optimal control problem.

Non-linear models

We adapt the method proposed by [3] to solve tracking problem for discrete time models. The outline of the method is the following: we replace the original problem (0.4) by a recursive sequence of problems, where the l-th one is defined by:

$$\begin{array}{rcl}
\min_{u_i^d} \mathcal{C}_i^{d,l}(b_i, x_{i,0}, u_i^d \mid \theta, U) & := & \left\| C X^{d,l}_{\theta,b_i,x_{i,0},u_i^d}(t_{in_i}) - y_{in_i} \right\|_2^2 \\
& + & \sum_{k=0}^{K_i-1} \triangle_k \left(\left\| C X^{d,l}_{\theta,b_i,x_{i,0},u_i^d}(t_k^d) - y_k^d \right\|_2^2 w_k + u_i(t_k)^T U u_i(t_k) \right)
\end{array}$$
$$\text{such that} \begin{cases} x_i(t_{k+1}^d) = \left(I_d + \Delta_k A_{\theta,b_i}(t_k^d, \overline{X}^{d,l-1}_{\theta,b_i}(t_k^d)) \right) x_i(t_k^d) + \Delta_k r_{\theta,b_i}(t_k^d) + B\Delta_k u_i(t_k) \\ x_i(0) = x_{i,0}. \end{cases} \tag{0.9}$$

where $\overline{X}^{d,l-1}_{\theta,b_i}$ is the solution of problem (0.9) at iteration $l-1$. Thus, for each l the matrix $A_{\theta,b_i}(t_k^d, \overline{X}^{d,l-1}_{\theta,b_i}(t_k^d))$ does not depends on x_i and the problem (0.9) is a Linear-Quadratic one. We use the results of section to construct the following algorithm:

1. Initialization phase: $\overline{X}^{u,d,0}_{\theta,b_i}(t_k^d) = x_{i,0}^{u,r}$ for all $k \in [\![0,\, n_i]\!]$ where $x_{i,0}^{u,r}$ is an arbitrary starting point for the unknown initial condition and $\overline{X}^{k,d,0}_{\theta,b_i}(t_k^d) = x_{i,0}^k$.

2. At iteration l: use proposition 0.1 to obtain $(R^l_{\theta,b_i},\, h^l_{\theta,b_i})$, $\overline{u}^{d,l}_{i,\theta,b_i}$,$\overline{X}^{d,l}_{\theta,b_i}$ and $g_i^l(b_i \mid \theta, \Delta, U)$.

3. If $\sum_{k=1}^{K_i} \left\| \overline{X^{d,l}_{\theta,b_i}}(t_k^d) - \overline{X^{d,l-1}_{\theta,b_i}}(t_k^d) \right\|_2^2 < \varepsilon_1$ and $\left| g_i^l(b_i \mid \theta, \Delta, U) - g_i^{l-1}(b_i \mid \theta, \Delta, U) \right| < \varepsilon_2$, then step 4; otherwise get back to step 2.

4. Set $(R_{\theta,b_i},\, h_{\theta,b_i}) = (R^l_{\theta,b_i},\, h^l_{\theta,b_i})$, $\overline{u}^d_{i,\theta,b_i} = \overline{u}^{d,l}_{i,\theta,b_i}$, $\overline{X}^d_{\theta,b_i} = \overline{X}^{d,l}_{\theta,b_i}$ and $g_i(b_i \mid \theta, \Delta, U) = g_i^l(b_i \mid \theta, \Delta, U)$.

Appendix C: Estimation results for $n = 20$

In this section, we resume examples of section 4 for $n = 20$. We explore the effect of small sample size on the relative accuracy of our method with respect to ML. In particular, we want to estimate the coverage rate of our 95%-confidence interval in a situation where the asymptotic justification for our variance estimator (and ML variance estimator as well) is questionable. Interestingly, the regularization brought by our method limits the degradation of estimation accuracy when we move from $n = 50$ to $n = 20$ comparing to ML in all examples below both in well and misspecified cases. Also, our confidence intervals generally have a better coverage rate and closer to 95% than the ML ones, the exception for Ψ^* being explained by the high variance of SAEM estimator for these parameters.

		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	$\widehat{\theta}^{ML}_{x_0}$	0.02	-0.01	0.02	0.02	0.89		0.01	-0.01	0.01	0.02	0.94	
	$\widehat{\theta}^{ML}_{x_{0,2}}$	0.03	-0.01	0.03	0.03	0.91		0.01	-0.01	0.01	0.02	0.88	
	$\widehat{\theta}^{ML}$	0.20	0.01	0.20	0.04	0.80		0.30	0.09	0.29	0.24	0.70	
	$\widehat{\theta}$	**0.02**	**-0.02**	**0.02**	**0.03**	**0.96**		**0.03**	**1e-3**	**0.03**	**0.24**	**1**	
θ_2	$\widehat{\theta}^{ML}_{x_0}$	0.02	0.02	0.02	0.001	0.94		5e-4	-5e-4	5e-4	4e-3	0.70	
	$\widehat{\theta}^{ML}_{x_{0,2}}$	0.02	0.02	0.02	3e-4	0.89		6e-4	-1e-3	6e-4	5e-4	0.65	
	$\widehat{\theta}^{ML}$	0.03	0.05	0.02	2e-3	0.69		0.03	0.07	0.03	0.002	0.62	
	$\widehat{\theta}$	**1e-4**	**2e-3**	**1e-4**	**1e-4**	**0.92**		**4e-4**	**6e-4**	**4e-4**	**4e-4**	**0.92**	
Ψ	$\widehat{\theta}^{ML}_{x_0}$	0.02	0.01	0.02	0.02	1	0.01	0.02	-0.02	0.02	0.02	1	0.01
	$\widehat{\theta}^{ML}_{x_{0,2}}$	0.03	-0.01	0.03	0.02	1	0.01	0.02	-0.03	0.02	0.02	1	0.02
	$\widehat{\theta}^{ML}$	0.24	0.26	0.14	0.11	1	0.15	0.46	0.37	0.33	0.05	1	0.18
	$\widehat{\theta}$	**0.03**	**-0.03**	**0.03**	**0.03**	**0.88**	**0.01**	**0.04**	**-0.05**	**0.03**	**0.02**	**0.92**	**0.01**

Tab. 0.1: Results of estimation for model (4.1) for $n = 20$

		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}	$\widehat{\theta}^{MS}_{S_i}$	2e-4	5e-3	1e-4	5e-5	0.74		1e-4	7e-3	7e-5	2e-5	0.72	
	$\widehat{\theta}_{ML}$	0.01	0.06	7e-3	1e-3	0.62		0.02	0.04	0.01	2e-4	0.50	
	$\widehat{\theta}_{S_i}$	**7e-5**	**1e-3**	**7e-5**	**7e-5**	**0.92**		**3e-5**	**2e-3**	**3e-5**	**3e-5**	**0.94**	
	$\widehat{\theta}$	**4e-3**	**2e-3**	**4e-3**	**4e-3**	**0.92**		**5e-4**	**8e-4**	**5e-4**	**5e-4**	**0.93**	
θ_{S_I}	$\widehat{\theta}^{MS}_{S_i}$	known						known					
	$\widehat{\theta}_{ML}$	0.02	0.07	0.01	1e-3	0.67		0.02	0.02	0.02	1e-3	0.50	
	$\widehat{\theta}_{S_i}$	**known**						**known**					
	$\widehat{\theta}$	**4e-4**	**5e-4**	**4e-4**	**6e-4**	**0.92**		**2e-3**	**-1e-3**	**2e-3**	**3e-3**	**0.92**	
θ_n	$\widehat{\theta}^{MS}_{S_i}$	2e-3	-5e-3	2e-3	1e-4	0.86		1e-3	-3e-4	1e-3	1e-3	0.89	
	$\widehat{\theta}_{ML}$	0.01	-3e-3	0.01	1e-3	0.82		5e-3	-5e-3	5e-3	1e-3	0.50	
	$\widehat{\theta}_{S_i}$	**1e-3**	**5e-3**	**1e-3**	**1e-3**	**0.92**		**1e-3**	**3e-3**	**1e-3**	**1e-3**	**0.90**	
	$\widehat{\theta}$	**1e-3**	**5e-3**	**1e-3**	**1e-3**	**0.90**		**1e-3**	**5e-3**	**1e-3**	**1e-3**	**0.90**	
Ψ	$\widehat{\theta}^{MS}_{S_i}$	0.05	-0.04	0.04	0.05	0.92	0.03	0.03	-0.05	0.02	0.04	0.91	0.02
	$\widehat{\theta}_{ML}$	0.06	-0.04	0.05	0.05	0.90	0.03	0.05	-0.07	0.05	0.05	0.87	0.02
	$\widehat{\theta}_{S_i}$	**0.03**	**-0.04**	**0.03**	**0.02**	**0.90**	**0.01**	**0.02**	**-0.05**	**0.02**	**0.02**	**0.93**	**0.01**
	$\widehat{\theta}$	**0.04**	**-0.04**	**0.04**	**0.04**	**0.90**	**0.02**	**0.03**	**-0.06**	**0.03**	**0.02**	**0.92**	**0.01**

Tab. 0.2: Results of estimation for model (4.3) for $n = 20$

Partially observed linear model

We present the estimation results in table 0.1.

Partially observed nonlinear model

We present the estimation results in table 0.2.

Antibody concentration evolution model

Estimation results are presented in table 0.3.

		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}	$\widehat{\theta}^{ML}_{\delta_S}$	known						known					
	$\widehat{\theta}_{ML}$	3.58	1.35	1.75	76.83	0.87		4.64	1.69	1.76	63.9	0.70	
	$\widehat{\theta}_{\delta_S}$	**known**						**known**					
	$\widehat{\theta}$	**1.10**	**-0.40**	**0.97**	**1.14**	**0.87**		**1.32**	**-0.70**	**0.92**	**0.60**	**0.77**	
θ_{ϕ_S}	$\widehat{\theta}^{ML}_{\delta_S}$	3e-3	0.02	3e-3	1e-3	0.89		1e-3	0.02	1e-3	1e-3	0.84	
	$\widehat{\theta}_{ML}$	0.02	-0.08	7e-3	0.43	0.90		0.02	-0.09	8e-3	0.37	0.85	
	$\widehat{\theta}_{\delta_S}$	**3e-3**	**-0.04**	**1e-3**	**2e-3**	**0.90**		**1e-3**	**-0.02**	**7e-4**	**3e-3**	**0.91**	
	$\widehat{\theta}$	**7e-3**	**-0.03**	**6e-3**	**4e-3**	**0.90**		**0.01**	**-0.01**	**0.01**	**2e-3**	**0.89**	
θ_{ϕ_L}	$\widehat{\theta}^{ML}_{\delta_S}$	0.01	0.02	0.01	6e-3	0.91		7e-3	0.02	7e-3	5e-3	0.87	
	$\widehat{\theta}_{ML}$	0.02	0.05	0.02	7e-3	0.84		0.01	0.04	0.01	6e-3	0.86	
	$\widehat{\theta}_{\delta_S}$	**4e-3**	**-0.01**	**4e-3**	**0.01**	**0.95**		**7e-3**	**7e-3**	**7e-3**	**9e-3**	**0.95**	
	$\widehat{\theta}$	**0.01**	**4e-3**	**0.01**	**0.01**	**0.85**		**0.06**	**-0.03**	**0.06**	**4e-3**	**0.91**	
$\theta_{\delta_{Ab}}$	$\widehat{\theta}^{ML}_{\delta_S}$	3e-3	-0.02	3e-3	1e-3	0.82		2e-3	-0.01	1e-3	1e-3	0.77	
	$\widehat{\theta}_{ML}$	0.01	-0.03	0.02	0.01	0.76		4e-3	-0.03	4e-3	1e-3	0.85	
	$\widehat{\theta}_{\delta_S}$	**1e-3**	**0.01**	**4e-4**	**1e-3**	**0.93**		**6e-4**	**7e-3**	**5e-4**	**1e-3**	**0.94**	
	$\widehat{\theta}$	**6e-4**	**0.01**	**5e-4**	**6e-4**	**0.86**		**8e-4**	**0.01**	**7e-4**	**6e-4**	**0.90**	
Ψ_{ϕ_S}	$\widehat{\theta}^{ML}_{\delta_S}$	0.04	0.05	0.03	0.13	1	0.15	0.09	0.05	0.09	2.04	1	0.21
	$\widehat{\theta}_{ML}$	0.16	0.05	0.15	0.28	1	0.18	0.15	0.11	0.15	1.22	1	0.25
	$\widehat{\theta}_{\delta_S}$	**0.02**	**0.02**	**0.03**	**0.05**	**0.93**	**0.10**	**0.03**	**-0.04**	**0.03**	**0.03**	**0.88**	**0.11**
	$\widehat{\theta}$	**0.04**	**-0.01**	**0.04**	**0.03**	**0.90**	**0.11**	**0.05**	**-0.06**	**0.04**	**0.03**	**0.85**	**0.15**
Ψ_{ϕ_L}	$\widehat{\theta}^{ML}_{\delta_S}$	0.05	0.03	0.05	0.08	1	0.30	0.08	-0.20	0.04	0.10	1	0.95
	$\widehat{\theta}_{ML}$	0.09	0.05	0.08	0.18	1	0.64	0.10	0.08	0.08	0.14	1	0.96
	$\widehat{\theta}_{\delta_S}$	**0.03**	**-0.03**	**0.03**	**0.03**	**0.86**	**0.10**	**0.04**	**-0.11**	**0.03**	**0.03**	**0.88**	**0.12**
	$\widehat{\theta}$	**0.04**	**-0.10**	**0.03**	**0.03**	**0.90**	**0.12**	**0.06**	**-0.16**	**0.03**	**0.03**	**0.90**	**0.15**
$\Psi_{\delta_{Ab}}$	$\widehat{\theta}^{ML}_{\delta_S}$	0.23	0.10	0.20	0.12	1	0.38	0.27	0.27	0.19	0.09	1	0.46
	$\widehat{\theta}_{ML}$	0.46	0.30	0.37	0.17	1	0.55	0.51	0.64	0.28	0.14	1	0.74
	$\widehat{\theta}_{\delta_S}$	**0.09**	**-0.21**	**0.04**	**0.11**	**0.85**	**0.03**	**0.11**	**-0.21**	**0.07**	**0.08**	**0.88**	**0.05**
	$\widehat{\theta}$	**0.18**	**-0.34**	**0.07**	**0.10**	**0.83**	**0.04**	**0.16**	**-0.31**	**0.07**	**0.08**	**0.88**	**0.05**

Tab. 0.3: Results of estimation for model (4.6) for $n = 20$.

Appendix D: Formal details for Variance-Covariance estimator

Before presenting the proof, let us introduce notations used in the following lemma:

1. The operator $\otimes$ defined by the relationship $\left(\frac{d^2}{d\gamma_1 d\gamma_2}Q\right)^T \otimes R = \sum_i R_i \frac{d^2}{d\gamma_1 d\gamma_2}Q_i$ for two same dimensional vectors Q and R,

2. $\overline{\partial_{\alpha_1,\ldots\alpha_l} f} = \sup_{(\theta,b_i,t,x)} \left\| \frac{\partial^l f_{\theta,b_i}}{\partial\alpha_1,\ldots\partial\alpha_l}(t,x) \right\|_2$ for $(\theta, b_i, t, x) \in \Theta \times \Theta_b \times [0,\, T] \times \chi$.

Theorem 0.3. *Under conditions 1-7, there is a model dependent lower bound λ such that if $\|U\|_2 > \lambda$ then the estimator $\left(\widehat{\theta}, \widehat{\delta}\right)$ is asymptotically normal and:*

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

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:*

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

Proof. From lemma 0.4, we can use the law of large number presented theorem 5.1.2 in Wang (1968) for the independent, possibly non identically distributed, random variables $h(\widehat{b}(\theta, \Delta(\delta)), \theta, \Delta(\delta), y_i)$ to derive the almost surely uniform convergence of

$$\frac{1}{n}\sum_i^n h(\widehat{b}(\theta, \Delta(\delta)), \theta, \Delta(\delta), y_i) \longrightarrow \lim_n \frac{1}{n}\sum_i^n \mathbb{E}\left[h(\widehat{b}(\theta, \Delta(\delta)), \theta, \Delta(\delta), y_i)\right]$$

on $\Theta \times \Omega$. This is in turn sufficient to derive the same result for $\frac{1}{n} G\left[\theta, \Delta(\delta) \mid \mathbf{y}\right]$ toward $\widetilde{G}\left[\theta, \Delta(\delta)\right]$. Now, by relying on condition 1 regarding the well-separateness property of $\left(\overline{\theta}, \overline{\delta}\right)$, we can apply theorem 5.3 in [7] to conclude that $\left(\widehat{\theta}, \widehat{\delta}\right) \longrightarrow \left(\overline{\theta}, \overline{\delta}\right)$ in probability when $n \longrightarrow \infty$.

Since $\left(\widehat{\theta}, \widehat{\delta}\right)$ is defined as an M-estimator, we can obtain an asymptotic Variance-Covariance estimator by using a Taylor expansion of equation $\nabla_{(\theta,\delta)} G\left[\widehat{\theta}, \Delta(\widehat{\delta}) \mid \mathbf{y}\right] = 0$. By developing the previous gradient expression, we have $\sum_{i=1}^n J(\widehat{\theta}, \widehat{\delta}, \mathbf{y_i}) = M_n(\widehat{\theta}, \widehat{\delta})$ where

$$M_n(\theta, \delta) = \frac{2}{\mathbb{P}(\theta, \Delta(\delta))} \frac{\partial \mathbb{P}(\theta, \Delta(\delta))}{\partial\,(\theta, \delta)} \frac{\sum_i h(\widehat{b}_i(\theta, \Delta(\delta)), \theta, \Delta(\delta), \mathbf{y_i})}{d^o \sum_i n_i + qn}$$

and the $(p+q')$ components of the vector valued function J for $1 \leq k \leq p$ are given by

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

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

$$J_k(\theta, \delta, y_i) = \frac{d}{d\delta_k} h(\widehat{b}_i(\theta, \Delta(\delta)), \theta, \Delta(\delta), y_i) - \frac{2}{d^o \frac{\sum_i n_i}{n} + q} Tr\left(\triangle(\delta)^{-1} \frac{\partial \triangle(\delta)}{\partial \delta_k}\right) h(\widehat{b}_i(\theta, \Delta(\delta)), \theta, \Delta(\delta), y_i).$$

In order to use asymptotic results, we want to consider quantities independents of n and n_i, the n_i being i.i.d by using the central limit theorem we get $\sqrt{n}\left(\frac{\sum_i n_i}{n} - \mathbb{E}\left[n_1\right]\right) \rightsquigarrow 0$, the last convergence in distribution is also a convergence in probability which can be re-expressed as $\frac{\sum_i n_i}{n} = \mathbb{E}\left[n_1\right] + \sqrt{n}^{-1} o_{p,n}(1)$. From this, we can introduce $\widetilde{J}$ given by $\widetilde{J_k}(\theta, \delta, y_i) = J_k(\theta, \delta, y_i)$ for $1 \leq k \leq q'$ and for $p+1 \leq k \leq p+q'$ by

$$\widetilde{J_k}(\theta, \delta, y_i) = \frac{d}{d\delta_k} h(\widehat{b}_i(\theta, \Delta(\delta)), \theta, \Delta(\delta), y_i) - \frac{2}{d^o \mathbb{E}\left[n_1\right] + q} Tr\left(\triangle(\delta)^{-1} \frac{\partial \triangle(\delta)}{\partial \delta_k}\right) h(\widehat{b}_i(\theta, \Delta(\delta)), \theta, \Delta(\delta), y_i)$$

which respects the approximation $J_k(\theta, \delta, y_i) = \widetilde{J_k}(\theta, \delta, y_i) + \sqrt{n}^{-1} o_{p,n}(1)$ and

$$\sum_{i=1}^{n} \widetilde{J}(\widehat{\theta}, \widehat{\delta}, \mathbf{y_i}) = \widetilde{M_n}(\widehat{\theta}, \widehat{\delta}) + \sqrt{n} o_{p,n}(1)$$

with $\widetilde{M_n}(\theta, \delta) = \frac{2}{\mathbb{P}(\theta, \Delta(\delta))} \frac{\partial \mathbb{P}(\theta, \Delta(\delta))}{\partial(\theta, \delta)} \frac{1}{n} \frac{\sum_i h(\widehat{b}_i(\theta, \Delta(\delta)), \theta, \Delta(\delta), \mathbf{y_i})}{d^o \mathbb{E}[n_1] + q}$. Now we want to proceed to the Taylor expansion of $K_n(\widehat{\theta}, \widehat{\delta}) = n^{-1} \sum_{i=1}^{n} \widetilde{J}(\widehat{\theta}, \widehat{\delta}, \mathbf{y_i})$ around $\left(\overline{\theta}, \overline{\delta}\right)$, but first we need to ensure $K_n \in C^2(\Theta_{\overline{\theta}}^{(1)} \times \Omega_{\overline{\delta}}^{(1)}, \mathbb{R}^{(p+q')})$ for a neighborhood $\Theta_{\overline{\theta}}^{(1)} \times \Omega_{\overline{\delta}}^{(1)}$ of $\left(\overline{\theta}, \overline{\delta}\right)$ almost surely for every sequence $\mathbf{y_i}$. From the definition of $\widetilde{J}$, it is obvious this holds if $(\theta, \delta) \longmapsto h(\widehat{b}_i(\theta, \Delta(\delta)), \theta, \Delta(\delta), \mathbf{y_i}) \in C^3(\Theta_{\overline{\theta}}^{(1)} \times \Omega_{\overline{\delta}}^{(1)}, \mathbb{R})$. From the conditions 6 and 7 on $f_{\theta, b_i}$and $\mathcal{C}_i$ invertibility and part 2) of lemma 0.6, we derive that it exists a neighborhood of $\widetilde{\Theta}_{\overline{\theta}} \times \widetilde{\Theta}_{\widehat{b}_i(\overline{\theta}, \Delta(\overline{\delta}))} \subset \Theta_{\overline{\theta}} \times \mathbb{R}^q$ such that $(\theta, b_i, t) \longmapsto \overline{X}_{\theta, b_i}(t) \in C^3(\widetilde{\Theta}_{\overline{\theta}} \times \widetilde{\Theta}_{\widehat{b}_i(\overline{\theta}, \Delta(\overline{\delta}))} \times \left[0,\, T\right], \mathbb{R}^d)$ and $(\theta, b_i, t) \longmapsto \overline{u}_{\theta, b_i}(t) \in C^3(\widetilde{\Theta}_{\overline{\theta}} \times \widetilde{\Theta}_{\widehat{b}_i(\overline{\theta}, \Delta(\overline{\delta}))} \times \left[0,\, T\right], \mathbb{R}^{d_u})$. From this, it is easy to see that $(\theta, b_i, \delta) \longmapsto h(b_i, \theta, \Delta(\delta), \mathbf{y_i})$ and $(\theta, b_i, \delta) \longmapsto g_i(b_i \mid \theta, \Delta(\delta), U)$ both belong to $C^3(\widetilde{\Theta}_{\overline{\theta}} \times \widetilde{\Theta}_{\widehat{b}_i(\overline{\theta}, \Delta(\overline{\delta}))} \times \Omega, \mathbb{R})$ almost surely for every sequence $\mathbf{y_i}$. Since $\nabla_{b_i} g_i(\widehat{b_i}\left(\theta, \Delta(\delta)\right) \mid \theta, \Delta(\delta), U) = 0$ and we assume$\frac{\partial^2}{\partial^2 b_i} g_i(\widehat{b_i}\left(\overline{\theta}, \Delta(\overline{\delta})\right) \mid \overline{\theta}, \Delta(\overline{\delta}), U)$ is of full rank (condition 6), we use the implicit function theorem to conclude about the existence of an open subset $\Theta_{\overline{\theta}}^{(1)} \times \Omega_{\overline{\delta}}^{(1)} \subset \widetilde{\Theta}_{\overline{\theta}} \times \Omega$ such that $(\theta, \delta) \longmapsto \widehat{b_i}\left(\theta, \delta\right) \in C^3(\Theta_{\overline{\theta}}^{(1)} \times \Omega_{\overline{\delta}}^{(1)}, \widetilde{\Theta}_{\widehat{b}_i(\overline{\theta}, \Delta(\overline{\delta}))})$. From this, classic results on the regularity of composed functions gives us $(\theta, \delta) \longrightarrow h(\widehat{b}_i(\theta, \Delta(\delta)), \theta, \Delta(\delta), \mathbf{y_i}) \in C^3(\Theta_{\overline{\theta}}^{(1)} \times \Omega_{\overline{\delta}}^{(1)}, \mathbb{R}^{p+q})$. Since $\left(\widehat{\theta}, \widehat{\delta}\right) \longrightarrow (\overline{\theta}, \overline{\delta})$, the event $\left(\widehat{\theta}, \widehat{\delta}\right) \in \Theta_{\overline{\theta}}^{(1)} \times \Omega_{\overline{\delta}}^{(1)}$ has a probability tending to one and we can use the Taylor expansion of $K_n(\widehat{\theta},\widehat{\delta})$ around $\left(\overline{\theta},\overline{\delta}\right)$ on $\Theta^{(1)}_{\overline{\theta}} \times \Omega^{(1)}_{\overline{\delta}}$ which gives us:

$$\frac{1}{n}\widetilde{M_n}(\widehat{\theta},\widehat{\delta})+\sqrt{n}^{-1}o_{p,n}(1) = K_n(\overline{\theta},\overline{\delta})+((\widehat{\theta},\widehat{\delta})-(\overline{\theta},\overline{\delta}))^T\frac{\partial K_n(\overline{\theta},\overline{\delta})}{\partial(\theta,\delta)}+((\widehat{\theta},\widehat{\delta})-(\overline{\theta},\overline{\delta}))^T\frac{\partial^2 K_n(\widetilde{\theta}_n,\widetilde{\delta}_n)}{\partial^2(\theta,\delta)}((\widehat{\theta},\widehat{\delta})-(\overline{\theta},\overline{\delta}))$$

with $\left(\widetilde{\theta}_n,\widetilde{\delta}_n\right)$ defined such that $\left\|(\widetilde{\theta}_n,\widetilde{\delta}_n)-(\overline{\theta},\overline{\delta})\right\| \leq \left\|(\widehat{\theta},\widehat{\delta})-(\overline{\theta},\overline{\delta})\right\|$. By rearranging, we get:

$$\sqrt{n}\left(\frac{\partial K_n(\overline{\theta},\overline{\delta})}{\partial(\theta,\delta)}^T+((\widehat{\theta},\widehat{\delta})-(\overline{\theta},\overline{\delta}))^T\frac{\partial^2 K_n(\widetilde{\theta}_n,\widetilde{\delta}_n)}{\partial^2(\theta,\delta)}\right)((\widehat{\theta},\widehat{\delta})-(\overline{\theta},\overline{\delta})) = -\sqrt{n}K_n(\overline{\theta},\overline{\delta})+\frac{1}{\sqrt{n}}\widetilde{M_n}(\widehat{\theta},\widehat{\delta})+o_{p,n}(1). \tag{0.10}$$

As before, by using the weak law of large numbers, we derive:

$$\frac{\sum_i h(\widehat{b}_i(\theta,\Delta(\delta)),\theta,\Delta(\delta),\mathbf{y_i})}{d^o\sum_i n_i + qn} \longrightarrow \left(\frac{1}{q+d^o\mathbb{E}\left[n_1\right]}\right)\lim_n \frac{1}{n}\sum_i^n \mathbb{E}\left[h(\widehat{b}(\theta,\Delta(\delta)),\theta,\Delta(\delta),y_i)\right]$$

in probability. From this and the continuity mapping theorem it is easy to see that

$$\widetilde{M_n}(\widehat{\theta},\widehat{\delta}) \longrightarrow \frac{2}{\mathbb{P}(\overline{\theta},\Delta(\overline{\delta}))}\frac{\partial\mathbb{P}(\overline{\theta},\Delta(\overline{\delta}))}{\partial\left(\theta,\delta\right)}\left(\frac{1}{q+d^o\mathbb{E}\left[n_1\right]}\right)\mathbb{E}_Y\left[h(\widehat{b}(\overline{\theta},\Delta(\overline{\delta})),\overline{\theta},\Delta(\overline{\delta}),Y)\right]$$

in probability thus $\widetilde{M_n}(\widehat{\theta},\widehat{\delta})$ is an uniformly tight sequence and so $\frac{1}{\sqrt{n}}\widetilde{M_n}(\widehat{\theta},\widehat{\delta}) = o_{p,n}(1)$, the continuity of $\mathbb{E}_Y\left[h(\widehat{b}(\overline{\theta},\Delta(\overline{\delta})),\overline{\theta},\Delta(\overline{\delta}),Y)\right]$ being ensured by the uniform convergence. From these results, we can use the approximation $-\sqrt{n}K_n(\overline{\theta},\overline{\delta}) + \frac{1}{\sqrt{n}}M_n(\widehat{\theta},\widehat{\delta}) + o_{p,n}(1) = -\sqrt{n}K_n(\overline{\theta},\overline{\delta}) + o_{p,n}(1)$ in the right-hand side of equation (0.10). Based on condition 2, if the initial conditions $\left\{x^*_{0,i}\right\}_{i\in[\![1,n]\!]} \in [\![1,\, n]\!]$ are i.i.d so are the $\widetilde{J}(\overline{\theta},\overline{\delta},\mathbf{y_i})$ and we derive $-\sqrt{n}K_n(\overline{\theta},\overline{\delta}) \rightsquigarrow N(0,B(\overline{\theta},\overline{\delta}))$ with $B(\overline{\theta},\overline{\delta}) = \lim_n \frac{1}{n}\left[\sum_i \widetilde{J}(\overline{\theta},\overline{\delta},\mathbf{y_i})\widetilde{J}(\overline{\theta},\overline{\delta},\mathbf{y_i})^T\right]$ according to the central limit theorem, otherwise with the Lindeberg-Feller theorem, bound conditions for $\frac{d}{d(\theta,\delta)}h(\widehat{b}(\overline{\theta},\Delta(\overline{\delta})),\overline{\theta},\Delta(\overline{\delta}),\mathbf{y_i})$ are proven in lemma 0.5. Now, regarding the left-hand side of (0.10), by using the convergence of $\left(\widetilde{\theta}_n,\widetilde{\delta}_n\right)$ toward $(\overline{\theta},\overline{\delta})$ we know the event $\left(\widetilde{\theta}_n,\widetilde{\delta}_n\right) \in \Theta^{(1)}_{\overline{\theta}} \times \Omega^{(1)}_{\overline{\delta}}$ has a probability tending to one. From this, we derive the matrix $\frac{\partial^2 K_n(\widetilde{\theta}_n,\widetilde{\delta}_n)}{\partial^2(\theta,\delta)}$, involving third derivatives of h which are continuous on $\Theta^{(1)}_{\overline{\theta}} \times \Omega^{(1)}_{\overline{\delta}}$, is bounded in probability, thus $((\widehat{\theta},\widehat{\delta})-(\overline{\theta},\overline{\delta}))^T\frac{\partial^2 K_n(\widetilde{\theta}_n,\widetilde{\delta}_n)}{\partial^2(\theta,\delta)} = o_{p,n}(1)$. From this, equation (0.10) can be re-expressed as:

$$\sqrt{n}\left(\frac{\partial K_n(\overline{\theta},\overline{\delta})}{\partial(\theta,\delta)}^T + o_{p,n}(1)\right)((\widehat{\theta},\widehat{\delta})-(\overline{\theta},\overline{\delta})) = -\sqrt{n}K_n(\overline{\theta},\overline{\delta}) + o_{p,n}(1)$$

and we can use the weak law of large numbers to get the convergence in probability of $\frac{\partial K_n(\overline{\theta},\overline{\delta})}{\partial(\theta,\delta)} = n^{-1}\sum_{i=1}^n\left[\frac{\partial\widetilde{J}(\overline{\theta},\overline{\delta},\mathbf{y_i})}{\partial(\theta,\delta)}\right] \rightarrow \lim_n \frac{1}{n}\sum_{i=1}^n\left[\frac{\partial\widetilde{J}(\overline{\theta},\overline{\delta},\mathbf{y_i})}{\partial(\theta,\delta)}\right] = A(\overline{\theta},\overline{\delta})$ with moment boundedness conditions on $\frac{d}{d(\theta,\delta)}h(\widehat{b}(\overline{\theta},\Delta(\overline{\delta})),\overline{\theta},\Delta(\overline{\delta}),\mathbf{y_i})$ proven in lemma 0.5. If the matrix $A(\overline{\theta},\overline{\delta})$ is non-singular, we can then use the multivariate version of Slutsky's lemma to conclude that:

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

□

Lemma 0.4. *Under conditions 3,4,5,* $(\theta,\delta) \longmapsto h(\widehat{b}_i(\theta,\Delta(\delta)),\theta,\Delta(\delta),\mathbf{y_i})$ *is uniformly bounded on* $\Theta \times \Omega$ *and has a finite moment of order two uniformly bounded on* $i \in [\![1,\, n]\!]$

Proof. By definition of our optimal control problem, we derive the following bound by sequentially using $\|\overline{u}_{\theta,b_i}\|^2_{U,L^2} > 0$, upper-bounding the minimum in b_i by the true subject specific parameter values b_i^*, upper-bounding the minimum in $x^u_{0,i}$ by the true initial condition value $x^*_{0,i}$ and upper-bounding in u_i by $u_i^0 = 0$:

$$\begin{aligned}
&h(\widehat{b}_i(\theta,\Delta(\delta)),\theta,\Delta(\delta),\mathbf{y_i}) = \left\|\Delta(\delta)\widehat{b}_i(\theta,\Delta(\delta))\right\|_2^2 + \textstyle\sum_j \left\|C\overline{X}_{\theta,\widehat{b}_i(\theta,\Delta(\delta))}(t_{ij}) - y_{ij}\right\|_2^2\\
&= \left\|\Delta(\delta)\widehat{b}_i(\theta,\Delta(\delta))\right\|_2^2 + \textstyle\sum_j \left\|C\overline{X}_{\theta,\widehat{b}_i(\theta,\Delta(\delta))}(t_{ij}) - y_{ij}\right\|_2^2\\
&\leq \left\|\Delta(\delta)\widehat{b}_i(\theta,\Delta(\delta))\right\|_2^2 + \textstyle\sum_j \left\|C\overline{X}_{\theta,\widehat{b}_i(\theta,\Delta(\delta))}(t_{ij}) - y_{ij}\right\|_2^2 + \left\|\overline{u}_{\theta,\widehat{b}_i(\theta,\Delta(\delta))}\right\|_{U,L^2}^2\\
&= g_i(\widehat{b}_i(\theta,\Delta(\delta)) \mid \theta,\Delta(\delta),U)\\
&= \min_{b_i}\left\{\|\Delta(\delta)b_i\|_2^2 + \min_{x^u_{0,i}}\min_{u_i}\left\{\textstyle\sum_j \|CX_{\theta,b_i,x_{0,i},u_i}(t_{ij}) - y_{ij}\|_2^2 + \|u_i\|_{U,L^2}^2\right\}\right\}\\
&\leq \|\Delta(\delta)b_i^*\|_2^2 + \min_{x^u_{0,i}}\min_{u_i}\left\{\textstyle\sum_j \|CX_{\theta,b_i^*,x_{0,i},u_i}(t_{ij}) - y_{ij}\|_2^2 + \|u_i\|_{U,L^2}^2\right\}\\
&\leq \|\Delta(\delta)b_i^*\|_2^2 + \min_{u_i}\left\{\textstyle\sum_j \left\|CX_{\theta,b_i^*,x^*_{0,i},u_i}(t_{ij}) - y_{ij}\right\|_2^2 + \|u_i\|_{U,L^2}^2\right\}\\
&\leq \|\Delta(\delta)b_i^*\|_2^2 + \textstyle\sum_j \left\|CX_{\theta,b_i^*,x^*_{0,i}}(t_{ij}) - y_{ij}\right\|_2^2
\end{aligned}$$

where $X_{\theta,b_i,x^*_{0,i}}$ is the solution of the original ODE. Since $y_{ij} = CX_{\theta,b_i^*,x^*_{0,i}}(t_{ij}) + \epsilon_{i,j}$, we get:

$$h(\widehat{b}_i(\theta,\Delta(\delta)),\theta,\Delta(\delta),\mathbf{y_i}) \leq \|\Delta(\delta)b_i^*\|_2^2 + \|C\|_2^2\sum_j \left\|X_{\theta,b_i^*,x^*_{0,i}}(t_{ij}) - X_{\theta^*,b_i^*,x^*_{0,i}}(t_{ij})\right\|_2^2 + \sum_j \|\epsilon_{i,j}\|_2^2\,.$$

Now, let us control the behavior of:

$$\begin{aligned}
X_{\theta,b_i^*,x^*_{0,i}}(t) - X_{\theta^*,b_i^*,x^*_{0,i}}(t) \quad &= \quad \textstyle\int_0^t (f_{\theta,b_i^*}(t,X_{\theta,b_i^*,x^*_{0,i}}(t)) - f_{\theta^*,b_i^*}(t,X_{\theta,b_i^*,x^*_{0,i}}(t))dt\\
&+ \quad \textstyle\int_0^t (f_{\theta^*,b_i^*}(t,X_{\theta,b_i^*,x^*_{0,i}}(t)) - f_{\theta^*,b_i^*}(t,X_{\theta^*,b_i^*,x^*_{0,i}}(t))dt
\end{aligned}$$

by triangular inequality, we obtain

$$\left\|X_{\theta,b_i^*,x^*_{0,i}}(t) - X_{\theta^*,b_i^*,x^*_{0,i}}(t)\right\|_2^2 \quad \leq \quad 2\overline{\partial_\theta f}T\,\|\theta-\theta^*\|_2^2 + 2\overline{\partial_x f}\,\textstyle\int_0^t \left\|X_{\theta,b_i^*,x^*_{0,i}}(t) - X_{\theta^*,b_i^*,x^*_{0,i}}(t)\right\|_2^2 dt$$

and from this we can use the Gronwall lemma to derive the following bound uniform in time:

$\left\|X_{\theta,b_i^*,x_{0,i}^*}(t) - X_{\theta^*,b_i^*,x_{0,i}^*}(t)\right\|_2^2 \leq 2\overline{\partial_\theta f}T\left\|\theta-\theta^*\right\|_2^2 e^{2\overline{\partial_x f}T}$ and finally obtain $h(\widehat{b_i}(\theta,\Delta(\delta)),\theta,\Delta(\delta),\mathbf{y_i}) \leq \left\|\Delta(\delta)b_i^*\right\|_2^2 + 2\left\|C\right\|_2^2 T\left\|\theta-\theta^*\right\|_2^2 n_i\overline{\partial_\theta f}e^{2\overline{\partial_x f}T} + \sum_j\left\|\epsilon_{i,j}\right\|_2^2$ since (δ,θ) belongs to a compact, we derive from this last inequality the uniform boundedness of $h(\widehat{b_i}(\theta,\Delta(\delta)),\theta,\Delta(\delta),\mathbf{y_i})$ on $\Theta\times\Omega$. Moreover, $\left\|\Delta(\delta)b_i^*\right\|_2^2$ and $\sum_j\left\|\epsilon_{i,j}\right\|_2^2$ follow χ^2 law and the $\{n_i\}_{i\in[\![1,\,n]\!]}$ are bounded so they both have moment of order two uniformly bounded. By hypothesis on the compact support of f_{θ,b_i} with respect to b, both $\overline{\partial_\theta f_{b_i^*}}$ and $e^{2\overline{\partial_x f_{b_i^*}}T}$ are bounded in probability, the same holds for $n_i\overline{\partial_\theta f_{b_i^*}}e^{2\overline{\partial_x f_{b_i^*}}T}$ and so the previous upper bound we obtain for $h(\widehat{b_i}(\theta,\Delta(\delta)),\theta,\Delta(\delta),\mathbf{y_i})$ has a uniformly bounded moment of order two. □

Lemma 0.5. *Under conditions 4,5,6,7 the second moment of* $\frac{dh}{d(\theta,\delta)}(\widehat{b_i}\left(\overline{\theta},\Delta(\overline{\delta})\right),\overline{\theta},\Delta(\overline{\delta}),\mathbf{y_i})$ *and* $\frac{d^2h}{d^2(\theta,\delta)}(\widehat{b_i}\left(\overline{\theta},\Delta(\overline{\delta})\right),\overline{\theta},\Delta(\overline{\delta}),\mathbf{y_i})$ *are uniformly bounded*

Proof. In the following, we denote $\eta=(\theta,\delta)$ and $\overline{\eta}=(\overline{\theta},\overline{\delta})$ and replace dependence for $\widehat{b_i}$ with respect to parameter by $\widehat{b_i}\left(\overline{\eta}\right)$. Also, we have to emphasize at some point of the lemma the dependence of the estimated initial conditions with respect to the parameters (θ,b), i.e $\widehat{x_0} := \widehat{x_0}(\overline{\theta},\widehat{b_i}\left(\overline{\eta}\right))$ and that for the sake of correctness the optimal trajectory and control should be written $\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})} := \overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta}),\widehat{x_0}(\overline{\eta},\widehat{b}_i(\overline{\eta}))}$ and $\overline{u}_{\overline{\theta},\widehat{b}_i(\overline{\eta})} := \overline{u}_{\overline{\theta},\widehat{b}_i(\overline{\eta}),\widehat{x_0}(\overline{\eta},\widehat{b}_i(\overline{\eta}))}$.

First of all, let us derive that $\mathbb{E}_{b_i}\left[\left\|\overline{u}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}\right\|_{U,L^2}^2\right]$ and $\mathbb{E}_{b_i}\left[\left\|\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t)\right\|_2^2\right]$ are finite. Similarly as the bound derived for h, we get:

$$
\begin{aligned}
&\left\|\overline{u}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}\right\|_{U,L^2}^2 + \left\|\Delta(\overline{\delta})\widehat{b}_i(\overline{\eta})\right\|_2^2 \\
&\leq \textstyle\sum_j \left\|C\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t_{ij}) - y_{ij}\right\|_2^2 + \left\|\overline{u}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}\right\|_{U,L^2}^2 + \left\|\Delta(\overline{\delta})\widehat{b}_i(\overline{\eta})\right\|_2^2 \\
&= \min_{b_i}\left\{\left\|\Delta(\overline{\delta})b_i\right\|_2^2 + \min_{x_{0,i}^u}\min_{u_i}\left\{\textstyle\sum_j\left\|CX_{\overline{\theta},b_i,x_{0,i},u_i}(t_{ij}) - y_{ij}\right\|_2^2 + \left\|u_i\right\|_{U,L^2}^2\right\}\right\} \\
&\leq \left\|\Delta(\overline{\delta})b_i^*\right\|_2^2 + \textstyle\sum_j\left\|CX_{\overline{\theta},b_i^*,x_{0,i}^*}(t_{ij}) - y_{ij}\right\|_2^2
\end{aligned}
$$

and as in the previous lemma, we derive from this inequality that $\mathbb{E}\left[\left\|\overline{u}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}\right\|_{L^2}^2\right]$ and $\mathbb{E}\left[\left\|\widehat{b}_i(\overline{\eta})\right\|_2^2\right]$ are finite. We bound now $\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}$:

$$
\begin{aligned}
&\left\|\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t) - X_{\overline{\theta},\widehat{b}_i(\overline{\eta}),\widehat{x_{0,i}}}(t)\right\|_2^2 \\
&= \left\|\textstyle\int_0^t f_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(s,\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(s)) - f_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(s,X_{\overline{\theta},\widehat{b}_i(\overline{\eta}),\widehat{x_{0,i}}}(s))ds + \textstyle\int_0^t \overline{u}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(s)ds\right\| \\
&\leq 2\textstyle\int_0^t \left\|f_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(s,\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(s)) - f_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(s,X_{\overline{\theta},\widehat{b}_i(\overline{\eta}),\widehat{x_{0,i}}}(s))\right\|_2^2 ds + 2\left\|\overline{u}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}\right\|_{L^2}^2 \\
&\leq 2\overline{\partial_x f}\textstyle\int_0^t \left\|\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(s) - X_{\overline{\theta},\widehat{b}_i(\overline{\eta}),\widehat{x_{0,i}}}(s)\right\|_2^2 ds + 2\left\|\overline{u}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}\right\|_{L^2}^2
\end{aligned}
$$

and by using the Gronwall lemma, we derive the bound $\left\|\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t) - X_{\overline{\theta},\widehat{b}_i(\overline{\eta}),\widehat{x_{0,i}}}(t)\right\|_2^2 \leq 2\left\|\overline{u}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}\right\|_{L^2}^2 e^{2\overline{\partial_x f}T}$. Now by using triangular inequality $\left\|\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t)\right\|_2^2 \leq \left\|\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t) - X_{\overline{\theta},\widehat{b}_i(\overline{\eta}),\widehat{x_{0,i}}}(t)\right\|_2^2 + \left\|X_{\overline{\theta},\widehat{b}_i(\overline{\eta}),\widehat{x_{0,i}}}(t)\right\|_2^2$

and condition 4 we derive $\mathbb{E}\left[\left\|\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t)\right\|_2^2\right]$ is finite. Now, we derive the expression of $\frac{dh}{d(\theta,\delta)}$, by differentiation we got:

$$\begin{aligned}
\frac{dh}{d\theta}(\widehat{b}_i(\overline{\eta}),\overline{\eta},\mathbf{y_i}) &= \frac{d}{d\theta}\left(\left\|\Delta(\overline{\delta})\widehat{b}_i(\overline{\eta})\right\|_2^2+\sum_j\left\|C\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t_{ij})-y_{ij}\right\|_2^2\right)\\
&= 2\frac{\partial\widehat{b}_i(\overline{\eta})}{\partial\theta}^T\Delta(\overline{\delta})^T\Delta(\overline{\delta})\widehat{b}_i(\overline{\eta})+\sum_j\left(C\frac{d}{d\theta}\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t_{ij})\right)^T\left(C\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t_{ij})-y_{ij}\right)
\end{aligned}$$

and similarly

$$\begin{aligned}
\frac{dh}{d\delta}(\widehat{b}_i(\overline{\eta}),\overline{\eta},\mathbf{y_i}) &= \frac{d}{d\delta}\left(\left\|\Delta(\overline{\delta})\widehat{b}_i(\overline{\eta})\right\|_2^2+\sum_j\left\|C\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t_{ij})-y_{ij}\right\|_2^2\right)\\
&= 2\frac{\partial\Delta(\overline{\delta})\widehat{b}_i(\overline{\eta})}{\partial\delta}^T\Delta(\overline{\delta})\widehat{b}_i(\overline{\eta})+2\sum_j\left(C\frac{d}{d\delta}\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t_{ij})\right)^T\left(C\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t_{ij})-y_{ij}\right)
\end{aligned}$$

with

$$\begin{aligned}
\frac{d}{d\theta}\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t) &= \frac{\partial}{\partial\theta}\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t)+\frac{\partial}{\partial b_i}\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t)\frac{\partial}{\partial\theta}\widehat{b}_i(\overline{\eta})\\
&+ \frac{\partial}{\partial x_0}\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t)\left(\frac{\partial}{\partial\theta}\widehat{x_0}(\overline{\theta},\widehat{b_i}\,(\overline{\eta}))+\frac{\partial}{\partial b_i}\widehat{x_0}(\overline{\theta},\widehat{b_i}\,(\overline{\eta}))\frac{\partial\widehat{b}_i(\overline{\eta})}{\partial\theta}\right)
\end{aligned}$$

and $\frac{d}{d\delta}\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t)=\frac{\partial}{\partial b_i}\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t)\frac{\partial}{\partial\delta}\widehat{b}_i(\overline{\eta})+\frac{\partial}{\partial x_0}\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t)\frac{\partial}{\partial b_i}\widehat{x_0}(\overline{\theta},\widehat{b_i}\,(\overline{\eta}))\frac{\partial\widehat{b}_i(\overline{\eta})}{\partial\delta}$. From this, it is clear we need prove that $\mathbb{E}\left[\left\|\frac{\partial}{\partial\theta}\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t)\right\|_2^2\right]$, $\mathbb{E}\left[\left\|\frac{\partial}{\partial b_i}\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t)\right\|_2^2\right]$, $\mathbb{E}\left[\left\|\frac{\partial}{\partial x_0}\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t)\right\|_2^2\right]$, $\mathbb{E}\left[\left\|\frac{\partial}{\partial\theta}\widehat{x_0}(\overline{\theta},\widehat{b_i}\,(\overline{\eta}))\right\|_2^2\right]$, $\mathbb{E}\left[\left\|\frac{\partial}{\partial b_i}\widehat{x_0}(\overline{\theta},\widehat{b_i}\,(\overline{\eta}))\right\|_2^2\right]$ and $\mathbb{E}\left[\left\|\frac{\partial}{\partial\eta}\widehat{b}_i(\overline{\eta})\right\|_2^2\right]$ are finite to derive the same result for $\frac{dh}{d(\theta,\delta)}(\widehat{b}_i(\overline{\eta}),\overline{\eta},\mathbf{y_i})$. For this, we firstly derive a suitable expression for $\frac{\partial}{\partial\eta}\widehat{b}_i(\overline{\eta})$. Thanks to condition 6 and 7, we can use the implicit function theorem which gives us $\frac{\partial}{\partial\eta}\widehat{b}_i(\overline{\eta})=-\frac{\partial^2}{\partial^2 b_i}g_i(\widehat{b}_i(\overline{\eta})\mid\overline{\eta},U)^{-1}\frac{\partial}{\partial\eta}\nabla_{b_i}g_i(\widehat{b}_i(\overline{\eta})\mid\overline{\eta},U)$ with

$$\begin{aligned}
&\frac{1}{2}\nabla_{b_i}g_i(b_i\mid\eta,U)=\Delta(\overline{\delta})^T\Delta(\overline{\delta})b_i\\
&+\sum_j\left(\frac{d}{db_i}C\overline{X}_{\theta,b_i}(t_{ij})\right)^T\left(C\overline{X}_{\theta,b_i}(t_{ij})-y_{ij}\right)+\int_0^T\left(\frac{d}{db_i}\overline{u}_{\theta,b_i}(t_{ij})\right)^T\overline{u}_{\theta,b_i}(t)dt\\
&\frac{1}{2}\frac{\partial^2}{\partial^2 b_i}g_i(b_i\mid\eta,U)=\sum_j\left(\frac{d^2}{d^2b_i}C\overline{X}_{\theta,b_i}(t_{ij})\right)^T\otimes\left(C\overline{X}_{\theta,b_i}(t_{ij})-y_{ij}\right)\\
&+\sum_j\left(C\frac{d}{db_i}\overline{X}_{\theta,b_i}(t_{ij})\right)^T C\frac{d}{db_i}\overline{X}_{\theta,b_i}(t_{ij})\\
&+\int_0^T\left(\frac{d^2}{d^2b_i}\overline{u}_{\theta,b_i}(t)\right)^T\otimes\overline{u}_{\theta,b_i}(t)dt+\int_0^T\left(\frac{d}{db_i}\overline{u}_{\theta,b_i}(t)\right)^T\frac{d}{db_i}\overline{u}_{\theta,b_i}(t)dt+\Delta(\overline{\delta})^T\Delta(\overline{\delta})\\
&\frac{1}{2}\frac{\partial}{\partial\delta}\nabla_{b_i}g_i(b_i\mid\eta,U)=\frac{\partial}{\partial\delta}\left(\Delta(\overline{\delta})^T\Delta(\overline{\delta})b_i\right)\\
&\frac{1}{2}\frac{\partial}{\partial\theta}\nabla_{b_i}g_i(b_i\mid\eta,U)=\sum_j\left(\frac{d^2}{d\theta db_i}C\overline{X}_{\theta,b_i}(t_{ij})\right)^T\otimes\left(C\overline{X}_{\theta,b_i}(t_{ij})-y_{ij}\right)\\
&+\sum_j\left(C\frac{d}{db_i}\overline{X}_{\theta,b_i}(t_{ij})\right)^T\left(C\frac{d}{d\theta}\overline{X}_{\theta,b_i}(t_{ij})\right)\\
&+\int_0^T\left(\left(\frac{d^2}{d\theta db_i}\overline{u}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t)\right)^T\otimes\overline{u}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}(t)+\left(\frac{d}{db_i}\overline{u}_{\theta,b_i}\right)^T\frac{d}{d\theta}\overline{u}_{\theta,b_i}(t)\right)d
\end{aligned}$$

and $\frac{d}{db_i}\overline{X}_{\theta,b_i}(t)=\frac{\partial}{\partial b_i}\overline{X}_{\theta,b_i,\widehat{x_0}(\theta,b_i)}(t)+\frac{\partial}{\partial x_0}\overline{X}_{\theta,b_i,\widehat{x_0}(\theta,b_i)}(t)\frac{\partial\widehat{x_0}(\theta,b_i)}{\partial b_i}$, $\frac{d}{db_i}\overline{u}_{\theta,b_i}(t)=\frac{\partial}{\partial b_i}\overline{u}_{\theta,b_i,\widehat{x_0}(\theta,b_i)}(t)+\frac{\partial}{\partial x_0}\overline{u}_{\theta,b_i,\widehat{x_0}(\theta,b_i)}(t)\frac{\partial\widehat{x_0}(\theta,b_i)}{\partial b_i}$, $\frac{d}{d\theta}\overline{X}_{\theta,b_i}(t)=\frac{\partial}{\partial\theta}\overline{X}_{\theta,b_i,\widehat{x_0}(\theta,b_i)}(t)+\frac{\partial}{\partial x_0}\overline{X}_{\theta,b_i,\widehat{x_0}(\theta,b_i)}(t)\frac{\partial\widehat{x_0}(\theta,b_i)}{\partial\theta}$ and $\frac{d}{d\theta}\overline{u}_{\theta,b_i}(t)=\frac{\partial}{\partial\theta}\overline{u}_{\theta,b_i,\widehat{x_0}(\theta,b_i)}(t)+\frac{\partial}{\partial x_0}\overline{u}_{\theta,b_i,\widehat{x_0}(\theta,b_i)}(t)\frac{\partial\widehat{x_0}(\theta,b_i)}{\partial\theta}$. This lead us in turn to control the behavior of the partial derivatives $\frac{\partial}{\partial b_i}\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}$, $\frac{\partial}{\partial\theta}\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}$, $\frac{\partial^2}{\partial^2 b_i}\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}$, $\frac{\partial^2}{\partial\theta\partial b_i}\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}$, $\frac{\partial}{\partial b_i}\overline{u}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}$, $\frac{\partial}{\partial\theta}\overline{u}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}$, $\frac{\partial^2}{\partial^2 b_i}\overline{u}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}$,

$\frac{\partial^2}{\partial\theta\partial b_i}\overline{u}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}, \frac{\partial}{\partial b_i}\widehat{x_0}\left(\overline{\theta},\widehat{b}_i(\overline{\eta})\right)$, $\frac{\partial}{\partial\theta}\widehat{x_0}\left(\overline{\theta},\widehat{b}_i(\overline{\eta})\right)$, $\frac{\partial^2}{\partial^2 b_i}\widehat{x_0}\left(\overline{\theta},\widehat{b}_i(\overline{\eta})\right)$, $\frac{\partial^2}{\partial\theta\partial b_i}\widehat{x_0}\left(\overline{\theta},\widehat{b}_i(\overline{\eta})\right)$, this is done in lemma 0.6 part 3). From this, the conclusion holds as well for $\frac{\partial}{\partial\eta}\widehat{b}_i(\overline{\eta})$ composed of products and sums of random variables of finite bounded moment of order two. The same results holds then for $\frac{d}{d\theta}\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}$ and $\frac{d}{d\delta}\overline{X}_{\overline{\theta},\widehat{b}_i(\overline{\eta})}$ and we can conclude for $\frac{dh}{d(\theta,\delta)}(\widehat{b}_i(\overline{\eta}),\overline{\eta},\mathbf{y_i})$. The proof for $\frac{d^2h}{d^2(\theta,\delta)}(\widehat{b}_i(\overline{\eta}),\overline{\eta},\mathbf{y_i})$ follows the same steps. □

Lemma 0.6. *Let us assume $(\theta,b,t,x) \longmapsto f_{\theta,b}(t,x)$ is $(k+1)$ continuously differentiable and bounded on a subset $\widetilde{\Theta}\times\mathbb{R}^q\times[0,\,T]\times\widetilde{\chi}$. 1)If $K_1(U) = 2T\left\|C^TC\right\|_2^2\left\|BU^{-1}B^T\right\|e^{6T\overline{\partial_x f}} < 1$ and $K_2(U) = \frac{\left\|BU^{-1}B^T\right\|}{1-K_1(U)}4T\overline{\partial_{xx}f}\sup_{(\theta,b,x_0)}\left\|C^T\left(C\overline{X}^{(j)}_{\theta,b,x_0}(1)-y_j\right)\right\|_2^2 e^{7T\overline{\partial_x f}} < 1$, then $\overline{u}_{\theta,b,x_0} = \arg\min_u\left\{\sum_j\|CX_{\theta,b,x_0}(t_j)-y_j\|_2^2+\|u\|^2_{U,L^2}\right\}$ and the related optimal trajectory $\overline{X}_{\theta,b,x_0}$ are k continuously differentiable on $\widetilde{\Theta}\times\mathbb{R}^q\times[0,\,T]\times\widetilde{\chi}$.*

2) For every $\left(\widetilde{\theta},\widetilde{b}_i\right)\in\widetilde{\Theta}\times\mathbb{R}^q$ such that $\frac{\partial^2\mathcal{C}_i}{\partial^2 b_i}(\widetilde{b}_i,\widehat{x_{i,0}}(\widetilde{\theta},\widetilde{b}_i),\overline{u}_{\widetilde{\theta},\widetilde{b}_i,\widehat{x_{i,0}}(\widetilde{\theta},\widetilde{b}_i)}\mid\widetilde{\theta},U)$ is of full rank and $\widehat{x_{i,0}}(\widetilde{\theta},\widetilde{b}_i)\in\widetilde{\chi}$, there is a neighborhood of $\widetilde{\Theta}_{\widetilde{\theta}}\times\widetilde{\Theta}_{\widetilde{b_i}}$ such that $(\theta,b_i)\longmapsto\widehat{x_{i,0}}\left(\theta,b_i\right)$, $\overline{X}_{\theta,b}$ and $\overline{u}_{\theta,b}$ are $(k-1)$ differentiable on $\widetilde{\Theta}_{\widetilde{\theta}}\times\widetilde{\Theta}_{\widetilde{b_i}}$.

3)If $K_3(U) = 2\left\|C^TC\right\|_2^2\left\|BU^{-1}B\right\|_2^2e^{2T\overline{\partial_x f}} < 1$, $K_4(U) = \frac{4T\left\|BU^{-1}B\right\|_2^2e^{6T\overline{\partial_x f}}}{1-K_3(U)}\left(2\left\|C^TC\right\|_2^2\overline{\partial_x f}+\frac{\overline{\partial_{xx}f}}{|p^0|}\right) < 1$ and $K_5(U) = \frac{4T}{1-K_3(U)}\left\|BU^{-1}B\right\|_2^2\left\|C^TC\right\|_2^2\overline{\partial_x f}e^{6T\overline{\partial_x f}} < 1$, then $\frac{\partial^k}{\partial^k(\theta,b)}\overline{X}_{\overline{\theta},b(\overline{\theta},\triangle(\overline{\delta}))}(t)$, $\frac{\partial^k}{\partial^k(\theta,b)}\overline{u}_{i,\overline{\theta},b(\overline{\theta},\triangle(\overline{\delta}))}(t)$ and $\frac{\partial^k}{\partial^k(\theta,b)}\widehat{x_{i,0}}\left(\overline{\theta},b(\overline{\theta},\triangle(\overline{\delta}))\right)$ have finite moment of order two.

Proof. We focus here on one subject, so we omit the dependence in i. Let us introduce the n_i state variables $X^{(j)}_{\theta,b,x_0}(t) = X_{\theta,b,x_0}(t_j+t(t_{j+1}-t_j))$ such that $X^{(j)}_{\theta,b,x_0}(0) = X_{\theta,b,x_0}(t_j)$ and $X^{(j)}_{\theta,b,x_0}(1) = X_{\theta,b,x_0}(t_j)$ to reformulate our control problem as a final time one:

$$\min_u\left\{\sum_j\left\|CX^{(j)}_{\theta,b,x_0}(1)-y_j\right\|_2^2+\sum_j\int_0^1 u^{(j)}(t)Uu^{(j)}(t)dt\right\}$$
$$\begin{cases}\frac{d}{dt}X^{(j)}_{\theta,b,x_0} = (t_{j+1}-t_j)f_{\theta,b}(t_j+t(t_{j+1}-t_j),X^{(j)}_{\theta,b,x_0})+Bu^{(j)}(t)\\ X^{(0)}_{\theta,b,x_0}(0) = x_0,\ X^{(j)}_{\theta,b,x_0}(0) = X^{(j-1)}_{\theta,b,x_0}(1)\\ u^{(j)}(0) = u^{(j-1)}(1).\end{cases}$$

Now we use the Pontryagin maximum principle [5] to characterize the dependence of the optimal control and trajectory with respect to (θ,b,x_0) and measurement noise. We formulate the Hamiltonian of the controlled system: $H_{\theta,b}(t,x,p,p^0,u) = \sum_j\left(p^{(j)}\right)^T\left((t_{j+1}-t_j)f_{\theta,b}(t_j+t(t_{j+1}-t_j),x^{(j)})+Bu^{(j)}(t)\right)+p^0\left(u^{(j)}\right)^TUu^{(j)}$ and derive the conjugate equation:

$$\begin{aligned}\dot{x}^{(j)} &= \frac{\partial H_{\theta,b}}{\partial p^{(j)}}(t,x,p,p^0,u) = (t_{j+1}-t_j)f_{\theta,b}(t_j+t(t_{j+1}-t_j),x^{(j)})+Bu^{(j)}(t)\\ \dot{p}^{(j)} &= -\frac{\partial H_{\theta,b}}{\partial x^{(j)}}(t,x,p,p^0,u) = -(t_{j+1}-t_j)\frac{\partial f_{\theta,b}}{\partial x}(t_j+t(t_{j+1}-t_j),x^{(j)})^Tp^{(j)}.\end{aligned}$$

The maximum condition on the Hamiltonian $H_{\theta,b}(t,x(t),p(t),p^0,\overline{u_{\theta,b}}(t)) = \max_u H_{\theta,b}(t,x(t),p(t),p^0,u)$

allows us to derive the following expression for the optimal control:

$$\nabla_{u^{(j)}} H_{\theta,b}(t, x(t), p(t), p^0, \overline{u}^{(j)}_{\theta,b,x_0}(t)) = B^T p^{(j)}(t) + 2p^0 U \overline{u}^{(j)}_{\theta,b,x_0}(t) = 0 \iff \overline{u}^{(j)}_{\theta,b,x_0}(t) = -\frac{U^{-1}B^T}{2p^0} p^{(j)}(t).$$

Now regarding the final time value for the adjoint system imposed by the transversality conditions, we get $p^j(1) = 2p^0 C^T \left(C\overline{X}^{(j)}_{\theta,b,x_0}(1) - y_j \right)$. We end up with following boundary value problem:

$$\begin{cases} \frac{d}{dt}\overline{X}^{(j)}_{\theta,b,x_0} = (t_{j+1} - t_j) f_{\theta,b}(t_j + t(t_{j+1} - t_j), \overline{X}^{(j)}_{\theta,b,x_0}) - \frac{1}{2p^0} BU^{-1}B^T p^{(j)}_{\theta,b,x_0}(t) \\ \frac{d}{dt}p^{(j)}_{\theta,b,x_0} = -(t_{j+1} - t_j) \frac{\partial f_{\theta,b}}{\partial x}(t_j + t(t_{j+1} - t_j), \overline{X}^{(j)}_{\theta,b,x_0})^T p^{(j)}_{\theta,b,x_0} \\ X^{(0)}_{\theta,b,x_0}(0) = x_0,\ X^{(j)}_{\theta,b,x_0}(0) = X^{(j-1)}_{\theta,b,x_0}(1) \\ p^{(j)}_{\theta,b,x_0}(1) = 2p^0 C^T \left(C\overline{X}^{(j)}_{\theta,b,x_0}(1) - y_j \right) \\ \overline{u}^{(j)}_{\theta,b,x_0}(t) = -\frac{1}{2p^0} U^{-1} B^T p^{(j)}_{\theta,b,x_0}(t),\ u^{(j)}_{\theta,b,x_0}(0) = u^{(j-1)}_{\theta,b,x_0}(1). \end{cases}$$

So, for two set of parameters (θ, b, x_0) and $\left(\theta', b', x_0^{'}\right)$ we get the inequality:

$$\begin{aligned} & \left\| \overline{X}^{(j)}_{\theta,b,x_0}(t) - \overline{X}^{(j)}_{\theta',b',x_0'}(t) \right\|_2^2 \\ & \leq 8T \left(\overline{\partial_b f} \left\| b - b' \right\|_2^2 + \overline{\partial_\theta f} \left\| \theta - \theta' \right\|_2^2 \right) + 4\overline{\partial_x f} T \int_0^t \left\| \overline{X}^{(j)}_{\theta,b,x_0}(s) - \overline{X}^{(j)}_{\theta',b',x_0'}(s) \right\|_2^2 ds \\ & + \left\| \frac{BU^{-1}B^T}{p^0} \right\| \int_0^t \left\| p^{(j)}_{\theta,b,x_0}(s) - p^{(j)}_{\theta',b',x_0'}(s) \right\|_2^2 ds + \left\| \overline{X}^{(j)}_{\theta,b,x_0}(0) - \overline{X}^{(j)}_{\theta',b',x_0'}(0) \right\|_2^2 \\ & \left\| p^{(j)}_{\theta,b,x_0}(t) - p^{(j)}_{\theta',b',x_0'}(t) \right\|_2^2 \\ & \leq 2 \left\| p^0 C^T C \right\|_2^2 \left\| \overline{X}^{(j)}_{\theta,b,x_0}(1) - \overline{X}^{(j)}_{\theta',b',x_0'}(1) \right\|_2^2 + 2T\overline{\partial_x f} \int_t^1 \left\| p^{(j)}_{\theta',b',x_0'}(s) - p^{(j)}_{\theta,b,x_0}(s) \right\|_2^2 ds \\ & + 8T \left(\overline{\partial_{bx} f} \left\| b - b' \right\|_2^2 + \overline{\partial_{\theta x} f} \left\| \theta - \theta' \right\|_2^2 \right) \\ & + 4T\overline{\partial_{xx} f} \int_t^1 \left\| p^{(j)}_{\theta',b',x_0'}(s) \right\|_2^2 \left\| \overline{X}^{(j)}_{\theta,b,x_0}(s) - \overline{X}^{(j)}_{\theta',b',x_0'}(s) \right\|_2^2 ds. \end{aligned}$$

which leads by using Gronwall lemma:

$$\begin{aligned} & \left\| \overline{X}^{(j)}_{\theta,b,x_0}(t) - \overline{X}^{(j)}_{\theta',b',x_0'}(t) \right\|_2^2 \\ & \leq \left\| \frac{BU^{-1}B^T}{p^0} \right\| \left\| p^{(j)}_{\theta,b,x_0} - p^{(j)}_{\theta',b',x_0'} \right\|_{L^2} e^{4T\overline{\partial_x f}} \\ & + \left(8T \left(\overline{\partial_b f} \left\| b - b' \right\|_2^2 + \overline{\partial_\theta f} \left\| \theta - \theta' \right\|_2^2 \right) + \left\| \overline{X}^{(j)}_{\theta,b,x_0}(0) - \overline{X}^{(j)}_{\theta',b',x_0'}(0) \right\|_2^2 \right) e^{4T\overline{\partial_x f}} \end{aligned}$$

by re-injecting in $\left\|p^{(j)}_{\theta,b,x_0}(t)-p^{(j)}_{\theta',b',x'_0}(t)\right\|_2^2$ and using Cauchy-Schwarz inequality we get:

$$\begin{aligned}
\left\|p^{(j)}_{\theta,b,x_0}(t)-p^{(j)}_{\theta',b',x'_0}(t)\right\|_2^2 &\leq 2\left\|C^TC\right\|_2^2\left\|BU^{-1}B^T\right\|\left\|p^{(j)}_{\theta,b,x_0}-p^{(j)}_{\theta',b',x'_0}\right\|_{L^2}e^{4T\overline{\partial_x f}}\\
&+ 2\left\|p^0C^TC\right\|_2^2 8T\left(\overline{\partial_b f}\left\|b-b'\right\|_2^2+\overline{\partial_\theta f}\left\|\theta-\theta'\right\|_2^2\right)e^{4T\overline{\partial_x f}}\\
&+ 2\left\|p^0C^TC\right\|_2^2\left\|\overline{X}^{(j)}_{\theta,b,x_0}(0)-\overline{X}^{(j)}_{\theta',b',x'_0}(0)\right\|_2^2 e^{4T\overline{\partial_x f}}\\
&+ 4T\overline{\partial_{xx} f}\int_t^1\left\|p^{(j)}_{\theta',b',x'_0}(s)\right\|_2^2 ds\int_t^1\left\|\overline{X}^{(j)}_{\theta,b,x_0}(s)-\overline{X}^{(j)}_{\theta',b',x'_0}(s)\right\|_2^2 ds\\
&+ 8T\left(\overline{\partial_{bx} f}\left\|b-b'\right\|_2^2+\overline{\partial_{\theta x} f}\left\|\theta-\theta'\right\|_2^2\right)\\
&+ 2T\overline{\partial_x f}\int_t^1\left\|p^{(j)}_{\theta',b',x'_0}(s)-p^{(j)}_{\theta,b,x_0}(s)\right\|_2^2 ds
\end{aligned}$$

in the same way, we obtain:

$$\begin{aligned}
\left\|p^{(j)}_{\theta',b',x'_0}(t)\right\|_2^2 &= \left\|p^{(j)}_{\theta',b',x'_0}(1)+\int_t^1(t_{j+1}-t_j)\frac{\partial f_{\theta',b'}}{\partial x}(t_j+s(t_{j+1}-t_j),\overline{X}^{(j)}_{\theta',b',x'_0})^Tp^{(j)}_{\theta',b',x'_0}(s)ds\right\|_2^2\\
&\leq \left\|p^0C^T\left(C\overline{X}^{(j)}_{\theta',b',x'_0}(1)-y_j\right)\right\|_2^2+T\overline{\partial_x f}\int_t^1\left\|p^{(j)}_{\theta',b',x'_0}(s)\right\|_2^2 ds
\end{aligned}$$

and $\left\|p^{(j)}_{\theta',b',x'_0}(t)\right\|_2^2\leq\sup_{(\theta,b,x_0)}\left\|p^0C^T\left(C\overline{X}^{(j)}_{\theta,b,x_0}(1)-y_j\right)\right\|_2^2 e^{T\overline{\partial_x f}}$ so the inequality for $\left\|p^{(j)}_{\theta,b,x_0}(t)-p^{(j)}_{\theta',b',x'_0}(t)\right\|_2^2$ can be re-expressed as:

$$\begin{aligned}
\left\|p^{(j)}_{\theta,b,x_0}(t)-p^{(j)}_{\theta',b',x'_0}(t)\right\|_2^2 &\leq 2\left\|C^TC\right\|_2^2\left\|BU^{-1}B^T\right\|\left\|p^{(j)}_{\theta,b,x_0}-p^{(j)}_{\theta',b',x'_0}\right\|_{L^2}e^{4T\overline{\partial_x f}}\\
&+ 2\left\|p^0C^TC\right\|_2^2\left(\left\|\overline{X}^{(j)}_{\theta,b,x_0}(0)-\overline{X}^{(j)}_{\theta',b',x'_0}(0)\right\|_2^2\right)e^{4T\overline{\partial_x f}}\\
&+ 16\left\|p^0C^TC\right\|_2^2 T\left(\overline{\partial_b f}\left\|b-b'\right\|_2^2+\overline{\partial_\theta f}\left\|\theta-\theta'\right\|_2^2\right)e^{4T\overline{\partial_x f}}\\
&+ 4T\overline{\partial_{xx} f}\sup_{(\theta,b,x_0)}\left\|p^0C^T\left(C\overline{X}^{(j)}_{\theta,b,x_0}(1)-y_j\right)\right\|_2^2\left\|\overline{X}^{(j)}_{\theta,b,x_0}-\overline{X}^{(j)}_{\theta',b',x'_0}\right\|_{L^2}^2 e^{T\overline{\partial_x f}}\\
&+ 8T\left(\overline{\partial_{bx} f}\left\|b-b'\right\|_2^2+\overline{\partial_{\theta x} f}\left\|\theta-\theta'\right\|_2^2\right)+2T\overline{\partial_x f}\int_t^T\left\|p^{(j)}_{\theta',b',x'_0}(s)-p^{(j)}_{\theta,b,x_0}(s)\right\|_2^2 ds
\end{aligned}$$

and by using Gronwall lemma:

$$\begin{aligned}
\left\|p^{(j)}_{\theta,b,x_0}(t)-p^{(j)}_{\theta',b',x'_0}\right\|_{L^2}^2 &\leq 2T\left\|C^TC\right\|_2^2\left\|BU^{-1}B^T\right\|\left\|p^{(j)}_{\theta,b,x_0}-p^{(j)}_{\theta',b',x'_0}\right\|_{L^2}e^{6T\overline{\partial_x f}}\\
&+ 16T\left\|p^0C^TC\right\|_2^2\left(\overline{\partial_b f}\left\|b-b'\right\|_2^2+\overline{\partial_\theta f}\left\|\theta-\theta'\right\|_2^2\right)e^{6T\overline{\partial_x f}}\\
&+ 2\left\|p^0C^TC\right\|_2^2\left\|\overline{X}^{(j)}_{\theta,b,x_0}(0)-\overline{X}^{(j)}_{\theta',b',x'_0}(0)\right\|_2^2 e^{6T\overline{\partial_x f}}\\
&+ 8T\left(\overline{\partial_{bx} f}\left\|b-b'\right\|_2^2+\overline{\partial_{\theta x} f}\left\|\theta-\theta'\right\|_2^2\right)e^{6T\overline{\partial_x f}}\\
&+ 4T\overline{\partial_{xx} f}\sup_{(\theta,b,x_0)}\left\|p^0C^T\left(C\overline{X}^{(j)}_{\theta,b,x_0}(1)-y_j\right)\right\|_2^2\left\|\overline{X}^{(j)}_{\theta,b,x_0}-\overline{X}^{(j)}_{\theta',b',x'_0}\right\|_{L^2}^2 e^{3T\overline{\partial_x f}}
\end{aligned}$$

so if U is chosen such that $K_1(U) = 2T \left\|C^T C\right\|_2^2 \left\|BU^{-1}B^T\right\| e^{6T\overline{\partial_x f}} < 1$, we derive:

$$\begin{aligned}\left\|p^{(j)}_{\theta,b,x_0} - p^{(j)}_{\theta',b',x_0'}\right\|^2_{L^2} \quad &\leq \quad \frac{2\left\|p^0 C^T C\right\|_2^2\left(8T\left(\overline{\partial_b f}\left\|b-b'\right\|_2^2+\overline{\partial_\theta f}\left\|\theta-\theta'\right\|_2^2\right)+\left\|\overline{X}^{(j)}_{\theta,b,x_0}(0)-\overline{X}^{(j)}_{\theta',b',x_0'}(0)\right\|_2^2\right)e^{6T\overline{\partial_x f}}}{1-K_1(U)} \\ &+ \quad \frac{8T\left(\overline{\partial_{bx} f}\left\|b-b'\right\|_2^2+\overline{\partial_{\theta x} f}\left\|\theta-\theta'\right\|_2^2\right)e^{6T\overline{\partial_x f}}}{1-K_1(U)} \\ &+ \quad \frac{4T\overline{\partial_{xx} f}\sup_{(\theta,b,x_0)}\left\|p^0 C^T\left(C\overline{X}^{(j)}_{\theta,b,x_0}(1)-y_j\right)\right\|_2^2\left\|\overline{X}^{(j)}_{\theta,b,x_0}-\overline{X}^{(j)}_{\theta',b',x_0'}\right\|^2_{L^2}e^{3T\overline{\partial_x f}}}{1-K_1(U)}\end{aligned}$$

by re-injecting this expression in the inequality ruling $\left\|\overline{X}^{(j)}_{\theta,b,x_0}(t) - \overline{X}^{(j)}_{\theta',b',x_0'}(t)\right\|_2^2$, we get:

$$\begin{aligned}&\left\|\overline{X}^{(j)}_{\theta,b,x_0}(t) - \overline{X}^{(j)}_{\theta',b',x_0'}(t)\right\|_2^2 \\ &\leq \left\|BU^{-1}B^T\right\| \frac{2\left\|C^T C\right\|_2^2\left(8T\left(\overline{\partial_b f}\left\|b-b'\right\|_2^2+\overline{\partial_\theta f}\left\|\theta-\theta'\right\|_2^2\right)+\left\|\overline{X}^{(j)}_{\theta,b,x_0}(0)-\overline{X}^{(j)}_{\theta',b',x_0'}(0)\right\|_2^2\right)}{1-K_1(U)}e^{10T\overline{\partial_x f}} \\ &+ \left\|\frac{BU^{-1}B^T}{p^0}\right\| \frac{8T\left(\overline{\partial_{bx} f}\left\|b-b'\right\|_2^2+\overline{\partial_{\theta x} f}\left\|\theta-\theta'\right\|_2^2\right)}{1-K_1(U)}e^{10T\overline{\partial_x f}} \\ &+ \left\|BU^{-1}B^T\right\| \frac{4T\overline{\partial_{xx} f}\sup_{(\theta,b,x_0)}\left\|C^T\left(C\overline{X}^{(j)}_{\theta,b,x_0}(1)-y_j\right)\right\|_2^2\left\|\overline{X}^{(j)}_{\theta,b,x_0}-\overline{X}^{(j)}_{\theta',b',x_0'}\right\|^2_{L^2}}{1-K_1(U)}e^{7T\overline{\partial_x f}} \\ &+ \left(8T\left(\overline{\partial_b f}\left\|b-b'\right\|_2^2 + \overline{\partial_\theta f}\left\|\theta-\theta'\right\|_2^2\right) + \left\|\overline{X}^{(j)}_{\theta,b,x_0}(0) - \overline{X}^{(j)}_{\theta',b',x_0'}(0)\right\|_2^2\right)e^{4T\overline{\partial_x f}}\end{aligned}$$

now if U is chosen such that $K_2(U) = \frac{\left\|BU^{-1}B^T\right\|}{1-K_1(U)} 4T\overline{\partial_{xx} f}\sup_{(\theta,b,x_0)}\left\|C^T\left(C\overline{X}^{(j)}_{\theta,b,x_0}(1) - y_j\right)\right\|_2^2 e^{7T\overline{\partial_x f}} < 1$, then the previous inequality becomes

$$\begin{aligned}&\left\|\overline{X}^{(j)}_{\theta,b,x_0} - \overline{X}^{(j)}_{\theta',b',x_0'}\right\|^2_{L^2} \\ &\leq \left\|BU^{-1}B^T\right\| \frac{2\left\|C^T C\right\|_2^2\left(8T\left(\overline{\partial_b f}\left\|b-b'\right\|_2^2+\overline{\partial_\theta f}\left\|\theta-\theta'\right\|_2^2\right)+\left\|\overline{X}^{(j)}_{\theta,b,x_0}(0)-\overline{X}^{(j)}_{\theta',b',x_0'}(0)\right\|_2^2\right)e^{12T^2\overline{\partial_x f}}}{(1-K_2(U))(1-K_1(U))} \\ &+ \left\|\frac{BU^{-1}B^T}{p^0}\right\| \frac{8T\left(\overline{\partial_{bx} f}\left\|b-b'\right\|_2^2+\overline{\partial_{\theta x} f}\left\|\theta-\theta'\right\|_2^2\right)}{(1-K_2(U))(1-K_1(U))}e^{10T\overline{\partial_x f}} \\ &+ \frac{\left(8T\left(\overline{\partial_b f}\left\|b-b'\right\|_2^2+\overline{\partial_\theta f}\left\|\theta-\theta'\right\|_2^2\right)+\left\|\overline{X}^{(j)}_{\theta,b,x_0}(0)-\overline{X}^{(j)}_{\theta',b',x_0'}(0)\right\|_2^2\right)}{1-K_2(U)}e^{4T\overline{\partial_x f}}.\end{aligned}$$

Since $\overline{X}^{(0)}_{\theta,b,x_0}(0) - \overline{X}^{(0)}_{\theta',b',x_0'}(0) = x_0 - x_0'$, we see from the previous inequality, it exists constants K_b, K_θ and K_{x_0} such that $\left\|\overline{X}^{(0)}_{\theta,b,x_0} - \overline{X}^{(0)}_{\theta',b',x_0'}\right\|^2_{L^2} \leq K_b\left\|b-b'\right\|_2^2 + K_\theta\left\|\theta-\theta'\right\|_2^2 + K_{x_0}\left\|x_0 - x_0'\right\|_2^2$ and so we derive the uniform continuity of $(\theta, b, x_0) \longrightarrow \left\|\overline{X}^{(0)}_{\theta,b,x_0}\right\|^2_{L^2}$ which in turn leads to the continuity of $(\theta, b, x_0) \longrightarrow \left\|p^{(0)}_{\theta,b,x_0}\right\|^2_{L^2}$. The differential constraints imposed by the conjugate equations turn this continuity in L^2-norm into the continuity of $(\theta, b, x_0, t) \longrightarrow \overline{X}^{(0)}_{\theta,b,x_0}(t)$ and $(\theta, b, x_0) \longrightarrow p^{(0)}_{\theta,b,x_0}(t)$. Now, for the general case of $j > 0$, we notice from the previous inequalities that $\left\|\overline{X}^{(j)}_{\theta,b,x_0} - \overline{X}^{(j)}_{\theta',b',x_0'}\right\|^2_{L^2} \leq K_b\left\|b-b'\right\|_2^2 + K_\theta\left\|\theta-\theta'\right\|_2^2 + K_{x_0}\left\|\overline{X}^{(j-1)}_{\theta,b,x_0}(1) - \overline{X}^{(j-1)}_{\theta',b',x_0'}(1)\right\|_2^2$ and $\left\|p^{(j)}_{\theta,b,x_0} - p^{(j)}_{\theta,b,x_0}\right\|^2_{L^2} \leq K_b'\left\|b-b'\right\|_2^2 + K_\theta'\left\|\theta-\theta'\right\|_2^2 + K_{x_1}'\left\|\overline{X}^{(j)}_{\theta,b,x_0}(1) - \overline{X}^{(j)}_{\theta',b',x_0'}(1)\right\|^2_{L^2} + K_{x_0}'\left\|\overline{X}^{(j-1)}_{\theta,b,x_0}(1) - \overline{X}^{(j-1)}_{\theta',b',x_0'}(1)\right\|$ since $\overline{X}^{(j)}_{\theta,b,x_0}(0) = \overline{X}^{(j-1)}_{\theta,b,x_0}(1)$. So, we can use recursive argument to conclude about the continuity of $(\theta, b, x_0, t) \longrightarrow \overline{X}^{(j)}_{\theta,b,x_0}(t)$ and $(\theta, b, x_0) \longrightarrow p^{(j)}_{\theta,b,x_0}(t)$ for all j, since the property holds for $j = 0$.

The dependence between $\overline{u}^{(j)}_{\theta,b,x_0}$ and $p^{(j)}_{\theta,b,x_0}$ allows us to conclude for the continuity of $\overline{u}^{(j)}_{\theta,b,x_0}$ as well. The proof remains the same for higher order of differentiation and thus it is omitted.

Now regarding the second part of the lemma, let us denote: $\overline{\mathcal{C}_i}(\theta, b_i, x_{i,0}) := \mathcal{C}_i(b_i, x_{i,0}, \overline{u}_{\theta,b_i,x_{i,0}} \mid \theta, U) = \sum_j \left\|C\overline{X}_{\theta,b_i,x_{i,0}}(t_{ij}) - y_{ij}\right\|_2^2 + \left\|\overline{u}_{\theta,b_i,x_{i,0}}\right\|_{U,L^2}^2$,thanks to the first part of the lemma, we know $\overline{\mathcal{C}_i}$ is k continuously differentiable on $\widetilde{\Theta} \times \mathbb{R}^q \times \widetilde{\chi}$. Since $\nabla_{x_{i,0}}\overline{\mathcal{C}_i}(\theta, b_i, \widehat{x_{i,0}}(\theta, b_i)) = 0$, we can use the implicit function theorem to derive that for each point $\left(\widetilde{\theta}, \widetilde{b_i}\right) \in \widetilde{\Theta} \times \mathbb{R}^q$ such that $\frac{\partial^2}{\partial^2 x_{i,0}}\overline{\mathcal{C}_i}(\widetilde{\theta}, \widetilde{b_i}, \widehat{x_{i,0}}\left(\widetilde{\theta}, \widetilde{b_i}\right))$ is of full rank it exist a neighborhood of $\widetilde{\Theta}_{\widetilde{\theta}} \times \widetilde{\Theta}_{\widetilde{b_i}}$ such that $(\theta, b_i) \longmapsto \widehat{x_{i,0}}(\theta, b_i)$ is $(k-1)$ differentiable on $\widetilde{\Theta}_{\widetilde{\theta}} \times \widetilde{\Theta}_{\widetilde{b_i}}$ and moreover:

$$\frac{\partial}{\partial\left(\theta, b_i\right)}\widehat{x_{i,0}}(\theta, b_i) = -\frac{\partial^2}{\partial^2 x_{i,0}}\overline{\mathcal{C}_i}(\theta, b_i, \widehat{x_{i,0}}(\theta, b_i))^{-1}\frac{\partial}{\partial\left(\theta, b_i\right)}\nabla_{x_{i,0}}\overline{\mathcal{C}_i}(\theta, b_i, \widehat{x_{i,0}}(\theta, b_i)).$$

From this, we also derive the $(k-1)$ differentiability of $\overline{X}_{\theta,b}$ and $\overline{u}_{\theta,b_i}$ by classic results on the regularity of composed functions. Since, we get:

$$\begin{array}{rcl}
\frac{\partial^2}{\partial^2 x_{i,0}}\overline{\mathcal{C}_i}(\theta, b_i, \widehat{x_{i,0}}(\theta, b_i)) & = & \sum_j \left(C\frac{\partial^2}{\partial^2 x_{i,0}}\overline{X}_{\theta,b_i,\widehat{x_{i,0}}(\theta,b_i)}(t_{ij})\right)^T \otimes \left(C\overline{X}_{\theta,b_i,\widehat{x_{i,0}}(\theta,b_i)} - y_{ij}\right) \\
& + & 2\sum_j \left(C\frac{\partial}{\partial x_{i,0}}\overline{X}_{\theta,b_i,\widehat{x_{i,0}}(\theta,b_i)}(t_{ij})\right)^T C\frac{\partial}{\partial x_{i,0}}\overline{X}_{\theta,b_i,\widehat{x_{i,0}}(\theta,b_i)}(t_{ij}) \\
& + & 2\int_0^T \left(\frac{\partial^2}{\partial^2 x_{i,0}}\overline{u}_{\theta,b_i,\widehat{x_{i,0}}(\theta,b_i)}(t)\right)^T \otimes \overline{u}_{\theta,b_i,\widehat{x_{i,0}}(\theta,b_i)}(t)dt \\
& + & 2\int_0^T \left(\frac{\partial}{\partial x_{i,0}}\overline{u}_{\theta,b_i,\widehat{x_{i,0}}(\theta,b_i)}(t)\right)^T \frac{\partial}{\partial x_{i,0}}\overline{u}_{\theta,b_i,\widehat{x_{i,0}}(\theta,b_i)}(t)dt
\end{array}$$

and

$$\begin{array}{rcl}
\frac{\partial}{\partial(\theta,b_i)}\nabla_{x_{i,0}}\overline{\mathcal{C}_i}(\theta, b_i, \widehat{x_{i,0}}(\theta, b_i)) & = & \sum_j \left(C\frac{\partial^2}{\partial^2(\theta,b_i)\partial x_{i,0}}\overline{X}_{\theta,b_i,\widehat{x_{i,0}}(\theta,b_i)}(t_{ij})\right)^T \otimes \left(C\overline{X}_{\theta,b_i,\widehat{x_{i,0}}(\theta,b_i)} - y_{ij}\right) \\
& + & 2\sum_j \left(C\frac{\partial}{\partial(\theta,b_i)}\overline{X}_{\theta,b_i,\widehat{x_{i,0}}(\theta,b_i)}(t_{ij})\right)^T C\frac{\partial}{\partial x_{i,0}}\overline{X}_{\theta,b_i,\widehat{x_{i,0}}(\theta,b_i)}(t_{ij}) \\
& + & 2\int_0^T \left(\frac{\partial^2}{\partial(\theta,b_i)\partial x_{i,0}}\overline{u}_{\theta,b_i,\widehat{x_{i,0}}(\theta,b_i)}(t)\right)^T \otimes \overline{u}_{\theta,b_i,\widehat{x_{i,0}}(\theta,b_i)}(t)dt \\
& + & 2\int_0^T \left(\frac{\partial}{\partial(\theta,b_i)}\overline{u}_{\theta,b_i,\widehat{x_{i,0}}(\theta,b_i)}(t)\right)^T \frac{\partial}{\partial x_{i,0}}\overline{u}_{\theta,b_i,\widehat{x_{i,0}}(\theta,b_i)}(t)dt
\end{array}$$

the behavior of $\frac{\partial}{\partial(\theta,b_i)}\widehat{x_{i,0}}(\theta, b_i)$ is controlled by the partial derivatives of $\overline{X}_{\theta,b_i,\widehat{x_{i,0}}(\theta,b_i)}$ and $\overline{u}_{\theta,b_i,\widehat{x_{i,0}}(\theta,b_i)}$, in particular if their moment of order two are bounded the same holds for $\frac{\partial}{\partial(\theta,b_i)}\widehat{x_{i,0}}(\theta, b_i)$.

Now, for the third part of the lemma, we establish the partial derivative of $\overline{X}^{(j)}_{\theta,b_i,x_0}$ and $p^{(j)}_{\theta,b_i,x_0}$ with respect to (θ, b_i, x_0) have finite moment of order 2. This in turn will be sufficient to conclude for the partial derivative of $\overline{u}_{\theta,b_i,x_0}$ and $\widehat{x_{i,0}}(\theta, b_i)$ and to conclude the proof of lemma 0.5. Now we differentiate the conjugate equation with respect to (b_i, θ):

$$
\begin{array}{lcl}
\frac{d}{dt}\frac{\partial}{\partial(\theta,b_i)}\overline{X}^{(j)}_{\theta,b_i,x_0} & = & (t_{j+1}-t_j)\frac{\partial}{\partial(\theta,b_i)}f_{\theta,b_i}(t_j+t(t_{j+1}-t_j),\overline{X}^{(j)}_{\theta,b_i,x_0}) \\
 & + & (t_{j+1}-t_j)\frac{\partial}{\partial x}f_{\theta,b_i}(t_j+t(t_{j+1}-t_j)),\overline{X}^{(j)}_{\theta,b_i,x_0})\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0} \\
 & - & \frac{BU^{-1}B^T}{2p^0}\frac{\partial}{\partial(b_i,\theta)}p^{(j)}_{\theta,b_i,x_0} \\
\frac{d}{dt}\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0} & = & (t_{j+1}-t_j)\frac{\partial}{\partial x}f_{\theta,b_i}(t_j+t(t_{j+1}-t_j)),\overline{X}^{(j)}_{\theta,b_i,x_0})\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0} \\
 & - & \frac{BU^{-1}B^T}{2p^0}\frac{\partial}{\partial x_0}p^{(j)}_{\theta,b_i,x_0} \\
\frac{d}{dt}\frac{\partial}{\partial(\theta,b_i)}p^{(j)}_{\theta,b_i,x_0} & = & -(t_{j+1}-t_j)\frac{\partial^2}{\partial(\theta,b_i)\partial x}f_{\theta,b_i}(t_j+t(t_{j+1}-t_j),\overline{X}^{(j)}_{\theta,b_i,x_0})^T p^{(j)}_{\theta,b_i,x_0} \\
 & - & (t_{j+1}-t_j)\left(\frac{\partial^2}{\partial^2 x}f_{\theta,b_i}(t_j+t(t_{j+1}-t_j),\overline{X}^{(j)}_{\theta,b_i,x_0})\frac{\partial}{\partial(\theta,b_i)}\overline{X}^{(j)}_{\theta,b_i,x_0}\right)^T p^{(j)}_{\theta,b_i,x_0} \\
 & - & (t_{j+1}-t_j)\frac{\partial}{\partial x}f_{\theta,b_i}(t_j+t(t_{j+1}-t_j),\overline{X}^{(j)}_{\theta,b_i,x_0})^T\frac{\partial}{\partial(\theta,b_i)}p^{(j)}_{\theta,b_i,x_0} \\
\frac{d}{dt}\frac{\partial}{\partial x_0}p^{(j)}_{\theta,b_i,x_0} & = & -(t_{j+1}-t_j)\left(\frac{\partial^2}{\partial^2 x}f_{\theta,b_i}(t_j+t(t_{j+1}-t_j),\overline{X}^{(j)}_{\theta,b_i,x_0})\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0}\right)^T p^{(j)}_{\theta,b_i,x_0} \\
 & - & (t_{j+1}-t_j)\frac{\partial}{\partial x}f_{\theta,b_i}(t_j+t(t_{j+1}-t_j),\overline{X}^{(j)}_{\theta,b_i,x_0})^T\frac{\partial}{\partial x_0}p^{(j)}_{\theta,b_i,x_0} \\
\frac{\partial}{\partial(\theta,b_i,x_0)}\overline{X}^{(j)}_{\theta,b_i,x_0}(0) & = & \frac{\partial}{\partial(\theta,b_i,x_0)}\overline{X}^{(j-1)}_{\theta,b_i,x_0}(1),\ \frac{\partial}{\partial(\theta,b_i)}\overline{X}^{(0)}_{\theta,b_i,x_0}(0)=0,\ \frac{\partial}{\partial x_0}\overline{X}^{(0)}_{\theta,b_i,x_0}(0)=I_d \\
\frac{\partial}{\partial(\theta,b_i,x_0)}p^{(j)}_{\theta,b_i,x_0}(1) & = & 2p^0C^TC\frac{\partial}{\partial(\theta,b_i,x_0)}\overline{X}^{(j)}_{\theta,b_i,x_0}(1).
\end{array}
$$

We get the inequalities:

$$
\begin{array}{lcl}
\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}(t)\right\|_2^2 & \leq & 4T\overline{\partial_{(b_i,\theta)}f}+4T\overline{\partial_x f}\int_0^t\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}(s)\right\|_2^2 ds \\
 & + & \left\|\frac{BU^{-1}B}{p^0}\right\|_2^2\left\|\frac{\partial}{\partial(b_i,\theta)}p^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}+\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}(0)\right\|_2^2 \\
\left\|\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0}(t)\right\|_2^2 & \leq & \left\|\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0}(0)\right\|_2^2+2T\overline{\partial_x f}\int_0^t\left\|\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0}(s)\right\|_2^2 ds \\
 & + & \left\|\frac{BU^{-1}B}{p^0}\right\|_2^2\left\|\frac{\partial}{\partial x_0}p^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}
\end{array}
$$

$$
\begin{array}{lcl}
\left\|\frac{\partial}{\partial(b_i,\theta)}p^{(j)}_{\theta,b_i,x_0}(t)\right\|_2^2 & \leq & \left\|2p^0C^TC\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}(1)\right\|_2^2+4T\overline{\partial_{(b_i,\theta)x}f}\int_t^1\left\|p^{(j)}_{\theta,b_i,x_0}(s)\right\|_2^2 ds \\
 & + & 4T\overline{\partial_{xx}f}\int_t^1\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}(s)\right\|\left\|p^{(j)}_{\theta,b_i,x_0}(s)\right\|_2^2 ds+2T\overline{\partial_x f}\int_t^1\left\|\frac{\partial}{\partial(b_i,\theta)}p^{(j)}_{\theta,b_i,x_0}\right\|_2^2 ds \\
 & \leq & 8T\left\|p^0C^TC\right\|_2^2\left(\overline{\partial_{(b_i,\theta)}f}+\overline{\partial_x f}\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}^2\right) \\
 & + & \left\|2p^0C^TC\right\|_2^2\left(\left\|\frac{BU^{-1}B}{p^0}\right\|_2^2\left\|\frac{\partial}{\partial(b_i,\theta)}p^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}+\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}(0)\right\|_2^2\right) \\
 & + & 4T\left(\overline{\partial_{xx}f}\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}^2+\overline{\partial_{(b_i,\theta)x}f}\right)\left\|p^{(j)}_{\theta,b_i,x_0}\right\|_{L^2} \\
 & + & 2T\overline{\partial_x f}\int_t^1\left\|\frac{\partial}{\partial(b_i,\theta)}p^{(j)}_{\theta,b_i,x_0}\right\|_2^2 ds \\
\left\|\frac{\partial}{\partial x_0}p^{(j)}_{\theta,b_i,x_0}(t)\right\|_2^2 & \leq & \left\|2p^0C^TC\frac{\partial}{\partial x}\overline{X}^{(j)}_{\theta,b_i,x_0}(1)\right\|_2^2 \\
 & + & 2T\left(\overline{\partial_{xx}f}\int_t^1\left\|p^{(j)}_{\theta,b_i,x_0}(s)\right\|_2^2 ds+\overline{\partial_x f}\int_t^1\left\|\frac{\partial}{\partial(b_i,\theta)}p^{(j)}_{\theta,b_i,x_0}\right\|_2^2 ds\right) \\
 & \leq & \left\|2p^0C^TC\right\|_2^2\left(2T\overline{\partial_x f}\int_0^t\left\|\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0}(s)\right\|_2^2 ds+\left\|\frac{BU^{-1}B}{p^0}\right\|_2^2\left\|\frac{\partial}{\partial x_0}p^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}\right) \\
 & + & \left\|2p^0C^TC\right\|_2^2\left\|\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0}(0)\right\|_2^2 \\
 & + & 2T\left(\overline{\partial_{xx}f}\int_t^1\left\|p^{(j)}_{\theta,b_i,x_0}(s)\right\|_2^2 ds+\overline{\partial_x f}\int_t^1\left\|\frac{\partial}{\partial(b_i,\theta)}p^{(j)}_{\theta,b_i,x_0}\right\|_2^2 ds\right)
\end{array}
$$

and from this and the Gronwall lemma, we end up with

$$\begin{array}{rcl}
\left\|\frac{\partial}{\partial(b_i,\theta)}p^{(j)}_{\theta,b_i,x_0}(t)\right\|_2^2 & \leq & 8T\left\|p^0C^TC\right\|_2^2\left(\overline{\partial_{(b_i,\theta)}f}+\overline{\partial_x f}\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}^2\right)e^{2T\overline{\partial_x f}}\\
 & + & \left\|2p^0C^TC\right\|_2^2\left(\left\|\frac{BU^{-1}B}{p^0}\right\|_2^2\left\|\frac{\partial}{\partial(b_i,\theta)}p^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}+\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}(0)\right\|_2^2\right)e^{2T\overline{\partial_x f}}\\
 & + & 4T\left(\overline{\partial_{xx}f}\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}^2+\overline{\partial_{(b_i,\theta)x}f}\right)\left\|p^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}e^{2T\overline{\partial_x f}}\\
\left\|\frac{\partial}{\partial x_0}p^{(j)}_{\theta,b_i,x_0}(t)\right\|_2^2 & \leq & \left\|2p^0C^TC\right\|_2^2\left(2T\overline{\partial_x f}\int_0^t\left\|\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0}(s)\right\|_2^2ds+\left\|\frac{BU^{-1}B}{p^0}\right\|_2^2\left\|\frac{\partial}{\partial x_0}p^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}\right)e^{2T\overline{\partial_x f}}\\
 & + & 2T\overline{\partial_{xx}f}\int_t^T\left\|p^{(j)}_{\theta,b_i,x_0}(s)\right\|_2^2dse^{2T\overline{\partial_x f}}+\left\|2p^0C^TC\right\|_2^2\left\|\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0}(0)\right\|_2^2e^{2T\overline{\partial_x f}}
\end{array}$$

so if we choose $K_3(U)=2\left\|C^TC\right\|_2^2\left\|BU^{-1}B\right\|_2^2e^{2T\overline{\partial_x f}}<1$, we obtain:

$$\begin{array}{rcl}
\left\|\frac{\partial}{\partial(b_i,\theta)}p^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}^2 & \leq & \frac{8T\left\|p^0C^TC\right\|_2^2\left(\overline{\partial_{(b_i,\theta)}f}+\overline{\partial_x f}\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}^2\right)e^{2T\overline{\partial_x f}}}{1-K_3(U)}\\
 & + & \frac{\left\|2p^0C^TC\right\|_2^2\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}(0)\right\|_2^2e^{2T^2\overline{\partial_x f}}}{1-K_3(U)}\\
 & + & \frac{4T\left(\overline{\partial_{xx}f}\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}^2+\overline{\partial_{(b_i,\theta)x}f}\right)\left\|p^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}e^{2T\overline{\partial_x f}}}{1-K_3(U)}\\
\left\|\frac{\partial}{\partial x_0}p^{(j)}_{\theta,b_i,x_0}(t)\right\|_{L^2} & \leq & \frac{2T\left(2\left\|p^0C^TC\right\|_2^2\overline{\partial_x f}\left\|\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}^2+\overline{\partial_{xx}f}\left\|p^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}\right)e^{2T\overline{\partial_x f}}}{1-K_3(U)}\\
 & + & \frac{\left\|2p^0C^TC\right\|_2^2\left\|\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0}(0)\right\|_2^2e^{2T\overline{\partial_x f}}}{1-K_3(U)}
\end{array}$$

and by re-injecting in the inequality ruling $\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}(t)\right\|_2^2$ and $\left\|\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0}(t)\right\|_2^2$ we get:

$$\begin{array}{rcl}
\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}(t)\right\|_2^2 & \leq & 4T\overline{\partial_{(b_i,\theta)}f}+4T\overline{\partial_x f}\int_0^t\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}(s)\right\|_2^2ds+\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}(0)\right\|_2^2\\
 & + & \left\|BU^{-1}B\right\|_2^2\frac{8T\left\|C^TC\right\|_2^2\left(\overline{\partial_{(b_i,\theta)}f}+\overline{\partial_x f}\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}^2\right)e^{2T\overline{\partial_x f}}}{1-K_3(U)}\\
 & + & \left\|BU^{-1}B\right\|_2^2\frac{\left\|2C^TC\right\|_2^2\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}(0)\right\|_2^2e^{2T\overline{\partial_x f}}}{1-K_3(U)}\\
 & + & \left\|\frac{BU^{-1}B}{p^0}\right\|_2^2\frac{4T\left(\overline{\partial_{xx}f}\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}^2+\overline{\partial_{(b_i,\theta)x}f}\right)\left\|p^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}e^{2T\overline{\partial_x f}}}{1-K_3(U)}\\
\left\|\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0}(t)\right\|_2^2 & \leq & \left\|\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0}(0)\right\|_2^2\left(1+\left\|BU^{-1}B\right\|_2^2\frac{\left\|2C^TC\right\|_2^2e^{2T\overline{\partial_x f}}}{1-K_3(U)}\right)\\
 & + & 2T\overline{\partial_x f}\int_0^t\left\|\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0}(s)\right\|_2^2ds\\
 & + & \left\|\frac{BU^{-1}B}{p^0}\right\|_2^2\frac{2T\left(2\left\|p^0C^TC\right\|_2^2\overline{\partial_x f}\left\|\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}^2+\overline{\partial_{xx}f}\left\|p^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}\right)e^{2T\overline{\partial_x f}}}{1-K_3(U)}
\end{array}$$

which leads to

$$\begin{aligned}
\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}\right\|^2_{L^2} &\leq \left(4T\overline{\partial_{(b_i,\theta)}f}+\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}(0)\right\|^2_2\right)e^{4T\overline{\partial_x f}}\\
&+ \left\|BU^{-1}B\right\|^2_2\frac{8T\left\|C^TC\right\|^2_2\left(\overline{\partial_{(b_i,\theta)}f}+\overline{\partial_x f}\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}\right\|^2_{L^2}\right)e^{6T\overline{\partial_x f}}}{1-K_3(U)}\\
&+ \left\|BU^{-1}B\right\|^2_2\frac{\left\|2C^TC\right\|^2_2\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}(0)\right\|^2_2e^{6T\overline{\partial_x f}}}{1-K_3(U)}\\
&+ \left\|\frac{BU^{-1}B}{p^0}\right\|^2_2\frac{4T\left(\overline{\partial_{xx}f}\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}\right\|^2_{L^2}+\overline{\partial_{(b_i,\theta)x}f}\right)\left\|p^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}e^{6T\overline{\partial_x f}}}{1-K_3(U)}\\
\left\|\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0}\right\|^2_{L^2} &\leq \left\|\frac{BU^{-1}B}{p^0}\right\|^2_2\frac{2T\left(2\left\|p^0C^TC\right\|^2_2\overline{\partial_x f}\left\|\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0}\right\|^2_{L^2}+\overline{\partial_{xx}f}\left\|p^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}\right)e^{6T\overline{\partial_x f}}}{1-K_3(U)}\\
&+ \left\|\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0}(0)\right\|^2_2\left(1+\left\|BU^{-1}B\right\|^2_2\frac{\left\|2C^TC\right\|^2_2e^{2T\overline{\partial_x f}}}{1-K_3(U)}\right)e^{4T\overline{\partial_x f}}.
\end{aligned}$$

So if U is chosen such that $K_4(U)=\frac{4T\left\|BU^{-1}B\right\|^2_2e^{6T\overline{\partial_x f}}}{1-K_3(U)}\left(2\left\|C^TC\right\|^2_2\overline{\partial_x f}+\frac{\overline{\partial_{xx}f}}{|p^0|}\right)<1$ and $K_5(U)=\frac{4T}{1-K_3(U)}\left\|BU^{-1}B\right\|^2_2\left\|C^TC\right\|^2_2\overline{\partial_x f}e^{6T\overline{\partial_x f}}<1$, then:

$$\begin{aligned}
\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}\right\|^2_{L^2} &\leq \frac{\left(4T\overline{\partial_{(b_i,\theta)}f}+\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}(0)\right\|^2_2\right)}{1-K_4(U)}e^{4T\overline{\partial_x f}}\\
&+ \left\|BU^{-1}B\right\|^2_2\left\|C^TC\right\|^2_2e^{6T\overline{\partial_x f}}\frac{\left(8T\left\|C^TC\right\|^2_2\overline{\partial_{(b_i,\theta)}f}+2\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(j)}_{\theta,b_i,x_0}(0)\right\|^2_2\right)}{(1-K_4(U))(1-K_3(U))}\\
&+ \left\|\frac{BU^{-1}B}{p^0}\right\|^2_2\frac{4T\overline{\partial_{(b_i,\theta)x}f}\left\|p^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}e^{6T\overline{\partial_x f}}}{(1-K_4(U))(1-K_3(U))}\\
\left\|\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0}\right\|^2_{L^2} &\leq \left\|\frac{BU^{-1}B}{p^0}\right\|^2_2\frac{2T\overline{\partial_{xx}f}\left\|p^{(j)}_{\theta,b_i,x_0}\right\|_{L^2}e^{6T\overline{\partial_x f}}}{(1-K_5(U))(1-K_3(U))}\\
&+ \frac{\left\|\frac{\partial}{\partial x_0}\overline{X}^{(j)}_{\theta,b_i,x_0}(0)\right\|^2_2}{(1-K_5(U))}\left(1+\left\|BU^{-1}B\right\|^2_2\frac{\left\|2C^TC\right\|^2_2e^{2T\overline{\partial_x f}}}{1-K_3(U)}\right)e^{4T\overline{\partial_x f}}.
\end{aligned}$$

Now let us focus on the parameter value $(\theta,b_i,x_0):=\left(\overline{\theta},\widehat{b_i}(\overline{\eta}),\widehat{x_0}(\overline{\theta},\widehat{b_i}(\overline{\eta})\right)$. We already show that $\left\|p^{(j)}_{\overline{\theta},\widehat{b_i}(\overline{\eta})}(t)\right\|^2_2\leq\sup_{(\theta,b,x_0)}\left\|p^0C^T\left(C\overline{X}^{(j)}_{\overline{\theta},\widehat{b_i}(\overline{\eta})}(1)-y_j\right)\right\|^2_2e^{T^2\overline{\partial_x f}}$ from which we derive $\left\|p^{(j)}_{\overline{\theta},\widehat{b_i}(\overline{\eta})}(t)\right\|^2_2$ has a finite moment of order 2, since we already establish the result for $\left\|p^0C^T\left(C\overline{X}^{(j)}_{\overline{\theta},\widehat{b_i}(\overline{\eta})}(1)-y_j\right)\right\|^2_2$ in the previous lemma. Since $\left\|\frac{\partial}{\partial(b_i,\theta)}\overline{X}^{(0)}_{\overline{\theta},\widehat{b_i}(\overline{\eta})}(0)\right\|^2_2=0$ and $\left\|\frac{\partial}{\partial x_0}\overline{X}^{(0)}_{\overline{\theta},\widehat{b_i}(\overline{\eta})}(0)\right\|^2_2=d$, we can use the previous inequality to show that $\left\|\frac{\partial}{\partial(b_i,\theta,x_0)}\overline{X}^{(0)}_{\overline{\theta},\widehat{b_i}(\overline{\eta}),}\right\|_{L^2}$ has a finite moment of order 2, from this it follows the same holds for $\left\|\frac{\partial}{\partial(b_i,\theta,x_0)}\overline{p}^{(0)}_{\overline{\theta},\widehat{b_i}(\overline{\eta})}\right\|_{L^2}$ and $\left\|\frac{\partial}{\partial(b_i,\theta,x_0)}\overline{u}^{(0)}_{\overline{\theta},\widehat{b_i}(\overline{\eta})}\right\|_{L^2}$ this in turn leads to the boundedness of moment of order 2 for $\left\|\frac{\partial}{\partial(b_i,\theta,x_0)}\overline{X}^{(0)}_{\overline{\theta},\widehat{b_i}(\overline{\eta})}(t)\right\|_2$ and $\left\|\frac{\partial}{\partial(b_i,\theta,x_0)}\overline{u}^{(0)}_{\overline{\theta},\widehat{b_i}(\overline{\eta})}(t)\right\|_2$. As in the previous part of the lemma, we can use a recursive argument to show boundedness of moments of order two of $\left\|\frac{\partial}{\partial(b_i,\theta,x_0)}\overline{X}^{(j)}_{\overline{\theta},\widehat{b_i}(\overline{\eta})}(t)\right\|_2$ and $\left\|\frac{\partial}{\partial(b_i,\theta,x_0)}\overline{u}^{(j)}_{\overline{\theta},\widehat{b_i}(\overline{\eta})}(t)\right\|_2$ for $j>0$. Since $\frac{\partial}{\partial(\theta,b_i)}\widehat{x_{i,0}}\,(\theta,b_i)=-\frac{\partial^2}{\partial^2x_{i,0}}\overline{\mathcal{C}_i}(\theta,b_i,\widehat{x_{i,0}}\,(\theta,b_i))^{-1}\frac{\partial}{\partial(\theta,b_i)}\nabla_{x_{i,0}}\overline{\mathcal{C}_i}(\theta,b_i,\widehat{x_{i,0}}\,(\theta,b_i))$, the same results holds for $\left\|\frac{\partial}{\partial(\theta,b_i)}\widehat{x_{i,0}}\left(\overline{\theta},\widehat{b_i}(\overline{\eta})\right)\right\|_2$. The proof is similar for the derivative of higher dimension. □

References

[1] M.D.S. Aliyu. *Nonlinear H-Infinity Control, Hamiltonian Systems and Hamilton-Jacobi Equations.* CRC Press, 2011.

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

[3] T. Cimen and S.P. Banks. Nonlinear optimal tracking control with application to super-tankers for autopilot design. *Automatica*, 40:1845–1863, 2004.

[4] Quentin Clairon. A regularization method for the parameter estimation problem in ordinary differential equations via discrete optimal control theory. *Journal of Statistical Planning and Inference*, 2020.

[5] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mischenko. *The Mathematical Theory of Optimal Processes.* Wiley-Interscience, 1962.

[6] E. Sontag. *Mathematical Control Theory: Deterministic finite-dimensional systems.* Springer-Verlag (New-York), 1998.

[7] A.W. van der Vaart. *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilities Mathematics. Cambridge University Press, 1998.