A Tracking Approach to Parameter Estimation in Linear Ordinary Differential Equations

Nicolas J-B. Brunel, Quentin Clairon

A Tracking Approach to Parameter Estimation in Linear Ordinary Differential Equations

Nicolas J-B. Brunel, Quentin Clairon

ENSIIE & Laboratoire de Mathématiques et Modélisation d'Evry, UMR CNRS 8071, Université d'Evry, France

Abstract

Ordinary Differential Equations are widespread tools to model chemical, physical, biological process but they usually rely on parameters which are of critical importance in terms of dynamic and need to be estimated directly from the data. Classical statistical approaches (nonlinear least squares, maximum likelihood estimator) can give unsatisfactory results because of computational difficulties and ill-posedness of the statistical problem. New estimation methods that use some nonparametric devices have been proposed to circumvent these issues. We present a new estimator that shares properties with Two-Step estimator and Generalized Smoothing (introduced by Ramsay et al. [34]). We introduce a perturbed model and we use optimal control theory for constructing a criterion that aims at minimizing the discrepancy with data and the model. Here, we focus on the case of linear Ordinary Differential Equations as our criterion has a closed-form expression that permits a detailed analysis. Our approach avoids the use of a nonparametric estimator of the derivative, which is one of the main cause of inaccuracy in Two-Step estimators. Moreover, we take into account model discrepancy and our estimator is more robust to model misspecification than classical methods. The discrepancy with the parametric ODE model correspond to the minimum perturbation (or control) to apply to the initial model. Its qualitative analysis can be informative for misspecification diagnosis. In the case of well-specified model, we show the consistency of our estimator and that we reach the parametric $\sqrt{n}-$ rate when regression splines are used in the first step.

1 Introduction

We consider a dynamical process defined by an Ordinary Differential Equation (ODE) with a known and fixed initial value

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

Such a model is called an Initial Value Problem (IVP). The state x is in $\mathbb{R}^d$ and θ is an unknown parameter, that belongs to a subset Θ of $\mathbb{R}^p$. f is a time-dependent vector field from $[0,T] \times \mathbb{R}^d \times \Theta$ to $\mathbb{R}^d$. This class of dynamical models are commonly used in physics, engineering, ecology,... [13, 30, 12, 17]. Let $t \mapsto X_{\theta^*}(t) = X^*(t)$ be the solution to the IVP (1) on $[0,T]$, for the true parameter set θ^*.

We want to estimate θ^* from noisy observations Y_i, $i = 1,\dots,n$ of the trajectory X^*, made at time t_i. Estimation can be done by classical estimators such as Nonlinear Least Squares (NLS), Maximum Likelihood Estimator (MLE) [27] or Bayesian approaches ([21],[14],[6] and [15] for example). Nevertheless, the statistical estimation of an ODE model by NLS leads to a difficult nonlinear estimation problem. These difficulties were pointed out by Ramsay et al. [34]: computational complexity comes from repeated ODE integrations required by the optimization algorithm; moreover, the usual criterion exhibits multiple local minima . Even though meta-heuristic methods can be proposed to circumvent the last issue, parameter estimation can be considered as an ill-posedness inverse problem [11], that needs alternatives to the classic statistical approaches.

Alternative statistical estimators have been developped or adapted to this particular framework, such as hierarchical Bayesian approaches [21, 33] and MCMC ([16]), Generalized Smoothing [34, 32, 10, 7] or Two-Step estimators [4, 5, 28, 18]. Recently variational approaches have also been developed [23].

As other two-step estimators, our method produces a minimum distance estimator [25] but it shares strong links with Generalized Smoothing approaches. Two-Step estimators were initiated by [39] and aims at minimizing a discrepancy measure between a nonparametric estimator $\hat{X}$ and quantities characterizing the differential models. Usually, a Two-Step method is defined by the following procedure:

1. Construct a nonparametric curve estimator $\widehat{X}$ from the data $(t_i, Y_i)_{1 \le i \le n}$,
2. Compute a model discrepancy measure $R(\hat{X}, \theta)$, such as the weighted L^2 distance
$$R(\hat{X}, \theta) = \int_0^T \left\| \dot{\hat{X}}(t) - f(t, \widehat{X}(t), \theta) \right\|^2 w(t) dt \tag{2}$$
3. Define the parameter estimator
$$\widehat{\theta}^{TS} = \arg\min_{\theta \in \Theta} R(\hat{X}, \theta) \tag{3}$$

These estimators have a good computational efficiency vs NLS and they avoid repeated ODE integration. In practice, the used criteria are also smoother and easier to optimize than the NLS criterion. Two-step estimators are consistent in general, but there is a trade-off with the statistical precision, and some care in the use of nonparametric estimate $\dot{\hat{X}}$ has to be taken in order to keep the parametric rate [4, 18]. The variance of TS estimator can be higher, but the use of other criterion, for instance based on a weak formulation of the differential

equation can give competitive alternatives, in particular in high dimensional parameter space or small sample size, see [5].

In the case of Generalized Smoothing [34], the solution X^* is approximated by a basis expansion that solves approximately the ODE model; hence, the parameter inference is performed by dealing with an imperfect model, as the collocation approximation of the ODE solution can be seen as a relaxation of the ODE model constraint, needed for taking into account some uncertainty about the model. Based on the Generalized Profiling approach, Hooker proposed a criteria that estimates the lack-of-fit through the estimation of a "forcing function" $t \mapsto u(t)$ in the ODE $\dot{x} - f(t, x, \tilde{\theta}) = u(t)$, where $\tilde{\theta}$ is a previous estimate obtained by Generalized Profiling [5]. The objective of this paper is to provide a parameter estimate and an approximate solution to the ODE that

- avoids the use a nonparametric estimate of the derivative $\dot{X}$ as in two step estimators,
- incorporates robustness in model specification and controls the quality of approximation,
- introduces the use of infinite dimensional optimization tools, exploiting the differential structure of the model.

One interest of the latter point is to avoid the use of series expansion for function estimation, and avoid some arbitrary practical choices about the basis. Moreover, infinite dimensional optimization tools give powerful characterization of the solutions that gives an additional insight in Generalized Smoothing.

Our method provides a consistent parametric estimator when the model is correct. We show that it is root-n consistent and asymptotically normal. At the same time, we get a discrepancy measure between the model and the data under the form of an optimal control u analogous to the forcing function in [19].

In the next section, we introduce the notations and we motivate our approach by discussing the Generalized Smoothing approach, and the link with Optimal Control Theory. In section 3, we show that the estimator is consistent under some regularity assumption about the model. Then in section 4, we show that we reach the root$-n$ rate using regression splines for $\widehat{X}$. Finally, we show the interest of our method on a toy model (and we make comparison with Nonlinear Least Squares and Generalized Smoothing) and we consider also a real data case, where a linear ODE is used for describing the isomerization reaction of α-Pinene.

2 Model and methodology

We introduce our statistical framework, and we recall the mechanics of the Generalized Smoothing estimator in the particular context of a linear ODE.

2.1 The statistical model and a Generalized Smoothing wrap-up

We observe a "true" trajectory X^* at n random times $0 = t_1 < t_2 \cdots < t_n = T$, such that we have n observations $(Y_1, \ldots, Y_n)$ defined as

$$Y_i = X^*(t_i) + \epsilon_i \tag{4}$$

where ϵ_i is the (random) observation error. We assume that there is a true parameter θ^* belonging to a subset Θ of $\mathbb{R}^p$, such that X^* is the unique solution of the linear ODE

$$\dot{x}(t) = A_\theta(t)x(t) + r_\theta(t) \tag{5}$$

with initial condition $X^*(0) = X_0^*$; where $t \mapsto A_\theta(t) \in \mathbb{R}^{d\times d}$ and $t \mapsto r_\theta(t) \in \mathbb{R}^d$. More generally, we denote X_θ the solution of (5) for a given θ, and initial condition X_0^*. We assume that the initial condition X_0^* is exactly known, and we want to infer θ^* from $(Y_1, \ldots, Y_n)$. In the linear case, Duhamel's formula gives a closed form expression for X_θ for $t \in [0, T]$

$$X_\theta(t) = \Phi_\theta(t, 0)X_0^* + \int_0^T \Phi_\theta(t, s)r_\theta(s)ds$$

where the matrix-valued function $\Phi_\theta : (t, s) \mapsto \Phi_\theta(t, s)$ is the so-called resolvant of the ODE. By definition, the resolvant is the solution of the homogenous ODE

$$\begin{aligned}\dot{\Phi}_\theta(t, s) &= A_\theta(t)\Phi_\theta(t, s)\\ \Phi_\theta(s, s) &= I_d\end{aligned}$$

The estimation of θ^* can be done straightforwardly with the Nonlinear Least Squares (NLS) estimator that minimizes

$$\sum_{i=1}^{n} \|Y_i - X_\theta(t_i)\|_2^2 .$$

In Generalized Smoothing (GS), parameter estimation is regularized by using an approximate solutions of the ODE (5), as GS takes advantage of the double interpretation of splines for smoothing data, and for numerical solving of ODE by collocation.

A basis expansion $\widehat{X}(t, \theta) = \widehat{\beta}(\theta)^T p(t)$ is computed for each θ, where $\widehat{\beta}(\theta)$ is obtained by minimizing in β the criterion

$$J_n(\beta|\theta, \lambda) = \sum_{i=1}^{n} \left\| Y_i - \widehat{\beta}^T p(t_i) \right\|_2^2 + \lambda \int_0^T \left\| \widehat{\beta}^T \dot{p}(t) - \left(A_\theta(t)\widehat{\beta}^T p(t) + r_\theta(t) \right) \right\|_2^2 dt \tag{6}$$

This first step is considered as profiling along the nuisance parameter β, whereas the estimation of the parameter of interest is obtained by minimizing the sum

of squared errors of the proxy $\hat{X}(t,\theta)$ by

$$\hat{\theta}^{GS} = \arg\min_{\theta} \sum_{i=1}^{n} \left\| Y_i - \hat{X}(t_i,\theta) \right\|^2 \tag{7}$$

Obviously, the estimator depends on the hyperparameter λ, that needs to be selected from the data in practice (some adaptive procedures have been proposed, see [22]). The essential difference with NLS is to replace the exact solution $X_\theta(\cdot)$ by an approximation $\hat{X}(\cdot,\theta)$ (that depends also on the data). This means that GS deals with 2 sources of errors: in addition to the classical statistical error (variance due to noisy data), there is an approximation error as $\hat{X}(\cdot,\theta)$ is a spline that does not solve exactly the ODE model (5). Indeed, collocation algorithms compute the coefficients of a B-spline expansion based on the relationships between $\hat{X}$ and its derivative $\dot{\hat{X}}$ evaluated on an appropriate grid of time points $0 = s_1 < s_2 < \cdots < s_p$, [2]. This gives a nonlinear system that is usually solved with a Newton algorithm, whose roots are the unknown coefficients of the basis expansion. The collocation schemes are essentially useful for solving Boundary Value Problems (instead of the classical Initial Value Problem).

For parameter estimation, the basis expansion is defined in a somehow arbitrary manner and the ODE constraint is not used as an equality constraint as it should be the case in a "normal" collocation scheme. Instead, the ODE equation is transformed into an inequality constraint defined on the interval $[0,T]$ and the model constraint is never set to 0 because of the trade-off with the data-fitting term $\sum_{i=1}^{n} \left\| Y_i - \widehat{\beta}^T p(t_i) \right\|_2^2$. For this reason, the ODE model (5) is not solved and it is useful to introduce the discrepancy term $\hat{u}_\theta(t) = \widehat{\beta}^T \dot{p}(t) - \left(A_\theta(t)\widehat{\beta}^T p(t) + r_\theta(t) \right)$ that corresponds to a model error. In fact, the proxy $\hat{X}(\cdot,\theta)$ satisfies a pertubed ODE $\dot{x} = A_\theta x + r_\theta + \hat{u}_\theta$. This forcing function $\hat{u}_\theta$ is an outcome of the optimization process and can be relatively hard to analyze or understand, as it depends on the basis expansion used and it depends also on the data via the minimization of $J_n(\beta|\theta,\lambda)$. Nevertheless, Hooker et al. have proposed goodness-of-fit tests based on this so-called "empirical forcing function" $\hat{u}_\theta$, as $\hat{u}_\theta$ are the residuals but at the derivative scale and not at the state scale [19, 20].

Based on these remarks, we introduce the pertubed linear ODE

$$\dot{x}(t) = A_\theta(t)x(t) + r_\theta(t) + u(t) \tag{8}$$

where the function $t \mapsto u(t)$ can be any function in L^2. The solution of the corresponding Initial Value Problem

$$\begin{aligned} \dot{x}(t) &= A_\theta(t)x(t) + r_\theta(t) + u(t) \\ x(0) &= X_0 \end{aligned}$$

is denoted $X_{\theta,u}$. Instead of using the spline proxy $\hat{X}(\cdot,\theta)$ for approximating X^*, we use the trajectories $X_{\theta,u}$ of the ODE (8) controlled by the function u.

2.2 The Tracking Estimator

Following the Generalized Smoothing approach, we look for a candidate $X_{\theta,u}$ that can minimize at the same time the discrepancy with the data and the norm $\|u\|_{L^2}$. Moreover, we replace the classical Sum of Squared Errors by a smoothed version $\int_0^T \left\|\hat{X}(t) - X_{\theta,u}(t)\right\|_2^2 dt$ based on a nonparametric proxy $\hat{X}$. Hence, we consider the subsequent cost function

$$C\left(\hat{X}; u, \theta, \lambda\right) = \int_0^T \left\|\hat{X}(t) - X_{\theta,u}(t)\right\|_2^2 dt + \lambda \int_0^T \|u(t)\|_2^2 \, dt \tag{9}$$

for a given $\lambda > 0$. Moreover, for each θ in Θ, we introduce the infimum function

$$S\left(\hat{X}; \theta, \lambda\right) = \inf_{u \in L^2} C\left(\hat{X}; u, \theta, \lambda\right) \tag{10}$$

obtained by "profiling" on the function u. The definition of S mimicks the minimization of $J_n(\beta|\theta, \lambda)$ but it is more involved as it is defined on infinite functional space, instead of a finite dimensional vector space. Finally, our estimator is defined by minimizing the same function S i.e

$$\widehat{\theta}^T = \arg\min_{\theta \in \Theta} S\left(\hat{X}; \theta, \lambda\right) \tag{11}$$

whereas the GS estimator minimizes a different criterion $\sum_{i=1}^n \left\|Y_i - \hat{X}(t_i, \theta)\right\|^2$. This means that in our methodology, we try to find a parameter θ that maintain a reasonable trade-off between model and data, whereas the Generalized Smoothing Estimator $\hat{\theta}^{GS}$ is dedicated to fit the data with the proxy $\hat{X}(\cdot, \theta)$, without considering the size of model error represented by $\hat{u}_\theta$.

Before going deeper into the interpretation and analysis of our estimator, we need to show that the criterion $S\left(\hat{X}; \theta, \lambda\right)$ is properly defined and that we can obtain a tractable expression for computations and for the theoretical analysis of (11). The existence of S is a direct consequence of the so-called Linear-Quadratic Theory (LQ Theory), which belongs to the broader field of Optimal Control Theory [26, 37, 29, 9]. In our case, we consider the control of linear ODE with a quadratic cost function that enables to have quite general and simple results. This is possible because we have replaced the discrete sum of squared errors by an integral criterion where the original data have been replaced by a nonparametric proxy $\hat{X}$. Thanks to that, we can use directly calculus of variations and optimal control [24, 9]. For completeness, we recall briefly in the appendix the main results of LQ Theory.

Theorem and Definition of $S(\zeta; \theta, \lambda)$**.** *Let $t \mapsto \zeta(t)$ be a function belonging to $H^1([0, T], \mathbb{R}^d)$ and $X_{\theta,u}$ be the solution to the controlled ODE (8).*
For any θ,λ, there exists a unique optimal control $\bar{u}_{\theta,\lambda}$ that minimizes the cost function

$$C(\zeta; u, \theta, \lambda) = \int_0^T \left\{\|\zeta(t) - X_{\theta,u}(t)\|_2^2 + \lambda \|u(t)\|_2^2\right\} dt \tag{12}$$

The control $\overline{u}_{\theta,\lambda}$ can be computed in a "closed-loop" form as

$$\overline{u}_{\theta,\lambda}(t) = \frac{E(t)}{\lambda}\left(X_{\theta,\overline{u}_{\theta,\lambda}}(t) - \zeta(t)\right) + \frac{h(t)}{\lambda} \tag{13}$$

where E and h are solutions of the Final Value Problems

$$\begin{cases} \dot{E}(t) = I_d - A_\theta(t)^T E(t) - E(t)A_\theta(t) - \frac{E(t)^2}{\lambda} \\ \dot{h}(t) = -A_\theta(t)^T h(t) - E(t)\left(A_\theta(t)\zeta(t) + r_\theta(t) - \dot{\zeta}(t)\right) - \frac{E(t)h(t)}{\lambda} \end{cases} \tag{14}$$

and $E(T) = 0$, $h(T) = 0$. For all $t \in [0,T]$, the matrix $E(t)$ is symetric, and the ODE defining the matrix-valued function $t \mapsto E(t)$ is called the Matrix Riccati Differential Equation of the ODE (8).

Finally, the Profiled Cost S has the closed form

$$S(\zeta;\theta,\lambda) \quad = \quad -\int_0^T \left\{ 2\left(A_\theta(t)\zeta(t) + r_\theta(t) - \dot{\zeta}(t)\right)^\top h(t) + \frac{\|h(t)\|^2}{\lambda} \right\} dt \tag{15}$$

The cost (12) is usually used for solving the so-called "Tracking Problem" that consists in finding the optimal control u to apply to the ODE (8) in order to reach a target trajectory $t \mapsto \zeta(t)$, see [37] for an excellent introduction. The estimation problem is then to determine the parameter θ so that the corresponding ODE need a small control u (in L^2 norm) in order to be close to the noisy trajectory $t \mapsto \hat{X}(t)$.

Remark 2.1. $H^1([0,T],\mathbb{R}^d) = \left\{ X \in L^2([0,T],\mathbb{R}^d) \mid \dot{X} \in L^2([0,T],\mathbb{R}^d) \right\}$ is the classical Sobolev space of L^2 (weakly) differentiable function, see [3]. The derivative is defined in the weak sense, so it allows us to consider non-parametric estimator with some (controlled) discontinuities. Of course, every differentiable functions belong to H^1 and the weak derivative coincides with the classic one.

Remark 2.2. We insist on the fact that $t \mapsto E(t), h(t)$ depends also on θ, λ and ζ because of their definition via equation (14). Nevertheless, we do not write it systematically for notational brievety. As mentioned in the theorem, it is possible to compute $X_{\theta,\overline{u}_{\theta,\lambda}}$ in a "closed-loop" form as we can solve in a preliminary stage the 2 equations (14) that gives the function E and h for all $t \in [0,T]$. Then, we just need to solve the ODE

$$\begin{aligned} &\dot{x}(t) = A_\theta(t)x(t) + r_\theta(t) + \tfrac{E(t)}{\lambda}\left(x(t) - \zeta(t)\right) + \tfrac{h(t)}{\lambda} \\ &x(0) = X_0 \end{aligned}$$

Remark 2.3. From equation (15), we see that S depends smoothly in θ and λ, as in ζ. This was not easy to see from the infimum definition (10), but as the minimum is reached, and attained for a known function, we can have even more information than in the Generalized Smoothing approach based on splines.

Remark 2.4. Our pertubated ODE framework permits to consider naturally the problem of model misspecification, when the true model is

$$\dot{x}(t) = A_\theta(t)x(t) + r_\theta(t) + v(t)$$

with $v \in L^2([0,\, T]\,, \mathbb{R}^d)$ is an unknown function. We do not provide any theoretical analysis for this kind of model misspecification, but we consider it in the Experiments section, in order to gain some insight. We will see in a simple example that our estimator allows us to have more accurate estimation than classical NLS estimator in that case. Moreover we can propose a proper correction term to add to the initial model to counteract the misspecification in order to lower the prediction error.

The next section is dedicated to the derivation of the regularity properties of S. Thanks to the use of a functional formulation and the associated Linear-Quadratic theory, we show the smoothness in ζ and θ, and compute directly the needed derivatives.

3 Consistency of the Tracking estimator

Under reasonable and practical assumptions, we can assert that the tracking estimator (11) is a consistent estimator of θ^* when the ODE model (5) is well-specified, and when we use a consistent nonparametric estimator $\hat{X}$. In practice, it is quite common to use a smoothing spline or a kernel smoother in order to smooth the data and estimates roughly the trajectory X^*. As the tracking estimator is an M-estimator, we can employ the classical approaches for consistency that relies on the regularity and convergence of the stochastic criterion $S\left(\hat{X}; \theta, \lambda\right)$ to the asymptotic criterion $S\left(X^*; \theta, \lambda\right)$. Hence, we need to show some regularity in ζ, uniformly in θ. Similarly, in order to compute the rate of convergence and the variance of the estimator, we will need to check the smoothness w.r.t θ.

3.1 Regularity properties of $S(\zeta; \theta, \lambda)$

We introduce some necessary assumptions about the ODE model in order to derive the needed regularity as well as the identifiability property. The conditions are

C1: $\theta^* \in \Theta$ a compact subset of $\mathbb{R}^p$

C2: The model is identifiable at $\theta = \theta^*$ i.e

$$\forall \theta \in \Theta\,;\, X_\theta = X_{\theta^*} \Longrightarrow \theta = \theta^*$$

C3: $\forall\, (t, \theta) \in [0\,,\, T] \times \Theta$, $(t, \theta) \longmapsto A_\theta(t)$ and $(t, \theta) \longmapsto r_\theta(t)$ are continuous.

C4: $\forall\, (t, \theta) \in [0,\, T] \times \Theta$, $(t, \theta) \longmapsto \frac{\partial A_\theta(t)}{\partial \theta}$ and $(t, \theta) \longmapsto \frac{\partial r_\theta(t)}{\partial \theta}$ are continuous

According to the context, $\|\cdot\|_2$ denotes the Euclidean norm in $\mathbb{R}^d$ ($\|X\|_2 = \sqrt{\sum_{i=1}^d X_i^2}$) or the Frobenius matrix norm ($\|A\|_2 = \sqrt{\sum_{i,j} |a_{i,j}|^2}$). We use also the functional norm in $L^2\left([0\,T]\,, \mathbb{R}^d\right)$ defined by $\|f\|_{L^2} = \sqrt{\int_0^T \|f(t)\|_2^2\, dt}$. Continuity and differentiability have to be understood w.r.t these previous norms.

For the computation of $S\left(\hat{X};\theta,\lambda\right)$ (and $S\left(X^*;\theta,\lambda\right)$), we need some additional notations. In particular, we recall that the Riccati equation $\dot{E} = I_d - A_\theta(t)^\top E - EA_\theta(t) - \frac{E^2}{\lambda}$ depends on the model (5) , but it does not depend on the data $\hat{X}$, whereas it is the case for h, as we have $\dot{h}(t) = -A_\theta(t)^T h(t) - E(t)\left(A_\theta(t)\zeta(t) + r_\theta(t) - \dot{\zeta}(t)\right) - \frac{E(t)h(t)}{\lambda}$. For this reason, we introduce the functions α and β defined by

$$\begin{cases} \alpha_\theta(t) = \left(A_\theta(t)^T + \frac{E_\theta(t)}{\lambda}\right) \\ \beta_\theta(t,\zeta) = E_\theta(t)\left(A_\theta(t)\zeta + r_\theta(t) - \dot{\zeta}\right) \end{cases}$$

We denote then $\widehat{h_\theta}$ the solution to the Final Value Problem

$$\begin{cases} \dot{h} = -\alpha_\theta(t)h - \beta_\theta(t,\widehat{X}) \\ h(T) = 0 \end{cases}$$

and h^* the solution corresponding to case $\zeta = X^*$. More generally, we will denote $t \mapsto h_\theta(t,\zeta)$ for any target trajectory ζ.

We introduce also the matrix-valued function $(t,s) \mapsto R_\theta(t,s)$ defined for all t, s in $[0,T]$, as the solution of the Initial Value Problem

$$\begin{cases} \dot{R}_\theta(t,s) = \alpha_\theta(T-t)R(t,s) \\ R_\theta(s,s) = I_d \end{cases} \tag{16}$$

and where the time has been reversed in the function α_θ. We show in the next proposition that $\forall \zeta \in H^1([0,\,T]),\ \theta \mapsto S(\zeta;\theta,\lambda)$ is well defined, i.e finite on Θ.

Proposition 3.1. *Under conditions 1 and 3 we have:*

$$\overline{A} = \sup_{\theta\in\Theta} \|A_\theta\|_{L^2} < +\infty$$

$$\overline{X} = \sup_{\theta\in\Theta} \|X_\theta\|_{L^2} < +\infty$$

$$\bar{E} = \sup_{\theta\in\Theta} \|E_\theta\|_{L^2} < +\infty$$

and

$$\forall \zeta \in H^1([0,\,T]),\ \bar{h_\zeta} = \sup_{\theta\in\Theta} \|h_\theta(.,\zeta)\|_{L^2} < +\infty$$

Hence, for all ζ in $H^1([0,\,T])$, the map $\theta \longmapsto S(\zeta;\theta,\lambda)$ is well defined on Θ (i.e $\sup_{\theta\in\Theta} \|S(\zeta;\theta,\lambda)\| < +\infty$).

Proof. $\bar{A} < +\infty$ exists as $(t,\theta) \mapsto A_\theta(t)$ is a continuous function on the compact set $[0,T] \times \Theta$. The existence and extension theorem for IVP solution of linear ODE ensures that $\forall \theta \in \Theta$, $\|X_\theta\|_{L^2} < +\infty$. Moreover, solutions are continuous

in (t,θ) if the vector field is continuous in (t,θ). By analogy with theorem A.1 in appendix, we know that

$$E_\theta^g(t) := \begin{pmatrix} E_\theta(.) & h_\theta(.,\zeta)^T \\ h_\theta(.,\zeta) & \alpha_\theta(.,\zeta) \end{pmatrix}$$

with $\alpha_\theta(t,\zeta) = \int_t^T \left(2\left(A_\theta(s)\zeta(s) - \dot{\zeta}(s) + r_\theta(s)\right)^T h_\theta(s,\zeta) + \frac{1}{\lambda} h_\theta(s,\zeta)^T h_\theta(s,\zeta)\right) ds$ is the ODE solution of the extended Riccati ODE

$$\begin{aligned} &\dot{E}_\theta^g(t) = W^1 - A_\theta^1(t)^t E_\theta^g(t) - E_\theta^g(t) A_\theta^1(t) - \tfrac{1}{\lambda} E_\theta^g(t)^2 \\ &E_\theta^g(T) = 0_{d+1,d+1} \end{aligned}$$

where $W_1 = \begin{pmatrix} I_d & 0 \\ 0 & 0 \end{pmatrix}$, $A_\theta^1(t) = \begin{pmatrix} A_\theta(t) & r_\theta^1(t) \\ 0 & 0 \end{pmatrix}$ and $r_\theta^1(t) = A_\theta(t)\zeta(t) + r_\theta(t) - \dot{\zeta}(t)$.

Because for all $\theta \in \Theta$, $A_\theta \in L^2\left([0,\,T], \mathbb{R}^{d\times d}\right)$ and $(A_\theta\zeta - \dot{\zeta} + r_\theta) \in L^2\left([0,\,T], \mathbb{R}^d\right)$ thanks to lemma B.1 in appendix, $E_\theta^g(t)$ is bounded and continuous in (t,θ). Hence h_θ, E_θ are bounded on $[0,T]\times\Theta$. We conclude for $\theta \mapsto S(\zeta;\theta,\lambda)$ thanks to norm inequality. □

We complete our analysis by showing that S is continuously differentiable in θ.

Proposition 3.2. *Under conditions C1-C3*

$$\forall X \in H^1([0,\,T]),\ \theta \longmapsto S(X;\theta,\lambda)$$

is continuous on Θ.
Under conditions C1-C4, S *is* C^1 *on* Θ.

Proof. Since

$$S(X;\theta,\lambda) = -\int_0^T \left(2\left(A_\theta(t)X(t) + r_\theta(t) - \dot{X}(t)\right)^T h_\theta(t,X) + \frac{1}{\lambda}\left\|h_\theta(t,X)\right\|^2\right) dt$$

Condition 3, jointly with proposition 1 and 4 in the supplementary materials give the continuity of $\theta \longmapsto (t \longmapsto A_\theta(t))$ and $(\theta, X) \longmapsto (t \longmapsto h_\theta(t,X))$ on Θ. This is enough to show the continuity of $\theta \longmapsto S(X;\theta,\lambda)$ on Θ. Moreover, the gradient w.r.t θ of $S(X;\theta,\lambda)$ is equal to:

$$\begin{aligned} \nabla_\theta S(X;\theta,\lambda) \;=\; & -2\int_0^T \tfrac{\partial\left(A_\theta(t).X + r_\theta(t)\right)}{\partial\theta}^T h_\theta(t,X)dt \\ & +2\int_0^T \tfrac{\partial h_\theta(t,X)}{\partial\theta}^T \left(A_\theta(t).X + r_\theta(t) - \dot{X} + \tfrac{1}{\lambda} h_\theta(t,X)\right) dt \end{aligned}$$

In addition to the previous proposition, condition 4 and proposition 7 in supplementary material gives the continuity of $(\theta, X) \longmapsto \left(t \longmapsto \frac{\partial(h_\theta(t,X))}{\partial\theta}\right)$ on Θ. This is enough to show the continuous differentiability of $S(X;\theta,\lambda)$ on Θ. □

The last regularity properties justifies the use of classical optimization method to retrieve the minimum of S.

In the next proposition, we show that the criteria $S(X;\theta,\lambda)$ can be expressed without using the derivative $\dot{X}$ (thanks to the knowledge of the initial condition). As a consequence, our estimator is less sensible to the nonparametric noise than classical Two-Step estimators.

Proposition 3.3. *Under conditions 1 and 2, $\forall X \in H^1([0,T])$ with $X(0) = X_0^*$, $S(X;\theta,\lambda)$ does not depend on $\dot{X}$, i.e it is continuous nonlinear integral of $t \mapsto X(t)$.*

Proof. We will show $S(X;\theta,\lambda)$ can be written using only X and not $\dot{X}$. First of all we will use Lemma (B.3) to get rid of $\dot{X}$ in $\int_0^T \dot{X}(t)^T h_\theta(t,X)dt$, it gives:

$$\begin{array}{rcl} \int_0^T \dot{X}(t)^T h_\theta(t,X)dt & = & F_{1,\theta}(X) + F_{2,\theta}(X) + F_{3,\theta}(X) \\ & & -X_0^{*T}\int_0^T R_\theta(T,T-s)E_\theta(s)r_\theta(s)ds \\ & & -\frac{1}{2}X_0^{*T}E_\theta(0)X_0^* \end{array} \tag{17}$$

with

$$\left\{ \begin{array}{l} F_{1,\theta}(X) = -X_0^T \int_0^T R_\theta(T,T-s)X(s)ds \\ F_{2,\theta}(X) = \int_0^T X(t)^T \left(\alpha_\theta(t)h_\theta(t,X)dt + (A_\theta(t)X(t) + r_\theta(t))\right)dt \\ F_{3,\theta}(X) = \frac{1}{2}\int_0^T X(t)^T \dot{E_\theta}(t)X(t)dt \end{array} \right.$$

And so we can write $S(X;\theta,\lambda)$ under the form

$$\begin{array}{rcl} S(X;\theta,\lambda) & = & -\int_0^T \left(2\left(A_\theta(t)X(t) + r_\theta(t)\right)^T h_\theta(t,X) + \frac{1}{\lambda}h_\theta(t,X)^T h_\theta(t,X)\right)dt \\ & & +F_{1,\theta}(X) + F_{2,\theta}(X) + F_{3,\theta}(X) \\ & & -X_0^{*T}\int_0^T R_\theta(T,T-s)E_\theta(s)r_\theta(s)ds \\ & & -\frac{1}{2}X_0^{*T}E_\theta(0)X_0^* \end{array}$$

since from Lemma (B.2) we have the affine dependence of $h_\theta(t,X)$ w.r.t X through the formula:

$$h_\theta(t,X) = \int_t^T R_\theta(T-t,T-s)X(s)ds + E_\theta(t)X(t) + \int_t^T R_\theta(T-t,T-s)E_\theta(s)r_\theta(s)ds$$

we see $S(X;\theta,\lambda)$ does not depend on $\dot{X}$. □

3.2 Consistency

As we have seen previously, conditions 1 and 3 ensure the existence of $S(\hat{X};\theta,\lambda)$ and $S(X^*;\theta,\lambda)$ for all $\theta \in \Theta$. We derive the consistency of $\hat{\theta}^T$ by showing the uniform convergence of the criterion $S\left(\hat{X};\theta,\lambda\right)$, and by insuring that θ^* is a

unique and isolated global minima of $S(X^*;\theta,\lambda)$. Condition 2 is then sufficient to show that $S(X^*;\theta,\lambda)$ characterizes well θ^*, as global unique minimum. Hence, identifiability and convergence in supremum norm are sufficient to imply consistency (theorem 5.7 in [38]).

Proposition 3.4. *For all X in $H^1([0,T])$, $S(X;\theta,\lambda)\geq 0$ and under conditions C1 and C2 we have*

$$S(X^*;\theta,\lambda)=0 \Longleftrightarrow \theta=\theta^*$$

Proof. If $\theta=\theta^*$, then $u\equiv 0$ is the cost which minimizes

$$C(X^*;u,\theta^*,\lambda)=\int_0^T \|X^*(t)-X_{\theta^*,u}(t)\|_2^2\,dt+\lambda\int_0^T\|u(t)\|_2^2\,dt$$

and in that case $S(X^*;\theta^*,\lambda)=\inf_{u\in L^2} C(X^*;u,\theta^*,\lambda)=0$.

Conversely, let θ^0 be such that $S(X^*;\theta^0,\lambda)=0$. By definition, this means that $\int_0^T \left\|X^*(t)-X_{\theta^0,u}(t)\right\|_2^2 dt+\lambda\int_0^T\|u(t)\|_2^2\,dt=0$. A consequence is that $u=0\,a.e$ and $X_{\theta^* u=0}(t)=X_{\theta^0,u=0}(t)\,a.e$; by the identifiability condition we get that $\theta^0=\theta^*$. □

Theorem 3.5. *Under conditions 1, 2, 3 and if $\widehat{X}$ is consistent in probability (in L^2-norm sense), we have*

$$\widehat{\theta}^T \xrightarrow{P} \theta^*$$

Proof. Using proposition B.4, we have

$$\begin{aligned}&|S(X;\theta,\lambda)-S(X^*;\theta,\lambda)|\\&\leq 2\left(\bar{A}\bar{h}+K_1+K_2\left\|\widehat{h}_\theta\right\|_{L^2}+K_3\left\|\widehat{X}\right\|_{L^2}\right)\left\|X^*-\widehat{X}\right\|_{L^2}\\&+\left(\bar{A}\left\|\widehat{X}\right\|_{L^2}+K_4+\tfrac{1}{\lambda}\left(\left\|\widehat{h}_\theta\right\|_{L^2}+\bar{h}\right)\right)\left\|h^*_\theta-\widehat{h}_\theta\right\|_{L^2}\end{aligned}$$

with

$$\begin{aligned}K_1&=\sqrt{d}\,\|X_0\|_2\,\bar{R}+\sqrt{d}\bar{A}\bar{X}+\sqrt{d}\bar{\dot{E}}\overline{X}\\K_2&=\sqrt{d}\left(\bar{A}+\tfrac{\bar{E}}{\lambda}\right)\\K_3&=\sqrt{d}\bar{A}+\sqrt{d}\bar{\dot{E}}\\K_4&=\sqrt{d}\left(\bar{A}+\tfrac{\bar{E}}{\lambda}\right)\bar{X}\end{aligned}$$

and

$$\begin{aligned}\bar{R}&=\sup_{\theta\in\Theta}\|R_\theta(T,T-.)\|_{L^2}\\\bar{\dot{E}}&=\sup_{\theta\in\Theta}\left\|\dot{E}_\theta\right\|_{L^2}\end{aligned}$$

by using the same notation as in proposition 3.1. Proposition B.5 permits to bound $\left\|h^*_\theta-\widehat{h}_\theta\right\|_{L^2}$ with $\left\|\widehat{X}-X^*\right\|_{L^2}$ as

$$\begin{aligned}&\left\|\widehat{h_\theta}-h^*_\theta\right\|_{L^2}\leq K_5\left\|\widehat{X}-X^*\right\|_{L^2}\\&\text{with}: K_5=\sqrt{d}\left(Tde^{\sqrt{d}\left(\bar{A}+\frac{\bar{E}}{\lambda}\right)T}+\overline{E}\right)\end{aligned}$$

We obtain

$$|S(X;\theta,\lambda) - S(X^*;\theta,\lambda)| \leq \left(\left(2K_2 + \tfrac{K_5}{\lambda}\right) \left\|\widehat{h_\theta}\right\|_{L^2} + \left(2K_3 + K_5\bar{A}\right) \left\|\widehat{X}\right\|_{L^2} + K_7 \right) \left\|X^* - \widehat{X}\right\|_{L^2}$$
$$\text{with: } K_7 = 2\left(\bar{A}\bar{h} + K_1\right) + K_5\left(K_4 + \tfrac{\bar{h}}{\lambda}\right)$$

We can control $\left\|\widehat{X}\right\|_{L^2} \leq \left\|\widehat{X} - X^*\right\|_{L^2} + \|X^*\|_{L^2}$, which proves that if is $\widehat{X}$ is consistent, then $\sup_{\theta\in\Theta} |S(X;\theta,\lambda) - S(X^*;\theta,\lambda)| = o_P(1)$. Application of the proposition 3.4 gives us the identifiability criteria. Hence we conclude by using the theorem 5.7 in [38]. □

4 Asymptotics of $\widehat{\theta}^T$

Our objective in this part is to derive the proper rate of convergence of the Tracking Estimator, as well as its asymptotic distribution. The properties of the estimator depends on the behavior of the nonparametric estimate $\hat{X}$ used for the approximation of X^*. In order to fix ideas, we consider a regression spline, with a B-Spline decomposition of dimension K (increasing with n). That is we consider that $\widehat{X}$ is defined as

$$\widehat{X}(t) = \sum_{k=1}^{K} \beta_{kK} p_{kK}(t) = \beta_K^T p_K(t)$$

where β_K is computed by least-squares. It is likely that we could derive the same kind of results for different estimates, such as Local Polynomial or Smoothing Splines, as they behave similarly asymptotically, and that we show that the Tracking Estimate can be approximated by a plug-in estimate of a specific linear functional of $\hat{X}$. We introduce additional regularity conditions needed for the asymptotics:

C5: The Hessian $\frac{\partial^2 S(X^*;\theta,\lambda)}{\partial\theta^T\partial\theta}$ is nonsingular at $\theta = \theta^*$.

C6: The observations (t_i, Y_i) are i.i.d with $Var(Y_i \mid t_i) = \sigma I_d$ with $\sigma < +\infty$

C7: Observations time t_i are uniformely distributed on $[0\,,\,T]$

C8: It exists $s \geq 1$ such that $t \longmapsto A_{\theta^*}(t)$, $t \longmapsto r_{\theta^*}(t)$ are $C^{s-1}\left([0\,,\,T], \mathbb{R}^d\right)$ and $\sqrt{n}K^{-s} \longrightarrow 0$ and $\frac{K^s}{n} \longrightarrow 0$

Under these additional conditions, we show that $\hat{\theta}^T$ reaches the parametric convergence rate, and that it is asymptotically normal. Our strategy consists in two stages:

Stage 1 (Proposition 4.3) We show that $\hat{\theta}^T - \theta^*$ behaves asymptotically as the difference $\Gamma(\widehat{X}) - \Gamma(X^*)$ where Γ is a continuous linear functional,

Stage 2 (Proposition 4.4) We prove that $\Gamma\left(\widehat{X} - X^*\right)$ is asymptotically normal for regression splines, based on the properties of plug-in estimators computed with series estimators and derived in [31].

Remark 4.1. Condition C5 is a classic feature for $M-$estimator to ensure local identifiability, here:

$$\begin{array}{rcl} \frac{\partial^2 S(X^*;\theta^*,\lambda)}{\partial\theta^T\partial\theta} & = & 2\int_0^T \frac{\partial(A_{\theta^*}(t)X^*+r_{\theta^*}(t))}{\partial\theta}^T \frac{\partial h^*_{\theta^*}(t)}{\partial\theta} + \frac{\partial h^*_{\theta^*}(t)}{\partial\theta}^T \frac{\partial(A_{\theta^*}(t)X^*+r_{\theta^*}(t))}{\partial\theta} dt \\ & + & \frac{2}{\lambda}\int_0^T \frac{\partial h^*_{\theta^*}(t)}{\partial\theta}^T \frac{\partial h^*_{\theta^*}(t)}{\partial\theta} dt \end{array}$$

that is why we only require $\forall\,(t,\theta) \in [0\,,\,T] \times \Theta$, $(t,\theta) \longmapsto A_\theta(t)$ and $(t,\theta) \longmapsto r_\theta(t)$ to be C^1 and not C^2

Remark 4.2. Condition C8 is a classic feature for non-parametric estimator to ensure optimal convergence rate of $\widehat{X}$ using bias-variance tradeoff.

Proposition 4.3. *Under conditions 1-5, we have :*

$$\widehat{\theta^T} - \theta^* = 2\frac{\partial^2 S(X^*;\theta^*,\lambda)}{\partial\theta^T\partial\theta}^{-1}\left(\Gamma(\widehat{X}) - \Gamma(X^*)\right) + o_P(1)$$

where $\Gamma : C\left([0\,,\,T]\,,\mathbb{R}^d\right) \to \mathbb{R}^p$ *is a linear functional defined by*

$$\Gamma(X) = \int_0^T \left(\frac{\partial\left(A_{\theta*}(t).X^*\right)}{\partial\theta} + \frac{1}{\lambda}\frac{\partial h_{\theta^*}(t,X^*)}{\partial\theta}\right)^T \left(\int_t^T R_{\theta^*}(T-t,T-s)X(s)ds\right) dt. \tag{18}$$

R_{θ^*} *is defined* by (16).

Proposition 4.4. *Under conditions 1-8 and by defining* Γ *as in proposition 4.3 we have that* $\Gamma(\widehat{X}) - \Gamma(X^*)$ *is asymptotically normal and* $\Gamma(\widehat{X}) - \Gamma(X^*) = O_P(n^{-1/2})$

To obtain the final result, we only have to combine the two previous propositions:

Theorem 4.5. *If* $\widehat{X}$ *is a regression spline and conditions C1-C8 are satisfied, then* $\widehat{\theta^T} - \theta^*$ *is asymptotically normal and*

$$\widehat{\theta^T} - \theta^* = O_P(n^{-1/2})$$

Remark 4.6. The asymptotic linear representation given by proposition 4.3 allows us to obtain an expression for the asymptotic variance. This latter is given by the formula (33) in appendix D as well as a plug-in consistent estimator.

5 Experiments

We use several simple test beds for evaluating the practical efficiency of the Tracking Estimator $\hat{\theta}^T$, and we compare it with the performance of NLS $\hat{\theta}^{NLS}$and of Generalized Smoothing $\hat{\theta}^{GS}$. The different models are linear in the states, and they can be linear and nonlinear w.r.t parameters. We use several sample size and several variance error for comparing robustness and efficiency for varying sample size and noise level.

5.1 Experimental design

For a given sample size n and noise level σ, we estimate the Mean Square Error and the mean Absolute Relative Error (ARE)

$$\mathbb{E}_{\theta^*}\left[\frac{\left|\theta^* - \widehat{\theta}\right|}{|\theta^*|}\right]$$

by Monte Carlo, based on $N_{MC} = 100$ runs. For each run, we simulate an ODE solution with a Runge-Kutta algorithm (ode45 in Matlab), and a centered Gaussian noise (with variance σ) is added, in order to obtain the Y_i's.

We compare also the quality of estimation of $\hat{\theta}^T$, $\widehat{\theta}^{GS}$ and $\widehat{\theta}^{NLS}$ based on their prediction quality. We compute the Prediction Error

$$\mathbb{E}_{\theta^*,\sigma}\left[\left\|\mathrm{Y}^* - \mathrm{X}_{\widehat{\theta}}\right\|^2_{\mathrm{L}^2}\right] \tag{19}$$

where

- Y^* is a new observation drawn from the true model (4),
- $X_{\widehat{\theta}}$ is the solution to the linear ODE (5) with parameter $\widehat{\theta}$.

It should be emphasized that parameter estimation and prediction error minimization are two different problems, although they are related. Parameter estimation is required when parameters are directly of interest, for instance for a deep understanding of the inner dynamics of the system. One put forward prediction error when the aim is only to quantitatively predict the system state. Our primary interest is parameter estimation but we also discuss prediction performance for the three methods.

The nonparametric estimate $\hat{X}$ required in the first step is a spline defined with a uniform knots sequence $\xi_k, k = 1, \ldots, K$. For each run and each state variables, the number of knots is selected by minimizing the GCV criterion, [36]. We discuss in the next section an automated method for selecting adaptively the hyperparameter λ.

5.2 Selection of λ

The criterion $S\left(\hat{X}; \theta, \lambda\right)$ is based on a balance between data fidelity and model fidelity. When $\lambda \to 0$, we can select any u in order to interpolate $\hat{X}$. In that case, θ has almost no influence on $S\left(\hat{X}; \theta, \lambda\right)$ value.
When $\lambda \to \infty$, the optimal perturbation $\overline{u} \longrightarrow 0$, and we get a NLS-like criterion where the observations Y_i's are replaced by the proxy $\widehat{X}$. Because of this dramatic influence, we propose to select λ by minimizing the Sum of Squared Errors

$$SSE(\lambda) = \sum_{i=1}^{n}\left(Y_i - X_{\widehat{\theta_\lambda}^T}(t_i)\right)^2$$

.

5.3 Gradient computation

For optimization purpose, we need to compute the gradient $\nabla_\theta S(X;\theta,\lambda)$ which involves both $\frac{\partial \widehat{h_\theta}}{\partial \theta}$ and $\frac{\partial E_\theta}{\partial \theta}$. These partial derivatives can be obtained by solving the sensitivity equations, that gives the function values at each time $t \in [0, T]$. Nevertheless, the size of the ODE to solve grows quickly, as the sensitivity systems is of size $(d^2 + d) \times p$, and it becomes a computational burden for the optimization process. For this reason, we use the adjoint method to compute gradient expression [8]. This method exploits the fact that we do not need a point-wise computation of $\frac{\partial \widehat{h_\theta}}{\partial \theta}$ and $\frac{\partial E_\theta}{\partial \theta}$ but only some integral of the derivatives. An explanation of this method for gradient and Hessian computation can be found in [1]. The Riccati ODE can be written in vector form (of size $D := d^2+d$)

$$\begin{aligned} \dot{Q_\theta} &= F(Q_\theta, \theta, t) \\ Q_\theta(T) &= 0 \end{aligned}$$

where F is the row formulation of the Riccati ODE vector field and the solution is the vector

$$Q_\theta(t) = \left(\widehat{h_\theta}^T, (E_\theta^r)^T\right)^T (t).$$

Hence, we can compute $\nabla_\theta S(X;\theta,\lambda)$ thanks to the formula

$$\nabla_\theta S(X;\theta,\lambda) = \int_0^T \{\tfrac{\partial g(Q_\theta(t),\theta,t)}{\partial Q} - P(t).\tfrac{\partial F}{\partial \theta}(Q_\theta(t),\theta,t)\}dt$$

with

$$g(Q_\theta,\theta,t) = -2\left(A_\theta(t)\widehat{X}(t) - \dot{\widehat{X}}(t) + r_\theta(t)\right)^T \widehat{h_\theta} - \frac{1}{\lambda}\|\widehat{h_\theta}\|^2$$

and P is the so-called adjoint vector of size $D = d^2 + d$, solution of the adjoint model

$$\begin{aligned} \dot{P}(t) &= \tfrac{\partial g(Q_\theta(t),\theta,t)}{\partial Q} - P(t).\tfrac{\partial F}{\partial Q}(Q_\theta,\theta,t) \\ P(0) &= 0 \end{aligned}$$

The computational details for $\frac{\partial g}{\partial \theta}, \frac{\partial g}{\partial Q}, \frac{\partial F}{\partial \theta}, \frac{\partial F}{\partial Q}$ are left in appendix B. The adjoint method is more efficient than the direct sensitivity approach, as we need to solve an ODE of size D, instead of a system of size $D \times p$, which is valuable when the number of parameters increases.

5.4 Simple scalar equation

5.4.1 Linear w.r.t parameter

We consider here the basic model linear in parameter

$$\dot{x} = ax \tag{20}$$

with initial condition equal to $X_0^* = 1$. It is the simplest model we can consider, here the solution of the Cauchy problem is $x(t) = X_0^* e^{at}$ and we will use this closed form for the NLS estimator. Here we have tested two different sample

size $n = 50$ and $n = 20$ (observations were uniformly distributed between 0 and 5) and two different noise level $\sigma = 2$ and $\sigma = 4$. The lambda values tested are the 40 values uniformly distributed between 10 and 400.

(n, σ)		MSE ($\times 10^{-5}$)	ARE ($\times 10^{-3}$)	Pred error
$(50, 2)$				
	$\widehat{\theta}^{T}$	1.19	3.2	4.52
	$\widehat{\theta}^{NLS}$	1.25	3.5	4.53
	$\widehat{\theta}^{GS}$	43	22.5	6.06
$(50, 4)$				
	$\widehat{\theta}^{T}$	3.42	5.3	9.05
	$\widehat{\theta}^{NLS}$	3.91	6.0	9.07
	$\widehat{\theta}^{GS}$	230	47.2	11.76
$(20, 2)$				
	$\widehat{\theta}^{T}$	2.56	4.6	4.58
	$\widehat{\theta}^{NLS}$	2.92	5.6	4.74
	$\widehat{\theta}^{GS}$	100	32.7	7.48
$(20, 4)$				
	$\widehat{\theta}^{T}$	8.82	7.8	9.07
	$\widehat{\theta}^{NLS}$	9.41	9.3	9.10
	$\widehat{\theta}^{GS}$	440	66.0	11.80

Tab. 1: Results obtained for the linear model

The obtained results are presented in table 1. In every cases, the Tracking estimator gives more precise estimation than NLS and GS estimators (both in term of MSE and ARE), but this improvement is not impressive as MSE is expressed at scale 10^{-5} and ARE at scale 10^{-3}. The differences are small among the different estimation methods and estimations are reliable in all cases (even for the GS approach). This example mainly illustrates tracking approach is a relevant estimation method and can compete with the most used methods for simple model.

5.4.2 Nonlinear w.r.t parameters

Well-specified model We consider a following time dependent model

$$\dot{x} = \frac{\theta_1}{\theta_2^2 + t} x. \tag{21}$$

that is non-linear in parameters. We test two sample sizes $n = 50$ and $n = 20$ (observations are uniformly distributed between 0 and 15) and two noise levels $\sigma = 2$ and $\sigma = 4$. The true parameter value is $\theta^* = (\theta_1^*, \theta_2^*) = (1.4, 1)$ and the initial condition is equal to $X_0^* = 1$. The sequence of lambda used is $\lambda^v = \left\{10^k\right\}_{-1 \leq k \leq 7}$.

(n, σ)		MSE ($\times 10^{-2}$)	ARE ($\times 10^{-2}$)	Pred error
$(50, 2)$				
	$\widehat{\theta}^{T}$	0.18	4.04	7.79
	$\widehat{\theta}^{NLS}$	0.18	4.06	7.79
	$\widehat{\theta}^{GS}$	4.88	19.94	46.27
$(50, 4)$				
	$\widehat{\theta}^{T}$	0.68	8.35	15.57
	$\widehat{\theta}^{NLS}$	0.68	8.35	15.57
	$\widehat{\theta}^{GS}$	9.15	30.10	67.53
$(20, 2)$				
	$\widehat{\theta}^{T}$	0.87	8.93	15.59
	$\widehat{\theta}^{NLS}$	0.87	8.95	15.59
	$\widehat{\theta}^{GS}$	11.11	33.43	82.57
$(20, 4)$				
	$\widehat{\theta}^{T}$	1.30	11.16	15.64
	$\widehat{\theta}^{NLS}$	1.33	11.22	15.63
	$\widehat{\theta}^{GS}$	11.23	32.79	62.79

Tab. 2: Results obtained for the non linear model

The results are presented in table 2. The Tracking and NLS estimators have equivalent performance in the well specified case. The GS approach gives a less precise estimation but it is still reliable, but the prediction error for GS is much more important than for the Tracking and NLS estimators. This illustrates the high sensitivity of ODE model w.r.t parameter and the need of accurate estimates for prediction, even for simple models.

Misspecified model The data is generated by a pertubed model

$$\dot{x} = \frac{\theta_1}{\theta_2^2 + t} x + \sin(t) \tag{22}$$

with $\theta_1^* = 1.4, \theta_2^* = 1$ and $X_0^* = 1$. Nevertheless, we still use the model (21) for the parameter estimation. We use the two sample size $n = 100$ and $n = 50$ (observations were uniformly distributed between 0 and 15) and the two noise levels $\sigma = 2$ and $\sigma = 4$. We use a sequence of hyperparameter $\lambda^v = \left\{10^k, 5 \times 10^k\right\}_{-4 \leq k \leq 0}$.

Moreover, we are interested in the use of the residual control $\overline{u}$ obtained along the parametric estimation in order to propose a "corrected" model:

$$\dot{x} = \frac{\theta_1}{\theta_2^2 + t} x + \overline{u} \tag{23}$$

for minimizing the "corrected" prediction error

$$\mathbb{E}_{\theta^*, \sigma} \left[\left\| \mathrm{Y} - \mathrm{X}_{\widehat{\theta}, \overline{\mathrm{u}}} \right\|_{\mathrm{L}^2}^2 \right] \tag{24}$$

When model misspecification is suspected, we can consider two trajectory predictors $X_{\widehat{\theta},0}$ and $X_{\widehat{\theta},\overline{u}}$ for a given λ. From the definition of the criterion S, the corrected trajectory $X_{\widehat{\theta},\overline{u}}$ is prone to give smaller prediction errors than the misspecified trajectory $X_{\widehat{\theta},0}$. This indicates that we should select the hyperparameter λ in a different way in the case of misspecification, if we want to use the correction $\overline{u}$ that depends also on λ. For this reason, we propose to select λ by minimizing the Corrected Sum of Squared Errors

$$CSSE(\lambda) = \sum_{i=1}^{n} \left(Y_i - X_{\widehat{\theta^T_\lambda},\overline{u}}(t_i) \right)^2 .$$

as a proxy for the prediction errror (24). Finally, we have two tracking estimators $\widehat{\theta}^T$ and $\widehat{\theta}^T_c$ based on two choices of hyperparameter (λ that minimizes SSE and λ_c that minimizes $CSSE$). The estimation of the perturbation (or control u) is done in two ways:

- for the Tracking estimator $\widehat{\theta^T}$, it is provided directly by the estimation process;
- for the NLS estimator $\widehat{\theta}^{NLS}$, we propose to estimate the perturbation based on the nonparametric proxy $\widehat{X}$

$$\overline{u}(t) = \dot{\widehat{X}}(t) - \frac{\widehat{\theta_1}}{\widehat{\theta_2}^2 + t} \widehat{X}(t).$$

For the Generalized Smoothing, we do not have to estimate the perturbation u, as the penalized spline $\widehat{X}(\cdot, \hat{\theta}^{GS})$ already contains the model misspecification.

Results are presented in table 3. The two first columns gives the parametric estimation performance in terms of MSE and ARE. The third column gives an estimation of (19) and the fourth one an estimation of (24).

We can see $\widehat{\theta^T}$ gives more accurate parametric estimation than the NLS estimator. The use of residual control in $X_{\widehat{\theta},\overline{u}}$ improves the prediction error in any case. At the contrary, the correction of the NLS estimate with the estimated control makes things worst, which can be explained by the use of the non-parametric estimator of the derivative. The Generalized Smoothing estimator $\widehat{\theta}^{GS}$ competes well with others approaches, that can be explained by the relaxation introduced by the collocation which makes the method robust in misspecification presence. However, we observe a fast drop in estimation precision w.r.t noise augmentation and poor performance for prediction purpose.

The corrected estimator $\widehat{\theta^c}^T$ gives higher precision than $\widehat{\theta^T}$ for small measurement error σ. We also notice the dramatic drop in prediction error by using the corrected model instead of the initial one. It is due to the fact we effectively take into account an exogeneous perturbation using $\overline{u}$ for parametric estimation. We simultaneously estimate the parametric part and the nonparametric

(n, σ)		MSE ($\times 10^{-2}$)	ARE ($\times 10^{-2}$)	Pred error	Corrected Pred error
$(100, 2)$					
	$\widehat{\theta^T}$	4.31	24.16	8.51	8.16
	$\widehat{\theta^c}^T$	2.88	18.52	16.94	8.03
	$\widehat{\theta^{NLS}}$	4.61	25.11	8.15	8.09
	$\widehat{\theta^{GS}}$	2.39	14.27	42.36	42.27
$(50, 2)$					
	$\widehat{\theta^T}$	4.35	24.05	8.84	8.20
	$\widehat{\theta^c}^T$	3.31	19.18	20.83	8.21
	$\widehat{\theta^{NLS}}$	4.62	24.89	8.20	8.27
	$\widehat{\theta^{GS}}$	4.64	19.87	55.75	55.63
$(100, 4)$					
	$\widehat{\theta^T}$	4.74	24.95	16.14	15.75
	$\widehat{\theta^c}^T$	4.36	21.75	25.23	15.83
	$\widehat{\theta^{NLS}}$	4.99	25.67	15.76	15.82
	$\widehat{\theta^{GS}}$	8.08	27.04	70.25	70.09
$(50, 4)$					
	$\widehat{\theta^T}$	4.83	24.77	16.51	15.84
	$\widehat{\theta^c}^T$	6.10	25.67	27.73	16.01
	$\widehat{\theta^{NLS}}$	5.05	25.49	15.83	16.03
	$\widehat{\theta^{GS}}$	9.18	29.54	82.82	82.72

Tab. 3: Estimation results for misspecified model

part of the model as it is the case. But we can not propose an adaptive way to detect when the nonparametric model correction $\overline{u}$ should be applied, as it is far beyond the scope of that paper.

Finally, we are also interested in the control $\overline{u}$ itself, even though we do not expect it gives us an estimation of the true control $u^*(t) = \sin(t)$ (because of identifiability issues). Its features can give hints about potential misspecification presence and qualitative informations about its nature. We plot in figure 1 the mean control obtained in the case $(n, \sigma) = (100, 2)$ in blue and the true control $u^*(t) = \sin(t)$ in green for the sake of comparison. Although, the scale is not the same, we can see that the estimated control exhibits some features of the u^*, such as oscillations, with the same approximate period. The pertubation could be used as an exploratory tool for analyzing and inferring the missing part in the dynamics.

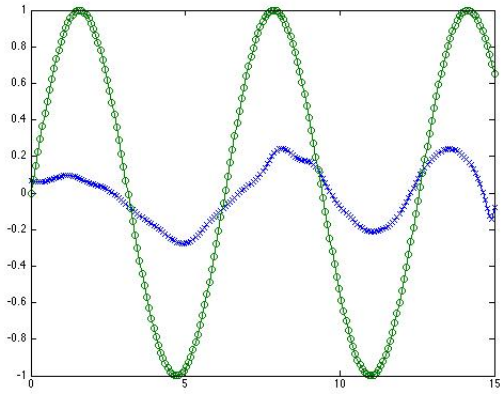

Fig. 1: Obtained mean residual control for $(n, \sigma) = (100, 2)$

5.5 $\alpha-$Pinene model

The "$\alpha-$Pinene model" is a model introduced in [35] for modeling the isomerization of $\alpha-$Pinene. It is an autonomous linear ODE in $\mathbb{R}^5$ with a sparse structure

$$\dot{X}(t) = \begin{pmatrix} -(\theta_1+\theta_2) & 0 & 0 & 0 & 0 \\ \theta_1 & 0 & 0 & 0 & 0 \\ \theta_2 & 0 & -(\theta_3+\theta_4) & 0 & \theta_5 \\ 0 & 0 & \theta_3 & 0 & 0 \\ 0 & 0 & \theta_4 & 0 & -\theta_5 \end{pmatrix} X(t) := A(\theta)X(t)$$

It is an Initial Value Problem, with known initial condition $X_0^* = (100, 0, 0, 0, 0)$. This estimation problem is still considered as cumbersome and many estimation methods fails to converge or converge to bad local solutions because of high correlation between θ_4 and θ_5. Before analyzing the real dataset, we perform a simulation study for evaluating the difficulty of the estimation problem, and benchmarking the several estimators.

5.5.1 Simulated data

The observation interval is $[0, 100]$ and the true parameter is $\theta^* = (5.93,\ 2.96,\ 2.05,\ 27.5, 4) \times 10^{-4}$. Because of dramatic differences in the order of magnitude of the state variables, the noise standard deviation has to be rescaled componentwise. Here for a given σ value the standard deviation applied to the state variable X_i is equal to $\frac{\sigma}{100} \times \frac{1}{T}\int_0^T X_i(t)dt$, $\frac{1}{T}\int_0^T X_i(t)dt$ being the X_i mean value on the observation interval. For the Tracking estimator, we use a sequence of hyperparameter $\lambda^v = \left\{10^k,\ 5 \times 10^k\right\}_{0 \leq k \leq 5}$.

(n, σ)		MSE ($\times 10^{-4}$)	ARE ($\times 10^{-2}$)	Pred error
$(100, 8)$				
	$\widehat{\theta}^{T}$	2.83	1.25	52.72
	$\widehat{\theta}^{NLS}$	2.75	0.87	51.71
	$\widehat{\theta}^{GS}$	3.21	1.72	54.18
$(50, 8)$				
	$\widehat{\theta}^{T}$	2.92	1.35	52.40
	$\widehat{\theta}^{NLS}$	3.14	1.24	52.20
	$\widehat{\theta}^{GS}$	7.05	2.97	54.54
$(100, 16)$				
	$\widehat{\theta}^{T}$	6.08	3.58	103.73
	$\widehat{\theta}^{NLS}$	14	4.8	103.38
	$\widehat{\theta}^{GS}$	15.4	4.06	106.88
$(50, 16)$				
	$\widehat{\theta}^{T}$	11	7.01	103.91
	$\widehat{\theta}^{NLS}$	26	8.47	103.58
	$\widehat{\theta}^{GS}$	25.85	5.36	107.91

Tab. 4: Results obtained for $\alpha-$Pinene

The results are presented in table 4. The Tracking and Generalized Smoothing estimators give more accurate parameter estimation than the Nonlinear Least Squares. In the last case, GS gives the best performance in terms of ARE. In this model, the Relative Error is especially relevant to quantify parameter precision because of important differences between the scale of θ_4 and the other parameters. As expected the difference in performance mainly comes from the estimation of the couple (θ_4, θ_5). This model shows that the NLS is appropriate for parameter estimation whereas the GS seems to favor the quality of prediction, showing that estimation and prediction are somehow two competing objectives. The Tracking estimator realizes a trade-off between these two objectives.

5.5.2 Real data analysis

We use the data coming from [13], and presented in table 5. They consist in simultaneous measures of the 5 components relative concentration at eight time steps. We compare our results with the previous estimation $\widehat{\theta_b} = 10^{-4} \times$ (0.593, 0.296, 0.205, 2.75, 0.4) obtained in [35], which provides a good data fitting. We use the Tracking method for the estimation of the parameter and of the residual control $\overline{u}$. In order to avoid numerical problems, we have divided the observation by 1000 and renormalized the parameters (as the system is autonomous).

For the non-parametric estimator $\widehat{X}$, we use only two nodes: one at the end and one at the beginning of the observation interval. We have also impose to

the non-parametric estimator to start from the known initial condition $X_0^* = (100, 0, 0, 0, 0)$. For the Tracking estimator, we use $\lambda = 10^k$, $k = 1, 2 \cdots, 11$ and $\lambda = 5 \times 10^k$, $k = 1, 2, 3, 4$ and we select the hyper-parameter that minimizes the SSE (named λ_{SSE}. We call the corresponding estimator $\widehat{\theta_{SSE}}^T$) . The estimation results are presented in table 6.

Times (min)	X_1	X_2	X_3	X_4	X_5
1230	88.35	7.3	2.3	0.4	1.75
3060	76.4	15.6	4.5	0.7	2.8
4920	65.1	23.1	5.3	1.1	5.8
7800	50.4	32.9	6	1.5	9.3
10680	37.5	42.7	6	1.9	12
15030	25.9	49.1	5.9	2.2	17
22620	14	57.4	5.1	2.6	21
36420	4.5	63.1	3.8	2.9	25.7

Tab. 5: Experimental data for $\alpha-$pinene model coming from Fuguitt &Hawkins

10^{-4}	θ_1	θ_2	θ_3	θ_4	θ_5	SSE
$\widehat{\theta_{SSE}}^T$	0.589	0.290	0.193	2.301	0.234	23.88
$\widehat{\theta_b}$	0.593	0.296	0.205	2.75	0.4	19.89

Tab. 6: Parameter estimates

We have obtained $\lambda_{SSE} = 100$. For $\lambda \geq 5000$, the estimated values are almost constant and equal to $\theta = (0.583, 0.295, 0.207, 2.259, 0.238)$. The first three estimated parameters are close to the estimates given in [35], but θ_4 and θ_5 are different; however, we obtain good predicted curves, similar to Rodriguez et al., see figure 2.

We can compute the minimal control $\overline{u}_{SSE}$ corresponding to $(\widehat{\theta_{SSE}}^T, \lambda_{SSE})$, and we can also compute the minimal control corresponding to $\theta = \widehat{\theta_b}$ for a given value of λ (that is $\overline{u}_{\widehat{\theta_b},\lambda}$) and compare its norm with $\overline{u}_{SSE}$ when $\lambda = \lambda_{SSE}$. The controls are represented in figure 3 where the curves plotted with $\times$ represent the minimal control obtained for $\widehat{\theta_b}$ and the curves plotted with $\circ$ are the control obtained with $\widehat{\theta_{SSE}}^T$. Here, the Tracking control is a five-dimensional vector, where each entry $\overline{u}_i$ corresponds to one state variable X_i. The control plotted in yellow corresponds to X_1, the one in black correspond to X_2, the one in green correspond to X_3, the one in blue correspond to X_4 and the one in red to X_5.

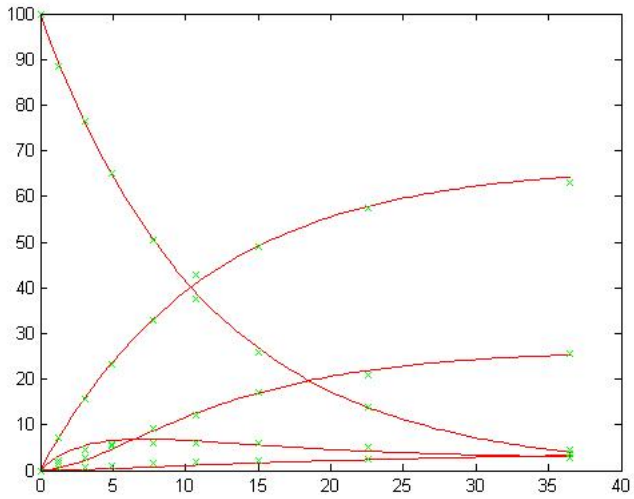

Fig. 2: Reconstructed curve for $\widehat{\theta^T_{SSE}}$ with data in green

As expected, the estimated control $\overline{u}_{SSE}$ is smaller in L^2- norm for $\widehat{\theta_{SSE}}^T$ than for $\widehat{\theta_b}$. Nonetheless according to 3, there is no clear difference between $\overline{u}_{\widehat{\theta_b},\lambda}$ and $\overline{u}_{SSE}$ except for their component related to X_5 (in red on the figure 3) and which is the state variable exclusively related to the parameters θ_4 and θ_5 (the most difficult parameters to estimate according to [35]. Even though the two resulting solutions $X_{\widehat{\theta_b}}$ and $X_{\widehat{\theta_{SSE}}^T}$ are close, the insight given by $\overline{u}_{\widehat{\theta_b},\lambda}$ and $\overline{u}_{SSE}$ shows stronger differences at the dynamic scale.

6 Conclusion

We have introduced a new estimation method for parameter estimation in linear ODE based on relaxation of the initial model. Similarly to the Generalized Smoothing estimator, we end up with a function $X_{\theta,u}$ that is an approximate solution of the ODE model of interest. The added perturbation u enables to take into account the noisy observations but also some uncertainty in the model. Quite remarkably, the trade-off between model and data discrepancy is formulated and solved by using Optimal Control Theory and this work is one of the first use of this theory for dealing with statistical estimation and optimization in infinite dimensional spaces. Moreover, this functional framework and the Riccati theory gives practical algorithms and theoretical insight in the properties of the estimator that permits a detailed analysis of the statistical properties. In addition to the parameter estimate, we obtain also directly an estimate of the

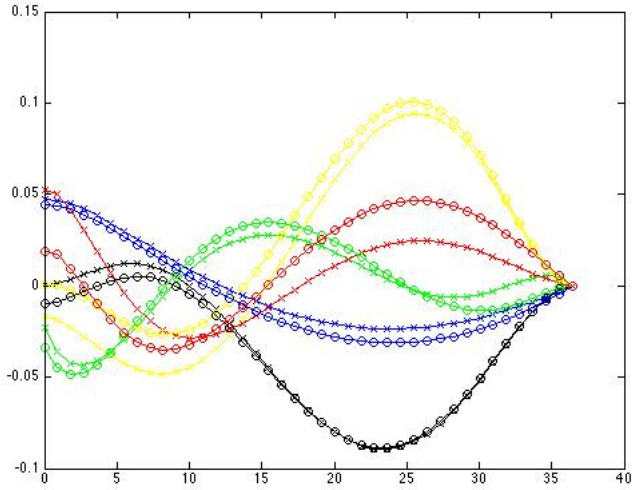

Fig. 3: Estimated control for $\widehat{\theta_b}$ with $\lambda = \lambda_{SSE}$ and $\widehat{\theta_{SSE}}^T$

model discrepancy through the perturbation u that can give hints for analyzing and discussing the relevancy of a parametric model. More can be done with the estimated u about model analysis, and complementary analysis about model testing could be done following the results given in [20].

In the experiments, we show that the Tracking estimator have similar or better performances than nonlinear least squares or generalized smoothing, even in the case of very simple models with closed-form expression. Hence, paradoxically, the use of perturbed model and of nonparametric estimators can ameliorate the statistical efficiency of standard estimates, even in well-specified cases. In the case of model misspecification, the differences are bigger, as the relaxation brought by λ gives us a more robust estimation method (comparing to NLS) which can deal with small model definition imperfection. Moreover, the optimal control obtained for a given parameter estimate allows us to minimize the prediction error by introducing a proper correction term to the initial model. This control can also be used as a qualitative tool to diagnose model misspecification.

However, we are aware of some limitations of our method: first, we assume that the initial condition is known. We can consider X_0^* as an additional parameter to estimate, and reformulate our approach for doing simultaneous observations. The second but the main limitation is the linear assumption about the ODE. Although linear ODEs are common in applications, numerous useful models are nonlinear and thus our methodology cannot be applied directly. Nevetheless, our work can be extended by using more general results of optimal control, such as Pontryagin Maximum Principle that can offer efficient characterization in the general nonlinear case.

Appendix

A Fundamentals Results of Optimal Control: Linear-Quadratic Theory

The "theorem and definition" in section 2.2 is a particular case of a more general theorem which ensures existence and uniqueness of optimal control for cost under the form:

$$C\left(t_0, u, \lambda\right) = z_u(T)^T Q z_u(T) + \int_{t_0}^{T} z_u(t)^T W(t) z_u(t) + u(t)^T U(t) u(t) dt \quad (25)$$

Theorem A.1. *Let $A \in L^2([0,\, T]\,, \mathbb{R}^{d\times d})$ and $B \in L^2([0,\, T]\,, \mathbb{R}^{d\times d})$ We consider z_u the solution of the following ODE:*

$$\dot{z_u}(t) = A(t) z_u(t) + B(t) u(t),\ z(t_0) = z_0$$

and we want to minimize the cost (25) defined on $L^2([0,\, T]\,, \mathbb{R}^d)$, with Q positive, $W \in L^\infty([0,\, T]\,, \mathbb{R}^{d\times d})$ positive matrix for all $t \in [0,\, T]$ and $U(t)$ definite positive matrix for all $t \in [0,\, T]$ respecting the coercivity condition:

$$\exists \alpha > 0\, s.t\, \forall u \in L^2([0,\, T]\,, \mathbb{R}^d)\ :\ \int_0^T u(t)^T U(t) u(t) dt \geq \alpha \int_0^T \|u(t)\|_2^2\, dt$$

It exists a unique optimal trajectory $z_{\bar{u}}$ associated to the unique optimal control $\overline{u}(t) = U^{-1}(t) E(t) B(t) z_{\overline{u}}(t)$ where E is the matrix solution of the Riccati ODE:

$$\begin{aligned} &\dot{E}(t) = W(t) - A(t)^t E(t) - E(t) A(t) - E(t) B(t) U(t)^{-1} B(t)^T E(t) \\ &E(T) = -Q \end{aligned}$$

and the minimal cost is equal to: $C\left(t_0, \overline{u}, \lambda\right) = -z_0^T E(t_0) z_0$.

B Proof & Intermediary results

B.1 $\theta \longmapsto S(\widehat{X};\theta,\lambda)$ and $\theta \longmapsto S(X^*;\theta,\lambda)$ properties

Lemma B.1. *Let us define E the solution of*

$$\begin{aligned} &\dot{E}(t) = W(t) - A(t)^t E(t) - E(t)A(t) - \tfrac{1}{\lambda}E(t)^2 \\ &E(T) = -Q \end{aligned} \tag{26}$$

with $A(t) \in L^2([0,\,T]\,,\mathbb{R}^{d\times d})$ Q bounded,$W \in L^\infty([0,\,T]\,,\mathbb{R}^{d\times d})$.

Then E is bounded on $[0\,,\,T]$.

Proof. (This proof is presented in Sontag's book "Mathematical Control Theory" [37] chapter 7 theorem 30)

By using theorem A.1 and if we define the quadratic cost:

$$C(t_0,u,\lambda) = x_u(T)^T Q x_u(T) + \int_{t_0}^{T} x_u(t)^T W(t) x_u(t) + \lambda \left\|u(t)\right\|_2^2 dt$$

with x_u the ODE solution of

$$\begin{aligned} &\dot{x_u}(t) = A(t)x_u(t) + u(t) \\ &x_u(t_0) = x_0 \end{aligned}$$

We know that

$$\min_u C(t_0,u,\lambda) = -x_0^T E(t_0) x_0$$

Let us reason by contradiction: if $E(t)$is not bounded then $\exists t_e \in [0\,,\,T]$ s.t $\lim_{t\to t_e+} \left\|E(t)\right\|_2 = +\infty$.

It implies:

$$\forall \alpha > 0\; \exists t_0 \in \,]t_e\,,\,T]\,,\; x_0 \in \mathbb{R}^d \,with\, \left\|x_0\right\|_2 = 1 \,s.t\, \left|x_0^T E(t_0) x_0\right| \geq \alpha \tag{27}$$

We also know it exists a unique optimal trajectory for the LQ problem on $[t_0,T]$ with $x(t_0) = x_0$ and the associated optimal cost is $-x_0^T E_\theta(t_0) x_0$. But by minimality of this cost it has to be majored by the cost $C(t_0,0,\lambda)$ i.e the cost associated to the control $u = 0$. We can see it exists a constant $D > 0$ such $C(t_0,0,\lambda)$ is majored by $D\left\|x_0\right\|_2^2$ and so:

$$\left|x_0^T E_\theta(t_0) x_0\right| \leq D$$

which contradict (27) and finish the proof. □

Lemma B.2. *$\forall (t,\theta)$ $h_\theta(t,.)$ is an affine function of X and can be written under the form:*

$$h_\theta(t,X) = N_\theta(t).X$$

With:

$$N_\theta(t).X := \int_t^T R_\theta(T-t,T-s)X(s)ds + E_\theta(t)X(t) + \int_t^T R_\theta(T-t,T-s)E_\theta(s)r_\theta(s)ds$$

with R_θ defined by (16).

Proof. Considering the backward ODE:

$$\begin{cases} \dot{h_{\theta,i}}(t,X) = \alpha_\theta(T-t)h_{\theta,i}(t,X) + \beta_\theta(T-t,X) \\ h_{\theta,i}(0,X) = 0 \end{cases}$$

We know thanks to Duhamel formula:

$$h_{\theta,i}(t,X) = \int_0^t R_\theta(t,s)\beta(T-s,X)ds$$

with R_θ defined by 16.
Hence:

$$\begin{array}{l} h_\theta(t,X) = h_i(T-t,X) = \int_0^{T-t} R_\theta(T-t,s)\beta(T-s,X)ds \\ = \int_t^T R_\theta(T-t,T-s)\beta_\theta(s,X)ds \end{array}$$

Taking the value of β and using integration by part we have:

$$\begin{array}{rcl} h_\theta(t,X) & = & \int_t^T \left(R_\theta(T-t,T-s)E_\theta(s)A_\theta(s) + \frac{d(R_\theta(T-t,T-s)E_\theta(s))}{ds} \right) X(s)ds \\ & + & E_\theta(t)X(t) + \int_t^T R_\theta(T-t,T-s)E_\theta(s)r_\theta(s)ds \end{array}$$

And using resolvant property we finally obtain:

$$\frac{d\left(R_\theta(T-t,T-s)E_\theta(s)\right)}{ds} = R_\theta(T-t,T-s)\left(I_p - E_\theta(s)A_\theta(s)\right)$$

So:

$$h_\theta(t,X) = \int_t^T R_\theta(T-t,T-s)X(s)ds + E_\theta(t)X(t) + \int_t^T R_\theta(T-t,T-s)E_\theta(s)r_\theta(s)ds \tag{28}$$

□

Lemma B.3. *Under conditions 1 and 2,* $\forall X \in H^1([0,T],\mathbb{R}^d)$ *with* $X(0) = x_0^*$ *we have*

$$\begin{array}{rcl} \int_0^T \dot{X}(t)^T h_\theta(t,X)dt & = & F_{1,\theta}(X) + F_{2,\theta}(X) + F_{3,\theta}(X) \\ & - & x_0^{*T} \int_0^T R_\theta(T,T-s)E_\theta(s)r_\theta(s)ds \\ & - & \frac{1}{2}x_0^{*T}E_\theta(0)x_0^* \end{array}$$

$$\textit{with:} \begin{cases} F_{1,\theta}(X) = -x_0^{*T}\int_0^T R_\theta(T,T-s)X(s)ds \\ F_{2,\theta}(X) = \int_0^T X(t)^T\left(\alpha_\theta(t)h_\theta(t,X) + A_\theta(t)X + r_\theta(t)\right)dt \\ F_{3,\theta}(X) = \frac{1}{2}\int_0^T X(t)^T \dot{E_\theta}(t)X(t)dt \end{cases}$$

Proof. Integration by part give us:

$$\begin{array}{l} \int_0^T \dot{X}(t)^T h_\theta(t,X)dt = \left[X(t)^T h_\theta(t,X)\right]_0^T + \int_0^T X(t)^T\left(\alpha_\theta(t)h_\theta(t,X) + \beta_\theta(t,X)\right)dt \\ = -x_0^{*T}h_\theta(0,X) + \int_0^T X(t)^T\left(\alpha_\theta(t)h_\theta(t,X) + E_\theta(t)\left(A_\theta(t)X + r_\theta(t)\right)\right)dt - \int_0^T X(t)^T E_\theta(t)\dot{X}(t)dt \end{array}$$

and

$$\int_0^T X(t)^T E_\theta(t)\dot{X}(t)dt = -\tfrac{1}{2}\left(x_0^{*T}E_\theta(0)x_0^* + \int_0^T X(t)^T \dot{E_\theta}(t)X(t)dt\right)$$

Moreover using affine nature of h w.r.t X and using the same notation as in proposition lemma B.2:

$$\begin{aligned} x_0^{*T}h_\theta(0,X) &= x_0^{*T}\int_0^T R_\theta(T,T-s)X(s)ds + x_0^{*T}E_\theta(0)x_0^* \\ &+ \quad x_0^{*T}\int_0^T R_\theta(T,T-s)E_\theta(s)r_\theta(s)ds \end{aligned}$$

Finally we obtain:

$$\begin{aligned} \int_0^T \dot{X}(t)^T h_\theta(t,X)dt &= -x_0^{*T}\int_0^T R_\theta(T,T-s)X(s)ds - \tfrac{1}{2}x_0^{*T}E_\theta(0)x_0^* \\ &+ \int_0^T X(t)^T\left(\alpha_\theta(t)h_\theta(t,X)dt + (A_\theta(t)X + r_\theta(t))\right)dt \\ &+ \tfrac{1}{2}\int_0^T X(t)^T\dot{E_\theta}(t)X(t)dt \\ &- x_0^{*T}\int_0^T R_\theta(T,T-s)E_\theta(s)r_\theta(s)ds \end{aligned}$$

□

B.2 Consistency Proof

In the following proposition B.4 we show $\left|S(\widehat{X};\theta,\lambda) - S(X^*;\theta,\lambda)\right|$ is controlled by the distance between $\widehat{X}$ and X^* and between $\widehat{h}$ and h^*. In proposition B.5 we show $\left\|\widehat{h_\theta} - h_\theta^*\right\|_{L^2}$ is uniquely controlled by $\left\|\widehat{X} - X^*\right\|_{L^2}$ the same will follow for $|S_\lambda(\theta) - S_\lambda^*(\theta)|$

Proposition B.4. *Under conditions 1 and 3, $\forall\theta\in\Theta$ we have:*

$$\begin{aligned} &\left|S(\widehat{X};\theta,\lambda) - S(X^*;\theta,\lambda)\right| \\ &\leq 2\left(\bar{A}\bar{h} + K_1 + K_2\left\|\widehat{h}_\theta\right\|_{L^2} + K_3\left\|\widehat{X}\right\|_{L^2}\right)\left\|X^* - \widehat{X}\right\|_{L^2} \\ &+ \left(\bar{A}\left\|\widehat{X}\right\|_{L^2} + K_4 + \tfrac{1}{\lambda}\left(\left\|\widehat{h}_\theta\right\|_{L^2} + \bar{h}\right)\right)\left\|h_\theta^* - \widehat{h}_\theta\right\|_{L^2} \end{aligned}$$

$$\textit{With}: \begin{cases} K_1 = \sqrt{d}\left\|X_0\right\|_2\bar{R} + d\bar{E}\bar{A}\bar{X} + \sqrt{d}\bar{\dot{E}}\overline{X} \\ K_2 = \sqrt{d}\left(\bar{A} + \frac{\bar{E}}{\lambda}\right) \\ K_3 = d\bar{E}\bar{A} + \sqrt{d}\bar{\dot{E}} \\ K_4 = \sqrt{d}\left(\bar{A} + \frac{\bar{E}}{\lambda}\right)\bar{X} \end{cases}$$

$$\textit{and:} \begin{array}{l} \bar{R} = \sup_{\theta\in\Theta}\left\|R_\theta(T,T-.)\right\|_{L^2} \\ \bar{\dot{E}} = \sup_{\theta\in\Theta}\left\|\dot{E}_\theta\right\|_{L^2} \end{array}$$

Proof. For the sake of notation we will consider the homogenous case i.e $r_\theta(t) = 0$

By triangular inequality we have:

$$\begin{aligned}
&\left|S(\widehat{X};\theta,\lambda) - S(X^*;\theta,\lambda)\right| \\
&\leq 2\left|\int_0^T \left(h_\theta^*(t)^T A_\theta(t) X^*(t) - \widehat{h}_\theta(t)^T A_\theta(t)\widehat{X}(t)\right) dt\right| \\
&+2\left|\int_0^T \left(\dot{\widehat{X}}(t)^T \widehat{h}_\theta(t) - \dot{X}^*(t)^T h_\theta^*(t)\right) dt\right| \\
&+\tfrac{1}{\lambda}\left|\int_0^T \left(h_\theta^*(t)^T h_\theta^*(t) - \widehat{h}_\theta(t)^T \widehat{h}_\theta(t)\right) dt\right|
\end{aligned}$$

Now we will separatly bound each of the three previous terms:
The first one:

$$\begin{aligned}
&\left|\int_0^T \left(h_\theta^*(t)^T A_\theta(t) X^*(t) - \widehat{h}_\theta(t)^T A_\theta(t)\widehat{X}(t)\right) dt\right| \\
&\leq \left|\int_0^T h_\theta^*(t)^T A_\theta(t)\left(X^*(t) - \widehat{X}(t)\right) dt\right| + \left|\int_0^T \left(h_\theta^*(t) - \widehat{h_\theta}(t)\right)^T A_\theta(t)\widehat{X}(t) dt\right| \\
&\leq \left\|h_\theta^{*T} A_\theta\right\|_{L^2} \left\|X^* - \widehat{X}\right\|_{L^2} + \left\|A_\theta \widehat{X}\right\|_{L^2} \left\|h_\theta^* - \widehat{h_\theta}\right\|_{L^2}
\end{aligned}$$

The last inequality has been obtained thanks to Cauchy-Schwarz inequality.
The second one inequality is a bit cumbersome in terms of computation. For the sake of clarity we left some computational details in lemma B.3 and we obtain with the same notation:

$$\begin{aligned}
\int_0^T \dot{\widehat{X}}(t)^T \widehat{h_\theta}(t) dt &= F_{1,\theta}(\widehat{X}) + F_{2,\theta}(\widehat{X}) + F_{3,\theta}(\widehat{X}) \\
&- x_0^{*T} E_\theta(0) x_0^*
\end{aligned}$$

and:

$$\begin{aligned}
\int_0^T \dot{X}(t)^{*T} h_\theta^*(t) dt &= F_{1,\theta}(X^*) + F_{2,\theta}(X^*) + F_{3,\theta}(X^*) \\
&- x_0^{*T} E_\theta(0) x_0^*
\end{aligned}$$

Hence we can formulate $S(\widehat{X};\theta,\lambda)$ without the derivative form expression and the last decomposition allows us to bound $\left|\int_0^T \left(\dot{\widehat{X}}(t)^T \widehat{h_\theta}(t) - \dot{X}^*(t)^T h_\theta^*(t)\right) dt\right|$ only with $\left\|\widehat{X} - X^*\right\|_{L^2}$ and $\left\|\widehat{h}_\theta - h_\theta^*\right\|_{L^2}$
By use of norm inequalities we obtain the following bounds:

$$\begin{aligned}
\left|F_{1,\theta}(\widehat{X}) - F_{1,\theta}(X^*)\right| &\leq \sqrt{d}\left\|X_0\right\|_2 \bar{R}\left\|\widehat{X} - X^*\right\|_{L^2} \\
\left|F_{2,\theta}(\widehat{X}) - F_{2,\theta}(X^*)\right| &\leq \sqrt{d}\left(\bar{A} + \tfrac{\bar{E}}{\lambda}\right)\left(\left\|\widehat{X} - X^*\right\|_{L^2}\left\|\widehat{h_\theta}\right\|_{L^2} + \bar{X}\left\|\widehat{h}_\theta - h_\theta^*\right\|_{L^2}\right) \\
&+ \sqrt{d}\bar{A}\left(\left\|\widehat{X}\right\|_{L^2} + \bar{X}\right)\left\|\widehat{X} - X^*\right\|_{L^2} \\
\left|F_{3,\theta}(\widehat{X}) - F_{3,\theta}(X^*)\right| &\leq \sqrt{d}\left\|\widehat{X} - X^*\right\|_{L^2} \dot{\bar{E}}\left(\left\|\widehat{X}\right\|_{L^2} + \overline{X}\right)
\end{aligned}$$

And we obtain for the second part:

$$\begin{aligned}
\left|\int_0^T \left(\dot{\widehat{X}}(t)^T \widehat{h}_\theta(t) - \dot{X}(t)^{*T} h_\theta^*(t)\right) dt\right| &\leq \left(K_1 + K_2\left\|\widehat{h_\theta}\right\|_{L^2} + K_3\left\|\widehat{X}\right\|_{L^2}\right)\left\|\widehat{X} - X^*\right\|_{L^2} \\
&+K_4\left\|\widehat{h_\theta} - h_\theta^*\right\|_{L^2}
\end{aligned}$$

$$\text{With: } \begin{cases} K_1 = \sqrt{d}\left\|X_0\right\|_2 \bar{R} + \sqrt{d}\bar{A}\bar{X} + \sqrt{d}\dot{\bar{E}}\overline{X} \\ K_2 = \sqrt{d}\left(\bar{A} + \frac{\bar{E}}{\lambda}\right) \\ K_3 = \sqrt{d}\bar{A} + \sqrt{d}\dot{\bar{E}} \\ K_4 = \sqrt{d}\left(\bar{A} + \frac{\bar{E}}{\lambda}\right)\bar{X} \end{cases}$$

For the third one we have:

$$\begin{aligned} &\left|\int_0^T \left(h_\theta^*(t)^T h_\theta^*(t) - \widehat{h_\theta}(t)^T \widehat{h_\theta}(t)\right) dt\right| \\ &= \left|\int_0^T \left(h_\theta^*(t)^T \left(h_\theta^*(t) - \widehat{h_\theta}(t)\right) - \widehat{h_\theta}(t)^T \left(\widehat{h_\theta}(t) - h_\theta^*(t)\right)\right) dt\right| \\ &\leq \left(\left\|\widehat{h_\theta}\right\|_{L^2} + \|h_\theta^*\|_{L^2}\right) \|h_\theta^* - h_\theta\|_{L^2} \end{aligned}$$

Hence by summing we finish the proof. □

Proposition B.5. *Under conditions 1 and 3* $\forall \theta \in \Theta$ *we have:*

$$\begin{aligned} &\left\|\widehat{h_\theta} - h_\theta^*\right\|_{L^2} \leq K_5 \left\|\widehat{X} - X^*\right\|_{L^2} \\ &\text{with : } K_5 = \sqrt{d}\left(Tde^{\sqrt{d}\left(\overline{A} + \frac{\overline{E}}{\lambda}\right)T} + \overline{E}\right) \end{aligned}$$

Proof. Thanks lemma B.2 we have the following affine dependance of h w.r.t X:

$$\widehat{h_\theta}(t) - h_\theta^*(t) = \int_t^T R_\theta(T-t, T-s)\left(\widehat{X}(s) - X^*(s)\right) ds + E_\theta(t)\left(\widehat{X}(t) - X^*(t)\right)$$

Taking the norm gives us:

$$\begin{aligned} \left\|\widehat{h_\theta}(t) - h_\theta^*(t)\right\|_2 &\leq \left\|\int_t^T R_\theta(T-t, T-s)\left(\widehat{X}(s) - X^*(s)\right) ds\right\|_2 \\ &+ \left\|E_\theta(t)\left(\widehat{X}(t) - X^*(t)\right)\right\|_2 \\ &\leq \sqrt{d}\left(\sqrt{T}de^{\sqrt{d}\left(\overline{A} + \frac{\overline{E}}{\lambda}\right)T}\left\|\widehat{X} - X^*\right\|_{L^2} + \|E_\theta(t)\|_2 \left\|\widehat{X}(t) - X^*(t)\right\|_2\right) \end{aligned}$$

Using condition C1 and C3 and the upper bound $\|R_\theta(T-t, T-s)\|_2 \leq de^{\sqrt{d}\left(\overline{A} + \frac{\overline{E}}{\lambda}\right)T}$ thanks to proposition 3 in supplementary material.

Finally we obtain:

$$\left\|\widehat{h_\theta} - h_\theta^*\right\|_{L^2} \leq \sqrt{d}\left(Tde^{\sqrt{d}\left(\overline{A} + \frac{\overline{E}}{\lambda}\right)T} + \|E_\theta\|_{L^2}\right)\left\|\widehat{X} - X^*\right\|_{L^2}$$

□

B.3 Asymptotic normality proof

The proof of continuity of some functionals useful for proposition 4.3 are left in the supplementary materials, as they require cumbersome computations and they does not provide particular insights in the mechanics of the proofs.

Proposition B.6. *Under conditions 1-5, we have :*

$$\widehat{\theta}^T - \theta^* = 2\frac{\partial^2 S(X^*;\theta^*,\lambda)}{\partial\theta^T\partial\theta}^{-1}\left(\Gamma(\widehat{X}) - \Gamma(X^*)\right) + o_P(1)$$

where $\Gamma : C\left([0\,,\,T]\,,\mathbb{R}^d\right) \to \mathbb{R}^p$ *is a linear functional defined by*

$$\Gamma(X) = \int_0^T \left(\frac{\partial\left(A_{\theta*}(t).X^*\right)}{\partial\theta} + \frac{1}{\lambda}\frac{\partial h_{\theta^*}(t,X^*)}{\partial\theta}\right)^T \left(\int_t^T R_{\theta^*}(T-t,T-s)X(s)ds\right)dt. \tag{29}$$

R_{θ^*} *is defined* by (16).

Proof. For the sake of notational simplicity here $\widehat{\theta}^T$ will be simply denoted $\widehat{\theta}$.

The first order optimal condition is

$$\nabla_\theta S(\hat{X};\widehat{\theta},\lambda) = 0$$

Equivalently, we have

$$\int_0^T \frac{\partial\left(A_{\widehat{\theta}}(t).\widehat{X} + r_{\widehat{\theta}}(t)\right)}{\partial\theta}^T h_{\widehat{\theta}}(t,\widehat{X}) + \frac{\partial h_{\widehat{\theta}}(t,\widehat{X})}{\partial\theta}^T \left(A_{\widehat{\theta}}(t).\widehat{X} + r_{\widehat{\theta}}(t) - \dot{\widehat{X}}\right) + \frac{1}{\lambda}\frac{\partial h_{\widehat{\theta}}(t,\widehat{X})}{\partial\theta}^T h_{\widehat{\theta}}(t,\widehat{X}) = 0 \tag{30}$$

We use the following decomposition for $A_{\widehat{\theta}}(t).\widehat{X} - \dot{\widehat{X}}$ and $h_{\widehat{\theta}}(t,\widehat{X})$:

$$\begin{aligned} A_{\widehat{\theta}}(t).\widehat{X} + r_{\widehat{\theta}}(t) - \dot{\widehat{X}} &= A_{\widehat{\theta}}(t)\left(\widehat{X} - X^*\right) + \frac{\partial\left(A_{\widetilde{\theta}}(t).X^* + r_{\widetilde{\theta}}(t)\right)}{\partial\theta}\left(\widehat{\theta} - \theta^*\right) + \left(\dot{X}^* - \dot{\widehat{X}}\right) \\ h_{\widehat{\theta}}(t,\widehat{X}) &= \frac{\partial\left(h_{\widetilde{\theta}}(t,\widehat{X})\right)}{\partial\theta}\left(\widehat{\theta} - \theta^*\right) + N_{\theta^*}(t).\left(\widehat{X} - X^*\right) \end{aligned}$$

with $\widetilde{\theta}$ being a random point between θ^* and $\widehat{\theta}$ and N defined as in lemma B.2.
By replacing in (30), we obtain:

$$\int_0^T H_1(t,\widehat{\theta},\widehat{X})dt\,\left(\widehat{\theta} - \theta^*\right) = \int_0^T H_2(t,\widehat{\theta},\widehat{X})\left(\widehat{X} - X^*\right) - \frac{\partial\left(h_{\widehat{\theta}}(t,\widehat{X})\right)}{\partial\theta}^T \left(\dot{X}^* - \dot{\widehat{X}}\right)dt \tag{31}$$

with

$$\begin{aligned} H_1(t,\widehat{\theta},\widehat{X}) &= \frac{\partial\left(A_{\widehat{\theta}}(t).\widehat{X} + r_{\widehat{\theta}}(t)\right)}{\partial\theta}^T \frac{\partial h_{\widetilde{\theta}}(t,\widehat{X})}{\partial\theta} + \frac{\partial\left(h_{\widehat{\theta}}(t,\widehat{X})\right)^T}{\partial\theta}\frac{\partial\left(A_{\widetilde{\theta}}(t).X^* + r_{\widetilde{\theta}}(t)\right)}{\partial\theta} + \frac{1}{\lambda}\frac{\partial\left(h_{\widehat{\theta}}(t,\widehat{X})\right)}{\partial\theta}^T \frac{\partial h_{\widetilde{\theta}}(t,\widehat{X})}{\partial\theta} \\ H_2(t,\widehat{\theta},\widehat{X}) &= \frac{\partial\left(A_{\widehat{\theta}}(t).\widehat{X} + r_{\widehat{\theta}}(t)\right)}{\partial\theta}^T N_{\theta^*}(t) + \frac{\partial\left(h_{\widehat{\theta}}(t,\widehat{X})\right)^T}{\partial\theta}A_{\widehat{\theta}}(t) + \frac{1}{\lambda}\frac{\partial\left(h_{\widehat{\theta}}(t,\widehat{X})\right)}{\partial\theta}^T N_{\theta^*}(t) \end{aligned}$$

Thanks to propositions in supplementary material, the following functionals

$$\begin{cases} D_1 : \theta \longmapsto (t \longmapsto A_\theta(t)) \\ D_2 : (\theta, X) \longmapsto \left(t \longmapsto \frac{\partial(A_\theta(t).X)}{\partial\theta}\right) \\ D_3 : (\theta, X) \longmapsto (t \longmapsto h_\theta(t,X)) \\ D_4 : (\theta, X) \longmapsto \left(t \longmapsto \frac{\partial(h_\theta(t,X))}{\partial\theta}\right) \end{cases}$$

are continuous on $\Theta \times L^2\left([0\,,\,T]\,,\mathbb{R}^d\right)$, and the continuous mapping theorem implies that $t \longmapsto H_1(t,\widehat{\theta},\widehat{X})$ and $t \longmapsto H_2(t,\widehat{\theta},\widehat{X})$ converge in probability in the L^2 sense to the function $t \longmapsto H_1(t,\theta^*,X^*)$ and $t \longmapsto H_2(t,\theta^*,X^*)$. So $\left\|H_1(.,\widehat{\theta},\widehat{X})\right\|_{L^2}$ converges in probability to $\|H_1(.,\theta^*,X^*)\|_{L^2}$ and so it is bounded. Finally, we have the convergence in probability of each entry of $\int_0^T H_1(t,\widehat{\theta},\widehat{X})dt$ to the corresponding entry to $\int_0^T H_1(t,\theta^*,X^*)dt$. Moreover, condition C5 assumes that the Hessian

$$\int_0^T H_1(t,\theta^*,X^*)dt = \frac{1}{2}\frac{\partial^2 S(X^*;\theta^*,\lambda)}{\partial\theta^T\partial\theta}$$

is nonsingular at $\theta=\theta^*$. Finally, we have

$$\int_0^T H_1(t,\widehat{\theta},\widehat{X})dt \xrightarrow{P} \frac{1}{2}\frac{\partial^2 S(X^*;\theta^*,\lambda)}{\partial\theta^T\partial\theta}$$

By an analogous reasoning, the asymptotic behavior of $\widehat{\theta}-\theta^*$ is given by

$$2\frac{\partial^2 S(X^*;\theta^*,\lambda)}{\partial\theta^T\partial\theta}^{-1}\left(\int_0^T H_2(t,\theta^*,X^*)\left(\widehat{X}-X^*\right)dt - \frac{\partial\left(h_{\theta^*}(t,X^*)\right)}{\partial\theta}^T\left(\dot{X}^*-\dot{\widehat{X}}\right)dt\right)$$

and Integration By Part gives

$$\begin{array}{rcl}\int_0^T \frac{\partial h_{\theta^*}(t,X^*)}{\partial\theta}^T\left(\dot{X}^*-\dot{\widehat{X}}\right)dt & = & \left[\frac{\partial h_{\theta^*}(t,X^*)}{\partial\theta}^T\left(X^*-\widehat{X}\right)\right]_0^T \\ & - & \int_0^T \frac{d}{dt}\left(\frac{\partial h_{\theta^*}(t,X^*)}{\partial\theta}^T\right)\left(X^*-\widehat{X}\right)dt\end{array}$$

But, as $\frac{\partial h(T,\theta^*,X^*)}{\partial\theta}=0$ and $\widehat{X}(0)=x_0^*$ we have:

$$\begin{array}{l}\int_0^T\left(H_2(t,\theta^*,X^*)+\frac{d}{dt}\left(\frac{\partial h_{\theta^*}(t,X^*)}{\partial\theta}^T\right)\right).X(t)dt \\ =\int_0^T\left(\frac{\partial(A_{\theta*}(t).X^*+r_{\theta^*}(t))}{\partial\theta}+\frac{1}{\lambda}\frac{\partial h_{\theta^*}(t,X^*)}{\partial\theta}\right)^T\left(h_{\theta^*}(t,X)-E_{\theta^*}(t).X(t)\right)dt\end{array}$$

Hence we can write

$$\widehat{\theta}-\theta^* = 2\frac{\partial^2 S(X^*;\theta^*,\lambda)}{\partial\theta^T\partial\theta}^{-1}\left(\Gamma(\widehat{X})-\Gamma(X^*)\right)+o_P(1)$$

with

$$\Gamma(X)=\int_0^T\left(\frac{\partial\left(A_{\theta*}(t).X^*\right)}{\partial\theta}+\frac{1}{\lambda}\frac{\partial h_{\theta^*}(t,X^*)}{\partial\theta}\right)^T\left(\int_t^T R_{\theta^*}(T-t,T-s)X(s)ds\right)dt$$

where R_{θ^*} is defined by (16). □

Proposition B.7. *Under conditions 1-8 and by defining* Γ *as in proposition B.6 we have that* $\Gamma(\widehat{X}) - \Gamma(X^*)$ *is asymptotically normal and* $\Gamma(\widehat{X}) - \Gamma(X^*) = O_P(n^{-1/2})$

Proof. This proposition is a direct consequence of Theorem 9 in [31]. The conditions to be satisfied are

1. (Y_i, t_i) are i.i.d with $Var(Y \mid t)$ bounded.
2. $E((Y - X^*(t))^4 \mid t)$ is bounded, and $Var(Y \mid t)$ is bounded away from 0.
3. The support of t is a compact interval on which t has a probability density function bounded away from 0.
4. There is $v(t)$ such that $E(v(t)v(t)^T)$ is finite and non-singular such that: $D(\Gamma)(X^*)(X^*) = E(v(t)X^*(t))$ and $D(\Gamma)(X^*)(p_{kK}) = E(v(t)p_{kK}(t))$ for all k and K and there is β_K with $E(\|v(t) - \beta_K p_K(t)\|_2^2) \to 0$
5. $X^*(t) = E(Y \mid t)$ is derivable of order s on the support of t.

Requirements 1,2,3 are direct consequences of conditions C6 and C7 (and the solution is always defined on $[0, T]$).

For the fourth requirement we will consider the monodimensional case $d = 1$. We know that Γ is linear and continuous on $L^2\left([0, T], \mathbb{R}^d\right)$ thanks to conditions C1 and C3-4 and hence differentiable with: $D(\Gamma)(X^*)(X) = \Gamma(X)$. By the Riesz-Frechet representation theorem we have: $v \in L^2([0, T], \mathbb{R})$ s.t $\Gamma(X) = \int_0^T v(t)X(t)dt$ which verify the three conditions of the forth requirement. Starting from the mono-dimensional case, multi-dimensional case can be made componentwise.

Requirement 5 is a simple consequence of the condition C8. □

C Gradient Computation : Adjoint Method & Sensitivity equation

C.1 Notation and partial derivative computation

For optimization purpose we need to compute the gradient of $S(\widehat{X};\theta,\lambda)$. For this we will present two methods: a direct approach using sensitivity equation and a second one using adjoint method.

C.1.1 Row vector notation for the vector field of the general Riccati equation

We will define the solution of the general Riccati equation in row formulation, we introduce

$$Q_\theta(t) = \left(\widehat{h_\theta}^T, (E_\theta^r)^T\right)^T (t)$$

with $E_\theta^r := \left(E_{\theta,1}^T, \cdots, E_{\theta,d}^T\right)^T$ the row formulation of E_θ, $E_{\theta,i}$ beeing the $i-th$ column of E_θ. It is a $D := d^2 + d$ sized function respecting the ODE :

$$\begin{array}{l} \dot{Q}_\theta = F(Q_\theta, \theta, t) \\ Q_\theta(T) = 0 \end{array}$$

by introducing the general vector field F:

$$F(Q_\theta,\theta,t) = \begin{pmatrix} G(Q_\theta,\theta,t) \\ H(Q_\theta,\theta) \end{pmatrix}$$

with G and H defined by:

$$\begin{array}{lcl} G(Q_\theta,\theta,t) & := & -\left(A_\theta(t)^T + \frac{E_\theta}{\lambda}\right)\widehat{h_\theta} - E_\theta\left(A_\theta(t)\widehat{X}(t) - \dot{\widehat{X}}(t) + r_\theta(t)\right) \\ H_{(j-1)d+i}(Q_\theta,\theta) & := & \delta_{i,j} - (A_{\theta,i}^T E_j + A_{\theta,j}^T E_{\theta,i} + \frac{1}{\lambda} E_{\theta,i}^T E_{\theta,j}) \end{array}$$

and $A_{\theta,i}$ beeing the $i-th$ column of A_θ.
We also introduce:

$$g(Q_\theta,\theta,t) = -2\left(A_\theta(t)\widehat{X}(t) - \dot{\widehat{X}}(t) + r_\theta(t)\right)^T \widehat{h_\theta} - \frac{1}{\lambda}\widehat{h_\theta}^T\widehat{h_\theta}$$

In order to write our system under the row form:

$$\begin{array}{l} S(\widehat{X};\theta,\lambda) := \int_0^T g(Q_\theta(t),\theta,t)dt \\ \left\{ \begin{array}{l} \dot{Q}_\theta = F(Q_\theta,\theta,t) \\ Q_\theta(T) = 0 \end{array} \right. \end{array} \tag{32}$$

For the next subsections we will drop dependence in θ for A_θ, r_θ, E_θ, $\widehat{h_\theta}$

C.1.2 Partial derivative of Riccati vector field

In order to compute sensitivity equation or adjoint model we need to compute $\frac{\partial g}{\partial \theta}(Q_\theta,\theta,t)$, $\frac{\partial g}{\partial Q}(Q_\theta,\theta,t)$, $\frac{\partial F}{\partial \theta}(Q_\theta,\theta,t)$ and $\frac{\partial F}{\partial Q}(Q_\theta,\theta,t)$

The computation for $\frac{\partial g}{\partial \theta}(Q_\theta,\theta,t)$, $\frac{\partial g}{\partial Q}(Q_\theta,\theta,t)$ is straightforward

$$\frac{\partial g}{\partial \theta}(Q_\theta,\theta,t) = -2\widehat{h}^T\left(\frac{\partial\left(A(t)\widehat{X}(t)\right)}{\partial\theta}+\frac{\partial r}{\partial\theta}(t)\right)$$

$$\frac{\partial g}{\partial Q}(Q_\theta,\theta,t) = \left(-2\left(A(t)\widehat{X}(t)-\dot{\widehat{X}}(t)+r(t)+\frac{\widehat{h}}{\lambda}\right)^T, 0_{1,d^2}\right)$$

For $\frac{\partial F}{\partial \theta}(h,E^r,\theta,t)$ and $\frac{\partial F}{\partial Q}(R_\theta,\theta,t)$ we obtain

$$\frac{\partial F}{\partial \theta}(Q_\theta,\theta,t) = \begin{pmatrix} \frac{\partial G}{\partial\theta}(Q_\theta,\theta,t) \\ \frac{\partial H}{\partial\theta}(Q_\theta,\theta) \end{pmatrix}$$

$$\frac{\partial F}{\partial Q}(Q_\theta,\theta,t) = \begin{pmatrix} -\left(A(t)^T+\frac{E}{\lambda}\right) & \frac{\partial G_i}{\partial E^r_j}(Q_\theta,\theta,t) \\ 0_{d^2,d} & \frac{\partial H(Q_\theta,\theta)}{\partial E^r} \end{pmatrix}$$

with:

$$\frac{\partial G_i}{\partial E^r_{(k-1)d+h}}(Q_\theta,\theta,t) = -\delta_{i,h}\left(\frac{\widehat{h}}{\lambda}+A(t)\widehat{X}(t)-\dot{\widehat{X}}(t)+r(t)\right)_k$$

$$\frac{\partial G}{\partial\theta}(Q_\theta,\theta,t) = -\left(h^T\frac{\partial A_i(t)}{\partial\theta}\right)_{1\le i\le d} - E\left(\frac{\partial\left(A(t)\widehat{X}(t)\right)}{\partial\theta}+\frac{\partial r(t)}{\partial\theta}\right)$$

We also need to compute $H(Q_\theta,\theta)$ partial derivative w.r.t E^r and θ.
We have:

$$\left(\frac{\partial H(Q_\theta,\theta)}{\partial E^r}\right)_{(j-1)d+i} = -\begin{pmatrix} 0 & A^t_j & 0 & A^t_i & 0 \end{pmatrix} - \frac{1}{\lambda}\begin{pmatrix} 0 & E^t_j & 0 & E^t_i & 0 \end{pmatrix}$$

Because:

- $\frac{\partial}{\partial E^r}\left(A^t_jE_i+A^t_iE_j\right) = \begin{pmatrix} 0 & A^t_j & 0 & A^t_i & 0 \end{pmatrix}$ where A^t_j is in $i-th$ position and A^t_i is in $j-th$ position.
- $\frac{1}{\lambda}\frac{\partial}{\partial E}\left(E^t_jE_i\right) = \begin{pmatrix} 0 & \frac{1}{\lambda}E^t_j & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & \frac{1}{\lambda}E^t_i & 0 & 0 \end{pmatrix}$ where E^t_j is in $i-th$ position and E^t_i is in $j-th$ position.

And:

$$\left(\frac{\partial H(Q_\theta,\theta)}{\partial\theta}\right)_{(j-1)d+i} = -E^t_i\frac{\partial A_j}{\partial\theta} - E^t_j\frac{\partial A_i}{\partial\theta}$$

- Because $\frac{\partial}{\partial\theta}\left(A^t_jE_i+A^t_iE_j\right) = E^t_i\frac{\partial A_j}{\partial\theta}+E^t_j\frac{\partial A_i}{\partial\theta}$ where $\frac{\partial A_i}{\partial\theta} = \left(\frac{\partial A_i}{\partial\theta_1}\cdots\frac{\partial A_i}{\partial\theta_p}\right)$ a $d\times p$ matrix

C.2 Gradient computation by sensitivity equation

By Gradient definition we have

$$\nabla_\theta S(\widehat{X};\theta,\lambda) = \int_0^T \frac{\partial g(Q_\theta(t),\theta,t)}{\partial Q}\frac{\partial Q_\theta(t)}{\partial \theta} + \frac{\partial g(Q_\theta(t),\theta,t)}{\partial \theta}dt$$

With $\frac{\partial Q_\theta(t)}{\partial \theta}$ solution of the sensitivity equation:

$$\frac{d}{dt}(\frac{\partial Q_\theta(t)}{\partial \theta}) = \frac{\partial F}{\partial Q}(Q_\theta(t),\theta,t)\frac{\partial Q_\theta(t)}{\partial \theta} + \frac{\partial F}{\partial \theta}(Q_\theta(t),\theta,t)$$

And we know that $Q_\theta(T) = 0$ so $\frac{\partial Q_\theta(T)}{\partial \theta} = 0$, hence we can obtain $\frac{\partial Q_\theta(t)}{\partial \theta}$ by solving the Cauchy problem:

$$\begin{array}{l} \frac{d}{dt}(\frac{\partial Q_\theta(t)}{\partial \theta}) = \frac{\partial F}{\partial Q}(Q_\theta(t),\theta,t)\frac{\partial Q_\theta(t)}{\partial \theta} + \frac{\partial F}{\partial \theta}(Q_\theta(t),\theta,t) \\ \frac{\partial Q_\theta(T)}{\partial \theta} = 0 \end{array}$$

C.3 Gradient computation by adjoint Method

Once again we have

$$\nabla_\theta S(\widehat{X};\theta,\lambda) = \int_0^T \frac{\partial g(Q_\theta(t),\theta,t)}{\partial Q}\frac{\partial Q_\theta(t)}{\partial \theta} + \frac{\partial g(Q_\theta(t),\theta,t)}{\partial \theta}dt$$

With $\frac{\partial Q_\theta(t)}{\partial \theta}$ solution of the sensitivity equation:

$$\frac{d}{dt}(\frac{\partial Q_\theta(t)}{\partial \theta}) = \frac{\partial F}{\partial Q}(Q_\theta(t),\theta,t)\frac{\partial Q_\theta(t)}{\partial \theta} + \frac{\partial F}{\partial \theta}(Q_\theta(t),\theta,t)$$

If we premultiply the right and left term of the previous ODE by the $D-$sized adjoint vector $P(t)$ and then integrate we obtain

$$\int_0^T P(t).\frac{d}{dt}(\frac{\partial Q_\theta(t)}{\partial \theta})dt = \int_0^T P(t).\frac{\partial F}{\partial Q}(Q_\theta(t),\theta,t)\frac{\partial Q_\theta(t)}{\partial \theta}dt + \int_0^T P(t).\frac{\partial F}{\partial \theta}(Q_\theta(t),\theta,t)dt$$

Integration by part gives us

$$\int_0^T P(t).\frac{d}{dt}(\frac{\partial Q_\theta(t)}{\partial \theta})dt = P(T).\frac{\partial Q_\theta(T)}{\partial \theta} - P(0).\frac{\partial Q_\theta(0)}{\partial \theta} - \int_0^T \dot{P}(t).\frac{\partial Q_\theta(t)}{\partial \theta}dt$$

We already know that $\frac{\partial Q_\theta(T)}{\partial \theta} = 0$ and if we take $P(0) = 0$ we obtain the variational relation:

$$\int_0^T \left(\dot{P}(t) + P(t).\frac{\partial F}{\partial Q}(Q_\theta(t),\theta,t)\right)\frac{\partial Q_\theta(t)}{\partial \theta}dt + \int_0^T P(t).\frac{\partial F}{\partial \theta}(Q_\theta(t),\theta,t)dt = 0$$

and by imposing:

$$\dot{P}(t) + P(t).\frac{\partial F}{\partial Q}(Q_\theta, \theta, t) = \frac{\partial g(Q_\theta(t), \theta, t)}{\partial Q}$$

we deduce that

$$\int_0^T \frac{\partial g(Q_\theta(t), \theta, t)}{\partial Q} \frac{\partial Q_\theta(t)}{\partial \theta} dt = -\int_0^T P(t).\frac{\partial Q}{\partial \theta}(Q_\theta(t), \theta, t) dt$$

and so

$$\nabla_\theta S(\widehat{X}; \theta, \lambda) = \int_0^T \frac{\partial g(Q_\theta(t), \theta, t)}{\partial \theta} - P(t).\frac{\partial F}{\partial \theta}(Q_\theta(t), \theta, t) dt$$

We propose here an alternative for gradient computation, we compute $\nabla_\theta S(\widehat{X}; \theta, \lambda)$ by considering:

$$\begin{array}{l} \nabla_\theta S(\widehat{X}; \theta, \lambda) = \int_0^T \frac{\partial g(Q_\theta(t), \theta, t)}{\partial \theta} - P(t).\frac{\partial F}{\partial \theta}(Q_\theta(t), \theta, t) dt \\ \dot{P}(t) = \frac{\partial g(Q_\theta(t), \theta, t)}{\partial Q} - P(t).\frac{\partial F}{\partial Q}(Q_\theta(t), \theta, t) \\ P(0) = 0 \end{array}$$

The interest here is computational, computing gradient by solving sensitivity equation drives us to solve a $D \times p$ ODE system. Here the adjoint system defining P is only of size D.

D Asymptotic variance expression

We know asymptotically $\widehat{\theta^T} - \theta^*$ behaves as:

$$2\frac{\partial^2 S(X^*;\theta^*,\lambda)}{\partial\theta^T\partial\theta}^{-1}\left(\Gamma(\widehat{X}) - \Gamma(X^*)\right)$$

with:

$$\frac{1}{2}\frac{\partial^2 S(X^*;\theta^*,\lambda)}{\partial\theta^T\partial\theta} =$$

$$\frac{\partial\left(A_{\theta^*}(t).X^*\right)}{\partial\theta}^T\frac{\partial h_{\theta^*}(t,X^*)}{\partial\theta}+\frac{\partial\left(h_{\theta^*}(t,X^*)\right)^T}{\partial\theta}\frac{\partial\left(A_{\theta^*}(t).X^*\right)}{\partial\theta}+\frac{1}{\lambda}\frac{\partial\left(h_{\theta^*}(t,X^*)\right)}{\partial\theta}^T\frac{\partial h_{\theta^*}(t,X^*)}{\partial\theta}$$

the hessian of the asymptotic criteria at $\theta = \theta^*$
and:

$$\Gamma(X) = \int_0^T\left(\frac{\partial\left(A_{\theta*}(t).X^*\right)}{\partial\theta} + \frac{1}{\lambda}\frac{\partial h_{\theta^*}(t,X^*)}{\partial\theta}\right)^T\left(\int_t^T R_{\theta^*}(T-t,T-s)X(s)ds\right)dt$$

A linear functional w.r.t to X so asymptotically:

$$Var(\widehat{\theta^T}) = 4\frac{\partial^2 S(X^*;\theta^*,\lambda)}{\partial\theta^T\partial\theta}^{-1}Var(\Gamma(\widehat{X}))\frac{\partial^2 S(X^*;\theta^*,\lambda)}{\partial\theta^T\partial\theta}^{-1}$$

If $\widehat{X}$ is a b-Splines basis decomposition estimator under the form $\widehat{X} = \sum_{i=1}^K \widehat{\beta}_{iK}p_{iK}(t)$ we can formulate Γ as a linear function w.r.t coefficients $\widehat{\beta}_{iK}$:

$$\Gamma(\widehat{X}) := P(\theta^*,X^*)\widehat{\beta}_K$$

with:

$$P_i(\theta,X) = \int_0^T\left(\frac{\partial\left(A_\theta(t).X\right)}{\partial\theta} + \frac{1}{\lambda}\frac{\partial h_\theta(t,X)}{\partial\theta}\right)^T\left(\int_t^T R_\theta(T-t,T-s)p_{iK}(s)ds\right)dt$$

the $i-$th columns
Finally the asymptotic variance of $\widehat{\theta^T}$ is equal to:

$$Var(\widehat{\theta^T}) = 4\frac{\partial^2 S(X^*;\theta^*,\lambda)}{\partial\theta^T\partial\theta}^{-1}P(\theta^*,X^*)Var(\widehat{\beta}_K)P(\theta^*,X^*)^T\frac{\partial^2 S(X^*;\theta^*,\lambda)}{\partial\theta^T\partial\theta}^{-1} \tag{33}$$

and we can use the consistent estimator:

$$\widehat{Var(\widehat{\theta^T})} = 4\frac{\partial^2 S(\widehat{X};\widehat{\theta^T},\lambda)}{\partial\theta^T\partial\theta}^{-1}P(\widehat{\theta^T},,\widehat{X})\widehat{Var(\widehat{\beta}_K)}P(\widehat{\theta^T},,\widehat{X})^T\frac{\partial^2 S(\widehat{X};\widehat{\theta^T},\lambda)}{\partial\theta^T\partial\theta}^{-1}$$

References

[1] E. Blayo, E. Cosme, M. Nodet, and A. Vidart. Introduction to data assimilation. 2011.

[2] C. De Boor. *A practical guide to Splines*, volume 27 of *Applied Mathematical Sciences*. Spinger, 2001.

[3] H. Brezis. *Functional Analysis*. Dunod, 1983.

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

[5] N. J-B. Brunel, Q. Clairon, and F. D'Alche-Buc. Parameter estimation of ordinary differential equations with orthogonality conditions. *JASA*, 109(205):173–185, 2014.

[6] Ben Calderhead and Mark Girolami. Estimating Bayes factors via thermodynamic integration and population MCMC. *Computational Statistics & Data Analysis*, 53(12):4028–4045, October 2009.

[7] D.A. Campbell and O. Chkrebtii. Maximum profile likelihood estimation of differential equation parameters through model based smoothing state estimates. *Mathematical Biosciences*, 2013.

[8] Y. Cao, S. Li, L. Petzold, and R. Serban. Adjiont sensitivity analysis for differential-algebraic equations: the adjoint dae system and its numerical solution. *SIAM J. on Scientific Computing*, 24(3):1076–1089, 2003.

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

[10] G. Hooker D.A. Campbell and K. B. McAuley. Parameter estimation in differential equation models with constrained states. *Journal of Chemometrics*, 26:322–332, 2011.

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

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

[13] R. E. Fuguitt and J.E. Hawkins. Rate of Thermal Isomerization of a-Pinene in the Liquid Phase. *J.A.C.S*, 319(39), 1947.

[14] A. Gelman, F. Bois, and J. Jiang. Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association*, 91, 1996.

[15] O. Ghasemi, M. Lindsey, T. Yang, N. Nguyen, Y. Huang, and Y. Jin. Bayesian parameter estimation for nonlinear modelling of biological pathways. *BMC Systems Biology*, 5, 2011.

[16] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. volume 73, pages 1–37, 2011.

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

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

[19] G. Hooker. Forcing function diagnostics for nonlinear dynamics. *Biometrics*, 65:928–936, 2009.

[20] G. Hooker and S. Ellner. Goodness of fit in nonlinear dynamics: Mis-specified rates or mis-specified states? Technical report, Cornell University, 2013. arXiv:1312.0294.

[21] Y. Huang and H. Wu. A bayesian approach for estimating antiviral efficacy in hiv dynamic models. *Journal of Applied Statistics*, 33:155–174, 2006.

[22] B. Hipszer T. V. Apanasovich I. Chervoneva, B. Freydin and J.I. Joseph. Estimation of nonlinear differential equation model for glucose-insulin dynamics in type i diabetic patients using generalized smoothing. *Annals of Applied Statistics(submitted)*, 2014.

[23] D. Kaschek and J. Timmer. A variational approach to parameter estimation in ordinary differential equations. *BMC Systems Biology*, 6:99, 2012.

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

[25] H.L. Koul. Weighted empiricals and linear models. *Hayward, CA: Institute of Mathematical Statistics*, 21:105–175, 1992.

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

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

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

[29] A.A. Milyutin and N.P. Osmolovskii. *Calculus of Variation and Optimal control*. Mathematical Monographs. American mathematical society, 1998.

[30] H.P. Mirsky, A.C. Liu, D.K. Welsh, S.A. Kay, and F.J. Doyle III. A model of the cell-autonomous mammalian circadian clock. *PNAS*, 106(27):11107–11112, July 2009.

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

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

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

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

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

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

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

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

[39] 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.