Parameter estimation of ODE's via nonparametric estimators

Nicolas Brunel[*]

June 15, 2007

Abstract

Ordinary differential equations (ODE's) are widespread models in physics, chemistry and biology. In particular, this mathematical formalism is used for describing the sets of interactions and the evolution of complex systems and it might consist of high-dimensional sets of coupled nonlinear differential equations. In this setting, we propose a general method for estimating the parameters indexing ODE's from times series. Our method is able to alleviate the computational difficulties encountered by the classical parametric methods. These difficulties are due to the implicit definition of the model. We propose the use of a nonparametric estimator of regression functions as a first-step in the construction of an M-estimator, and we show the consistency of the derived estimator under general conditions. In the case of spline estimators, we provide asymptotic normality, and we derive the rate of convergence, which is not the usual $\sqrt{n}$-rate for parametric estimators. This rate depends on the smoothness of the differential equation. Some perspectives of refinements of this new family of parametric estimators are given.

Key words: Consistency, Ordinary Differential Equation, Splines, Nonparametric regression, Parametric estimation, M-estimator.

1 Introduction

Ordinary differential equations are used for the modelling of dynamic processes in physics, engineering, chemistry, biology,etc. In particular, such a

[*]EURANDOM P.O. Box 513, 5600 MB Eindhoven, The Netherlands
nicolas.brunel@ibisc.fr

formalism is used for the description of regulatory networks (for example between competing species in biology), or of cell regulatory systems e.g. the temporal evolution of the concentrations of some biochemical species (mRNA, proteins) involved in biological functions inside the cell [5]. Usually, the model for the state variables $x = (x_1, \dots, x_d)^\top$ consists in an initial value problem

$$\left\{ \begin{array}{lcl} \dot{x}(t) & = & F(x(t), \theta),\ \forall t \in [0,1], \\ x(0) & = & x_0, \end{array} \right. \tag{1}$$

where F is a vector field from $\mathbb{R}^d$ to $\mathbb{R}^d$, $d \in \mathbb{N}$, and $\theta \in \Theta$, Θ being a subset of a Euclidean space. When data are available such as a time series, we are interested in the problem of estimation of the coefficients parametrizing the ODE. In principle, this may be done by some classical parametric estimators, usually the least squares estimator [13] or the Maximum Likelihood estimator (MLE). Different estimators have been derived in order to take into account some particular features of the differential equation such as special boundary values (there exists a function g linking the values at the boundary i.e. $g(x(0), x(1)) = 0$ instead of the simple initial value problem), or random initial values or random parameters [4]. Otherwise, there may be some variations on the observational process such as noisy observation times that necessitate the introduction of appropriate minimization criteria [12].

Despite their satisfactory theoretical properties, the efficiency of these estimators may be dramatically degraded in practice by computational problems that arise from the implicit and nonlinear definition of the model. Indeed, these estimators give rise to nonlinear optimization problems that necessitate the approximation of the solution of the ODE and the exploration of the (usually high-dimensional) parameter space. Hence, we have to face possibly numerous local optima and a huge computation time. Instead of considering the estimation of θ straightforwardly as a parametric problem, it may be useful to look at it as the estimation of a univariate regression function $t \mapsto x(t)$ that belongs to the (finite dimensional) family of functions satisfying (1). So we may use tools from functional estimation in order to derive a proxy for the solution of the ODE and derive estimates of the parameters from this. Similar attempts of getting a smooth approximation of the solution without solving the ODE were made by Madar *et al.* [14] or Varah [20] with cubic splines (and a well-chosen sequence of knots). Different spline estimators were proposed by Ramsay and Silverman [15] based on the fact that smoothing splines are obtained by solving the trade-off between adequacy to data and smoothness measured by some linear differential operators. It was extended more recently by Ramsay *et al.* [16] to the case of nonlinear differential operators. Moreover, this functional point of view enables one to use prior knowledge on the solutions of the ODE such as positivity or boundedness whereas it is difficult to exploit the strictly parametric form. Indeed, it implies that we have a thorough knowledge of the influence of the parameters on the qualitative behavior of the solutions of (1), which is rarely the case. In this paper, we exploit this interpretation of estimation of ODE's as functional estimation, so that we are able to obtain a general estimation procedure, by exploiting numerous results from nonparametric regression theory.

In the next section, we introduce the statistical model and we define our so-called two-step estimator of θ. We show that under broad conditions this estimator is consistent, and we give some straightforward extensions of this estimator to different models. In section 3, we review some useful definitions and results of spline theory which are useful for understanding the properties of the spline estimator derived in section 4. We derive then the rate of convergence of the parametric estimator in this particular case. In the last section, we give some simulation results obtained with the classical Lotka-Volterra's population model coming from biology. In conclusion, we give some possible extensions of this work.

2 Two-step estimator

2.1 Statistical model

We want to estimate the parameter θ of the ordinary differential equation (1) from noisy observations at n points in $[0,1]$, $0 \leq t_1 < \cdots < t_n \leq 1$,

$$y_i = x(t_i) + \epsilon_i,\ i = 1, \ldots, n, \tag{2}$$

where the ϵ_i's are i.i.d centered random variables. The ODE is indexed by a parameter $\theta \in \Theta \subset \mathbb{R}^p$ with initial value x_0; the true parameter value is θ^* and the corresponding solution of (1) is x^*.

The vector field defining the ODE is a function $F : \mathcal{X} \times \Theta \rightarrow \mathbb{R}^d$ ($\mathcal{X} \subset \mathbb{R}^d$) of class C^m w.r.t x for every θ and with $m \geq 1$. It is a Lipschitz function so that we have existence and uniqueness of a solution x_{θ,x_0} to (1) on a neighborhood of 0 for each θ and x_0; and we assume that every corresponding solution can be defined on $[0,1]$. Hence, the solutions x_{θ,x_0} belong to $\mathcal{C}^{m+1}([0,1],\mathbb{R}^d)$. Moreover, we suppose also that F is a smooth function in θ so that each solution x_{θ,x_0} depends smoothly[1] on the parameters θ and x_0. Then, we suppose that F is of class C^1 in θ for every x. Let f_Σ be the density of the noise ϵ, then the log-likelihood in the i.i.d case is

$$l(\theta, x_0, \Sigma) = \sum_{i=1}^{n} \log f_\Sigma(y_i - x_{\theta,x_0}(t_i)) \tag{3}$$

and the model that we want to identify is parametrized by $(\theta, x_0, \Sigma) \in \Theta \times \mathcal{X} \times \mathcal{S}^+$ for instance when the noise is centered Gaussian with covariance matrix Σ ($\mathcal{S}^+$ is the set of symmetric positive matrices). An alternative parametrization is $(\theta, x_{\theta,x_0}, \Sigma) \in \Theta \times \mathcal{F} \times \mathcal{S}^+$, with $\mathcal{F}$ the set of functions that solve (1) for some θ and x_0, thanks to the injective mapping between initial conditions and a solution.

In most applications, we are not really interested in the initial conditions but rather in the parameter θ, so that x_0 or x_{θ,x_0} can be viewed as a nuisance parameter like the covariance matrix Σ of the noise. We want to define estimators of the "true" parameters (x^*, θ^*) ($x^* = x_{\theta^*,x_0^*}$) that will be denoted by $(\hat{x}_n, \hat{\theta}_n)$. The estimation problem appears as a standard parametric problem that can be dealt with by the classical theory in order to provide good estimators (with good properties, e.g. $\sqrt{n}$-consistency) such as the Maximum Likelihood Estimator (MLE). Indeed, from the smoothness properties of F, the log-likelihood $l(\theta, x_0)$ is at least C^1 w.r.t (θ, x_0) so that we can define the score $s(\theta, x_0) = (\frac{\partial l}{\partial \theta}^\top \frac{\partial l}{\partial x_0}^\top)^\top$. If $s(\theta, x_0)$ is square integrable under the true probability$P_{(x^*,\theta^*)}$, we can claim under weak conditions (e.g. theorem 5.39 [19]) that the MLE is an asymptotically efficient estimator. The difficulty of this approach is then essentially practical because of the implicit dependence of x on the parameter (θ, x_0), which prohibits proper maximization of $l(\theta, x_0)$. Indeed, derivative-based methods like Newton-Raphson are not easy to handle then and evaluation of the likelihood necessitates the integration of the ODE, which becomes a burden when we have to explore a huge parameter space. Moreover, the ODE's proposed for modelling may be expected to give a particular qualitative behavior which can be easily interpreted in terms of systems theory, e.g. convergence to an equilibrium state or oscillations. Typically, these qualitative properties of ODE are hard to control and involve bifurcation analysis [11] and may necessitate a mathematical knowledge which is not always accessible for huge systems. Moreover, boundedness of the solution x^* ($a \leq x^*(t) \leq b$, with $a, b \in \mathbb{R}^d$) may be difficult to use

[1] if F depends smoothly on x and θ then the solution depends on the parameter by the same order of smoothness, see Anosov & Arnold, Dynamical systems, p.17.

during the estimation via the classical device of a constraint optimization. Hence, these remarks motivate us to consider the estimation of an ODE as a functional estimation and use flexible methods coming from nonparametric regression from which we could derive a likely parameter for the ODE.

2.2 Principle

We use consistent nonparametric estimators of the solution x^* and its derivative $\dot{x}^*$ in order to derive a fitting criterion for the ODE and subsequently the M-estimator of θ^* corresponding to the criterion. We denote by $\|f\|_q = \left(\int_0^1 |f(t)|^q dt\right)^{1/q}$,$0 < q \leq \infty$, the L^q norm on the space of Lebesgue integrable functions on $[0,1]$. By using classical nonparametric regression estimators, we can construct consistent estimators $\hat{x}_n$ and $\hat{\dot{x}}_n$ of x^* and $\dot{x}^*$ (actually we will obtain the estimator of the derivative by deriving $\hat{x}_n$ so that $\hat{\dot{x}}_n = \dot{\hat{x}}_n$) i.e. $\|\hat{x}_n - x^*\|_q = o_P(1)$ and $\|\hat{\dot{x}}_n - \dot{x}^*\|_q = o_P(1)$. We may choose as criterion function to minimize $R_n^q(\theta) = \|\dot{\hat{x}}_n - F(\hat{x}_n, \theta)\|_q$ from which we derive the two-step estimator

$$\hat{\theta}_n \quad = \quad \arg\min_{\theta} R_n^q(\theta). \tag{4}$$

Thanks to the previous convergence results and under additional suitable conditions to be specified below, we can show that $R_n^q(\theta) \to R^q(\theta) = \|\dot{x}^* - F(x^*, \theta)\|_q$ in probability, and that this discrepancy measure enables us to construct a consistent estimator $\hat{\theta}_n$. Note that there are no computational difficulties now as there are in the straightforward parametric model approach.

We are left with two choices of practical and theoretical importance: the choice of q and the choice of the nonparametric estimator. In this paper, we focus on the one hand on $q = 2$ (so that the optimization program (4) can be processed as a nonlinear least squares regression) and on the other hand we consider splines with a number of knots depending on the number of observations n. It is likely that some other families of nonparametric estimators such as smoothing splines, kernels or wavelets could be used, depending on the performance or the type of constraints we want.

2.3 Consistency

We show that the minimization of $R_n^q(\theta)$ gives a consistent estimator for θ. We introduce the asymptotic criterion $R^q(\theta) = \|F(x^*, \theta^*) - F(x^*, \theta)\|_q$

derived from R^q_n and we make the additional assumption:

$$\forall \epsilon > 0, \inf_{\|\theta-\theta^*\|\geq\epsilon} R^q(\theta) > R^q(\theta^*), \tag{5}$$

which may be viewed as an identifiability criterion for the model.

Proposition 2.1. *We suppose there exists a compact set $\mathcal{K} \subset \mathcal{X}$ such that $\forall\, \theta \in \Theta, \forall\, x_0 \in \mathcal{X}, \forall t \in [0,1]$, $x_{\theta,x_0}(t)$ is in $\mathcal{K}$. Moreover we suppose that uniformly in $\theta \in \Theta$, $F(\cdot,\theta)$ is $K-$ Lipschitz on $\mathcal{K}$. If $\hat{x}_n$ and $\hat{\dot{x}}_n$ are consistent, and $\hat{x}_n(t) \in \mathcal{K}$ almost surely, then we have*

$$\sup_{\theta\in\Theta} |R^q_n(\theta) - R^q(\theta)| = o_P(1).$$

Moreover, if the identifiability condition (5) is fulfilled the two-step estimator is consistent, i.e.

$$\hat{\theta}_n - \theta^* = o_P(1).$$

Proof. In order to show the convergence of $|R^q_n(\theta)-R^q(\theta)| = |\|\hat{\dot{x}}_n - F(\hat{x}_n,\theta)\|_q - \|F(x^*,\theta) - F(x^*,\theta^*)\|_q|$, we make the following decomposition

$$\begin{aligned} |R^q_n(\theta) - R^q(\theta)| &\leq \| \left(\hat{\dot{x}}_n - F(\hat{x}_n,\theta)\right) + (F(x^*,\theta) - F(x^*,\theta^*)) \|_q \\ &\leq \|\hat{\dot{x}}_n - F(x^*,\theta^*)\|_q + \|F(\hat{x}_n,\theta) - F(x^*,\theta)\|_q. \end{aligned} \tag{6}$$

Since all the solutions $x_{\theta,x_0}(t)$ and $\hat{x}_n(t)$ stay in $\mathcal{K} \subset \mathcal{X}$, and $x \mapsto F(x,\theta)$ are $K-$ Lipschitz uniformly in θ, we obtain for all $\theta \in \Theta$

$$\|F(\hat{x}_n,\theta) - F(x^*,\theta)\|_q \leq K \left(\int_0^1 |\hat{x}_n(t) - x^*(t)|^q dt \right)^{1/q} = K\|\hat{x}_n - x^*\|_q. \tag{7}$$

Together, (6) and (7) imply

$$\begin{aligned} \sup_{\theta\in\Theta} |R^q_n(\theta) - R^q(\theta)| &\leq \|\hat{\dot{x}}_n - F(x^*,\theta^*)\|_q + \sup_{\theta\in\Theta} \|F(\hat{x}_n,\theta) - F(x^*,\theta)\|_q \\ &\leq \|\hat{\dot{x}}_n - F(x^*,\theta^*)\|_q + K\|\hat{x}_n - x^*\|_q. \end{aligned}$$

and consequently, by the consistency of $\hat{x}_n$ and $\dot{\hat{x}}_n$,

$$\sup_{\theta\in\Theta} |R^q_n(\theta) - R^q(\theta)| = o_P(1).$$

With the additional identifiability condition (5) for the vector field F, Theorem 5.7 of [19] implies that the estimator $\hat{\theta}_n$ converges in probability to θ^*.

□

In the case $q = 2$, the Hessian of $R^2(\theta)$ at $\theta = \theta^*$ is

$$J^* = \int_0^1 \left(D_\theta F(x^*(t), \theta^*)\right)^\top D_\theta F(x^*(t), \theta^*) dt.$$

Nonsingularity of J^* enables one to have a local identifiability criterion; indeed, the criterion behaves like a positive definite quadratic form on a neighborhood $\mathcal{V}(\theta^*)$ of θ^* so that condition (5) is true on $\mathcal{V}(\theta^*)$.

Remark 2.1.

The principle of the two-step estimation is the same when the ODE is nonautonomous, i.e. the vector field depends also on time. Moreover, proposition 2.1 can be adapted to $R_n^q(\theta) = \int_0^1 |\dot{\hat{x}}_n(t) - F(\hat{x}_n(t), t; \theta)|^q dt$, *and* $\hat{\theta}_n$ *is consistent provided that the asymptotic criterion* $R^q(\theta) = \int_0^1 |\dot{x}^*(t) - F(x^*(t), t; \theta)|^q dt$ *has property (5).*

Remark 2.2.

The estimator proposed can be easily extended to cases where several variables are not observed. Indeed, if the differential system (1) is partially linear in the following sense

$$\left\{ \begin{array}{lcl} \dot{u} & = & G(u, v; \eta) \\ \dot{v} & = & H(u; \eta) + Av \end{array} \right. \tag{8}$$

with $x = (u^\top \, v^\top)^\top$, $u \in \mathbb{R}^{d_1}$, *being observed,* $v \in \mathbb{R}^{d_2}$ *being unobserved, and* $d_1 + d_2 = d$ *(the initial conditions are* $x_0 = (u_0^\top \, v_0^\top)^\top$*), i.e.* $x(t_i)$ *is replaced by* $u(t_i)$ *in (2) (the noise* ϵ_i *being then* d_1*-dimensional). We want to estimate the parameter* $\theta = (\eta, A)$ *when* H *is a nonlinear function and* A *is a matrix, so we can take advantage of the linearity in* v *in order to derive an estimator for* v*. We can derive a nonparametric estimator for* v *by using* $\hat{u}_n$ *and the fact that* $t \mapsto v(t)$ *must be the solution of the non-homogeneous linear ODE* $\dot{v} = Av + H(\hat{u}_n; \eta)$, $v(0) = v_0$*. The solution of this ODE is given by Duhamel's formula [8]*

$$\forall t \in [0, 1],\ \hat{v}_n(t) = \exp(tA)\, v_0 + \int_0^t \exp\left((t-s)A\right) H(\hat{u}_n(s); \theta) ds, \tag{9}$$

which then can be plugged into the criterion $R_n^q(\theta)$. This estimator depends explicitly on the initial condition v_0 which must be estimated at the same time

$$(\hat{\theta}, \hat{v}_0) = \arg \min_{(\eta, A, v_0)} R_n^q(\eta, A, v_0) = \left\| \dot{\hat{u}}_n - F(\hat{u}_n, \hat{v}_n; \eta) \right\|_q .$$

As previously, if H is uniformly Lipschitz the integral $\int_0^t \exp\left((t-s)A\right) H(\hat{u}_n; \theta) ds$ converges (uniformly) in probability in the L^q sense to $\int_0^t \exp\left((t-s)A\right) H(u^*; \theta) ds$ as soon as $\hat{u}_n$ does, hence $R_n^q(\theta, A, v_0)$ converges also uniformly to the asymptotic criterion

$$R^q(\theta, v_0) = \left\| \dot{u}^* - F(u^*, v^*; \theta) \right\|_q .$$

The estimator $(\hat{\theta}, \hat{v}_0)$ is consistent as soon as $R^q(\theta, v_0)$ verifies the identifiability criterion (5).

Remark 2.3.

If the observation times $t_1, \ldots, t_n$ are realizations of i.i.d. random variables $(T_1, \ldots, T_n)$ with common c.d.f Q, the nonparametric estimators $\hat{x}_n$, as the one used before, are relevant candidates for the definition of the two-step estimator since they are still consistent under some additional assumptions on Q.

As in the setting considered by Lalam and Klaassen [12], the observation times may be observed with some random errors $\tau_i = t_i + \eta_i, i = 1, \ldots, n$, (the η_i's being some white noise) so we have to estimate x from the noisy bivariate measurements (τ_i, y_i). Consistent nonparametric estimators have been proposed for the so-called "errors-in-variables" regression and some examples are kernel estimators [6] and splines estimators [10] (in the L^2 sense). Hence, we can define exactly the same criterion function R_n^2 and derive a consistent parametric estimator.

3 Splines theory

In order to give a better insight into the properties of the nonparametric estimator proposed in section 4, we recall some properties of splines and B-splines. Indeed, the good statistical behavior of the estimator we propose, is based on our ability to approximate the solution x^* with a finite number of known functions. Moreover, the particular properties of the B-splines basis enable us to derive efficient algorithms for practical implementation of the related approximation method and give eventually computationally efficient statistical procedures. For the sake of completeness, we recall some facts about piecewise polynomial functions and splines, borrowed from De Boor [2].

We consider a sequence of increasing points (sometimes called breakpoints) $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{\ell+1})$ and we denote by $\mathbb{P}_{\boldsymbol{\xi},k}$ the set of piecewise polynomial functions (pp functions) of degree $k-1$ (or of order k), i.e., $f \in \mathbb{P}_{\boldsymbol{\xi},k} \Leftrightarrow \forall i = 1, \ldots, \ell,\ f_{|]\xi_i,\xi_{i+1}[}$ is a polynomial of degree at most $k-1$. $\mathbb{P}_{\boldsymbol{\xi},k}$ is a vector space of dimension equal to ℓk. A pp function is in general neither continuously differentiable nor continuous. Nevertheless, it is useful to introduce the j^{th} derivative of a pp function f denoted by $D^j f$ which is the piecewise derivative having possibly some jump[2] $\Delta_{\xi_i} f^{(j)}$ at the breakpoint ξ_i ($f^{(j)}$ is the j^{th} derivative of f on each open interval $]\xi_s, \xi_{s+1}[$, $s = 1, \ldots, \ell$). A pp function is usually described by the values of its right derivatives $\left(D^{(j-1)} f(\xi_i^+)\right)$, $i = 1, \ldots, \ell, j = 1, \ldots, k$. Of course, we are particularly interested in pp functions with regularities at some knots described by some integers $\boldsymbol{\nu} = (\nu_i)_{i=2,\ldots,\ell}$

$$\forall i \in [2, \ldots, l], \forall j \in [1, \ldots, \nu_i],\ D^{(j-1)} f(\xi_i^+) = D^{(j-1)} f(\xi_i^-).$$

This vector space is denoted $\mathbb{P}_{\boldsymbol{\xi},k,\boldsymbol{\nu}}$. When $\nu_i = \nu$ for all $i = 2, \ldots, \ell$, we denote $\mathbb{P}_{\boldsymbol{\xi},k,\nu} = C^\nu \cap \mathbb{P}_{\boldsymbol{\xi},k}$. There exist several representations of pp functions but we will focus on the B-splines decomposition, which uses the finite dimensional linear property of spaces of pp functions. The B-splines are a nearly orthogonal basis for pp functions which are defined (among others) by the following recurrence equation from a nondecreasing sequence of knots $\boldsymbol{\tau} = (\tau_i)_{i \in I}$ (I is a possibly infinite set of indices, τ may contain identical knots):

$$\forall i \in I,\ B_{i,1}(x) = 1_{[\tau_i, \tau_{i+1}]}(x)$$

and

$$\forall i \in I,\ \forall k \geq 2,\ B_{i,k}(x) = \frac{x - \tau_i}{\tau_{i+k-1} - \tau_i} B_{i,k-1}(x) + \frac{\tau_{i+k} - x}{\tau_{i+k} - \tau_{i+1}} B_{i+1,k-1}.$$

The B-splines possess the following useful properties:

Property 3.1.

1. *Compact support:* $\forall x \notin [\tau_i, \tau_{i+k}]$, $B_i(x) = 0$ *and* $\forall x \in]\tau_i, \tau_{i+k}[$, $B_i(x) > 0$.

[2] The jump of a function h at point ξ is $\Delta_\xi h = h(\xi^+) - h(\xi^-)$.

2. *Only the k B-splines $B_{j-k+1}, \ldots, B_j$ are nonzero on $[\tau_j, \tau_{j+1}]$.*

3. $\forall x \in [\tau_r, \tau_s]$, $\sum_i B_i(x) = \sum_{i=r+1-k}^{s-1} B_i(x) = 1$.

A spline function of order k with knot sequence $\boldsymbol{\tau}$ is defined as a linear combination of B-splines of order k for the knot sequence $\boldsymbol{\tau}$. The generated vector space is denoted $\mathbb{S}_{k,\boldsymbol{\tau}}$. The link between $\mathbb{S}_{k,\boldsymbol{\tau}}$ and $\mathbb{P}_{\boldsymbol{\xi},k,\nu}$ is given by the following theorem from Curry and Schoenberg, see De Boor p. 113.

Theorem 3.1.

For a given strictly increasing sequence $\boldsymbol{\xi} = (\xi_i)_{i=1,\ldots,\ell+1}$ and a given nonnegative integer sequence $\nu = (\nu_i)_{i=2,\ldots,\ell}$ with $\nu_i \le k$ we define

$$K \triangleq k + \sum_{i=2}^{\ell}(k - \nu_i) = k\ell - \sum_{i=2}^{\ell} \nu_i = \dim \mathbb{P}_{\xi,k,\boldsymbol{\nu}} \qquad (10)$$

and a nondecreasing sequence $\boldsymbol{\tau} = (\tau_i)_{i=1,\ldots,K+k}$ with $\tau_1 \le \cdots \le \tau_k \le \xi_1$ and $\xi_{\ell+1} \le \tau_{K+1} \le \cdots \le \tau_{K+k}$ and for each $i = 2, \ldots, \ell$, ξ_i appears exactly $k - \nu_i$ times in the sequence $\boldsymbol{\tau}$.

Then the sequence $B_1, \ldots, B_K$ of B-splines of order k for the knot sequence $\boldsymbol{\tau}$ is a basis for $\mathbb{P}_{\boldsymbol{\xi},k,\boldsymbol{\nu}}$ considered as functions on $[\tau_k, \tau_{K+1}]$.

Hence, fewer knots means more continuity and we have the equation:

$$\text{number of continuity conditions at } \xi + \text{number of knots at } \xi = k. \qquad (11)$$

In particular, a k-fold knot is a point with no continuity. At the opposite, no knot at a point enforces k continuity conditions so that the two polynomial pieces that meet at this point are identical. There are $\sum_{i=2}^{\ell}(k-\nu_i)$ additional interior knots τ_i between ξ_2 and ξ_ℓ and there are also k initial and k final knots outside the interval $[\xi_1, \xi_{\ell+1}]$. These $2k$ knots are usually chosen in the following convenient way $\tau_1 = \cdots = \tau_k = \xi_1$ and $\tau_{K+1} = \cdots = \tau_{K+k} = \xi_{\ell+1}$, which means that $\nu_1 = \nu_{\ell+1} = 0$. The B-representation of $f \in \mathbb{P}_{\boldsymbol{\xi},k,\nu}$ is given by K and k, the vector of knots $(\tau_i)_{i=1,\ldots,K+k}$ and the coefficients $(\alpha_i)_{i=1,\ldots,K}$ so that

$$\forall x \in [\tau_k, \tau_{K+1}],\ f(x) = \sum_{i=1}^{K} \alpha_i B_i(x).$$

Splines are mainly used for their approximation capability, which can be measured for instance in the sup norm sense with $\operatorname{dist}(g,\, \mathbb{S}_{k,\boldsymbol{\tau}}) = \inf_{s\in\mathbb{S}_{k,\boldsymbol{\tau}}} \|g-$

$s\|_\infty$, g being a smooth function. For a given knots sequence $\boldsymbol{\tau} = (\tau_1, \dots, \tau_{K+k})$ (in $[0,1]$), a usual $k-$th order spline approximation is defined thanks to a sequence of points t_i, $i = 1, \dots, K$, by

$$\forall t \in [0,1],\ Ag(t) = \sum_{i=1}^{K} g(t_i) B_i(t).$$

If g is C^m and $m \leq k-1$, we have for all $\boldsymbol{\tau} = (\tau_i)_{i=1..K+k}$ such as $\tau_1 = \dots = \tau_k = 0$ and $\tau_{K+1} = \dots = \tau_{K+k} = 1$

$$\operatorname{dist}(g,\, \mathbb{S}_{k,\boldsymbol{\tau}}) \leq C_{k,m} |\boldsymbol{\tau}|^m \|g^{(m)}\|_\infty \tag{12}$$

with $|\boldsymbol{\tau}|$ being the mesh size of the partition.

4 Spline-based estimators

We describe in this section a two-step method based on the scheme given in section 2, but in the particular case of splines estimators and using the least squares criterion R_n^2. From section 3, it appears that splines possess interesting computational properties and will result in rapid algorithms which is the main motivation of the present research. Moreover, this spline-based estimator is linear which enables one to derive rather straightforwardly the properties of the two-step estimator under broad and intuitive assumptions. Finally, splines are used in the collocation method [2] for the numerical integration of ODE's, so it links the statistical estimation and the approximation problems and it may provide a better insight into the statistical problem to the numerical analysis community.

4.1 Definition of the estimator

We have n observations $y_1, \dots, y_n$ corresponding to noisy observations of the solution of the ODE (1) at times $t_1, \dots, t_n$. We introduce Q_n the empirical distribution of the sampling times and we suppose that this empirical distribution converges to a distribution function Q (which possesses a density q). We construct our estimator $\hat{x}_n$ of x^* as a function in the spline space $\mathbb{S}_{k,\boldsymbol{\tau}}$[3]. We consider a breakpoint sequence $\boldsymbol{\xi} = (\xi_1, \dots, \xi_{L_n+1})$ of size L_n, so $\dim(\mathbb{S}_{k+1,\boldsymbol{\tau}}) = K_n = (k+1-\nu)L_n + \nu$ is allowed to depend on the number of observations (and the knots sequence $\boldsymbol{\tau}$ is of size $K_n + k + 1$). The

[3] If $\nu = 2$ for each internal breakpoint, the estimator is a C^1 function which is a minimal requirement for the solution of an ODE if F is continuous. In order to be coherent with the smoothness property of our model, we can take $\nu = m+1$, so that the function is C^m.

breakpoint sequence is chosen such that $\max_{1\leq i\leq L_n} |h_{i+1} - h_i| = o(L_n^{-1})$, $|\boldsymbol{\xi}|/\min_i h_i \leq M$ where $h_i = (\xi_{i+1} - \xi_i)$ and $|\boldsymbol{\xi}| = \sup_i h_i$ is the mesh size of $\boldsymbol{\xi}$. As a consequence, we have $|\boldsymbol{\xi}| = O(L_n^{-1})$. Like in Zhou *et al.* [22], we suppose that we have convergence of Q_n towards Q at a rate controlled by the mesh size, i.e.

$$\sup_{t\in[0,1]} |Q_n(t) - Q(t)| = o(|\boldsymbol{\xi}|). \tag{13}$$

The estimator $\hat{x}_n$ we consider is written componentwise

$$\forall i = 1, \ldots, d,\ \hat{x}_{n,i} = \sum_{k=1}^{K_n} c_{ik} B_k \tag{14}$$

or in matrix form $\hat{x}_n = C_n \mathbf{B}$ with the vector-valued function $\mathbf{B} = (B_1, \ldots, B_{K_n})^\top$ and the $d \times K_n$ coefficient matrix $C_n = (c_{ik}^n)_{1\leq i,k\leq d,K_n}$ (and column vectors $\mathbf{c}_{i,n} = (c_{i1}, \ldots, c_{iK_n})^\top \in \mathbb{R}^{K_n}$). We stress the fact that all the components $\hat{x}_{n,i}$ are approximated via the same space, although it may be inappropriate in some practical situations but it enables to keep simple expressions for the estimator. The fact that we look for a function in the vector space spanned by B-splines, puts emphasis on the regression interpretation of the first step of our estimating procedure. The estimation of the parameter C_n can be cast into the classical multivariate regression setting

$$\mathbf{Y}_n = \mathbf{B}_n C_n^\top + \epsilon_n, \tag{15}$$

where $\mathbf{Y}_n = (\mathbf{Y}_1 \ldots \mathbf{Y}_d)$ is the $n \times d$ matrix of observations, ϵ is the $n \times d$ matrix of errors, $C_n^\top$ is the $K_n \times d$ matrix of coefficients and $\mathbf{B}_n = (B_j(t_i))_{1\leq i\leq n, 1\leq j\leq K_n}$ is the design matrix. We look for a function close to the data in the L^2 sense, i.e. we estimate the coefficient matrix C_n by least-squares

$$\hat{\mathbf{c}}_{i,n} = \arg\min_{\mathbf{c}\in\mathbb{R}^{K_n}} \sum_{j=1}^{n} (y_{ij} - \mathbf{B}(t_j)^\top \mathbf{c})^2, i = 1, \ldots, d,$$

which gives the least squares estimator $\hat{\mathbf{C}}_n = (\mathbf{B}_n^\top \mathbf{B}_n)^+ \mathbf{B}_n^\top \mathbf{Y}$ where $(\cdot)^+$ denotes the Moore-Penrose inverse. We have

$$\forall i \in \{1, \ldots, d\}, \forall t \in [0,1],\ \hat{x}_{i,n}(t) = \mathbf{B}^\top(t) \hat{\mathbf{c}}_{i,n},$$

where $\hat{\mathbf{c}}_i = (\mathbf{B}_n^\top \mathbf{B}_n)^+ \mathbf{B}_n^\top \mathbf{Y}_i$. Finally, we introduce the projection matrix $P_{B,n} = \mathbf{B}_n (\mathbf{B}_n^\top \mathbf{B}_n)^+ \mathbf{B}_n^\top$. We will use the notation $x \lesssim y$ to denote that there exists a constant $M > 0$ such that $x \leq My$.

General results given by Huang in [9] ensure that $\hat{x}_n \xrightarrow{L^2} x^*$ in probability for sequences of suitably chosen approximating spaces $\mathbb{S}_{k+1,\boldsymbol{\xi},\nu}$ with an increasing number of knots. Indeed, corollary 1 in [9] enables us to claim that if the observation times are random with $Q(B) \geq c\lambda(B)$ ($0 < c \leq 1$ and $\lambda(\cdot)$ is the Lebesgue measure on $[0,1]$), the function x^* is in the Besov space $B^{\alpha}_{2,\infty}$ (with $k \geq \alpha - 1$) and the dimension grows such that $\lim_n \frac{K_n \log K_n}{n} = 0$ then

$$\frac{1}{n}\sum_{i=1}^{n}(y_{ij} - \hat{x}_i(t_j))^2 + \|\hat{x}_i - x_i^*\|_2 = O_P(\frac{K_n}{n} + K_n^{-2\alpha}).$$

Moreover, the optimal rate $O_P(n^{-2\alpha/(2\alpha+1)})$ (given by Stone [17]) is reached for $K_n \sim n^{1/(2\alpha+1)}$. For this nonparametric estimator, it is possible to construct a consistent two-step estimator $\hat{\theta}_n$ by minimization of $R_n^2(\theta)$.

4.2 Asymptotics

We give in this part the rate of convergence of the estimator $\hat{\theta}_n$. In order to derive the asymptotics, we use linearization techniques based on Taylor's expansion and we use the fact that the estimator depends linearly on the observations. We need to have a precise picture of the evolution of the basis $(B_1, \ldots, B_{K_n})$ as $K_n \to \infty$ and particularly the asymptotic behavior of the empirical covariance $G_{K_n,n} = \frac{1}{n}(\mathbf{B}_n^\top \mathbf{B}_n)$ and of the (theoretical) covariance $G_{K_n} = \int_0^1 \mathbf{B}(t)\mathbf{B}(t)^\top dQ(t)$. So we recall some useful technical properties on B-splines (of order k) when the knot sequence $\boldsymbol{\tau}$ is such that:

($*$) $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{n+k})$ is such that $\tau_1 = \cdots = \tau_k = 0$ and $\tau_{n+1} = \cdots = \tau_{n+k} = 1$, and $\tau_i < \tau_{i+k}$, $i = 1, \ldots, n$.

Property 4.1. *Under condition ($*$), for $k > 1, \exists\, c_0(k)$, $\forall s \in \mathbb{S}_{k,\boldsymbol{\tau}}$, s.t.*

$$c_0(k)\left(\sum_{i=1}^{K} a_i^2(\tau_i - \tau_{i-k})\right)^{1/2} \leq \left(\int_0^1 s^2(x)dx\right)^{1/2} \leq \left(\sum_{i=1}^{K} a_i^2(\tau_i - \tau_{i-k})\right)^{1/2}. \tag{16}$$

Inequalities (16) were derived by De Boor [2] in order to assess the well-posedness property of the B-splines basis. Zhou *et al.* [22] have given some refinements of these inequalities for the study of spline estimators for regression functions.

Property 4.2. *If $\boldsymbol{\tau}$ satisfies $(*)$, there exist constants $0 < c_1 \leq c_2 < \infty$ (independent of n or K_n) such that for any $s \in \mathbb{S}_{k,\boldsymbol{\tau}}$*

$$(c_1 + o_n(1))|\boldsymbol{\xi}| \left(\sum_{i=1}^{K_n} a_i^2 \right) \leq \int_0^1 s^2(x) dQ_n(x) \leq (c_2 + o_n(1))|\boldsymbol{\xi}| \left(\sum_{i=1}^{K_n} a_i^2 \right). \tag{17}$$

This property gives a bound on the eigenvalues of the covariance matrices:

Property 4.3. *If $\boldsymbol{\tau}$ satisfies $(*)$, there exist constants $0 < c_1 \leq c_2 < \infty$ (independent of n or K_n) such that for any $s \in \mathbb{S}_{k,\boldsymbol{\tau}}$*

$$(c_1 + o_n(1))|\boldsymbol{\xi}| \leq \lambda_{min} G_{K_n,n} \leq \lambda_{max} G_{K_n,n} \leq (c_2 + o_n(1))|\boldsymbol{\xi}|. \tag{18}$$

with $\lambda_{min} G_{K_n,n}$, $\lambda_{max} G_{K_n,n}$ being respectively the lowest and the highest eigenvalues of the empirical covariance matrix $G_{K_n,n}$ of the basis $(B_1, \ldots, B_{K_n})$.

Eventually, we have the following asymptotic behavior if $K_n = o(n)$

$$\forall t \in]0,1],\ \mathbf{B}^\top(t) \left(\mathbf{B}_n^\top \mathbf{B}_n \right)^{-1} \mathbf{B}(t) = \frac{1}{n} \mathbf{B}(t)^\top G_{K_n}^{-1} \mathbf{B}(t) + o(\frac{1}{n|\boldsymbol{\xi}|}). \tag{19}$$

We are interested now in the asymptotic behavior of $\Gamma(\hat{x}_n)$ where Γ is a linear functional $\Gamma(x) = \int_0^1 A(s)^\top x(s) ds$ with $s \mapsto A(s)$ a function in $C^m([0,1], \mathbb{R}^d)$. If $x = C^\top \mathbf{B}$, $\Gamma(x) = \sum_{i=1}^d \mathbf{c}_i^\top \boldsymbol{\gamma}_i = Trace(C^\top \boldsymbol{\gamma})$ with $\boldsymbol{\gamma}_i$ the columns of the $K \times d$ matrix $\boldsymbol{\gamma} = \int_0^1 \mathbf{B}(s) A^\top(s) ds$. Hence, the asymptotic behavior is derived directly from the asymptotics of $\hat{\mathbf{C}}_n$ and of matrix $\boldsymbol{\gamma}$. By using the results from Andrews [1], we will derive the asymptotic normality of this functional. For simplicity, we consider only the case $d = 1$, the extension to higher dimensions is cumbersome but straightforward. If the variance of the noise is σ^2, the variance of $\hat{\mathbf{c}}_n$ is

$$V(\hat{\mathbf{c}}_n) = \sigma^2 (\mathbf{B}_n^\top \mathbf{B}_n)^+ \tag{20}$$

and the variance of the estimator of the functional is

$$V_n = V(\Gamma(\hat{x}_n)) = \sigma^2 \boldsymbol{\gamma}_n^\top (\mathbf{B}_n^\top \mathbf{B}_n)^+ \boldsymbol{\gamma}_n. \tag{21}$$

Proposition 4.1.

Let $(\boldsymbol{\xi}_n)_{n\geq 1}$ be a sequence of knot sequences of length L_n and let K_n be the dimension of the associated spline spaces $\mathbb{S}_{k,\boldsymbol{\tau}_n}$. We suppose that $L_n \to \infty$ (equivalently $K_n \to \infty$ or $|\boldsymbol{\xi}_n| \to 0$) such that $n|\boldsymbol{\xi}_n| \to \infty$. If $\Gamma(x) = \int_0^1 A(s)x(s)ds$ with $A : [0,1] \to \mathbb{R}$ is C^m and x^* is C^{m+1} then:

(i) $\Gamma(\hat{x}_n) - \Gamma(x) = O_P(n^{-1/2})$ *and* $\sqrt{n}(\Gamma(\hat{x}_n) - \Gamma(x))$ *is asymptotically normal,*

(ii) $\forall t \in [0,1]$, $\hat{x}_n(t) - x(t) = O_P(n^{-1/2}|\boldsymbol{\xi}_n|^{-1/2})$,

(iii) $V(\hat{x}_n(t))^{-1/2}(\hat{x}_n(t) - x(t))$ *is asymptotically normal,* $t \in [0,1]$.

Proof. In order to prove the asymptotic normality of $\Gamma(\hat{x}_n) - \Gamma(x)$, we check the assumptions of theorem 2.1 of [1]. Assumption A is satisfied because the ϵ_i's are i.i.d. with finite variance. For assumption B, since A is C^m, the functional is continuous with respect to the Sobolev norm (or simply the sup norm). Moreover, it is possible to construct a spline $\tilde{A} = \sum_{i=1}^{K_n} \alpha_i B_i = \boldsymbol{\alpha}_n^\top \mathbf{B} \in \mathbb{S}_{k,\boldsymbol{\tau}_n}$ such that $\|A - \tilde{A}\|_\infty = O(|\boldsymbol{\xi}|^m)$ if $k \geq m$ (distance to spline space is given by (12)) and we have the approximation $|\Gamma_A(x) - \Gamma_{\tilde{A}}(x)| = |\int_0^1 (A - \tilde{A})(s)x(s)ds| \lesssim |\boldsymbol{\xi}|^m \|x\|_\infty$. Hence, it suffices to look at the case $A = \boldsymbol{\alpha}_n^\top \mathbf{B}$ because $\Gamma_A(x) - \Gamma_{\tilde{A}}(x)$ will tend to zero at faster rate than $n^{1/2}$. We introduce the vectors $\gamma_n = (\Gamma_A(B_1) \dots \Gamma_A(B_{K_n}))^\top$, so we have $\gamma_n^\top \gamma_n = \boldsymbol{\alpha}^\top G_{K_n}^\top G_{K_n} \boldsymbol{\alpha} \geq \lambda_{min}^2 G_{K_n} \times \|\boldsymbol{\alpha}\|_2^2$. From (18), we get $\gamma_n^\top \gamma_n \gtrsim |\boldsymbol{\xi}| \|\boldsymbol{\alpha}\|_2^2$. Inequality (17) ensures that $\gamma_n^\top \gamma_n$ is bounded away from 0 because

$$|\boldsymbol{\xi}| \|\boldsymbol{\alpha}\|_2^2 \gtrsim \int_0^1 A^2(s) dQ_n(s)$$

hence $\liminf_n \gamma_n^\top \gamma_n > 0$ and assumption B is checked.

From (19), we get the behavior of the diagonal entries of $P_{B,n}$:

$$\forall i \in [1..K_n],\ (P_{B,n})_{ii} = \frac{1}{n} \mathbf{B}(\xi_i)^\top G_{K_n}^{-1} \mathbf{B}(\xi_i) + o\left((n|\boldsymbol{\xi}|)^{-1}\right) \qquad (22)$$

we see that assumption C(ii) is true because $\mathbf{B}(\xi_i)^\top G_{K_n}^{-1} \mathbf{B}(\xi_i) \leq c_1 \|\mathbf{B}(\xi_i)\|_2^2 |\boldsymbol{\xi}|^{-1}$ and $\|\mathbf{B}(\xi_i)\|_2^2 \leq k$ (because the B-splines are bounded by 1 and only k of them are strictly positive) ensure that $\max_i (P_{B,n})_{ii} = O((n|\boldsymbol{\xi}|)^{-1}) \to 0$. It is clear that $\mathbf{B}_n$ is of full rank for n large enough.

We know from (12) that there exists a sequence $(\tilde{x}_n) \in \mathbb{S}_{k,\boldsymbol{\tau}_n}$ such that $\|\tilde{x}_n - x^*\|_\infty = O_P(|\boldsymbol{\xi}|^{m+1})$ hence

$$n^{1/2} \|\tilde{x}_n - x^*\|_\infty \to 0.$$

If we use again the spline approximation of the function A, we derive the following expression for

$$\gamma_n^\top \left(\frac{\mathbf{B}^\top\mathbf{B}}{n}\right)^{-1}\gamma_n = \boldsymbol{\alpha}^\top G_K^\top G_{K,n}^{-1} G_K \boldsymbol{\alpha}.$$

From (18) we have $\boldsymbol{\alpha}^\top G_K^\top G_{K,n}^{-1} G_K\boldsymbol{\alpha} \gtrsim \boldsymbol{\alpha}^\top G_K \boldsymbol{\alpha}$. As for $\gamma_n^\top\gamma_n$, we have

$$\liminf_n \gamma_n^\top \left(\frac{\mathbf{B}^\top\mathbf{B}}{n}\right)^{-1}\gamma_n > 0,$$

which remains true when A is any smooth function in C^m.

According to Andrews, we can conclude $V_n^{-1/2}(\Gamma(\hat{x}_n)-\Gamma(x^*)) \rightsquigarrow N(0,1)$. We obtain an equivalent of the rate of convergence by the same approximation as above

$$\begin{aligned} V_n &= \gamma_n^\top \left(\mathbf{B}^\top\mathbf{B}\right)^{-1}\gamma_n \\ &= \frac{1}{n}\boldsymbol{\alpha}^\top G_K G_{K,n}^{-1} G_K\boldsymbol{\alpha} \\ &\simeq \frac{1}{n}\boldsymbol{\alpha}^\top G_K\boldsymbol{\alpha} \end{aligned}$$

i.e. $V_n \sim \frac{\|\alpha\|^2|\boldsymbol{\xi}|}{n}$ by (18) and we obtain finally that $V_n \sim n^{-1}$.

The technique used by Andrews for his theorem 2.1 gives also asymptotic normality of $\hat{x}_n(t) = \mathbf{B}(t)^\top\hat{\mathbf{c}}_{i,n}$. We have then $\forall t \in [0,1]$, $V(\hat{x}_n(t)) = \sigma^2\mathbf{B}(t)^\top(\mathbf{B}_n^\top\mathbf{B}_n)^+\mathbf{B}(t)$ and from (19) we get $V(\hat{x}_n(t)) = \frac{\sigma^2}{n}\mathbf{B}(t)^\top G_{K_n}^{-1}\mathbf{B}(t) + o(\frac{1}{n|\boldsymbol{\xi}|})$, so that $V(\hat{x}_n(t)) \sim \frac{C}{n|\boldsymbol{\xi}|}$ from lemma 6.6 in [22].

□

By linearizing the criterion R_n^2, we show that the two-step estimator is a simple functional of the spline estimator. We introduce the differentials of F at (x,θ) w.r.t. θ and x and we denote them $D_\theta F(x,\theta)$ and $D_x F(x,\theta)$ respectively. For short, we adopt the following notation for the functions: $D_\theta F(x^*,\theta^*) = D_\theta F^*$ and $D_x F(x^*,\theta^*) = D_x F^*$.

Theorem 4.1.

Let F a C^m vector field w.r.t (θ,x) $(m \geq 1)$, such that $D_\theta F$ and $D_x F$ are Lipschitz w.r.t (θ,x). We suppose that the Hessian J^* of the asymptotic criterion $R^2(\theta)$ evaluated at θ^* is nonsingular. We suppose that the conditions of proposition 2.1 are satisfied and that the knots of the spline estimators are such that $n|\boldsymbol{\xi}_n| \to 0$, then the two-step estimator $\hat{\theta}_n = \arg\min R_n^2(\theta)$ is:

(i) *asymptotically normal,*

(ii) $(\hat{\theta}_n - \theta^*) = O_P(n^{-1/2}|\boldsymbol{\xi}_n|^{-1/2})$.

Moreover, the optimal rate of convergence for the Mean Square Error is obtained for $K_n = O(n^{1/(2m+3)})$ *and we have then* $(\hat{\theta}_n - \theta^*) = O_P(n^{-(m+1)/(2m+3)})$.

Proof. $\nabla_\theta R_n^2(\hat{\theta}_n) = 0$ implies that $\int_0^1 \left(D_\theta F(\hat{x}_n(t), \hat{\theta}_n)\right)^\top (\dot{\hat{x}}_n(t) - F(\hat{x}_n(t), \hat{\theta}_n))dt = 0$. We remove dependence on t and n for notational convenience and introduce F^* and $F(\hat{x}, \theta^*)$ which gives

$$\int_0^1 \left(D_\theta F(\hat{x}, \hat{\theta})\right)^\top ((\dot{\hat{x}} - \dot{x^*}) + F^* - F(\hat{x}, \theta^*) + F(\hat{x}, \theta^*) - F(\hat{x}, \hat{\theta}))dt = 0$$

and

$$\int_0^1 \left(D_\theta F(\hat{x}, \hat{\theta})\right)^\top \left((\dot{\hat{x}} - \dot{x^*}) + D_x F(\widetilde{x^*}, \theta^*)(x^* - \hat{x}) + D_\theta F(\hat{x}, \widetilde{\theta^*})(\theta^* - \hat{\theta})\right) dt = 0$$

with $\widetilde{x^*}$ and $\widetilde{\theta^*}$ being random points between x^* and $\hat{x}$, and θ^* and $\hat{\theta}$ respectively. We introduce $\hat{A} = D_\theta F(\hat{x}, \hat{\theta})$, and an asymptotic expression for $(\theta^* - \hat{\theta})$ is

$$\begin{aligned}(\theta^* - \hat{\theta}) \int_0^1 \hat{A}^\top D_\theta F(\hat{x}, \theta^*)dt &= -\int_0^1 \hat{A}^\top (\dot{\hat{x}} - \dot{x^*})dt \\ &\quad - \int_0^1 \hat{A}^\top D_x F(\widetilde{x^*}, \theta^*)(x^* - \hat{x})dt.\end{aligned}$$

It suffices to consider the convergence in law of the random integral $H_n = \int_0^1 (D_\theta F^*)^\top \left((\dot{\hat{x}} - \dot{x^*}) + D_x F^*(x^* - \hat{x})\right) dt$ because the random variable

$$G_n = \int_0^1 \hat{A}^\top \left((\dot{\hat{x}} - \dot{x^*}) + D_x F(\widetilde{x^*}, \theta^*)(x^* - \hat{x})\right) dt$$

is such that $G_n - H_n \to 0$ in probability (in the L^2 sense), moreover we have the convergence in probability of $\int_0^1 \hat{A}^\top D_\theta F(\hat{x}, \theta^*)dt$ to J^*.

Indeed, we consider the map $\mathcal{D} : (x, \theta) \mapsto (t \mapsto D_\theta F(x(t), \theta))$ defined on $C([0,1], \mathcal{K}) \times \Theta$ (with the product Hilbert metric) with values in $C([0,1], \mathbb{R}^{n \times p})$ (with the L^2 norm $\|A\|_2 = \int_0^1 Tr(A^\top(t)A(t))dt)$. Since $D_\theta F$ is Lipschitz in (x, θ), the functional map $\mathcal{D}$ is a continuous map, and we can claim by the continuous mapping theorem that the random functions

$t \mapsto D_\theta F(\hat{x}(t), \hat{\theta})$ and $t \mapsto D_\theta F(\hat{x}(t), \theta^*)$ converge in probability (in the L^2sense) to $D_\theta F^*$. As a consequence, $\left\| D_\theta F(\hat{x}, \hat{\theta}) \right\|_2$ converges in probability to $\|D_\theta F^*\|_2$ so it is also bounded, and $\|D_\theta F(\hat{x}, \theta^*) - D_\theta F^*\|_2 \to 0$ in probability. This statement is also true for all entries of these (function) matrices, which enables to claim that all entries of the matrix

$$\int_0^1 \left(D_\theta F(\hat{x}, \hat{\theta}) \right)^\top (D_\theta F(\hat{x}, \theta^*) - D_\theta F^*) \, dt$$

tend to zero in probability (by applying the Cauchy-Schwarz inequality componentwise). Moreover, we have convergence in probability of each entry of $\int_0^1 \left(D_\theta F(\hat{x}, \hat{\theta}) \right)^\top D_\theta F^* dt$ to the corresponding entry of $\int_0^1 (D_\theta F^*)^\top D_\theta F^* dt$ (consequence of the convergence of $D_\theta F(\hat{x}, \hat{\theta})$ to $D_\theta F^*$ in the L^2 sense), which implies finally that

$$\int_0^1 \left(D_\theta F(\hat{x}, \hat{\theta}) \right)^\top D_\theta F(\hat{x}, \theta^*) dt \xrightarrow{P} J^*$$

By the same arguments and by using the fact that $D_x F$ is also Lipschitz in (x, θ), we have convergence of the matrix $G_n - H_n$ to 0 in probability. The asymptotic behavior of $(\hat{\theta}_n - \theta^*)$ is then given by the random integral

$$J^{*-1} \int_0^1 (D_\theta F^*)^\top \left((\dot{\hat{x}} - \dot{x^*}) + D_x F^* (x^* - \hat{x}) \right) dt. \tag{23}$$

We can write it also as $\Gamma(\hat{x}) - \Gamma(x^*)$ by introducing the $\mathbb{R}^d$ -valued continuous linear functional defined by

$$\Gamma(x) = \int_0^1 \left(B(s) - \frac{d}{ds} A(s) \right) x(s) ds + B(1)x(1) - B(0)x(0)$$

with $s \mapsto A(s) = D_x F(x^*(s), \theta^*)^\top$ and $s \mapsto B(s) = A(s) D_\theta F(x^*(s), \theta^*)$ being (at least) C^{m-1}functions. From proposition 4.1, we may claim the asymptotic normality $A_n(\Gamma(\hat{x}_n) - \Gamma(x^*)) \rightsquigarrow N(0, I_d)$ where A_n is a properly chosen normalizing sequence (normality is extended from scalar functional to multidimensional functional by the Cramér-Wold device). Moreover, we know that $\Gamma(\hat{x}_n) - \Gamma(x^*) = O_P(n^{-1/2}) + O_P(n^{-1/2} |\boldsymbol{\xi}_n|^{-1/2}) = O_P(n^{-1/2} |\boldsymbol{\xi}_n|^{-1/2})$, because the estimation of $x^*(t)$ is done at a slower rate.

Now to determine the optimal rate of convergence in the mean square sense, we need to use the Bias - Variance decomposition for the evaluation functional $\|\hat{\theta}_n - \theta^*\|^2 = O_P\left((E(\hat{x}_n(t)) - x(t))^2 \right) + O_P(Var(\hat{x}_n(t)))$.

Theorem 2.1 of [22] gives $E(\hat{x}_n(t)) - x^*(t) = O(|\boldsymbol{\xi}_n|^{m+1})$ (because x^* is C^{m+1}) and $Var(\hat{x}_n(t)) = O_P(n^{-1}|\boldsymbol{\xi}_n|^{-1})$ so the optimal rate is reached for $|\boldsymbol{\xi}_n| = O(n^{-1/(2m+3)})$ and is $O(n^{-(2m+2)/(2m+3)})$.

□

Remark 4.1.

The asymptotic result given for the deterministic observational times $0 \leq t_1 < \cdots < t_n \leq 1$ *remains true when they are replaced by realizations of some random variables* $T_1, \ldots, T_n$ *as long as the assumptions of the two previous propositions are true with probability one. Andrews gives some conditions (theorem 2) in order to obtain this. It turns out that in the case of* $T_1, \ldots, T_n$ *i.i.d. random variables drawn from the distribution* Q*, it suffices to have* $K_n^4 \lesssim n^r$ *with* $0 < r < 1$*. In particular, as soon as* $m \geq 1$*, the conclusion of proposition 4.1 holds with probability one for the optimal rate* $K_n = n^{1/(2m+3)}$.

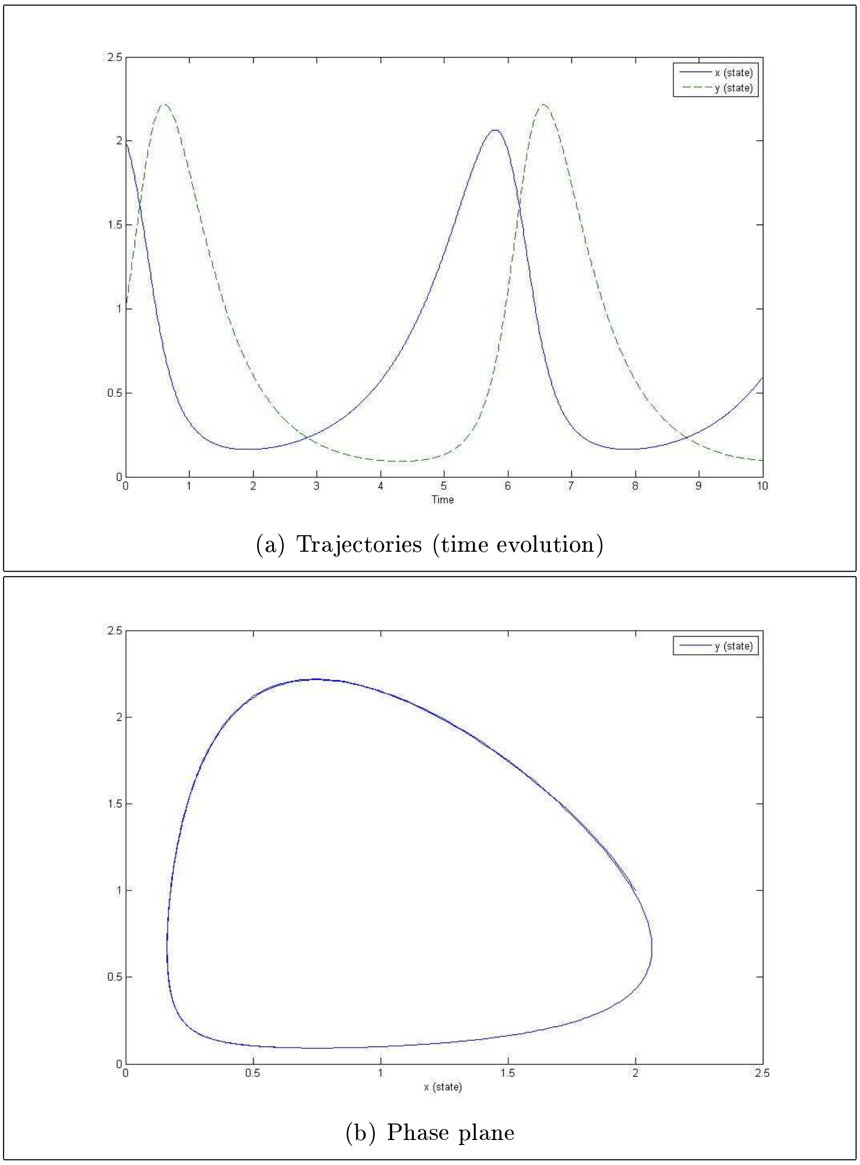
5 Experiments

The Lotka-Volterra equation is a standard model for the evolution of prey-predator populations. It is a planar ODE

$$\begin{cases} \dot{x}_1 & = ax_1 - bx_1x_2 \\ \dot{x}_2 & = -cx_2 + dx_1x_2 \end{cases} \tag{24}$$

whose behavior is well-known [8]. Despite its simplicity, it exhibits convergence to limit cycles which is one of the main features of nonlinear dynamical systems, which has usually a meaningful interpretation. Due to this simplicity and the interpretability of the solution, it is often used in biology (population dynamics or phenomena with competing species), but the statistical estimation of the parameter $\theta = (a,\, b,\, c,\, d)^\top$ has not been extensively addressed. Nevertheless, Varah (1982) illustrates spline-based method (with natural cubic splines and knots chosen by an expert) on the same model as (24). Froda *et al.* (2005) [7] have considered another original method exploiting the fact that the orbit may be a closed curve for some values of the parameters.

For this benchmark example, we study the behavior of the two-step estimator corresponding to the criterion $R_n^2(\theta)$. A challenging problem in the construction of the estimator is the usual problem of the selection of the number of knots during the spline estimation (which was left to the practitioner in Varah's paper). A similar problem arises also in Ramsay's method based on smoothing splines where one has to choose properly the trade-off constant λ during the minimization of the penalized fitting criterion $\sum_{i=1}^{n} |y_i - \hat{x}(t_i)|_2^2 + \lambda \|\dot{\hat{x}} - F(\hat{x}, \theta)\|_2^2$. The classical optimal value given by cross-validation (chapter 3 [21]) is not directly relevant in this case, so it is also a parameter left to the modeler. The nonparametric estimation relies on the choice of the sequence of knots, and we take a uniform grid $\boldsymbol{\tau}$ here. Nevertheless, the present result (theorem 4.1) is not practical and does not enable us to select a correct number of knots. As suggested before, one can think of an extension of the celebrated GCV, but in our setting this problem of knots selection seems more naturally dealt by the free-knot splines [3, 18]. We do not propose here a knots selection procedure for the practioner, but an adhoc one, based on the ability to approximate the function of interest by splines. In order to do this, we study and choose an arbitrary nondecreasing sequence of number of knots K_n by graphical arguments relying on the approximation of x^* by its L^2 projection on $\mathbb{S}_4^{K_n}$ the space of cubic splines that are C^2 with K_n uniformly spaced knots. The projection is denoted by $P_{K_n} x^*$.



(a) Trajectories (time evolution)

(b) Phase plane

Figure 1: Solution of Lotka-Volterra system in the phase plane.

In our experiment, we consider the system with $a = 1$, $b = 1.5$, $c = 1.5$, $d = 2$ and $x(0) = 1$, $y(0) = 2$. As shown in Figure 1, the solution is attracted by a periodic solution and is observed on the time interval $[0, 10]$ which corresponds roughly to 2 periods (and the trace in the phase plane is nearly a

n	50	100	200	400	600	800	1000	1200	1400	1600	1800	2000	5000
K_n	9	9	12	13	14	15	16	17	18	19	20	20	20
$R_n^2(\tilde{\theta}_{K_n})$	12	6.8	5.9	5.4	3.4	2.9	2.9	2.0	2.2	2.0	1.9	1.7	0.7

Table 1: Number of knots and minima of the criterion R_n^2

closed curve). With a Monte-Carlo study (based on $N_{mc} = 1000$ independent drawings), we show the asymptotic properties of the two-step estimator in the case of a homoscedastic Gaussian noise with $\sigma = 0.4$ ($y_i = x^*(t_i) + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma^2 I_2)$) by considering different sample size $n = 50$, 100, 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000 and 5000 (the observation times are uniformly drawn in $[0, 10]$).

The L^2 distance between the solution x^* and its spline approximation is diminishing with the number of the knots K_n but this is not monotone as we can see from figure 2 (a), because the spaces $\mathbb{S}_4^K$, $K > 1$ are not nested. We introduce $\tilde{\theta}_{K_n}$ the minimizer of the criterion $R_n^2(\theta)$ computed with $P_{K_n}x^*$ and we give the evolution of $\|\tilde{\theta}_n - \theta^*\|^2$ in Figure 2 (b). This is another way to evaluate the convergence of the criterion R_n^2 to the asymptotic one R^2 (and in fact this is a most important characteristic of the convergence of an M-estimator). A striking feature is that the dimension of K_n is not an indicator of the quality of approximation: for instance, we have a good approximation of x^* by $P_{K_n}x^*$ for $K_n = 12$ (local minimum of the curve in Figure 2 (a)) which is better than for $K_n = 13$ and we have $\tilde{\theta}_{K_n} \approx \theta^*$. Despite this peculiar behavior, Figure 2 shows that for $K_n \geq 20$, we have a very good estimate of x^* and $\tilde{\theta}_{K_n} \approx \theta^*$, moreover there is no noticeable difference between $K_n = 20, \ldots, 50$. Hence, the discrepancy between $\tilde{\theta}_{K_n}$ and θ^* will introduce a bias in the estimation, which is finally a by-product of the bias in the nonparametric regression. The selected number of knots and the mean values of the selected minima $\min_\theta R_n(\theta)$ are shown in Table 1.

The choice of the knots is done according to the features enhanced in equation (2): indeed, we choose the knots according to the approximating

n	Number of knots	Mean $(\hat{a}, \hat{b}, \hat{c}, \hat{d})$	Standard deviation
50	9	(0.7, 1.22, 1.35, 1.81)	(0.29, 0.31, 0.39, 0.48)
100	9	(0.73, 1.28, 1.50, 1.99)	(0.20, 0.21, 0.27, 0.35)
200	12	(0.93, 1.46, 1.41, 1.92)	(0.17, 0.18, 0.20, 0.26)
400	13	(0.88, 1.36, 1.32, 1.77)	(0.12, 0.11, 0.13, 0.18)
600	14	(0.88, 1.39, 1.51, 2.02)	(0.10, 0.10, 0.13, 0.18)
800	15	(0.98, 1.49, 1.42, 1.93)	(0.10, 0.10, 0.11, 0.15)
1000	16	(0.91, 1.41, 1.43, 1.91)	(0.08, 0.08, 0.10, 0.14)
1200	17	(0.98, 1.48, 1.50, 2.03)	(0.08, 0.08, 0.09, 0.13)
1400	18	(0.98, 1.48, 1.44, 1.94)	(0.08, 0.07, 0.09, 0.13)
1600	19	(0.96, 1.46, 1.47, 1.97)	(0.08, 0.07, 0.09, 0.13)
1800	20	(1.00, 1.50, 1.48, 1.98)	(0.07, 0.07, 0.08, 0.12)
2000	20	(1.00, 1.50, 1.48, 1.98)	(0.07, 0.06, 0.08, 0.12)
5000	20	(1.00, 1.49, 1.49, 1.99)	(0.05, 0.04, 0.05, 0.08)

Table 2: Mean and standard deviation of the two step estimator $\hat{\theta}_n$

power of the corresponding spline space. The leading principle is to avoid a small space or a space that behaves worse than a smaller one. Hence we do not take $K_n \leq 8$ because the distance between $\mathbb{S}_4^K$ and x^* is too big. Moreover, we do not use neither $K_n = 10, 11$ because the approximation is worse than with $K_n = 9$ (the same for $K_n = 13$).

The computation of $\hat{\theta}_n$ is done by a Nelder-Mead algorithm starting from the true value $\theta^* = (1, 1.5, 1.5, 2)^\top$: this enables us to stay in a relevant part of the parameter space, hence we can avoid the bias due to the numerical determination of the estimator. Nevertheless, this local algorithm gives spurious minima in less than 1% of the simulations because of the spikiness of the function R_n^2: in this case the results are discarded from the statistics. The performance of the estimator (mean and standard deviation) is gathered in table 2, which illustrates the convergence in quadratic mean of the estimator.

We check the asymptotic normality of the estimator by performing a Kolmogorov Smirnov (KS) test on each component of the standardized residuals $r_n = \Sigma_n^{-1}(\hat{\theta}_n - \theta^*)$. We control also the normality of the biased residuals $r_n^b = \Sigma_n^{-1}(\hat{\theta}_n - \bar{\theta}_n)$ where $\bar{\theta}_n = \frac{1}{N_{mc}} \sum_{i=1}^{N_{mc}} \hat{\theta}_n^{(i)}$ and $\hat{\theta}_n^{(i)}$ is the estimate of the i^{th} Monte Carlo simulation, and we store the p-values of the KS test for the

n	$P(U > ks(r_n))$	$P(U > ks(r_n^b))$
50	(0, 0, 0, 0)	(0.22, 0.0009, 0.022, 0.0048)
100	(0, 0, 0, 0)	(0.13, 0.08, 0.11, 0.35)
200	(0, 0, 0, 0)	(0.54, 0.05, 0.98, 0.87)
400	(0, 0, 0, 0)	(0.85, 0.18, 0.99, 0.62)
600	(0, 0, 0, 0.74)	(0.61, 0.59, 0.65, 0.84)
800	(0, 0, 0, 0.0001)	(0.75, 0.26, 0.81, 0.31)
1000	(0, 0, 0, 0)	(0.46, 0.01, 0.73, 0.57)
1200	(0, 0, 0, 0)	(0.50, 0.97, 0.95, 0.62)
1400	(0, 0, 0, 0.0001)	(0.74, 0.21, 0.76, 0.86)
1600	(0, 0, 0, 0.0921)	(0.65, 0.87, 0.18, 0.27)
1800	(0.7, 0.3, 0, 0.33)	(0.21, 0.98, 0.71, 0.28)
2000	(0.97, 0.08, 0, 0.74)	(0.98, 0.59, 0.35, 0.85)
5000	(0.33, 0, 0, 0.29)	(0.97, 0.45, 0.30, 0.12)

Table 3: P-values of the Kolmogorov-Smirnov (componentwise) test for asymptotic normality (U is the Kolmogorov-Smirnov statistic). In this table, 0 means lower than 10^{-4} and values lower than 0.05 implies rejection of the normality assumption with 95% confidence.

two residuals in Table 3 (we denote $ks(r_n)$ and $ks(r_n^b)$ the values of the KS statistic). We may conclude from Table 3 that the convergence to normality of the residuals r_n is quite slow and is not attained for n as big as 5000 (but it is true for 2 components as soon as $n \geq 1800$). In fact, this is partly due to the KS test we use, because it is clear from table 2 that the bias tends to zero; nevertheless, the difference between the Monte Carlo sample and the true parameter remains significant (despite it is less than 0.1) because we have a huge sample size. Indeed, the rejection of the normality of the estimator stems from the bias, and we can verify that we have asymptotic normality of the estimator by applying KS test for r_n^b. Moreover, the normality is rapidly reached, since the normality assumption cannot be rejected as soon as $n \geq 100$ (most of the p-values are indeed greater than 0.05).

6 Conclusion

We have proposed a new family of parametric estimators of ODE's relying on nonparametric estimators, which are simpler to compute than straightforward parametric estimators such as MLE or LSE. The construction of this parametric estimator puts emphasis on the regression interpretation of the ODE's estimation problem, and on the link between a parameter of the ODE and an associated function. By using an intermediate functional proxy, we expect to gain information and precision on likely value of the parameters. We do not have studied the effect of using shape or inequality constraints of the estimator $\hat{x}_n$ but it might be valuable information for the inference of complex models, either by shortening the computation time (it gives more suitable initial conditions) or by accelerating the rate of convergence of the estimator thanks to restriction to smaller sets of admissible parameter values.

We have particularly studied the case $R_n^2(\theta)$, but other M-estimators such as the one obtained from $R_n^1(\theta)$ may possess interesting theoretical and practical properties such as robustness. This could be particularly useful in the case of noisy data which can give oscillating estimates of the derivatives of the function.

We have only considered spline-based estimators, we have derived the asymptotic normality of the two-step estimator, and we have determined the optimal rate as $n^{-(m+1)/(2m+3)}$ which is obtained for an appropriately growing sequence of knots. We have touched on the effective selection of the number of knots in section 5 and a necessary theoretical and practical development is the construction of a data-driven methodology to determine the number of knots. A more general problem of knots selection might be addressed by the use of a free-knots spline estimator where the number and the location of the knots is determined from the data [3, 18]. This type of estimator is much more flexible and may help in reducing the observed bias in the experiments for small n. Eventually, our two-step estimator can be improved to a $\sqrt{n}$-consistent and even a parametrically efficient estimator by additional steps. This will be pursued within a more general framework elsewhere.

Acknowledgement. *The author would like to thank Chris Klaassen (EURANDOM and Universiteit van Amsterdam) for giving many comments who greatly improved the quality of this paper and Florence d'Alché-Buc (Université d'Evry, France) for giving interest in the estimation of ODE's from noisy data. The paper was written during a visit at EURANDOM partly supported by the PASCAL network.*

References

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

345, 1991.

[2] C. de Boor. *A practical guide to splines*, volume 27. Springer-Verlag, 1979.

[3] I. Dimatteo, C.R. Genovese, and R.E. Kass. Bayesian curve-fitting with free-knot splines. *Biometrika*, 88(4):1055–1071, 2001.

[4] S. Donnet and A. Samson. Estimation of parameters in incomplete data models defined by dynamical systems. 2006. Submitted to Journal of Statistical Planning and Inference.

[5] C.P. Fall, E.S. Marland, J.M. Wagner, and J.J. Tyson, editors. *Computational Cell Biology.* Interdisciplinary applied mathematics. Springer, 2002.

[6] J. Fan and Y.K. Truong. Nonparametric regression with errors in variable. *Annals of Statistics*, 21(4):1900–1925, 1993.

[7] S. Froda and G. Colavita. Estimating predator-prey systems via ordinary differential equations with closed orbits. *Australian & New-Zealand Journal of Statistics*, 47(2):235–254, 2005.

[8] M.W. Hirsch, S. Smale, and R. Devaney. *Differential equation, Dynamical Systems and an Introduction to Chaos*, volume 60 of *Pure and Applied Mathematical series.* Elsevier Academic Press, 2nde edition edition, 2003.

[9] J.Z. Huang. Asymptotics for polynomial spline regression under weak conditions. *Statistics and probability letters*, 65:207–216, 2003.

[10] J-Y Koo and K-W Lee. B-splines estimation of regression functions with errors in variable. *Statistics and probability letters*, 40:57–66, 1998.

[11] Y.A. Kuznetsov. *Elements of Applied Bifurcation Theory.* Springer-Verlag, New York, 2004.

[12] N. Lalam and C. Klaassen. Pseudo-maximum likelihood estimation for differential equations. Technical Report 2006-18, Eurandom, 2006.

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

[14] 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.

[15] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis.* Springer series in statistics. Springer, 1997.

[16] J.O. Ramsay, G. Hooker, J. Cao, and D. Campbell. Estimating differential equations. *submitted*, 2006. (available from http://ego.psych.mcgill.ca/misc/fda/DIFEpaper.pdf).

[17] C.J. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10(4):1040–1053, 1982.

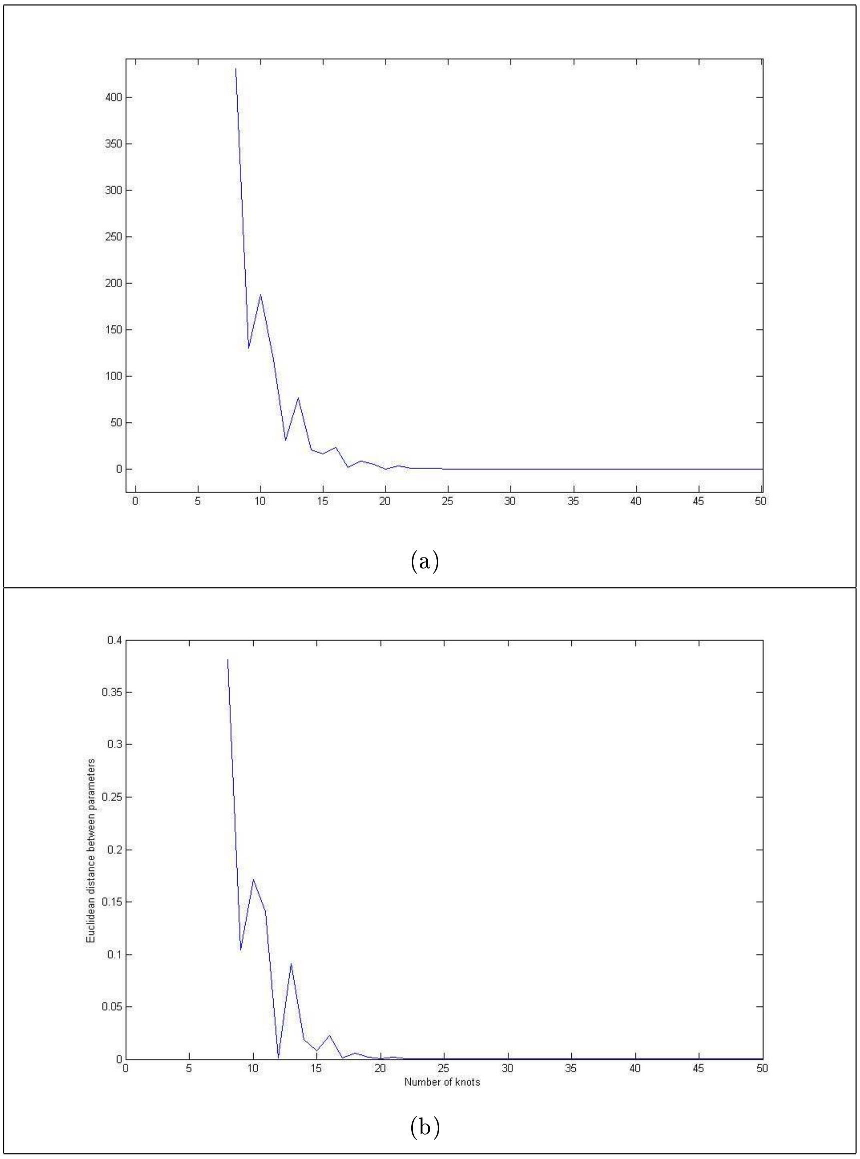
[18] C.J. Stone and J.Z. Huang. Free knot splines in concave extended linear modeling. *Journal of Statistical Planning and Inference*, 108:219–253, 2002.

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

[20] 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.

[21] G. Wahba. *Spline models for observational data*, volume 59. SIAM, 1990.

[22] S. Zhou, X. Shen, and D.A. Wolfe. Local asymptotics for regression splines and confidence regions. *Annals of Statistics*, 26(5):1760–1782, 1998.



(a)

(b)

Figure 2: (a) L^2 Distance between x^* and its spline approximations (w.r.t the number of knots)
(b) Euclidean distance between $\tilde{\theta}_n$ and θ^*