Parametric Estimation of Ordinary Differential Equations with Orthogonality Conditions

Nicolas J.-B. Brunel, Quentin Clairon, Florence d'Alché-Buc

Parametric Estimation of Ordinary Differential Equations with Orthogonality Conditions

Nicolas J-B. Brunel[1], Quentin Clairon[2], Florence d'Alché-Buc[3,4]

May, 04 2013

Abstract

Differential equations are commonly used to model dynamical deterministic systems in applications. When statistical parameter estimation is required to calibrate theoretical models to data, classical statistical estimators are often confronted to complex and potentially ill-posed optimization problem. As a consequence, alternative estimators to classical parametric estimators are needed for obtaining reliable estimates. We propose a gradient matching approach for the estimation of parametric Ordinary Differential Equations observed with noise. Starting from a nonparametric proxy of a true solution of the ODE, we build a parametric estimator based on a variational characterization of the solution. As a Generalized Moment Estimator, our estimator must satisfy a set of orthogonal conditions that are solved in the least squares sense. Despite the use of a nonparametric estimator, we prove the root-n consistency and asymptotic normality of the Orthogonal Conditions estimator. We can derive confidence sets thanks to a closed-form expression for the asymptotic variance. Finally, the OC estimator is compared to classical estimators in several (simulated and real) experiments and ODE models in order to show its versatility and relevance with respect to classical Gradient Matching and Nonlinear Least Squares estimators. In particular, we show on a real dataset of influenza infection that the approach gives reliable estimates. Moreover, we show that our approach can deal directly with more elaborated models such as Delay Differential Equation (DDE).

Key-words: Gradient Matching, Nonparametric statistics, Methods of Moments, Plug-in Property,

Variational formulation, Sobolev Space.

1 Introduction

1.1 Problem position and motivations

Differential Equations are a standard mathematical framework for modeling dynamics in physics, chemistry, biology, engineering sciences, etc and have proved their efficiency in describing the real world. Such models are defined thanks to a time-dependent vector field $\boldsymbol{f}$, defined on the state-space $\mathcal{X} \subset \mathbb{R}^d$ and that depends on a parameter $\theta \in \Theta \subset \mathbb{R}^p$, $d, p \geq 1$. The vector field is then a function from $[0,1] \times \mathcal{X} \times \Theta$ to $\mathbb{R}^d$. If $\phi(t)$ is the current state of the system, the time evolution is given by the following Ordinary Differential Equation, defined for $t \in [0,1]$ by:

$$\dot{\phi}(t) = \boldsymbol{f}(t, \phi(t), \theta) \tag{1.1}$$

where dot indicates derivative with respect to time. An important task is then the estimation of the parameter θ from real data. [30] proposed a significant improvement to this statistical problem, and gave motivations for further statistical studies. We are interested in the definition and in the optimality of a statistical procedure for the estimation of the parameter θ from noisy observations $y_1, \ldots, y_n \in \mathbb{R}^d$ of a solution at times $t_1 < \cdots < t_n$.

Most works deal with Initial Value Problems (IVP), i.e. with ODE models having a given (possibly unknown) initial value $\phi(0) = \phi_0$. There exists then a unique solution $\phi(\cdot, \phi_0, \theta)$ to the ODE (1.1) defined on the interval $[0,1]$, that depends smoothly on ϕ_0 and θ.

The estimation of θ is a classical problem of nonlinear regression, where we regress y on the time t.

[1] ENSIIE & Laboratoire Statistique et Génome, Université d'Evry Val d'Essonne, UMR CNRS 8071 - USC INRA - FRANCE
[2] Laboratoire Analyse et Probabilités, Université d'Evry Val d'Essonne - FRANCE
[3] Laboratoire IBISC, Université d'Evry Val d'Essonne - FRANCE
[4] INRIA-Saclay, LRI, Université Paris Sud, UMR CNRS 8623 - FRANCE

If ϕ_0 is known, the Nonlinear Least Square Estimator $\hat{\theta}^{NLS}$ (NLSE) is obtained by minimizing

$$Q_n^{LS}(\theta) = \sum_{i=1}^{n} |y_i - \phi(t_i, \phi_0, \theta)|^2 \tag{1.2}$$

where $|\cdot|$ is the classical Euclidean norm. The NLSE, Maximum Likelihood Estimator or more general M-estimators [36] are commonly used because of their good statistical properties (root-n consistency, asymptotic efficiency), but they come with important computational difficulties (repeated ODE integrations and presence of multiple local minima) that can decrease their interest. We refer to [30] for a detailed overview of the previous works in this field. An adapted NLS estimator (dedicated the specific difficulties of ODEs) is also introduced and studied in [43].

Global optimization methods are then often used, such as simulated annealing, evolutionary algorithms ([22] for a comparison of such methods). Other classical estimators are obtained by interpreting noisy ODEs as state-space models: filtering and smoothing techniques can be used for parameter inference [9], which can provide estimates with reduced computational complexity [29, 17, 16].

Nevertheless, the difficulty of the optimization problem is the outward sign of the ill-posedness of the inverse problem of ODE parameter estimation, [12]. Hence some improvements on classical estimation have been proposed by adding regularization constraints in an appropriate way.

Starting from different methods used for solving ODEs, different estimators can be developed based on a mixture of nonparametric estimation and collocation approximation. This gives rise to Gradient Matching (or Two-Step) estimators that consists in approximating the solution ϕ with a basis expansion $\{B_1, \ldots, B_K\}$, such as cubic splines. The rationale is to estimate nonparametrically the solution ϕ by $\hat{\phi} = \sum_{k=1}^{L} \hat{c}_k B_k$ so that we can also estimate the derivative $\dot{\hat{\phi}}$. An estimator of θ can be obtained by looking for the parameter that makes $\hat{\phi}$ satisfy the differential equation (1.1) in the best possible manner. Two different methods have been proposed, based on a L^2 distance between $\dot{\hat{\phi}}$ and $\boldsymbol{f}(t, \hat{\phi}, \theta)$: The first one, called the *two-step method*, was originally proposed by [38], and has been particularly developed in (bio)chemical engineering [20, 39, 28]. It avoids the numerical integration of the ODE and usually gives rise to simple optimization program and fast procedures that usually performs well in practice. The statistical properties of this two stage estimator (and several variants) have been studied in order to understand the influence of nonparametric techniques to estimate a finite dimensional parameter

[8, 19, 14]. While keeping the same kind of numerical approximation of the solution, [30] proposed a second method based on the generalized smoothing approach for determining at the same time the parameter θ and the nonparametric estimation $\hat{\phi}$. The essential difference between these two approaches is that the nonparametric estimator in the generalized smoothing approach is computed adaptively with respect to the parametric model, whereas two-step estimators are "model-free smoothing".

We introduce here a new estimator that can be seen as an improvement and a generalization of the previous two-step estimators. It uses also a nonparametric proxy $\hat{\phi}$, but we modify the criterion used to identify the ODE parameter (i.e. the second step). The initial motivations are

- to get a closed-form expression for the asymptotic variance and confidence sets,
- to reduce sensitivity to the estimation of the derivative in Gradient Matching approaches,
- to take into account explicitly time-dependent vector field, with potential discontinuities in time.

The most notable feature of the proposed method is the use of a variational formulation of the differential equations instead of the classical point-wise one, in order to generate conditions to satisfy. This formulation is rather general and can cover a greater number of situations: we come up with a generic class of estimator of Differential Equations (e.g Ordinary, Delay, Partial, Differential-Algebraic), that can incorporate relatively easily prior knowledge about the true solution. In addition to the versatility of the method, the criterion is built in order to offer computational tractability, that implies that we can give a precise description of the asymptotics and give the bias and variance of the estimator. We also give a way to ameliorate adaptively our estimator and to compute asymptotic confidence intervals.

First, we introduce the statistical ODE-based model and main assumptions, we motivate and describe our estimator, and show its consistency. Then, we provide a detailed description of the asymptotics, by proving its root-n consistency and asymptotic normality. Based on the asymptotic approximation, we give a closed-form expression of the asymptotic variance, and we address the problem of obtaining the best variance through the choice of an appropriate weighting matrix. Finally, we provide some insights into the practical behavior of the estimator through simulations and by considering two real-data examples. The objective of the experiments parts is to show the interest of OC with respect to the nonlinear least squares and classical gradient matching estimators.

1.2 Examples

We motivate our work in detail by presenting two models that are relatively common and simple but that nevertheless causes difficulties for estimation.

1.2.1 Ricatti ODE

The (scalar) Ricatti equation is defined by a quadratic vector field $f(t,x) = a(t)x^2 + b(t)x + c(t)$ where $a(\cdot), b(\cdot), c(\cdot)$ are time-varying functions. This equation arises naturally in control theory for solving linear-quadratic control problem [35], or in mathematical finance, in the analysis of stochastic interest rate models [7]. We consider one of the simplest case where a is constant, $b = 0$ and $c(t) = c\sqrt{t}$. The objective is to estimate parameters a, c from the noisy observations $y_i = \phi(t_i) + \epsilon_i$ for $t_i \in [0, 14]$. Here the true parameters are $a = 0.11$, $c = 0.09$ and $\phi_0 = -1$, and one can see the solution and simulated observations in figure 1.1. Although the solution ϕ is smooth in the parameters, there exists no closed form and simulations are required for implementing NLS and classical approaches. The hard part in this equation is due to the extreme sensitivity of the squared term in the vector field: for small differences in the parameters or initial condition, the solution can explode before reaching the final time $T = 14$ 1.1. Explosions are not due to numerical instability but to the failure of (theoretical) existence of a global solution on the entire interval (e.g the *tangent* function is solution of $\dot{\phi} = \phi^2 + 1$, $\phi(0) = 0$ and explodes at $t = \frac{\pi}{2}$). The explosions have to be handled in estimation algorithms and this slows down the exploration of the parameter space (which can be difficult for high-dimensional state or parameter spaces). Nevertheless, we show in the experiment part that NLS or Gradient Matching can do well for parameter estimation, but some additional difficulties does appear when the time-dependent function $c(\cdot)$ has abrupt changes. We consider the case where $c(t) = c\sqrt{t} - d'\mathbb{1}_{[T_r,T]}$, T_r is a change-point time, with $d' > 0$. This situation is classical (e.g in engineering) where some input variables $t \mapsto u(t)$ modify the evolution of the system $\dot{\phi} = f(t, \phi(t)) + u(t)$ (typically it can be the introduction of a new chemical species in a reactor at time T_r), see figure 1.1. The Cauchy-Lipschitz theory for existence and uniqueness of solutions to time-discontinuous ODE is extended straightforwardly with measure theoretic arguments [35]. The (generalized) solution is defined almost everywhere and belongs to a Sobolev space. For sake of completeness, we provide a generalized version of the Cauchy-Lipschitz theorem for IVP in

Supplementary Material I. This abrupt change causes some difficulties in estimating non-parametrically the solution and its derivative, which can make Gradient Matching less precise. We consider then the estimation of the two additional parameters d' and T_r. Hence, the parameter estimation problem can be seen as a change-point detection problem, where the solution ϕ still depends smoothly in the parameters. Nevertheless, in the case of the joint estimation of a, c, d' and T_r, the particular influence of the parameter T_r makes the problem much more difficult to deal with for classical approaches as it is suggested by the objective functions in *Supplementary Material II.* The variational formulation for model estimation gives a seamless approach for estimating models which possess time discontinuities.

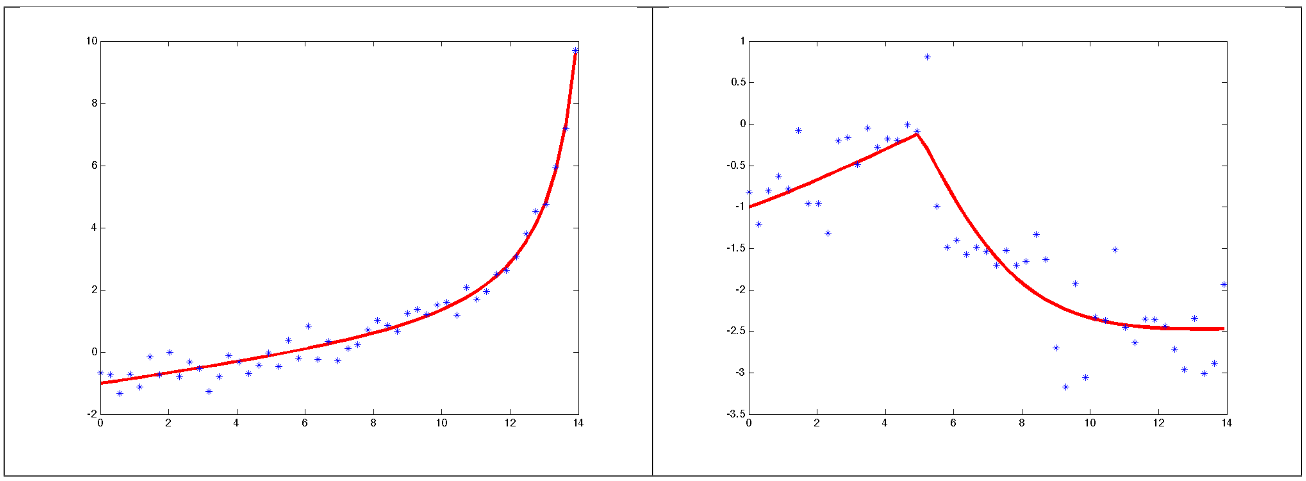

Figure 1.1: Solutions of Ricatti ODE with noisy observations ($n = 50$, $\sigma = 0.4$). Left figure is smooth time-dependent ODE. Right figure has a change-point at time $T_r = 5$ ($d' = 1$).

1.2.2 Dynamics of Blowfly populations

The modeling of the dynamics of population is a classical topic in ecology an more generally in biology. Differential Equations can describe very precisely the mechanics of evolution, with birth, death and migration effects. The case of single-species models is the easiest case to consider, as interactions with rest of the world can be limited, and the acquisition of reliable data is easier. In the 50s, Nicholson measured quite precisely the dynamics of a blowfly population, known as Nicholson's experiments [26]. The data are relatively hard to model, and it is common to use Delay Differential Equation (DDE) whose general form is $\dot{N}(t) = f\left(N(t), N(t-\tau), \theta\right)$, in order to account for the almost chaotic behavior

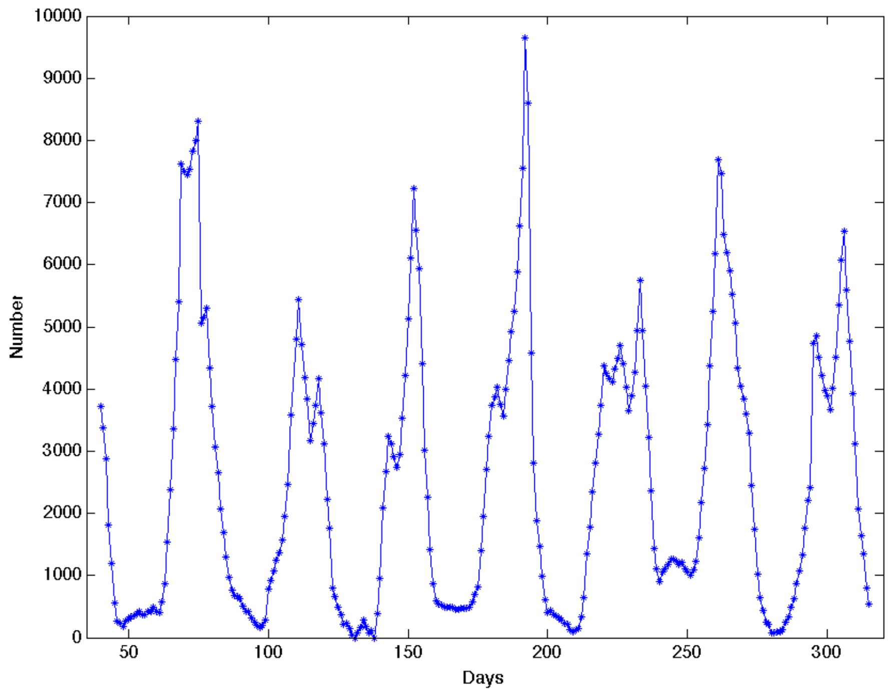

Figure 1.2: Blowfly Data, collected by Nicholson

of the data, see figure 1.2. Nicholson's dataset is now a classical benchmark for evaluating time series algorithms due its intrinsic complexity. Nevertheless, the following DDE is commonly acknowledged as a correct model [11, 23]:

$$\dot{N} = PN(t-\tau)\exp\left(-N(t-\tau)/N_0\right) - \delta N(t) \tag{1.3}$$

whose parameter fitting (of P, N_0, δ) remains delicate. In particular classical NLS are difficult to use in this setting as the initial condition, which is a function defined on $[-\tau, 0]$, is unknown. Alternative solutions, such as Gradient Matching or Bayesian Methods (based on ABC, [41]) give reliable estimates that reproduce the observed dynamics without estimation of the initial condition. These aforementioned methods use particular statistics or functions of the model that provides high-level information on the parameters. The Orthogonal Conditions estimator has a similar approach for dealing with the estimation of Differential Equations.

2 Differential Equation Model and Gradient Matching

2.1 ODE models and Gradient Matching

For ease of readability, we focus on a two-dimensional system of ODEs. In our case, as there is no computational and theoretical differences between the situation $d = 2$ and $d > 2$, there is no lack of generality by this assumption. We consider noisy observations $Y_1, \ldots, Y_n \in \mathbb{R}^2$ of the function ϕ^* measured at random times $t_1 < \cdots < t_n \in [0,1]$:

$$Y_i = \phi^*(t_i) + \epsilon_i \tag{2.1}$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d with $E(\epsilon_i) = 0$ and $V(\epsilon_i) = \sigma^2 I_2$. We suppose that the regression function ϕ^* belongs to the Sobolev space $H^1 = \{u \in L^2([0,1]) \ |\dot{u} \in L^2([0,1])\}$, and ϕ^* is a solution to the parametrized Ordinary Differential Equation (1.1), i.e. there exists a *true* parameter $\theta^* \in \Theta \subset \mathbb{R}^p$ such that for $t \in [0,1]$ almost everywhere (a.e.)

$$\dot{\phi}^*(t) = \boldsymbol{f}\left(t, \phi^*(t), \theta^*\right) \tag{2.2}$$

where $\boldsymbol{f} = (f_1, f_2)$ is a vector field from $[0,1] \times \mathcal{X} \times \Theta$ to $\mathbb{R}^2$, where $\mathcal{X} \subset \mathbb{R}^2$.

The statistical problem can be seen as a noisy version of a parametrized Multipoint Boundary-Value Problem (MBVP, [4]). MBVP deals with the existence, uniqueness and computation of a solution ϕ^* to equation (1.1), with general boundary conditions $\phi^*(t_1) = y_1, \ldots, \phi^*(t_n) = y_n$, $n \geq 2$. Obviously, MBVP is a much more difficult problem than the classical Initial Value Problem although some theoretical results do exist in some restricted cases ([3, 27] and references therein). On the computational side, numerous algorithms such as collocation, multiple shooting,... have been proposed to solve general Boundary Value Problems, [2]. Among them, the 2 points Boundary Value Problem (BVP) where $G\left(\phi^*(0), \phi^*(1)\right) = 0$ with G a given function, is one of the most common and important one, as it arises in numerous applications (physics, control theory,...). We emphasize that a convenient way to deal theoretically and computationally with BVP, in particular linear second order differential ODEs, is not based on an adaptation of the IVP theory, but it rather involves elaborated concepts from functional analysis such as weak derivative, variational formulation and Sobolev spaces [10]. If we denote the inner product of L^2 as $\forall \varphi, \psi \in L^2([0,1])$, $\langle \varphi, \psi \rangle = \int_0^1 \varphi(t)\psi(t)dt$, the weak derivative of the function g in H^1 is not defined point-wise but as the function $\dot{g} \in L^2$ satisfying $\langle \dot{g}, \varphi \rangle = -\langle g, \dot{\varphi} \rangle$, for all function φ in C^1 with support included in $]0,1[$ (denoted $C_C^1(]0,1[)$). Of course, if $t \mapsto \phi(t, x_0, \theta)$ is a C^1 function on $]0,1[$, the classical derivative $\dot{\phi}$ is also the weak derivative. We introduce then the (weak) variational formulation of the ODE (1.1): a weak solution g to (1.1) is a function in H^1 such that $\forall \varphi \in C_C^1(]0,1[)$

$$\int_0^1 \boldsymbol{f}(t, g(t), \theta)\varphi(t)dt + \int_0^1 g(t)\dot{\varphi}(t)dt = 0 \tag{2.3}$$

This variational formulation is the key of the Finite Elements Method which is the reference approach for solving Boundary Value Problems and Partial Differential Equations, [6]. In the case of ODEs, this formulation is not well used for computing solutions, because the geometry of the (1-D) interval $]0,1[$ is simple, and it is easy to build a spline approximation by collocation that solves approximately the ODE. Nevertheless, the characterization (2.3) is useful for the statistical inference task, as it enables to give necessary conditions for a good estimate. In particular, we emphasize that we do not solve the ODE, but we want to identify a parameter θ indexing the vector field $\boldsymbol{f}$. Hence, we develop a new algorithmic approach, different from the one used for solving the direct problem.

2.2 Definition

We define a new gradient matching estimator based on (2.3): starting from a nonparametric estimator $\hat{\phi}$, computed from the observations (t_i, y_i), $i = 1, \ldots, n$, we want to find the parameter θ that minimizes the discrepancy between the parametric derivative $t \mapsto \boldsymbol{f}\left(t, \hat{\phi}(t), \theta\right)$ and a nonparametric estimate of the derivative, e.g. $\dot{\hat{\phi}}$. A classical discrepancy measure is the L^2 distance, that gives rise to the two-step estimator $\hat{\theta}^{TS}$ defined as $\hat{\theta}^{TS} = \arg\min_{\theta \in \Theta} R_{n,w}(\theta)$ where

$$R_{n,w}(\theta) = \int_0^1 |\dot{\hat{\phi}}(t) - \boldsymbol{f}\left(t, \hat{\phi}(t), \theta\right)|^2 w(t)dt. \tag{2.4}$$

This estimator is consistent for several usual nonparametric estimators [8, 19, 14], but the use of a positive weight function w vanishing at the boundaries ($w(0) = w(1) = 0$) is needed to get the classical parametric root-n rate. The importance of the weight function w for the asymptotics of $\hat{\theta}^{TS}$ is assessed by theorem 3.1 in [8]. Indeed, if w does not vanish at the boundaries, then $\hat{\theta}^{TS}$ does not have a root-n rate, because the asymptotics is then dominated by the nonparametric estimates $\hat{\phi}(0)$ and $\hat{\phi}(1)$. The usefulness of such weighting function is well acknowledged in nonparametric or semiparametric estimation. For instance, the so-called weighted average derivative is based on a similar weight function in order to get estimators with parametric rate in partial index models [25].

The use of a nonparametric proxy (instead of a solution to be computed) gives the opportunity to consider parameter estimation in f_1 and in f_2 separately. For this reason and ease of readability, we consider only the estimation of the parameter θ_1 when $\boldsymbol{f}$ can be written $\boldsymbol{f}(t,x,\theta) = (f_1(t,x,\theta_1), f_2(t,x,\theta_2))^\top$ and $\theta = (\theta_1,\theta_2)^\top$ ($\theta_i \in \mathbb{R}^{p_i}$ and $p_1 + p_2 = p$). The joint estimation of $\theta = (\theta_1,\theta_2)^\top$ can be done by stacking the observations into a single column: there is no consequence on the asymptotics, but the estimator covariance matrix has to be slightly modified in order to take into account the correlations between the two equations f_1 and f_2. Having said that, we write simply $f = f_1$ and $\theta = \theta_1$ and we consider only one equation $\dot{x}_1 = f(t,x,\theta)$. We use a nonparametric estimator $\hat{\phi} = (\hat{\phi}_1, \hat{\phi}_2)$ of $\phi^* : [0,1] \to \mathbb{R}^2$.

Starting from (2.3), a reasonable estimator $\hat{\theta}$ should satisfy the weak formulation

$$\forall \varphi \in C_C^1\left(]0,1[\right), \int_0^1 f\left(t, \hat{\phi}(t), \hat{\theta}\right)\varphi(t)dt + \int_0^1 \hat{\phi}_1(t)\dot{\varphi}(t)dt = 0. \tag{2.5}$$

The vector space $C_C^1\left(]0,1[\right)$ is not tractable for variational formulation, and one prefers Hilbert space with a structure related to L^2. In our case, we use $H_0^1 = \{h \in H^1 | h(0) = h(1) = 0\}$ which has a simple description within L^2: an orthonormal basis is given by the sine functions $t \mapsto \sqrt{2}\sin(\ell\pi t), \ell \geq 1$ and we have

$$H_0^1 = \left\{ \sum_{\ell=1}^{\infty} a_\ell \sqrt{2}\sin\left(\ell\pi t\right) \middle| \sum_{\ell=1}^{\infty} \ell^2 a_\ell^2 < \infty \right\} \tag{2.6}$$

Hence, it suffices to consider a countable number of orthogonal conditions (2.5) defined, for instance, with the test functions $\varphi_\ell = \sqrt{2}\sin(\ell\pi t)$, $\forall \ell \geq 1$:

$$\mathcal{C}_\ell(\theta) : \int_0^1 f\left(t, \hat{\phi}(t), \hat{\theta}\right)\varphi_\ell(t)dt + \int_0^1 \hat{\phi}(t)\dot{\varphi}_\ell(t)dt = 0. \tag{2.7}$$

More generally, we consider a family of orthonormal functions $\varphi_\ell \in H_0^1$, with $\ell \geq 1$, and we introduce the vector space $\mathcal{F} = \overline{span\{\varphi_\ell, \ell \geq 1\}}$. The vector space $\mathcal{F}$ may not be necessarily dense in H_0^1, as the functions φ_ℓ could be chosen for computational tractability or because of a natural interpretation (for instance B-splines, polynomials, wavelets, ad-hoc functions, ...). For this reason, we introduce the orthogonal decomposition of $H_0^1 = \mathcal{F} \oplus \mathcal{F}^\perp$, where $\mathcal{F}^\perp = \{g \in H_0^1 | \langle g, \varphi \rangle = 0, \varphi \in \mathcal{F}\}$, and we can have $\mathcal{F} \neq H_0^1$. In general, an estimator $\hat{\theta}$ satisfying $\mathcal{C}_\ell(\hat{\theta})$ for $\ell \geq 1$ also approximately satisfies (2.5). However in practice, we will use a finite set of orthogonal constraints defined by L test functions $(L > p)$. In order to discuss the influence of the choice of $\mathcal{F}$ and of finite dimensional subspace spanned by $\varphi_1, \ldots, \varphi_L$ we introduce the nonlinear operator $\mathcal{E} : (g, \theta) \mapsto \mathcal{E}(g, \theta)$, such that $t \mapsto \mathcal{E}(g, \theta)(t) = f(t, g(t), \theta)$.

For all θ in Θ and g in H^1, the Fourier coefficients of $\mathcal{E}(g,\theta) - \dot{g}$ in the basis $(\varphi_\ell)_{\ell \geq 1}$ are $e_\ell(g, \theta) = \langle \mathcal{E}(g, \theta) - \dot{g}, \varphi_\ell \rangle = \langle \mathcal{E}(g,\theta), \varphi_\ell \rangle + \langle g, \dot{\varphi}_\ell \rangle$, and we introduce the vectors in $\mathbb{R}^L$ $\boldsymbol{e}_L(g,\theta) = (e_\ell(g,\theta))_{\ell=1..L}$ and $\boldsymbol{e}_L^*(\theta) = (e_\ell(\phi^*, \theta))_{\ell=1..L}$. Finally, our estimator is defined by minimizing the quadratic form $Q_{n,L}(\theta) = \left| \boldsymbol{e}_L(\hat{\phi}, \theta) \right|^2$:

$$\hat{\theta}_{n,L} = \arg\min_{\theta \in \Theta} Q_{n,L}(\theta). \tag{2.8}$$

$\hat{\theta}_{n,L}$ is the parameter that "almost" vanishes the first L Fourier coefficients in the orthogonal decomposition of $H_0^1 = \mathcal{F} \oplus \mathcal{F}^\perp$:

$$\mathcal{E}(g,\theta) - \dot{g} \quad = \quad \boldsymbol{E}_L(g,\theta) + \boldsymbol{R}_L(g,\theta) + \boldsymbol{E}_{\mathcal{F}}^\perp(g,\theta)$$

with $\boldsymbol{E}_L(g,\theta) = \sum_{\ell=1}^L e_\ell(g,\theta) \varphi_\ell$, $\boldsymbol{R}_L(g,\theta) = \sum_{\ell > L} e_\ell(g,\theta)\varphi_\ell$ and $\boldsymbol{E}_{\mathcal{F}}^\perp(g,\theta) \in \mathcal{F}^\perp$.

The function $\boldsymbol{E}_{\mathcal{F}}^\perp(\phi^*, \theta)$ represents the behavior of $\mathcal{E}(g,\theta) - \dot{g}$ at the boundaries of the interval $[0,1]$. As $\hat{\phi}$ approaches ϕ^* asymptotically in supremum norm, the objective function $Q_{n,L}(\theta)$ is close to $Q_L^*(\theta) = \|\boldsymbol{E}_L(\phi^*, \theta)\|_{L^2}^2$. The discriminative power of $Q_L^*(\theta)$ can be analyzed locally around its global minimum θ_L^*, as it behaves approximately as the quadratic form $Q_L^*(\theta) \approx (\theta - \theta_L^*)^\top \mathbf{J}_{\theta,L}^{*\top} \mathbf{J}_{\theta,L}^* (\theta - \theta_L^*)$ where $\mathbf{J}_{\theta,L}^*$ is the matrix in $\mathbb{R}^{L \times p}$ with entries $\int_0^1 f_{\theta_j}(t, \phi^*(t), \theta_L^*)\varphi_\ell(t)dt$, for $j = 1, \ldots, p$, $\ell = 1, \ldots, L$.

2.3 Boundary Conditions and Construction of Orthogonal Conditions

The construction of the orthogonal conditions $e_\ell(\theta)$ exposed in the previous section is generic and can be proposed for numerous types of Differential Equations, in particular for Ordinary and Delay Differential Equations. Moreover, similar orthogonal conditions could be also derived for solutions of PDEs with a relevant set of test functions φ, but this extension is beyond the scope of the present paper. A process for deriving "regular" orthogonal conditions, (i.e that gives rise to root-n consistent estimator, as it is shown in section 4) is to use conditions $\mathcal{C}_\ell(\theta)$ with an integral expression $\int_0^1 h_\ell\left(t, \hat{\phi}(t), \theta\right) dt$. The function $h_\ell : (t, x, \theta) \longrightarrow \mathbb{R}$ must be smooth and must satisfy the remarkable identity $\int_0^1 h_\ell\left(t, \phi^*(t), \theta^*\right) = 0$. The variational formulation generates functions $h_\ell(t, x, \theta) = \left(f\left(t, x, \theta\right) \varphi_\ell(t) - \dot{\varphi}_\ell(t) x\right)$ whereas the classical Gradient Matching considers a single function $h(t, x, y, \theta) = \left\|f\left(t, x, \theta\right) - y\right\|^2 \varphi(t)$, and the variable y is evaluated along the derivative $\dot{\phi}(t)$. The asymptotic analysis shows that the dependency in y can be removed and that h' behaves in fact as a function $h(t, x, \theta)$.

The OC framework then generalizes the classical TS estimator and gives ways to ameliorate it. Among other, the use of the boundary vanishing function φ implies an information loss close to the boundaries. This loss can be sensible in terms of estimation quality, and should be avoided when the boundary values are known. For instance, for an IVP with known initial condition $\phi(0) = \phi_0$, we can derive an orthogonal condition that takes into account the knowledge of ϕ_0. By direct computation, we have

$$\begin{aligned}\int_0^1 h(t, \phi(t), \theta) dt &= \int_0^1 f(t, \phi(t), \theta)\varphi(t) dt - [\phi(1)\varphi(1) - \phi(0)\varphi(0)] \\ &\quad + \int_0^1 \phi(t)\dot{\varphi}(t) dt.\end{aligned}$$

If $\phi(1)$ is unknown, but $\phi(0)$ is known, it suffices to take φ such that $\varphi(1) = 0$ and $\varphi(0) \neq 0$. The orthogonal condition still have the same expression $h(t, x, \theta)$. The same adaptation can be done when boundary values of the derivative are known (called Neumann's condition), for instance $\dot{\phi}(1) = \phi_1'$ is known. Indeed, the ODE gives a relationship between the second order derivative $\ddot{\phi}$ and the state ϕ, as $\ddot{\phi}(t) = \partial_x f(t, \phi, \theta) f(t, \phi, \theta)$. By choosing φ such that $\varphi(0) = 0$ and by Integration By Part, the following identity

$$\varphi(1)\phi_1' = \int_0^1 \partial_x f(t,\phi,\theta) f(t,\phi,\theta)\varphi(t)dt + \int_0^1 f(t,\phi,\theta)\dot{\varphi}(t)dt$$

gives a new condition that exploits the behavior of the solution at the boundary. Obviously, these conditions can be successfully used if the nonparametric proxy satisfies the boundary conditions of interest. At the contrary, it seems rather difficult to integrate such information about the boundary within the criterion $R_{n,w}(\theta)$. The orthogonal conditions introduced in the previous section are a direct exploitation of the ODE model, and the introduction of the space $\mathcal{F}$ is a way to deal with the problem of the choice of the number of conditions and their type. Nevertheless, it would be useful to introduce model specific conditions $h(t,\phi(t),\theta)$ which are known to have a vanishing integral for $\theta=\theta^*$. Our estimator can be thought as a Generalized Method of Moments estimator, but where Moments do characterize curves and not probability distributions. A similar idea has been developed recently in the context of functional data analysis [18].

3 Consistency of the Orthogonal Conditions estimator

In order to obtain precise results with closed-form expression for the bias and variance estimators, we consider series estimators, i.e. estimators expressed as $\hat{\phi}_j = \sum_{k=1}^K \hat{c}_{k,j} p_{kK} = \hat{\mathbf{c}}_j \boldsymbol{p}^K$, where $\boldsymbol{p}^K = (p_{1K}\ ,\ \ldots, p_{kK})$ is a vector of approximating functions and the coefficients $\hat{\mathbf{c}}_j = (\hat{c}_{k,j})_{k=1..K}$ are computed by least squares. For notational simplicity, we use the same functions (and the same number K) for estimating ϕ_1^* and ϕ_2^*. We denote $P^K = (p_{kK}(t_i))_{1\le i,k\le n,K}$ the design matrix and $\mathbf{Y}_j = (y_{i,j})_{i=1..n}$ the vectors of observations. Hence, the estimated coefficients $\hat{\mathbf{c}}_j = \left(P^{K\top}P^K\right)^\dagger P^{K\top}\mathbf{Y}_j$ (where $\dagger$ denotes a generalized inverse) gives rise to the so-called hat matrix $H = P^K\left(P^{K\top}P^K\right)^\dagger P^{K\top}$ and the vector of smoothed observations is $\hat{\boldsymbol{\phi}}_j = H\mathbf{Y}_j$, $j=1,2$. One can typically think of regression splines, [32]. We introduce now the conditions required for the definition and consistency of our estimator.

Condition C1: **(a)** Θ is a compact set of $\mathbb{R}^p$ and θ^* is an interior point of Θ, $\mathcal{X}$ is an open subset of $\mathbb{R}^2$; **(b)** $(t,x)\mapsto f(t,x,\theta^*)$ is L^2-Lipschitz and L^2-Caratheodory (see *Supplementary Material I*, section 1).

Condition C2: **(a)** (Y_i, t_i) are i.i.d. with variance $V(Y|T=t) = \Sigma_\epsilon = \sigma^2 I_2$; **(b)** For every K, there is a nonsingular constant matrix B such that for $P^K = B_p^K(t)$; ***(i)*** the smallest eigenvalue of $E\left[P^K(T)P^K(T)^\top\right]$ is bounded away from zero uniformly in K and ***(ii)*** there is a sequence of constants $\zeta_0(K)$ satisfying $\sup_t \left|P^K(t)\right| \leq \zeta_0(K)$ and $K = K(n)$ such that $\zeta_0(K)^2 K/n \longrightarrow 0$ as $n \longrightarrow \infty$; **(c)** There are $\alpha, \mathbf{c}_{1,K}, \mathbf{c}_{2,K}$ such that $\left\|\phi_j^* - p^K \mathbf{c}_{j,K}\right\|_\infty = \sup_{[0,1]} \left|\phi_j^*(t) - p^K(t)^\top \mathbf{c}_{j,K}\right| = O(K^{-\alpha})$.

Condition C3: There exists $D > 0$, such that the D-neighborhood of the solution range $\mathcal{D} = \{x \in \mathbb{R}^2 | \exists t \in [0,1], |x - \phi^*(t)| < D\}$ is included in $\mathcal{X}$ and f is C^2 in (x, θ) on $\mathcal{D} \times \Theta$ for t in $[0,1]$ a.e. Moreover, the derivatives of f w.r.t x and θ (with obvious notations) f_x, f_θ, f_{xx}, $f_{x\theta}$ and $f_{\theta\theta}$ are L^2 uniformly bounded on $\mathcal{D} \times \Theta$ by L^2 functions $\bar{h}_x$, $\bar{h}_\theta$, $\bar{h}_{x\theta}$,$\bar{h}_{xx}$ and $\bar{h}_{\theta\theta}$ (respectively).

Condition C4: Let $(\varphi_\ell)_{\ell \geq 1}$ be an orthonormal sequence of C^1 functions in H_0^1.

Condition C5: θ^* is the unique global minimizer of $Q_{\mathcal{F}}^*$ and $\inf_{|\theta - \theta^*| > \epsilon} Q_{\mathcal{F}}^*(\theta) > 0$.

Condition C6: There exists L_0 such that for $L \geq L_0$, $\mathbf{J}_{\theta,L}(g, \theta)$ is full rank in a neighborhood of (ϕ^*, θ^*).

Condition **C1** gives the existence and uniqueness of a solution ϕ^* in H^1 to the IVP for $\theta = \theta^*$ and $x(0) = \phi^*(0)$. If f is continuous in t and x, then the derivative $\dot{\phi}^*(t) = f(t, \phi^*(t), \theta^*)$ can be defined on $]0,1[$ and is also continuous, see appendix A. More generally, **C1** does apply when there is a discontinuous input variable, such as in the Ricatti example described in section 1.2.1.

Under condition **C2** (satisfied among others by regression splines with $\zeta_0(K) = \sqrt{K}$), it is known that the series estimator $\hat{\phi}_j$ are consistent estimators of ϕ_j^* for usual norms, in particular $\left\|\hat{\phi}_j - \phi_j^*\right\|_\infty = O_P\left(\zeta_0(K)\left(\sqrt{K/n} + K^{-\alpha}\right)\right)$ (theorem 1, [24]). If ϕ^* is C^s and we use splines then $\alpha = s$ and $\left\|\hat{\phi} - \phi^*\right\|_\infty = O_P\left(K/\sqrt{n} + K^{1/2-s}\right)$.

Condition **C3** is here to control the continuity and regularity of the function $\mathcal{E}$ involved in the inverse problem. Moreover, it provides uniform control needed for stochastic convergence.

Condition **C4** is a sufficient condition for deriving independent conditions $\mathcal{C}_\ell(\theta)$, and normalization is useful only to avoid giving implicitly more weight to a condition w.r.t. the other conditions.

Condition **C5** means that θ^* is a global and isolated minima of $Q^*_{\mathcal{F}}(\theta)$, which is standard in M-estimation [37], but can be hard to check in practice. Indeed, the parametric identifiability of ODE models can be hard to show, even for small systems. No general and practical results do exist for assessing the identifiability of an ODE model [21]: it is useful to discriminate between ODE identifiability, statistical identifiability and practical identifiability. The latter being the most useful but almost impossible to check a priori. The essential meaning of condition **C5** is that the addition of more and more orthogonal conditions should lead to a perfect and univocal estimation of the true parameter. From our experience and by numerical computations, we can check that $Q^*_L(\theta)$ has a unique minima in θ^* in a region of interest, for L big enough (usually $L \simeq 2 \times p$). The natural criterion for estimating θ and for identifiability analysis is

$$Q^*(\theta) = \left\|\mathcal{E}\left(\phi^*, \theta\right) - \mathcal{E}\left(\phi^*, \theta^*\right)\right\|^2_{L^2}$$

but $\left\|\boldsymbol{E}^{\perp}_{\mathcal{F}}\left(\phi^*, \theta\right)\right\|^2_{L^2}$ is withdrawn and we use the quadratic form $Q^*_{\mathcal{F}}(\theta)$ in order to avoid boundary effects. This is needed in order to get a parametric rate of convergence, as in the original two-step criterion (2.4). As a consequence, we lose a piece of information brought by the trajectory $t \mapsto \phi^*(t)$ and we have to be sure that the parameter θ has a low influence on $\left\|\boldsymbol{E}^{\perp}_{\mathcal{F}}\left(\phi^*, \theta\right)\right\|^2_{L^2}$. A favorable case is that it is almost constant on Θ, so that Q^* and $Q^*_{\mathcal{F}}$ are essentially the same functions, with the same global minimum and the same discriminating power. In practice, we can check that **C5** is approximately satisfied by computing numerically the criterion $\boldsymbol{E}_{L'}\left(\phi(\cdot, \hat{\theta}_{n,L}), \theta\right)$, in a neighborhood of $\hat{\theta}_{n,L}$, for $L' \geq L$.

Finally, Condition **C6** is about the influence of the number of test functions used. We use only the first L Fourier coefficients of $\mathcal{E}(g, \theta) - \dot{g}$ to identify the parameter θ, but this might not be sufficient to discriminate between two parameters θ and θ'. In a way, we perform dimension reduction but we need to be sure that we have an exact recovery when L goes to infinity: we expect that the global minimum θ^*_L of $|\mathbf{e}^*_L(\theta)|^2$ is close to the global minimum $\theta*$ of $Q^*_{\mathcal{F}}(\theta) = \left\|\boldsymbol{E}_{\mathcal{F}}\left(\phi^*, \theta\right)\right\|^2_{L^2}$ (found under condition **C5**). We introduce the Jacobian matrices $\mathbf{J}_{\theta,L}\left(g, \theta\right)$ in $\mathbb{R}^{L \times p}$ with entries $\int_0^1 f_{\theta_j}(t, g(t), \theta)\varphi_\ell(t)dt$ and $\mathbf{J}_{x,L}\left(g, \theta\right)$ in $\mathbb{R}^{L \times d}$ with entries $\int_0^1 f_{x_i}(t, g(t), \theta)\varphi_\ell(t)dt$. For this reason, we suppose that $\mathbf{J}^*_{\theta,L}$ is full rank, so that $Q^*_L(\theta)$ is locally strictly convex, with a unique local minimum θ^*_L.

The Jacobian matrix introduced in condition **C6** is classical in sensitivity analysis (in ODE models). Usually, the sensitivity matrix used is the Jacobian of the least squares criterion (similar to $\mathbf{J}_{\theta,L}\left(\phi(\cdot,\hat{\theta}),\theta\right)$); it enables to check a posteriori the identifiability of the parameter θ. Conversely, local non-identifiable parameter (*sloppy parameters*, [15]) can be detected in that case.

Theorem 3.1. *If conditions **C1** to **C6** are satisfied, then*

$$\hat{\theta}_{n,L} - \theta_L^* = O_P(1)$$

and the bias $\boldsymbol{B}_L = \theta_L^* - \theta^*$ tends to zero as $L \to \infty$.

In particular, if we use the sine basis and if $\mathcal{E}\left(\phi^*,\theta\right)$ is in H^1 for all θ, then $\boldsymbol{B}_L = o\left(\frac{1}{L}\right)$.

Remark 3.1. *The convergence rate of the bias $\boldsymbol{B}_L$ can be refined according to the test functions φ_ℓ: if we use B-splines, the bias is controlled by the meshsize $\Delta = \max_{j>1}(\tau_j - \tau_{j-1})$ of the sequence of knots τ_j, $j = 1,\ldots,L$ defining the spline spaces, see section 6 in [34].*

Remark 3.2. *In practice, we have $\boldsymbol{B}_L = 0$ for medium-size L, around $2 \times d \times p$.*

4 Asymptotics

We give a precise description of the asymptotics of $\hat{\theta}_{n,L}$ (rate, variance and normality), by exploiting the well-known properties of series estimators. We consider the linear case, then we extend the obtained results to general nonlinear ODEs. We show in a preliminary step that the asymptotics of $\hat{\theta}_{n,L} - \theta_L^*$ are directly related to the behavior of $\mathbf{e}_L(\hat{\phi},\theta^*)$, which is a classical feature of Moment Estimators.

4.1 Asymptotic representation for $\hat{\theta}_n - \theta_L^*$

From the definition (2.8) of $\hat{\theta}_{n,L}$ and differentiability of f, the first order optimality condition is

$$\mathbf{J}_{\theta,L}\left(\hat{\phi},\hat{\theta}_{n,L}\right)^\top \mathbf{e}_L\left(\hat{\phi},\hat{\theta}_{n,L}\right) = 0 \tag{4.1}$$

from which we derive an asymptotic representation for $\hat{\theta}_{n,L}$, by linearizing $\mathbf{e}_L\left(\hat{\phi},\hat{\theta}_{n,L}\right)$ around θ_L^*. We need to introduce the matrix-valued function defined on $\mathcal{D}\times\theta$ such that $\boldsymbol{M}_L(g,\theta) = \left[\mathbf{J}_{\theta,L}\left(g,\theta\right)^\top \mathbf{J}_{\theta,L}\left(g,\theta\right)\right]^{-1}$

$\mathbf{J}_{\theta,L}(g,\theta)^\top$, and proposition 4.1 shows that $\boldsymbol{M}_L(\hat{\phi},\hat{\theta}_{n,L})$ is also a consistent estimator of $\boldsymbol{M}_L^*$.

Proposition 4.1. *If conditions **C1-C6** are satisfied, then*

$$\left[\mathbf{J}_{\theta,L}\left(\hat{\phi},\hat{\theta}_{n,L}\right)^\top \widetilde{\mathbf{J}}_L\right]^{-1}\mathbf{J}_{\theta,L}\left(\hat{\phi},\hat{\theta}_{n,L}\right)^\top \xrightarrow{P} \boldsymbol{M}_L^* = \left[\mathbf{J}_{\theta,L}^{*\top}\mathbf{J}_{\theta,L}^*\right]^{-1}\mathbf{J}_{\theta,L}^{*\top} \tag{4.2}$$

where the matrix $\widetilde{\mathbf{J}}_L$ is the Jacobian $\mathbf{J}_{\theta,L}$ evaluated at a point $\widetilde{\theta}$ between θ^* and $\hat{\theta}_{n,L}$. Moreover, we have

$$\hat{\theta}_{n,L}-\theta_L^* = -\boldsymbol{M}_L^*\mathbf{e}_L(\hat{\phi},\theta_L^*)+o_P(1). \tag{4.3}$$

4.2 Linear differential equations

We consider the parametrized linear ODE defined as

$$\begin{cases}\dot{x}_1 &= a(t,\theta_1)x_1+b(t,\theta_1)x_2\\ \dot{x}_2 &= c(t,\theta_2)x_1+d(t,\theta_2)x_2\end{cases} \tag{4.4}$$

where $a(\cdot,\theta)$, $b(\cdot,\theta)$, $c(\cdot,\theta)$, $d(\cdot,\theta)$ are in L^2. We focus only on the estimation of the parameter $\theta=\theta_1$ involved in the first equation $\dot{x}_1 = a(t,\theta)x_1+b(t,\theta)x_2$ and we suppose that we have two series estimators $\hat{\phi}_1 = \boldsymbol{p}_K^\top\hat{\mathbf{c}}_1$ and $\hat{\phi}_2 = \boldsymbol{p}_K^\top\hat{\mathbf{c}}_2$ satisfying condition **C2**. The orthogonal conditions are simple linear functionals of the estimators $e_\ell(\hat{\phi},\theta) = \left\langle\hat{\phi}_1,\dot{\varphi}_\ell+a(\cdot,\theta)\varphi_\ell\right\rangle+\left\langle\hat{\phi}_2,b(\cdot,\theta)\varphi_\ell\right\rangle$. Hence the asymptotic behavior of the empirical orthogonal conditions relies on the plug-in properties of $\hat{\phi}_1$ and $\hat{\phi}_2$ into the linear forms $T_\rho : x\mapsto\int_0^1\rho(t)x(t)dt$ where ρ is a smooth function. Moreover, the linearity of series estimator makes the orthogonal conditions $\mathbf{e}_L(\hat{\phi},\theta)$ easy to compute as

$$\mathbf{e}_L(\hat{\phi},\theta) = \mathbf{A}(\theta)\hat{\mathbf{c}}_1+\mathbf{B}(\theta)\hat{\mathbf{c}}_2 \tag{4.5}$$

where $\mathbf{A}(\theta)$ and $\mathbf{B}(\theta)$ are matrices in $\mathbb{R}^{L\times K}$ with entries $A_{\ell,k}(\theta) = \int_0^1\left(a(t,\theta)\varphi_\ell(t)+\dot{\varphi}_\ell(t)\right)p_{kK}(t)dt$ and $B_{\ell,k}(\theta) = \int_0^1\left(b(t,\theta)\varphi_\ell(t)\right)p_{kK}(t)dt$. The gradient of $\mathbf{e}_L(\hat{\phi},\theta)$ is $\mathbf{J}_{\theta,L}\left(\hat{\phi},\theta\right) = \partial_\theta\mathbf{A}(\theta)\hat{\mathbf{c}}_1+\partial_\theta\mathbf{B}(\theta)\hat{\mathbf{c}}_2$ where $\partial_\theta\mathbf{A}(\theta)$ and $\partial_\theta\mathbf{B}(\theta)$ are straightforwardly computed by permuting differentiation and integration. Although $\mathbf{e}_L(\hat{\phi},\theta)$ depends linearly on the observations, we have to take care of the asymptotics as we are in a nonparametric framework and K grows with n. The behavior of linear functionals $T_\rho(\hat{\phi})$ for several nonparametric estimators (kernel regression, series estimators, orthogonal series) is well known [1, 5, 13, 24], and in generality it can be shown that such linear forms can be estimated with the classical root-n rate and that they are asymptotically normal under quite general conditions. In the particular case of series estimators, we rely on theorem 3 of [24] that ensures the root-n consistency and the asymptotic normality of the plugged-in estimators $T_\rho(\hat{\phi}_j)$, $j = 1, 2$ under almost minimal conditions. We will give in the next section the precise assumptions required for root-n consistency of linear and nonlinear functional of the series estimator. Moreover, the variance of $\hat{\theta}_{n,L}$ has a remarkable expression

$$V_{e,L}(\theta) = V\left(\mathbf{e}_L(\hat{\phi}, \theta)\right) = \mathbf{A}(\theta) V\left(\hat{\mathbf{c}}_1\right) \mathbf{A}(\theta)^\top + \mathbf{B}(\theta) V(\hat{\mathbf{c}}_2) \mathbf{B}(\theta)^\top. \tag{4.6}$$

We remark that there is no covariance term between $\hat{\mathbf{c}}_1$ and $\hat{\mathbf{c}}_2$ since we assume that $V\left(Y|T = t\right)$ is diagonal (assumption **C2**), but in all generality, we should add $2\mathbf{A}(\theta)\mathrm{cov}(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)\mathbf{B}(\theta)^\top$. We can use the classical estimates of the variance of $\hat{\mathbf{c}}_1$ and $\hat{\mathbf{c}}_2$ to compute an estimate of $V_{e,L}(\theta)$

$$\widehat{V_{e,L}(\theta)} = \mathbf{A}(\theta) \widehat{V\left(\hat{\mathbf{c}}_1\right)} \mathbf{A}(\theta)^\top + \mathbf{B}(\theta) \widehat{V(\hat{\mathbf{c}}_2)} \mathbf{B}(\theta)^\top \tag{4.7}$$

Thanks to proposition 4.1, we can estimate the asymptotic variance of the estimator $\hat{\theta}_{n,L}$ with the consistent estimator $\hat{\boldsymbol{M}}_L = \boldsymbol{M}_L(\hat{\phi}, \hat{\theta}_{n,L})$ and we estimate $V\left(\hat{\theta}_{n,L}\right)$ by $\widehat{V\left(\hat{\theta}_{n,L}\right)} = \hat{\boldsymbol{M}}_L \widehat{V\left(\mathbf{e}_L(\hat{\phi}, \hat{\theta}_{n,L})\right)} \hat{\boldsymbol{M}}_L^\top$. From the asymptotic normality of the plug-in estimate, we can derive confidence balls with level $1 - \alpha$. For instance, for each parameter θ_i, $i = 1, \ldots, p$:

$$IC(\theta_i; 1-\alpha) = \left[\left(\hat{\theta}_{n,L}\right)_i \pm q_{1-\frac{\alpha}{2}} \widehat{V\left(\hat{\theta}_{n,L}\right)}_{ii}^{1/2}\right]$$

where $q_{1-\alpha/2}$ is the quantile of order $1 - \frac{\alpha}{2}$ of a standard Gaussian distribution. Nevertheless, we recall that these confidence intervals might be affected by the bias of $\hat{\theta}_{n,L}$ depending on L.

4.3 Nonlinear differential equations

We give here general results for the asymptotics of $e_\ell(\hat{\phi},\theta)$ when the functional is linear or not in $\hat{\phi}$. In [24], the root-n consistency and asymptotic normality is obtained if the functional $g \mapsto e_\ell(g,\theta)$ has a continuous Fréchet derivative $De_\ell(g,\theta)$ with respect to the norm $\|\cdot\|_\infty$. If $x \mapsto f(t,x,\theta)$ is twice continuously differentiable for $t \in [0,1]$ a.e. in x and θ in Θ, then we can compute easily its Fréchet derivative for $g \in H^1$ in the uniform ball $\|g-\phi^*\|_\infty \leq D$. For all $h \in H^1$ such $\|g+h-\phi^*\|_\infty \leq D$, we have

$$e_\ell(g+h,\theta) - e_\ell(g,\theta) = \langle f_x(\cdot,g,\theta)\,h,\varphi_\ell\rangle + \langle h,\dot{\varphi}\rangle + \left\langle h^\top f_{xx}(\cdot,\tilde{g},\theta)\,h,\varphi_\ell\right\rangle \tag{4.8}$$

by a Taylor expansion around g. As in the linear case, we introduce the tangent linear operator $\mathcal{A}_g(\theta) : u \mapsto \dot{u} - a_g(t,\theta)u$ with $a_g(t,\theta) = f_{x_1}(t,g(t),\theta)$ and the function $b_g(t,\theta) = f_{x_2}(t,g(t),\theta)$. For all θ, the Fréchet derivative of $e_\ell(g,\theta)$ (w.r.t to the uniform norm) is the linear operator $h = (h_1,h_2) \mapsto De_\ell(g,\theta).h = \langle h_1,\dot{\varphi}_\ell + a_g(t,\theta)\varphi_\ell\rangle + \langle h_2, b_g(\cdot,\theta)\varphi_\ell\rangle$ and satisfies for all $\theta \in \Theta$

$$|e_\ell(g+h,\theta) - e_\ell(g,\theta) - De_\ell(g,\theta).h| \leq C\,\|h\|_\infty^2$$

because f_{xx} is uniformly dominated on $\mathcal{D}\times\Theta$. Moreover, for all ϵ (with $0<\epsilon<D$), for all g,g' such that $\|g-\phi^*\|_\infty, \|g'-\phi^*\|_\infty \leq \epsilon$, we have

$$\begin{aligned}
|De_\ell(g,\theta).h - De_\ell(g',\theta).h| &\leq \int_0^1 h(t)^\top f_{xx}(t,\tilde{g}(t),\theta)\,(g(t)-g'(t))\,\varphi_\ell(t)dt \\
&\leq C\,\|h\|_\infty \|g-g'\|_\infty
\end{aligned}$$

with C, a constant independent of θ, ϵ and g,g' (because f_{xx} is uniformly dominated).

As in the linear case, we need to evaluate $De_\ell(g,\theta)$ on the basis $\boldsymbol{p}^K$. We denote $\mathbf{A}(g,\theta)$ and $\mathbf{B}(g,\theta)$ the matrices in $\mathbb{R}^{L\times K}$ with entries $\int_0^1 a_g(t,\theta)\varphi_\ell(t)p_{kK}dt$ and $\int_0^1 b_g(t,\theta)\varphi_\ell(t)p_{kK}dt$ (respectively) and we have the approximation

$$\mathbf{e}_L(\hat{\phi},\theta) = \mathbf{e}_L(\phi^*,\theta) + \mathbf{A}(\phi^*,\theta)\hat{\mathbf{c}}_1 + \mathbf{B}(\phi^*,\theta)\hat{\mathbf{c}}_2 + O\left(\|h\|_\infty^2\right). \tag{4.9}$$

We can derive the asymptotic variance of $\mathbf{e}_L(\hat{\phi},\theta)$ from (4.9)

$$V_{e,L}(\theta) \quad = \quad \mathbf{A}(\phi^*,\theta)V\left(\hat{\mathbf{c}}_1\right)\mathbf{A}(\phi^*,\theta)^\top + \mathbf{B}(\phi^*,\theta)V(\hat{\mathbf{c}}_2)\mathbf{B}(\phi^*,\theta)^\top \tag{4.10}$$

and we can get an estimate $\widehat{V_{e,L}(\theta)}$ from the data as in the linear case.

In order to assess the previous discussion and for deriving the root-n rate of our estimator, we introduce the following two conditions:

Condition C7: **(a)** The times $T_1,\ldots,T_n$ have a density π w.r.t. Lebesgue measure such $0 < c < \pi < C < \infty$; **(b)** $E\left[\epsilon^4\right] < \infty$.

Condition C8: For $\ell = 1,\ldots,L$, $\theta\in\Theta$, there exists $\tilde{\beta}_{K_\ell}$ in $\mathbb{R}^{K_\ell}$ with $\|\frac{f_x(\cdot,\phi^*,\theta)\varphi_\ell+\dot{\varphi}_\ell}{\pi} - \tilde{\beta}_{K_\ell}^\top p^{K_\ell}\|_{L^2} \longrightarrow 0$.

Conditions **C7** and **C8** are similar to the assumptions given in [24]. Condition **C8** is here to ensure that the Fréchet derivative $De_\ell(\phi^*,\theta)$ that drives the asymptotic rate of $e_\ell(g,\theta)$ (see equation 4.8) can be well approximated in the basis $\boldsymbol{p}^K$ as the nonparametric proxy. Then the linearized nonlinear functional of the nonparametric estimator is well approximated by a linear combination of the regression coefficients. When we use B-splines with uniform knot sequence, condition **C8** can be replaced by the simpler condition **C9**:

Condition C9: **(a)** The series estimator is a regression spline with a uniform knot sequence $(\tau_{1,K},\ldots,\tau_{N_K,K})$ defining the spline basis $\boldsymbol{p}^K$ satisfies $\max_i|\tau_{i+1,K}-\tau_{i,K}| \longrightarrow 0$ as $K\longrightarrow\infty$; **(b)** For all $\theta\in\Theta$, for $\ell = 1\ldots L$, $v_\ell : t\mapsto \frac{f_x(t,\phi^*(t),\theta)\varphi_\ell(t)+\dot{\varphi}_\ell(t)}{\pi(t)}$ is C^1.

Theorem 4.1. *If either the following conditions are satisfied:*

$\boldsymbol{p}^K$ **is a general series estimators** *Under conditions **C1-C8** and if f is a linear vector field or, f is a nonlinear vector field and K is chosen such that* $\frac{\zeta_0(K)^4K^2}{n}\longrightarrow 0$

$\boldsymbol{p}^K$ **is a uniform knot splines** *Under conditions **C1-C2(a),C3-C7,C9** and if f is a linear vector field and* $\frac{K^2}{n}\longrightarrow 0$, *or f is a nonlinear vector field and* $\frac{K^4}{n}\longrightarrow 0$

Then $\hat{\theta}_{n,L}$ is such that

$$\sqrt{n}\left(\hat{\theta}_{n,L} - \theta_L^*\right) \rightsquigarrow N(0, \mathbf{V}_L^*) \tag{4.11}$$

with

$$\mathbf{V}_L^* = \mathbf{M}_L^* \mathbf{V}_{e,L}^* \mathbf{M}_L^{*\top}. \tag{4.12}$$

where $\mathbf{V}_{e,L}^* = V_{e,L}(\theta_L^*)$. The asymptotic variance can be estimated by $\hat{\boldsymbol{M}}_L \widehat{V_{e,L}(\hat{\theta}_{n,L})} \hat{\boldsymbol{M}}_L^\top \xrightarrow{P} \mathbf{V}_L^*$. In particular, if we use regression splines and $t \mapsto f(t, \phi^*(t), \theta)$ is C^s on $[0,1]$ with $s \geq 3$, then (4.11) holds with K such that $\sqrt{n}K^{-s} \to 0$ and $n^{-1}K^4 \to 0$.

Moreover, if $L = L(n) \longrightarrow \infty, n \longrightarrow \infty$ is chosen such that the bias $\boldsymbol{B}_{L(n)} = O(n^{-1/2})$, then we have

$$\hat{\theta}_{n,L(n)} - \theta^* = O_P(n^{-1/2}). \tag{4.13}$$

In particular, this is the case when the test functions φ_ℓ are the sine basis, and $L(n) = O(n^\alpha)$ with $\alpha > 1/2$.

This theorem is a direct application of theorem 3 in [24] that claims the root-n consistency and asymptotic normality of general (nonlinear) plug-in estimators. The main steps of the proof are given in *Supplementary Material I*.

5 Experiments

5.1 Description of the setting

We compare the NLS estimator $\hat{\theta}^{NLS}$, the Two-Step Estimator (TS) $\hat{\theta}^{TS}$ and the OC estimator $\hat{\theta}^{OC}$ for varying sample sizes ($n = 400$, 200, 50) and varying noise levels (high and small). We consider 3 different ODEs with different mathematical structure: the α-pinene ODE (linear in state and in parameter), the Ricatti ODE (nonlinear in state, linear in parameter) and the FitzHugh-Nagumo ODE (nonlinear in state and in parameter). Experiments on these three different models provide a good idea of the behavior of the different estimators in terms of the robustness, consistency and efficiency. It helps also in assessing the quality of the linear approximation for the asymptotics (in particular for the computation of the covariance matrix).

In the simulations, the noise is homoscedastic and Gaussian, so that the NLS are asymptotically efficient. Hence, the settings $n = 200$ or $n = 400$ indicates the efficiency loss of the Gradient Matching estimators whereas the small size setting ($n = 50$) gives some information on the small sample case, where the asymptotic approximations cannot be assessed.

As the standard reference method, the Sum of Squared Errors (SSE) is minimized by a Levenberg-Marquardt algorithm using 20 starting points centered around the true parameter value θ^*, and we retain the best minimum. The solution of the ODE is computed by a Runge-Kutta algorithm of order 4, implemented in the Matlab function *ode45*. Hence, we expect that we obtain the true NLS estimator, and that the estimated variance is the true best one.

The Gradient Matching estimators (TS and OC) use the same regression spline, decomposed on a B-spline basis with a uniform knots sequence $\xi_k, k = 1, \ldots, K$. For each dataset (and each dimension), the number of knots is selected by minimizing the GCV criterion, [32]. For the plain TS estimator, we use a piecewise affine weight function with $w(0) = w(1) = 0$, as in [8].

The Orthogonal Conditions are defined with the sine basis or B-Splines basis. We have to face with the practical problem of finding the best number of conditions L, that depends on the model and on $\hat{\phi}$. In each setting, we have fixed a minimum and a maximum number of conditions $\boldsymbol{L}_{min}$ and $\boldsymbol{L}_{max} \leq 2 \times d \times p$ and we select the OC estimator $\hat{\theta}_{n,L}$ that gives the smallest prediction error (i.e that minimizes the SSE):

$$\hat{\theta}^{OC} = \arg \min_{\boldsymbol{L}_{min} \leq L \leq \boldsymbol{L}_{max}} \sum_{i=1}^{n} \left\| y_i - \phi(t_i, \hat{\phi}_0, \hat{\theta}_{n,L}) \right\|^2$$

where $\hat{\phi}_0 = \hat{\phi}(0)$ is the nonparametric estimate of the initial condition.

We use Monte Carlo simulations, based on $N_{MC} = 500$ independent draws for comparing the estimators. We compute their Mean Squared Errors $\left\| \hat{\theta} - \theta^* \right\|^2$. The accuracy of the estimator is roughly estimated by the trace of the covariance matrices of the estimators, denoted $Tr\left(V(\hat{\theta})\right)$. Moreover, the reliability of the estimates (and asymptotic approximation) is evaluated with the coverage probabilities of the 95% confidence ellipse (except in the case of TS because there is no closed-form for asymptotic variance). For the NLS, the asymptotic variance is computed via the Matlab function *nlinfit*. A more detailed analysis of the experiments (including coverage probabilities of confidence sets) are given in

Supplementary Materials II: Experiments, Tables and Figures.

5.2 α-pinene

A linear ODE with constant coefficients is written $\dot{x} = \boldsymbol{A}x$, where $\boldsymbol{A}^\top = (A_1|\ldots|A_d)$. For $i = 1,\ldots,d$, the weak formulation gives the identity $\boldsymbol{Y}_i^\varphi = \boldsymbol{X}^\varphi A_i$ to be satisfied, where $\boldsymbol{X}^\varphi$ is a $d \times L$ matrix with entries $\langle x_k, \varphi_\ell \rangle$ and$\boldsymbol{Y}_i^\varphi$ is a vector in $\mathbb{R}^L$ with entries equal to $-\langle x_i, \dot{\varphi}_\ell \rangle$. For illustration, we consider the *α-pinene ODE* used in [22] for the comparison of several global optimization algorithms:

$$\begin{cases} \dot{x}_1 &= -(\theta_1+\theta_2)x_1 \\ \dot{x}_2 &= \theta_1 x_1 \\ \dot{x}_3 &= \theta_2 x_1 - (\theta_3+\theta_4)x_3 + \theta_5 x_5 \\ \dot{x}_4 &= \theta_3 x_3 \\ \dot{x}_5 &= \theta_4 x_3 - \theta_5 x_5 \end{cases} \tag{5.1}$$

The true parameter to be estimated from a completely observed trajectory on $[0, 100]$ is $\theta^* = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)^\top$. As this ODE is linear and time-invariant, we have a closed-form for the solution $\phi^*(t, \theta, \phi_0) = e^{tA}\phi_0$ that can be directly used for the computation of the NLS estimator.

The test functions used for the OC estimators are B-Splines (with uniform knots sequence) $\varphi_\ell, \ell = 1,\ldots,L$ with compact support included in $]0, 20[$. We consider a varying number of conditions L, i.e $2 \leq L \leq 15$. Finally, we have two settings for the estimation of θ: when the initial condition ϕ_0 is known (and equal to $(100, 0, 0, 0, 0)^\top$ as in [31]), and when ϕ_0 is unknown and needs to be estimated (for NLS).

5.2.1 Known initial condition

For the OC and TS estimator, we constrain the spline estimator $\hat{\phi}$ to satisfy the condition $\hat{\phi}(0) = \phi_0$ (by adding a linear constraint to the classical least-squares minimization). Moreover, following section 2.3, we integrate the knowledge of the initial condition by adding a test function φ_0 which is a B-spline with $\varphi_0(0) \neq 0$. Hence, we define 2 differents OC estimators, respectively, $\hat{\theta}^{OC,0}$ and $\hat{\theta}^{OC}$ that uses or not (resp.) the knowledge of the initial condition.

$\times 10^{-2}$	MSE					$Tr\left(V(\hat{\theta})\right)$		
(n,σ)	TS	OC	OC,0	NLS		OC	OC,0	NLS
$(400,3)$	0.72	0.05	0.04	0.02		0.04	0.04	0.02
$(400,8)$	2.28	0.22	0.25	0.10		0.95	1.20	0.12
$(200,3)$	1.19	0.27	0.30	0.03		0.09	0.13	0.03
$(200,8)$	2.95	0.44	0.37	0.18		2.66	2.68	0.27
$(50,3)$	2.39	0.27	0.26	0.16		1.37	1.58	0.16
$(50,8)$	4.54	1.03	0.93	0.68		7.96	7.27	1.68

Table 5.1: MSE, Asymptotic Variance for α-pinene model with known Initial Condition

5.2.2 Unknown initial condition

In this case, the NLS needs to estimate the initial condition as well, whereas it is not needed for Gradient Matching estimators and we have the same estimates (for $\hat{\theta}^{TS}$ and $\hat{\theta}^{OC}$) as in the previous section. In this setting, we consider another OC estimator that uses information about the other boundary $T = 100$. Indeed, we know that the $\alpha-$pinene network converges to a stationary point, that is almost reached at time $T = 100$. Hence the boundary condition $\dot{\phi}^*(100) = 0$ can be used for estimation (section 2.3): if φ_1 is a test function with $\varphi_1(100) \neq 0$, we have $A^2 < \phi^*, \varphi_1 > +A < \phi^*, \dot{\varphi}_1 >= 0$. This gives an additional condition to be satisfied for the OC estimator, which is denoted as $\hat{\theta}^{OC,1}$, see section 2.3.

$\times 10^{-2}$	MSE					$Tr\left(V(\hat{\theta})\right)$		
(n,σ)	TS	OC	OC,1	NLS		OC	OC,1	NLS
$(400,3)$	0.25	0.11	0.11	0.07		0.10	0.10	0.06
$(400,8)$	1.07	0.85	0.56	0.50		1.06	0.82	0.61
$(200,3)$	0.6	0.37	0.23	0.14		0.25	0.20	0.14
$(200,8)$	1.64	1.42	0.83	1.34		2.36	1.64	1.54
$(50,3)$	1.33	1.31	0.80	0.69		1.63	1.02	0.76
$(50,8)$	3.64	2.11	1.79	1.96		5.34	2.20	4.38

Table 5.2: MSE, Asymptotic Variance for α-pinene model with unknown initial conditions

5.3 Ricatti Equation

The true ODE is $\dot{\phi} = a\phi^2 + c\sqrt{t} - d'\mathbb{1}_{[T_r;14]}$, with $a^* = 0.11$, $c^* = 0.09$, $d^* = 2$ and $\phi_0 = -1$, for $t \in [0, 14]$. For all φ in C^1 with $\varphi(0) = \varphi(14) = 0$, we have $< \phi, \dot{\varphi} > +a < \phi^2, \varphi > +c < \sqrt{t}, \varphi >$

$-d'\left(\tilde{\varphi}(14) - \tilde{\varphi}(T_r)\right) = 0$ where $\tilde{\varphi}$ is the antiderivative of φ.

When T_r is known, we use a cubic B-splines basis with 3 knots at T_r, meaning that $\hat{\phi}$ can have a discontinuous derivative at time T_r (hence the curve estimation from noisy data is pretty correct at T_r). The curve is mainly flat for $t \in [0, T_r]$ and after T_r, one can observe a linear behavior: 3 knots are used to estimate the curve, and their positions are selected manually.

When T_r is unknown, it is required to estimate $\theta = (a, c, d', T_r)$. The OC is no more linear in parameters, but $\hat{\theta}^{OC}$ can be computed by solving the general nonlinear program. The Two-Step estimator fails to estimate T_r because the derivative of the solution is badly estimated when T_r is unknown. OC estimators still give reliable estimates as it uses only $\widehat{\phi}$ in the criterion. Some care has to be taken for the knots selection because of unknown T_r: when $n = 200$, 400 we use a uniform grid of 15 knots on [0, 14]. For $n = 50$, we have used 8 knots uniformly located on [0, 14]. Nevertheless, the nonparametric estimates are too rough to obtaining any correct estimate $\hat{\theta}^{TS}$.

Concerning NLS, we were not able to solve the optimization problem and we cannot give Monte Carlo statistics for the evaluation of NLS. NLS collapses in practice because the optimization problem is hard (severely ill-posed problem). Indeed, the Levenberg-Marquardt algorithm becomes very sensitive to initial conditions and gives different solutions for very close starting values, even in the neighborhood of the true value θ^*. Moreover, we have to face with the problem of explosion of the solutions during the optimization process. In particular, this problem is very delicate because we have to chose (a, c) so that the (potential) explosion of the solution can be balanced by a proper choice of d' and T_r. Probably, NLS would benefit from a specific optimization algorithm that could exploit the particular properties of the ODE, but this is out of the scope of the paper.

$\times 10^{-2}$	MSE				$Tr\left(V\left(\hat{\theta}\right)\right)$	
(n,σ)	TS	OC	NLS		OC	NLS
$(400,0.2)$	0.18	0.27	0.58		1.76	0.10
$(400,0.4)$	0.78	1.21	0.94		2.56	0.38
$(200,0.2)$	0.33	0.87	0.57		2.85	0.25
$(200,0.4)$	1.12	2.69	1.12		5.64	0.98
$(50,0.2)$	1.03	1.30	1.54		4.70	1.00
$(50,0.4)$	3.80	4.43	3.94		8.89	4.08

Table 5.3: MSE , $Tr\left(V\left(\hat{\theta}\right)\right)$ for Parameter estimation for Ricatti Equation with known T_r

$\times 10^{-2}$	$MSE(\widehat{a})$	$MSE(\widehat{c})$	$MSE(\widehat{d'})$	$MSE(\widehat{T_r})$
(n,σ)	OC	OC	OC	OC
$(400,0.2)$	0.09	0.00	2.54	1.39
$(400,0.4)$	0.29	0.01	4.27	3.54
$(200,0.2)$	0.21	0.00	4.08	3.18
$(200,0.4)$	0.61	0.01	11.96	6.93
$(50,0.4)$	0.64	0.02	11.20	14.25
$(50,0.4)$	0.77	0.01	17.18	19.40

$\times 10^{-2}$	MSE	$Tr\left(V(\hat{\theta})\right)$
(n,σ)	OC	OC
$(400,0.2)$	4.01	3.97
$(400,0.4)$	8.11	8.02
$(200,0.2)$	7.47	7.35
$(200,0.4)$	19.51	18.94
$(50,0.2)$	26.10	5.14
$(50,0.4)$	37.36	9.49

Table 5.4: MSE, Sum Empirical Variance for Parameter estimation for Ricatti with unknown T_r

6 Real data analysis

6.1 Influenza virus growth and migration model

We consider the ODE model introduced in Wu et. al [42] for the growth and migration of influenza virus-specific effector CD8+ T cells, among lymph node (T_E^m), spleen (T_E^s), and lung (T_E^l) of mice. After a model selection process, it turns out that the following model

$$\begin{cases} \frac{d}{dt}X_1 = \rho_m D^m(t-\tau) - \gamma_{ms} \\ \frac{d}{dt}X_2 = \rho_s D^m(t-\tau) - \gamma_{sl} + \gamma_{ms}e^{(X_1-X_2)} \\ \frac{d}{dt}X_3 = \gamma_{sl}e^{(X_2-X_3)} - \delta_l \end{cases} \tag{6.1}$$

is credible for representing the dynamics of the observations. Model (6.1) is written in log-scale (i.e with $X_1 = log(T_E^m)$, $X_2 = log(T_E^s)$ and $X_3 = log(T_E^l)$), and the parameter $\theta = (\rho_m, \rho_s, \delta_l, \gamma_{ms}, \gamma_{sl})^T$ has to be estimated from the data. The function D and the delay are known (estimated from the data).

The available data are the variables T_E^m, T_E^s and T_E^l for six different subjects and are measured at times $T = [0, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 24]$. Following Wu et al., we stabilize the variance by a log transformation, hence we consider directly the variables X_i, $i = 1, 2, 3$. We assume that each subject share the same true parameter θ^* and the same initial conditions: at each time point, we compute the mean of the log-measurement (over the subjects) as pseudo-observations.

We estimate D^m with a spline smoother computed with cubic B-Splines and GCV selection for the knots. As in Wu et al, the nonparametric proxy is a regression spline $\widehat{X} = \left(\widehat{X_1}, \widehat{X_2}, \widehat{X_3}\right)$ defined on $[5, 14]$; we do not consider earlier times since the influenza specific CD8+ T cells are not produced before. Since we have a small number of observations, the choice of the knots for the cubic splines is done manually.

Nevertheless for the parameter estimation, we have tested several estimates $\hat{X}$ (with different knots locations), and different number of tests functions L: we selected $L = 3$ or $L = 4$. The corresponding estimators are denoted $\hat{\theta}_3^{OC}$ and $\hat{\theta}_4^{OC}$. Moreover, in order to improve the accuracy , we have used a weighted version of the OC estimator, similar to the classical "Generalized Methods of Moments" (this procedure is detailed in section 5 of *Supplementary Material I*). The quality of the estimator is evaluated by the SSE:

$$SSE = \sum_{s=1}^{6}\sum_{d=1}^{3}\sum_{i=1}^{N}\left(y_{i,d,s} - \phi_d(t_i, \widehat{\theta}, \widehat{X}(0))\right)^2$$

where $y_{i,d,s}$ is the observation at time t_i for the s-th subject for the transformed variable X_d. As suggested in Wu et al, we use the OC estimates as initial values for NLS estimation. For both estimates, we obtain the same estimator which is then simply denoted as $\hat{\theta}^{NLS}$. We provide three different estimates $\hat{\theta}_3^{OC}$, $\hat{\theta}_4^{OC}$ and $\hat{\theta}^{NLS}$; we mention also $\tilde{\theta}^{ref}$, which is the estimate obtained in Wu et al [42].

	$\hat{\theta}_3^{OC}$	$\hat{\theta}_4^{OC}$	$\hat{\theta}^{NLS}$	$\tilde{\theta}^{ref}$
ρ_m	2.9e-5	2.7e-5	1.5e-5	$1.6e-5$
ρ_s	4.1e-5	4.7e-5	4.1e-5	$4.5e-5$
δ_l	2.0	3.4	3.7	3.96
γ_{ms}	0.39	0.35	0.15	0.157
γ_{sl}	0.72	0.81	0.47	0.49
RMSE	13.5	13.9	9.0	9.5

	$\hat{\theta}_3^{OC}$		$\hat{\theta}_4^{OC}$		$\hat{\theta}^{NLS}$	
	Low. Bound	Up. Bound	Low. Bound	Up. Bound	Low. Bound	Up. Bound
ρ_m	2.1e-5	3.7e-5	1.9e-5	3.4e-5	0.7e-0.5	2.4e-0.5
ρ_s	0.7e-5	7.4e-5	0.9e-5	8.4e-5	3.4e-0.5	4.8e-0.5
δ_l	-1.11	5.21	-0.28	7.21	2.59	4.93
γ_{ms}	0.27	0.50	0.24	0.46	0.03	0.26
γ_{sl}	-0.10	1.55	-0.14	1.76	0.39	0.55

Table 6.1: Estimates, RMSE and the 95% confidence intervals for different L and estimators.

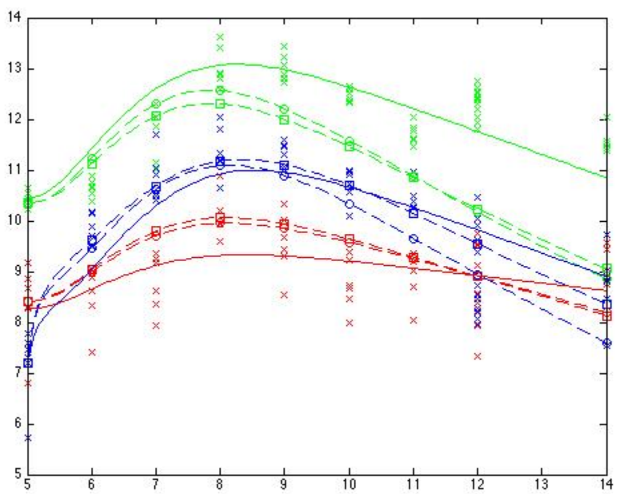

Figure 6.1: Influenza model, Estimated curves for X_1 (red), X_2 (green), X_3 (blue); $\times$: observations, $\square$: solution for $\hat{\theta}_1^{OC}$, $\circ$: solution for $\hat{\theta}_2^{OC}$, solid line: solution with $\hat{\theta}^{NLS}$.

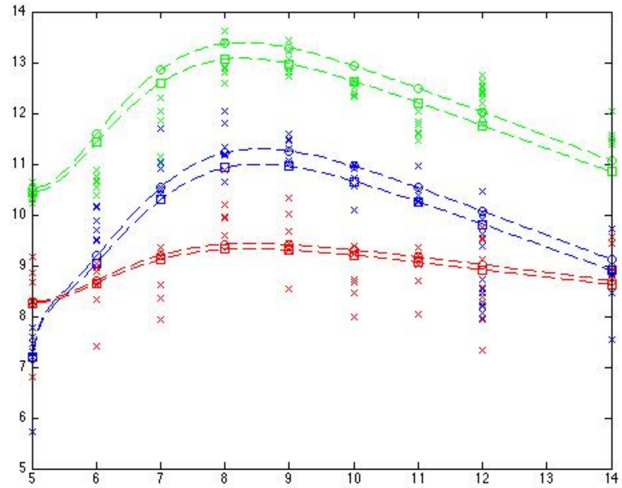

Figure 6.2: Influenza model, Estimated curves for X_1 (red), X_2 (green), X_3 (blue); $\square$ solution obtained with OC+NLS, $\circ$ solution obtained with $\tilde{\theta}^{ref}$.

6.2 Blowfly model

The Delay Differential Equation (1.3) was proposed by Gurney et al [40] to model the dynamics of a population of blowflies, from the Nicholson's blowfly data [26]. These data consists of 350 counts taken every two days during between day $40 = T_0$ and day $315 = T_1$. As Gurney did, we take $\tau = 14.8$ days and our aim is to estimate $\theta = (P, N_0, \delta)$. The orthogonal conditions derived from the weak form is $\forall \varphi \in C_c^1\left(]a, b[\right)$,

$$\int_a^b N(u)\dot{\varphi}(u)du + P\int_{a-\tau}^{b-\tau} N(u)e^{-\frac{N(u)}{N_0}}\varphi(u+\tau)du - \delta\int_a^b N(u)\varphi(u)du = 0$$

where $[a, b]$ has to be chosen such that: $[a, b]$,$[a-\tau, b-\tau] \subset [T_0, T_1]$. Due to a change in the dynamics, we have used only the first 180 observations, see [33]. For the nonparametric estimation, we have used 42 knots located between $t = 40$ and $t = 220$. Preliminary tests and comparisons suggests to use the sine basis for the test function φ_ℓ, and we use $2 \leq L \leq 15$. A simulation is given in figure 6.3

	$L = 11$	$L = 9$	$L = 12$
P	7.81	7.52	7.91
N_0	381.8	385.9	377.7
δ	0.154	0.153	0.154
RSSE	1.7136e+03	1.7557e+03	1.7990e+03

	$L = 11$		$L = 9$		$L = 12$	
O.C	Low. Bound	Up. Bound	Low. Bound	Up. Bound	Low. Bound	Up. Bound
P	5.80	9.81	5.64	9.40	5.0416	10.77
N_0	303.62	459.94	306.59	465.38	289.36	465.98
δ	0.10	0.20	0.11	0.19	0.10	0.20

Table 6.2: Estimates, RSSE and 95% confidence intervals for different L

7 Discussion

Among the simulated models we considered (α-pinene, Ricatti), the NLS estimator is often the best estimator in the asymptotic case (and small noise case) in terms of MSE for the parameters. Nevertheless, in some complex case such as unknown initial conditions for α-pinene (with small sample size or high noise level), or Ricatti equation (with known or unknown change point T_r), then TS and OC can offer better statistical performances. The α-pinene model shows the interest of using information on the boundaries in OC (as introduced in section 2.1). Moreover, simulations show that OC can improve on classical TS although it uses only (partial information) about (weak) derivatives. The fact that the NLS can be caught up, even in the very favorable case of a closed-form solution and starting values (for NLS optimization) close to the true parameter indicates that the introduction of Functions Moments offers a competitive estimator to the direct classical for complex case. In the latter case of Ricatti, the TS approaches is uniformly better than NLS, whereas OC is not systematically better than NLS. Ricatti Equation is striking, as it shows that good proxies $\hat{\phi}$ gives a lot of information: when T_r is known, the reconstruction of the solution and its derivative is excellent, which gives a clear advantage to the plain TS. Nevertheless, when T_r is unknown the derivative estimation is of poor quality around T_r, and the TS estimator is unstable and cannot be computed. The same situation occurs for NLS, because of some lack of identifiability and dramatic changes in derivative estimation which makes the optimization algorithms inefficient. For the influenza dataset analysis, the two OC estimators give correct parameter estimates from real and sparse data (the simulated ODE have a correct qualitative behavior). When used as starting for NLS, both estimates give the same NLS estimator, which improves (obviously) the

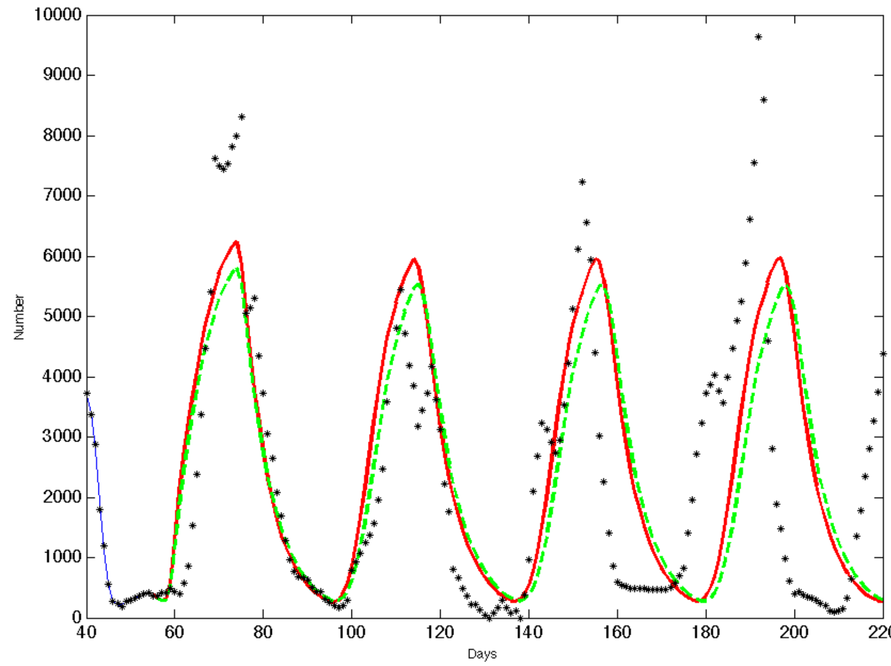

Figure 6.3: Solution N of the Nicholson's DDE simulated with the OC estimator (computed with $L = 11$ conditions - continuous red line). The NLS solution is given by the dashed green curve. The initial function is estimated between day 40 and 55 and the simulation starts after day 55. Drift between data and simulations comes from a chaotic behavior and uncertainty in initial condition (and parameters)

SSE and still gives an estimator closer to the estimates given Wu et al (and same qualitative behavior for the solution). We consider the (self-)consistency of the OC estimates as an indication for reliability of the OC approach. More generally, OC can be used for initializing a NLS estimator, which is often a critical problem in nonlinear regression. In our case, we found a slightly better estimate (for RSS) w.r.t the original paper by Wu et al. For the Delay Differential Equation modeling the blowfly dataset, we insist on the ease of implementation of the method, that avoids the semiparametric estimation of the initial condition. Moreover it provides an estimate close to the posterior mean obtained by ABC: $P^{ABC} = 7.39$, $N_0^{ABC} = 365.03$ and $\delta^{ABC} = 0.15$. With a varying number of Orthogonal Conditional, we can assess the self-consistency of our estimate. Moreover, the posterior mean is always in the 95% confidence set computed for OC.

Acknowledgments

This work was funded by two ANR projects GD2S (ANR-05-MMSA-0013-01) and ODESSA (ANR-09-SYSC-009-01) and received a partial support from the Analysis of Object Oriented Data program (2010-2011) in Statistical and Applied Mathematical Sciences Institute (SAMSI), USA.

References

[1] D. K. Andrews. Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica*, 59(2):307–345, 1991.

[2] U.M. Ascher, R.M.M. Mattheij, and R.D. Russell. *Numerical Solutions of Boundary Value Problems for Ordinary Differential Equations*, volume 478. Prentice Hall New Jersey, 1988.

[3] R. Bellman. A note on the identification of linear systems. *Proc. Amer. Math. Soc.*, 17:68–71, 1966.

[4] R. Bellman, H. Kagiwada, and R. Kalaba. Orbit determination as a multi-point boundary-value problem and quasilinearization. *PNAS*, 48:1327–1329, 1962.

[5] P.J. Bickel and Y. Ritov. Nonparametric estimators which can be plugged-in. *Annals of Statistics*, 31(4):4, 2003.

[6] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer, 2008.

[7] D. Brigo and F. Mercurio. *Interest Rate Models - Theory and Practice*. Springer Finance. Springer-Verlag, 2nd edition edition, 2006.

[8] N. J-B. Brunel. Parameter estimation of ode's via nonparametric estimators. *Electronic Journal of Statistics*, 2:1242–1267, 2008.

[9] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer-Verlag, 2005.

[10] J.B. Conway. *A course in functional analysis*, volume 96. Springer, 1990.

[11] S.P. Ellner and J. Guckenheimer. *Dynamic Models in Biology*. Number vol. 13 in Princeton Paperbacks. Princeton University Press, 2006.

[12] Hein W Engl, Christoph Flamm, Philipp Kügler, James Lu, Stefan Müller, and Peter Schuster. Inverse problems in systems biology. *Inverse Problems*, 25(12), 2009.

[13] L. Goldstein and K. Messer. Optimal plug-in estimators for nonparametric functional estimation. *The annals of statistics*, 20(3):1306–1328, 1992.

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

[15] Ryan N Gutenkunst, Joshua J Waterfall, Fergal P Casey, Kevin S Brown, Christopher R Myers, and James P Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol*, 3(10):e189, 10 2007.

[16] E. L Ionides, A. Bhadra, Y. Atchade, and A. A King. Iterated filtering. *Annals of Statistics*, 39:1776–1802, 2011.

[17] E. L. Ionides, C. Breto, and A. A. King. Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 103:18438–18443, 2006.

[18] G.M. James. Curve alignment by moments. *Annals of Applied Statistics*, 1(2):480–501, 2007.

[19] H Liang and H. Wu. Parameter estimation for differential equation models using a framework of measurement error in regression models. *Journal of the American Statistical Association*, 103(484):1570–1583, December 2008.

[20] J. Madar, J. Abonyi, H. Roubos, and F. Szeifert. Incorporating prior knowledge in cubic spline approximation - application to the identification of reaction kinetic models. *Industrial and Engineering Chemistry Research*, 42(17):4043–4049, 2003.

[21] H. Miao, X. Xia, A. S. Perelson, and H. Wu. On identifiability of nonlinear ode models and applications in viral dynamics. *SIAM Review*, 53:3–39, 2011.

[22] C.G. Moles, P. Mendes, and J.R. Banga. Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Research*, 13:2467–2474, 2003.

[23] B. Murray. *Mathematical Biology, Vol. 1: An Introduction, 3E*. Springer (India) Pvt. Ltd., 2004.

[24] W. K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79:147–168, 1997.

[25] W.K. Newey and T.M. Stoker. Efficiency of weighted average derivative estimators and index models. *Econometrica*, 61(5):1199–1223, 1993.

[26] A. J. Nicholson. The self-adjustement of population to change. *Cold Spring Harbor Symposia on Quantitative Biology*, 22:153–173, 1957.

[27] T. Ojika and W. Welsh. A numerical method for the solution of multi-point problems for ordinary differential equations with integral constraints. *Journal of Mathematical Analysis and Applications*, 72:500–511, 1979.

[28] A. A. Poyton, M.S. Varziri, K.B. McAuley, P.J. McLellan, and J.O. Ramsay. Parameter estimation in continuous-time dynamic models using principal differential analysis. *Computers and Chemical Engineering*, 30:698–708, 2006.

[29] M. Quach, N. Brunel, and F. d'Alche Buc. Estimating parameters and hidden variables in non-linear state-space models based on odes for biological networks inference. *Bioinformatics*, 23(23):3209–3216, 2007.

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

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

[32] D. Ruppert, M.P. Wand, and R.J. Carroll. *Semiparametric regression*. Cambridge series on statistical and probabilistic mathematics. Cambridge University Press, 2003.

[33] Y. Seifu S. P. Ellner and R.H. Smith. Fitting Population Dynamic Models To Time-Series Data By Gradient Matching. *Ecology*, 83(8):2256–2270, 2002.

[34] L. Schumaker. *Spline Functions: Basic Theory.* Cambridge University Press, 3rd edition, 2007.

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

[36] S. van de Geer. *Empirical processes in M-estimation.* Cambridge University Press, 2000.

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

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

[39] E.O. Voit and J. Almeida. Decoupling dynamical systems for pathway identification from metabolic profiles. *Bioinformatics*, 20(11):1670–1681, 2004.

[40] S. P. Bkythe W. S. C.Gurney and R.L. Nisbet. Nicholson's blowflies revisited. *Nature*, 287:17–21, 1980.

[41] S.N. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104, 2010.

[42] H. Wu, A. Kumar, H. Miao, J. Holden-Wiltse, T.R. Mosmann, A.M. Livingstone, G.T. Belz, A.S. Perelson, M.S. Zand, and D.J Topham. 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, 2011.

[43] H. Xue, H. Miao, and H. Wu. 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(4):2351–2387, 2010.