Electronic Journal of Statistics
Vol. 9 (2015) 2903–2949
ISSN: 1935-7524
DOI: 10.1214/15-EJS1086

A tracking approach to parameter estimation in linear ordinary differential equations

Nicolas J. B. Brunel* and **Quentin Clairon**

LaMME – Laboratoire de Mathématiques et Modélisation d'Evry, UMR CNRS 8071
ENSIIE & Université d'Evry Val d'Essonne, 91025 Evry, France
e-mail: nicolas.brunel@ensiie.fr; quentin.clairon@univ-evry.fr

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-posed 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 estimators and Generalized Smoothing (introduced by Ramsay et al. [37]). Our estimation method relies on a relaxation and penalization scheme to regularize the inverse problem. We introduce a perturbed model and we use optimal control theory for constructing a criterion that aims at minimizing the discrepancy between data and the original 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 causes of inaccuracy in Two-Step estimators. Regarding the theoretical asymptotic behavior of our estimator, we show its consistency and that we reach the parametric $\sqrt{n}$-rate when regression splines are used in the first step. We consider the estimation of two models possessing *sloppy parameters*, which usually makes the estimation of ODE models an ill-posed problem in applications [20, 41] and shows the efficiency of the Tracking estimator. Quite interestingly, our relaxation scheme makes the estimator robust to some kind of model misspecification, as shown in simulations.

Keywords and phrases: Parameter estimation, ordinary differential equations, optimal control, Riccati differential equation, smoothing, plug-in property, asymptotic statistics.

Received September 2014.

Contents

*Corresponding author.

1. Introduction

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

$$\begin{cases} \dot{x} & = f(t,x,\theta) \\ x(0) & = x_0 \end{cases} \tag{1.1}$$

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,. . . [14, 34, 13, 18].

Let $t \mapsto X_{\theta^*}(t) = X^*(t)$ be the solution to the IVP (1.1) on $[0,T]$, for the true parameter set θ^*.

We want to estimate θ^* from noisy observations Y_i, $i = 1,\ldots,n$ 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) [30] or Bayesian approaches ([25, 15, 7, 17] and [16] 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 [37]. There is a computational complexity coming from repeated ODE integrations and from the computation of the gradient that are required by the optimization algorithm. Moreover, the usual criterion for NLS exhibits multiple local minima due to the strong non-linearity of $\theta \longmapsto X_\theta$ and to structural identifiability issues [1, 32]. It is then hard to start the optimization process or to assess the quality of the minima retained. These situations are often aggravated by "practical" identifiability issue that we detail below.

If $H(X_\theta) = \sum_{i=1}^{n} \|y_i - X_\theta(t_i)\|^2$ is the usual sum of squared residuals, the computation of the NLS estimator $\hat{\theta}^{NLS}$ is usually obtained by a Gauss-Newton algorithm or a variant such as the Levenberg-Marquardt algorithm. The sensitivity matrix $S(t_k,\theta) = \left(\frac{\partial X_\theta(t_k)}{\partial \theta_j}\right)_{j=1,\ldots,p}$ plays a critical role in finding the optimum, as its bad conditioning can make the local inverse problem ill-posed. The statistical importance of the sensitivity matrix is emphasized by the Fisher Information Matrix that depends on the matrix $S(t,\theta)S(t,\theta)^\top$. A low-rank sensitivity matrix gives rise to high correlations between parameters even with big sample size. Unfortunately, this situation arises frequently with models used in chemichal engineering and biology: the implicit influence of the states X_θ in the parameter θ, with sparse sampling, are conditions that favor the appearance of such degeneracy. The matrix $S(t,\widehat{\theta}^{NLS})S(t,\widehat{\theta}^{NLS})^\top$ is often nearly singular with dramatic differences in the order of magnitudes of its eigenvalues $\lambda_1,\ldots,\lambda_p$ that makes the parameter not identifiable in practice and gives rise to *sloppy parameters* [20, 41]. Hence, parameter estimation is often an ill-posed inverse problem [11] where regularization and approximation technics can help.

We are interested by specific estimators developed for the estimation of differential equation models based on nonparametric devices such as Gradient Matching estimators [5, 6, 31, 19] or Generalized Smoothing [37, 36, 10, 8].

Gradient Matching is a two stage procedure that uses a preliminary nonparametric curve estimator $\widehat{X}$ from the data $(t_i, Y_i)_{1\leq i\leq n}$. A minimum distance estimator [28] is obtained by minimizing the weighted L^2 distance $\int_0^T \|\dot{\widehat{X}}(t) - f(t,\widehat{X}(t),\theta)\|^2 w(t)dt$ with respect to θ. This simple estimator initiated by [44] (see variants [5, 19]) has a good computational efficiency while being consistent with a parametric rate of convergence. Nevertheless this method is not efficient, and the variance of Gradient Matching estimators are in general higher, in particular because of the use of nonparametric estimates of the derivative.

In the case of Generalized Smoothing [37], 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 relaxation on the ODE model constraint, needed for taking into account some uncertainty about the model. Based on Generalized Smoothing, Hooker [22] 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.

Generalized Smoothing and Gradient Matching shares the fact they deal with a function $\tilde{X}(\cdot, \tilde{\theta})$ that solves approximately the original ODE model (1.1) with estimate $\tilde{\theta}$, while being close to the data. In these methods, we do not know *a priori* the differential equation solved by the approximate solution $\tilde{X}(\cdot, \tilde{\theta})$ and the perturbed model is known only *a posteriori,* by introducing the forcing function.

A critical point is to understand and control the influence of the model approximation on the parameter estimates. For Generalized Smoothing, this theoretical and practical issue is adressed via the selection of an hyperparameter λ, whereas for Gradient Matching, it is adressed by using consistent estimators close to the solution (and its derivative).

We propose to invert the usual point of view and to define a Two-Step procedure that mimicks Generalized Smoothing. We do this by introducing a forced model and using optimal control theory, in order

- to avoid the use of a nonparametric estimate of the derivative $\dot{X}$,
- to control explicitely the model discrepancy of the "approximate solutions" $\tilde{X}(\cdot, \theta)$.

Our method provides a consistent parametric estimator when the model is correct, that shares similarities with Nonlinear Least Squares and Generalized Smoothing estimators. We show that it is root-n consistent and asymptotically normal. At the same time, we obtain a discrepancy measure between the model and the data having the form of a forcing function u, similar to the one introduced in [22].

An originality of that work is to use infinite dimensional optimization tools in a statistical framework, that is unusual even in the context of ODE estimation (except when dealing with Design of Experiments). Thanks to that, we do not use a finite dimensional approximation for $\tilde{X}(\cdot, \theta)$, and we avoid the approximation error usually encountered in that framework. Remarkably, the consistency of our method does not depend critically on the asymptotics of hyperparameters. We obtain a lower variance for our estimation procedure in practice, in particular thanks to a simple and easy to implement procedure for hyperparameter selection.

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 assumptions about the model. Then in section 4, we show that we reach the root$-n$ rate using regression splines for $\widehat{X}$. Finally, we compare our method with Nonlinear Least Squares and Generalized Smoothing on two realistic testbed models that have sloppy parameters, and we discuss the main differences between our approach and Generalized Smoothing.

We also consider the case of model misspecification in order to illustrate the ability of our approach to deal with this essential problem in practice. Finally, our experiments are completed by a real data analysis, obtained from the literature for ease of comparison and reproductibility.

2. Model and methodology

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

2.1. The statistical model and Generalized Smoothing

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

$$Y_i = X^*(t_i) + \epsilon_i$$

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

$$\dot{x}(t) = A_\theta(t)x(t) + r_\theta(t) \tag{2.1}$$

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 (2.1) 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 Generalized Smoothing (GS), parameter estimation is regularized by using an approximate solution of the ODE (2.1), as GS takes advantage of the double interpretation of splines for smoothing data, and for numerical solving of ODE by collocation. A basis expansion $\widehat{X}_\lambda(t, \theta) = \widehat{\beta}_\lambda(\theta)^T p(t)$ is computed for each θ, where $\widehat{\beta}_\lambda(\theta)$ is obtained by minimizing in β the criterion

$$J_n(\beta|\theta, \lambda) = \sum_{i=1}^{n} \left\|y_i - \beta^T p(t_i)\right\|_2^2 + \lambda \int_0^T \left\|\beta^T \dot{p}(t) - \left(A_\theta(t)\beta^T p(t) + r_\theta(t)\right)\right\|_2^2 dt \tag{2.2}$$

This first step (*inner optimization*) is profiling along the nuisance parameter β, whereas the estimation of the parameter of interest is obtained in the *middle optimization* by minimizing the sum of squared errors of the proxy $\hat{X}_\lambda(t, \theta)$

$$\hat{\theta}_\lambda^{GS} = \arg\min_\theta \sum_{i=1}^{n} \left\|y_i - \hat{X}_\lambda(t_i, \theta)\right\|^2 \tag{2.3}$$

The estimator depends on the hyperparameter λ, that needs to be selected during the *outer optimization*: the objective can be to minimize the sum of squared errors while controlling the discrepancy between the exact solution $X_{\hat{\theta}_{\lambda}^{GS}}$ and its approximation $\hat{X}_{\lambda}(\cdot,\hat{\theta}_{\lambda}^{GS})$ [37], to detect stability in the parameter estimates $\hat{\theta}_{\lambda}^{GS}$ [36], or to minimize the prediction error of $\hat{X}_{\lambda}(\cdot,\hat{\theta}_{\lambda}^{GS})$ [26].

The essential difference with NLS is the replacement of the exact solution X_{θ} by an approximation $\hat{X}_{\lambda}(\cdot,\theta)$ (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 (2.1). Indeed, collocation algorithms compute the coefficients of a B-spline expansion based on the relationships between $\widehat{X}$ and its derivative $\dot{\widehat{X}}$ evaluated on an appropriate grid of time points $0 = s_1 < s_2 < \cdots < s_p = T$, [3]. 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 (basis functions or size of the basis) 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 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 $H\left(\hat{X}_{\lambda}(\cdot,\theta)\right) = \sum_{i=1}^{n}\left\|y_i - \hat{X}_{\lambda}(t_i,\theta)\right\|_2^2$. For this reason, the ODE model (2.1) is not solved and it is useful to introduce the discrepancy term $\hat{u}_{\theta,\lambda}(t) = \dot{\hat{X}}_{\lambda}(t,\theta) - \left(A_{\theta}(t)\hat{X}_{\lambda}(t,\theta) + r_{\theta}(t)\right)$ that corresponds to a model error. In fact, the proxy $\hat{X}_{\lambda}(\cdot,\theta)$ satisfies the perturbed ODE $\dot{x} = A_{\theta}x + r_{\theta} + \hat{u}_{\theta,\lambda}$. This forcing function $\hat{u}_{\theta,\lambda}$ is an outcome of the optimization process and can be relatively hard to analyze, as it depends on the basis expansion used and on the data via the minimization of $J_n(\beta|\theta,\lambda)$. Nonetheless, the forcing function can be used for model selection: Hooker et al. have proposed goodness-of-fit tests based on this so-called “empirical forcing function” $\hat{u}_{\theta,\lambda}$, as $\hat{u}_{\theta,\lambda}$ is the residual at the derivative scale, but not at the state scale [24, 23].

We now detail our estimation method that relies on a classical relaxation and penalization scheme for regularizing the inverse problem. Based on the GS approach, we relax the ODE constraint (2.1) by introducing a perturbed version of the equation:

$$\dot{x}(t) = A_{\theta}(t)x(t) + r_{\theta}(t) + u(t) \tag{2.4}$$

where the function $t \mapsto u(t)$ can be any function in L^2. The function u is the residual of the regression of the derivative $\dot{X}$ on the exact model $A_{\theta}X + r_{\theta}$ and contains potentially several sources of error: model uncertainty, parameter uncertainty and random measurement errors. As in a classical regression, the objective is to minimize the norm of the residuals $\|u\|_{L^2}^2$ while being close to the data; the novelty of our approach is to deal with the residuals at the derivative level, instead of the state level as it is classically done.

Our analysis relies on solutions to the corresponding Initial Value Problem

$$\begin{cases} \dot{x}(t) = A_\theta(t)x(t) + r_\theta(t) + u(t) \\ x(0) = x_0^* \end{cases}$$

that exist as soon as A_θ is locally bounded on $[0, T]$ (see appendix C in [40]). We denote these functions $X_{\theta,u}$. Instead of using the spline proxy $\hat{X}_\lambda(\cdot, \theta)$ for approximating X^*, we use the trajectories $X_{\theta,u}$ of the ODE (2.4) 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 data misfit, and the model misfit represented by $u = \dot{X}_{\theta,u} - (A_\theta X_{\theta,u} + r_\theta)$. Instead of using the classical Sum of Squared Errors $H(X_{\theta,u})$, we use a smooth version based on a nonparametric proxy $\hat{X}$: $\int_0^T \left\| \hat{X}(t) - X_{\theta,u}(t) \right\|_2^2 dt$. Hence, we consider the subsequent cost function

$$C\left(\hat{X}; u, \theta, \lambda\right) = \int_0^T \left\| \hat{X}(t) - X_{\theta,u}(t) \right\|_2^2 dt + \lambda \int_0^T \|u(t)\|_2^2 \, dt \tag{2.5}$$

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

$$S\left(\hat{X}; \theta, \lambda\right) = \inf_{u \in L^2} C\left(\hat{X}; u, \theta, \lambda\right) \tag{2.6}$$

obtained by "profiling" on the function u. Finally, our estimator is defined by minimizing the same function S i.e

$$\widehat{\theta}_\lambda^T = \arg\min_{\theta \in \Theta} S\left(\hat{X}; \theta, \lambda\right) \tag{2.7}$$

The criterion $C\left(\hat{X}; u, \theta, \lambda\right)$ is (almost) the same as the criterion $J_n(\beta|\theta, \lambda)$, with the hyperparameter λ making the balance between data and model fidelity. The optimization step (2.6) is the same as the optimization $J_n(\beta|\theta, \lambda)$, and it gives rise to a similar state approximate solution $X_{\theta,\bar{u}_\lambda}$ that depends both on the data, the model and λ. Nevertheless, our estimator possesses two essential differences with Generalized Smoothing:

1. If we consider that $\sum_{i=1}^n \|y_i - X(t_i)\|^2 \simeq \int_0^T \|\hat{X}(t) - X(t)\|^2 dt$, then the first step of GS and Tracking solve the same problem. The difference between these two approaches comes from the way the optimization problem is effectively solved. Whereas the optimization problem has to be solved in $H^1 = \{X \in L^2 | \dot{X} \in L^2\}$ (see [4] for the definition and properties of Sobolev spaces), the Tracking approach uses the fact that any function X in H^1 is a solution of the perturbed ODE. The perturbation u is by

definition equal to $\dot{X} - (A_\theta X + r_\theta)$, so the optimization is effectively performed on H^1. In the case of Generalized Smoothing, the optimization is performed by using a B-splines expansion, which may induce some approximation error as the problem is constrained to a finite dimensional vector space instead of H^1. The use of a perturbed model enables to explore a bigger space during optimization. In particular, the GS solution $\hat{X}_\lambda(\cdot,\theta)$ can be re-written as $X_{\theta,\hat{u}_{\theta,\lambda}}$, which means that the profiling step (2.6) encompasses at the same time the GS and the NLS candidates. When the ODE model is well-specified, the Tracking approach does not suffer from the bias caused by model approximation whereas it is a known limitation of GS (see Olhede's comment of [37] about the influence and choice of the basis).

2. During the *middle optimization*, the Tracking estimator minimizes approximately the penalized least squares $H(X_{\theta,\bar{u}_{\theta,\lambda}}) + \lambda\int_0^T \|\bar{u}_{\theta,\lambda}(t)\|_2^2\,dt$, whereas GS minimizes the usual least squares criterion $H\left(X_{\theta,\hat{u}_{\theta,\lambda}}\right)$ without taking into account the model discrepancy $\|\hat{u}_{\theta,\lambda}\|_{L^2}^2$. Consequently, $X_{\hat{\theta}_\lambda^{GS},\hat{u}_{\theta,\lambda}}$ can be far from the true ODE solution $X_{\hat{\theta}_\lambda^{GS}}$, which changes the influence of λ on the parameter estimator. In particular, there is a risk of overfitting with GS (i.e. a "big" $\hat{u}_{\theta,\lambda}$ with a small $H\left(X_{\hat{\theta}_\lambda^{GS},\hat{u}_{\theta,\lambda}}\right)$) that can induce a high bias and variance. The presence of a "big" $\hat{u}_{\theta,\lambda}$ can be detected with a careful selection of λ (by comparing for instance $\hat{X}$ and ODE solutions). We show briefly in section 5.3.3 how the functions $\lambda \mapsto \hat{\theta}_\lambda^{GS}$, $\hat{\theta}_\lambda^{T}$ differ.

These two remarks put emphasis on the need to control simultaneously λ and K as n tends to infinity for generalized smoothing. In order to ensure the consistency and root-n rate of the GS estimator, Qi and Zaho [36] need to assume that $K = K(n)$ and $\lambda = \lambda(n)$, see theorems 3.2 and 3.3. In this work, the Tracking estimator is proven to be root-n consistent in section 3 and 4 for any λ, as soon as λ is positive. This means that there is no need to consider the case of a data-dependent hyperparameter $\lambda = \lambda_n$ for the asymptotic analysis (moreover, there no other hyperparameter to select in Tracking).

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 (2.7). 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 [29, 40, 33, 9]. In our case, we consider the control of a 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 [27, 9]. For completeness, we recall briefly in the appendix A the main results of LQ Theory.

Theorem 2.1 (Theorem and Definition of $S(\zeta;\theta,\lambda)$)**.** *Let $t \mapsto \zeta(t)$ be a function belonging to the Sobolev space $H^1([0,\,T],\mathbb{R}^d)$ and $X_{\theta,u}$ be the solution to the controlled ODE (2.4).*
For any θ,λ, there exists an unique optimal control $\overline{u}_{\theta,\lambda}$ that minimizes the cost function

$$C(\zeta;u,\theta,\lambda) = \int_0^T \left\{ \|\zeta(t) - X_{\theta,u}(t)\|_2^2 + \lambda \|u(t)\|_2^2 \right\} dt \tag{2.8}$$

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

$$\overline{u}_{\theta,\lambda}(t) = \frac{E(t)}{\lambda}\left(X_{\theta,\overline{u}_{\theta,\lambda}}(t) - \zeta(t)\right) + \frac{h(t)}{\lambda} \tag{2.9}$$

where E and h are solutions of the Final Value Problems

$$\left\{\begin{array}{l} \dot{E}(t) = I_d - A_\theta(t)^T E(t) - E(t)A_\theta(t) - \frac{E(t)^2}{\lambda} \\ \dot{h}(t) = -A_\theta(t)^T h(t) - E(t)\left(A_\theta(t)\zeta(t) + r_\theta(t) - \dot{\zeta}(t)\right) - \frac{E(t)h(t)}{\lambda} \end{array}\right. \tag{2.10}$$

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 (2.4).

Finally, the Profiled Cost S has the closed form

$$S(\zeta;\theta,\lambda) = -\int_0^T \left\{ 2\left(A_\theta(t)\zeta(t) + r_\theta(t) - \dot{\zeta}(t)\right)^\top h(t) + \frac{\|h(t)\|^2}{\lambda} \right\} dt \tag{2.11}$$

The cost (2.8) is usually used for solving the so-called "Tracking Problem" that consists in finding the optimal control u to apply to the ODE (2.4) in order to reach a target trajectory $t \mapsto \zeta(t)$, see [40] for an excellent introduction. The estimation problem is then to determine the parameter θ so that the corresponding ODE needs the smallest control u (in L^2 norm) in order to reach the noisy trajectory $t \mapsto \hat{X}(t)$.

Remark 2.1. *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 (2.10). 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 (2.10) that gives the functions E and h for all $t \in [0,T]$. Then, we just need to solve the ODE*

$$\left\{\begin{array}{l} \dot{x}(t) = A_\theta(t)x(t) + r_\theta(t) + \frac{E(t)}{\lambda}(x(t) - \zeta(t)) + \frac{h(t)}{\lambda} \\ x(0) = x_0^* \end{array}\right.$$

Remark 2.2. *From equation (2.11), we see that S depends smoothly in θ and λ, as in ζ. This was not easy to see from the infimum definition (2.6), 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 2.3. *The pertubed 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 perform simulations in order to get some insight. We will see in a simple example that our estimator gives a more accurate estimate than NLS.

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 LQ theory, we can 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 (2.7) is a consistent estimator of θ^* when the ODE model (2.1) is well-specified, and when we use a consistent nonparametric estimator $\widehat{X}$. 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 $S(\widehat{X};\theta,\lambda)$ to the asymptotic criterion $S(X^*;\theta,\lambda)$. 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 (2.1), 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 the case $\zeta = X^*$. More generally, we 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

$$\begin{cases} \dot{R_\theta}(t,s) = \alpha_\theta(T-t)R(t,s) \\ R_\theta(s,s) = I_d \end{cases} \tag{3.1}$$

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:*

$$\begin{aligned} \overline{X} &= \sup\nolimits_{\theta\in\Theta} \|X_\theta\|_{L^2} < +\infty \\ \bar{E} &= \sup\nolimits_{\theta\in\Theta} \|E_\theta\|_{L^2} < +\infty \end{aligned}$$

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. $\sup_\theta \|A_\theta\|_{L^2} = \bar{A} < +\infty$ exists as $(t,\theta) \mapsto A_\theta(t)$ is a continuous function on $[0,T]\times\Theta$ compactness. 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, we know that

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

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

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

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 $r_\theta^1(t) = A_\theta(t)X(t) + r_\theta(t) - \dot{X}(t)$.

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

We complete our analysis by showing that S is C^1 on 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 and $\Theta \times L^2\left([0,\, T], \mathbb{R}^d\right)$ respectively. 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:

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

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 $\Theta \times L^2\left([0,\, T], \mathbb{R}^d\right)$. This is enough to show the continuous differentiability of $S(X;\theta,\lambda)$ on Θ. □

The last regularities 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 a continuous nonlinear integral of $t \mapsto X(t)$.*

Proof. We show $S(X;\theta,\lambda)$ can be written using only X and not $\dot{X}$. First of all we 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:

$$\begin{aligned}\textstyle\int_0^T \dot{X}(t)^T h_\theta(t,X)dt = {} & F_{1,\theta}(X) + F_{2,\theta}(X) + F_{3,\theta}(X) \\ & - x_0^{*T} \textstyle\int_0^T R_\theta(T, T-s)E_\theta(s)r_\theta(s)ds \\ & - \tfrac{1}{2} x_0^{*T} E_\theta(0)x_0^*\end{aligned} \tag{3.2}$$

with

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

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

$$\begin{aligned} S(X;\theta,\lambda) = {} & -\textstyle\int_0^T \left(2\left(A_\theta(t)X(t) + r_\theta(t)\right)^T h_\theta(t,X) + \frac{1}{\lambda} h_\theta(t,X)^T h_\theta(t,X)\right) dt \\ & + F_{1,\theta}(X) + F_{2,\theta}(X) + F_{3,\theta}(X) \\ & - x_0^{*T} \textstyle\int_0^T R_\theta(T, T-s)E_\theta(s)r_\theta(s)ds \\ & - \tfrac{1}{2} x_0^{*T} E_\theta(0)x_0^* \end{aligned}$$

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 $S\left(X^*;\theta,\lambda\right)$. Condition 2 is then sufficient to show that $S\left(X^*;\theta,\lambda\right)$ characterizes well θ^*, as a global unique minimum. Hence, identifiability and convergence in supremum norm are sufficient to imply consistency (theorem 5.7 in [43]).

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

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

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

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 3.1. *Under conditions 1, 2, 3 and if $\widehat{X}$ is consistent in probability (in L^2-norm sense), and all $\lambda > 0$, we have*

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

Proof. Using proposition B.1, we have

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

with

$$\begin{array}{l}
K_1 = \sqrt{d}\,\|x_0^*\|_2\, \bar{R} + \sqrt{d}\bar{A}\bar{X} + \sqrt{d}\bar{\dot{E}}\overline{X} \\
K_2 = \sqrt{d}\left(\bar{A} + \frac{\bar{E}}{\lambda}\right) \\
K_3 = \sqrt{d}\bar{A} + \sqrt{d}\bar{\dot{E}} \\
K_4 = \sqrt{d}\left(\bar{A} + \frac{\bar{E}}{\lambda}\right)\bar{X}
\end{array}$$

and

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

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

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

We obtain

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

We can control $\left\|\widehat{X}\right\|_{L^2} \leq \left\|\widehat{X} - X^*\right\|_{L^2} + \|X^*\|_{L^2}$, which proves that if $\widehat{X}$ 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 [43]. □

Remark 3.1. *The initial condition x_0^* is assumed to be known. In practice, we can estimate it with the initial value of the non-parametric estimator $\widehat{X}(0)$. The criterion to use is then the same; because of the smooth dependence in the initial condition, we observe in practice that the tracking estimator $\widehat{\theta}_\lambda^T$ remains consistent.*

Remark 3.2. *For ensuring the consistency of $\widehat{\theta}_\lambda^T$, the smoothing parameter λ is only required to be nonnegative and we have no condition on its asymptotic behavior, whereas GS needs that the hyperparameter λ tends to infinity for removing the bias as it is usually done in smoothing. This low sensitivity in λ is a direct consequence of avoiding a finite basis decomposition approach for relaxing the constraint imposed by the ODE (2.1). Obvisously, the hyperparameter λ does influence the bias and variance of the $\widehat{\theta}_\lambda^T$ (see the next section on the asymptotics), but obtaining the precise influence of λ is beyond the scope of the paper.*

Remark 3.3. *In GS, the size of the basis expansion K^{GS} is critical for ensuring the identifiability of θ^*. Indeed, we need to be sure that the identity $\int_0^T \left|X^*(t) - \widehat{X}_\lambda(t,\theta)\right|^2 dt = 0$ implies that $\theta = \theta^*$ when $\widehat{X}_\lambda(\cdot,\theta)$ is a finite basis decomposition. This property is stronger than the structural identifiability of the model. The theoretical analysis of GS performed in [36] shows that it is needed to control a specific distance between X_θ and $\hat{X}_\lambda(\cdot,\theta)$, that can be done in practice by knots selection. The tracking approach avoids these difficulty, and rely only on the structural identifiability of the model, as $S(X^*;\theta,\lambda) = 0$ if and only if $\theta = \theta^*$, thanks to proposition 3.4.*

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

Our objective 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: There 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 $\hat{\theta}^T_\lambda$ reaches the parametric convergence rate, and that it is asymptotic normal. Our strategy consists in two stages:

Stage 1 (Prop 4.1) We show that $\hat{\theta}^T_\lambda - \theta^*$ behaves asymptotically as the difference $\Gamma(\widehat{X}) - \Gamma(X^*)$ where Γ is a continuous linear functional,

Stage 2 (Prop 4.2) 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 [35].

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

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

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 4.1. *Under conditions 1-5, we have:*

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

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

$$\Gamma(X) = \int_0^T \left(\frac{\partial\left(A_{\theta*}(t).X^*\right)}{\partial\theta} + \frac{1}{\lambda}\frac{\partial h_{\theta^*}(t,X^*)}{\partial\theta}\right)^T \left(\int_t^T R_{\theta^*}(T-t, T-s)X(s)ds\right)dt. \tag{4.1}$$

R_{θ^*} *is defined* by (3.1).

Proposition 4.2. *Under conditions 1-8 and by defining Γ as in proposition 4.1 we have that $\Gamma(\widehat{X}) - \Gamma(X^*)$ is asymptotically normal and $\Gamma(\widehat{X}) - \Gamma(X^*) = O_P(n^{-1/2})$.*

The root-n rate and asymptotic normality is obtained by combining the two previous propositions. Quite remarkably, we do not need to have $\lambda \longrightarrow \infty$ but only $\lambda > 0$ to claim

Theorem 4.1. *If $\widehat{X}$ is a regression spline and conditions C1-C8 are satisfied, then $\widehat{\theta}_\lambda^T - \theta^*$ is asymptotically normal and*

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

Remark 4.3. *The asymptotic linear representation given by proposition 4.1 gives a closed form expression for the asymptotic variance (D.1) given in appendix D. The expression obtained depends on λ, but it remains difficult to analyze it. Nevertheless, we can derive a plug-in estimate of the variance as it is needed for the computation of confidence intervals.*

5. Experiments

We evaluate the practical efficiency and illustrate our results by comparing the Tracking estimator $\hat{\theta}_\lambda^T$, with NLS $\hat{\theta}^{NLS}$ and Generalized Smoothing $\hat{\theta}^{GS}$ on two (relatively) small real models used in chemical engineering: the methanation reaction model and the isomerization of α-Pinene. We test several sample sizes and variance errors, and we are interested in comparing the respective bias and variance of the estimators, and also in evaluating the influence of the tracking estimator with respect to the nonparametric smoother $\hat{X}$ and the hyperparameter λ.

5.1. Experimental design

For a given sample size n and noise level σ, we estimate the Mean Square Error (bias and variance) 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, the observations are obtained by computing the ODE solution with a Runge-Kutta algorithm (ode45 in Matlab), then by adding a centered Gaussian noise (with variance σ^2) .

The nonparametric estimate needed for the Tracking estimator is a regression spline built with B-splines on a uniform knot sequence $\xi_k, k = 1, \dots, K$. For each run and each state variable, the number of knots is selected by minimizing GCV [39]. Tracking and Generalized Smoothing requires automated methods for selecting adaptively the hyperparameter λ: this is introduced and discussed in the next section.

5.2. Hyperparameter selection

For GS, we use the selection method for λ presented in [10, 36]: the value of λ is increased until the approximate solution $\hat{X}_\lambda(\cdot, \hat{\theta}^{GS}_\lambda)$ starts to differ significantly from the exact solution $X_{\widehat{\theta_\lambda}}$, i.e the difference starts to increase. This procedure permits to control the approximation error due to the use of B-splines, although the number of knots K^{GS} used remains high. In our experiments, K^{GS} is equal to the number of observations available, and it is fixed. The selected GS estimator is then denoted simply $\hat{\theta}^{GS}$.

In sections 3 and 4, we did not discuss the selection of λ for the tracking estimator, as it does not have the same theoretical importance as for smoothing. In practice, $\widehat{\theta_\lambda}$ is significantly affected by λ as it balances model and data fidelity. When $\lambda \rightarrow 0$, any u can be selected and we do overfitting and $X_{\theta,u}$ interpolates the nonparametric estimate $\widehat{X}$. At the opposite, when $\lambda \rightarrow \infty$, we have $\overline{u}_{\theta,\lambda} \longrightarrow 0$ and the criterion $S(\hat{X}; \theta, \lambda)$ is similar to the NLS.

We propose then a simple procedure for selecting λ based on the sum of squared errors. We know that the perturbed solution $X_{\hat{\theta}^T_\lambda, \bar{u}_{\theta,\lambda}}$ has a tendency to do overfitting w.r.t $X_{\hat{\theta}^T_\lambda}$ because of the presence of the control $\bar{u}_{\theta,\lambda}$ (when the model is well-specified). Hence, we use

$$\hat{\lambda} = \arg\min_{\lambda>0} H\left(X_{\hat{\theta}^T_\lambda}\right) \tag{5.1}$$

and the adaptive Tracking estimator used in practice is defined as $\hat{\theta}^T = \hat{\theta}^T_{\hat{\lambda}}$.

This *outer optimization* procedure bridges the gap with the NLS estimator that considers directly the minimization of $H(X_\theta)$. Alternatively, we can see equation (5.1) as selecting the hyperparameter that minimizes the prediction error of the estimated exact model. It is well-known that cross-validation, bootstrap,... give much better estimation of the prediction error, and that we could take advantage of these resampling methods. Nevertheless, the computational cost of the optimization of S can be high, and we prefer considering the direct use of H. Despite its simplicity and the above critics, we show in subsection 5.3.3 that our choice is sensible. Moreover, our experiments show that the procedure equation (5.1) gives competitive estimates, even if they might be suboptimal.

5.3. Optimization algorithms & gradient computation for S

As mentionned in the introduction or in the previous section, the optimization is computationally demanding, and some care has to be taken because of the nonlinearity of the criterion.

The NLS problem is solved with a Levenberg-Marquardt algorithm (function 'nlinfit' in Matlab), while GS is solved by the Gauss-Newton algorithm derived in [37] (Matlab code available on Giles Hooker's webpage[1]). The Tracking estimator is found with a trust-region algorithm (function ‘fminunc’ in Matlab),

[1] http://faculty.bscb.cornell.edu/~hooker/profile_webpages/

that uses the gradient $\nabla_\theta S(X;\theta,\lambda)$. For this, we need to compute both $\frac{\partial \widehat{h_\theta}}{\partial\theta}$ and $\frac{\partial E_\theta}{\partial\theta}$. We discuss briefly an efficient way to compute these derivatives that avoids the use of the sensitivity equations. Indeed, the sensitivity equations requires solving an ODE system of size $(d^2+d)\times p$, that grows quickly with the dimension of the original ODE.

We use instead an adjoint method for the computation of the gradients. This is a classical approach in data assimilation (see [2] for example) that reduces the size of the differential equation to solve. Indeed, we take advantage of the fact that we need only to compute the integrals or L^2 inner products of $\frac{\partial \widehat{h_\theta}}{\partial\theta}$ and $\frac{\partial E_\theta}{\partial\theta}$ (and not pointwise). If we introduce $Q_\theta = \left(\widehat{h_\theta}^T, (E_\theta^r)^T\right)^T$, the Riccati ODE is written row-wise as

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

where F is the row formulation of the Riccati ODE vector field. The gradient $\nabla_\theta S(X;\theta,\lambda)$ is such that

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

with

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

The function P is of dimension $\left(d^2+d\right)$, and is called the adjoint vector. It is solution of the ODE:

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

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. Here P is obtained by solving a d^2+d size ODE system, that is much smaller than the initial sensitivity equations, as it does not depend on the number of parameters.

5.3.1. *The model and comparison of estimators*

The "Methanation reaction" model is a linear autonomous ODE in $\mathbb{R}^4$, that is nonlinear w.r.t parameters. We have

$$A_\theta = \begin{pmatrix} -\frac{V+V'+F_0^{C0}/W}{\beta C^{C0}/W+C^{COl}} & 0 & 0 & 0 \\ \frac{V+V'}{\beta C^{H_20}/W} & -\frac{V+V'+v_5}{\beta C^{H_20}/W} & 0 & \frac{v_5}{\beta C^{H_20}/W} \\ \frac{V'}{\beta C^{CO_2}/W} & 0 & -\frac{V'+v_6}{\beta C^{CO_2}/W} & \frac{v_6}{\beta C^{CO_2}/W} \\ 0 & \frac{v_5}{C^{O_s}} & \frac{v_6}{C^{O_s}} & -\frac{v_5+v_6}{C^{O_s}} \end{pmatrix}$$

and

$$r_\theta = \left(\frac{F_i^{C0} z_i^{CO}}{\beta C^{C0}/W + C^{COl}}, 0, 0, 0\right)^\top.$$

Table 1
Known parameters in Methanation Reaction Model

F_i^{CO}/F_0^{CO}	z_i^{CO}	V/V'	C^j	W	β
1.31	0.132	12.4	0.02/0.01/0.01	0.744	206.1

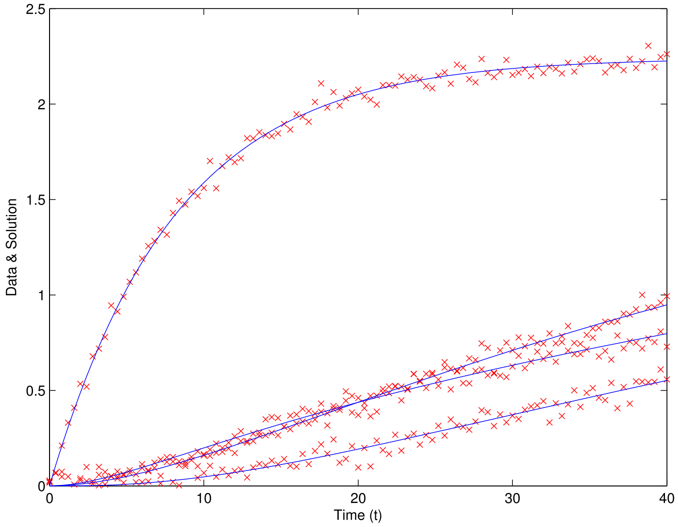


Fig 1. *Methanation Model: True solution X^* and simulated sample* $(n, \sigma) = (100, 0.004)$.

This model has been introduced in [21] for describing the dynamics of carbon monoxide and hydrogen methanation over a supported nickel catalyst by transient isotopic tracer in a gradientless circulating reactor. The state X is defined as $X^\top = \left(X^{CO}, X^{H_2O}, X^{CO_2}, X^{O_s}\right)$, and represents the quantity of the chemical species involved in the reaction. A constant inlet CO flow rate with constant and known fraction of isotope ^{18}O is introduced within the reactor; the fraction of ^{18}O present in oxygen atoms for each component is measured at different timeframe using a mass spectrometer. In the model, $X^j(t)$ represents the measured fraction of ^{18}O present in oxygen atoms of the chemical species j at time t. Some parameters are already known (see table table 1), and we have to estimate only the parameters $\theta = \left(C^{COl}, C^{Os}, v_5, v_6\right)$. For the simulations, the true parameter θ^* is the estimate provided in [21], i.e $\theta^* = (0.1,\ 11.1,\ 0.35,\ 0.008)$. The initial condition is known and equals to $x_0^* = (0,\ 0,\ 0,\ 0)$. The true solution X^* and simulated observations are plotted in figure 1.

For this parameter value and for $n = 100$, the matrix $S(\theta^*)^T S(\theta^*)$ has eigenvectors:

$$U(\theta^*) = \begin{pmatrix} 0.18 & 0.98 & 0.02 & 0 \\ -0.98 & 0.18 & -0.02 & -0.01 \\ -0.02 & -0.02 & 0.99 & 0.02 \\ 0.02 & 0 & -0.02 & 0.99 \end{pmatrix}$$

TABLE 2
Methanation Model: Comparison of the precision of Tracking, NLS, GS estimators

(n, σ)		$Bias(\widehat{\theta})$ ×10^{-2}	$Tr\left(V(\widehat{\theta})\right)$×$10^{-2}$	MSE×10^{-2}	ARE ×10^{-2}
$(100, 0.002)$	$\widehat{\theta}^T$	1.37	10.04	10.05	68.19
	$\widehat{\theta}^{NLS}$	1.03	11.46	11.47	63.26
	$\widehat{\theta}^{GS}$	6.76	38.92	39.35	77.98
$(50, 0.002)$	$\widehat{\theta}^T$	2.41	20.39	20.42	104.15
	$\widehat{\theta}^{NLS}$	2.82	20.59	20.66	83.74
	$\widehat{\theta}^{GS}$	12.85	37.80	39.17	95.08
$(100, 0.004)$	$\widehat{\theta}^T$	13.11	36.62	37.73	119.58
	$\widehat{\theta}^{NLS}$	8.10	44.91	45.32	119.33
	$\widehat{\theta}^{GS}$	7.40	197.32	97.63	144.84
$(50, 0.004)$	$\widehat{\theta}^T$	6.97	62.15	62.39	172.17
	$\widehat{\theta}^{NLS}$	9.15	67.74	68.19	184.41
	$\widehat{\theta}^{GS}$	4.74	222.30	222.41	208.84

with eigenvalues $\sigma(\theta^*) = \begin{pmatrix} 0.03 & 0.2 & 2.59 & 39.5 \end{pmatrix}$. Each eigenvector is essentially associated to a single parameter (U_1 for C^{Os}, U_2 for C^{COl}, U_3 for v_5, U_4 for v_6). The ratio $\frac{\lambda_4}{\lambda_1}$ is of order 10^3, which indicates a bad-conditionning of the estimation problem due to an important instability and sensibility with respect to noise. As a consequence, we have a high variance for statistical estimation (for NLS). Moreover, we can detect a correlation between C^{COl}, C^{Os} whereas (v_5, v_6) can be estimated well (almost independently). We consider that we are in presence of sloppy parameters when the eigenvalue ratio is about 10^4; this means that the estimation problem is moderately "ill-posed".

We have conducted 4 Monte Carlo experiments two sample sizes $n = 100$ and $n = 50$ (observations are uniformely sampled the time interval $[0, 40]$), with 2 noise levels $\sigma = 0.002$ and 0.004, that are presented in table 2.

For the lowest level of noise ($\sigma = 0.002$), $\widehat{\theta}^T$ and $\widehat{\theta}^{NLS}$ show similar results in terms of bias, variance and MSE, but NLS has a smaller ARE. For $\sigma = 0.004$, $\widehat{\theta}^T$ has a smaller variance than $\widehat{\theta}^{NLS}$, but the bias is not systematically higher. Nevertheless, the Tracking estimator gives the best estimates in terms of MSE and ARE. In every cases, the GS estimator has a higher variance, MSE and ARE.

If we look at the bias of the 3 estimators, it is hard to identify a trend, as it depends on the nonlinearity of the problem, and of the approximation used. The bias can be either bigger or smaller, depending of the circumstances, but in particular, we see that the NLS remains biased (even when $n = 100$).

The lowest variance is always reached by Tracking, showing that we can improve on NLS by an appropriate regularization. Hence, despite the similar construction of GS and Tracking, the Tracking estimator seems to behave more like the NLS. We investigate in the next subsection the differences between $\widehat{\theta}^T = \hat{\theta}^T_{\hat{\lambda}}$ and $\widehat{\theta}^{GS} = \widehat{\theta}^{GS}_{\hat{\lambda}}$, that explain such different performances.

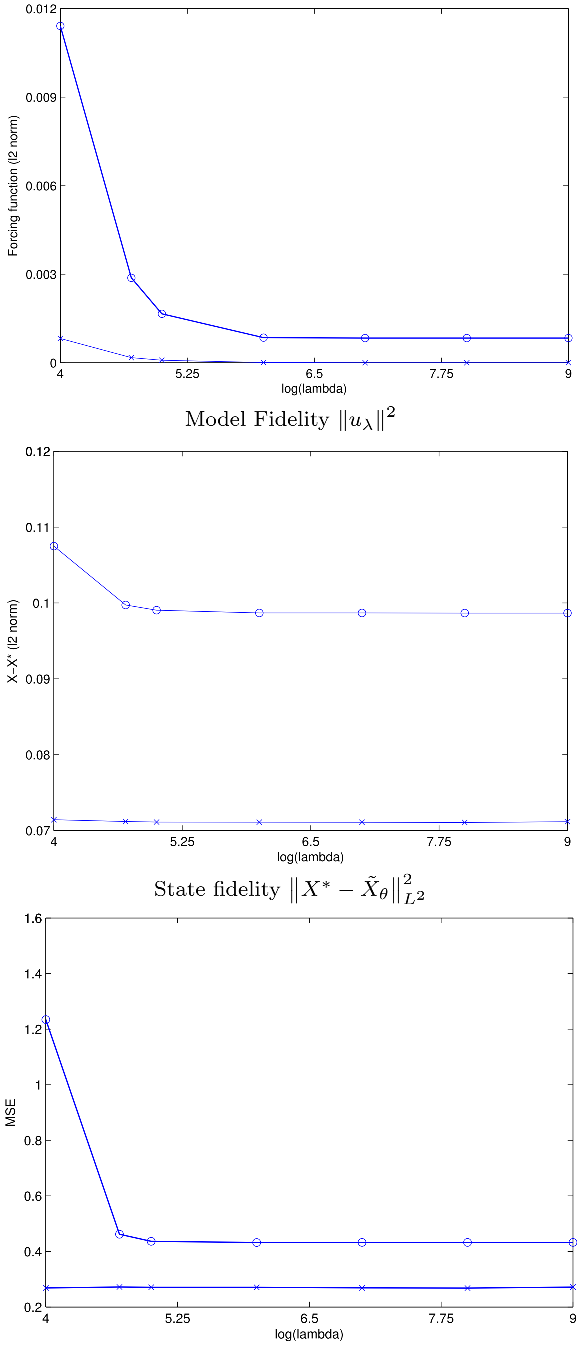
5.3.2. *Influence of the hyperparameter λ: comparison of GS and Tracking*

We compare the behavior of $\lambda \mapsto \hat{\theta}_{\lambda}^{T}$ and $\lambda \mapsto \hat{\theta}_{\lambda}^{GS}$ in the case of the methanation model, when the sample size is $n = 20$ and the noise level is $\sigma = 0.002$. We are mainly interested in the case of low sample size, because the different estimators are all consistent and asymptotically normal, and their characteristic features become less distinguishable when n is high. Moreover, the adaptive selection of λ is particularly critical for small n, and it is informative to understand the rationale of the selection method proposed in 5.3.3. Hence, we are interested in how λ maintains the trade-off for Generalized Smoothing and Tracking:

1. Model fidelity (norm of the forcing function): $\lambda \longmapsto \left\|\hat{u}_{\hat{\theta}_{\lambda}^{GS},\lambda}\right\|_{L^2}^2$ and $\lambda \longmapsto \left\|\bar{u}_{\hat{\theta}_{\lambda}^{T},\lambda}\right\|_{L^2}^2$
2. State fidelity (Observation fidelity): $\lambda \longmapsto \left\|X^* - \hat{X}_{\lambda}\left(\cdot, \hat{\theta}_{\lambda}^{GS}\right)\right\|_{L^2}^2$ and $\lambda \longmapsto \left\|X^* - X_{\hat{\theta}_{\lambda}^{T},\bar{u}_{\lambda}}\right\|_{L^2}^2$
3. Parameter fidelity (Estimator accuracy): $\lambda \mapsto \left\|\theta^* - \hat{\theta}_{\lambda}^{GS}\right\|^2$and $\lambda \mapsto \left\|\theta^* - \hat{\theta}_{\lambda}^{T}\right\|^2$

Based on $N_{MC} = 100$ runs, we plot the mean curves evaluated at λ in $\{10^k, 5 \times 10^k\}_{k \in [\![4,9]\!]}$ in figure 2. We can see that the Tracking approach has a higher fidelity to the model, state and parameter for any value of λ. Concerning model fidelity, there is a fast decrease in the model misfit (in particular for GS) and it stays constant for $\log \lambda \geq 6$. Whereas the model error vanishes for Tracking, the model error stays positive for Generalized Smoothing. This remaining error is a consequence of the splines approximation which is not good enough. During the *middle optimization*, Tracking minimizes $\left\|\hat{X} - X_{\hat{\theta}_{\lambda}^{T},\bar{u}_{\lambda}}\right\|_{L^2}^2 + \lambda \left\|\bar{u}_{\hat{\theta}_{\lambda}^{T},\lambda}\right\|_{L^2}^2$, and tends to find a parameter which induces a low model misfit (that vanishes if λ is too big). The state fidelity for Tracking is flat and very small: it starts with a reasonable nonparametric estimate, and the graph indicates that Tracking can always find a function $X = X_{\theta,u}$ close to the true function X^*, and such that u is not big. This comes from the fact the optimization is done on the space H^1; at the contrary, despite good approximation properties of B-splines, it always remains an error $\hat{X}_{\lambda}\left(\cdot, \hat{\theta}_{\lambda}^{GS}\right)$ that gives rise to a plateau. This graph justifies the need of increasing the B-splines basis with λ as assumed by Qi and Zhao [36]. Finally, the state error due to the use of approximate solutions generates a bias for parameter estimation: this bias remains constant for Tracking and lower than Generalized Smoothing (figure 2 (c)). For GS, the bias drops quite fast thanks to the drop in model misfit and then it remains constant.

The classical fitting procedure with GS consists in a sequential identification of the parameters called the parameter cascade: made of an *inner optimization* (β), a *middle optimization* (θ) and an *outer optimization* (λ). In GS, the outer optimization is done for selecting the parameter that predicts correctly the observations, but also for controlling the model error. In the case of Tracking, the control of the model error is done during the *middle optimization* because it minimizes the penalized least-squares. Then the objective of the *outer optimization* in Tracking is only to fit the exact model.



Model Fidelity $\|u_\lambda\|^2$

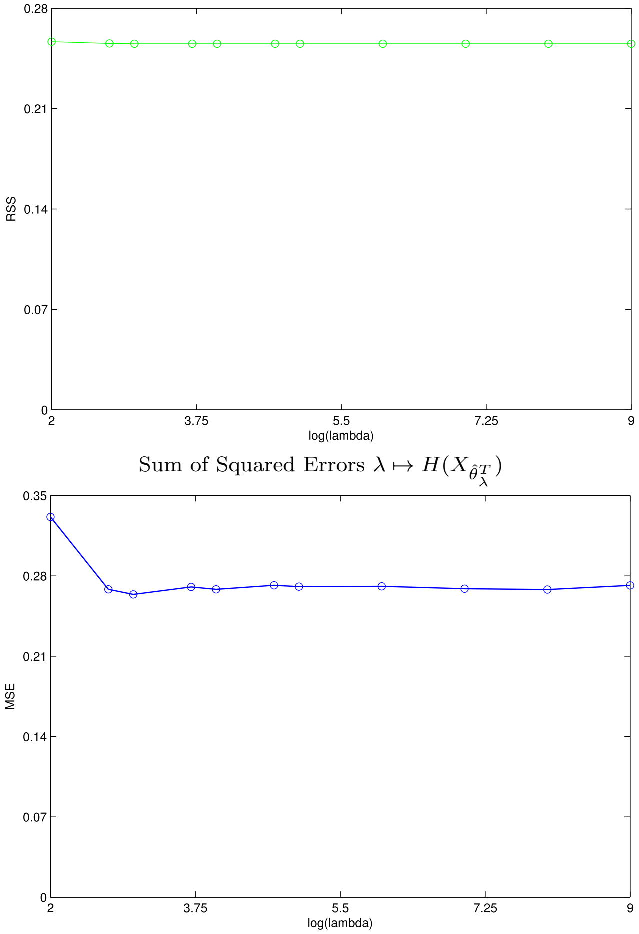
State fidelity $\left\|X^* - \tilde{X}_\theta\right\|_{L^2}^2$

Parameter fidelity $\left\|\theta^* - \hat{\theta}_\lambda\right\|^2$

FIG 2. *Methanation Reaction Model: Influence of* λ *for Model fidelity, State fidelity and Parameter fidelity. (*$\circ$*): Generalized Smoothing; (*$\ast$*) is Tracking.*

5.3.3. *Adaptive selection for hyperparameter and influence of the nonparametric estimate for Tracking*

We complete the analysis made in the previous subsection by justifying empirically the selection rule (*outer optimization*) introduced in section 5.2. Figure 3 shows that minimizing $H\left(X_{\widehat{\theta_\lambda}^T}\right)$ enables to locate the minimum of the MSE function $\lambda \mapsto \left\|\theta^* - \hat{\theta}_\lambda^T\right\|^2$. The two functions present parallel evolutions, that is a decrease followed by a stabilization around a constant value and presenting some minor ripples. Our selection criteria for λ is a relevant way for selecting an estimator corresponding to a low MSE value.



Sum of Squared Errors $\lambda \mapsto H(X_{\hat{\theta}_\lambda^T})$

Parameter fidelity $\lambda \mapsto \|\theta^* - \hat{\theta}_\lambda^T\|^2$

Fig 3. *Methanation model: Joint influence of λ on Sum of Squared Errors H and Parameter Estimation (Squared Distance) for the Tracking estimator $\hat{\theta}_\lambda^T$.*

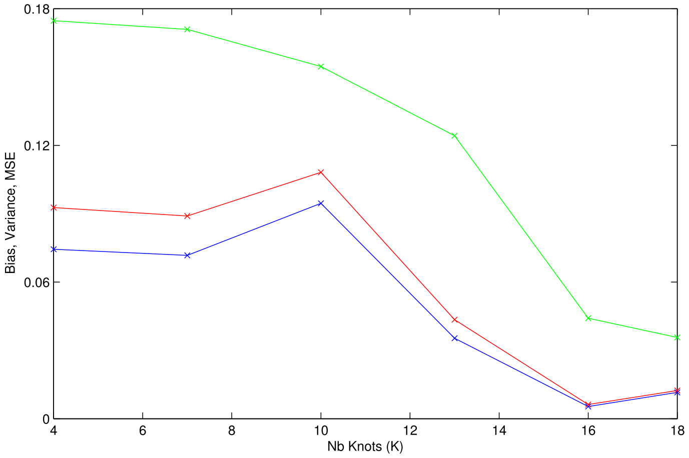


FIG 4. *Influence of nonparametric proxy $\hat{X}$ (number of knots K): $K \longmapsto \|Bias(\widehat{\theta}^T)\|_2$ (green), $K \longmapsto \|Var(\widehat{\theta}^T)\|_2$ (blue), $K \longmapsto MSE(\widehat{\theta}^T)$ (red).*

The tracking estimator might be also influenced - indirectly - by the hyperparameter selection involved in the nonparametric proxy $\hat{X}$. In particular, the selection of the number of knots or their location can be influential. It is well-known that Gradient Matching estimators and other two-step estimators depends strongly of the choice of these hyperparameters. Similarly, Chervoneva et al. [26] showed that the perfomances of Generalized Smoothing can be improved by selecting adaptively the knots placement and numbers, i.e. by minimizing the approximation error. Tracking is also influenced by the quality of approximation of $\hat{X}$ but this relation is easy to handle.

We have evaluated by Monte Carlo the influence of the number of knots when $\hat{X}$ is a regression spline ($N_{MC} = 100$ runs with the sample size $n = 20$ and noise level $\sigma = 0.002$). For each run, we have estimated θ^* for several choice of knots number K for $\widehat{X}$ (from 4 to 18, uniformely sampled on $[0, 40]$). The figure 4 represents the estimated MSE obtained after the N_{MC} runs for different number of knots.

One can observe a dramatic decrease of the MSE as long as K increases (except for $K = 10$, indicating that the knot location is important), that can nearly vanish when K is big enough. When K is big, we do undersmoothing that enables to catch the main patterns of the data, i.e of the state. A major drawback of undersmoothing is the presence of spurious oscillations and high variance, in particular for the estimation of the derivative. As stated in proposition 3.3, the Tracking approach does not depend on an estimator of the derivative, as it should be the case for Gradient-Matching. Moreover, in the *middle optimization*, we penalize H with $\|\bar{u}_{\theta,\lambda}\|^2_{L^2}$ i.e with the norm of the derivative of the approximate solution $X_{\theta,u}$: as a consequence we remove the additional oscillations of the data.

The results in 4 shows that $\hat{X}$ can influence the parameter estimates, but the estimation does not need to be efficient and is less sensitive to the bias variance trade-off as Gradient Matching (the best estimate is obtained for $K = 18$, whereas $n = 20$). Hence we can select the nonparametric smoother $\hat{X}$ by minimizing GCV (or other approximate optimal methods that might do undersmoothing) and obtaining good estimates.

5.4. $\alpha-$Pinene model

This linear model is introduced in [38] as a benchmark for nonlinear optimization, and it is used for modeling the isomerization of $\alpha-$Pinene. The model is autonomous and homogeneous ($r_\theta = 0$) with

$$A_\theta = \begin{pmatrix} -(\theta_1+\theta_2) & 0 & 0 & 0 & 0 \\ \theta_1 & 0 & 0 & 0 & 0 \\ \theta_2 & 0 & -(\theta_3+\theta_4) & 0 & \theta_5 \\ 0 & 0 & \theta_3 & 0 & 0 \\ 0 & 0 & \theta_4 & 0 & -\theta_5 \end{pmatrix} \tag{5.2}$$

The initial condition is known and equal to $x_0^* = (100, 0, 0, 0, 0)$ and the true parameter value is $\theta^* = (5.93,\, 2.96,\, 2.05,\, 27.5,\, 4) \times 10^{-4}$. The estimation of θ^* is still considered as cumbersome and many estimation methods fail to converge or converge to bad local solutions because of difficulty to accurately estimate θ_4 and θ_5. This can be explained by the high correlation between the parameters [38]. This issue can be analyzed with the sensitivity matrix and the Fisher Information matrix $S(\theta^*)^T S(\theta^*)$. When $n = 100$, the 5 eigenvectors are given by

$$U^* = \begin{pmatrix} 0.00 & 0.01 & 0.13 & 0.87 & -0.46 \\ 0.00 & 0.03 & 0.18 & 0.43 & 0.88 \\ 0.01 & -0.4 & 0.89 & -0.18 & -0.08 \\ 0.98 & -0.14 & -0.07 & 0.02 & 0.01 \\ 0.16 & 0.90 & 0.38 & -0.09 & -0.06 \end{pmatrix}$$

and the corresponding spectrum is $\sigma(\theta^*) \simeq 10^7 \times \big(0.003\ 0.033\ 0.418\ 0.487\ 4.575\big)$. The ratio $\frac{\lambda_5}{\lambda_1} \simeq 1.5 \times 10^4$ indicates that we have again sloppy parameters, that makes the parameter estimation an ill-posed inverse problem. In particular, the parameters $(\theta_4,\, \theta_5)$ are associated to the directions with the lowest eigenvalues.

We perform Monte Carlo simulations when the observed time range is $[0,\, 100]$. Because of different orders of magnitude for the state variables, we rescale the standard deviation of the measurement error componentwise. Here for a given reference σ value for the noise, the standard deviation of the noise for variable X_i is equal to $\frac{\sigma}{100} \times \frac{1}{T}\int_0^T X_i(t)dt$ (see figure 5). For computing the Tracking estimator, we select λ among $\left\{10^k,\, 5\times 10^k\right\}_{1\le k\le 3}$.

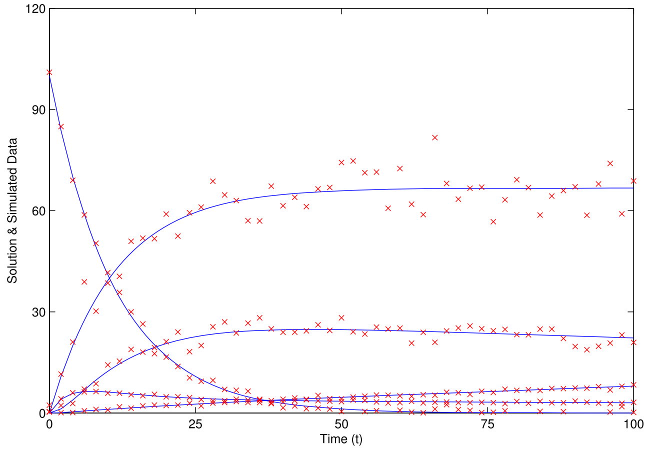


FIG 5. *Well-specified α-pinene Model: True solution X^* and noisy observations y_i with $n = 50, \sigma = 8$.*

5.4.1. *Well-specified model*

Despite the difficulty of the estimation, the estimators $\widehat{\theta}^{NLS}$, $\widehat{\theta}^{GS}$ and $\widehat{\theta}^{T}$ still provide consistent estimates, see the bias, variance and ARE in table 3. As one can expect, the approximate approaches ($\widehat{\theta}^{GS}$ and $\widehat{\theta}^{T}$) give more biased estimation than $\widehat{\theta}^{NLS}$, but $\widehat{\theta}^{GS}$ and $\widehat{\theta}^{T}$ have smaller variance and Mean Square Error than $\widehat{\theta}^{NLS}$.

This "efficiency" of Generalized Smoothing and Tracking can be explained by the fact that the biased estimators can bypass the theoretical limitations defined by the Fisher Information Matrix. Hence, the negative effect of sloppy

TABLE 3
Well-specified α−pinene: Parameter estimation accuracy for Tracking, Generalized Smoothing and Nonlinear Least Squares

(n,σ)		$Bias(\widehat{\theta})\ \times 10^{-2}$	$Tr\left(V(\widehat{\theta})\right) \times 10^{-4}$	MSE$\times 10^{-4}$	ARE $\times 10^{-2}$
$(100, 4)$	$\widehat{\theta}^{T}$	0.05	0.3	0.31	4.80
	$\widehat{\theta}^{NLS}$	0.03	0.85	0.86	7.82
	$\widehat{\theta}^{GS}$	0.05	0.38	0.38	8.52
$(100, 8)$	$\widehat{\theta}^{T}$	0.46	1.15	1.28	11
	$\widehat{\theta}^{NLS}$	0.08	2.12	2.13	13
	$\widehat{\theta}^{GS}$	0.21	1.93	1.95	21
$(50, 4)$	$\widehat{\theta}^{T}$	0.45	0.59	0.74	7.86
	$\widehat{\theta}^{NLS}$	0.05	1.44	1.44	10.25
	$\widehat{\theta}^{GS}$	0.26	1.01	1.03	17.73
$(50, 8)$	$\widehat{\theta}^{T}$	0.28	3.02	3.05	17
	$\widehat{\theta}^{NLS}$	0.04	6.96	6.96	23
	$\widehat{\theta}^{GS}$	0.41	3.54	3.61	36

parameters can be reduced thanks to the biased approach. Comparing $\widehat{\theta}^{GS}$ and $\widehat{\theta}^{T}$ only, we remark that Tracking gives the smallest MSE and ARE. From the experiments it is hard to say what approach is less biased, but the Tracking has a systematically a lower variance than GS. Moreover, the ARE indicates that for Tracking the relative accuracy is enhanced for each parameter (in particular for θ_4 that is bigger than the other parameters).

5.4.2. Misspecified model

Model misspecification is a classical but important limitation during parameter estimation. Ordinary Differential Equations are mechanistic models, and the bias induced by some misspecification can be particularly misleading in the role of each variable. Although it is hard to anticipate the effect of model error during the estimation process, several paper starts to grasp this difficulty, by showing that integrating a possible error can ameliorate significantly the quality of estimation, see [12, 42].

The Generalized Smoothing and the Tracking procedures are prone to cope with model misspecification, because they consider approximate solutions to exact model, or conversely, exact solutions to approximate model. In order to support this idea, we consider that the true α-pinene model is pertubed with a forcing function v, i.e X^* is such that

$$\dot{X}^*(t) = A_{\theta^*} X^*(t) + v(t) \tag{5.3}$$

with $v(t) = 0.1 \sin(\frac{\pi}{50}t) \times (1\,1\,1\,1\,1)^\top$. Based on the same framework as section 5.4.1, we have used NLS, GS and Tracking for estimating θ^*, based on the assumption that $\dot{X}^*(t) = A_{\theta^*} X^*(t)$ is the right model.

The quality of the estimators estimated by Monte Carlo are gathered on table 4 and support the claims of [12]: the NLS estimator gives the worst estimator,

Table 4

Misspecified α−pinene: Parameter estimation accuracy for Tracking, Generalized Smoothing and Nonlinear Least Squares

(n,σ)		$Bias(\widehat{\theta})\ \times 10^{-2}$	$Tr\left(V(\widehat{\theta})\right)\times 10^{-4}$	MSE$\times 10^{-4}$	ARE $\times 10^{-2}$
(100, 4)	$\widehat{\theta}^{T}$	0.73	0.38	0.63	0.24
	$\widehat{\theta}^{NLS}$	4.40	1.03	7.16	1.06
	$\widehat{\theta}^{GS}$	1.53	3.44	4.01	0.62
(100, 8)	$\widehat{\theta}^{T}$	0.81	1.02	1.28	0.27
	$\widehat{\theta}^{NLS}$	4.45	3.13	9.40	1.08
	$\widehat{\theta}^{GS}$	1.48	7.86	8.42	0.66
(50, 4)	$\widehat{\theta}^{T}$	0.75	1.32	1.55	0.26
	$\widehat{\theta}^{NLS}$	4.83	3.22	11.00	1.08
	$\widehat{\theta}^{GS}$	1.44	3.09	3.83	0.68
(50, 8)	$\widehat{\theta}^{T}$	0.75	2.97	3.22	0.31
	$\widehat{\theta}^{NLS}$	4.74	10.00	18.00	1.13
	$\widehat{\theta}^{GS}$	1.11	4.83	5.17	0.50

with the biggest bias, MSE and ARE. The estimator $\hat{\theta}^T$ and $\hat{\theta}^{GS}$ have always a smaller and nearly constant bias accross the different experiments. The bias remains smaller for Tracking, as $\tilde{X}_\theta$ is not limited to be a spline; the variance is also smaller, because we avoid overfitting by minimizing the penalized least squares instead of the function H only for GS. This analysis shows that the use of approximate models can robustify the statistical estimation while being consistent in the case of well-specified models. There are then the method of choice for dealing with real data.

5.4.3. *Real data analysis*

We finish our analysis of α-pinene by fitting the model on a real data set retrieved from [14], see table (5). The state variables are all observed at eight times steps, and the initial conditions are known and equal to x_0^*.

As a benchmark, we use the estimate provided by [38] which corresponds to the parameter value θ^* used in the previous simulations. This solution X_{θ^*} fits well the data and we are interested in computing the estimate $\hat{\theta}^T$, and analyzing the forcing functions $\overline{u}^T$. For computational efficiency, we have reparametrized the time in order to divide the observation time by 1000: this does not change anything as the system is autonomous and it avoids the difficulties generated by long-term integration.

Because the data are sparse, the nonparametric estimator $\widehat{X}$ is selected "by hand" in order to catch the mean feature of the data. We take a spline with nodes only at the boundaries and we use the constraint $\hat{X}(0) = x_0^*$ during estimation. Finally, the hyperparameter λ is selected by the rule (5.1) among the candidates $\lambda = 10^k$, $k = 1, 2 \cdots, 11$ and $\lambda = 5 \times 10^k$, $k = 1, 2, 3, 4$. We obtain $\hat{\lambda} = 100$ but the tracking estimator $\hat{\theta}_\lambda^T$ reaches a plateau for $\lambda \geq 5000$ (and the estimates are quite stable along the regularization path). We see $\widehat{\theta_1}^T$, $\widehat{\theta_2}^T$, $\widehat{\theta_3}^T$ are close to the estimation of [38], but the sloppy parameters $\left(\hat{\theta}_4^T, \hat{\theta}_5^T\right)$ differs from the estimates (θ_4^*, θ_5^*); anyway, the tracking estimates provides a good fit to the data (see the sum of squared errors in table 6 and figure 6).

We can also compare the model discrepancy for the two different estimates by comparing the corresponding forcing functions. We compute a forcing function

TABLE 5
$\alpha-$pinene: experimental data set from Fuguitt & Hawkins

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

TABLE 6
α-pinene: Estimates in the real cata case

10^{-4}	θ_1	θ_2	θ_3	θ_4	θ_5	$H\left(X_\theta\right)$
θ^*	0.593	0.296	0.205	2.75	0.4	19.89
$\hat{\theta}^T_{\hat{\lambda}}$	0.589	0.290	0.193	2.301	0.234	23.88
$\hat{\theta}^T_{5000}$	0.583	0.295	0.207	2.259	0.238	25.29

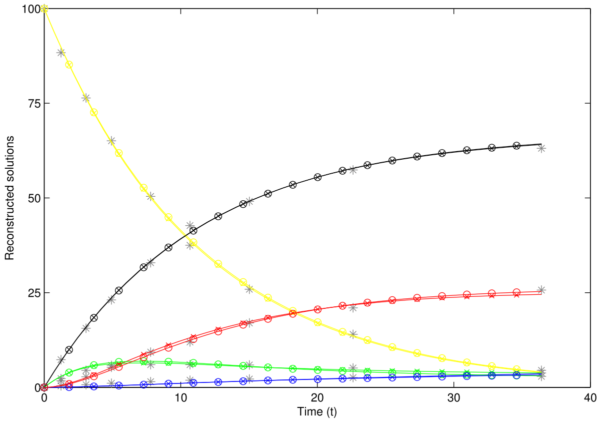


FIG 6. *α-pinene: real data and estimated curves with Nonlinear Least Squares and Tracking. ($\circ$): NLS solution; ($\times$): Tracking (exact) solution.*

$\overline{u}^*$ associated to the NLS estimate θ^* by the closed looped formula (2.9) with $\theta = \theta^*$ and $\lambda = \hat{\lambda} = 100$. The controls are represented in figure 7, where the forcing function $\overline{u}^*$ is ploted with $\times$, and $\overline{u}^T$ is ploted with $\circ$. The controls are vector functions in $\mathbb{R}^5$, and each entry $\overline{u}_i$ correspond to one state variable X_i. The function ploted 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.

We obtain that $\left\|\overline{u}^T\right\|^2_{L^2} \leq \left\|\overline{u}^*\right\|^2_{L^2}$, but there is no significant differences between the two forcing functions for the component 1 to 4. The only important difference is for X_5 (red curve on on the figure 7), which is the state variable exclusively linked to parameters θ_4 and θ_5 (the most difficult parameters to estimate according to [38] because of high correlation). According to our analysis, this indicates that the NLS estimate needs a bigger model correction with a forcing function, in order to compensate a bigger estimation error for θ_4 and θ_5. This suggests that our model estimate might be more reliable than the NLS.

6. Discussion

We have introduced an estimator that have a smaller variance than NLS and Generalized Smoothing, and better MSE and ARE in almost all the cases considered. The Tracking estimator does improve on NLS (and GS), when the model is particularly ill-posed. This is because we can bypass the Fisher Information

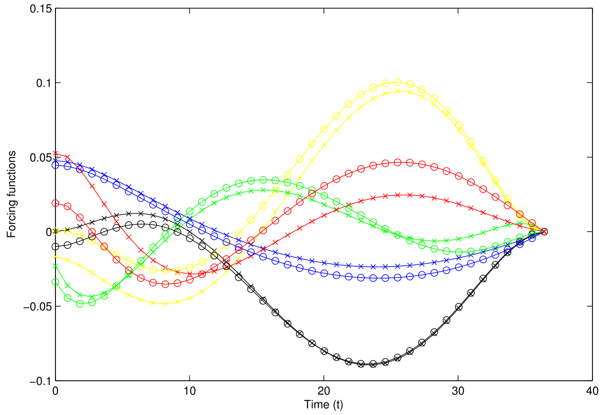


FIG 7. *α-pinene: Forcing functions for Nonlinear Least Squares and Tracking in the real data case. ($\circ$): NLS Forcing function $\bar{u}^*$; ($\times$): Tracking forcing function $\bar{u}_{\hat{\theta}}$.*

matrix limit; moreover, we reduce the overfitting that one can have with Generalized Smoothing in the *middle step*, as we do a better control of the model discrepancy (with the optimization of S instead of H). Moreover the selection of λ makes the Tracking estimator close to NLS, and avoid the use of complex selection procedure that increases the variance.

We insist on the fact that estimation problems for ODE appear even for small models (quite low dimensional, and with a small number of parameters), and we think that the Tracking approach can be helpful for any model (as some simulations on simple systems without sloppy parameter has given similar performances to NLS) - but at a higher computational cost. Nevertheless, the practical performances of NLS here can be misleading as all the estimates has been computed by starting the optimization algorithms with the true parameter θ^*. This information is very valuable for NLS and we clearly overestimate the practical performances of NLS whereas Generalized Smoothing still provides good estimate with poor starting values (or NLS need to rely on elaborated optimization procedures). The Tracking estimator can also cope with relatively bad initial guesses.

An advance provided by the Tracking and Generalized Smoothing is the ability to deal properly with model misspecification and to provide guidance and tools for analyzing model misfit with the forcing functions. Our brief analysis in the real data case indicates that we can assess the influence of parameters, and discuss the relevancy of the model. In order to pursue that task, we need to derive more elaborate results on the behavior u in the case of model misspecification, and developping testing tools and methodology.

Our theoretical results suffer from practical limitations, as we consider fully observed models, which are linear, with known initial conditions. Nevertheless, the same analysis with other optimal control tools remains possible with adapted criterion to optimize and characterization of optimal solutions (e.g. Pontryagin Maximum Principle).

Parameter estimation and optimal control: Appendix

Appendix A: Fundamental Results of Optimal Control: Linear-Quadratic Theory

The previous theorem is a particular case of a more general theorem which ensure existence and uniqueness of optimal control for cost under the form:

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

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:

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

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:

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

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

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

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

Appendix B: Proof & Intermediary results

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

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

$$\begin{array}{l} \dot{E}(t) = W(t) - A(t)^t E(t) - E(t) A(t) - \frac{1}{\lambda} E(t)^2 \\ E(T) = -Q \end{array} \tag{B.1}$$

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" [40] 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 we have:

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

Let us reason by contradiction, at the contrary we assumed that $\exists t_e \in [0\,,\,T]$ s.t $\lim_{t \to t_e +} \left\| E(t) \right\|_2 = +\infty$. It implies:

$$\forall \alpha > 0 \, \exists t_0 \in \left]t_e\,,\,T\right]\,,\, x_0 \in \mathbb{R}^d \, with \, \left\| x_0 \right\|_2 = 1 \, s.t \, \left| x_0^T E(t_0) x_0 \right| \geq \alpha \tag{B.2}$$

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 (B.2) 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:

$$\begin{aligned} N_\theta(t).X \; :=& \int_t^T R_\theta(T-t, T-s) X(s) ds + E_\theta(t) X(t) \\ &+ \int_t^T R_\theta(T-t, T-s) E_\theta(s) r_\theta(s) ds \end{aligned}$$

with R_θ defined by (3.1).

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:

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

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:

$$\begin{aligned} h_\theta(t,X) = & \int_t^T R_\theta(T-t,T-s)X(s)ds + E_\theta(t)X(t) \\ & + \int_t^T R_\theta(T-t,T-s)E_\theta(s)r_\theta(s)ds \end{aligned} \tag{B.3}$$

□

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*

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

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

Proof. Integration by part give us:

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

and

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

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

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

Finally, we obtain:

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

□

B.2. Consistency Proof

In the following proposition B.1 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.2 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.1. *Under conditions 1 and 3, $\forall \theta \in \Theta$ we have:*

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

$$With : \begin{cases} K_1 = \sqrt{d}\left\|x_0^*\right\|_2 \bar{R} + d\bar{E}\bar{A}\bar{X} + \sqrt{d}\bar{\dot{E}}\overline{X} \\ K_2 = \sqrt{d}\left(\bar{A} + \frac{\bar{E}}{\lambda}\right) \\ K_3 = d\bar{E}\bar{A} + \sqrt{d}\bar{\dot{E}} \\ K_4 = \sqrt{d}\left(\bar{A} + \frac{\bar{E}}{\lambda}\right)\bar{X} \end{cases}$$

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

Proof. By triangular inequality we have:

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

Now we 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 B.3 and we obtain with the same notation:

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

and:

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

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:

$$\begin{aligned}
\left|F_{1,\theta}(\widehat{X}) - F_{1,\theta}(X^*)\right| &\leq \sqrt{d}\left\|x_0^*\right\|_2 \bar{R}\left\|\widehat{X} - X^*\right\|_{L^2} \\
\left|F_{2,\theta}(\widehat{X}) - F_{2,\theta}(X^*)\right| &\leq \sqrt{d}\left(\bar{A} + \frac{\bar{E}}{\lambda}\right)\left(\left\|\widehat{X} - X^*\right\|_{L^2}\left\|\widehat{h_\theta}\right\|_{L^2} + \bar{X}\left\|\widehat{h}_\theta - h_\theta^*\right\|_{L^2}\right) \\
&\qquad + \sqrt{d}\bar{A}\left(\left\|\widehat{X}\right\|_{L^2} + \bar{X}\right)\left\|\widehat{X} - X^*\right\|_{L^2} \\
\left|F_{3,\theta}(\widehat{X}) - F_{3,\theta}(X^*)\right| &\leq \sqrt{d}\left\|\widehat{X} - X^*\right\|_{L^2}\bar{\dot{E}}\left(\left\|\widehat{X}\right\|_{L^2} + \overline{X}\right)
\end{aligned}$$

and we obtain for the second part:

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

$$\text{with: } \begin{cases}
K_1 = \sqrt{d}\left\|x_0^*\right\|_2 \bar{R} + \sqrt{d}\bar{A}\bar{X} + \sqrt{d}\bar{\dot{E}}\overline{X} \\
K_2 = \sqrt{d}\left(\bar{A} + \frac{\bar{E}}{\lambda}\right) \\
K_3 = \sqrt{d}\bar{A} + \sqrt{d}\dot{\bar{E}} \\
K_4 = \sqrt{d}\left(\bar{A} + \frac{\bar{E}}{\lambda}\right)\bar{X}
\end{cases}$$

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

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

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:

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

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

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

□

B.3. Asymptotic normality proof

The demonstration of continuity of some functionals useful for proposition 4.1 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.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 $\Gamma\,:\,C\left([0\,,\,T]\,,\mathbb{R}^d\right)\,\rightarrow\mathbb{R}^p$ *is a linear functional defined by*

$$\Gamma(X) = \int_0^T\left(\frac{\partial\left(A_{\theta*}(t).X^*\right)}{\partial\theta} + \frac{1}{\lambda}\frac{\partial h_{\theta^*}(t,X^*)}{\partial\theta}\right)^T\left(\int_t^T R_{\theta^*}(T-t,T-s)X(s)ds\right)dt. \tag{B.4}$$

R_{θ^*} *is defined* by (3.1).

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

The first order optimal condition is

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

Equivalently, we have

$$\int_0^T \frac{\partial\left(A_{\widehat{\theta}}(t).\widehat{X} + r_{\widehat{\theta}}(t)\right)}{\partial\theta}^T h_{\widehat{\theta}}(t,\widehat{X}) + \frac{\partial h_{\widehat{\theta}}(t,\widehat{X})}{\partial\theta}^T\left(A_{\widehat{\theta}}(t).\widehat{X} + r_{\widehat{\theta}}(t) - \dot{\widehat{X}}\right)$$
$$+ \frac{1}{\lambda}\frac{\partial h_{\widehat{\theta}}(t,\widehat{X})}{\partial\theta}^T h_{\widehat{\theta}}(t,\widehat{X}) = 0 \tag{B.5}$$

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

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

with $\widetilde{\theta}$ being a random point between θ^* and $\widehat{\theta}$ and N defined as in B.2. By replacing in (B.5), we obtain:

$$\int_0^T H_1(t,\widehat{\theta},\widehat{X})dt\ \left(\widehat{\theta} - \theta^*\right) = \int_0^T H_2(t,\widehat{\theta},\widehat{X})\left(\widehat{X} - X^*\right)$$
$$- \frac{\partial\left(h_{\widehat{\theta}}(t,\widehat{X})\right)}{\partial\theta}^T\left(\dot{X}^* - \dot{\widehat{X}}\right)dt \tag{B.6}$$

with

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

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

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

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

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

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 (3.1). □

Proposition B.4. *Under conditions 1-8 and by defining* Γ *as in proposition B.3 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 [35]. 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 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. □

Appendix 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 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 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 $E^r_\theta := \left(E^T_{\theta,1}, \cdots, E^T_{\theta,d}\right)^T$ 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:

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

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:

$$\begin{array}{l} S(\widehat{X}; \theta, \lambda) := \int_0^T g(Q_\theta(t), \theta, t)dt \\ \left\{ \begin{array}{l} \dot{Q}_\theta = F(Q_\theta, \theta, t) \\ Q_\theta(T) = 0 \end{array} \right. \end{array} \tag{C.1}$$

For the next subsections we 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

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

with:

$$\begin{aligned}
&\tfrac{\partial G_i}{\partial E^r_{(k-1)d+h}}(Q_\theta,\theta,t) = -\delta_{i,h}\left(\tfrac{\widehat{h}}{\lambda} + A(t)\widehat{X}(t) - \dot{\widehat{X}}(t) + r(t)\right)_k \\
&\tfrac{\partial G}{\partial \theta}(Q_\theta,\theta,t) = -\left(h^T \tfrac{\partial A_i(t)}{\partial \theta}\right)_{1\leq i\leq d} - E\left(\tfrac{\partial\big(A(t)\widehat{X}(t)\big)}{\partial \theta} + \tfrac{\partial r(t)}{\partial \theta}\right)
\end{aligned}$$

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

$$\left(\frac{\partial H(Q_\theta,\theta)}{\partial E^r}\right)_{(j-1)d+i} = -\begin{pmatrix} 0 & A_j^t & 0 & A_i^t & 0 \end{pmatrix} - \frac{1}{\lambda}\begin{pmatrix} 0 & E_j^t & 0 & E_i^t & 0 \end{pmatrix}$$

Because:

- $\frac{\partial}{\partial E^r}\left(A_j^t E_i + A_i^t E_j\right) = \begin{pmatrix} 0 & A_j^t & 0 & A_i^t & 0 \end{pmatrix}$ where A_j^t is in $i-th$ position and A_i^t is in $j-th$ position.
- $\frac{1}{\lambda}\frac{\partial}{\partial E}\left(E_j^t E_i\right) = \begin{pmatrix} 0 & \frac{1}{\lambda}E_j^t & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & \frac{1}{\lambda}E_i^t & 0 & 0 \end{pmatrix}$ where E_j^t is in $i-th$ position and E_i^t is in $j-th$ position.

And:

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

- Because $\frac{\partial}{\partial \theta}\left(A_j^t E_i + A_i^t E_j\right) = E_i^t\frac{\partial A_j}{\partial \theta} + E_j^t\frac{\partial A_i}{\partial \theta}$ 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:

$$\begin{aligned}
&\tfrac{d}{dt}(\tfrac{\partial Q_\theta(t)}{\partial \theta}) = \tfrac{\partial F}{\partial Q}(Q_\theta(t),\theta,t)\tfrac{\partial Q_\theta(t)}{\partial \theta} + \tfrac{\partial F}{\partial \theta}(Q_\theta(t),\theta,t) \\
&\tfrac{\partial Q_\theta(T)}{\partial \theta} = 0
\end{aligned}$$

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

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

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:

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

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.

Appendix 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:

$$\begin{aligned}\frac{1}{2}\frac{\partial^2 S(X^*;\theta^*,\lambda)}{\partial\theta^T\partial\theta} = & \frac{\partial\left(A_{\theta^*}(t).X^*\right)}{\partial\theta}^T\frac{\partial h_{\theta^*}(t,X^*)}{\partial\theta}\\ & + \frac{\partial\left(h_{\theta^*}(t,X^*)\right)^T}{\partial\theta}\frac{\partial\left(A_{\theta^*}(t).X^*\right)}{\partial\theta}\\ & + \frac{1}{\lambda}\frac{\partial\left(h_{\theta^*}(t,X^*)\right)}{\partial\theta}^T\frac{\partial h_{\theta^*}(t,X^*)}{\partial\theta}\end{aligned}$$

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:

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

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] R. Bellman and K.J. Astrom. On structural identifiability. *Mathematical Biosciences*, 7:329–339, 1970.

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

[3] C. De Boor. *A practical guide to Splines*, volume 27 of *Applied Mathematical Sciences*. Spinger, 2001. MR1900298

[4] H. Brezis. *Functional Analysis*. Dunod, 1983. MR0697382

[5] N. J.-B. Brunel. Parameter estimation of ode's via nonparametric estimators. *Electronic Journal of Statistics*, 2:1242–1267, 2008. MR2471285

[6] 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. MR3180555

[7] B. Calderhead and M. Girolami. Estimating Bayes factors via thermodynamic integration and population MCMC. *Computational Statistics & Data Analysis*, 53(12):4028–4045, October 2009. MR2744303

[8] D.A. Campbell and O. Chkrebtii. Maximum profile likelihood estimation of differential equation parameters through model based smoothing state estimates. *Mathematical Biosciences*, 2013. MR3132050

[9] F. Clarke. *Functional Analysis, Calculus of Variations and Optimal Control*. Graduate Texts in Mathematics. Springer-Verlag London, 2013. MR3026831

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

[12] J. Brynjarsdottir and A. O'Hagan. Learning about physical parameters: The importance of model discrepancy. *Inverse Problems*, 30:24, 2014. MR3274591

[13] C.P. Fall, E.S. Marland, J.M. Wagner, and J.J. Tyson, editors. *Computational Cell Biology*. Interdisciplinary applied mathematics. Springer, 2002. MR1911592

[14] 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.

[15] 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.

[16] 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.

[17] M. Girolami and B. Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. volume 73, pages 1–37, 2011. MR2814492

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

[19] 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. MR2948913

[20] J.J. Waterfall, F.P. Casey, K.S. Brown, C.R. Myers, R.N. Gutenkunst, and J.P. Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLoS Computational Biology*, 3:e189, 2007. MR2369325

[21] J. Happel, I. Suzuki, P. Kokayeff, and V. Fthenakis. Multiple isotope tracing of methanation over nickel catalyst. *Journal of Catalysis*, 65:59–77, 1980.

[22] G. Hooker. Forcing function diagnostics for nonlinear dynamics. *Biometrics*, 65:928–936, 2009. MR2649866

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

[24] G. Hooker and S.P Ellner. Goodness of fit in nonlinear dynamics: Misspecified rates or misspecified states? *Annals of Applied Statistics*, 9(2):754–776, 2015. MR3371334

[25] Y. Huang and H. Wu. A bayesian approach for estimating antiviral efficacy in hiv dynamic models. *Journal of Applied Statistics*, 33:155–174, 2006. MR2223142

[26] 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. MR3262538

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

[28] H.L. Koul. Weighted empiricals and linear models. *Hayward, CA: Institute of Mathematical Statistics*, 21:105–175, 1992. MR1218395

[29] R.V. Gamkrelidze, L.S. Pontryagin, V.G. Boltyanskii, and E.F. Mischenko. *The Mathematical Theory of Optimal Processes*. Wiley-Interscience, 1962. MR0166037

[30] Z. Li, M.R. Osborne, and T. Prvan. Parameter estimation of ordinary differential equations. *IMA Journal of Numerical Analysis*, 25:264–285, 2005. MR2126204

[31] 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. MR2504205

[32] H. Miao, X. Xia, A.S. Perelson, and H. Wu. On identifiability of nonlinear ode models and applications in viral dynamics. *SIAM Review*, 53:3–39, 2011. MR2785878

[33] A.A. Milyutin and N.P. Osmolovskii. *Calculus of Variation and Op-*

timal control. Mathematical Monographs. American mathematical society, 1998. MR1641590

[34] 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.

[35] W.K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79:147–168, 1997. MR1457700

[36] 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. MR2589327

[37] 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. MR2368570

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

[39] D. Ruppert, M.P. Wand, and R.J. Carroll. *Semiparametric regression*. Cambridge series on statistical and probabilistic mathematics. Cambridge University Press, 2003. MR1998720

[40] E. Sontag. *Mathematical Control Theory: Deterministic finite-dimensional systems*. Springer-Verlag (New-York), 1998. MR1640001

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

[42] R. Tuo and C.F.J. Wu. Efficient calibration for imperfect computer models. *Annals of Statistics*, 2015. MR3405596

[43] A.W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilities Mathematics. Cambridge University Press, 1998. MR1652247

[44] 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. MR0651865