



**HAL**  
open science

## Parametric Estimation of Ordinary Differential Equations with Orthogonality Conditions

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

► **To cite this version:**

Nicolas J.-B. Brunel, Quentin Clairon, Florence d'Alché-Buc. Parametric Estimation of Ordinary Differential Equations with Orthogonality Conditions. 2012. hal-00713368v1

**HAL Id: hal-00713368**

**<https://hal.science/hal-00713368v1>**

Preprint submitted on 30 Jun 2012 (v1), last revised 28 Oct 2014 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# 1 Parametric Estimation of Ordinary Differential Equations with 2 Orthogonality Conditions

3 Nicolas J-B. Brunel<sup>1</sup>, Quentin Clairon<sup>2</sup>, Florence d'Alché-Buc<sup>3,4</sup>

## 4 Abstract

5 Differential equations are commonly used to model dynamical deterministic systems in appli-  
6 cations. When statistical parameter estimation is required to calibrate theoretical models to data,  
7 classical statistical estimators are often confronted to complex and potentially ill-posed optimization  
8 problem. As a consequence, alternative estimators to classical parametric estimators are needed for  
9 obtaining reliable estimates. We propose a gradient matching approach for the estimation of para-  
10 metric Ordinary Differential Equations observed with noise. Starting from a nonparametric proxy of  
11 a true solution of the ODE, we build a parametric estimator based on a variational characterization  
12 of the solution. As a Generalized Moment Estimator, our estimator must satisfy a set of orthogonal  
13 conditions that are solved in the least squares sense. Despite the use of a nonparametric estimator,  
14 we prove the root- $n$  consistency and asymptotic normality of the Orthogonal Conditions estimator.  
15 We can derive confidence sets thanks to a closed-form expression for the asymptotic variance, and  
16 we give a practical way to optimize the variance by adaptive reweighting. Finally, we compare  
17 our estimator in several experiments in order to show its versatility and relevance with respect to  
18 classical Gradient Matching and Nonlinear Least Squares estimators.

19 **Key-words:** Gradient Matching, Non-parametric statistics, Plug-in Property, Variational formulation,  
20 Sobolev Space.

---

<sup>1</sup> ENSIIE & Laboratoire Statistique et Génome, Université d'Évry Val d'Essonne, UMR CNRS 8071 - USC INRA - FRANCE

<sup>2</sup> Laboratoire Analyse et Probabilités, Université d'Évry Val d'Essonne - FRANCE

<sup>3</sup> Laboratoire IBISC, Université d'Évry Val d'Essonne - FRANCE

<sup>4</sup> INRIA-Saclay, LRI, Université Paris Sud, UMR CNRS 8623 - FRANCE

# 1 Introduction

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. A classical model is the Ordinary Differential Equation defined for  $t \in [0, 1]$ ,

$$\dot{x}(t) = \mathbf{f}(t, x(t), \theta) \tag{1.1}$$

where dot indicates derivative with respect to time.  $\mathbf{f}$  is a time-dependent vector field from  $\mathcal{X} \subset \mathbb{R}^d$  to  $\mathbb{R}^d$  which is parametrized by a parameter  $\theta \in \Theta \subset \mathbb{R}^p$ ,  $d, p \geq 1$ . An important task is then the estimation of the parameter  $\theta$  from real data. [21] 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  $\theta$  from noisy observations  $y_1, \dots, y_n \in \mathbb{R}^d$  of a solution at times  $t_1 < \dots < t_n$ .

Most works deal with Initial Value Problems (IVP), i.e. with ODE models having a given (possibly unknown) initial value  $x(0) = x_0$ . There exists then a unique solution  $\phi(\cdot, x_0, \theta)$  of (1.1) defined on the interval  $[0, 1]$ , that depends smoothly on  $x_0$  and  $\theta$ . The estimation of  $\theta$  is a classical problem of nonlinear regression, where we regress  $y$  on the time  $t$ . If  $x_0$  is known, the Nonlinear Least Square Estimator  $\hat{\theta}^{NLS}$  (NLSE) is obtained by minimizing

$$Q_n^{LS}(\theta) = \sum_{i=1}^n |y_i - \phi(t_i, x_0, \theta)|^2 \tag{1.2}$$

where  $|\cdot|$  is the classical Euclidean norm. The NLSE, Maximum Likelihood Estimator or more general M-estimators [23] 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 [21] for a detailed overview of the previous works in this field and of some of the problems encountered in estimating ODE. Global optimization methods are then often used, such as simulated annealing, evolutionary algorithms ([15] for a comparison of such methods). Other classical estimators are obtained by interpreting noisy ODEs as state-space models: filtering and smoothing technics can be used for parameter

1 inference [5], which can provide estimates with reduced computational complexity [20, 11, 10].

2 Nevertheless, the difficulty of the optimization problem is the outward sign of the illposedness of the  
3 inverse problem of ODE parameter estimation, [6]. Hence some improvements on classical estimation  
4 have been proposed by regularizing the statistical inference in an appropriate way. Starting from different  
5 methods used for solving ODEs, different estimators can be developed based on a mixture of nonpara-  
6 metric estimation and collocation approximation. This gives rise to Gradient Matching (or Two-Step)  
7 estimators that consists in approximating the solution  $\phi$  with a basis expansion, such as cubic splines.  
8 The rationale is to estimate nonparametrically the solution  $\phi$  by  $\hat{\phi} = \sum_{k=1}^L \hat{c}_k B_k$  so that we can also  
9 estimate the derivative  $\dot{\hat{\phi}}$ . An estimator of  $\theta$  can be obtained by looking for the parameter that makes  
10  $\hat{\phi}$  satisfy the differential equation (1.1) in the best possible manner. Two different methods have been  
11 proposed, based on a  $L^2$  distance between  $\dot{\hat{\phi}}$  and  $\mathbf{f}(t, \hat{\phi}, \theta)$ : The first one, called the *two-step method*, was  
12 originally proposed by [25], and has been particularly developed in (bio)chemical engineering [13, 26, 18].  
13 It avoids the numerical integration of the ODE and usually gives rise to simple optimization program  
14 and fast procedures that usually performs well in practice. The statistical properties of this two stage  
15 estimator (and several variants) have been studied in order to understand the influence of nonparametric  
16 technics to estimate a finite dimensional parameter [4, 12, 8, 27]. While keeping the same kind of numer-  
17 ical approximation of the solution, [21] proposed a second method based on the generalized smoothing  
18 approach for determining at the same time the parameter  $\theta$  and the nonparametric estimation  $\hat{\phi}$ . The  
19 essential difference between these two approaches is that the nonparametric estimator in the generalized  
20 smoothing approach is computed adaptively with respect to the parametric model, whereas two-step  
21 estimators are “model-free smoothing”.

22 We introduce here a new estimator that can be seen as an improvement of the previous two-step  
23 estimators. It uses also a nonparametric proxy  $\hat{\phi}$ , but we modify the criterion used to identify the  
24 ODE parameter (i.e. the second step). The initial motivation is to get a closed-form expression for  
25 the asymptotic variance and confidence sets. The most notable feature of the proposed method is the  
26 use of a variational characterization of the solution of the differential equations instead of the classical  
27 point-wise one. This characterization is general and can cover a greater number of situations. Thanks to  
28 its computational tractability, we can give a precise description of the asymptotics and give the bias and

1 variance of the estimator. We also give a way to ameliorate adaptively our estimator and to compute  
2 asymptotic confidence intervals.

3 First, we give the ODE model and main assumptions, we introduce our estimator, and show its  
4 consistency. Then, we provide a detailed description of the asymptotics, by proving its root- $n$  consistency  
5 and asymptotic normality. Based on the asymptotic approximation, we give a closed-form expression of  
6 the asymptotic variance, and we address the problem of obtaining the best variance through the choice  
7 of an appropriate weighting matrix. Finally, we provide some insights into the practical behavior of the  
8 estimator through simulations. The objective of the experiments part is to show its interest with respect  
9 to the nonlinear least squares and classical gradient matching estimators.

## 10 2 Differential Equation Model and Gradient Matching

### 11 2.1 ODE models and Gradient Matching

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

$$Y_i = \phi^*(t_i) + \epsilon_i \quad (2.1)$$

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

$$\dot{\phi}^*(t) = \mathbf{f}(t, \phi^*(t), \theta^*) \quad (2.2)$$

20 where  $\mathbf{f} = (f_1, f_2)$  is a vector field from  $[0, 1] \times \mathcal{X} \times \Theta$  to  $\mathbb{R}^2$ , where  $\mathcal{X} \subset \mathbb{R}^2$ . The classical Cauchy-  
21 Lipschitz theory of Initial Value Problem (IVP) can be extended to rougher vector field in time  $t$  (see

1 theorem 3.4 in [17]) under the following assumptions

2 **IVP (a)**  $\mathbf{f}$  is  $L^2$ -Caratheodory, i.e.  $x \mapsto \mathbf{f}(t, x, \theta)$  is continuous for  $t$  a.e in  $[0, 1]$ ,  $x \mapsto \mathbf{f}(t, x, \theta)$  is  
3 measurable for all  $x \in \mathcal{X}$  and  $\forall c > 0, \exists h_c(\cdot, \theta) \in L^2 \setminus |x| \leq c \implies |\mathbf{f}(t, x, \theta)| \leq h_c(t, \theta)$ ,

4 **IVP (b)**  $\mathbf{f}$  is  $L^2$ -Lipschitz, i.e.  $\exists a(\cdot, \theta) \in L^2 / \forall x, x' \in \mathcal{X}, |\mathbf{f}(t, x, \theta) - \mathbf{f}(t, x', \theta)| \leq a(t, \theta) |x - x'|$ .

5 In that case, for each  $\theta \in \Theta$  and  $x_0 \in \mathcal{X}$ , we can guarantee that there exists a unique solution in  $H^1$  to  
6 the IVP

$$\begin{cases} \dot{x}(t) &= \mathbf{f}(t, x(t), \theta) \\ x(0) &= x_0 \end{cases} \quad (2.3)$$

7 As a generalized solution is in  $H^1$  and satisfies the ODE only almost everywhere, it motivates the  
8 use of weak derivative, that give a neat way to express the notion of *a.e* solution. Moreover, weak  
9 derivatives have been proved to be a very convenient way to solve differential equation models as it  
10 permits to introduce quite naturally variational formulation, in particular for Boundary Value Problems  
11 and Partial Differential Equations (a typical example is Galerkin approximation, [3]). If we denote the  
12 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  $g$   
13 in  $H^1$  is not defined point-wise but as the function  $\dot{g} \in L^2$  satisfying  $\langle \dot{g}, \psi \rangle = -\langle g, \dot{\psi} \rangle$ , for all function  
14  $\psi$  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  
15 on  $]0, 1[$ , the classical derivative  $\dot{\phi}$  is also the weak derivative.

16 We introduce then the weak form or variational formulation of the ODE (2.3). A weak solution  $g$  to  
17 (2.3) is a function in  $H^1$  such that  $\forall \varphi \in C_C^1(]0, 1[)$

$$\int_0^1 \mathbf{f}(t, g(t), \theta)\varphi(t)dt + \int_0^1 g(t)\dot{\varphi}(t)dt = 0 \quad (2.4)$$

18 In our case, a generalized solution is a weak solution and we know that there exists at least one weak  
19 solution to (2.4). For the statistical inference task, we use only the variational formulation as a necessary  
20 condition.

## 1 2.2 Definition

2 We define a new gradient matching estimator based on (2.4): starting from a nonparametric estimator  
 3  $\hat{\phi}$ , computed from the observations  $(t_i, y_i)$ ,  $i = 1, \dots, n$ , we want to find the parameter  $\theta$  that minimizes  
 4 the discrepancy between the parametric derivative  $t \mapsto \mathbf{f}(t, \hat{\phi}(t), \theta)$  and a nonparametric estimate of  
 5 the derivative, e.g.  $\dot{\hat{\phi}}$ . A classical discrepancy measure is the  $L^2$  distance, that gives rise to the two-step  
 6 estimator  $\hat{\theta}^{TS}$  defined as  $\hat{\theta}^{TS} = \arg \min_{\theta \in \Theta} R_{n,w}(\theta)$  where

$$R_{n,w}(\theta) = \int_0^1 |\dot{\hat{\phi}}(t) - \mathbf{f}(t, \hat{\phi}(t), \theta)|^2 w(t) dt. \quad (2.5)$$

7 This estimator is consistent for several usual nonparametric estimators [4, 12, 8], but the use of a  
 8 positive weight function  $w$  vanishing at the boundaries ( $w(0) = w(1) = 0$ ) is needed to get the classical  
 9 parametric root- $n$  rate. This approach has the computational advantage of decoupling each equation  
 10 and enables to have simple estimation procedures. For this reason and ease of readability, we consider  
 11 only the estimation of the parameter  $\theta_1$  when  $\mathbf{f}$  can be written  $\mathbf{f}(t, x, \theta) = (f_1(t, x, \theta_1), f_2(t, x, \theta_2))^\top$  and  
 12  $\theta = (\theta_1, \theta_2)^\top$  ( $\theta_i \in \mathbb{R}^{p_i}$  and  $p_1 + p_2 = p$ ). The joint estimation of  $\theta = (\theta_1, \theta_2)^\top$  can be done by stacking  
 13 the observations into a single column: there is no consequence on the asymptotics, but the estimator  
 14 covariance matrix has to be slightly modified in order to take into account the correlations between the  
 15 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  
 16 equation  $\dot{x}_1 = f(t, x, \theta)$ . We use a nonparametric estimator  $\hat{\phi} = (\hat{\phi}_1, \hat{\phi}_2)$  of  $\phi^* : [0, 1] \rightarrow \mathbb{R}^2$ .

17 Starting from (2.4), a reasonable estimator  $\hat{\theta}$  should satisfy the weak formulation

$$\forall \varphi \in C_C^1(]0, 1[), \int_0^1 f(t, \hat{\phi}(t), \hat{\theta}) \varphi(t) dt + \int_0^1 \hat{\phi}_1(t) \dot{\varphi}(t) dt = 0. \quad (2.6)$$

18 The vector space  $C_C^1(]0, 1[)$  is not tractable for variational formulation, and one prefers Hilbert space  
 19 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  
 20 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  
 21 we have

$$H_0^1 = \left\{ \sum_{\ell=1}^{\infty} a_\ell \sqrt{2} \sin(\ell\pi t) \mid \sum_{\ell=1}^{\infty} \ell^2 a_\ell^2 < \infty \right\} \quad (2.7)$$

1 Hence, it suffices to consider a countable number of orthogonal conditions (2.6) defined, for instance,  
 2 with the test functions  $\varphi_\ell = \sqrt{2} \sin(\ell\pi t)$ ,  $\forall \ell \geq 1$ :

$$\mathcal{C}_\ell(\theta) : \int_0^1 f(t, \hat{\phi}(t), \hat{\theta}) \varphi_\ell(t) dt + \int_0^1 \hat{\phi}(t) \dot{\varphi}_\ell(t) dt = 0. \quad (2.8)$$

3 More generally, we consider a family of orthonormal functions  $\varphi_\ell \in H_0^1$ , with  $\ell \geq 1$ , and we introduce  
 4 the vector space  $\mathcal{F} = \overline{\text{span}\{\varphi_\ell, \ell \geq 1\}}$ . The vector space  $\mathcal{F}$  may not be necessarily dense in  $H_0^1$ , as  
 5 the functions  $\varphi_\ell$  could be chosen for computational tractability or because of a natural interpretation  
 6 (for instance B-splines, polynomials, wavelets, ad-hoc functions, ...). For this reason, we introduce the  
 7 orthogonal decomposition of  $H_0^1 = \mathcal{F} \oplus \mathcal{F}^\perp$  where we can have  $\mathcal{F} \neq H_0^1$ . In general, an estimator  $\hat{\theta}$   
 8 satisfying  $\mathcal{C}_\ell(\hat{\theta})$  for  $\ell \geq 1$  also approximately satisfies (2.6). However in practice, we will use a finite set  
 9 of orthogonal constraints defined by  $L$  test functions ( $L > p$ ).

10 In order to discuss the influence of the choice of  $\mathcal{F}$  and of finite dimensional subspace spanned by  
 11  $\varphi_1, \dots, \varphi_L$  we introduce the nonlinear operator  $\mathcal{E} : (g, \theta) \mapsto \mathcal{E}(g, \theta)$ , such that  $t \mapsto \mathcal{E}(g, \theta)(t) =$   
 12  $f(t, g(t), \theta)$ .

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

$$\hat{\theta}_{n,L} = \arg \min_{\theta \in \Theta} Q_{n,L}(\theta). \quad (2.9)$$

17  $\hat{\theta}_{n,L}$  is the parameter that “almost” vanishes the first  $L$  Fourier coefficients in the orthogonal decompo-  
 18 sition of  $H_0^1 = \mathcal{F} \oplus \mathcal{F}^\perp$ :

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

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

20 The function  $\mathbf{E}_{\mathcal{F}}^\perp(\phi^*, \theta)$  represents the behavior of  $\mathcal{E}(g, \theta) - \dot{g}$  at the boundaries of the interval  
 21  $[0, 1]$ . As  $\hat{\phi}$  approaches  $\phi^*$  asymptotically in supremum norm, the objective function  $Q_{n,L}(\theta)$  is close  
 22 to  $Q_L^*(\theta) = \|\mathbf{E}_L(\phi^*, \theta)\|_{L^2}^2$ . The discriminative power of  $Q_L^*(\theta)$  can be analyzed locally around  $\hat{\theta}$ , as



1 it behaves approximately as the quadratic form  $Q_L^*(\theta) \approx (\theta - \theta^*)^\top \mathbf{J}_{\theta,L}^{*\top} \mathbf{J}_{\theta,L}^* (\theta - \theta^*)$  where  $\mathbf{J}_{\theta,L}^*$  is the  
 2 matrix in  $\mathbb{R}^{L \times p}$  with entries  $\int_0^1 f_{\theta_j}(t, \phi^*(t), \theta^*) \varphi_\ell(t) dt$ , for  $j = 1, \dots, p$ ,  $\ell = 1, \dots, L$ .

### 3 Consistency of the Orthogonal Conditions estimator

4 In order to obtain precise results with closed-form expression for the bias and variance estimators,  
 5 we consider series estimators, i.e. estimators expressed as  $\hat{\phi}_j = \sum_{k=1}^K \hat{c}_{k,j} p_{kK} = \hat{\mathbf{c}}_j \mathbf{p}^K$ , where  $\mathbf{p}^K =$   
 6  $(p_{1K}, \dots, p_{kK})$  is a vector of approximating functions and the coefficients  $\hat{\mathbf{c}}_j = (\hat{c}_{k,j})_{k=1..K}$  are computed  
 7 by least squares. For notational simplicity, we use the same functions (and the same number  $K$ ) for  
 8 estimating  $\phi_1^*$  and  $\phi_2^*$ . We denote  $P^K = (p_{kK}(t_i))_{1 \leq i, k \leq n, K}$  the design matrix and  $\mathbf{Y}_j = (y_{i,j})_{i=1..n}$  the  
 9 vectors of observations. Hence, the estimated coefficients  $\hat{\mathbf{c}}_j = (P^{K\top} P^K)^\dagger P^{K\top} \mathbf{Y}_j$  (where  $\dagger$  denotes a  
 10 generalized inverse) gives rise to the so-called hat matrix  $H = P^K (P^{K\top} P^K)^\dagger P^{K\top}$  and the vector of  
 11 smoothed observations is  $\hat{\phi}_j = H \mathbf{Y}_j$ ,  $j = 1, 2$ . One can typically think of regression splines, [22]. We  
 12 introduce now the conditions required for the definition and consistency of our estimator.

13 **Condition C1** (a)  $\Theta$  is a compact set of  $\mathbb{R}^p$  and  $\theta^*$  is an interior point of  $\Theta$ ,  $\mathcal{X}$  is an open subset of  
 14  $\mathbb{R}^2$ ; (b)  $(t, x) \mapsto f(t, x, \theta^*)$  is  $L^2$ -Lipschitz and  $L^2$ -Caratheodory.

15 **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  
 16 is a nonsingular constant matrix  $B$  such that for  $P^K = B_p^K(t)$ ; (i) the smallest eigenvalue of  
 17  $E [P^K(T) P^K(T)^\top]$  is bounded away from zero uniformly in  $K$  and (ii) there is a sequence of  
 18 constants  $\zeta_0(K)$  satisfying  $\sup_t |P^K(t)| \leq \zeta_0(K)$  and  $K = K(n)$  such that  $\zeta_0(K)^2 K/n \rightarrow 0$  as  
 19  $n \rightarrow \infty$ ; (c) There are  $\alpha, \mathbf{c}_{1,K}, \mathbf{c}_{2,K}$  such that  $\|\phi_j^* - p^K \mathbf{c}_{j,K}\|_\infty = \sup_{[0,1]} |\phi_j^*(t) - p^K(t)^\top \mathbf{c}_{j,K}| =$   
 20  $O(K^{-\alpha})$ .

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

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

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

2 **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  $(\phi^*, \theta^*)$ .

3 Condition **C1** gives the existence and uniqueness of a solution  $\phi^*$  in  $H^1$  to the IVP for  $\theta = \theta^*$  and  
4  $x(0) = \phi^*(0)$ . If  $f$  is continuous in  $t$  and  $x$ , then the derivative  $\dot{\phi}^*(t) = f(t, \phi^*(t), \theta^*)$  can be defined on  
5  $]0, 1[$  and is also continuous.

6  
7 Under condition **C2** (satisfied among others by regression splines with  $\zeta_0(K) = \sqrt{K}$ ), it is known  
8 that the series estimator  $\hat{\phi}_j$  are consistent estimators of  $\phi_j^*$  for usual norms, in particular  $\left\| \hat{\phi}_j - \phi_j^* \right\|_{\infty} =$   
9  $O_P\left(\zeta_0(K)\left(\sqrt{K/n} + K^{-\alpha}\right)\right)$  (theorem 1, [16]). If  $\phi^*$  is  $C^s$  and we use splines then  $\alpha = s$  and  $\left\| \hat{\phi} - \phi^* \right\|_{\infty} =$   
10  $O_P\left(K/\sqrt{n} + K^{1/2-s}\right)$ .

11  
12 Condition **C3** is here to control the continuity and regularity of the function  $\mathcal{E}$  involved in the inverse  
13 problem. Moreover, it provides uniform control needed for stochastic convergence.

14  
15 Condition **C4** is a sufficient condition for deriving independent conditions  $\mathcal{C}_\ell(\theta)$ , and normalization  
16 is useful only to avoid giving implicitly more weight to a condition w.r.t. the other conditions. A prin-  
17 cipled choice of weights and orthogonal functions will be discussed in section 4.4.

18  
19 Condition **C5** is standard in M-estimation [24], but can be hard to check in practice. Indeed, the  
20 parametric identifiability of ODE models can be hard to show, even for small systems [14]. Moreover,  
21 the natural criterion for estimating  $\theta$  and for identifiability analysis is

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

22 but  $\|\mathbf{E}_{\mathcal{F}}^{\perp}(\phi^*, \theta)\|_{L^2}^2$  is withdrawn and we use the quadratic form  $Q_{\mathcal{F}}^*(\theta)$  in order to avoid boundary effects.  
23 This is needed in order to get a parametric rate of convergence, as in the original two-step criterion (2.5).  
24 As a consequence, we lose a piece of information brought by the trajectory  $t \mapsto \phi^*(t)$  and we have to be  
25 sure that the parameter  $\theta$  has a low influence on  $\|\mathbf{E}_{\mathcal{F}}^{\perp}(\phi^*, \theta)\|_{L^2}^2$ . A favorable case is that it is almost

1 constant on  $\Theta$ , so that  $Q^*$  and  $Q_{\mathcal{F}}^*$  are essentially the same functions, with the same global minimum  
 2 and the same discriminating power.

3 Finally, Condition **C6** is about the influence of the truncation. We use only the first  $L$  Fourier coeffi-  
 4 cients of  $\mathcal{E}(g, \theta) - \dot{g}$  to identify the parameter  $\theta$ , but this might not be sufficient to discriminate between  
 5 two parameters  $\theta$  and  $\theta'$ . In a way, we perform dimension reduction but we need to be sure that we  
 6 have an exact recovery when  $L$  goes to infinity. More precisely, we expect that the minimum of  $|\mathbf{e}_L^*(\theta)|^2$   
 7 found under condition **C5** is also a minima of  $Q_{\mathcal{F}}^*(\theta) = \|\mathbf{E}_{\mathcal{F}}(\phi^*, \theta)\|_{L^2}^2$ . We have to be sure that  $\theta^*$  is a  
 8 global and isolated minima of  $Q_{\mathcal{F}}^*(\theta)$ . More generally, we can introduce the Jacobian matrices  $\mathbf{J}_{\theta, L}(g, \theta)$   
 9 in  $\mathbb{R}^{L \times p}$  with entries  $\int_0^1 f_{\theta_j}(t, g(t), \theta) \varphi_\ell(t) dt$  and  $\mathbf{J}_{x, L}(g, \theta)$  in  $\mathbb{R}^{L \times d}$  with entries  $\int_0^1 f_{x_i}(t, g(t), \theta) \varphi_\ell(t) dt$ .  
 10 For this reason, we suppose that  $\mathbf{J}_{\theta, L}^*$  is full rank, so that  $Q_{\mathcal{F}}^*(\theta)$  is locally strictly convex, with a unique  
 11 local minimum  $\theta^*$ .

12

13 **Theorem 1.** *If conditions **C1** to **C6** are satisfied, then*

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

14 *and the bias  $r_L$  tends to zero as  $L \rightarrow \infty$ . If we use the sine basis and if  $\mathcal{E}(\phi^*, \theta)$  is in  $H^1$  for all  $\theta$ , then*  
 15  *$r_L = o\left(\frac{1}{L}\right)$ .*

## 16 4 Asymptotics

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

## 1 4.1 Asymptotic representation for $\hat{\theta}_n - \theta$

2 From the definition (2.9) of  $\hat{\theta}_{n,L}$  and differentiability of  $f$ , the first order optimality condition is

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

3 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  $\theta_L^*$ . We  
 4 need to introduce the matrix-valued function defined on  $\mathcal{D} \times \theta$  such that  $\mathbf{M}_L(g, \theta) = \left[ \mathbf{J}_{\theta,L}(g, \theta)^\top \mathbf{J}_{\theta,L}(g, \theta) \right]^{-1}$   
 5  $\mathbf{J}_{\theta,L}(g, \theta)^\top$ , and proposition 2 shows that  $\mathbf{M}_L(\hat{\phi}, \hat{\theta}_{n,L})$  is also a consistent estimator of  $\mathbf{M}_L^*$ .

6 **Proposition 2.** *If conditions C1-C6 are satisfied, then*

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

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

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

## 8 4.2 Linear differential equations

9 We consider the parametrized linear ODE defined as

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

10 Since the ODE is linear, conditions **IVP(a)** and **IVP(b)** are satisfied as soon as the functions  $a(\cdot, \theta^*)$ ,  
 11  $b(\cdot, \theta^*)$ ,  $c(\cdot, \theta^*)$ ,  $d(\cdot, \theta^*)$  are in  $L^2$ , regardless of the initial conditions, [17]. We focus only on the estimation  
 12 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  
 13 we have two series estimators  $\hat{\phi}_1 = \mathbf{p}_K^\top \hat{\mathbf{c}}_1$  and  $\hat{\phi}_2 = \mathbf{p}_K^\top \hat{\mathbf{c}}_2$  satisfying condition **C2**. The orthogonal  
 14 conditions are simple linear functionals of the estimators  $e_\ell(\hat{\phi}, \theta) = \left\langle \hat{\phi}_1, \dot{\varphi}_\ell + a(\cdot, \theta)\varphi_\ell \right\rangle + \left\langle \hat{\phi}_2, b(\cdot, \theta)\varphi_\ell \right\rangle$ .  
 15 Hence the asymptotic behavior of the empirical orthogonal conditions relies on the plug-in properties  
 16 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  $\rho$  is a smooth function. Moreover, the

1 linearity of series estimator makes the orthogonal conditions  $\mathbf{e}_L(\hat{\phi}, \theta)$  easy to compute as

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

2 where  $\mathbf{A}(\theta)$  and  $\mathbf{B}(\theta)$  are matrices in  $\mathbb{R}^{L \times K}$  with entries  $A_{\ell,k}(\theta) = \int_0^1 (a(t, \theta)\varphi_\ell(t) + \dot{\varphi}_\ell(t)) p_{kK}(t) dt$   
3 and  $B_{\ell,k}(\theta) = \int_0^1 (b(t, \theta)\varphi_\ell(t)) p_{kK}(t) dt$ . The gradient of  $\mathbf{e}_L(\hat{\phi}, \theta)$  is  $\mathbf{J}_{\theta,L}(\hat{\phi}, \theta) = \partial_\theta \mathbf{A}(\theta)\hat{\mathbf{c}}_1 + \partial_\theta \mathbf{B}(\theta)\hat{\mathbf{c}}_2$   
4 where  $\partial_\theta \mathbf{A}(\theta)$  and  $\partial_\theta \mathbf{B}(\theta)$  are straightforwardly computed by permuting differentiation and integration.  
5 Although  $\mathbf{e}_L(\hat{\phi}, \theta)$  depends linearly on the observations, we have to take care of the asymptotics as we  
6 are in a nonparametric framework and  $K$  grows with  $n$ . The behavior of linear functionals  $T_\rho(\hat{\phi})$  for  
7 several nonparametric estimators (kernel regression, series estimators, orthogonal series) is well known  
8 [1, 2, 7, 16], and in generality it can be shown that such linear forms can be estimated with the classical  
9 root- $n$  rate and that they are asymptotically normal under quite general conditions. In the particular  
10 case of series estimators, we rely on theorem 3 of [16] that ensures the root- $n$  consistency and the  
11 asymptotic normality of the plugged-in estimators  $T_\rho(\hat{\phi}_j)$ ,  $j = 1, 2$  under almost minimal conditions.  
12 We will give in the next section the precise assumptions required for root- $n$  consistency of linear and  
13 nonlinear functional of the series estimator. Moreover, the variance of  $\hat{\theta}_{n,L}$  has a remarkable expression

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

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

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

17 Thanks to proposition 2, we can estimate the asymptotic variance of the estimator  $\hat{\theta}_{n,L}$  with the con-  
18 sistent estimator  $\hat{\mathbf{M}}_L = \mathbf{M}_L(\hat{\phi}, \hat{\theta}_{n,L})$  and we estimate  $V(\hat{\theta}_{n,L})$  by  $\widehat{V(\hat{\theta}_{n,L})} = \hat{\mathbf{M}}_L V\left(\mathbf{e}_L(\hat{\phi}, \hat{\theta}_{n,L})\right)\hat{\mathbf{M}}_L^\top$ .  
19 From the asymptotic normality of the plug-in estimate, we can derive confidence balls or confidence

1 with level  $1 - \alpha$ . For instance, for each parameter  $\theta_i$ ,  $i = 1, \dots, p$ :

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

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

4

### 5 4.3 Nonlinear differential equations

6 We give here general results for the asymptotics of  $e_\ell(\hat{\phi}, \theta)$  when the functional is linear or not in  $\hat{\phi}$ .  
 7 In [16], the root- $n$  consistency and asymptotic normality is obtained if the functional  $g \mapsto e_\ell(g, \theta)$  has  
 8 a continuous Fréchet derivative  $De_\ell(g, \theta)$  with respect to the norm  $\|\cdot\|_\infty$ . If  $x \mapsto f(t, x, \theta)$  is twice  
 9 continuously differentiable for  $t \in [0, 1]$  a.e. in  $x$  and  $\theta$  in  $\Theta$ , then we can compute easily its Fréchet  
 10 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  
 11 have

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

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

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

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

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

1 with  $C$ , a constant independent of  $\theta$ ,  $\epsilon$  and  $g, g'$  (because  $f_{xx}$  is uniformly dominated).

2 As in the linear case, we need to evaluate  $De_\ell(g, \theta)$  on the basis  $\mathbf{p}^K$ . We denote  $\mathbf{A}(g, \theta)$  and  $\mathbf{B}(g, \theta)$   
 3 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  
 4 have the approximation

$$\mathbf{e}_L(\hat{\phi}, \theta) = \mathbf{e}_L(\phi^*, \theta) + \mathbf{A}(\phi^*, \theta) \hat{\mathbf{c}}_1 + \mathbf{B}(\phi^*, \theta) \hat{\mathbf{c}}_2 + O(\|h\|_\infty^2). \quad (4.8)$$

5 We can derive the asymptotic variance of  $\mathbf{e}_L(\hat{\phi}, \theta)$  from (4.8)

$$V_{e,L}(\theta) = \mathbf{A}(\phi^*, \theta) V(\hat{\mathbf{c}}_1) \mathbf{A}(\phi^*, \theta)^\top + \mathbf{B}(\phi^*, \theta) V(\hat{\mathbf{c}}_2) \mathbf{B}(\phi^*, \theta)^\top \quad (4.9)$$

6 and we can get an estimate  $\widehat{V}_{e,L}(\theta)$  from the data as in the linear case. We need the two additional  
 7 conditions for deriving the root- $n$  rate, that comes from [16] (Condition C8 is essential for root- $n$  rate):

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

10 **Condition C8** There exists  $\tilde{\beta}_K$  in  $\mathbb{R}^K$  with  $E_T \left[ \left| f_x(T, \phi^*(T), \theta) \varphi_\ell(T) - \tilde{\beta}_K^\top p^K(T) \right|^2 \right] \rightarrow 0$ .

11 **Theorem 3.** Under conditions **C1** to **C8** and if  $f$  is a linear ODE or,  $f$  is a nonlinear ODE and  
 12  $\frac{\zeta_0(K)^4 K^2}{n} \rightarrow 0$  then for all  $\theta \in \Theta$ ,  $\sqrt{n} \left( \mathbf{e}_L(\hat{\phi}, \theta) - \mathbf{e}_L(\phi^*, \theta) \right) \rightsquigarrow N(0, V_{e,L}(\theta))$  with  $V_{e,L}(\theta)$  given by  
 13 (4.9), and we denote  $\mathbf{V}_{e,L}^* = V_{e,L}(\theta^*)$ . Moreover,  $\hat{\theta}_{n,L}$  is such that

$$\sqrt{n} \left( \hat{\theta}_{n,L} - \theta_L^* \right) \rightsquigarrow N(0, \mathbf{V}_L^*) \quad (4.10)$$

14 with

$$\mathbf{V}_L^* = \mathbf{M}_L^* \mathbf{V}_{e,L}^* \mathbf{M}_L^{*\top}. \quad (4.11)$$

15 The asymptotic variance can be estimated as  $\hat{\mathbf{M}}_L V_{e,L}(\hat{\theta}_{n,L}) \hat{\mathbf{M}}_L^\top \xrightarrow{P} \mathbf{V}_L^*$ . In particular, if we use re-  
 16 gression splines and  $t \mapsto f(t, \phi^*(t), \theta)$  is  $C^s$  on  $[0, 1]$  with  $s \geq 3$ , then (4.10) holds with  $K$  such that  
 17  $\sqrt{n} K^{-s} \rightarrow 0$  and  $n^{-1} K^4 \rightarrow 0$ .

1 Theorem 3 is a direct application of theorem 3 in [16] that claims the root- $n$  consistency and asymptotic  
 2 normality of general (nonlinear) plug-in estimators.

### 3 4.4 From optimal optimal weighting to a practical estimation algorithm

4 Theorem (3) is of practical interest as it provides a closed-form expression for the asymptotic variance  
 5 of  $\hat{\theta}_{n,L}$  for general ODE. We can make a parallel between the orthogonal condition estimator and the  
 6 weighted nonlinear least-squares:

$$\hat{\theta}_c^{WLS} = \arg \min \sum_{i=1}^n w(t_i) |y_i - \phi(t_i, \theta_c)|^2$$

7 where  $w(\cdot)$  is a (positive) weight function and  $\theta_c = (x_0, \theta)$ .  $\hat{\theta}_c^{WLS}$  is consistent and  $\sqrt{n} \left( \hat{\theta}_c^{WLS} - \theta_c^* \right) \rightsquigarrow$   
 8  $N(0, \mathbf{V}^{AWLS})$  under classical regularity assumptions on  $\mathbf{f}$ , see [19]. The asymptotic variance  $\mathbf{V}^{AWLS}$  is  
 9 directly computed from the sensitivity equations, and the optimal weight function  $w(\cdot)$  is proportional  
 10 to the variance function  $\sigma^2(\cdot)$ , meaning that the unweighted least-squares estimator is optimal in the  
 11 homoscedastic case. It is clear that  $\mathbf{V}_L^* \neq \mathbf{V}^{AWLS}$ , which means that we are sure that the orthogonal  
 12 condition estimator is not an efficient estimator (at least in the Gaussian case, because NLS and MLE  
 13 are the same in that case). A striking difference between  $\mathbf{V}_L^*$  and  $\mathbf{V}^{AWLS}$  is that the least squares  
 14 involves the Jacobian of the solution w.r.t. the initial values and parameters, whereas the orthogonal  
 15 conditions involve the Jacobian of the vector field. It is then hard to compare these two matrices in  
 16 generality and to evaluate the loss of efficiency. Nevertheless, we can compare the influence of a weight  
 17 matrix for the minimization of the orthogonal conditions. Indeed, if we introduce a positive definite  
 18 matrix  $W$  in  $\mathbb{R}^{L \times L}$ , we can define the weighted criterion

$$Q_{n,L}^W(\theta) = \mathbf{e}_L(\hat{\phi}, \theta)^\top W \mathbf{e}_L(\hat{\phi}, \theta)$$

19 and the corresponding estimator

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



1 The results of theorems 1 and 3 are then still true under straightforward adaptations. In particular,  
2  $\hat{\theta}_{n,L}^W$  is consistent and asymptotically normal with asymptotic variance  $\mathbf{M}_{W,L}^* \mathbf{V}_{e,L}^* \mathbf{M}_{W,L}^{*\top}$  where  $\mathbf{M}_{W,L}^* =$   
3  $[\mathbf{J}_L^{*\top} W \mathbf{J}_L^*]^{-1} \mathbf{J}_L^{*\top} W$ . Eventually, we can then ask for the best weighting matrix  $W$  giving the smallest  
4 asymptotic variance. This is a classical result for Generalized Moment Estimators that we recall in the  
5 following proposition (see for instance section 3.6 in [9]):

6 **Proposition 4. Optimal weighting matrix**

7 *The minimal asymptotic variance for  $\hat{\theta}_{n,L}^W$  is obtained with  $W^{opt} = \mathbf{V}_{e,L}^{*-1}$  and*

$$\mathbf{V}_L^{opt}(\theta^*) = (\mathbf{J}_{\theta,L}^{*\top} \mathbf{V}_{e,L}^{*-1} \mathbf{J}_{\theta,L}^*)^{-1} \quad (4.12)$$

8 Even if we have an homoscedastic model, we have an interest in using a weighted estimator. Inter-  
9 estingly, the problem of choosing the best weighting matrix is directly related to the choice of the best  
10 set of test functions  $\varphi_1, \dots, \varphi_L$ . Indeed, the diagonalization of  $\mathbf{V}_{e,L}^* = \mathbf{U} \Lambda \mathbf{U}^\top$  permits the introduction  
11 of the eigenvalues  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_L)$  (with  $\lambda_1 \geq \dots \geq \lambda_L \geq 0$ ) and  $U = (\mathbf{u}_1 | \dots | \mathbf{u}_L)$  is an orthogonal  
12 matrix. If  $\Lambda$  is nonsingular,  $Q_{n,L}^{opt}(\theta) \doteq Q_{n,L}^{W^{opt}}(\theta) = \widetilde{\mathbf{e}}_L^\top(\hat{\phi}, \theta) \Lambda^{-1} \widetilde{\mathbf{e}}_L(\hat{\phi}, \theta)$ , with  $\widetilde{\mathbf{e}}_L(\hat{\phi}, \theta) = \mathbf{U}^\top \mathbf{e}_L(\hat{\phi}, \theta)$ .  
13 By linearity, the new orthogonal conditions  $\widetilde{\mathbf{e}}_L(\hat{\phi}, \theta)$  can be written as orthogonal conditions with the  
14 test functions  $\psi_1, \dots, \psi_L$  derived from the eigenvectors of  $\mathbf{V}_{e,L}^*$  as  $\psi_\ell = \sum_{k=1}^L u_{k\ell} \varphi_k$ . The use of the opti-  
15 mal weighting matrix is then equivalent to choose the best test functions in  $\mathcal{F}_L$  and the diagonalization  
16 show that some care must be taken (typically  $L$  have to be reasonably small to avoid  $\lambda_L \approx 0$ ).

17 Finally, as the optimal weight  $W^{opt}$  depends on the true unknown parameter, the practical use of  
18 the previous analysis is hard to apply. We suggest to approximate  $\mathbf{V}_{e,L}^{*-1}$  by computing a sequence of  
19 weighted estimates, as described in **Algorithm 1**.

20 This is known as iterated Generalized Method of Moments (GMM), and consists in successive re-  
21 weighting, by using the consistent estimator of the variance  $\widehat{V_{e,L}(\hat{\theta}_{n,L})}$ , [9]. There is no theoretical  
22 guarantee for the (numerical) convergence of this algorithm but if the model is correct, this sequence of  
23 iterations tends to ameliorate the quality of the GMM estimator. Possibly, during this stage, it could  
24 be necessary to select appropriately  $L$ , as  $\widehat{V_{e,L}(\hat{\theta}_{n,L}^{(k)})}$  can be close to be nonsingular at an iteration  $k$ .

---

**Algorithm 1** Iteratively Reweighted Orthogonal Conditions

---

**Require:**  $\hat{\phi}, \theta_{init} \in \Theta, \epsilon > 0,$

Compute the unweighted estimator

$$\hat{\theta}_{n,L}^{(0)} = \arg \min_{\theta} Q_{n,L}(\theta).$$

Compute the asymptotic covariance of  $\mathbf{e}_L(\hat{\phi}, \hat{\theta}_{n,L}^{(0)})$  with equation (4.7).

**while**  $|\hat{\theta}_{n,L}^{(k)} - \hat{\theta}_{n,L}^{(k-1)}| > \epsilon$  **do**

$$\begin{cases} W^{(k)} &= \left( V_{e,L}(\hat{\theta}_{n,L}^{(k)}) \right)^{-1} \\ \hat{\theta}_{n,L}^{(k+1)} &= \arg \min_{\theta} Q_{n,L}^{W^{(k)}}(\theta) \end{cases}. \quad (4.13)$$

**end while**

---

## 1 5 Experiments

2 We apply the Orthogonal Condition (OC) estimators to first order ODEs linear and nonlinear in state.  
3 We consider situations where the Orthogonal Conditions estimator has a closed-form, or when the  
4 classical NLS estimator or Two-Step estimator may collapse. We compare systematically the NLS  
5 estimator  $\hat{\theta}^{NLS}$  and Two-Step Estimator (TS)  $\hat{\theta}^{TS}$  to the OC estimator  $\hat{\theta}_{n,L}$  and the weighted OC  
6 estimator  $\hat{\theta}_{n,L}^W$  for varying sample sizes ( $n = 400, 200, 50$ ) and varying noise level (high and small). This  
7 gives a reasonable picture of the robustness, sensitivity and efficiency of the different estimators. For  
8 the NLS estimator, we compute a one-step ameliorated estimator by using  $\hat{\theta}_{n,L}^W$  as a starting point. The  
9 estimator obtained should be efficient and should give an indication about the best attainable variance,  
10 and the efficiency loss of the Gradient Matching methods.

11 Indeed, we consider a Gaussian and homoscedastic noise, such that the NLS should be nearly effi-  
12 cient. This optimality is asymptotic and the case  $n = 200$  or  $n = 400$  should give indications on the  
13 efficiency loss, whereas the small size case ( $n = 50$ ) gives some relevant information on the complex  
14 small sample case, where the asymptotic approximation cannot be assessed. In particular, the NLS does  
15 not necessarily provide the best estimator.

16 The nonparametric estimate for the gradient matching estimators (TS and OC) is the same regression  
17 spline, decomposed on B-splines with a uniform knot sequence  $\xi_k, k = 1, \dots, K$ . For each dataset (and  
18 each dimension), the number of knots is selected by minimizing the GCV criterion, [22]. For the plain

1 TS estimator, we use the same weight function as in [4], and we consider two variations for the OC  
 2 estimator depending on the weight matrix: a uniform weight matrix  $W = I_L$ , and the optimal weighting  
 3 matrix  $W^{opt}$  computed as described in Algorithm 1. The Orthogonal Conditions are defined with the  
 4 sine basis and the default number of conditions is  $L = 2 \times p \times d$ .

5 Finally, we compute the bias and variance of the different estimators, the confidence intervals and  
 6 their coverage probabilities by Monte Carlo simulations (based on  $N_{MC} = 100$  independent draws).

## 7 5.1 Linear ODE

8 The classical linear constant ODE can be written as  $\dot{x} = \mathbf{A}x + \mathbf{h}(t)$  with constant coefficients  $\mathbf{A}$  with  
 9  $\mathbf{A}^\top = (A_1 | \dots | A_d)$  and  $\mathbf{h} = (h_1(t), \dots, h_d(t))$ . For each  $i$ ,  $\dot{x}_i = \sum_{k=1}^d a_{ik}x_k + h_i = A_i^\top x + h_i$ . The weak  
 10 form gives rise to a linear matrix equality for  $i = 1, \dots, d$  as  $\mathbf{Y}_i^\varphi = \mathbf{X}^\varphi A_i + \mathbf{H}_i^\varphi$ : the matrix involved is  
 11  $\mathbf{X}^\varphi$  is  $d \times L$  with entries  $\langle x_k, \varphi_\ell \rangle$ ,  $\mathbf{Y}_i^\varphi$  and the vectors  $\mathbf{H}_i^\varphi$  are in  $\mathbb{R}^L$  with respectively entries equal to  
 12  $-\langle x_i, \dot{\varphi}_\ell \rangle$  and  $\langle h_i, \varphi_\ell \rangle$ . For illustration, we consider the  $\alpha$ -pinene (biochemical) network used in [15] for  
 13 the comparison of several global optimization algorithms:

$$\left\{ \begin{array}{l} \dot{x}_1 = -(\theta_1 + \theta_2)x_1 \\ \dot{x}_2 = \theta_1 x_1 \\ \dot{x}_3 = \theta_2 x_1 - (\theta_3 + \theta_4)x_3 + \theta_6 x_5 \\ \dot{x}_4 = \theta_3 x_3 \\ \dot{x}_5 = \theta_4 x_3 + \theta_5 x_5 \end{array} \right. \quad (5.1)$$

14 The initial condition is  $x_0 = (100, 0, 0, 0, 0)$  and is considered unknown. In that case, the system  
 15 converges to a stationary point. The global quality can be analyzed from the Mean Square Errors in  
 16 table (5.1): with large sample case and small noise the estimators are equivalent (with error around  
 17 0.1), but in the small sample case ( $n = 50$ ), the best estimators are TS and OCOpt, and the worst is  
 18 NLS. NLS and OC are roughly equivalent, and from this example, it is clear that there is a benefit in  
 19 using relevant weight  $W$  in the OC conditions. Despite the truncation, the optimal weighting enables to  
 20 recover some information with respect to TS, as it identifies the most relevant directions (conditions) to  
 21 use for measuring the model discrepancy. We do not detail the mean and variance for each parameter,

$(n; \sigma)$	$\hat{\theta}_{TS}$	$\hat{\theta}_{OC}$	$\hat{\theta}_{OCopt}$	$\hat{\theta}_{NLS}$
(400; 50)	0.1055	0.0851	0.0907	0.1171
(400; 100)	0.1325	0.2170	0.1372	0.2443
(200; 50)	0.1139	0.2021	0.1183	0.1964
(200; 100)	0.1655	0.3553	0.2615	0.3515
(50; 50)	0.2034	0.5684	0.3698	0.4494
(50; 100)	0.5289	0.7130	0.5473	0.8042

Table 5.1: Mean Square Error for  $\theta$  for Alpha-Pinene ODE (5.1)

$(n; \sigma)$	$\theta_1$			$\theta_2$			$\theta_3$		
	OC	OCopt	NLS	OC	OCopt	NLS	OC	OCopt	NLS
(400; 50)	0.9600	0.9700	0.9588	0.9900	1.0000	0.9278	0.9600	0.9600	0.9691
(400; 100)	0.9700	0.9796	0.9072	0.9800	0.9898	0.8866	0.9700	0.9694	0.9381
(200; 50)	0.9800	0.9800	0.8866	1.0000	1.0000	0.9175	0.9600	0.9600	0.9278
(200; 100)	0.9381	0.9579	0.9333	0.9897	0.9789	0.8889	0.9691	0.9684	0.9444
(50; 50)	0.9474	0.9579	0.8795	0.9684	0.9789	0.8434	0.9579	0.9579	0.9398
(50; 100)	0.8681	0.9176	0.7571	0.9341	0.9647	0.8286	0.8462	0.8824	0.8143

Table 5.2: Coverage Probability of 95% Asymptotic Confidence Interval for Alpha-Pinene ODE (5.1)

- 1 but we consider only the coverage probabilities of the the confidence intervals given in tables (5.2, 5.3).
- 2 We see that the asymptotic approximation obtained for the parameters of is really good for  $\theta_1, \theta_2, \theta_3,$
- 3 and are degraded for  $\theta_4, \theta_5, \theta_6$ . Finally, we see that the behavior of OCopt is similar to NLS, and better
- 4 than simple OC.

5

6

$(n; \sigma)$	$\theta_4$			$\theta_5$			$\theta_6$		
	OC	OCopt	NLS	OC	OCopt	NLS	OC	OCopt	NLS
(400; 50)	0.7800	0.8400	0.9278	0.7700	0.8400	0.9072	0.8200	0.8500	0.9175
(400; 100)	0.6300	0.6939	0.8763	0.6400	0.7245	0.8866	0.6600	0.7449	0.8763
(200; 50)	0.6800	0.7500	0.9278	0.6700	0.7500	0.9278	0.7000	0.7800	0.9278
(200; 100)	0.5464	0.6526	0.8889	0.5773	0.6842	0.8889	0.7113	0.8000	0.8778
(50; 50)	0.5474	0.6737	0.8795	0.7263	0.8632	0.8795	0.7158	0.7895	0.9036
(50; 100)	0.6154	0.7294	0.7714	0.8352	0.9765	0.8143	0.6813	0.7412	0.7857

Table 5.3: Coverage Probability of 95% Asymptotic Confidence Interval for Alpha-Pinene ODE (5.1)

## 1 5.2 Nonlinear ODEs

### 2 5.2.1 Ricatti equation

3 Ricatti ODE is a family of quadratic ODE that arises commonly in Control Theory. In this example,  
4 we consider the simple constant coefficient Ricatti equation

$$\dot{x} = ax^2 + b\sqrt{t} \quad (5.2)$$

5 with  $a = 0.12$ ,  $b = 0.2$  and  $x_0 = -1$ , for  $t \in [0, 5]$ . Because of the squared term  $x^2$ , the solution can  
6 explode on  $[0, 5]$ , typically if  $a$  is too big. This is a problem for NLS, because the parameter optimization  
7 can give rise to irrelevant parameter candidate with diverging solution before  $T = 5$ . This problem does  
8 not appear for OC estimator, because we use only the weak form:  $\forall \varphi \in C_C^1([0, 1])$ ,

$$\langle x, \dot{\varphi} \rangle + a \langle x^2, \varphi \rangle + b \langle \sqrt{t}, \varphi \rangle = 0. \quad (5.3)$$

9 As the Ricatti equation is linear in the parameter, the OC estimator is computed by solving a similar  
10 linear system as in the previous linear example. The essential difference is that the matrix  $\mathbf{X}^\varphi$  involves  
11 nonlinear functionals of  $\hat{\phi}$ . The Mean Square Errors show that NLS and TS provide better estimation  
12 than OC and OC opt. For all the estimators, the bias is similar (and  $b$  is notably harder to estimate  
13 than  $a$ ), but the estimated asymptotic variance is smaller for NLS, and is smaller than for OC and OC  
14 opt. As a consequence, the coverage probability are better for OC and OCopt than NLS (in particular  
15 for  $b$ , see tables 5.6, 5.7). This is classical short-come of NLS to give overoptimistic variance estimators.

$(n; \sigma)$	$\hat{a}_{TS}$	$\hat{a}_{OC}$	$\hat{a}_{OCopt}$	$\hat{a}_{NLS}$
(400; 0.2)	0.0007	0.0061	0.0060	0.0007
(400; 0.4)	0.0028	0.0119	0.0115	0.0029
(200; 0.2)	0.0015	0.0105	0.0106	0.0016
(200; 0.4)	0.0047	0.1215	0.0646	0.0055
(50; 0.2)	0.0062	0.0350	0.0349	0.0077
(50; 0.4)	0.0213	0.1755	0.0831	0.0310

Table 5.4: Mean Square Error estimated by Monte Carlo for  $a^* = 0.12$  for Ricatti ODE (5.2)

$(n; \sigma)$	$\hat{b}_{TS}$	$\hat{b}_{OC}$	$\hat{b}_{OCopt}$	$\hat{b}_{NLS}$
(400; 0.2)	0.0000	0.0001	0.0001	0.0000
(400; 0.4)	0.0000	0.0001	0.0001	0.0000
(200; 0.2)	0.0000	0.0001	0.0001	0.0000
(200; 0.4)	0.0001	0.0011	0.0007	0.0001
(50; 0.2)	0.0001	0.0003	0.0003	0.0001
(50; 0.4)	0.0003	0.0025	0.0011	0.0004

Table 5.5: Mean Square Error estimated by Monte Carlo for  $b^* = 0.2$  for Ricatti ODE (5.2)

1

$(n; \sigma)$	Estimated Asymptotic Variance			Coverage Probability (95%)		
	$\hat{V}(\hat{a}_{OC})$	$\hat{V}(\hat{a}_{OCopt})$	$\hat{V}(\hat{a}_{NLS})$	OC	OC opt	NLS
(400; 0.2)	0.0047	0.0047	0.0007	0.95	0.93	0.89
(400; 0.4)	0.0147	0.0145	0.0028	0.95	0.97	0.90
(200; 0.2)	0.0120	0.0120	0.0014	0.94	0.93	0.80
(200; 0.4)	0.6987	0.0797	0.0054	0.93	0.93	0.90
(50; 0.2)	0.0658	0.2945	0.0150	0.89	0.87	0.80
(50; 0.4)	2.9971	0.1584	0.0231	0.89	0.84	0.85

Table 5.6: Estimated Variance and Coverage Probabilities for  $a$ , for Ricatti ODE (5.2) with  $a^* = 0.12$ .

2

$(n; \sigma)$	Estimated Asymptotic Variance			Coverage Probability (95%)		
	$\hat{V}(\hat{b}_{OC})$	$\hat{V}(\hat{b}_{OCopt})$	$\hat{V}(\hat{b}_{NLS})$	OC	OC opt	NLS
(400; 0.2)	0.0001	0.0001	0.0000	0.96	0.97	0.82
(400; 0.4)	0.0002	0.0002	0.0000	0.97	0.98	0.89
(200; 0.2)	0.0001	0.0001	0.0000	0.97	0.97	0.84
(200; 0.4)	0.0058	0.0009	0.0001	0.97	0.97	0.89
(50; 0.2)	0.0006	0.0032	0.0106	0.96	0.96	0.82
(50; 0.4)	0.0363	0.0020	0.0004	0.95	0.95	0.88

Table 5.7: Estimated Variance and Coverage Probabilities for  $a$ , for ODE (5.2) with  $b^* = 0.2$

3

#### 4 5.2.2 FitzHugh-Nagumo

5 The FitzHugh-Nagumo is a nonlinear two-dimensional ODE introduced for modeling neurons. For well-  
6 chosen sets of parameters and initial conditions, it exhibits a periodic behavior, with typical oscillations

1 corresponding to a limit cycle

$$\begin{cases} \dot{V} = c \left( V - \frac{V^3}{3} + R \right) \\ \dot{R} = -\frac{1}{c} (V - a + bR) \end{cases} \quad (5.4)$$

2 The true parameters are  $a^* = b^* = 0.2$  and  $c^* = 3$  and  $x_0 = (V_0, R_0) = (-1, 1)$ , and are taken from  
 3 [21] where it was introduced as a benchmark for parameter estimation in ODEs. The noise levels we  
 4 consider are  $\sigma = 0.15, 0.3$ . In all the settings we consider (large or small sample size), the NLS provides  
 5 the smallest MSE, and it is clear that OC improves on TS, and that the optimal weight provides an  
 6 even better estimator OCopt, see table 5.8.

7 The detailed analysis of the bias show that the NLS have the smallest bias, and that the parameter  
 8  $c$  is the hardest to estimate. The most essential feature is that in most of the case, the asymptotic  
 9 variance computed by the NLS is smaller than for OC, which gives significant differences between the  
 10 coverage probabilities for  $a, b, c$  in particular in the small sample case. The confidence sets obtained  
 11 by OC are rather credible, with coverage probability around 90% in almost any case (except in table  
 12 5.11). As the choice of the optimal weighted estimator tends to reduce the variance, we get narrower  
 13 confidence sets with lower coverage probability.

$\times 10^{-3}$	$a$				$b$				$c$			
$(n; \sigma)$	TS	OC	OCopt	NLS	TS	OC	OCopt	NLS	TS	OC	OCopt	NLS
(400; 0.15)	0.1	0.1	0.1	0.0	0.8	3.2	1.3	0.5	32.4	13.2	4.7	0.2
(400; 0.3)	0.3	0.6	0.4	0.1	4.1	23.5	14.2	1.6	236.1	41.6	18.2	0.3
(200; 0.15)	0.2	0.4	0.2	0.0	2.0	23.7	2.4	0.9	110.4	29.2	9.1	0.5
(200; 0.3)	0.6	1.0	0.6	0.2	5.8	26.4	9.1	3.7	519.1	82.0	34.0	1.8
(50; 0.15)	0.3	0.6	0.3	0.1	3.1	36.7	9.0	1.6	228.4	97.1	35.1	1.3
(50; 0.3)	1.3	1.6	1.0	0.3	12.4	51.1	44.1	7.4	893.0	796.9	404.8	3.8

Table 5.8: Mean Square Error computed by Monte Carlo for FitzHugh-Nagumo ODE (5.4).

14

15

16

17

$(n; \sigma)$	Estimated Asymptotic Variance			Coverage Probability		
	$\hat{V}(\hat{a}_{OC})$	$\hat{V}(\hat{a}_{OCopt})$	$\hat{V}(\hat{a}_{NLS})$	OC	OC opt	NLS
(400; 0.15)	0.0001	0.0001	0.0000	0.9600	0.9200	0.7600
(400; 0.3)	0.0011	0.0002	0.0000	0.9394	0.8750	0.7292
(200; 0.15)	0.0004	0.0001	0.0000	0.8800	0.8163	0.7551
(200; 0.3)	0.0063	0.0004	0.0001	0.9375	0.8780	0.6098
(50; 0.15)	0.0577	0.0002	0.0000	0.9400	0.8876	0.6180
(50; 0.3)	0.7871	0.0008	0.0002	0.9043	0.9143	0.6377

Table 5.9: Estimated Variance and Coverage Probabilities for  $a$ , for FitzHugh-Nagumo ODE (5.4)

$(n; \sigma)$	Estimated Asymptotic Variance			Coverage Probability		
	$\hat{V}(\hat{b}_{OC})$	$\hat{V}(\hat{b}_{OCopt})$	$\hat{V}(\hat{b}_{NLS})$	OC	OC opt	NLS
(400; 0.15)	0.0039	0.0007	0.0002	0.9500	0.9000	0.6400
(400; 0.3)	0.1516	0.0058	0.0008	0.8889	0.8333	0.5521
(200; 0.15)	0.0568	0.0018	0.0003	0.9400	0.8980	0.5408
(200; 0.3)	2.6567	0.0088	0.0012	0.8958	0.9146	0.5000
(50; 0.15)	39.2617	0.0048	0.0006	0.9000	0.8315	0.6180
(50; 0.3)	116.2820	0.0299	0.0024	0.8936	0.8857	0.4928

Table 5.10: Estimated Variance and Coverage Probabilities for  $b$ , for FitzHugh-Nagumo ODE (5.4)

## 6 Discussion

The experiments done in the previous section show that the Orthogonal Conditions estimator is a reliable estimator that can compete with nonlinear least squares. Indeed, the OC estimators are used as starting point for NLS, so that the NLS improved estimator must be better than OC. When we consider the MSE, it is often the case (except for alpha-pinene), which justifies the common practice of using Gradient Matching for determining interesting initial values for NLS. Nevertheless, one can note that this improvement is rather limited with respect to OC (or OCopt). In particular, this shows that the projection and the use of limited number  $L$  of test functions  $\varphi_\ell$  is sufficient for estimating parameters, and that the bias  $r_L$  introduced in theorem 1 is in fact quite small. When performing simulations, it is not easy to identify the influence of the quality of the nonparametric proxy in the global quality of a Gradient Matching Estimator (typically OC), in particular when the sample size is small. Hence, some additional work should be done for the construction of good (or the best) nonparametric proxy to use in a Gradient Matching approach in order to control the influence of this step in estimation.

Moreover, the asymptotic variance and the Gaussian approximation have proved to be effective



$(n; \sigma)$	Estimated Asymptotic Variance			Coverage Probability		
	$\hat{V}(\hat{c}_{OC})$	$\hat{V}(\hat{c}_{OCopt})$	$\hat{V}(\hat{c}_{NLS})$	OC	OC opt	NLS
(400; 0.15)	0.0122	0.0012	0.0000	0.9300	0.7300	0.6100
(400; 0.3)	0.0827	0.0058	0.0002	0.9495	0.7500	0.7083
(200; 0.15)	0.0401	0.0024	0.0001	0.9500	0.6531	0.6735
(200; 0.3)	0.5445	0.0122	0.0003	0.7812	0.7195	0.5488
(50; 0.15)	15.8363	0.0060	0.0002	0.8300	0.5730	0.6067
(50; 0.3)	3.1898	0.1295	0.0008	0.6809	0.8000	0.6667

Table 5.11: Estimated Variance and Coverage Probabilities for  $c$ , for FitzHugh-Nagumo (5.4)

1 for quantifying the uncertainty about the OC parameters, through classical confidence which is often  
2 delicate for Gradient Matching estimators. From the simulation, it is clear that these estimation are over-  
3 optimistic, which gives too narrow confidence sets: this is a classical problem for nonlinear regression,  
4 where NLS tends also to give unreliable small confidence sets. As the variance  $\mathbf{V}_L^*$  is a bit bigger,  
5 we can have wider and possibly more relevant confidence sets. Conversely, the OC estimator does not  
6 use errors-in-variables technics as in [12], that can provide huge variances and confidence sets. Hence,  
7 the question of the “best” set estimation in ODEs is still hard to close, as the classical linearization  
8 approaches used for asymptotic approximation can be limited in Differential Equations.

9 It should be noticed that the use of an optimal weight in OC can significantly improve the estimator  
10 (by reducing the bias and the variance); but some care must be taken in the iterations as the estimated  
11 variance  $\mathbf{V}(e_L)$  can be singular. A simple device to avoid numerical divergence is to reduce the number  
12 of orthogonal conditions.

13 We have shown that there is a real interest in defining more elaborated criterion in Gradient Matching  
14 estimators for statistical tractability. In particular, variational characterization is quite new in statis-  
15 tical estimation, and gives computationally and statistically relevant estimators with a broad field of  
16 applicability that still needs to be explored.

## 17 Acknowledgments

18 This work was funded by two ANR projects GD2S (ANR-05-MMSA-0013-01) and ODESSA (ANR-  
19 09-SYSC-009-01) and received a partial support from the Analysis of Object Oriented Data program  
20 (2010-2011) in Statistical and Applied Mathematical Sciences Institute (SAMSI), USA.

# 1 A Proofs

## 2 Theorem 1

3 The classical proof for the consistency of an  $M$ -estimator  $\hat{\theta}_{n,L} = \arg \min_{\theta \in \Theta} Q_{n,L}(\theta)$  such as the orthog-  
 4 onal conditions estimator relies basically on two stages: the uniform convergence of  $Q_{n,L}(\theta)$  towards  
 5  $Q_L^*(\theta) = \sum_{\ell=1}^L |e_\ell(\phi^*, \theta)|^2$  for  $\theta$  in  $\Theta$ , and the fact that the true parameter  $\hat{\theta}_{n,L}$  is a unique isolated  
 6 global maximum (by theorem 5.7 in [24]). In a first step, we assume that  $Q_L^*(\theta)$  has a unique global  
 7 minimum denoted  $\theta_L^*$ , and we will show that this is indeed the case, and that  $\theta_L^*$  is not very far from  
 8 the true parameter  $\theta^*$ .

9 We have to show that  $\sup_{\theta} |Q_{n,L}(\theta) - Q_L^*(\theta)| \rightarrow 0$  as  $n \rightarrow \infty$ . From the simple additive expression  
 10 of  $Q_{n,L}(\theta)$ , we see that it suffices to show the uniform convergence of  $|e_\ell(\hat{\phi}, \theta)|^2$  to  $|e_\ell(\phi^*, \theta)|^2$ . From  
 11 the inequality  $|a^2 - b^2| \leq |a - b|(|a| + |b|)$ , we have

$$\left| |e_\ell(\hat{\phi}, \theta)|^2 - |e_\ell(\phi^*, \theta)|^2 \right| \leq |e_\ell(\hat{\phi}, \theta) - e_\ell(\phi^*, \theta)| \left( |e_\ell(\hat{\phi}, \theta)| + |e_\ell(\phi^*, \theta)| \right).$$

12 As  $\hat{\phi}$  is such that for  $j = 1, 2$  we have  $\|\hat{\phi}_j - \phi_j^*\|_\infty = O_P\left(\zeta_0(K) \left(\sqrt{K/n} + K^{-\alpha}\right)\right)$  this means that  
 13  $\hat{\phi}(t) \in \mathcal{D} = \{x \in \mathbb{R}^2 \mid \exists t \in [0, 1], |x - \phi^*(t)| \leq D\}$  with a probability tending to 1. This shows that  
 14 with a probability tending to 1, we have  $|f(t, \hat{\phi}(t), \theta)| \leq h(t, \theta) \leq H(t)$  (because  $f$  is uniformly  $L^2$ -  
 15 Caratheodory), moreover  $\hat{\phi}$  is also bounded by  $M + D > 0$  (because  $\phi^*$  is bounded as a continuous  
 16 function on  $[0, 1]$ ) and

$$\begin{aligned} |e_\ell(\hat{\phi}, \theta)| + |e_\ell(\phi^*, \theta)| &\leq \left| \langle \mathcal{E}(\hat{\phi}, \theta), \varphi_\ell \rangle \right| + \left| \langle \hat{\phi}, \dot{\varphi}_\ell \rangle \right| + |\langle \mathcal{E}(\phi^*, \theta), \varphi_\ell \rangle| + |\langle \phi^*, \dot{\varphi}_\ell \rangle| \\ &\leq 2 \|H\|_{L^2} + \|\dot{\varphi}_\ell\|_{L^2} (2M + D) \end{aligned}$$

17 Hence  $|e_\ell(\hat{\phi}, \theta)| + |e_\ell(\phi^*, \theta)|$  is uniformly bounded in probability, so that the uniform convergence of  
 18 the criterion boils down to the uniform convergence of  $|e_\ell(\hat{\phi}, \theta) - e_\ell(\phi^*, \theta)|$ .

19 We can re-write  $e_\ell(\hat{\phi}, \theta) - e_\ell(\phi^*, \theta) = \langle \mathcal{E}(\hat{\phi}, \theta) - \mathcal{E}(\phi^*, \theta), \varphi_\ell \rangle + \langle \hat{\phi} - \phi^*, \dot{\varphi}_\ell \rangle$  and it is clear that  
 20 the second right-hand side term converges uniformly to zero in probability for all  $\ell \geq 1$ . Consequently,  
 21 we just have to check that  $\mathcal{E}(\hat{\phi}, \theta) - \mathcal{E}(\phi^*, \theta)$  converges uniformly in  $\theta$  to 0 in probability.

1 First of all, we remark that  $g \mapsto \langle \mathcal{E}(g, \theta), \varphi_\ell \rangle$  is a continuous function from  $B_\infty(\phi^*, D) = \{g \in$   
2  $C([0, 1]) \mid \|g - \phi^*\|_\infty \leq D\}$  to  $\mathbb{R}$  (w.r.t the sup-norm), because

$$\begin{aligned} |\langle \mathcal{E}(g, \theta), \varphi_\ell \rangle - \langle \mathcal{E}(g', \theta), \varphi_\ell \rangle| &\leq \langle a(\cdot, \theta) |g - g'|, \varphi_\ell \rangle \\ &\leq \langle a(\cdot, \theta), \varphi_\ell \rangle \|g - g'\| \end{aligned}$$

3 with  $a(\cdot, \theta) \in L^2$  for all  $\theta$ , as the vector field is  $L^2$ -Lipschitz. By the continuous mapping theorem,  
4 we get the point-wise convergence of  $\langle \mathcal{E}(\hat{\phi}, \theta) - \mathcal{E}(\phi^*, \theta), \varphi_\ell \rangle$ , and the hard part consist in the uniform  
5 convergence. This can be proven by controlling the oscillations in  $\theta$  of the process  $\mathcal{E} = (\mathcal{E}(\theta))_{\theta \in \Theta} =$   
6  $\left( \langle \mathcal{E}(\hat{\phi}, \theta) - \mathcal{E}(\phi^*, \theta), \varphi_\ell \rangle \right)_{\theta \in \Theta}$  and by using theorem 18.11 in [24], as convergence in probability towards  
7 a constant is equivalent to weak convergence. In order to show that the process  $\mathcal{E}$  converges weakly to  
8 0 in the space  $C(\Theta)$  of continuous functions on  $\Theta$  equipped with the supremum norm, we have to check  
9 that for all  $k$  and all  $\theta_1, \dots, \theta_k$  in  $\Theta$

$$(\mathcal{E}(\theta_1), \dots, \mathcal{E}(\theta_k)) \rightsquigarrow (0 \dots 0) \tag{A.1}$$

10 and that for all  $\epsilon, \alpha > 0$  there exists a partition of  $\Theta_1, \dots, \Theta_K$  of  $\Theta$  such that

$$\limsup_{n \rightarrow \infty} P \left( \sup_k \sup_{\theta, \theta' \in \Theta_k} |\mathcal{E}(\theta) - \mathcal{E}(\theta')| \geq \alpha \right) \leq \epsilon. \tag{A.2}$$

11 The first condition (A.1) is a direct consequence of the point-wise convergence in probability of  $\mathcal{E}(\theta)$  in  
12  $\Theta$ . Concerning the second condition, from we have for  $t$  in  $[0, 1]$  a.e.

$$\begin{aligned} \left| f \left( t, \hat{\phi}(t), \theta \right) - f \left( t, \hat{\phi}(t), \theta' \right) \right| &\leq \sum_{i=1}^p \left| f_{\theta_i} \left( t, \hat{\phi}(t), \tilde{\theta}_n(t) \right) \right| \left| \theta_i - \theta'_i \right| \\ &\preceq \bar{\mathbf{a}}'(t) \left| \theta - \theta' \right| \end{aligned}$$

13 with a probability tending to 1 ( $\tilde{\theta}_n(t)$  being a parameter between  $\theta$  and  $\theta'$ ). As a consequence the

1 following inequality

$$|\mathcal{E}(\theta) - \mathcal{E}(\theta')| \preceq \left\langle \bar{\mathbf{a}}' \left| \theta - \theta' \right|, |\varphi_\ell| \right\rangle \preceq \left| \theta - \theta' \right|$$

2 is true with a probability tending to 1. Since  $\Theta$  is a compact set, it is possible to find a finite partition  
 3  $\Theta_1, \dots, \Theta_K$  of  $\Theta$  such that the diameter of  $\Theta_i$  is smaller than an arbitrary  $\alpha$  independently of  $n$ . This  
 4 ensures that condition (A.2) is also satisfied and the uniform convergence of  $Q_{n,L}(\theta)$  to  $Q_L^*(\theta)$  can be  
 5 claimed. Finally, from lemma 5, we see that there exists an isolated global minimum  $\theta_L^*$  of  $Q_L^*(\theta)$ , i.e.  
 6 such that for every  $\epsilon > 0$

$$\inf_{\theta/|\theta-\theta^*| \geq \epsilon} Q_L^*(\theta) > Q_L^*(\theta_L^*)$$

7 As a consequence,  $\hat{\theta}_{n,L}$  converges to  $\theta_L^*$  in probability. Moreover, lemma 5 asserts that the bias between  
 8  $\theta_L^*$  and  $\theta^*$  is of order  $r_L$ , which is the (uniform) approximation error for the set of function  $(\mathcal{E}(\phi^*, \theta))_{\theta \in \Theta}$ .  
 9 If the envelope function is sufficiently regular (in  $H^1$ ), and we use the sine basis we get that the uniform  
 10 approximation error is  $o(\frac{1}{L})$ .

11 **Lemma 5.** *Under condition C1 et C5, there exists  $L_0$  such that for all  $L > L_0$ ,  $Q_L^*$  has a local minimum*  
 12  *$\theta_L^*$  such that for all  $\epsilon > 0$ ,  $\inf_{|\theta-\theta_L^*| > \epsilon} Q_L^*(\theta) > 0$ . Moreover, for all  $\epsilon' > 0$ , there exists  $L'_0 > L_0$  such that*  
 13 *for all  $L > L'_0$*

$$|\theta_L^* - \theta^*| \leq \epsilon'.$$

14 *In particular, if  $\bar{h}_{\theta\theta}$  is in  $H^1$  and  $(\varphi_\ell)_{\ell \geq 1}$  is the sine basis, then  $\theta_L^* - \theta^* = o(\frac{1}{L})$ .*

15 *Proof.* The first step is to show that  $Q_L^*(\theta)$  has a local isolated minimum  $\theta_L^*$  close to  $\theta^*$ , and that this  
 16 distance can be made arbitrarily small. We introduce first some useful notations. Since  $(\varphi_\ell)_{\ell \geq 1}$  is an or-  
 17 thonormal basis in  $L^2([0, 1])$ , we have  $\mathbf{E}_{\mathcal{F}}(\phi^*, \theta) = \sum_{\ell \geq 1} e_\ell(\phi^*, \theta) \varphi_\ell$  and  $\partial_\theta \mathbf{E}_{\mathcal{F}}(\phi^*, \theta) = \sum_{\ell \geq 1} \partial_\theta e_\ell(\phi^*, \theta) \varphi_\ell$   
 18 where for  $j = 1, \dots, p$ ,

$$\partial_{\theta_j} e_\ell(\phi^*, \theta) = \langle \partial_{\theta_j} \mathcal{E}(\phi^*, \theta), \varphi_\ell \rangle = \langle f_{\theta_j}(\cdot, \phi^*, \theta), \varphi_\ell \rangle$$

19 because we can differentiate under the integral sum ( $f$  is uniformly  $L^2$ -Lipschitz in  $\mathcal{D} \times \Theta$ ) and we  
 20 have also the Hessian matrix defined element wise as  $\partial_{\theta_i \theta_j} \mathbf{E}_{\mathcal{F}}(\phi^*, \theta) = \sum_{\ell \geq 1} \partial_{\theta_i \theta_j} e_\ell(\phi^*, \theta) \varphi_\ell$  with  
 21  $\partial_{\theta_i \theta_j} e_\ell(\phi^*, \theta) = \langle f_{\theta_i \theta_j}(\cdot, \phi^*, \theta), \varphi_\ell \rangle$ , because  $f_\theta(t, x, \theta)$  is also uniformly  $L^2$ -Lipschitz in  $\mathcal{D} \times \Theta$ . First

1 of all, we remark that  $\mathcal{E}(\phi^*, \theta)$  is bounded by  $\bar{\mathbf{h}} \in L^2$  (uniformly in  $\theta$ , condition **C2**), which implies that  
2  $\|\mathbf{E}_L(\phi^*, \theta)\|_{L^2}^2$  converges also uniformly to  $\|\mathbf{E}_{\mathcal{F}}(\phi^*, \theta)\|_{L^2}^2$ . Let  $\epsilon' > 0$ , since  $\theta^*$  is a local strict minimum  
3 of  $Q_{\mathcal{F}}$ , there exists  $L_0$  such that for all  $L > L_0$ ,  $Q_L^*$  has a strict local minima such that  $|\theta_L^* - \theta^*| \leq \epsilon'$ .  
4 Moreover, by uniform convergence, we have also that  $\inf_{|\theta - \theta_L^*| > \epsilon} Q_L^*(\theta) > 0$ .

5 Now, we relate  $\theta_L^* - \theta^*$  to the approximation quality of  $\mathbf{E}_{\mathcal{F}}(\phi^*, \theta)$  by the basis  $(\phi_\ell)_{\ell \geq 1}$ . We remark  
6 first that we have  $|\mathbf{E}_{\mathcal{F}}(\phi^*, \theta)| \leq 2\bar{\mathbf{h}}(t)$ : this implies that for all  $\ell$ , and all  $\theta$  we have  $|e_\ell^*(\theta)| \leq 2\bar{\mathbf{h}}_\ell$  with  
7  $\bar{\mathbf{h}}_\ell = \langle \bar{\mathbf{h}}, \varphi_\ell \rangle$ . The global rate of convergence of the series  $\sum_\ell e_\ell^*(\theta)$  is controlled uniformly by the rate  
8 of  $\sum_{\ell \geq 1} H_\ell^2$ , denoted  $r_L^2 = \sum_{\ell > L} H_\ell^2$ . By orthogonality, we can write

$$\|\mathbf{E}_{\mathcal{F}}(\phi^*, \theta)\|_{L^2}^2 = \|\mathbf{E}_L(\phi^*, \theta)\|_{L^2}^2 + \|\mathbf{R}_L(\phi^*, \theta)\|_{L^2}^2$$

9 which means that is  $\|\mathbf{E}_L(\phi^*, \theta)\|_{L^2}^2$  is a perturbation of the function  $\theta \mapsto \|\mathbf{E}_{\mathcal{F}}(\phi^*, \theta)\|_{L^2}^2$  by the function  
10  $\theta \mapsto -\|\mathbf{R}_L(\phi^*, \theta)\|_{L^2}^2$ . This perturbation  $\|\mathbf{R}_L(\phi^*, \theta)\|_{L^2}^2$  is uniformly dominated by  $r_L$ , hence it becomes  
11 possible to relate the two minima.

12 From assumption **C5**, we know  $\|\mathbf{R}_L(\phi^*, \theta)\|_{L^2}^2$  is differentiable and we compute a series decomposition  
13 of its gradient thanks to  $\partial_\theta \mathbf{R}_L(\phi^*, \theta) = \sum_{\ell > L} \partial_\theta e_\ell(\phi^*, \theta) \varphi_\ell$ :

$$\partial_\theta \|\mathbf{R}_L(\phi^*, \theta)\|_{L^2}^2 = 2\mathbf{R}_L(\phi^*, \theta) \partial_\theta \mathbf{R}_L(\phi^*, \theta).$$

14 and we recall that  $\mathbf{R}_L(\phi^*, \theta)$  converges uniformly to 0 and  $\partial_\theta \mathbf{R}_L(\phi^*, \theta)$  is uniformly bounded in  $\theta$  on  $\Theta$   
15 (as a continuous function on the compact set  $\Theta$ ). Starting from this last remark, we use the Implicit  
16 Function Theorem to the continuously differentiable function  $G(\epsilon, \theta) = \partial_\theta \|\mathbf{E}_{\mathcal{F}}(\phi^*, \theta)\|_{L^2}^2 - 2\epsilon \partial_\theta \mathbf{R}_L(\phi^*, \theta)$ ,  
17 in order to get a Taylor expansion. We denote  $\theta_\epsilon$  the solution to  $G(\epsilon, \theta_\epsilon) = 0$ , and we remark that in particular  
18  $G(0, \theta^*) = 0$ . Thus, there exists  $\epsilon_0, \delta_0 > 0$  and a function  $\psi : ]-\epsilon_0, \epsilon_0[ \rightarrow B(\theta^*, \delta_0)$  such that  $\psi(\epsilon) = \theta_\epsilon$   
19 (i.e.  $G(\epsilon, \psi(\epsilon)) = 0$ ). We can also compute the first order variation of  $\psi$ :  $\psi(\epsilon) = \psi(0) + \epsilon\psi'(0) + o(\epsilon)$   
20 where

$$\psi'(0) = -2 \left( \partial_{\theta\theta} \|\mathbf{E}_{\mathcal{F}}(\phi^*, \theta)\|_{L^2}^2 \right)^{-1} \partial_\theta \mathbf{R}_L(\phi^*, \theta^*).$$

21 Since  $\mathbf{R}_L(\phi^*, \theta)$  converges uniformly to 0, there exists  $L_0 > 0$ , such that for  $L > L_0$ ,  $|\mathbf{R}_L(\phi^*, \theta)| \leq \epsilon_0$   
22 so that we can apply the linearization above to  $G(\mathbf{R}_L(\phi^*, \theta), \theta) = \partial_\theta \|\mathbf{E}_{\mathcal{F}}(\phi^*, \theta)\|_{L^2}^2 - \partial_\theta \|\mathbf{R}_L(\phi^*, \theta)\|_{L^2}^2$ .

1 We obtain that the minima  $\theta_L^*$  and  $\theta^*$  are such that

$$\theta_L^* = \theta^* + \mathbf{R}_L(\phi^*, \theta) 2 (\partial_{\theta\theta} \|\mathbf{E}_{\mathcal{F}}(\phi^*, \theta)\|_{L^2}^2)^{-1} \partial_{\theta} \mathbf{R}_L(\phi^*, \theta^*) + o(r_L).$$

2 This implies that  $|\theta_L^* - \theta^*| = O(r_L)$ . If we use the sine basis, this means that  $\mathbf{P}_{\mathcal{F}H} = \sum_{\ell} \bar{\mathbf{h}}_{\ell} \sqrt{2} \sin(\pi \ell t)$   
 3 but it is also in  $H_0^1$  (as it is in  $H^1$ ), then we have  $\sum_{\ell \geq 1} \ell^2 \bar{\mathbf{h}}_{\ell}^2$  and  $r_L^2 = o(\frac{1}{L^2})$ .  $\square$

### 4 Proposition 3

5 If conditions **C1-C6** are satisfied, then

$$\left[ \mathbf{J}_L(\hat{\phi}, \hat{\theta}_{n,L})^{\top} \tilde{\mathbf{J}}_L \right]^{-1} \mathbf{J}_L(\hat{\phi}, \hat{\theta}_{n,L})^{\top} \xrightarrow{P} \mathbf{M}_L^* = [\mathbf{J}_{\theta,L}^{*\top} \mathbf{J}_{\theta,L}^*]^{-1} \mathbf{J}_{\theta,L}^{*\top} \quad (\text{A.3})$$

6 and we have

$$\hat{\theta}_{n,L} - \theta_L^* = -\mathbf{M}_L^* \mathbf{e}_L(\hat{\phi}, \theta^*) + o_{\mathbb{P}}(1). \quad (\text{A.4})$$

7 The first order condition implies that  $\hat{\theta}_{n,L}$  satisfies (4.1). We develop a Taylor expansion of  $f(t, \hat{\phi}(t), \theta)$   
 8 around  $\theta^*$  of order 1 for  $t$  a.e. in  $[0, 1]$ :  $e_{\ell}(\hat{\phi}, \theta)$  as  $e_{\ell}(\hat{\phi}, \theta^*) + \left\langle f_{\theta}(\cdot, \hat{\phi}, \tilde{\theta}), \varphi_{\ell} \right\rangle$  where  $\tilde{\theta}$  is on straight  
 9 line between  $\theta^*$  and  $\hat{\theta}_{n,L}$ . We can write it in vector form

$$\mathbf{e}_L(\hat{\phi}, \hat{\theta}_{n,L}) = \mathbf{e}_L(\hat{\phi}, \theta^*) + \tilde{\mathbf{J}}_{\theta,L}(\hat{\theta}_{n,L} - \theta_L^*). \quad (\text{A.5})$$

10  $\tilde{\mathbf{J}}_{\theta,L}$  is a matrix  $\mathbb{R}^{L \times p}$  with entries  $\left\langle f_{\theta_i}(\cdot, \hat{\phi}, \tilde{\theta}), \varphi_{\ell} \right\rangle$ . Thus, if we premultiply (A.5) by  $\mathbf{J}_{\theta,L}(\hat{\phi}, \hat{\theta}_{n,L})^{\top}$ ,  
 11 we get the asymptotic expansion

$$0 = \mathbf{J}_{\theta,L}(\hat{\phi}, \hat{\theta}_{n,L})^{\top} \mathbf{e}_L(\hat{\phi}, \theta^*) + \mathbf{J}_{\theta,L}(\hat{\phi}, \hat{\theta}_{n,L})^{\top} \tilde{\mathbf{J}}_{\theta,L}(\hat{\theta}_{n,L} - \theta_L^*). \quad (\text{A.6})$$

12 This shows that the key results for relating the behavior of  $(\hat{\theta}_{n,L} - \theta_L^*)$  to the behavior of  $\mathbf{e}_L(\hat{\phi}, \theta^*)$  is  
 13 the convergence in probability of  $\mathbf{J}_{\theta,L}(\hat{\phi}, \hat{\theta}_{n,L})^{\top}$ . Indeed, if of to prove the convergence of true since the

1 matrix  $\mathbf{J}_L^{*\top} \mathbf{J}_L^*$  is nonsingular and

$$\left[ \mathbf{J}_{\theta,L} \left( \hat{\phi}, \hat{\theta}_{n,L} \right)^\top \tilde{\mathbf{J}}_L \right]^{-1} \mathbf{J}_{n,L}(\hat{\theta}_{n,L})^\top$$

2 converges in probability to  $[\mathbf{J}^* \mathbf{J}^*]^{-1} \mathbf{J}^{*\top} = \mathbf{M}$ . Indeed,  $\mathbf{J}_{n,L}(\theta)$  converges uniformly to  $\mathbf{J}_L(\theta)$  because the  
3 functions  $\frac{\partial}{\partial \theta_i} f$  are uniformly Lipschitz, thus this implies that  $\hat{\theta}_{n,L} - \theta_L^* = -\mathbf{M}_L \mathbf{e}_{n,L}(\theta^*) + o_{\mathbb{P}}(1)$ .

## 4 References

- 5 [1] D. K. Andrews. Asymptotic normality of series estimators for nonparametric and semiparametric  
6 regression models. *Econometrica*, 59(2):307–345, 1991.
- 7 [2] P.J. Bickel and Y. Ritov. Nonparametric estimators which can be plugged-in. *Annals of Statistics*,  
8 31(4):4, 2003.
- 9 [3] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer,  
10 2008.
- 11 [4] N. J-B. Brunel. Parameter estimation of ode’s via nonparametric estimators. *Electronic Journal of*  
12 *Statistics*, 2:1242–1267, 2008.
- 13 [5] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer-Verlag, 2005.  
14 Localisation : Nicolas.
- 15 [6] Hein W Engl, Christoph Flamm, Philipp Kügler, James Lu, Stefan Müller, and Peter Schuster.  
16 Inverse problems in systems biology. *Inverse Problems*, 25(12), 2009.
- 17 [7] L. Goldstein and K. Messer. Optimal plug-in estimators for nonparametric functional estimation.  
18 *The annals of statistics*, 20(3):1306–1328, 1992.
- 19 [8] S. Gugushvili and C.A.J. Klaassen. Root-n-consistent parameter estimation for systems of ordinary  
20 differential equations: bypassing numerical integration via smoothing. *Bernoulli*, to appear, 2011.
- 21 [9] A.R. Hall. *Generalized Method of Moments*. Oxford University Press, 2005.

- 1 [10] E. L. Ionides, A. Bhadra, Y. Atchade, and A. A. King. Iterated filtering. *Annals of Statistics*,  
2 39:1776–1802, 2011.
- 3 [11] E. L. Ionides, C. Breto, and A. A. King. Inference for nonlinear dynamical systems. *Proceedings of*  
4 *the National Academy of Sciences*, 103:18438–18443, 2006.
- 5 [12] H. Liang and H. Wu. Parameter estimation for differential equation models using a framework  
6 of measurement error in regression models. *Journal of the American Statistical Association*,  
7 103(484):1570–1583, December 2008.
- 8 [13] J. Madar, J. Abonyi, H. Roubos, and F. Szeifert. Incorporating prior knowledge in cubic spline ap-  
9 proximation - application to the identification of reaction kinetic models. *Industrial and Engineering*  
10 *Chemistry Research*, 42(17):4043–4049, 2003.
- 11 [14] H. Miao, X. Xia, A. S. Perelson, and H. Wu. On identifiability of nonlinear ode models and  
12 applications in viral dynamics. *SIAM Review*, 53:3–39, 2011.
- 13 [15] C.G. Moles, P. Mendes, and J.R. Banga. Parameter estimation in biochemical pathways: a com-  
14 parison of global optimization methods. *Genome Research*, 13:2467–2474, 2003.
- 15 [16] W. K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of*  
16 *Econometrics*, 79:147–168, 1997.
- 17 [17] D. O’Regan. *Existence theory for nonlinear ordinary differential equations*. Mathematics and its  
18 applications. Kluwer, 1997.
- 19 [18] A. A. Poyton, M.S. Varziri, K.B. McAuley, P.J. McLellan, and J.O. Ramsay. Parameter estimation  
20 in continuous-time dynamic models using principal differential analysis. *Computers and Chemical*  
21 *Engineering*, 30:698–708, 2006.
- 22 [19] L. Pronzato. Optimal experimental design and some related control problems. *Automatica*, 44:303–  
23 325, 2008.



- 1 [20] M. Quach, N. Brunel, and F. d'Alche Buc. Estimating parameters and hidden variables in non-linear  
2 state-space models based on odes for biological networks inference. *Bioinformatics*, 23(23):3209–  
3 3216, 2007.
- 4 [21] J.O. Ramsay, G. Hooker, J. Cao, and D. Campbell. Parameter estimation for differential equations:  
5 A generalized smoothing approach. *Journal of the Royal Statistical Society (B)*, 69:741–796, 2007.  
6 To appear.
- 7 [22] D. Ruppert, M.P. Wand, and R.J. Carroll. *Semiparametric regression*. Cambridge series on statis-  
8 tical and probabilistic mathematics. Cambridge University Press, 2003.
- 9 [23] S. van de Geer. *Empirical processes in M-estimation*. Cambridge University Press, 2000.
- 10 [24] A.W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilities  
11 Mathematics. Cambridge University Press, 1998.
- 12 [25] J. M. Varah. A spline least squares method for numerical parameter estimation in differential  
13 equations. *SIAM J.sci. Stat. Comput.*, 3(1):28–46, 1982.
- 14 [26] E.O. Voit and J. Almeida. Decoupling dynamical systems for pathway identification from metabolic  
15 profiles. *Bioinformatics*, 20(11):1670–1681, 2004.
- 16 [27] H. Xue, H. Miao, and H. Wu. Sieve estimation of constant and time-varying coefficients in nonlinear  
17 ordinary differential equation models by considering both numerical error and measurement error.  
18 *Annals of Statistics*, 38(4):2351–2387, 2010.