A regularization method for the parameter estimation problem in ordinary differential equations via discrete optimal control theory

Quentin Clairon

A regularization method for the parameter estimation problem in ordinary differential equations via discrete optimal control theory

March 19, 2020

Quentin Clairon[1]
[1] University of Bordeaux, Inria Bordeaux Sud-Ouest, Inserm, Bordeaux Population Health Research Center, SISTM Team, UMR 1219

Abstract

We present a parameter estimation method in Ordinary Differential Equation (ODE) models. Due to complex relationships between parameters and states the use of standard techniques such as nonlinear least squares can lead to the presence of poorly identifiable parameters. Moreover, ODEs are generally approximations of the true process and the influence of misspecification on inference is often neglected. Here, we propose a method based on discrete optimal control theory to regularize the ill posed problem of parameter estimation in this context. We describe how the estimation problem can be turned into a control one and present the numerical methods used to solve it. We show convergence of our estimators in the parametric and well-specified case. We test and compare our method with existing approaches on numerical experiments with models containing poorly identifiable parameters and with various sources of model misspecification. They illustrate the regularization brought by our approach to the problem comparing to exact methods such as Non-linear least squares. Moreover, this discrete optimal control based procedure is computationally less intensive and more accurate in sparse sample case than the one based on continuous control techniques. We finally test our approach on a real data example.

Keywords: Ordinary differential equation; discrete optimal control; parametric estimation; semi parametric estimation; model uncertainty

1 Introduction

We are interested by parameter estimation in Ordinary Differential Equation (ODE) models of the form

$$\begin{cases} \dot{x}(t) = f\left(t, x(t), \theta, \vartheta(t)\right) \\ x(0) = x_0 \end{cases} \tag{1}$$

where the state x is in $\mathbb{R}^d$, f is a vector field from $[0,\, T] \times \mathbb{R}^d \times \Theta \times \Theta_f$ to $\mathbb{R}^d$, θ is a parameter that belongs to a subset Θ of $\mathbb{R}^p$, ϑ is a functional parameter from $[0, T]$ to $\Theta_f \subseteq \mathbb{R}^{d_f}$ and x_0 is the initial condition that belongs to a subset χ of $\mathbb{R}^d$. ODEs are much used in practice as they provide an efficient framework for analyzing and predicting complex systems (see eg [Fall et al., 2002, Goldbeter, 1997, Mirsky et al., 2009, Wu et al., 2014]). In particular, there has recently been focus on joint use of ODE models and control theory methods for the purpose of optimal treatment design [Guo and Sun, 2012, Agusto and Adekunle, 2014, Zhang and Xu, 2016].

Our aim is to estimate the true parameters, denoted θ^* and ϑ^*, starting from data $y_1, \ldots, y_n$, that are realizations of an observation process for $i = 1, \ldots, n$

$$Y_i = CX^*(t_i) + \epsilon_i \tag{2}$$

on the observation interval $[0,\, T]$ where $X^* := X_{\theta^*, \vartheta^*, x_0^*}$ is the solution of (1) for $\theta = \theta^*$, $\vartheta = \vartheta^*$ and $x_0 = x_0^*$, C is a $d^{'} \times d$ observation matrix and ϵ_i is centered observation noise. That is, we want to estimate (θ^*, ϑ^*) starting from discrete, partial and noisy observations of X^* at observation times $0 = t_1 < t_2 \cdots < t_n = T$. In absence of ϑ^*, estimation of θ^* is a standard parametric nonlinear regression problem and can be solved by classical methods such as Nonlinear Least Squares (NLS), Maximum Likelihood Estimation (MLE), or Bayesian Inference [Esposito and Floudas, 2000, Li et al., 2005, Rodriguez-Fernandez et al., 2006, Wu et al., 2010]. However, in the case of ODE models, there is a risk of an ill-posed inverse problem [Engl et al., 2009, Stuart, 2010].

To explain why, let us denote as X_{θ, x_0} the solution to (1). The Fisher information matrix which controls the Cramer-Rao bound is proportional to $\mathcal{I}_n\left(\theta, x_0\right) = \sum_{i=1}^{n} \left(C \frac{\partial X_{\theta, x_0}(t_i)}{\partial(\theta, x_0)}\right)^T C \frac{\partial X_{\theta, x_0}(t_i)}{\partial(\theta, x_0)}$. Instabilities

in estimation arise when the matrices $C\frac{\partial X_{\theta,x_0}(t_i)}{\partial(\theta,x_0)}$ are badly-conditioned because in this case the inverse problem is very sensitive to any source of perturbations and the objective function (NLS or MLE criteria) is nearly flat around its minimum. This practical identifiability problem can be measured by computing the spectrum $\mu_1 \geq \cdots \geq \mu_p$ of $\mathcal{I}_n(\theta, x_0)$ and is associated to a weak condition number $\kappa(\mathcal{I}_n) = \frac{\mu_1}{\mu_p}$. The problem arises in part from the observation process, the sparsity and location of the observation times and also from the need to estimate the nuisance parameter x_0^*. Complication in ODEs also arises due to the complex geometry of the manifold $\{CX_{\theta,x_0}, \theta \in \Theta, x_0 \in \chi\}$ induced by the mapping $(\theta, x_0) \longmapsto CX_{\theta,x_0}$ where there can be a small number (in comparison with p) of important directions of variation very skewed from the original parameter axes [Gutenkunst et al., 2007, Transtrum et al., 2011, Transtrum et al., 2015]. This situation is termed sloppiness and leads to a regular and widespread distribution of the eigenvalues $\mu_1, \ldots, \mu_p$ with no clear one to one correspondence between the eigenvectors of $\mathcal{I}_n(\theta, x_0)$ and the original ODE parametrization. Numerous ODEs used for example in systems biology [Gutenkunst et al., 2007] and neuroscience [Leary et al., 2015] have been identified as sloppy. Sloppiness is a phenomenon arising from interactions between intrinsic system properties and the experimental design, it is due to the sparse and block structure of $C\frac{\partial X_{\theta,x_0}(t_i)}{\partial(\theta,x_0)}$ with highly correlated entries [Tonsing et al., 2014]. Since we cannot clearly distinguish important parameters from the others, there is no clear mechanism to suppress irrelevant parameters in the model. Moreover, methods based on optimal experimental design to circumvent sloppiness can lead to experiments which push the system in a state where the assumed model is no longer valid. This can cause model error problems when trying to estimate parameters from the new data set and reduce model predictive ability [White et al., 2016]. Despite that sloppiness and practical identifiability are not rigorously the same problem, the former often induces the latter by making some subset of parameters unidentifiable. Thus, there is a need to improve estimation methods which use the existing data without resorting to new experiments.

Another issue in ODE parameter estimation comes from the fact that the selected model can suffer from model misspecification issues. By resuming the terminology of [Kennedy and OHagan, 2001], we refer to model misspecification when the ODE model suffers from 1/ Model inadequacy: discrepancy between

the mean model response and real world process. ODEs are derived by approximations, simplification of interactions and omission of external factors influence can cause such discrepancy. 2/ Residual variability issues: many biological processes are known to be stochastic and the justification of deterministic modeling comes from the approximation of stochastic processes by ODE solutions see [Kurtz, 1970, Kurtz, 1978, Gillespie, 2000, Kampen, 1992]. Hence, inference of the parameters has to be done while recognizing that the model is false [Kirk et al., 2016, Brynjarsdottir and O'Hagan, 2014].

In this work, we propose a new estimation procedure to address these challenges, based on an approximate solution of the original ODE. The use of approximate solutions for statistical inference, such as the two-step approaches [Varah, 1982, Gugushvili and Klaassen, 2011, Liang et al., 2010, Brunel and D'Alche-Buc, 2014, Dattner, 2015], Generalized Profiling (GP) [G. Hooker and Earn, 2011, Ramsay et al., 2007] or in a Bayesian framework [Chkrebtii et al., 2016, Jaeger and Lambert, 2011], has already proven to be useful for regularizing the inverse problem of parameter estimation. In presence of poorly identifiable parameters, the appeal of such methods are their ability to bypass the Cramer-Rao bound which imposes to exact methods a dramatic increase of estimator variance. In case of model misspecification, they can improve estimation accuracy for they relax the constraint imposed by the ODE model and then account for model discrepancy in the criteria to optimize [Brynjarsdottir and O'Hagan, 2014].

Our proposed method presents similarities with the ones introduced in [Brunel and Clairon, 2015, Clairon and Brunel, 2019, Clairon and Brunel, 2018], where an approximation $X_{\theta,x_0,u}$ is a solution of the perturbed ODE $\dot{x}(t) = f(t, x(t), \theta) + Bu(t)$ where the perturbation $t \mapsto Bu(t)$ captures different sources of model misspecification. After a pre-smoothing step to obtain a nonparametric curve estimator $\widehat{Y}$, the estimator $\left(\widehat{\theta}, \widehat{x_0}\right)$ is defined as the minimizer of the cost $C_\lambda(\theta, x_0, u) = \left\|CX_{\theta,x_0,u} - \widehat{Y}\right\|_{L^2}^2 + \lambda \|u\|_{L^2}^2$ profiled on the possible perturbations u: $\left(\widehat{\theta}, \widehat{x_0}\right) = \arg\min_{(\theta,x_0)} S(\theta, x_0)$, where $S(\theta, x_0) = \min_u C_\lambda(\theta, x_0, u)$. This estimator, called the Tracking Estimator (TE), is thus defined as the parameter which needs the smallest perturbation u in order to track $\widehat{Y}$, the balance between the two contrary objectives of data fidelity (i.e. $\left\|CX_{\theta,x_0,u} - \widehat{Y}\right\|_{L^2}^2$) and original model fidelity (i.e. $\|u\|_{L^2}^2$) is done through the choice of an hyperparameter λ. For each value (θ, x_0), the optimal control problem $\min_u C_\lambda(\theta, x_0, u)$ is solved by using the Pontryagin

maximum principle [Pontryagin et al., 1962]. In comparison with GP and NLS, the TE generally has a lower variance and mean square error with the difference in performance even more marked in the presence of model misspecification. In the parametric case and for well-specified models, the TE is consistent with a $\sqrt{n}$- convergence rate under very mild model regularity conditions and provided $\lambda > \overline{\lambda_1}$, with $\overline{\lambda_1}$ a positive model dependent bound. Another attractive feature of the tracking framework is the seamless estimation of finite-dimensional and time-varying parameters. The estimation of ϑ is turned into an optimal control problem and estimator $\hat{\vartheta}$ is a by-product of θ^* estimation which does not require the use of standard approximations such as sieves or basis expansions [Xue et al., 2010, G. Hooker and Earn, 2011, Wang et al., 2014]. However, there are two main limitations for the method presented in [Clairon and Brunel, 2018]. First, the computational time: solving the optimal control problem by using the Pontryagin maximum principle leads to a boundary value problem (BVP) for each new (θ, x_0) value and x_0^* has to be estimated as nuisance parameter. Second, the method requires a nonparametric estimator which can be biased in sparse data case, this bias can then be spread to the parametric estimation. Here, while we still consider an optimal control based approach, we change the cost function C_λ as well as the numerical procedure used to solve the related optimal control problem. We rely on discrete control theory and a numerical method inspired by [Cimen and Banks, 2004b]. This allows us to construct a method which:

1. replaces the BVP by a sequence of finite difference equations which is numerically solved significantly faster,

2. does not require a pre-smoothing step, we can deal with sparse data cases which are consistent with most real observation framework,

3. can be easily adapted to avoid estimation of x_0^* if it is not required.

In order to define our estimators, we present in the next section the optimal control problem required to introduce our functional criteria and describe our approach for semi-parametric estimation. In section 3, we derive the numerical procedures. In section 4, we study the asymptotic behavior of our estimators. In section 5, we use Monte Carlo experiments to compare the Tracking, NLS and GP estimators on ODE

examples from chemistry and biology with both well-specified and misspecified models. The discrete control theory based method allows us to obtain more accurate estimates than GP and NLS. We also investigate differences between the method developed here and the one in [Clairon and Brunel, 2018]. We emphasize the advantage of using this discrete based one in terms of computational time and estimation accuracy, in particular for sparse sample cases. In Section 6, we consider parameter estimation with real data in a model used to study microbiotal population evolution.

2 Model and methodology

We recall the aim of this work is to define estimators of (θ^*, ϑ^*) as minimizers of functional criteria. First, we derive them in the parametric case where there is no functional parameter ϑ^*.

2.1 Formal parametric estimator definition

We denote by X_{θ,x_0} the solution of the Initial Value Problem (IVP):

$$\begin{cases} \dot{x}(t) = f(t, x(t), \theta) \\ x(0) = x_0. \end{cases} \tag{3}$$

First, we introduce a pseudo-linear and perturbed version of model (3):

$$\begin{cases} \dot{x}(t) = A_\theta(x(t), t)x(t) + Bu(t) \\ x(0) = x_0 \end{cases} \tag{4}$$

where the function $t \mapsto Bu(t)$ is a linear perturbation, B is a $d \times d_u$ matrix and u is in $L^2\left([0,T], \mathbb{R}^{d_u}\right)$. Here, the matrix A_θ is defined by the relation $A_\theta(x(t), t)x(t) = f(t, x(t), \theta)$, this formulation is crucial for solving in a computationally efficient way the optimal control problem defining our estimators. Linear models already fit in this formalism with $A_\theta(t) = A_\theta(x(t), t)$. For nonlinear models, the pseudo-linear representation is not unique but always exists [Cimen and Banks, 2004b]. We denote $X_{\theta,x_0,u}$ the solution of the perturbed ODE (4).

Now, we introduce the cost function required to define our estimators:

$$C_{(T,U)}(Y;\theta,x_0,u)=\sum_{i=0}^{n}\|CX_{\theta,x_0,u}(t_i)-y_i\|_2^2+\int_0^T u(t)^T U u(t)dt \tag{5}$$

where U is a symmetric definite positive matrix used as a weighting parameter balancing the amount of model and data fidelity. For each (θ,x_0) in $\Theta\times\chi$, we define the profiled cost:

$$S_{(n,U)}(Y;\theta,x_0) := \inf_u C_{(T,U)}(Y;\theta,x_0,u) \tag{6}$$

on the set of possible perturbations u, in the case we want to bypass x_0^* estimation, we also introduce

$$S^{CI}_{(n,U)}(Y;\theta) := \min_{x_0}\inf_u C_{(T,U)}(Y;\theta,x_0,u) \tag{7}$$

the profiled cost function on x_0 in addition to u. From these criteria, the estimators are defined as:

$$\left(\widehat{\theta}_U^T,\widehat{x_{0,U}}^T\right) := \arg\min_{(\theta,x_0)\in\Theta\times\chi} S_{(n,U)}(Y;\theta,x_0), \tag{8}$$

and

$$\widehat{\theta}_U^{T,CI} := \arg\min_{\theta\in\Theta} S^{CI}_{(n,U)}(Y;\theta) \tag{9}$$

i.e. as the parameter values giving the trajectory $X^d_{\theta,x_0,u}$ needing the smallest perturbation in order to be close to the observed data on $[0,T]$. Our method relaxes the original inverse problem by allowing a small divergence from the assumed model (1). The addition of u in the model followed by its norm penalization corresponds to the addition of a Tikhonov regularization term. The regularization brought by it is ensured in [Engl et al., 1996] theorem 10.12 which concludes to the smooth dependence of the regularized solution with respect to uncertainty in observations. Moreover, as pointed out in [Engl et al., 1996] chapter 5, this smooth dependence also works when the uncertainty is on the model. In the Bayesian paradigm the link between optimizing $C_{(T,U)}$ and statistical inference in functional spaces has been made in [Stuart, 2010]. The minimizer of $C_{(T,U)}$ can be seen as a MAP estimator corresponding to a prior chosen to be a centered Gaussian measure and with a covariance operator determined by U (according to theorem 3.5 and Corollary 3.10 in [Dashti et al., 2013]). That is, before having access to the observations, the original ODE is assumed

to be the most likely model for the system. In this case, the regularization brought by $\int_0^T u(t)^T U u(t) dt$ can be derived from the robustness of the posterior measure with respect to model misspecification and its smooth dependence with respect to data (respectively theorem 4.6 and theorem 4.2 in [Stuart, 2010]).

Remark 1. *One pseudo-linear representation A_θ we can easily derive for ODE (3) is obtained by dividing f componentwise by each state variable. It has to be noted that, by superposition principle, if $A_{1,\theta}$ and $A_{2,\theta}$ are two acceptable representations, so is $\alpha_1 A_{1,\theta} + \alpha_2 A_{2,\theta}$ with $\alpha_1 + \alpha_2 = 1$. In order to exploit this nonuniqueness as an additional degree of freedom see [Cimen, 2008] section 6.*

2.2 ϑ estimation

For this, let us introduce the state $x_e = (x, z_1, z_2)$ in $\mathbb{R}^{d+2d_f}$, the matrices:

$$A^e_\theta(x_e(t), t) = \begin{pmatrix} A_\theta(x(t), z_1(t), t) & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad B_{ext} = \begin{pmatrix} I_d & 0_{d,d_f} \\ 0_{d_f,d} & 0_{d_f} \\ 0_{d_f,d} & I_{d_f} \end{pmatrix}$$

and the perturbed solution $X^e_{\theta,x^e_0,u}$ of the extended ODE:

$$\begin{cases} \dot{x_e}(t) = A^e_\theta(x_e(t), t) x_e(t) + B_{ext} u(t) \\ x_e(0) = x^e_0. \end{cases} \tag{10}$$

Here, u is split into two parts, $u = (u_1,\, u_2)$ and $X^e_{\theta,x^e_0,u}$ is solution of

$$\begin{aligned} \dot{x} &= A(t, x, z_1, \theta) x + u_1 \\ \dot{z}_1 &= z_2 \\ \dot{z}_2 &= u_2 \end{aligned} \tag{11}$$

and z_1 plays the role of ϑ and z_2 of $\dot{\vartheta}$. Since we get a state variable estimator $\widehat{X^e}$ as a byproduct of θ^* estimation, we can define $\widehat{\vartheta} = \widehat{z_1}$. Let us introduce $U = \begin{pmatrix} \lambda_1 I_d & 0_{d,d_f} \\ 0_{d_f,d} & \lambda_2 I_{d_f} \end{pmatrix}$ and the cost

$$C_{(T,U)}(Y; \theta, x^e_0, u) = \sum_{i=0}^{n} \left\| C X^d_{\theta,x_0,u}(t_i) - y_i \right\|_2^2 + \lambda_1 \int_0^T \|u_1(t)\|_2^2 dt + \lambda_2 \int_0^T \|u_2(t)\|_2^2 dt. \tag{12}$$

Here $\lambda_1 \int_0^T \|u_1(t)\|_2^2 dt$ is used to quantify model discrepancy as in the parametric case and since $u_2 = \ddot{z}_1$, the last term in $C_{(T,U)}$ is the standard penalty used for functional estimation. Thus, a good choice of hyperparameters for cost (12) would be a large value for λ_1 (in order to select a small u_1), and λ_2 tending to 0 when the sample size n grows, as for standard nonparametric estimation.

Remark 2. *The state extension required for semi-parametric estimation involves the addition of new initial conditions* $\left(\vartheta(0), \dot{\vartheta}(0)\right)$. *However, if we profile on* x_0 *our approach does not add nuisance parameters.*

3 Tractable form for $S_{(n,U)}$ and $S^{CI}_{(n,U)}$

In this subsection, we derive tractable expressions for $S_{(n,U)}$ and $S^{CI}_{(n,U)}$. We start with linear ODEs then we extend the derived methods to nonlinear models by following [Cimen and Banks, 2004b]. For this, we need to specify the discrete optimal control problem we want to solve, that is the point of the next subsection.

3.1 A discrete optimal control problem framework

To proceed to parametric estimation, we resort on discrete optimal control theory. For this, we need a discrete version of the ODE (4) as well as of the cost (5). The discretization is made at $m+1$ time points $\left\{t_j^d\right\}_{0\leq j\leq m}$ with $t_0^d = 0$ and $t_m^d = T$. Letting $\Delta_j = t_{j+1}^d - t_j^d$ being the mesh size between two discretization time-points and $u = (u_0, \ldots, u_{m-1})$ the set of discrete values taken by the control at each time step, the discretized model is:

$$\begin{cases} x(t_{j+1}^d) = \left(I_d + \Delta_j A_\theta(x(t_j^d), t_j^d)\right) x(t_j^d) + B\Delta_j u_j \\ x(0) = x_0. \end{cases} \tag{13}$$

The set of discretization times has to contain the observation times i.e. $\{t_i\}_{0\leq i\leq n} \subset \left\{t_j^d\right\}_{0\leq j\leq m}$ but can be bigger, this is an important feature of the discretization scheme to accurately estimate $X_{\theta,x_0,u}$ even when the observations are sparse on $[0, T]$. We denote $X^d_{\theta,x_0,u}(t_j^d)$, the solution of (13) for the parameter θ, initial condition x_0 and the perturbation u at time t_j^d.

The cost (5) is discretized by replacing the integral $\int_0^T u(t)^T U u(t) dt$ by the Riemann sum corresponding to the discretization grid:

$$\begin{aligned} C^d_{(T,U)}(Y;\theta,x_0,u) &= \sum_{i=0}^{n} \left\| CX^d_{\theta,x_0,u}(t_i) - y_i \right\|_2^2 + \sum_{j=0}^{m-1} \triangle_j {u_j}^T U u_j \\ &= \left\| CX^d_{\theta,x_0,u}(t_n) - y_n \right\|_2^2 + \sum_{j=0}^{m-1} \triangle_j \left(\left\| CX^d_{\theta,x_0,u}(t^d_j) - \mathbf{y}_j \right\|_2^2 w_j + {u_j}^T U u_j \right) \end{aligned} \tag{14}$$

where

- $w_j = 1_{\left\{\exists t_i \text{ s.t } t_i = t^d_j\right\}} / \triangle_j$ i.e. $w_j = 1/\triangle_j$ if t^d_j is the observation time t_i, otherwise $w_j = 0$,
- $\mathbf{y}_j$ is equal to y_i if $t^d_j = t_i$ and 0 otherwise.

The weights w_j and the set of extended data $\{\mathbf{y}_i\}$ are introduced to have a vector of observation which has the same length as the discretization grid $\left\{t^d_j\right\}_{0 \le j \le m}$. This allows us to estimate the integral term in (5) with an arbitrary precision while fitting in the framework of discrete optimal control theory. To compute $S_{(n,U)}$ and $S^{CI}_{(n,U)}$ in practice we need to solve the problem:

$$\begin{aligned} &\min_u C^d_{(T,U)}(Y;\theta,x_0,u) \\ &\text{such that } x(t^d_{j+1}) = \left(I_d + \Delta_j A_\theta(x(t^d_j), t^d_j)\right) x(t_i) + B\Delta_j u_j \text{ and } x(0) = x_0. \end{aligned} \tag{15}$$

The problem (15) is a tracking problem where the aim is to find the smallest control possible to apply to a given dynamical system in order to track a signal. For linear models, these problems have been efficiently solved as they fit into the framework of discrete linear-quadratic problems, which ensures the existence and uniqueness of the solution and gives a computationally efficient way to find it. For non-linear models, [Cimen and Banks, 2004b] proposes an iterative method to solve continuous time tracking problems, the main idea being to replace the original problem by a sequence of linear-quadratic ones. We use the same method adapted to discrete models.

Remark 3. *Instead of an Euler scheme leading to (13), other integration methods can be chosen. As soon as it is an explicit scheme giving rise to a state-space equation of the form $z(t+1) = Q(t, z(t))z(t) + Bu(t)$ ([Hairer and Wanner, 1996, Hairer et al., 1993]), the presented procedure for solving the optimal control is still applicable.*

3.2 Linear models

Here $A_\theta(t) = A_\theta(x,t)$ in (3). For a given initial condition x_0, linear-quadratic theory ensures the existence and uniqueness of the optimal control $\overline{u^d_{\theta,x_0}} = \arg\min_{u\in L_a} C^d_{(T,U)}(Y;\theta,x_0,u)$ and that $\inf_{u\in L_a} C^d_{(T,U)}(Y;\theta,x_0,u)$ can be computed by solving a discrete final value problem, denoted the Riccati equation. Moreover, $\inf_{u\in L_a} C^d_{(T,U)}(Y;\theta,x_0,u)$ is a quadratic form with respect to x_0, the profiling is straightforward. Interestingly, the formal computation used to derive $S^{CI}_{(n,U)}(Y;\theta)$ follows the same step as the deterministic Kalman Filter state estimator derivation [Sontag, 1998]. The formal computational details are left in *supplementary materials*.

Proposition 4. *For (θ,x_0) in $\Theta\times\chi$, $S_{(n,U)}(Y;\theta,x_0)$ and $S^{CI}_{(n,U)}(Y;\theta)$ are equals to:*

$$S_{(n,U)}(Y;\theta,x_0) = x_0^T R^d_{\theta,0} x_0 + 2h^d_{\theta,0}(Y)^T x_0 + y_n^T y_n + \sum_{j=0}^{m-1} \triangle_j \left(w_j \mathbf{y}_j^T \mathbf{y}_j - h^d_{\theta,j+1}(Y)^T BG(R^d_{\theta,j+1}) B^T h^d_{\theta,j+1}(Y) \right) \tag{16}$$

and

$$S^{CI}_{(n,U)}(Y;\theta) = -h^d_{\theta,0}(Y)^T \left(R^d_{\theta,0}\right)^{-1} h^d_{\theta,0}(Y) + y_n^T y_n + \sum_{j=0}^{m-1} \triangle_j \left(w_j \mathbf{y}_j^T \mathbf{y}_j - h^d_{\theta,j+1}(Y)^T BG(R^d_{\theta,j+1}) B^T h^d_{\theta,j+1}(Y) \right) \tag{17}$$

with $G(R^d_{\theta,j+1}) := \left[U + \triangle_j B^T R^d_{\theta,j+1} B\right]^{-1}$ and $(R^d_{\theta,j},\, h^d_{\theta,j}(Y))$ for $1 \le j \le m$, the solution of the discrete Riccati equation:

$$\begin{array}{rcl} R^d_{\theta,j} & = & R^d_{\theta,j+1} + \triangle_j w_j C^T C + \Delta_j \left(R^d_{\theta,j+1} A_\theta(t^d_j) + A_\theta(t^d_j)^T R^d_{\theta,j+1} \right) + \triangle_j^2 A_\theta(t^d_j)^T R^d_{\theta,j+1} A_\theta(t^d_j) \\ & - & \triangle_j (I_d + \triangle_j A_\theta(t^d_j)^T) R^d_{\theta,j+1} BG(R^d_{\theta,j+1}) B^T R^d_{\theta,j+1} (I_d + \triangle_j A_\theta(t^d_j)) \\ h^d_{\theta,j}(Y) & = & h^d_{\theta,j+1}(Y) - \triangle_j w_j C^T \mathbf{y}_j + \triangle_j A_\theta(t^d_j)^T h^d_{\theta,j+1}(Y) \\ & - & \triangle_j (I_d + \triangle_j A_\theta(t^d_j)^T) R^d_{\theta,j+1} BG(R^d_{\theta,j+1}) B^T h^d_{\theta,j+1}(Y) \end{array} \tag{18}$$

with final condition $(R^d_{\theta,m},\, h^d_{\theta,m}(Y)) = (C^T C,\, -C^T y_n)$. The optimal control $\overline{u^d_{\theta,x_0}}$ minimizer of the cost (14) is unique and equal to:

$$\overline{u^d_{\theta,x_0,j}} = -G(R^d_{\theta,j+1}) B^T \left(R^d_{\theta,j+1} \left(I_d + \triangle_j A_\theta(t^d_j)\right) \overline{X^d_{\theta,x_0}}(t_j) + h^d_{\theta,j+1}(Y) \right) \tag{19}$$

where $\overline{X^d_{\theta,x_0}}$ *is the optimal trajectory, the solution of*

$$\left\{\begin{array}{lcl} \overline{X^d_{\theta,x_0}}(t^d_{j+1}) & = & \left(I_d + \triangle_j A_\theta(t^d_j)\right)\overline{X^d_{\theta,x_0}}(t^d_j) \\ & - & \triangle_j BG(R^d_{\theta,j+1})B^T\left(R^d_{\theta,j+1}\left(I_d+\triangle_j A_\theta(t^d_j)\right)\overline{X^d_{\theta,x_0}}(t^d_j) + h^{d,l}_{\theta,j+1}(Y)\right) \\ \overline{X^d_{\theta,x_0}}(0) & = & x_0. \end{array}\right. \tag{20}$$

The optimal control $\overline{u^d_\theta}$ *and optimal trajectory* $\overline{X^d_\theta}$ *such that* $S^{CI}_{(n,U)}(Y;\theta) = \min_{x_0}\inf_{u\in L_a} C^d_{(T,U)}(Y;\theta,x_0,u) = C^d_{(T,U)}(Y;\theta,\overline{X^d_\theta}(0),\overline{u^d_\theta})$ *are still given by equations (19) and (20) but with initial condition* $\overline{X^d_\theta}(0) = -\left(R^d_{\theta,0}\right)^{-1}h_{\theta,0}$.

3.3 Non-linear models

Here, we adapt the solving method proposed by [Cimen and Banks, 2004b] for discrete time models. Let us detail the procedure for $S_{(n,U)}$ computation, we replace the original problem (15) by a recursive sequence of linear-quadratic control problems, with iteration l defined by

$$\begin{array}{l} \min_u C^{d,l}_{(T,U)}(Y;\theta,x_0,u) := \left\|CX^l_{\theta,x_0,u}(t^d_m) - y_n\right\|^2_2 + \sum_{j=0}^{m-1}\triangle_j\left(\left\|CX^l_{\theta,x_0,u}(t^d_j) - \mathbf{y}_j\right\|^2_2 w_j + u_j^T U u_j\right) \\ \text{such that } x(t^d_{j+1}) = (I_d + \triangle_j A^l_\theta(t^d_j))x_j(t^d_j) + \triangle_j B_\sigma u_j \text{ and } x(0) = x_0 \end{array} \tag{21}$$

where $A^l_\theta(t^d_j) := A_\theta(\overline{X^{l-1}_{\theta,x_0}}(t^d_j), t^d_j)$ and $A^0_\theta(t^d_j) := A_\theta(x_0,t^d_j)$. Here $\overline{X^{l-1}_{\theta,x_0}}$ is the optimal trajectory corresponding the optimal control problem (21) at iteration $l-1$. For each l, we use the previous proposition to compute the solution of the Riccati equation $(R^{d,l}_\theta,\, h^{d,l}_\theta(Y))$, the optimal control $\overline{u^{d,l}_{\theta,x_0}}$, the trajectory $\overline{X^{d,l}_{\theta,x_0}}$ and the profiled cost value $S^l_{(n,U)}(Y;\theta,x_0)$. Moreover, the sequences $\left\{R^{d,l}_\theta,\, h^{d,l}_\theta(Y)\right\}_{l\in\mathbb{N}}$, $\left\{\overline{u^{d,l}_{\theta,x_0}}\right\}_{l\in\mathbb{N}}$, $\left\{\overline{X^{d,l}_{\theta,x_0}}\right\}_{l\in\mathbb{N}}$ and $\left\{S^l_{(n,U)}(Y;\theta)\right\}_{l\in\mathbb{N}}$ are uniformly convergent in l [Cimen and Banks, 2004b, Cimen and Banks, 2004a]. Thus, we can propose the following algorithm to compute $(R^d_\theta,\, h^d_\theta(Y))$, $\overline{u^d_{\theta,x_0}}$, $\overline{X^d_{\theta,x_0}}$ and $S_{(n,U)}(Y;\theta,x_0)$.

1. Initialization: $\overline{X^{d,0}_{\theta,x_0}}(t^d_j) = x_0$, $A^0_\theta(t^d_j) = A_\theta(x_0,t^d_j)$ for all $j\in[\![0,\, m]\!]$.
2. At iteration l: use Proposition 4 to obtain $(R^{d,l}_\theta,\, h^{d,l}_\theta(Y))$, $\overline{u^{d,l}_{\theta,x_0}}$, $\overline{X^{d,l}_{\theta,x_0}}$, $S^l_{(n,U)}(Y;\theta,x_0)$.
3. If $\left|S^l_{(n,U)}(Y;\theta,x_0) - S^{l-1}_{(n,U)}(Y;\theta,x_0)\right| < \varepsilon_1$ and $\sum_{j=1}^m\left\|\overline{X^{d,l}_{\theta,x_0}}(t^d_j) - \overline{X^{d,l-1}_{\theta,x_0}}(t^d_j)\right\|^2_2 < \varepsilon_2$ with $(\epsilon_1,\, \epsilon_2)$ two strictly positive constants, then step 4; otherwise return to step 2.

4. Set $(R_\theta^d,\, h_\theta^d(Y)) = (R_\theta^{d,l},\, h_\theta^{d,l}(Y))$, $\overline{u_{\theta,x_0}^d} = \overline{u_{\theta,x_0}^{d,l}}$, $\overline{X_{\theta,x_0}^d} = \overline{X_{\theta,x_0}^{d,l}}$, $S_{(n,U)}(Y;\theta,x_0) = S_{(n,U)}^l(Y;\theta,x_0)$.

For $S_{(n,U)}^{CI}$ the procedure is similar, only the initialization step has the be replaced: the initial state $\overline{X_\theta^{d,0}}$ has to be chosen and $A_\theta^0(t_j^d) = A_\theta(\overline{X_\theta^{d,0}}(t_j^d), t_j^d)$. We see in Section 5 what choice we made in practice.

4 Asymptotic analysis

Here we assume the discretization grid is the set of observation time points i.e. $\left\{t_j^d\right\} = \{t_i\}$ which are regularly spaced so $\triangle_i = \triangle = \frac{T}{n}$, and i.i.d $\epsilon_i \sim N(0, \sigma^2 I_{d'})$, we also consider U depends on n and can be written $U = U'/\triangle$ with U' positive definite. Since $\arg\min_{(\theta,x_0)} \triangle S_{(n,U'/\triangle)}^l(Y;\theta,x_0) = \arg\min_{(\theta,x_0)} S_{(n,U)}^l(Y;\theta,x_0)$ and $\arg\min_\theta \left\{\min_{x_0} \triangle S_{(n,U'/\triangle)}^l(Y;\theta,x_0)\right\} = \arg\min_\theta \left\{\min_{x_0} S_{(n,U)}^l(Y;\theta,x_0)\right\}$, we focus on $\triangle S_{(n,U'/\triangle)}^l(Y;\theta,x_0)$ instead of $S_{(n,U)}^l(Y;\theta,x_0)$ for the purpose of asymptotic analysis of $\left(\widehat{\theta}_U^T,\, \widehat{x_{0,U}}^T\right)$ and $\theta_U^{T,CI}$. The proofs are given in *supplementary materials*.

4.1 Asymptotic analysis of $\left(\widehat{\theta}_U^T,\, \widehat{x_{0,U}}^T\right)$ in parametric case

4.1.1 Required conditions

First, we introduce the asymptotic counterpart of $\triangle S_{(n,U'/\triangle)}^l(Y;\theta,x_0)$ when $n \longrightarrow \infty$ and $l \longrightarrow \infty$. In this asymptotic framework, we have access to the true continuous signal $t \longrightarrow Y^*(t) = CX_{\theta^*,x_0^*}(t)$ and so we can define the continuous cost:

$$C_{(T,U')}^\infty(\theta,x_0,u) \;=\; d'\sigma^2 + \int_0^T \left(\left\|CX_{\theta,x_0,u}^\infty(t) - Y^*(t)\right\|_2^2 + u(t)^T U' u(t)\right) dt, \tag{22}$$

its profiled couterpart, $S_{U'}^\infty(\theta,x_0) := \inf_u C_{(T,U')}^\infty(\theta,x_0,u)$, the associated ODE

$$\begin{cases} \dot{X_{\theta,x_0,u}^\infty} = A_\theta(\overline{X_{\theta,x_0}^\infty}(t),t)X_{\theta,x_0,u}^\infty + Bu(t) \\ X_{\theta,x_0,u}^\infty(0) = x_0 \end{cases} \tag{23}$$

and Riccati equation

$$\begin{cases} \dot{R_\theta^\infty}(t) = -C^TC - A_\theta(\overline{X_\theta^\infty}(t),t)^T R_\theta^\infty(t) - R_\theta^\infty(t)A_\theta(\overline{X_\theta^\infty}(t),t) + R_\theta^\infty(t)BU'^{-1}B^T R_\theta^\infty(t) \\ \dot{h_\theta^\infty}(t) = C^TY^*(t) - A_\theta(\overline{X_\theta^\infty}(t),t)^T h_\theta^\infty(t) + R_\theta^\infty(t)BU'^{-1}B^T h_\theta^\infty(t) \end{cases} \tag{24}$$

with $\left(R_\theta^l(T),\, h_\theta^l(T)\right) = (0_{d,d},\, 0_{d,1})$. Now, we present the conditions required for asymptotic analysis.

Condition C1: For all $t \in [0,\, T]$ and for all $\theta \in \Theta$, $x \longmapsto A_\theta(x,t)$ has a compact support Λ.

Condition C2: For all $x \in \Lambda$, $\theta \longmapsto A_\theta(x,.)$ is continuous on Θ and $\forall \theta \in \Theta$, $(x,t) \longmapsto A_\theta(x,t)$ is continuous on $\Lambda \times [0,\, T]$.

Condition C3: Matrix B has independent columns.

Condition C4: The parameters (θ^*, x_0^*) belong to the interior of $\Theta \times \chi$.

Condition C5: The solution X_{θ,x_0} of (4) for $u = 0$ is such that if $CX_{\theta,x_0}(t) = CX_{\theta^*,x_{0^*}}(t)$ for all $t \in [0,\, T]$ then $(\theta, x_0) = (\theta^*, x_{0^*})$.

Condition C6: For all $x \in \Lambda$, $\theta \longmapsto A_\theta(x,.)$ is twice differentiable on Θ and for all $\theta \in \Theta$, $(x,t) \longmapsto \frac{\partial A_\theta(x,t)}{\partial \theta}$ and $(x,t) \longmapsto \frac{\partial^2 A_\theta(x,t)}{\partial^2 \theta}$ are continuous on $\Lambda \times [0,\, T]$.

Condition C7: The asymptotic hessian matrix $\frac{\partial^2 S_{U'}^\infty(\theta^*, x_0^*)}{\partial^2(\theta, x_0)}$ is nonsingular.

Conditions C1, C2, C3 are required for the uniform convergence of $R_\theta^{d,l}$, $h_\theta^{d,l}$ to R_θ^∞, h_θ^∞ and $S_{(U',n)}^l$ to $S_{U'}^\infty$. Conditions C4 and C5 ensure (θ^*, x_0^*) is a well-separated minimum of $S_{U'}^\infty$ and conditions C6 and C7 guarantee that the asymptotic variance-covariance of θ^* is non singular.

4.1.2 Consistency

The estimator $\left(\widehat{\theta}_U^T,\, \widehat{x_{0,U}}^T\right)$ is defined as an M-estimator, so for consistency we need to show $S_{U'}^\infty(\theta, x_0)$ has a global well-separated minimum at $(\theta,\, x_0) = (\theta^*,\, x_0^*)$ and that $S_{(n,U')}^l(Y; \theta, x_0)$ converges uniformly to $S_{U'}^\infty(\theta, x_0)$ on $\Theta \times \chi$. This is the point of the next two propositions.

Proposition 5. *Under conditions C1 to C5,* $(\theta^*,\, x_0^*)$ *is the unique global minimizer of* $S_{U'}^\infty(\theta, x_0)$ *on* $\Theta \times \chi$.

Proposition 6. *Under conditions C1 to C5,*

$$\sup_{(\theta,x_0)\in\Theta\times\chi} \left| S_{U'}^\infty(\theta, x_0) - \triangle S_{(n,U'/\triangle)}^l(Y; \theta, x_0) \right| = o_l(1) + o_{p,n}(1).$$

From this, we use Theorem 5.7 in [van der Vaart, 1998] to conclude about the consistency.

Theorem 7. *Under conditions C1 to C5,* $\left(\widehat{\theta}_U^T,\ \widehat{x_{0,U}}^T\right) \longrightarrow (\theta^*,\ x_0^*)$ *in probability when* $(l,n) \longrightarrow \infty$.

Remark 8. *Interestingly, in [Clairon and Brunel, 2018], for the weighting matrix under the form* $U' = \lambda' I_{d_u}$, *consistency proof for* $\widehat{\theta}_U^T$ *requires the lower bound condition* $\lambda' > \overline{\lambda_1}$ *with* $\overline{\lambda_1}$ *a positive model-dependent bound. Here, we just need to have* U' *positive definite.*

4.1.3 Asymptotic normality

We show the asymptotic normality with $\sqrt{n}$-convergence rate in two steps. First, we derive a linear asymptotic representation of $\left(\widehat{\theta}_U^T,\ \widehat{x_{0,U}}^T\right) - (\theta^*,\ x_0^*)$ through a second order Taylor expansion of $(\theta,\ x_0) \longmapsto \triangle S^l_{(n,U'/\triangle)}(Y;\theta,x_0)$. Second, we approximate this linear asymptotic representation in order to make explicit its dependence with respect to measurement noise.

Proposition 9. *Under conditions C1 to C6, we have:*

$$-\nabla_{(\theta,x_0)}(\triangle S^l_{(n,U'/\triangle)}(Y;\theta^*,x_0^*) = (\frac{\partial^2 S^\infty_{U'}(\theta^*,x_0^*)}{\partial^2\,(\theta,x_0)} + o_{p,n}(1) + o_l(1))\left(\widehat{\theta}_U^T - \theta^*,\ \widehat{x_{0,U}}^T - x_0^*\right).$$

Proposition 10. *Under conditions C1 to C6, we have*

$$-\nabla_{(\theta,x_0)}(\triangle S^l_{(n,U'/\triangle)}(Y;\theta^*,x_0^*) = \left(\triangle\sum_{j=0}^{n}\epsilon_j^T\right)(K^\infty_{(\theta^*,x_0^*)} + o_l(1) + o_n(1)) + L\left(\triangle\sum_{j=0}^{n}\epsilon_j\right) + o_{p,n}(\sqrt{\triangle}) + o_l(1)$$

with $K^\infty_{(\theta,x_0)} = 2CBU'^{-1}B^T\int_0^T \frac{\partial h^\infty_{\theta^*}(t)}{\partial(\theta,x_0)}dt$ *and* $L = \begin{pmatrix} 0_{d',p} & -2C \end{pmatrix}^T$.

From this, we use $\frac{\partial^2 S^\infty_{U'}(\theta^*,x_0^*)}{\partial^2(\theta,x_0)}$ nonsingularity and the central limit theorem to obtain the following.

Theorem 11. *Under conditions C1 to C7 and if* $l = O_n(\sqrt{\Delta})$, $\left(\widehat{\theta}_U^T,\ \widehat{x_{0,U}}^T\right)$ *is asymptotically normal and* $\left(\widehat{\theta}_U^T,\ \widehat{x_{0,U}}^T\right) - (\theta^*,x_0^*) = o_{p,n}(n^{-\frac{1}{2}})$.

The required conditions on l for consistency and $\sqrt{n}$-convergence rate are necessary only for non-linear systems. For linear models, we use (18) to compute $S^l_{(n,U')}(Y;\theta)$ and we take $o_l(1) = 0$.

4.2 Asymptotic analysis of $\widehat{\theta}_U^{T,CI}$ for linear models in parametric case

For the asymptotic analysis of $\widehat{\theta}_U^{T,CI}$, we restrict to the linear models. Since A_θ does not depend on x, there is no need to consider asymptotics in l. The conditions are shown below.

Condition L1: For all $\theta \in \Theta$, $t \longmapsto A_\theta(t)$ is continuous on $[0, T]$.

Condition L2: $\theta \longmapsto A_\theta$ is continuous on Θ.

Condition L3: For all $\theta \in \Theta$, $R_\theta^\infty(0)$ is nonsingular.

Condition L4: The true parameter θ^* belongs to the interior of Θ.

Condition L5: The solution X_{θ,x_0} of (4) for $u = 0$ is such that if $CX_{\theta,x_0}(t) = CX_{\theta^*,x_0^*}(t)$ for all $t \in [0, T]$ then $(\theta, x_0) = (\theta^*, x_0^*)$.

Condition L6: $\theta \longmapsto A_\theta$ is C^2 on Θ.

Condition L7: The asymptotic hessian matrix $\frac{\partial^2 S_{U'}^{CI}(\theta^*)}{\partial^2 \theta}$ is nonsingular.

The proofs follow the same steps as in the previous sections, hence we just present the theorems, they are also detailed in *supplementary materials*.

Theorem 12. *Under conditions LC1 to LC5, we have* $\widehat{\theta}_U^{T,CI} \longrightarrow \theta^*$ *in probability when* $n \longrightarrow \infty$.

Theorem 13. *Under conditions LC1 to LC7,* $\widehat{\theta}_U^{T,CI}$ *is asymptotically normal and* $\widehat{\theta}_U^{T,CI} - \theta^* = o_{p,n}(n^{-\frac{1}{2}})$.

Remark 14. *The difficulty in deriving the asymptotic behavior of* $\widehat{\theta}_U^{T,CI}$ *in all generality comes from the initialisation point* x_0^r *required by the algorithm. So far, we have been unable to analyze the mapping* $Q_\theta : x_0^r \longmapsto \overline{X_\theta}(., x_0^r)$ *where* $\overline{X_\theta}(., x_0^r)$ *is the trajectory given by the algorithm in the limit case* $n = \infty$ *and* $l = \infty$. *If for* $\theta = \theta^*$, *the true trajectory* X^* *is a global attractor of* Q_{θ^*}, *the demonstrations will be completed, but our attempts to prove it remain unfruitful.*

5 Experiments

We use Monte-Carlo simulations on different models, for several numbers of measures n and corrupted with measurement noise of different magnitudes. We compare four estimators: $\hat{\theta}_U^T$ and $\hat{\theta}_U^{T,CI}$, the nonlinear least square (NLS) estimator $\widehat{\theta}^{NLS}$ and the generalized profiling (GP) estimator $\widehat{\theta}^{GP}$ introduced in [Ramsay et al., 2007]. The latter is the regularization method of reference for the estimation problem in ODEs. We compare $\hat{\theta}_U^T$, $\hat{\theta}_U^{T,CI}$, $\widehat{\theta}^{NLS}$, $\widehat{\theta}^{GP}$ on models facing practical identifiability problems in correctly and misspecified frameworks. For a given choice of (n, σ), we compute:

1. The variance $V(\widehat{\theta}_i)$ for each element θ_i of θ to analyze how each estimator behaves specifically for the components suffering from identifiability issues.
2. The estimator variance-covariance norm $\left\| V\left(\widehat{\theta}\right) \right\|_2$ to analyze how each estimator behaves for the whole parameter set.
3. The componentwise mean square error $\text{MSE}(\widehat{\theta}_i) = \left| \theta_i^* - \widehat{\mathbb{E}}\left[\widehat{\theta}_i\right] \right|^2 + V(\widehat{\theta}_i)$ and the global $\text{MSE}(\widehat{\theta}) = \sum_{i=1}^p \left| \theta_i^* - \widehat{\mathbb{E}}\left[\widehat{\theta}_i\right] \right|^2 + \left\| V\left(\widehat{\theta}\right) \right\|_2$ to measure estimator accuracy, in particular its degradation when facing misspecification.

These quantities are obtained by Monte Carlo procedure based on a number of N_{MC} trials specific to each experimental design. We choose N_{MC} large enough to have the Monte Carlo Error of $\text{MSE}(\widehat{\theta}_i)$ lower than 5% for each θ_i (see [Koehler et al., 2009] for details). For each run, the observations are obtained by integrating the ODE with a Runge-Kutta algorithm (ode function in R), with added centered Gaussian noise of variance σ^2. Parameters have different orders of magnitude, so results are given for normalized estimated values $\widehat{\theta}./\theta^*$, ./ being the componentwise division.

The GP method uses an approximate solution $\tilde{X}_\theta^\lambda$ of the ODE defined as the spline basis decomposition minimizing $\sum_{i=1}^n \left\| y_i - C\tilde{X}_\theta^\lambda(t_i) \right\|^2 + \lambda \left\| \frac{d}{dt}\tilde{X}_\theta^\lambda - f\left(\cdot, \tilde{X}_\theta^\lambda, \theta\right) \right\|_{L^2}^2$. GP requires a selection method for the knots location and the hyperparameter λ. The knots location is specific to each example and λ is selected by using the method presented in [Qi and Zhao, 2010]: the value of λ is increased until $\left\| X_{\hat{\theta}_\lambda^{GP}, \widetilde{x_0}} - \tilde{X}_{\hat{\theta}_\lambda^{GP}} \right\|_{L^2}^2$

starts increasing, that is when $\tilde{X}_{\hat{\theta}_\lambda^{GP}}$ starts to differ significantly from the exact solution $X_{\hat{\theta}_\lambda^{GP},\widetilde{x_0}}$ where $\widetilde{x_0} = \tilde{X}^{\lambda}_{\hat{\theta}_\lambda^{GP}}(0)$.

For $\widehat{\theta}_U^T$ and $\widehat{\theta}_U^{T,CI}$, we need to select the discretization grid $\left\{t_j^d\right\}_{0\leq j\leq m}$ and U. For the grid, we take $m = k_n n$ points and we place uniformly k_n discretization points between two observation times. We choose k_n large enough to correctly estimate the ODE solution. For U, we consider scalar matrices $U = \lambda I_d$. When λ tends to ∞, the criteria tends to the NLS one, if λ tends to 0, the criteria leads to interpolate the observations without any effect of θ. Hence the need for an adaptive selection method. For our approach, we use the forward cross-validation method presented in [G. Hooker and Earn, 2011]. We split $[0, T]$ into H subintervals $[t_h,\, t_{h+1}]$, such that $t_1 = 0$ and $t_H = T$ and we denote $X_\theta(., t_h, x_h)$ the solution of:

$$\begin{cases} \dot{x}(t) = f\left(t, x(t), \theta\right) \\ x(t_h) = x_h \end{cases} \tag{25}$$

defined on the interval $[t_h,\, t_{h+1}]$. The forward cross-validation uses the causal relation imposed to the data by the ODE to quantify the prediction error for the estimator $\hat{\theta}_U^{T,CI}$ (or equivalently $\hat{\theta}_U^T$):

$$\text{ERRPRED}(\lambda) = \sum_{h=1}^{H} \sum_{\{t_i \in [t_h,\, t_{h+1}]\}} \left\| y_i - C X_{\hat{\theta}_U^{T,CI}}(t_i, t_h, \overline{X^d_{\hat{\theta}_U^{T,CI}}}(t_h)) \right\|_2^2 .$$

The rationale of this selection method is the following: if λ is too small, $C\overline{X^d_{\hat{\theta}_U^{T,CI}}}(t_h)$ will be close to the observation y_h but not to the actual ODE solution, and the solution of Eq. (25) will diverge from the observations on $[t_h,\, t_{h+1}]$. If λ is too large, $\overline{X^d_{\hat{\theta}_U^{T,CI}}}(t_h)$ will be close to the original ODE solution but far from y_h and it will lead again to a large value for ERRPRED(λ). Thus, a proper value for λ which minimizes ERRPRED(λ) will be chosen between these two extreme cases. In the simulations, we use $H = 2$ subintervals. We denote $\hat{\theta}^T$ and $\hat{\theta}^{T,CI}$ the values minimizing ERRPRED among the set of tested weighing parameters.

Regarding the initial state $\overline{X_\theta^{d,0}}$ required by the algorithm presented in Section 3.3 when we profile on x_0, we take the measured value for the observed state variables and made simple but model specific choices for the unobserved ones; they will be detailed in the following examples. More refined choices may

be possible for the unobserved state variables but these simple strategies worked well in practice (i.e. no problem of convergence) for all examples given below.

5.1 $\alpha-$Pinene model

We begin with a linear ODE considered in [Rodriguez-Fernandez et al., 2006] and used for modeling the isomerization of $\alpha-$Pinene:

$$\begin{cases} \dot{x}_1 = -(\theta_1 + \theta_2)x_1 \\ \dot{x}_2 = \theta_1 x_1 \\ \dot{x}_3 = \theta_2 x_1 - (\theta_3 + \theta_4)x_3 + \theta_5 x_5 \\ \dot{x}_4 = \theta_3 x_3 \\ \dot{x}_5 = \theta_4 x_3 - \theta_5 x_5 \end{cases} \tag{26}$$

on the observation interval $[0,\, T] = [0,\, 100]$. Here the expression of A_θ is unique and straigthforward to derive. We set $x_0^* = (100,\, 0,\, 0,\, 0,\, 0)$ and $\theta^* = (5.93\,, 2.96,\, 2.05,\, 27.5,\, 4) \times 10^{-2}$. We plot in Figure 1 the solution of (26) corresponding to θ^* and an example of simulated observations.

In [Rodriguez-Fernandez et al., 2006], model (26) is used as a benchmark estimation comparison as many approaches fail to converge due to the difficulty of estimating θ_4^* and θ_5^* because of the high correlation between them. We select $k_n = 50$ and λ among the set $\left\{10^i,\, 5 \times 10^i\right\}_{0 \leq i \leq 2}$. For GP, we use 50 knots uniformly reparted on $[0,\, T]$.

Influence of measurement noise We consider one sample size $n = 10$. The level of noise is specific to each state variable to take into account their different order of magnitude. Each state X_i is corrupted by a measurement noise of standard deviation $\frac{\sigma}{100} \times \|X_i\|_{L^2}$, three levels for σ are tested ($\sigma = 5$, $\sigma = 10$ and $\sigma = 15$). We choose $N_{MC} = 200$ for $\sigma = 5$ and $\sigma = 10$ but we need to take $N_{MC} = 500$ for $\sigma = 15$. Results are presented in Table 1. For θ_4 and θ_5, we observe that $\widehat{\theta}^T$ and $\widehat{\theta}^{T,CI}$ give the smallest variance. Our approximate method regularizes the estimation of parameters facing a practical identifiability problem in comparison with NLS. Moreover, we notice the same pattern for $\|V(\theta)\|_2$ which takes into account covariance among parameters. However, TE and GP are methods based on approximated solutions and so

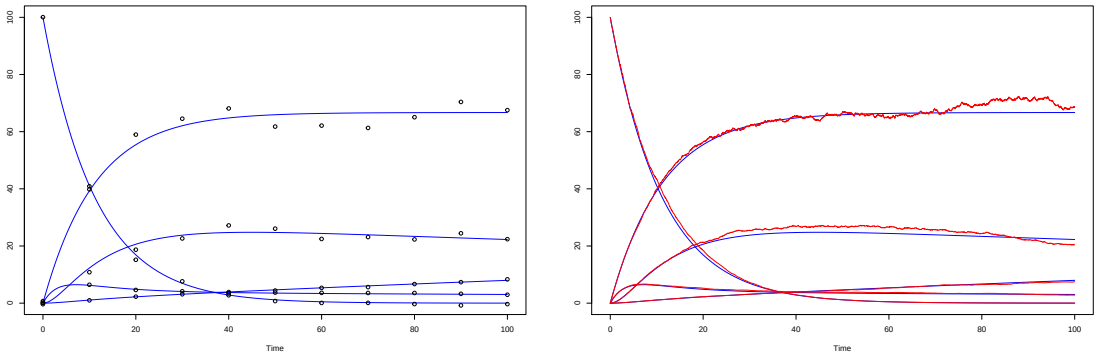

Figure 1: Left: Solution of (26) (blue) and noisy observations for $\sigma = 5$ (circle). Right: Solution of (26) (blue) and a realization of $dX = A_\theta X dt + c_t.X dt$ (red) for $\sigma_c^2 = 0.004$.

produce biased estimates. That is why we estimated the mean square error to verify that the price to pay to decrease the variance is not too high in terms of bias. Our methods have lower global mean square error than NLS which indicate a reasonable bias. GP on the other hand can have a very large mean square error. The reason, already been discussed in [Clairon and Brunel, 2018, Brunel and Clairon, 2015], is linked to the limited ability of $\tilde{X}_\theta^\lambda$ to approach the true solution. In contrast, for our method the mesh size can be arbitrarily small and thus $X_{\theta,u}^d$ can be arbitrarily close to the original ODE model.

Influence of model misspecification We set $(n, \sigma) = (10, 5)$ and the observations are now generated by using the stochastically perturbed model $dX = A_\theta X dt + c_t.X dt$ with c_t a random vector of length 5 with independant components $c_{j,t}$ such that $c_{j,t} \sim N(0, \sigma_c^2)$, the product $c_t.X$ is componentwise. We still estimate θ^* by using model (26), which is now a deterministic approximation of the true process. We plot in Figure 1 the solution of (26) and one realization of its perturbed version for the sake of comparison.

This experimental design has been chosen to mimic a real case of data analysis for chemical processes where the deterministic reaction rate equations are used as an approximation of stochastic differential equations [Gillespie, 2000]. We study the effect of misspecification by varying the value of σ_c^2 and results are presented in Table 2, here setting $N_{MC} = 200$ for every σ_c^2 values was enough to obtain accurate

Table 1: MSE and Variance (in parenthesis) for $\alpha-$Pinene model (26) and $n = 10$.

σ	$\times 10^{-2}$	θ_1	θ_2	θ_3	θ_4	θ_5	θ
5	$\widehat{\theta}^{T,CI}$	0.3 (0.2)	0.5 (0.4)	3.7 (0.5)	1.9 (1.2)	4.1 (2.4)	9.6 (3.8)
	$\widehat{\theta}^{T}$	0.3 (0.2)	0.5 (0.4)	3.8 (3.7)	2.0 (1.4)	5.1 (3.0)	11.0 (4.6)
	$\widehat{\theta}^{NLS}$	0.3 (0.2)	0.5 (0.5)	3.9 (3.6)	2.1 (1.6)	5.9 (3.5)	11.9 (5.3)
	$\widehat{\theta}^{GP}$	0.3 (0.2)	0.9 (0.8)	6.2 (4.0)	10.6 (10.5)	21.5 (18.7)	34.1 (28.9)
10	$\widehat{\theta}^{T,CI}$	0.9 (0.8)	1.5 (1.4)	4.7 (1.6)	6.3 (6.3)	20.2 (13.5)	30.4 (20.4)
	$\widehat{\theta}^{T}$	0.8 (0.7)	1.5 (1.4)	4.5 (1.1)	8.9 (8.6)	26.6 (18.2)	39.4 (27.3)
	$\widehat{\theta}^{NLS}$	0.8 (0.7)	1.7 (1.6)	4.4 (1.1)	10.2 (10.0)	30.8 (21.1)	49.5 (31.6)
	$\widehat{\theta}^{GP}$	0.6 (0.5)	3.8 (3.5)	8.4 (5.2)	15.9 (15.7)	29.4 (27.7)	48.3 (43.1)
15	$\widehat{\theta}^{T,CI}$	2.1 (1.8)	2.9 (2.8)	6.6 (3.3)	10.1 (9.0)	24.3 (20.1)	39.0 (31.5)
	$\widehat{\theta}^{T}$	1.8 (1.7)	3.2 (3.2)	6.3 (2.7)	12.1 (12.0)	31.0 (26.2)	47.8 (38.9)
	$\widehat{\theta}^{NLS}$	1.8 (1.7)	3.8 (3.8)	6.4 (2.6)	14.9 (14.9)	38.9 (32.8)	58.7 (48.6)
	$\widehat{\theta}^{GP}$	1.4 (1.3)	3.9 (3.3)	14.5 (10.1)	28.1 (27.7)	62.1 (53.4)	94.4 (80.6)

Monte-Carlo estimate. In comparison to NLS, our estimators limit the drop in estimation accuracy due to misspecification effect. This illustrates the benefits of taking into account model discrepancy for estimation in presence of model error.

Influence of sample size The interest of considering discrete optimal control theory over a continuous framework as in [Clairon and Brunel, 2019, Clairon and Brunel, 2018] is clear in terms of running time and accuracy in small sample case. To illustrate that we plot in figure 2, the evolution of θ^T (circle), $\theta^{T,CI}$ (square) and θ^K, the estimator presented in [Clairon and Brunel, 2019] for linear ODEs with unknown x_0, when the sample size n increases from 10 to 50. For θ^T and $\theta^{T,CI}$, k_n is selected such that the discretization grid is always made of 500 points. One see clearly the advantage brought in terms of computational time by the discrete approach. Moreover, even though the bias and MSE are getting closer for every methods when n increases, the benefits of using θ^T, $\theta^{T,CI}$ comparing to θ^K is clear when n is low. This drop in accuracy for θ^K comes from the committed error during the required nonparametric estimation of the

Table 2: MSE and Variance (in parenthesis) for misspecified $\alpha-$Pinene model and $(n, \sigma) = (10, 5)$.

σ_c^2	$\times 10^{-2}$	θ_1	θ_2	θ_3	θ_4	θ_5	θ
0.004	$\widehat{\theta}^{T,CI}$	0.3 (0.2)	0.7 (0.4)	7.3 (0.5)	2.6 (0.9)	2.0 (1.8)	12.2 (3.1)
	$\widehat{\theta}^{T}$	0.3 (0.2)	0.7 (0.4)	0.7 (0.4)	2.9 (0.1)	2.1 (2.0)	12.7 (3.3)
	$\widehat{\theta}^{NLS}$	0.2 (0.2)	0.7 (0.4)	7.5 (0.4)	3.1 (1.1)	2.3 (2.2)	13.2 (3.6)
	$\widehat{\theta}^{GP}$	0.2 (0.2)	1.8 (1.3)	8.1 (3.6)	10.1 (9.9)	17.1 (15.8)	31.9 (25.6)
0.006	$\widehat{\theta}^{T,CI}$	0.8 (0.2)	0.6 (0.4)	0.7 (0.5)	1.4 (1.3)	9.4 (1.5)	13.2 (4.1)
	$\widehat{\theta}^{T}$	0.6 (0.2)	1.2 (0.3)	0.7 (0.3)	2.0 (1.9)	12.5 (3.7)	16.4 (5.7)
	$\widehat{\theta}^{NLS}$	0.5 (0.2)	1.9 (0.3)	0.9 (0.3)	2.5 (2.2)	14.2 (4.4)	19.0 (6.7)
	$\widehat{\theta}^{GP}$	0.5 (0.3)	1.9 (1.2)	4.5 (3.9)	20.1 (19.6)	45.6 (37.6)	67.0 (56.7)
0.008	$\widehat{\theta}^{T,CI}$	0.9 (0.2)	0.7 (0.4)	16.3 (0.3)	6.2 (0.8)	3.4 (1.5)	27.0 (2.4)
	$\widehat{\theta}^{T}$	1.1 (0.2)	0.4 (0.3)	17.8 (0.2)	9.4 (0.9)	6.7 (1.7)	34.9 (2.7)
	$\widehat{\theta}^{NLS}$	1.1 (0.2)	0.4 (0.3)	18.1 (0.2)	11.3 (0.9)	8.9 (1.7)	39.0 (2.7)
	$\widehat{\theta}^{GP}$	0.9 (0.4)	1.5 (1.09)	19.2 (4.7)	50.8 (48.6)	98.0 (86.2)	164 (135)

curve.

5.2 Repressilator model

We present the Repressilator model proposed in [Elowitz and Leibler, 2000] for the study of a genetic regulation network. It is made of a feedback loop of 3 couples (mRNA, protein), denoted $(r_i, p_i)_{1 \leq i \leq 3}$, in which each protein inhibits the next gene transcription in the loop:

$$\left\{ \begin{array}{rcl} \dot{r_i} & = & \frac{v_i k_{i,[i+1]}^n}{p_{[i+1]}^n + k_{i,[i+1]}^n} - k_i^g r_i \\ \dot{p_i} & = & k_i r_i - k_i^p p_i. \end{array} \right. \tag{27}$$

We aim to estimate $\theta^* = \left(v_1^*, v_2^*, v_3^*, k_{1,2}^*, k_{2,3}^*, k_1^{g*}, k_2^{g,*} k_3^{g*}, k_1^{p*}, k_2^{p*}, k_3^{p*}\right) = (50, 100, 80, 50, 30, 1, 1, 1, 1, 2, 3)$ with initial conditions $\left(r_{1,0}^*, r_{2,0}^*, r_{3,0}^*\right) = (60, 20, 6)$ and $\left(p_{1,0}^*, p_{2,0}^*, p_{3,0}^*\right) = (18, 27, 1)$. We consider that only the mRNA concentrations are measured on $[0, T] = [0, 20]$ and for structural identifiability reasons we set $(k_{3,1}, k_1, k_2, k_3, n) = (40, 5, 6, 7, 3)$ and consider $\left(p_{1,0}^*, p_{2,0}^*, p_{3,0}^*\right)$ are known. We plot in Figure 3 the

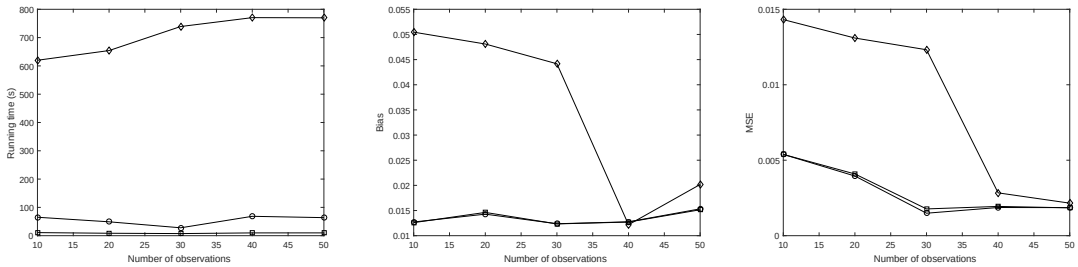

Figure 2: Average running time, bias and MSE for θ^T (circle), $\theta^{T,CI}$ (square) and θ^K (diamond).

solution of (27), for $\theta = \theta^*$. Here, we choose, $A_\theta(r,p,t) = \begin{pmatrix} -k_1^g & 0 & 0 & 0 & 0 & 0 & \frac{v_1 k_{1,2}^n}{p_2^n + k_{1,2}^n} \\ 0 & -k_2^g & 0 & 0 & 0 & 0 & \frac{v_3 k_{2,3}^n}{p_3^n + k_{2,3}^n} \\ 0 & 0 & -k_3^g & 0 & 0 & 0 & \frac{v_1 k_{3,1}^n}{p_1^n + k_{3,1}^n} \\ k_1 & 0 & 0 & -k_1^p & 0 & 0 & 0 \\ 0 & k_2 & 0 & 0 & -k_2^p & 0 & 0 \\ 0 & 0 & k_3 & 0 & 0 & -k_3^p & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$ where a constant artificial state variable $Z = 1$ has been added. This model has been identified as sloppy in [Gutenkunst et al., 2007], the eigendecomposition of $\mathcal{I}_n(\theta^*, x_0^*)$ for $n = 25$ indicates the subset of parameters $\theta_1^* = \left(v_1^*, v_2^*, v_3^*, k_{1,2}^*, k_{2,3}^*\right)$ corresponds to the lowest eigenvalues. Henceforth, we separate θ^* into θ_1^* and $\theta_2^* = \left(k_1^{g*}, k_2^{g*}, k_3^{g*}, k_1^{p*}, k_2^{p*}, k_3^{p*}\right)$ for presenting the estimation results and in particular analyze how the methods behave with θ_1^*. The results presented for the variance (resp. mean square error) for θ_1 and θ_2 denote the sum of the variance (resp. mean square error) of θ_1 and θ_2 components. We take $k_n = 20$ and select λ among $\{10, 20, 50, 100, 200\}$. For the unobserved part of $\overline{X_\theta^{d,0}}$ when profiling on initial condition, we choose $\left(p_{1,0}^*, p_{2,0}^*, p_{3,0}^*\right)$ on the whole observation interval as initial guess.

Influence of measurement noise We take $n = 25$ and consider three levels of measurement noise ($\sigma = 1$, $\sigma = 1.5$ and $\sigma = 2$). Results are presented in Table 3 (left). We were unable to obtain results for GP because of an important number of algorithmic failures during simulations (almost 80% of the runs) due to practical identifiability issues. Indeed, GP requires the introduction of nuisance parameters β

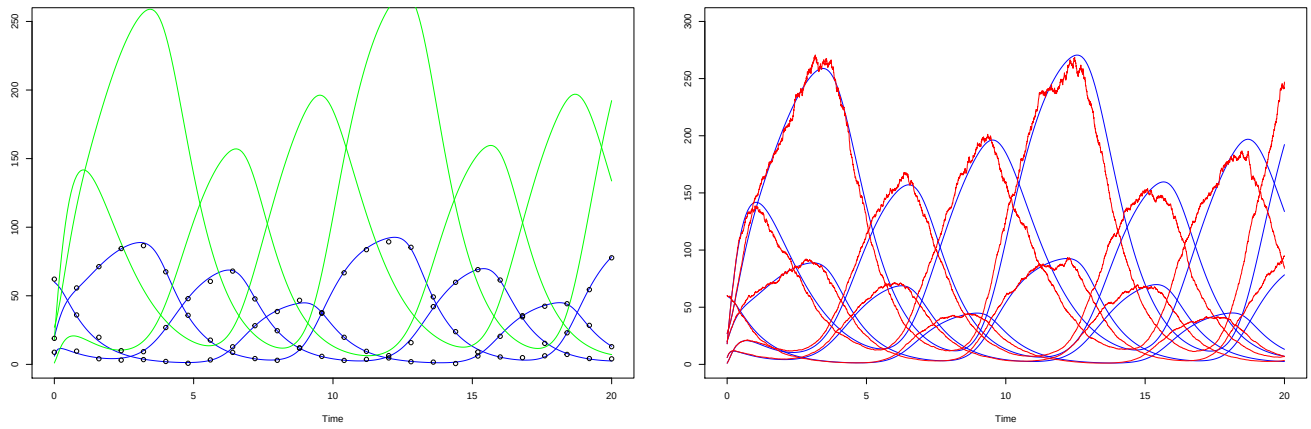

Figure 3: Left: Solution of (27) with proteins in green, mRNAs in blue and noisy observations for $\sigma = 1$ (circle). Right: Solution of (27) (blue) and a realization of (28) (red) for $\sigma_c^2 = 0.5$.

needed for obtaining a smooth curve estimator $\tilde{X}_\theta^\lambda$ which can lead to overfitting with diverging parameter estimates. In a partially observed framework, even for a $\widehat{\theta^{GP}}$ value far from θ^*, the observed part of the smooth curve $\tilde{X}^\lambda_{\widehat{\theta^{GP}}}$ can remain close to the observations because the parameters $\widehat{\beta}_\lambda$ can counteract the effects of $\widehat{\theta^{GP}}$. Our method improves the estimation of the subset of sloppy parameters. Moreover, our method globally improves the committed error when all parameters are simultaneously estimated, which is the recommended procedure in sloppy models [Gutenkunst et al., 2007].

Influence of model misspecification We set $(n, \sigma) = (25, 1)$ and the observations are now generated by a stochastically perturbed version of the original ODE:

$$\begin{cases} dr_i &= \left(\frac{v_i k^n_{i,[i+1]}}{p^n_{[i+1]} + k^n_{i,[i+1]}} - k^g_i r_i \right) dt + c_t r_i dt \\ dp_i &= (k_i r_i - k^p_i p_i)\, dt + c_t p_i dt \end{cases} \tag{28}$$

with $c_t \sim N(0, \sigma_c^2)$. We plot in Figure 3 the solution of (27) and one realization of (28) for the sake of comparison. Results are presented in Table 3 (right), they confirm the advantages of using an estimation method based on a relaxation of the original model in the presence of model error.

Table 3: MSE and Variance (in parenthesis) for Repressilator model (27). Left: Well-specified model for $n = 25$. Right: Misspecified case for $(n, \sigma) = (25, 1)$.

σ	$\times 10^{-2}$	θ_1	θ_2	θ
1	$\widehat{\theta}^{T,CI}$	1.22 (0.98)	0.75 (0.68)	1.17 (0.90)
	$\widehat{\theta}^{T}$	0.90 (0.70)	0.60 (0.51)	0.92 (0.65)
	$\widehat{\theta}^{NLS}$	1.61 (1.60)	1.15 (1.14)	1.61 (1.59)
1.5	$\widehat{\theta}^{T,CI}$	1.98 (1.73)	1.29 (1.23)	1.89 (1.63)
	$\widehat{\theta}^{T}$	1.41 (1.26)	1.01 (0.91)	1.36 (1.09)
	$\widehat{\theta}^{NLS}$	4.66 (4.59)	2.86 (2.84)	4.78 (4.70)
2	$\widehat{\theta}^{T,CI}$	3.19 (2.77)	2.42 (2.29)	3.15 (2.59)
	$\widehat{\theta}^{T}$	2.62 (2.28)	1.92 (1.77)	2.74 (2.26)
	$\widehat{\theta}^{NLS}$	6.11 (6.05)	4.39 (4.34)	5.97 (5.87)

σ_c^2	$\times 10^{-2}$	θ_1	θ_2	θ
0.5	$\widehat{\theta}^{T,CI}$	3.33 (2.94)	2.10 (2.02)	2.96 (2.48)
	$\widehat{\theta}^{T}$	2.22 (1.98)	1.70 (1.60)	2.36 (2.01)
	$\widehat{\theta}^{NLS}$	7.62 (7.52)	5.78 (5.57)	9.04 (8.78)
1	$\widehat{\theta}^{T,CI}$	2.33 (2.89)	2.16 (2.00)	3.20 (2.60)
	$\widehat{\theta}^{T}$	2.50 (2.26)	1.87 (1.73)	2.30 (1.91)
	$\widehat{\theta}^{NLS}$	8.20 (8.18)	6.51 (6.35)	10.03 (9.85)
1.5	$\widehat{\theta}^{T,CI}$	4.87 (4.68)	3.33 (3.29)	4.35 (4.12)
	$\widehat{\theta}^{T}$	3.11 (2.87)	2.12 (1.99)	2.78 (2.41)
	$\widehat{\theta}^{NLS}$	19.21 (18.53)	25.1 (22.2)	30.88 (27.31)

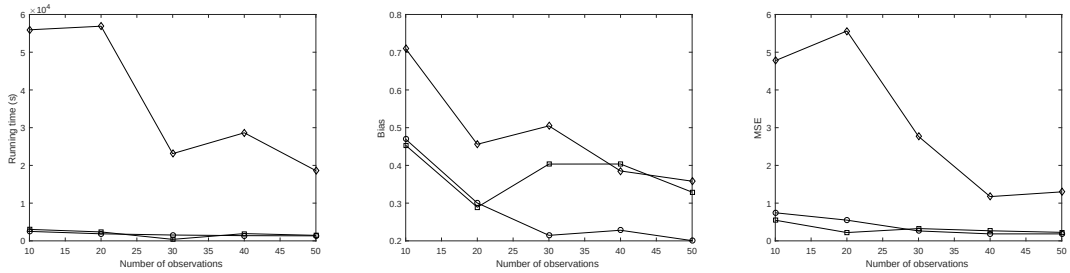

Figure 4: Average running time, bias and MSE for θ^T (circle), $\theta^{T,CI}$ (square) and θ^P (diamond).

Influence of sample size We compare the evolution of θ^T, $\theta^{T,CI}$ and θ^P, the estimator presented in [Clairon and Brunel, 2018] for nonlinear ODEs, for a varying sample size n, the results are presented in figure 4. For θ^T and $\theta^{T,CI}$, we selected a discretization grid of 300 points. We draw similar conclusions as in the α-Pinene model case.

5.3 FitzHugh-Nagumo with a functional parameter

We consider a modified version of the FitzHugh-Nagumo model proposed by [FitzHugh, 1961] to study the membrane potential evolution of neurons:

$$\left\{ \begin{array}{rcl} \dot{V} &=& c\left(V - \frac{V^3}{3} + R\right) \\ \dot{R} &=& -\frac{1}{c}\left(V - a(t) + bR\right), \end{array} \right. \tag{29}$$

V is the neuron membrane potential, R the synaptic conductance and we assume that only V is observed on $[0,\, T] = [0,\, 20]$. Here, the original parameter a^* is replaced by the function $a^*(t) = 0.2\left(1 + \sin(\frac{t}{5})\right)$ and the other parameters are set to $(b^*, c^*) = (0.2, 3)$ and $x_0^* = (V_0^*,\, R_0^*) = (-1, 1)$. Our aim is to compare the variational approach presented in Section 2.2 with a classic basis decomposition method for the simultaneous estimation of $\theta^* = (b^*,\, c^*)$ and $\vartheta^* = a^*$. For our semiparametric estimation method, we need to modify the matrix A_θ^e in the cost (10) comparing to the expression we derive in section 2.2: $A_\theta^e(V, R, z_1, z_2, t) = \begin{pmatrix} c\left(1 - V^2/3\right) & cR & 0 & 0 \\ -\frac{1}{c} & -b & \frac{1}{c} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ for having z_1 as the estimator of a^* and z_2 as its second derivative. This modification is required to absorb the exogenous term $1/c$ in ODE (29). We take $k_n = 20$ and select (λ_1, λ_2) among $\{(0.005, 0.005), (0.01, 0.01), (0.05, 0.05)\}$. For the unobserved variables of $\overline{X_\theta^{d,0}}$ when profiling on initial condition, we set their values to one on the whole observation interval. For GP, we put one knot at each observation time. To estimate a^* with NLS and GP, we use the finite basis approximation $\widehat{a}(t) \simeq \sum_{i=1}^{K_r} \beta_{r,K_r} p_i(t)$, here $\{p_i\}_i$ is a B-Spline basis with a uniform knot sequence. The additional K_r parameters $(\beta_{r,1},\, \ldots,\, \beta_{r,K_r})$ are estimated by introducing the extended set $\theta^{ext} = (\theta,\, \beta_{r,1},\, \ldots,\, \beta_{r,K_r})$ and K_r is selected by minimizing the Akaike Information: $\mathrm{AIC}(\widehat{\theta}^{ext}) = n \log\left(\frac{\sum_i (y_i - C X_{\widehat{\theta}^{ext}, \widehat{x_0}}(t_i))^2}{n}\right) + 2K_r$, where $\widehat{x_0}$ is the standard initial condition estimator for NLS and $\widehat{x_0} = \tilde{X}^{\lambda}_{\hat{\theta}^{GP}_{\lambda}}(0)$ for GP. For an estimator $\widehat{a}$, we quantify its accuracy by computing Monte-Carlo estimator of the integrated version of the variance: $V(\widehat{a}) = \int_0^T \left(\mathbb{E}\left[\widehat{a}^2(t)\right] - \mathbb{E}\left[\widehat{a}(t)\right]^2\right) dt$ and mean square error: $\mathrm{M}^f(\widehat{a}) = \int_0^T \mathbb{E}\left[\left(\widehat{a}(t) - a^*(t)\right)^2\right] dt$.

Table 4: Results for FHN model (29) and $n = 50$.

σ	$\times 10^{-2}$	b	c	θ	a
0.03	$\widehat{\theta}^{T,CI}$	3.88 (2.74)	0.08 (0.08)	3.90 (2.76)	4.11 (3.93)
	$\widehat{\theta}^{T}$	3.92 (2.76)	0.09 (0.08)	3.94 (2.77)	4.08 (3.91)
	$\widehat{\theta}^{NLS}$	5.68 (5.25)	0.05 (0.04)	5.70 (5.53)	5.31 (4.98)
	$\widehat{\theta}^{GP}$	3.17 (2.84)	0.20 (0.01)	3.36 (2.84)	6.86 (2.41)
0.05	$\widehat{\theta}^{T,CI}$	8.10 (6.87)	0.15 (0.12)	8.14 (6.88)	9.97 (9.21)
	$\widehat{\theta}^{T}$	8.11 (6.96)	0.12 (0.10)	8.20 (6.98)	9.91 (9.16)
	$\widehat{\theta}^{NLS}$	8.84 (7.98)	0.06 (0.05)	8.49 (7.89)	21.5 (16.1)
	$\widehat{\theta}^{GP}$	13.9 (13.9)	0.17 (0.04)	14.0 (13.6)	18.6 (16.7)

Influence of measurement noise We set $n = 50$ and consider two levels of measurement noise ($\sigma = 0.03$ and $\sigma = 0.05$). Results are presented in Table 4. Depending on the noise level, our estimators give better or equivalent estimation than NLS for θ but always outperform NLS for functional estimation. The finite basis decomposition used to replace a leads to use an approximated version of the original model for the estimation. This induced misspecification can explain the drop in accuracy for the NLS and GP estimators. Moreover, as pointed out in [Clairon and Brunel, 2018], the selection of a basis and knot location for semiparametric estimation is complicated and model-specific. In our case, the extension of the parametric estimation method to the semi parametric framework is straightforward for hyperparameters selection.

Influence of model misspecification We set $(n, \sigma) = (50, 0.03)$ and the observations are now a realization of the hypoelliptic stochastic differential equation:

$$\begin{cases} dV_t &= c\left(V_t - \frac{V^3}{3} + R_t\right) dt \\ dR_t &= -\frac{1}{c}\left(V_t - a(t) + bR_t\right) dt + \sigma_r dW_t \end{cases} \tag{30}$$

with W_t a Wiener process and σ_r a diffusion parameter but θ^* is still estimated by assuming the deterministic model (30) is true. This model has been proposed to include different sources of noise acting on R_t see [Lindner and Schimansky-Geier, 1999]. Results are presented in Table 5. Once again, our methods give

Table 5: MSE and Variance (in parenthesis) for misspecified FHN model (30) and $(n, \sigma) = (50, 0.03)$.

σ_τ^2	$\times 10^{-2}$	b	c	θ	a
0.1	$\widehat{\theta}^{T,CI}$	4.32 (3.76)	0.11 (0.10)	4.33 (3.78)	5.04 (4.91)
	$\widehat{\theta}^{T}$	4.34 (3.67)	0.11 (0.10)	4.37 (3.68)	5.64 (5.55)
	$\widehat{\theta}^{NLS}$	7.43 (6.77)	0.09 (0.08)	7.49 (6.81)	10.22 (7.76)
	$\widehat{\theta}^{GP}$	65.8 (65.4)	0.14 (0.11)	65.8 (65.4)	43.9 (42.8)
0.15	$\widehat{\theta}^{T,CI}$	4.74 (3.60)	0.11 (0.10)	4.68 (3.62)	4.95 (4.82)
	$\widehat{\theta}^{T}$	4.54 (3.52)	0.11 (0.10)	4.55 (3.54)	6.99 (4.78)
	$\widehat{\theta}^{NLS}$	7.75 (7.60)	0.14 (0.13)	7.81 (7.64)	10.34 (9.32)
	$\widehat{\theta}^{GP}$	73.7 (73.7)	0.25 (0.15)	73.8 (73.7)	61.3 (58.7)

better results than NLS and GP. The difference is even more striking here, possibly due to the accumulated effect of the different source of misspecifications on GP and NLS. For our approaches, the cost (12) takes into account a model discrepancy term expected to mitigate the effect of misspecification on estimation.

6 Real data case analysis

We focus on a model discussed by [Stein et al., 2013] to study the impact on a microbiota ecosystem of the interaction between an antibiotic treatment and a pathogen inoculation:

$$\dot{x}_i = \mu_i x_i + x_i \sum_{j=1}^{11} M_{i,j} x_j + x_i s_i v(t) \tag{31}$$

for $i = 1, \ldots, 11$. Each state variable x_i quantifies the presence of one microbial species and $t \mapsto v(t)$ describes the perturbation due to clindamycin administration. Regarding the parameter set $(\mu_i,\, M_{i,j},\, s_i)_{1 \leq i,j \leq 11}$, μ_i is the growth term for x_i, $M_{i,j}$ the interaction effect of x_j on x_i and s_i the susceptibility of x_i to the antibiotic treatment. The names of the microbial species as well as the values of $(\mu_i,\, M_{i,j},\, s_i)_{1 \leq i,j \leq 11}$ are provided by [Stein et al., 2013] (Figure 2). The acquired data are divided in three groups of three subjects. Group 1 was exposed to the pathogen (here, *C. difficile*), Group 2 received a single dose of clindamycin and Group 3 received clindamycin and was exposed to *C. difficile* the day after. We focus on Group 3

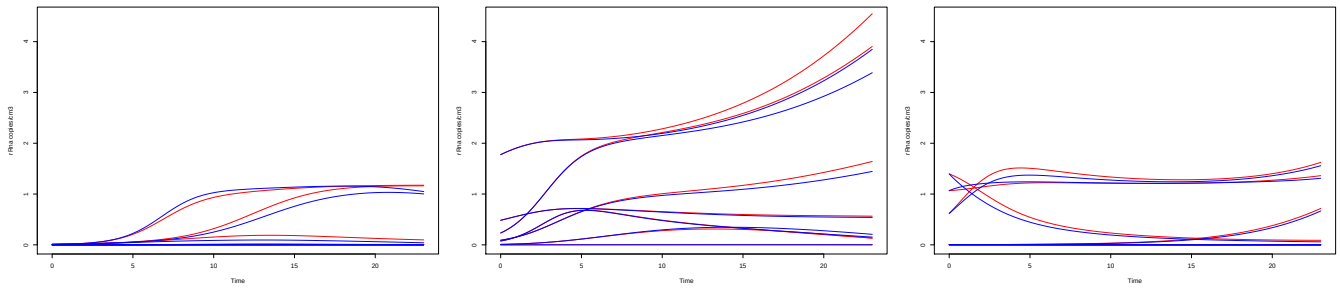

Figure 5: Solutions of (31) (blue) and (32) (red) for different initial conditions.

for which the perturbation is $v(t) = 1_{\{t=0\}}$. In this group, some microbiotal species have limited impacts on the whole ecosystem evolution, they correspond to x_6, x_7, x_8 and x_{10}. That is why for parameter estimation we use the following simplified model only composed of the 7 remaining state variables:

$$\dot{x_i^s} = \mu_i^s x_i^s + x_i^s \sum_{j=1}^{7} M_{i,j}^s x_j^s + x_i^s s_i^s v(t) \tag{32}$$

where $x_i^s = x_i$ for $1 \leq i \leq 5$, $x_6^s = x_9$ and $x_7^s = x_{11}$. The parameters $\left\{\mu_i^s,\, M_{i,j}^s,\, s_i^s\right\}$ are defined and linked to the previous parametrization accordingly. For comparison, we plot in Figure 5 the prediction made for state variables x_1, x_2, x_3, x_4, x_5, x_9 and x_{11} respectively by the ODE (31) and (32) for three initial condition values corresponding to subjects in Group 3.

Here, we estimate $M_{i,j}$ for they give the nature of interaction between species. The different (μ_i, s_i) are known and we estimate the $p = 31$ parameters: $\theta = (\{M_{i,1}\}_{i\in\{1,3,4,5\}}, \{M_{i,2}\}_{i\in\{2,3,4,5\}}, \{M_{i,3}\}_{i\in\{1,2,3\}}, \{M_{i,4}\}_{i\in\{1,2,4\}}, \{M_{i,5}\}_{i\in\{2,3,4,5\}}, \{M_{i,9}\}_{i\in\{3,5\}}, \{M_{i,11}\}_{i\in\{1,2,3,4,5\}}, \{M_{11,i}\}_{i\in\{1,2,3,5,9,11\}})$. The other interaction terms are unidentifiable in practice when we use only data coming from Group 3.

Before starting our real data analysis we have to specify our assumptions about the measurement noise structure. As explained in [Stein et al., 2013], for the same concentration profil each time point corresponds to a different mouse but which are all biological replicates, uniformly treated but separately housed. Hence, we assume independance of measurement noise between observations over time. Still, we have to take into account the difference in magnitude over time which exists for a same concentration profile as exemplified

in figure 5. For this, we rely on a log-transform of the observation to have the measurement noise increasing with state amplitude. For estimation, our method then uses the ODE:

$$\dot{\tilde{x}}_i^s = \mu_i^s + \sum_{j=1}^{7} M_{i,j}^s e^{\tilde{x_j^s}} + s_i^s v(t) \tag{33}$$

followed by $\tilde{x_i^s} = \log(x_i^s)$, the log-transform of the original state-variables. Still, there is the possibility that the noise have different order of magnitude specific to each $\tilde{x_i^s}$. To legimitate the use of our method in this context, we perform in the next subsection simulation analysis to compare it with NLS and GP in presence of such noise structure.

6.1 Preliminary results on simulated data

Estimation is based on the observation of three subjects with initial conditions $x_{1,0}^*$, $x_{2,0}^*$ and $x_{3,0}^*$. Here, $\widehat{\theta}^T$ have 52 components so to save computational time, we restrict our approach to $\widehat{\theta}^{T,CI}$. We split θ into two subgroups θ_1 and θ_2 according to the difficulty encountered by NLS to estimate them. To identify θ_1, the set of parameters poorly estimated by NLS, we rely once again on the eigendecomposition of $V^*D^*(V^*)^{-1} = \mathcal{I}_n\left(\theta^*, x_{1,0}^*, x_{2,0}^*, x_{3,0}^*\right)$ for $n = 25$ where $D^* = diag(\mu_1, \ldots, \mu_{52})$ is the matrix composed of the eigenvalues μ_i sorted in increasing order and each column $V_{\cdot,i}^*$ of V^* is the eigenvector related to μ_i. The associated condition number is $\kappa(\mathcal{I}_n) \simeq 8 \times 10^{-10}$ which indicates an ill-posed problem for NLS. Moreover, we have $\frac{\mu_{25}}{\mu_{52}} = 2 \times 10^{-6}$, thus the first 25 eigenvectors correspond to directions of weak change for the NLS criteria. For each parameter θ^j in θ, we compute $F(\theta^j) = \frac{\sum_{i=1}^{25}\left(V_{j,i}^*\right)^2}{\sum_{i=1}^{52}\left(V_{j,i}^*\right)^2}$ to quantify the impact of θ^j on NLS criteria. By doing so we identify 12 parameters such that $F(\theta^j) > 0.63$ which will compose θ_1. The choice of threshold for the eigenvalue rank and $F(\theta^j)$ value is somewhat arbitrary, but we will see in simulations the error for $\widehat{\theta}^{NLS}$ comes mainly from estimation of θ_1. The results presented for the variance (resp. mean square error) for θ_1 and θ_2 will denote the sum of the variance (resp. mean square error) of θ_1 and θ_2 components. We also compare the ability of the estimators to find the orientation of the interaction graph, we estimate $\mathrm{I}(\widehat{\theta}) = \frac{1}{p}\mathbb{E}_{\theta^*}\left[\sum_{i=1}^{p} 1_{\left\{sign(\widehat{\theta}_i)=sign(\theta_i^*)\right\}}\right]$, the expected fraction of correctly retrieved interaction. We select $k_n = 20$ and λ among $\{1, 2, 5\}$.

Table 6: MSE and Variance (in parenthesis) for Microbiota model (32) and $n = 25$.

σ		θ_1	θ_2	θ		$I(\widehat{\theta}_1)$	$I(\widehat{\theta}_2)$	$I(\widehat{\theta})$
0.01	$\widehat{\theta}^{T,CI}$	0.14 (0.10)	0.04 (0.03)	0.10 (0.06)		1	1	1
	$\widehat{\theta}^{NLS}$	3.29 (2.23)	1.28 (1.23)	2.05 (1.94)		0.97	1	0.99
	$\widehat{\theta}^{GP}$	0.15 (0.11)	0.20 (0.19)	0.18 (0.14)		1	1	1
0.02	$\widehat{\theta}^{T,CI}$	0.71 (0.62)	0.15 (0.15)	0.46 (0.36)		0.99	1	1
	$\widehat{\theta}^{NLS}$	7.54 (7.39)	2.19 (2.15)	4.97 (4.75)		0.94	0.99	0.97
	$\widehat{\theta}^{GP}$	1.40 (0.97)	0.80 (0.78)	1.22 (1.07)		1	0.98	1
0.03	$\widehat{\theta}^{T,CI}$	5.66 (1.98)	1.81 (1.41)	4.54 (1.43)		0.93	0.99	0.97
	$\widehat{\theta}^{NLS}$	9.15 (8.66)	2.61 (2.46)	8.22 (4.40)		0.92	0.99	0.96
	$\widehat{\theta}^{GP}$	7.67 (3.47)	2.50 (1.80)	6.21 (3.31)		0.93	0.98	0.95

Influence of measurement noise We consider one sample size $n = 25$. To make the level of noise state variable specific, each $\tilde{x_i^s}$ is corrupted by a measurement noise of standard deviation $\frac{\sigma}{n} \times \sum_{j=0}^{n} \left\| \tilde{x_i^s}(t_j) \right\|$, three levels for σ are tested ($\sigma = 0.03$, $\sigma = 0.06$ and $\sigma = 0.09$). Results are presented in Table 6. For θ_1 and θ_2, both regularization methods outperform NLS. Still, we obtain more accurate estimation than GP.

Influence of model misspecification We choose n and σ as before but the observations are now generated by using the log-transform of the model (31). We are interested in quantifying robustness of the different estimators with respect to misspecification due to neglected interactions, a common feature in the study of biological networks. In particular, we want to measure the ability of our estimator to retrieve the true interactions between two state variables despite the presence of unmeasured coufounders. Results are presented in Table 7. The situation is quite similar to the well-specified case but with the additional feature that the capacity of $\widehat{\theta}^{NLS}$ to retrieve the true interaction graph is more affected by model misspecification than $\widehat{\theta}^{T,CI}$.

Table 7: MSE and Variance (in parenthesis) for misspecified Microbiota model for $n = 25$.

σ		θ_1	θ_2	θ		$I(\widehat{\theta_1})$	$I(\widehat{\theta_2})$	$I(\widehat{\theta})$
0.01	$\widehat{\theta}^{T,CI}$	5.84 (1.76)	1.15 (0.63)	5.67 (1.06)		0.88	0.99	0.96
	$\widehat{\theta}^{NLS}$	11.95 (1.69)	3.19 (0.59)	14.01 (1.18)		0.87	0.98	0.94
	$\widehat{\theta}^{GP}$	7.05 (2.69)	2.60 (0.32)	10.01 (1.02)		0.89	0.98	0.97
0.02	$\widehat{\theta}^{T,CI}$	7.79 (4.15)	1.55 (1.59)	7.77 (3.44)		0.87	0.98	0.96
	$\widehat{\theta}^{NLS}$	13.23 (3.24)	3.88 (1.11)	14.99 (2.17)		0.81	0.94	0.89
	$\widehat{\theta}^{GP}$	12.21 (3.04)	3.02 (1.31)	11.23 (2.37)		0.83	0.94	0.91
0.03	$\widehat{\theta}^{T,CI}$	9.73 (5.03)	2.21 (0.97)	8.69 (3.41)		0.87	0.98	0.94
	$\widehat{\theta}^{NLS}$	20.00 (9.50)	5.75 (3.04)	20.35 (6.76)		0.81	0.94	0.89
	$\widehat{\theta}^{GP}$	14.10 (8.60)	4.05 (2.78)	13.34 (5.53)		0.83	0.94	0.92

6.2 Real data analysis

In this section, we estimate θ in model (32) starting from the data available in [Stein et al., 2013] for the Group 3 presented in Figure 6. The original observation interval was [0, 23] but here we restrict to [0, 16] since no data are available on]16, 23[and a first estimation attempt with [0, 23] leads to poor data fitting of the optimal trajectories $\overline{X^d_{\widehat{\theta}}}$. We selected $k_n = 40$ for it corresponds to a mesh size small enough to accurately estimate the ODE perturbed model without being too computationally intensive. We choose $\lambda = 1$, for it is small enough to account for model error presence and leads to accurate estimation comparing to NLS when some interactions are neglected according to the previous section results.

Despite the use of a simplified model and the presence of outliers which render difficult a good data fitting of $\overline{X^d_{\widehat{\theta}^{T,CI}}}$ (see Figure 6), we obtained a graph orientation close to the one obtained in [Stein et al., 2013] with only 7 out of the 31 estimated interaction parameters having a different sign (see Table 8). This confirms the benefit of using the approximated methods for real data analysis, where model uncertainty presence is the rule rather than the exception.

Our methodology copes with model misspecification by limiting its effect on estimation. However, our approach may also be useful for checking misspecification presence by analyzing the optimal control

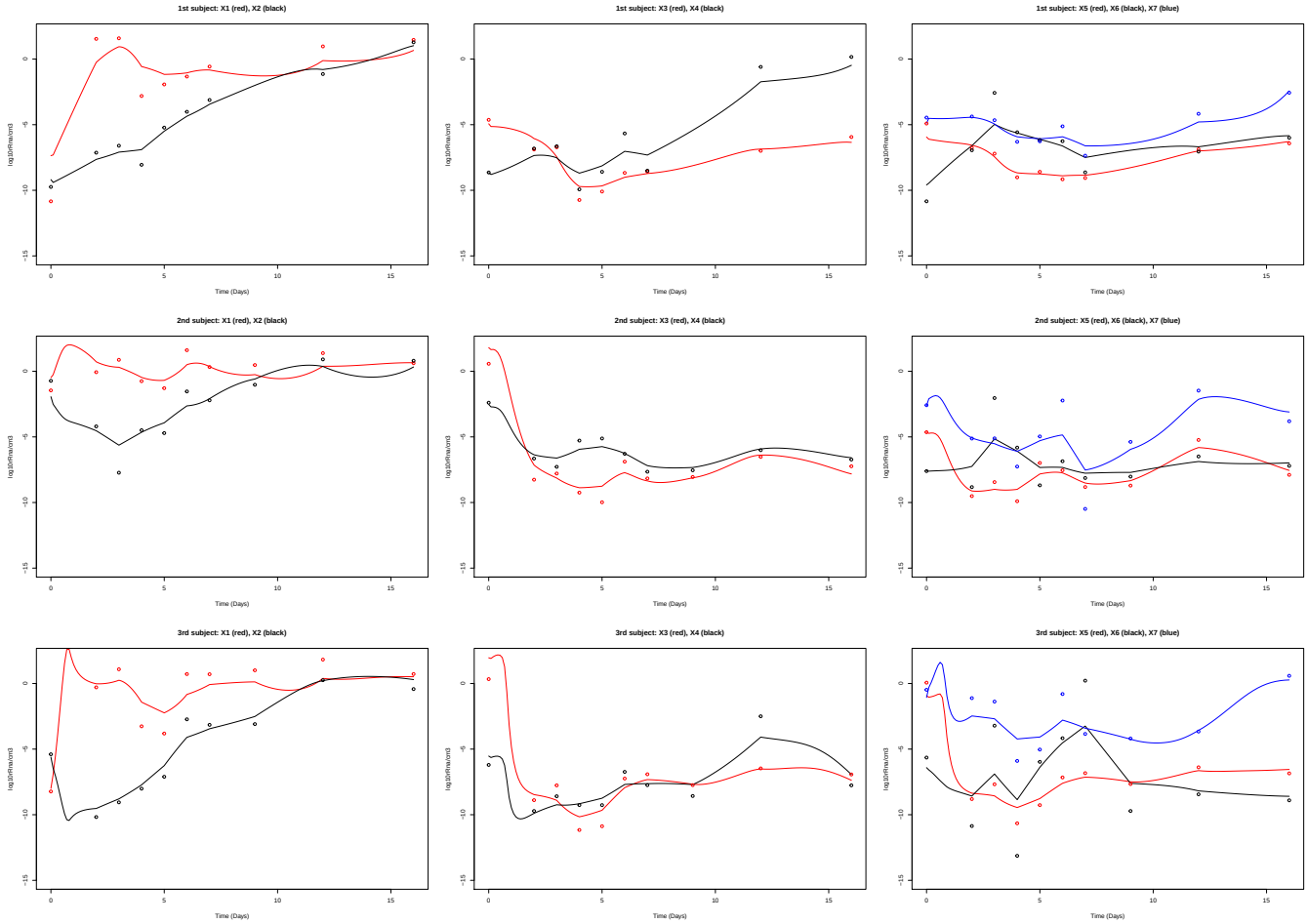

Figure 6: Log transformed observations and reconstructed $\overline{X^d_{\widehat{\theta}^{T,CI}}}$ for subject 1, 2, 3 in group 3 .

Table 8: Scaled values for [Stein et al., 2013] estimator $\widehat{\theta}^{Stein}$ and $\widehat{\theta}^{T,CI}$

$\widehat{\theta}^{Stein}$	-0.2	0.14	0.22	-0.18	-0.10	-0.19	0.14	-0.05	0.17	-0.04	-0.10	-0.16	-0.15	-0.83	-0.18	-0.16
$\widehat{\theta}^{T,CI}$	-3.29	1.77	0.29	1.79	-0.43	0.41	0.54	0.45	0.02	-0.15	-0.51	-1.16	-0.04	0.47	-1.52	0.92
$\widehat{\theta}^{Stein}$	-0.22	-0.71	0.30	0.16	-0.27	-0.20	-0.21	-0.40	0.11	-0.37	0.28	0.25	0.08	0.32	-0.38	
$\widehat{\theta}^{T,CI}$	-0.17	-0.57	1.56	0.85	-1.72	1.21	-0.78	-1.34	0.80	-0.34	1.02	-0.60	0.45	0.13	-0.64	

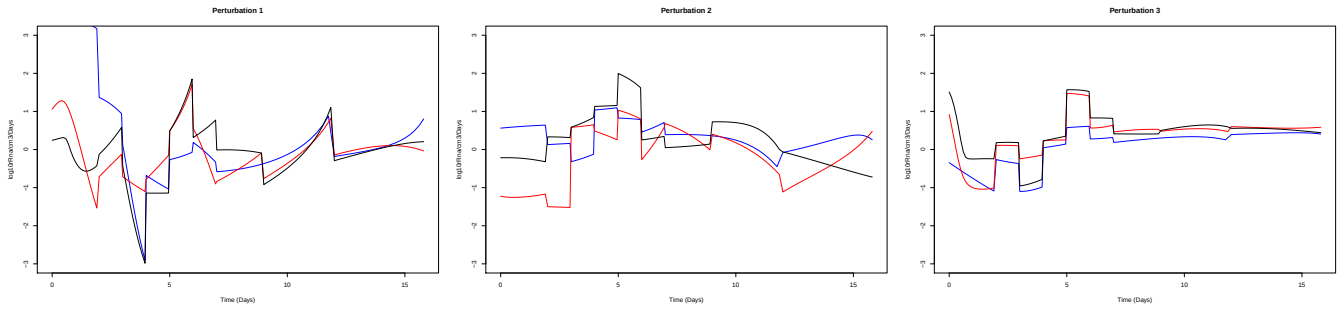

Figure 7: 1st, 2nd and 3rd component of $\overline{u^d_{\widehat{\theta}^{T,CI}}}$ for subject 1 (blue), 2 (red), 3 (black) in group 3.

$\overline{u^d_{\widehat{\theta}^{T,CI}}}$ values, which can be seen as residuals quantifying the discrepancy between the model and the actual system dynamic. We present in Figure 7 for each subject the first three components of $\overline{u^d_{\widehat{\theta}^{T,CI}}}$. One can see there are shared patterns, for example in the first graph around $t = 10$ days, where the optimal controls present the same behavior for all subjects. Such features indicate that some deterministic elements of the actual system dynamic have been missed by the assumed model.

7 Conclusion

This work develops a method based on discrete optimal control theory to regularize the problem of parameter estimation in ODEs comparing to the use of classic methods such as NLS. This regularization method is more computationally efficient and suffers from a lesser performance drop than the one presented in [Clairon and Brunel, 2019, Clairon and Brunel, 2018] in sparse sample cases. We also expose how we can easily profile on the initial conditions to avoid the estimation of additional nuisance parameters. The experimental and real data analyses confirm the good performance of our method in comparison with NLS and GP for small sample case, where the asymptotic analysis results do not hold.

An under-exploited feature of the method so far is the obtained optimal control. Here, we only use it for a qualitative based analysis in the real data case, but we suspect that a full analysis of $\overline{u^d_{\widehat{\theta}^{T,CI}}}$ maybe be useful to construct a statistical test of misspecification at the derivative level, which is more relevant

for such models than the test based on residuals [Hooker and Ellner, 2015]. This is a subject for further work. A second point worth exploring in the future is the extension to mixed effect model in which several subjects are observed and despite that they present different trajectories it is assumed their dynamics are ruled by the same evolution law. It means each subject i follows the equation $\dot{X} = f(t, X, \theta_i)$ where f is common to the whole population but θ_i is an individual parameter defined as the realization of a random variable following a law p depending on a population parameter θ i.e. $\theta_i \sim p(\theta)$. For these models, dedicated methods are necessary to incorporate inter-patient correlation in the estimation process [Raftery and Bao, 2010, Donnet and Samson, 2006, M. Lavielle and Mentre., 2011, M.Prague et al., 2013, Wang et al., 2014]. For our method, it would be interesting to consider mixed-effect on the estimated optimal controls $\overline{u_i^d}$ to take into account correlations on the committed model error among the individuals.

Acknowledgements

Quentin Clairon was supported by EPSRC award EP/M015637/1.

Software

The R code used to implement our optimal control based estimation methods is available on request at quentin.clairon@u-bordeaux.fr .

References

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

[Brunel and Clairon, 2015] Brunel, N. J.-B. and Clairon, Q. (2015). A tracking approach to parameter estimation in linear ordinary differential equations. *Electronic Journal of Statistics*, 9:2903–2949.

[Brunel and D'Alche-Buc, 2014] Brunel, N. J.-B. and D'Alche-Buc, Q. C. F. (2014). Parameter estimation of ordinary differential equations with orthogonality conditions. *Journal of the American Statistical Association*, 109:173–185.

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

[Chkrebtii et al., 2016] Chkrebtii, O. A., Campbell, D. A., Calderhead, B., and Girolami, M. (2016). Bayesian solution uncertainty quantification for differential equations. *Bayesian Analysis*, 11:1239–1267.

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

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

[Cimen and Banks, 2004b] Cimen, T. and Banks, S. (2004b). Nonlinear optimal tracking control with application to supertankers for autopilot design. *Automatica*, 40:1845–1863.

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

[Clairon and Brunel, 2019] Clairon, Q. and Brunel, N. J.-B. (2019). Tracking for parameter and state estimation in possibly misspecified partially observed linear ordinary differential equations. *Journal of Statistical Planning and Inference*, 199:188–206.

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

[Dattner, 2015] Dattner, I. (2015). Optimal rate of direct estimators in systems of ordinary differential equations linear in functions of the parameters. *Electronic Journal of Statistics*, 9:1939–1976.

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

[Elowitz and Leibler, 2000] Elowitz, M. and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338.

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

[Engl et al., 1996] Engl, H. W., Hanke, M., and Neubauer, A. (1996). *Regularization of inverse problems*, volume 375. Springer Science & Business Media.

[Esposito and Floudas, 2000] Esposito, W. and Floudas, C. (2000). Deterministic global optimization in nonlinear optimal control problems. *Journal of Global Optimization*, 17:97–126.

[Fall et al., 2002] Fall, C., Marland, E., Wagner, J., and Tyson, J., editors (2002). *Computational Cell Biology*. Interdisciplinary applied mathematics. Springer.

[FitzHugh, 1961] FitzHugh, R. (1961). Impulses and physiological states in theoretical models of nerve membrane. *Biophysical Journal*, 6:445–466.

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

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

[Goldbeter, 1997] Goldbeter, A. (1997). *Biochemical Oscillations and Cellular Rhythms: The Molecular Bases of Periodic and Chaotic Behaviour*. Cambridge University Press.

[Gugushvili and Klaassen, 2011] Gugushvili, S. and Klaassen, C. (2011). Root-n-consistent parameter estimation for systems of ordinary differential equations: bypassing numerical integration via smoothing. *Bernoulli*, to appear.

[Guo and Sun, 2012] Guo, B. and Sun, B. (2012). 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.

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

[Hairer et al., 1993] Hairer, E., Norsett, S., and Wanner, G. (1993). *Solving Ordinary Differential Equations I.* Springer Series in Computational Mathematics.

[Hairer and Wanner, 1996] Hairer, E. and Wanner, G. (1996). *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems.* Springer Series in Computational Mathematics.

[Hooker and Ellner, 2015] Hooker, G. and Ellner, S. P. (2015). Goodness of fit in nonlinear dynamics: Mis-specified rates or mis-specified states? Technical report, The Annals of Applied Statistics.

[Jaeger and Lambert, 2011] Jaeger, J. and Lambert, P. (2011). Bayesian generalized profiling estimation in hierarchical linear dynamic systems. *Discussion Paper*.

[Kampen, 1992] Kampen, N. V. (1992). *Stochastic Process in Physics and Chemistry*. Elsevier.

[Kennedy and OHagan, 2001] Kennedy, M. C. and OHagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society*, 63:425–464.

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

[Koehler et al., 2009] Koehler, E., Brown, E., and Haneuse, S. (2009). On the assessment of monte carlo error in simulation-based statistical analyses. *Journal of the American Statistical Association*, 63:155–162.

[Kurtz, 1970] Kurtz, T. (1970). Solutions of ordinary differential equations as limits of pure jump markov processes. *Journal of Applied Probability*, 7:49–58.

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

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

[Li et al., 2005] Li, Z., Osborne, M., and Prvan, T. (2005). Parameter estimation of ordinary differential equations. *IMA Journal of Numerical Analysis*, 25:264–285.

[Liang et al., 2010] Liang, H., Miao, H., and Wu, H. (2010). Estimation of constant and time-varying dynamic parameters of hiv infection in a nonlinear differential equation model. *Annals of Applied Statistics*, 4:460–483.

[Lindner and Schimansky-Geier, 1999] Lindner, B. and Schimansky-Geier, L. (1999). Analytical approach to the stochastic fitzhugh-nagumo system and coherence resonance. *Physical Review*, 60:7270–7276.

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

[Mirsky et al., 2009] Mirsky, H., Liu, A., Welsh, D., Kay, S., and III, F. D. (2009). A model of the cell-autonomous mammalian circadian clock. *Proceedings of the National Academy of Sciences*, 106(27):11107–11112.

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

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

[Qi and Zhao, 2010] Qi, X. and Zhao, H. (2010). Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations. *The Annals of Statistics*, 1:435–481.

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

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

[Rodriguez-Fernandez et al., 2006] Rodriguez-Fernandez, M., Egea, J., and Banga, J. R. (2006). Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. *BioMed Central*.

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

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

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

[Tonsing et al., 2014] Tonsing, C., Timmer, J., and Kreutz, C. (2014). Cause and cure of sloppiness in ordinary differential equation models. *Physical Review*, 90:023303.

[Transtrum et al., 2015] Transtrum, M., Machta, B., Brown, K., Daniels, B., Myers, C., and Sethna, J. (2015). Sloppiness and emergent theories in physics, biology, and beyond. *arXiv:1501.07668v1*, page 15.

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

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

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

[Wang et al., 2014] Wang, L., Cao, J., Ramsay, J., Burger, D., Laporte, C., and Rockstroh, J. (2014). Estimating mixed-effects differential equation models. *Statistics and Computing*, 24:111–121.

[White et al., 2016] White, A., Tolman, M., Thames, H. D., Withers, H. R., Mason, K. A., and M.K.Transtrum (2016). The limitations of model-based experimental design and parameter estimation in sloppy systems. *Public Library of Science Computational Biology*, 12:1–26.

[Wu et al., 2010] Wu, H., Kumar, A., Miao, H., Holden-Wiltse, J., Mosmann, T., Livingstone, A., Belz, G., Perelson, A., Zand, M., and Topham, D. (2010). Modeling of influenza-specific cd8+ t cells during the primary response indicates that the spleen is a major source of effectors. *The Journal of Immunology*, 187(9):4474–4482.

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

[Xue et al., 2010] Xue, H., Miao, H., and Wu, H. (2010). Sieve estimation of constant and time-varying coefficients in nonlinear ordinary differential equation models by considering both numerical error and measurement error. *Annals of Statistics*, 38:2351–2387.

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