



State and Parameter Estimation of Partially Observed Linear Ordinary Differential Equations with Deterministic Optimal Control

Quentin Clairon, Nicolas J-B. Brunel

► To cite this version:

Quentin Clairon, Nicolas J-B. Brunel. State and Parameter Estimation of Partially Observed Linear Ordinary Differential Equations with Deterministic Optimal Control. 2014. hal-01078173

HAL Id: hal-01078173

<https://hal.science/hal-01078173>

Preprint submitted on 28 Oct 2014

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

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



Distributed under a Creative Commons Attribution 4.0 International License

State and Parameter Estimation of Partially Observed Linear Ordinary Differential Equations with Deterministic Optimal Control

Quentin Clairon, Nicolas J-B. Brunel

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

Abstract

Ordinary Differential Equations are a simple but powerful framework for modeling complex systems. Parameter estimation from times series can be done by Nonlinear Least Squares (or other classical approaches), but this can give unsatisfactory results because the inverse problem can be ill-posed, even when the differential equation is linear.

Following recent approaches that use approximate solutions of the ODE model, we propose a new method that converts parameter estimation into an optimal control problem: our objective is to determine a control and a parameter that are as close as possible to the data. We derive then a criterion that makes a balance between discrepancy with data and with the model, and we minimize it by using optimization in functions spaces: our approach is related to the so-called Deterministic Kalman Filtering, but different from the usual statistical Kalman filtering.

We show the root- n consistency and asymptotic normality of the estimators for the parameter and for the states. Experiments in a toy model and in a real case shows that our approach is generally more accurate and more reliable than Nonlinear Least Squares and Generalized Smoothing, even in misspecified cases.

Keywords: Ordinary Differential Equation, Optimal Control, Parameter Estimation, Smoothing, Riccati Equation, M-estimation.

1 Introduction

Ordinary Differential Equations (ODE) are a widely used class of mathematical models in biology, physics, engineering, ... Indeed, it is a relatively simple but powerful framework for expressing the main mechanisms and interactions of potentially complex systems. It is often a reference framework in population dynamics and epidemiology [13], virology [29], or in genetics for describing gene regulation networks [26, 40]. The model takes the form $\dot{x} = f(t, x, \theta)$, where f is a vector field, x is the state, and θ is a parameter that can be partly known. The parameter θ is often of high interest, as it represents rates of changes, phenomenological constants needed for interpretability and analysis of the system. Typically, θ can be related to the sensitivity of a variable with respect to other variables.

Hence, the parameter estimation of ODEs from experimental data is a long-standing statistical subject that have been addressed with many different tools. Estimation can be done with classical estimators such as Nonlinear Least Squares (NLS) and Maximum Likelihood Estimator (MLE) [24, 39, 31] or Bayesian approaches [21, 15, 16, 9]. Nevertheless, the statistical estimation of an ODE model by NLS leads to a difficult nonlinear estimation problem. Some difficulties were pointed out by Ramsay et al. [33] such as computational complexity, due to ODE integration and nonlinear optimization. These difficulties are in fact reminiscent of intrinsic difficulties in the parameter estimation problem, that makes it an ill-posed inverse problem, that needs some regularization [14, 36].

Alternative statistical estimators have been developed to deal with this particular framework, such as Generalized Smoothing [33, 32, 12, 10] or Two-Step estimators [38, 5, 25, 17, 6]. Two-step estimators use a nonparametric estimator \hat{X} and aim at minimizing quantities characterizing the differential models, such as the weighted L^2 distance $\int_0^T \left\| \dot{\hat{X}}(t) - f(t, \hat{X}(t), \theta) \right\|^2 w(t) dt$. These estimators have a good computational efficiency as they avoid repeated ODE integration. In practice, the used criteria are also smoother and easier to optimize than the NLS criterion. Two-step estimators are consistent in general, but there is a trade-off with the statistical precision, and some care in the use of nonparametric estimate \hat{X} has to be taken in order to keep a parametric rate [5, 17].

In the case of Generalized Smoothing [33], 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. Based on

the Generalized Profiling approach, Hooker proposed a criteria that estimates the lack-of-fit through the estimation of a “forcing function” $t \mapsto u(t)$ in the ODE $\dot{x} - f(t, x, \tilde{\theta}) = u(t)$, where $\tilde{\theta}$ is a previous estimate obtained by Generalized Profiling.

In [7], the authors have proposed a two-step estimator for linear models, that avoids the use of \hat{X} and introduces a forcing function without the finite basis decomposition by using control theory. The principle is to transform the estimation problem into a control problem: we have to find the best (or smallest) control u such that the ODE is close to the data. The limitations of the results provided in [7] were the restriction to fully observed system with known initial condition. The objective of this paper is to provide a similar two-step estimate that permits the estimation of θ without knowing x_0 , that deals with the partially observed case and provides state estimates.

One interest of the approach used is to deal directly with the optimization in a function space without using of series expansion for function estimation. Moreover, infinite dimensional optimization tools give a powerful characterization of the solutions, useful in practice. This work can be seen as an extension of the previous one [7], aiming to use control theory result for parameter inference. We deal now with the partially observed case with unknown initial condition, that gives rise to a methodology close to the so-called “Deterministic Kalman Filter”. Indeed, in that paper, we assume that the system is linear, with a linear observation function.

Our method provides a consistent parametric estimator when the model is correct. We show that it is root- n consistent and asymptotically normal. At the same time, we get a discrepancy measure between the model and the data under the form of an optimal control u analogous to the forcing function in [19], and we show that we can estimate the final and initial conditions and hence all the states if needed, in particular the hidden ones.

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 investigate the existence and regularity of our new criterion; in particular, we derive necessary and sufficient conditions for defining our approach in partially observed case. We show that the estimator is consistent under some regularity assumption about the model. Then in section 4, we show that we reach the root- n rate using regression splines for \hat{Y} the nonparametric estimator of the observed signal. We derive then the consistency of the state estimator derived. Finally, we show the interest of our method on

a toy model and in a real model used in chemical engineering, by a comparison with Nonlinear Least Squares and Generalized Smoothing.

2 Model and methodology

We introduce first the statistical ODE model of interest, and the basic notations for defining our estimator. We relate this work to the Generalized Smoothing estimator and the Tracking estimator.

2.1 Model and Notations

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

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

where ϵ_i is a random noise and C is the observation matrix of size $d' \times d$.

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

$$\dot{x}(t) = A_\theta(t)x(t) + r_\theta(t) \quad (1)$$

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, x_0} the solution of (1) for a given θ , and initial condition x_0 . We assume that x_0^* and θ^* are unknown, and that they must be estimated from the data (y_1, \dots, y_n) . The parameter θ^* is the main parameter of interest, whereas the initial condition is considered as a nuisance parameter, needed essentially for the computation of candidate trajectories X_{θ, x_0} .

For linear equations, a central role is played by the solutions of the homogeneous ODE

$$\dot{x}(t) = A_\theta(t)x(t). \quad (2)$$

Indeed, for each s in $[0, T]$, we denote $t \mapsto \Phi_\theta(t, s)$ the solution to the matrix ODE (2), with initial condition I_d at time s (i.e $\Phi_\theta(s, s) = I_d$). The function $(t, s) \mapsto \Phi_\theta(t, s)$ is a $d \times d$ matrix valued function, called the resolvent of the ODE. It permits to give an explicit dependence of the solutions of (1) in r_θ and

the initial condition x_0 , thanks to Duhamel's formula:

$$X_{\theta, x_0}(t) = \Phi_\theta(t, 0)x_0 + \int_0^t \Phi_\theta(t, s)r_\theta(s)ds.$$

A consistent and classical method for the estimation of θ^* is Nonlinear Least Squares (NLS), that minimizes

$$\sum_{i=1}^n \|Y_i - CX_{\theta, x_0}(t_i)\|_2^2.$$

A classical alternative is Generalized Smoothing (GS), that uses approximate solutions of the ODE (1). GS replaces the solutions X_{θ, x_0} by splines that smooth data and solve approximately the ODE with a penalty based on the ODE model. A basis expansion $\hat{X}(t, \theta) = \hat{\beta}(\theta)^T p(t)$ is computed for each θ , where $\hat{\beta}(\theta)$ is obtained by minimizing in β the criterion

$$J_n(\beta|\theta, \lambda) = \sum_{i=1}^n \|Y_i - C\beta^T p(t)\|_2^2 + \lambda \int_0^T \|\beta^T \dot{p}(t) - (A_\theta(t)\beta^T p(t) + r_\theta(t))\|_2^2 dt \quad (3)$$

This first step is considered as profiling along the nuisance parameter β , whereas the estimation of the parameter of interest is obtained by minimizing the sum of squared errors of the proxy $\hat{X}(t, \theta)$:

$$\hat{\theta}^{GS} = \arg \min_{\theta} \sum_{i=1}^n \|Y_i - C\hat{X}(t_i, \theta)\|^2 \quad (4)$$

In practice, the hyperparameter λ needs to be selected from the data with adaptive procedures, see [11].

The essential difference with NLS is the replacement of the exact solution X_{θ, x_0} by the approximation $\hat{X}(\cdot, \theta)$ (that depends also on the data). This change induces a new source of error in the estimation of the true trajectory $t \mapsto X^*(t)$ as the functions $\hat{X}(\cdot, \theta)$ are splines that do not solve exactly the ODE model (1). The ODE constraint is relaxed into an inequality constraint defined on the interval $[0, T]$. The model constraint is never set to 0 because of the trade-off with the data-fitting term $\sum_{i=1}^n \|Y_i - C\beta^T p(t)\|_2^2$. For this reason, the ODE model (1) is not solved and it is useful to introduce the discrepancy term $\hat{u}_\theta(t) = \beta^T \dot{p}(t) - (A_\theta(t)\beta^T p(t) + r_\theta(t))$ that corresponds to a model error. In fact, the proxy $\hat{X}(\cdot, \theta)$ satisfies the perturbed ODE $\dot{x} = A_\theta x + r_\theta + \hat{u}_\theta$. This forcing

function \hat{u}_θ is an outcome of the optimization process and can be relatively hard to analyze or understand, but its analysis provides a good insight into the relevancy of the model [19, 20].

Based on these remarks, we introduce the perturbed linear ODE

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

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

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

is denoted $X_{\theta, x_0, u}$. Instead of using the spline proxy $\hat{X}(\cdot, \theta)$ for approximating X^* , we use the trajectories $X_{\theta, x_0, u}$ of the ODE (5) controlled by the additional functional parameter u .

In [7], the same perturbed model is introduced but the cost function is simpler as the observation matrix C is the identity, and the initial condition is fixed. In that framework, an M-estimator for θ is proposed, based on the optimization of the criterion

$$\tilde{S}(\hat{Y}; x_0, \theta, \lambda) = \inf_{u \in L^2} \{ \|\hat{Y} - X_{\theta, x_0, u}\|_{L^2}^2 + \|u\|_{L^2}^2 \}. \quad (6)$$

The proper definition of \tilde{S} and the derivation of its properties were obtained by using some classical results of Optimal Control Theory. Essentially, the computation of \tilde{S} corresponds to the classical "tracking problem" that can be solved by the Linear-Quadratic theory (LQ theory). LQ theory solves the minimization problem in L^2 of the cost function

$$C(u) = \|X_{\theta, x_0, u}(t)\|_{L^2}^2 + \|u(t)\|_{L^2}^2 + X_{\theta, x_0, u}(T)^\top Q X_{\theta, x_0, u}(T) \quad (7)$$

The criteria \tilde{S} used for parameter estimation is associated to the *value function* defined in Optimal Control as $S(t, x) = \inf\{C(u) | X_{\theta, x, u}(t) = x\}$. The value function plays a critical role in the analysis of optimal control problems, typically for the computation of an optimal policy. Under regularity assumptions, the value function S is the solution of the Hamilton-Jacobi-Bellman Equation, which is a first order Partial Differential Equation [1]. Quite remarkably, for a

linear ODE with a quadratic cost such as (7), the value function is a quadratic form in the state x , i.e $S(t, x) = -x^\top E(t)x$, where $E(t)$ is the solution of a matrix ODE (the Riccati equation), which makes its computation very tractable in practice.

LQ theory can be adapted for tracking of an output signal $\hat{Y} = CX^* + \epsilon$ with a perturbed linear ODE, see chapter 7 in [35]. When we do not know the initial condition, some adaptations are required. Indeed, as the initial condition can have a strong influence on the optimal control and the optimal cost; it seems much harder to solve the control problem when the initial condition is not known: the current state $x(t)$ is unknown and all the admissible trajectories must be considered. Nevertheless, this problem is solved by the Deterministic Kalman Filter (DKF) by using the fact that the value function S is a quadratic form on the state.

We show in the next section that the Deterministic Kalman Filtering (DKF) is well adapted for developing parameter estimation, as it enables to profile on x_0 , considered as a nuisance parameter. In a two-step approach, it is critical as we need to control the influence of the nonparametric estimate of \hat{Y} on the convergence rate. As we use $\hat{Y}(0)$ as a proxy for Cx_0^* , we need to show that the rate of the two-step estimator is not polluted by the use of nonparametric estimates of the boundary conditions, and that we keep a parametric rate for θ^* and x_0^* . This property was carefully checked in [5, 6, 25]; in that paper, as we do not use implicitly or explicitly the derivative of the nonparametric estimate, the mechanics of the proof are different.

In the next section, we give some details on LQ theory and on the criterion S . The classical costs in optimal control consist of an integral term plus a penalty term on the final state, such as $X_{\theta, x_0, u}(T)^\top Q X_{\theta, x_0, u}(T)$. A preliminary time-reversing transformation is used for introducing properly the initial state in the cost C , rather than the final state. In a second step, we derive the criterion S , and we give a tractable expression for estimation. Finally, we discuss the importance of identifiability and observability in the definition on our criterion.

2.2 The Deterministic Kalman Filter and the profiled cost

Following the Tracking estimator, we look for a candidate $X_{\theta, x_0, u}$ that minimizes at the same time the discrepancy with the data and the size of the perturbations

$\|u\|_{L^2}$. We consider nearly the same cost as in [7]

$$\tilde{C}(\hat{Y}; x_0, u, \theta, \lambda) = \int_0^T \left\| \hat{Y}(t) - CX_{\theta, x_0, u}(t) \right\|_2^2 dt + \lambda \int_0^T \|u(t)\|_2^2 dt \quad (8)$$

for given $\lambda > 0$. We can also add a positive quadratic form $x_0^T Q x_0$, where Q is a positive symmetric matrix Q . This additional term permits to introduce easily some prior knowledge on x_0 such that we have a cost defined as

$$C(\hat{Y}; u, x_0, \theta, \lambda) = x_0^T Q x_0 + \tilde{C}(\hat{Y}; x_0, u, \theta, \lambda).$$

Moreover, the matrix Q avoids some technical problems in the definition of our criterion S .

For each θ in Θ , we denote

$$S(\hat{Y}; \theta, \lambda) = \inf_{x_0} \left\{ \inf_{u \in L^2} C(\hat{Y}; x_0, u, \theta, \lambda) \right\} \quad (9)$$

obtained by “profiling” on the function u and then in the initial condition x_0 . The function $\tilde{S}(\hat{Y}; x_0, \theta, \lambda) = \inf_{u \in L^2} C(\hat{Y}; x_0, u, \theta, \lambda)$ is the criterion used in the case of fixed and known initial conditions. Our approach is rather “natural” as we simply profile the regularized criterion $x_0^T Q x_0 + \tilde{S}(\hat{Y}; x_0, \theta, \lambda)$.

The definition of S mimics the minimization of $J_n(\beta|\theta, \lambda)$ except that GS uses a discretized solution, defined on a B-splines basis. Nevertheless, our estimator possesses two other essential differences with Generalized Smoothing. As it was already mentioned in [7], we define our estimator as the global minimum of the profiled cost:

$$\hat{\theta}^K = \arg \min_{\theta \in \Theta} S(\hat{Y}; \theta, \lambda) \quad (10)$$

whereas the GS estimator minimizes a different criterion $\sum_{i=1}^n \|Y_i - C\hat{X}(t_i, \theta)\|^2$. This means that in our methodology, we try to find a parameter θ that maintain a reasonable trade-off between model and data, whereas the Generalized Smoothing Estimator $\hat{\theta}^{GS}$ is dedicated to fit the data with the proxy $\hat{X}(\cdot, \theta)$, without considering the size of model error represented by \bar{u}_θ . Another important difference is in the way we deal with the unobserved part of the system. For simplicity, let us consider that we observe only the first $k < p$ components of X , such that the state vector can be written $X = (X^{obs}, X^{unobs})$. For Generalized Smoothing, both functions X^{obs} and X^{unobs} are decomposed in a B-spline basis, and the corresponding coefficients β^{obs} and β^{unobs} are obtained by mini-

mizing $J_n(\beta^{obs}, \beta^{unobs}|\theta, \lambda)$. Because β^{unobs} does not have to make a trade-off between the data and the ODE model, the estimated missing part $\hat{X}^{unobs}(\cdot, \theta)$ is the exact solution to the ODE (5). At the contrary, even in the case of partial observations, the perturbed solution $X_{\theta, x_0, u}$ is used for estimating the missing states and a perturbation exists for each component. Consequently, the estimated hidden states are not solution of the initial ODE. We think that this is an advantage for state and parameter estimation with respect to Generalized Smoothing (and NLS) because it avoids to rely too strongly on a uncertain model during estimation. This uncertainty can be caused by errors in parameter estimation, or it can be due model misspecification, such as the presence of a forcing function u^* . In our experiments, we show that imposing model uncertainty for the unobserved variables is beneficial for error prediction.

Before going deeper into the interpretation and analysis of our estimator, we need to show that the criterion $S(\hat{Y}; \theta, \lambda)$ is properly defined and that we can obtain a tractable expression for computations and for the theoretical analysis of (10). We use the Deterministic Kalman Filter (DKF) to obtain a closed-form expression for the minimal cost w.r.t the control u and x_0 (9).

The initial aim of the DKF is to propose an estimation of the final state $X^*(T)$ by making a balance between the information brought by the noisy signal \hat{Y} and the ODE model (see [35] for an introduction). We recall the two steps necessary for the filter construction, more details are given in appendix:

1. For a given initial condition x_0 , we find the minimum cost thanks the fundamental theorem in LQ Theory (presented in A.1),
2. We minimize the quadratic form w.r.t the final condition.

We give now the main theorem of that section about the existence of the criterion defined in equation (9).

Theorem and Definition of $S(\zeta; \theta, \lambda)$. *Let $t \mapsto \zeta(t)$ be a function belonging to $L^\infty([0, T], \mathbb{R}^{d'})$ and $X_{\theta, x_0, u}$ be the solution to the controlled ODE (5). For any θ in Θ , $\lambda > 0$, $Q > 0$, there exists a unique optimal control $\bar{u}_{\theta, \lambda}$ and initial condition \widehat{x}_0 that minimizes the cost function*

$$C(\zeta; u, x_0, \theta, \lambda) = x_0^T Q x_0 + \int_0^T \left\{ \|\zeta(t) - C X_{\theta, x_0, u}(t)\|_2^2 + \lambda \|u(t)\|_2^2 \right\} dt \quad (11)$$

The optimal control $\bar{u}_{\theta,\lambda}$ is

$$\bar{u}_{\theta,\lambda}(t) = \frac{1}{\lambda} (E_{\theta}(t)X_{\theta,\widehat{x}_0,\bar{u}_{\theta,\lambda}}(t) + h_{\theta}(t,\zeta)) \quad (12)$$

where E_{θ} and h_{θ} are solutions of the Initial Value Problem

$$\begin{cases} \dot{E}_{\theta}(t) = C^T C - A_{\theta}^T E_{\theta} - E_{\theta} A_{\theta} - \frac{1}{\lambda} E_{\theta}^2, \\ \dot{h}_{\theta}(t,\zeta) = -\alpha_{\theta}(t)h_{\theta}(t,\zeta) - \beta_{\theta}(t,\zeta) \end{cases} \quad (13)$$

with $(E_{\theta}(0), h_{\theta}(0, \zeta)) = (Q, 0)$. The functions α_{θ} and β_{θ} are defined by

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

For all $t \in [0, T]$, the matrix $E_{\theta}(t)$ is symmetric, and the ODE defining the matrix-valued function $t \mapsto E_{\theta}(t)$ is called the *Matrix Riccati Differential Equation* of the ODE (5).

Finally, the Profiled Cost S has the closed form:

$$\begin{aligned} S(\zeta; \theta, \lambda) &= \int_0^T (\|\zeta(t)\|^2 - 2r_{\theta}(t)^T h_{\theta}(t, \zeta) - \frac{1}{\lambda} \|h_{\theta}(t, \zeta)\|^2) dt \\ &\quad - h_{\theta}(T, \zeta)^T E_{\theta}(T)^{-1} h_{\theta}(T, \zeta). \end{aligned} \quad (14)$$

and the final state is estimated by

$$X_{\theta,\widehat{x}_0,\bar{u}_{\theta,\lambda}}(T) = -E_{\theta}(T)^{-1} h_{\theta}(T, \zeta). \quad (15)$$

Remark 2.1. The functions $t \mapsto (E(t), h(t))$ are classically called the *adjoint model*. They depend also on θ , λ and ζ because of their definition via equation (13). Nevertheless, we do not write it systematically for notational brevity. As mentioned in the theorem, it is possible to compute $X_{\theta,\widehat{x}_0,\bar{u}_{\theta,\lambda}}$ in a “closed-loop” form as we can solve in a preliminary stage the adjoint model (13) that gives the function E and h for all $t \in [0, T]$. Thanks to equation (12), the closed-form expression of the optimal control $\bar{u}_{\theta,\lambda}$ can be plugged into (5). We can compute $X_{\theta,\widehat{x}_0,\bar{u}_{\theta,\lambda}}$ by solving the following Final Value Problem:

$$\begin{cases} \dot{x}(t) = \left(A_{\theta}(t) + \frac{E(t)}{\lambda} \right) x(t) + r_{\theta}(t) + \frac{h(t, \widehat{Y})}{\lambda} \\ x(T) = -E(T)^{-1} h(T, \widehat{Y}). \end{cases} \quad (16)$$

The estimate of $\widehat{x}_0 = X_{\theta, \widehat{x}_0, \bar{u}_{\theta, \lambda}}(0)$ of the initial condition is simply the initial value of the Backward ODE (16). Then by using $X_{\theta, \widehat{x}_0, \bar{u}_{\theta, \lambda}}$, we can compute effectively the control \bar{u}_{θ} thanks to (12).

The existence of the criterion S and the fundamental expression (14) heavily relies on the nonsingularity of the final value of the Riccati solution $E_{\theta}(T)$. In particular, the final state is estimated by $X_{\theta, \widehat{x}_0, \bar{u}_{\theta, \lambda}}(T) = -E_{\theta}(T)^{-1}h_{\theta}(T, \zeta)$, and it is then critical to identify the assumptions that could prevent E_{θ} to be singular. Our "Theorem and Definition" is legitimate (and proved in the appendix), because the assumption $Q > 0$ ensures that $E_{\theta}(t)$ is nonsingular for all t in $[0, T]$. Moreover, the matrix Q can be thought as a kind of prior for helping the state inference. In our basic definition of the cost (2.2), we put a prior on the norm of the initial condition and our regularization penalizes "huge" solutions. Nevertheless, we can have a more refined prior and use a preliminary guess $\mu \in \mathbb{R}^d$. The modification of the criterion is straightforward by setting

$$C_{\mu}(\zeta; u, x_0, \theta, \lambda) = (x_0 - \mu)^T Q (x_0 - \mu) + \|\zeta(t) - CX_{\theta, x_0, u}(t)\|_{L^2}^2 + \lambda \|u(t)\|_{L^2}^2.$$

By re-parameterizing the initial conditions with $y_0 = x_0 - \mu$ and exploiting the relation $X_{\theta, x_0 - \mu, u}(t) = X_{\theta, x_0 - \mu, u}(t) - \Phi_{\theta}(t, 0)\mu$ (consequence of the linearity of the ODE), we get that

$$\inf_{x_0} \inf_u C_{\mu}(\zeta; x_0, u, \theta, \lambda) = S(\zeta - C\Phi_{\theta}(\cdot, 0)\mu; \theta, \lambda).$$

At the opposite, it might be inappropriate in some circumstances to impose such kind of information for the initial condition. This can be the case if the number of observations tends to infinity and \hat{Y} becomes quite close to the truth. Another situation is when the initial conditions of the unobserved part are largely unknown. Hence, we extend our estimator to the case $Q = 0$, that corresponds also to our framework for studying the asymptotics of $\hat{\theta}^K$. In order to derive relevant and tractable conditions for ensuring the existence of S , we need to ensure that only one trajectory, with a unique initial condition (or final condition), is the global minimum of $C(\zeta; u, x_0, \theta, \lambda)$. The nonsingularity of $E_{\theta}(t)$ is in fact related to the concept of observability in control theory. In the next proposition, we will pave the way to the assumptions on C and the vector field A_{θ} that can guarantee the general existence of our method.

Proposition 2.2. *For a given parameter $\theta \in \Theta$ and observation matrix C , the*

properties 1 and 2 are equivalent:

1. The system outputs $Y(t) = C\Phi(t, 0)x_0$ satisfy

$$\int_0^T \|C\Phi_\theta(t, 0)x_{0,1} - C\Phi_\theta(t, 0)x_{0,2}\|_2 dt = 0 \implies x_{0,1} = x_{0,2} \quad (17)$$

2. The (final) observability matrix $O_\theta(T)$ is nonsingular

$$O_\theta(T) = \int_0^T \Phi_\theta(t, 0)^\top C^\top C \Phi_\theta(t, 0) dt \quad (18)$$

If one of the properties is satisfied, then $E_\theta(T)$ is nonsingular and S is defined for $Q \geq 0$.

An important feature of that proposition is that the criterion does not depend on r_θ . Moreover, if C is full rank, the matrix $E_\theta(T)$ is always nonsingular for all θ in Θ . The criterion 1 means that for a given θ , any solution $X_{\theta, x_{1,0}}$ and $X_{\theta, x_{2,0}}$ of (1) can be distinguished by their partial observation $Y_\theta^i(t) := CX_{\theta, x_{i,0}}(t)$, $i = 1, 2$. The matrix C "gives" enough information about the system so that the observed part is sufficient to uniquely characterize the whole system's state.

The next section is dedicated to the derivation of the regularity properties of S . Thanks to the different possible expressions for the criterion S , we can show the smoothness in ζ and θ , and compute directly the needed derivatives.

3 Consistency of the Deterministic Kalman Filter Estimator

3.1 Properties of the criterion $S(\hat{Y}; \theta, \lambda)$

We have a tractable expression of the cost function $S(\hat{Y}; \theta, \lambda)$ for a given θ , but we still need to derive the properties of $\theta \mapsto S(\hat{Y}; \theta, \lambda)$ and $\theta \mapsto S(Y^*; \theta, \lambda)$ on Θ , and shows some convergence properties. First of all, we need to ensure the existence of $S(\hat{Y}; \theta, \lambda)$; this is the case if the non-parametric estimator \hat{Y} belongs to $L^\infty([0, T], \mathbb{R}^{d'})$ (more explanations are given in appendix A). We show that for all Y in $L^\infty([0, T], \mathbb{R}^{d'})$, the function $\theta \mapsto S(Y; \theta, \lambda)$ is well defined and C^1 on Θ , under some regularity and identifiability assumptions, detailed below:

C1: Θ is a compact subset of \mathbb{R}^p and θ^* is in the interior $\mathring{\Theta}$,

C2a: $Q = 0$ and for all θ in Θ , $O_\theta(T)$ is nonsingular,

C2b: The model is identifiable at (θ^*, x_0^*) i.e

$$\forall (\theta, x_0) \in \Theta \times \mathcal{X}; CX_{\theta, x_0} = CX_{\theta^*, x_0^*} \implies (\theta, x_0) = (\theta^*, x_0^*),$$

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

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

Condition 2 is about identifiability condition: condition 2a is needed for the existence of the criterion S , and is related to the identifiability of the initial condition. But C2a is not sufficient, and we need Condition 2b for structural identifiability, based on the joint identifiability at (θ^*, x_0^*) . We require that the observed output CX_{θ^*, x_0^*} can be generated on by the couple (θ^*, x_0^*) . The identifiability problem of systems can be difficult (more than observability). For linear system, several approaches can be used, such as Laplace Transform [2], or Power Expansions [30], see [27] for a review. So far, most of existing methods are poorly used because they rely on (heavy) formal computations, which limit their interest to low dimensional system. Nonetheless, progress in automatic formal computation has promoted new methods based on differential algebra and the Ritt's algorithm, that improves identifiability checking, [22, 8, 23].

According to the context, the norm $\|\cdot\|_2$ will denote 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 the functional norm in $L^2([0, T], \mathbb{R}^{d'})$ defined by: $\|f\|_{L^2} = \sqrt{\int_0^T \|f(t)\|_2^2 dt}$. Continuity and differentiability have to be understood according to these norms.

Proposition 3.1. *Under conditions 1, 2a and 3 we have:*

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

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

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

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

and

$$\forall Y \in L^\infty([0, T], \mathbb{R}^{d'}), h_Y^- = \sup_{\theta \in \Theta} \|h_\theta(\cdot, Y)\|_{L^2} < +\infty$$

Hence, for all Y in $L^\infty([0, T], \mathbb{R}^{d'})$, the map $\theta \mapsto S(Y; \theta, \lambda)$ is well defined on Θ (i.e $\sup_{\theta \in \Theta} S(Y; \theta, \lambda) < +\infty$)

We have shown that for all Y in $L^\infty([0, T], \mathbb{R}^{d'})$, the maps $\theta \mapsto S(Y; \theta, \lambda)$ is well defined and so are $\theta \mapsto S(\hat{Y}; \theta, \lambda)$ and $\theta \mapsto S(Y^*; \theta, \lambda)$ as long as the non-parametric estimator \hat{Y} are well-defined on $[0, T]$.

Proposition 3.2. *Under conditions 1, 2, 3*

$$\forall Y \in L^\infty([0, T], \mathbb{R}^{d'}), \theta \mapsto S(Y; \theta, \lambda)$$

is continuous on Θ . Under conditions 1, 2a, 3, 4 it is C^1 on Θ .

In proposition 3.1 we have shown that our criteria $\theta \mapsto S(Y; \theta, \lambda)$ is well defined (i.e $0 \leq S(Y; \theta, \lambda) < +\infty$) and here we have demonstrated (using regularity assumptions on the model) that our finite and asymptotic criteria are continuous or even C^1 on Θ . Theses regularity properties justify the use of classical optimization method to retrieve the minimum of $S(\hat{Y}; \cdot, \lambda)$.

3.2 Consistency

We show the consistency of the parameter estimator $\hat{\theta}^K$ when the model is well-specified. As already mentioned, we have defined an M -estimator, and we can prove the consistency (see [37]), by showing

1. the uniform convergence of $S(\hat{Y}; \theta, \lambda)$ to $S(Y^*; \theta, \lambda)$ on Θ ,
2. θ^* is the unique global minimum of the asymptotic criterion $S(Y^*; \theta, \lambda)$ on Θ .

The second point is assessed in proposition 3.3, and it is related to the structure identifiability of the model provided by condition 2b.

Proposition 3.3. *Under conditions 1, 2a, 2b, we have:*

$$S(Y^*; \theta, \lambda) = 0 \iff \theta = \theta^*$$

Point 1 is proved by studying the regularity of the map $(\zeta, \theta) \mapsto S(\zeta; \theta, \lambda)$ and by obtaining appropriate controls of the variations by $\hat{Y} - Y^*$, see *Supplementary Materials*. Theorem (3.4) can be claimed with some generality on the nonparametric proxy \hat{Y} .

Theorem 3.4. *Under conditions 1, 2a, 2b, 3 and if \hat{Y} is consistent in probability in $L2$, then $\hat{\theta}^K \xrightarrow{P} \theta^*$.*

4 Asymptotics of $\hat{\theta}^K$

The aim of this section is to derive the rate of convergence and asymptotic law of $\hat{\theta}^K$. For this reason, we need more precise assumptions on \hat{Y} . The way we proceed is based on the plug-in properties of nonparametric estimates, when the functionals of interest are relatively smooth. In the case of series expansion, these properties are well understood [28, 4]. We focus here on regression splines, as they are well-used in practice and relatively simple to study, although more refined nonparametric estimators can be used in the same context, such as Penalized Splines. We assume that \hat{Y} has a B-Spline expansion

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

where β_K is computed by linear least-squares, and the dimension K increases with n . We introduce then additional regularity conditions on the ODE model, and on the distribution of observations:

- C5: $\forall (t, \theta) \in [0, T] \times \Theta$, $(t, \theta) \mapsto \frac{\partial^2 A_\theta}{\partial \theta^T \partial \theta}$ and $(t, \theta) \mapsto \frac{\partial r_\theta^2}{\partial \theta^T \partial \theta}$ are continuous,
- C6: $\frac{\partial^2 S(Y^*; \theta^*, \lambda)}{\partial \theta^T \partial Y}$ is nonsingular,
- C7: The observations (t_i, Y_i) are i.i.d with $\text{Var}(Y_i | t_i) = \sigma I_{d'}$ with $\sigma < +\infty$,
- C8: The observation times t_i are uniformly distributed on $[0, T]$,
- C9: It exists $s \geq 1$ such that $t \mapsto A_{\theta^*}(t)$, $t \mapsto r_{\theta^*}(t)$ are $C^{s-1}([0, T])\mathbb{R}^d$ and $\sqrt{n}K^{-s} \rightarrow 0$, $\frac{K^4}{n} \rightarrow 0$
- C10: The meshsize $\max_i |\tau_{i+1, K+1} - \tau_{i, K}| \rightarrow 0$ when $K \rightarrow \infty$

The proofs of the rate and asymptotic normality are somewhat technical, and they are relegated in the *Supplementary Materials*. We obtain a parametric convergence rate, and the asymptotic normality, by using two facts:

1. $\hat{\theta}^K - \theta^*$ behaves like the difference $\Gamma(\hat{Y}) - \Gamma(Y^*)$, where Γ is a linear functional,

2. if Γ is smooth enough, $\Gamma(\hat{Y} - Y^*)$ is asymptotically normal in the case of regression splines.

Conditions C5 and C6 ensures the sufficiency of second order optimality conditions for the criteria S . Conditions C7 to C10 are sufficient for the consistency of \hat{Y} , as well as for the consistency and the asymptotic normality of the plug-in estimators of linear functionals.

Theorem 4.1. *If conditions C1-C10 are satisfied, then $\hat{\theta}^K - \theta^* = O_P(n^{-1/2})$ and $\hat{\theta}$ is asymptotically normal.*

5 State Estimation

Once the unknown model has been estimated with $\hat{\theta}^K$, we focus on the problem of state estimation. From the definition (2.2), the criterion S is built with an estimation of the state based on the solution of the perturbed ODE $X_{\hat{\theta}, \hat{x}_0, \hat{u}}$. The estimate of the initial condition \hat{x}_0 is derived from a Final Value Problem with the final state $X_{\hat{\theta}, \hat{x}_0, \hat{u}}(T) = -[E_{\hat{\theta}}(T)]^{-1} h_{\hat{\theta}}(T, \hat{Y})$.

The state estimate $t \mapsto X_{\hat{\theta}, \hat{x}_0, \hat{u}}(t)$ that we have used, is different from the state estimation classically done when using the Deterministic Kalman Filter. The classical DKF state estimate is $\hat{X}^{DKF}(t) = -[E_{\hat{\theta}}(t)]^{-1} h_{\hat{\theta}}(t, \hat{Y}_{[0,t]})$, and it corresponds to the best estimate of $X^*(t)$ computed from the available information at time t , $\hat{Y}_{[0,t]} = \{\hat{Y}(s), s \in [0, t]\}$. Whereas the estimate can be very bad at the beginning for small t , the quality of $\hat{X}^{DKF}(t)$ improves as we get more data. A remarkable feature of the filter $t \mapsto \hat{X}^{DKF}(t)$ is that it can be computed recursively with an Ordinary Differential Equation $\dot{X}(t) = A_{\theta}X(t) + r_{\theta}(t) + L_{\theta, \lambda}(t)(CX(t) - \hat{Y}(t))$. The matrix $L_{\theta, \lambda}$ is the continuous counterpart of the classical Kalman Gain Matrix, derived from the Filtering Riccati Differential Equation, see page 313 in [35]. In that recursive form, the Deterministic Kalman Filter is somehow similar to the Kalman-Bucy Filter, which is the continuous version of the usual Kalman Filter. Nevertheless, there is a huge difference in the assumptions because the Kalman-Bucy Filter assumes that $X(t)$ is a Stochastic Differential Equation, driven by a Brownian Motion $W(t)$. This means that the deterministic perturbation $u(t)$ is replaced by a random perturbation $\sigma dW(t)$. The state estimate is then different from the one we consider as it can be shown that the filter is the solution of a stochastic differential equation driven by the stochastic process $(CX(t) - \hat{Y}(t))$, see for instance [3]. The state estimate $t \mapsto X_{\hat{\theta}, \hat{x}_0, \hat{u}}(t)$ is solution of the perturbed ODE,

with the control \bar{u} computed from all the data $\{\hat{Y}(s), s \in [0, T]\}$: hence, our state estimation is based on Kalman Smoothing and not on Filtering, as we have a backward integration step. In the rest of that section, we show that the estimator $X_{\hat{\theta}, \widehat{x_0}, \bar{u}}(t)$ is also a consistent estimator of the state $X^*(t)$. In order to do that, we show first that $\hat{X}(T) = -E_{\hat{\theta}}(T)^{-1}h_{\hat{\theta}}(T, \hat{Y})$ is a consistent estimator of the final state.

5.1 Final state estimation

In a way, the consistency of the final state estimator is a rather obvious conclusion. The Deterministic Kalman Filter is initially designed for getting the best possible estimate of the final state, starting from any initial condition x_0 . It is then normal that we have a good estimator of $X^*(T)$ when \hat{Y} is close to Y^* and $\hat{\theta}^K$ is close to θ^* .

Proposition 5.1. *We assume that conditions C1-C4 are satisfied and that \hat{Y} is a consistent estimator of Y^* . Then, the final state estimator $\hat{X}(T) = -E_{\hat{\theta}}(T)^{-1}h_{\hat{\theta}}(T, \hat{Y})$ converges in probability to $X^*(T)$.*

Proof. We show first that the true final state value is reached for $Y = Y^*$ and $\theta = \theta^*$ i.e $X^*(T) = -E_{\theta^*}(T)^{-1}h_{\theta^*}(T, Y^*)$. We recall that

$$S(Y; \theta, \lambda) = \inf_{x_0 \in \mathbb{R}^d} \left(\inf_{u \in L^2} C(Y; x_0, u, \theta, \lambda) \right),$$

and that $S(Y^*; \theta^*, \lambda) = 0$. The identifiability condition 2b implies that the reconstructed state is the exact one. In our case, the minimum is reached when the optimal control \bar{u} is equal to 0, i.e.

$$\bar{u}_{\theta^*, \lambda}(t) = \frac{1}{\lambda} (E_{\theta^*}(t)X^*(t) + h_{\theta^*}(t, Y^*))$$

which implies that $X^*(T) = -E_{\theta^*}(T)^{-1}h_{\theta^*}(T, Y^*)$ ($E_{\theta^*}(T)$ is nonsingular). We can decompose the difference $\hat{X}(T) - X^*(T)$:

$$\begin{aligned} \hat{X}(T) - X^*(T) &= E_{\theta^*}(T)^{-1}h_{\theta^*}(T, Y^*) - E_{\hat{\theta}}(T)^{-1}h_{\hat{\theta}}(T, \hat{Y}) \\ &= E_{\theta^*}(T)^{-1} \left(h_{\theta^*}(T, Y^*) - h_{\hat{\theta}}(T, \hat{Y}) \right) \\ &+ (E_{\theta^*}(T)^{-1} - E_{\hat{\theta}}(T)^{-1}) h_{\hat{\theta}}(T, \hat{Y}) \end{aligned}$$

The convergence will come from the consistency of $h_{\hat{\theta}}(T, \hat{Y})$ and $E_{\hat{\theta}}(T)^{-1}$:

$$\begin{aligned}
\left| \hat{X}(T) - X^*(T) \right| &\leq \sqrt{d} \|E_{\theta^*}(T)^{-1}\|_2 \left\| h_{\theta^*}(T, Y^*) - h_{\hat{\theta}}(T, \hat{Y}) \right\|_2 \\
&+ \sqrt{d} \left\| h_{\hat{\theta}}(T, \hat{Y}) \right\|_2 \|E_{\theta^*}(T)^{-1} - E_{\hat{\theta}}(T)^{-1}\|_2
\end{aligned}$$

The two right-hand side terms can be controlled easily by the $\hat{Y} - Y^*$ and $\hat{\theta} - \theta^*$, as it is shown in Lemma B.2 and B.3. We end up with the following inequalities:

$$\begin{aligned}
\|h_{\theta}(t, Y) - h_{\theta'}(t, Y')\|_2 &\leq K_6 e^{L_1 \frac{\overline{E_{\lambda}}}{\lambda}} \|Y - Y'\|_{L^2} \\
&+ (K_7 + K_8 \overline{E_{\lambda}}) \left(K_4 + \frac{K_5}{\lambda} \overline{E_{\lambda}} e^{L_1 \frac{\overline{E_{\lambda}}}{\lambda}} \right) e^{2L_1 \frac{\overline{E_{\lambda}}}{\lambda}} \|\theta - \theta'\| \\
&+ \left(K_9 e^{2L_1 \frac{\overline{E_{\lambda}}}{\lambda}} + K_{10} e^{L_1 \frac{\overline{E_{\lambda}}}{\lambda}} \right) \overline{E_{\lambda}} \|\theta - \theta'\| \\
\|E_{\theta}^{-1}(T) - E_{\theta'}^{-1}(T)\|_2 &\leq \left(\frac{K_{12}}{\lambda} + K_{11} \right) e^{K_{13} + \frac{K_{14}}{\lambda}} \|\theta - \theta'\| \\
\|E_{\theta}^{-1}(T)\| &\leq \frac{K_{15}}{\lambda} \\
\|h_{\theta}(T, Y)\|_2 &\leq \sqrt{T} d^2 \|C\|_2 e^{\sqrt{d} \left(\overline{A} + \frac{\overline{E_{\lambda}}}{\lambda} \right) T} \|Y\|_{L^2} + \sqrt{d \overline{E_{\lambda}} \overline{r_{\theta}}}
\end{aligned}$$

Under our conditions, we have $(\hat{\theta}, \hat{Y}) \rightarrow (\theta^*, Y^*)$, which implies that $\hat{X}(T)$ converges also in probability. \square

By plug-in principle, we can also derive the asymptotic normality and the rate of $\hat{X}(T)$ as described in the next proposition.

Proposition 5.2. *Under conditions C1-C10, the final state estimator $\hat{X}(T)$ is asymptotically normal and*

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

Proof. We have the following decomposition:

$$\begin{aligned}
\hat{X}(T) - X^*(T) &= E_{\theta^*}(T)^{-1} h_{\theta^*}(T, Y^*) - E_{\hat{\theta}}(T)^{-1} h_{\hat{\theta}}(T, \hat{Y}) \\
&= E_{\theta^*}(T)^{-1} \left(h_{\theta^*}(T, Y^*) - h_{\theta^*}(T, \hat{Y}) \right) \\
&+ E_{\theta^*}(T)^{-1} \left(h_{\theta^*}(T, \hat{Y}) - h_{\hat{\theta}}(T, \hat{Y}) \right) \\
&+ (E_{\theta^*}(T)^{-1} - E_{\hat{\theta}}(T)^{-1}) h_{\hat{\theta}}(T, \hat{Y})
\end{aligned}$$

According to Theorem 7 in [28] \hat{Y} is a consistent estimator of Y^* hence using proposition 5.1 and continuous mapping theorem we have:

$$\hat{X}(T) - X^*(T) = E_{\theta^*}(T)^{-1} \left(h_{\theta^*}(T, Y^*) - h_{\theta^*}(T, \hat{Y}) \right) + o_p(1)$$

Using the linear representation for h_{θ^*} we obtain:

$$E_{\theta^*}(T)^{-1} \left(h_{\theta^*}(T, Y^*) - h_{\theta^*}(T, \hat{Y}) \right) = -E_{\theta^*}(T)^{-1} \int_0^T R_{\theta^*}(T, s) C^T \left(\hat{Y}(s) - Y^*(s) \right) ds$$

We define

$$H(t, \theta).Y = E_{\theta}(T)^{-1} R_{\theta}(T, t) C^T Y(t)$$

the linear form such that

$$-E_{\theta^*}(T)^{-1} \left(h_{\theta^*}(T, Y^*) - h_{\theta^*}(T, \hat{Y}) \right) = \int_0^T \left(H(s, \theta^*).Y^* - H(s, \theta^*).\hat{Y} \right) ds$$

As for the normality of $\hat{\theta}^K$, we can use theorem 9 in [28] in order to obtain the asymptotic normality of $\int_0^T \left(H(s, \theta^*).Y^* - H(s, \theta^*).\hat{Y} \right) ds$ with \sqrt{n} -rate. \square

5.2 Estimation of the states on $[0, T]$ and influence of λ

We can estimate the trajectory $X^*(t)$ with the smoothed trajectory $t \mapsto X_{\hat{\theta}, \hat{x}_0, \bar{u}}(t)$ or with the exact model $t \mapsto X_{\hat{\theta}, \hat{x}_0, 0}$, without the perturbation \bar{u} . We need then to have a better understanding of the quality of these two estimates, and in particular of the relevancy of \hat{x}_0 , defined as the unknown initial condition of the Final Value Problem (16). We have profiled the initial condition in the definition of S , in order to separate the estimation of θ^* from the estimation of the initial condition. Nevertheless, the estimation of the states is a by-product of the parameter estimation, and the remaining point in our analysis is to ensure that \hat{x}_0 is really a good estimator for x_0^* . This is the case, and we will show more generally that $X_{\hat{\theta}, \hat{x}_0, \bar{u}}$ is a good estimator of the trajectory X^* . Quite remarkably, the consistency of $X_{\hat{\theta}, \hat{x}_0, \bar{u}}$ is the first result that relies on an assumption on the hyperparameter λ . This is due to the fact that \bar{u} is a perturbation computed for tracking \hat{Y} , while taking into account the model uncertainty estimated by $\hat{\theta}$ instead of θ^* . The convergence of \hat{Y} to Y^* and the identifiability conditions 2a and 2b (plus regularity conditions) are sufficient to ensure the convergence of $\hat{\theta}$ to θ^* , without particular assumptions on λ . This is possible because the true model $X_{\theta, x_0, 0}$ is included into the perturbed model $X_{\theta, x_0, u}$.

If λ is not big enough, the size of the perturbation $\|\bar{u}\|_{L^2}^2$ is not highly constrained in the cost function S , and we can have overfitting: the estimator $CX_{\hat{\theta}, \hat{x}_0, \bar{u}}$ can be quite close to \hat{Y} with a “big” \bar{u} that makes $X_{\hat{\theta}, \hat{x}_0, \bar{u}}$ far from

of X^* . This problem can be even more important, if we have errors on $\hat{\theta}$, because \bar{u} will have to compensate the errors in the parameter estimation. In that case, we cannot guarantee to have a consistent estimate for x_0 , if we don't have $\lambda \rightarrow \infty$. Indeed, the trajectory $X_{\hat{\theta}, \hat{x}_0, \bar{u}}$ is the solution to the perturbed initial value problem

$$\begin{cases} \dot{x}(t) = \left\{ A_{\hat{\theta}}(t) + \frac{1}{\lambda} E_{\hat{\theta}, \lambda}(t) \right\} x(t) + r_{\hat{\theta}}(t) + \frac{1}{\lambda} h(t, \hat{Y}) \\ x(T) = -E_{\hat{\theta}, \lambda}(T)^{-1} h(T, \hat{Y}) \end{cases} \quad (19)$$

Because of the convergence of $(\hat{\theta}, \hat{Y}) \rightarrow (\theta^*, Y^*)$, we can ensure the convergence to the right trajectory if we control λ .

Proposition 5.3. *Under conditions C1-C10 and if $\lambda_n \rightarrow \infty$, then*

$$X_{\hat{\theta}, \hat{x}_0, \bar{u}}(t) \rightarrow X^*(t)$$

for all $t \in [0, T]$. Moreover,

$$\hat{x}_0 - x_0^* = O_P(n^{-1/2}).$$

Proof. We first need to show that for all $\theta \in \Theta$, and ζ , the functions $E_{\theta, \lambda}$ and $h_{\theta, \lambda}(\cdot, \zeta)$ are bounded (as they converge) when $\lambda \rightarrow \infty$. As $E_{\theta, \lambda}$ is solution of the matrix equation $\dot{E} = C^\top C - A^\top E - EA - \frac{1}{\lambda} E^2$ that depends smoothly in λ^{-1} ; hence $E_{\theta, \lambda} \rightarrow E_{\theta, \infty}$ defined as the solution of the linear matrix ODE $\dot{E} = C^\top C - A^\top E - EA$ (with $E(0) = 0$). Moreover, $h_{\theta, \lambda}(\cdot, \zeta)$ is solution of the linear ODE $\dot{h} = -\alpha_{\theta, \lambda} h - \beta_{\theta, \lambda}(\cdot, \zeta)$ with $\alpha_{\theta, \lambda} \rightarrow A_\theta^\top$ and $\beta_{\theta, \lambda}(\cdot, \zeta) \rightarrow \beta_{\theta, \infty} = C^\top \zeta + E_{\theta, \infty} r_\theta$. As the dependency in λ^{-1} is smooth, the solution $h_{\theta, \lambda}(\cdot, \zeta)$ converges to $h_{\theta, \infty}$, solution of $\dot{h} = -\alpha_{\theta, \infty} h - \beta_{\theta, \infty}(\cdot, \zeta)$. Additionally, the dependency in (λ, ζ) is smooth on $\mathbb{R}^+ \times L^2$ and $h_{\theta, \lambda}(\cdot, \zeta)$ converges to $h_{\theta, \infty}(\cdot, Y^*)$ as (λ, ζ) tends to (∞, Y^*) . This means that if (θ, Y) converges to (θ^*, Y^*) as $\lambda \rightarrow \infty$, then $X_{\theta, x_0, \lambda}$ converges to the solution of the final value problem

$$\begin{cases} \dot{x}(t) = A_{\theta^*}(t)x(t) + r_{\theta^*}(t) \\ x(T) = X_{\theta^*, x_0^*, 0}(T) \end{cases} \quad (20)$$

as $\lambda^{-1} E_{\theta, \lambda}$ and $\lambda^{-1} h_{\theta, \lambda}(t, Y)$ tends to zero when $\lambda \rightarrow \infty$, and if $x(T) \rightarrow X^*(T)$. Because of the uniqueness of the solutions to Initial or Final Value Problem, we have $X_{\theta, x_0, \lambda} \rightarrow X^*$ for all $t \in [0, T]$.

In proposition 5.1, under conditions C1-C8, we have shown that $(\hat{\theta}_n^K, \hat{Y}_n)$ converges in probability to (θ^*, Y^*) for all λ on $\Theta \times L^2$. By the continuous mapping theorem applied to $(\hat{\theta}_n^K, \hat{Y}_n, \lambda_n)$, with $\lambda_n \rightarrow \infty$, we have $X_{\hat{\theta}, \hat{x}_0, \bar{u}_{\lambda_n}}(t) \rightarrow X^*(t)$ for all $t \in [0, T]$ in probability. In particular, we obtain the convergence of \hat{x}_0 to x_0^* .

The asymptotic normality and root-n rate of \hat{x}_0 comes from the asymptotic normality and rates of $\hat{\theta}$ and $\hat{X}(T)$. If $\psi_{\theta, \lambda}(t, 0)$ is the resolvent of the ODE $\dot{x}(t) = \{A_\theta(t) + \frac{1}{\lambda} E_{\theta, \lambda}(t)\} x(t)$, we have a closed form for the smoother

$$X_{\hat{\theta}, \hat{x}_0, \bar{u}_{\lambda_n}}(t) = \psi_{\hat{\theta}, \lambda}(t, 0) \hat{x}_0 + \psi_{\hat{\theta}, \lambda} \int_0^t \left[\psi_{\hat{\theta}, \lambda}(s, 0) \right]^{-1} \left\{ r_{\hat{\theta}}(s) + \frac{1}{\lambda} h_{\hat{\theta}, \lambda_n}(s, \hat{Y}) \right\} ds$$

When we evaluate at $t = T$, we obtain the following formula for the initial state $\hat{x}_0 = \left[\psi_{\hat{\theta}, \lambda}(T, 0) \right]^{-1} \hat{X}(T) - \int_0^T \left[\psi_{\hat{\theta}, \lambda}(s, 0) \right]^{-1} \left\{ r_{\hat{\theta}}(s) + \frac{1}{\lambda_n} h_{\hat{\theta}, \lambda_n}(s, \hat{Y}) \right\} ds$. Hence, \hat{x}_0 is a smooth transformation of $(\hat{\theta}, \hat{X}(T))$, and we can conclude by the parametric delta-method. \square

5.3 Choice of λ and cross-validation

Our theoretical analysis shows that when n tends to infinity, we have a family of good estimates $(\hat{\theta}_{\lambda_n}^K, \widehat{x_{0, \lambda_n}})$, with $\lambda_n \rightarrow \infty$. The remaining question is to define an appropriate selection procedure for λ , that could be used in practice with a finite number of observations (y_1, \dots, y_n) . A straightforward way of selecting λ is to use a cross-validation selection procedure. Indeed, our criterion $S(\hat{Y}; \theta, \lambda)$ is based on a balance between data fidelity and model fidelity, and a rough analysis shows that when $\lambda \rightarrow 0$, we can select any u in order to interpolate \hat{Y} and θ has almost no influence on $S(\hat{Y}; \theta, \lambda)$. Whereas when $\lambda \rightarrow \infty$, the optimal perturbation $\bar{u} \rightarrow 0$, and we get a NLS-like criterion where the observations Y_i 's are replaced by the proxy \hat{Y} .

A good hyperparameter λ_n should give a good estimate of the states $X^*(t)$ (and of the output $Y^*(t)$), even if we are only interested in parameter estimation. Anyway, if we want to use the minimization of prediction error for selecting λ and $\hat{\theta}_\lambda^K$, we need to have a good estimate of the initial condition x_0 as it is necessary for computing the predictions. We propose then to select λ by minimizing the Sum of Squared Errors

$$SSE(\lambda) = \sum_{i=1}^n \left\| Y_i - C X_{\hat{\theta}_\lambda^K, \widehat{x_{0, \lambda}}, 0}(t_i) \right\|_2^2. \quad (21)$$

Moreover, this criterion gives a way to reduce the influence of the nonparametric estimate \hat{Y} , as we use the original noisy data. This is the selection procedure that we implemented in the experiments part.

6 Experiments

We use two test beds for evaluating the practical efficiency of the deterministic Kalman filter estimator $\hat{\theta}^K$; we compare it with the NLS estimator $\hat{\theta}^{NLS}$ and the estimator obtained by Generalized Smoothing $\hat{\theta}^{GS}$. The two models are linear in the states, and they can be linear or nonlinear w.r.t parameters. We use several sample size and several variance error for comparing robustness and efficiency.

6.1 Experimental design

For a given sample size n and noise level σ , we estimate the Mean Square Error and the mean Absolute Relative Error (ARE) $\mathbb{E}_{\theta^*} \left[\frac{|\theta^* - \hat{\theta}|}{|\theta^*|} \right]$ by Monte Carlo, based on $N_{MC} = 100$ runs. For each run, we simulate an ODE solution with a Runge-Kutta algorithm (ode45 in Matlab), and a centered Gaussian noise (with variance σ) is added, in order to obtain the Y_i 's. We compare the accuracy of the 3 parameters $\hat{\theta}^K$, $\hat{\theta}^{GS}$ and $\hat{\theta}^{NLS}$, but we are also interested in their mean prediction error defined as

$$E_P(\hat{X}) = \mathbb{E}_{(Y_1, \dots, Y_n)} \left[\mathbb{E}_{\theta^*, \sigma} \left[\left\| Y^* - \hat{X} \right\| \right] \right] \quad (22)$$

where Y^* is a new observation generated with the parameters $(\theta^*, x_0^*, \sigma)$, and \hat{X} is an estimator of the trajectory, based on one of the three estimates $\hat{\theta}^K$, $\hat{\theta}^{GS}$ and $\hat{\theta}^{NLS}$. For the three estimators, the initial condition is estimated consistently:

NLS: $\widehat{x_0}$ is obtained simultaneously with the parameter estimation (as an additional parameter),

Kalman: $\widehat{x_0}$ and λ_n are selected as described in section 5.3,

Generalized Smoothing: $\widehat{x_0}$ is the initial value of the estimated curve corresponding to the estimated parameter $\hat{\theta}^{GS}$, with smoothing parameter λ_n selected adaptively as described in [33].

We insist on the fact that parameter estimation and prediction are two different statistical tasks, that are evaluated by different criteria. Parameter estimation is required when the parameter has an interest by itself or when the model has an explicative purpose, whereas the prediction error is dedicated to estimation of the state X , in the most efficient way. Our primary interest is parameter estimation but we also discuss prediction for the three methods; as we have seen in section 5, parameter estimation and state estimation are tightly related in particular for the selection of λ . We will consider two possible estimators for the state: a parametric estimator $X_{\hat{\theta}, \hat{x}_0}$ and a smoothed (or corrected) estimator $X_{\hat{\theta}, \hat{x}_0, \bar{u}}$. For the NLS estimator, the parametric and corrected state estimator are the same, whereas $X_{\hat{\theta}^{GS}, \hat{x}_0}$ and $\hat{X}(\cdot, \hat{\theta}^{GS})$ are different, as $X_{\hat{\theta}, \hat{x}_0}$ differs from $X_{\hat{\theta}, \hat{x}_0, \bar{u}}$.

The two test beds are partially observed models with one missing state variable. We compare the ability of the different methods to accurately reconstruct the hidden state. Thus, we compute for each estimator the L^2 -distance between the true missing state and the obtained reconstruction after parameter estimation:

$$\Delta(\hat{X}^{unobs}) = \mathbb{E} \left[\left\| X_{\hat{\theta}^*, x_0^*}^{unobs} - \hat{X}^{unobs} \right\|_{L^2} \right] \quad (23)$$

The nonparametric estimate \hat{Y} is a regression spline, with a B-spline basis defined on a uniform knot sequence $\xi_k, k = 1, \dots, K$. For each run and each state variables, the number of knots is selected by minimizing the GCV criterion, [34]. For optimizing the criterion S , we use the Matlab function 'fminunc' that implements a trust region algorithm for which gradient expression is required. The computation of the gradient of S w.r.t the parameter θ is computationally involved and is based on the sensitivity equations of the ODE model. The computational details are left in appendix C.

6.2 Toy Examples: Partially Observed ODE in 3 D

We consider the autonomous ODE

$$\begin{cases} \dot{x}_1 = -(k_1 + k_2)x_1 \\ \dot{x}_2 = k_1x_1 \\ \dot{x}_3 = k_2x_1 \end{cases} \quad (24)$$

where we observe only the variables x_2 and x_3 . Using the notation introduced in this paper, we have

$$A_\theta = \begin{pmatrix} -(k_1 + k_2) & 0 & 0 \\ k_1 & 0 & 0 \\ k_2 & 0 & 0 \end{pmatrix}$$

and

$$C = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

With that model, we show that the conditions introduced in the statistical analysis are workable on some simple models, in particular the conditions for identifiability C2a and C2b that needs to be checked. In the case of autonomous system (i.e when A_θ and r_θ do not depend on time), a simple sufficient and necessary criteria is the so-called Kalman criterion:

Proposition 6.1. *In the case of an autonomous model, the matrix $E_\theta(T)$ is nonsingular if and only if the matrix*

$$K_{A,C} = \begin{pmatrix} C \\ CA_\theta \\ \vdots \\ CA_\theta^{d-1} \end{pmatrix} \quad (25)$$

has a rank equals to d .

The matrix $K_{A,C}$ is usually called the Kalman matrix. In order to define properly our criterion S , we need to check that condition 2b is also satisfied (joint identifiability of θ and x_0). For this model, the analysis is relatively easy and we can use the characterisation proposed by [30] based on the power series expansion. As, the Kalman matrix (25) is

$$\begin{pmatrix} C \\ CA_\theta \\ CA_\theta^2 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ k_1 & 0 & 0 \\ k_2 & 0 & 0 \\ -k_1(k_1 + k_2) & 0 & 0 \\ -k_2(k_1 + k_2) & 0 & 0 \end{pmatrix},$$

the Kalman condition is fulfilled (i.e the matrix rank is 3) if $k_1 \neq 0$ or $k_2 \neq 0$. Hence, C2a holds for all relevant cases ($k_1 = 0$ or $k_2 = 0$ correspond to the case where x_2 and x_3 variations are disconnected from x_1 which makes the model useless for explanation or prediction purposes).

For condition C2b, we use the result shown by Pohjanpalo et al. If the model is $\dot{x} = f(t, x, \theta)$ and the observation function is $h(t, \theta, x)$, condition C2b is satisfied if the nonlinear system

$$h^{(j)}(x_0, \theta, x) = a_j(x_0) \quad j = 0, 1, \dots \quad (26)$$

has a unique solution θ . Pohjanpalo et al. showed that for linear autonomous system, this condition is sufficient and necessary. In our case, the equation (26) can be written as

$$CA_\theta^j X_0 = a_j \quad j = 0, 1, \dots$$

Since the initial condition $X_0 = (X_{0,1}, X_{0,2}, X_{0,3})$ is unknown, we have to consider the extended parameter $\theta = (k_1 k_2, X_0^T)$. The equations for $j = 0$

$$CX_0 = a_j \quad j = 0, 1, \dots$$

allow us to identify $X_{0,2} = a_{0,2}, X_{0,3} = a_{0,3}$. For $j = 1$, we have

$$\begin{cases} k_1 X_{0,1} &= a_{1,2} \\ k_2 X_{0,1} &= a_{1,3} \end{cases}$$

and the solutions are $X_{0,1} = \frac{a_{1,3}}{k_2}$ and $k_1 = \frac{a_{1,2}}{a_{1,3}} k_2$. Finally, we have a unique solution for k_2 , if we consider the additional equation (26) for $j = 2$. In that case, the system

$$\begin{cases} -k_1(k_1 + k_2)X_{0,1} &= a_{2,2} \\ -k_2(k_1 + k_2)X_{0,1} &= a_{2,3} \end{cases}$$

has a unique solution $k_2 = -\frac{a_{2,3}}{a_{1,2} + a_{1,3}}$.

Well-specified model (Toy Model 1) We test two sample sizes $n = 200$ and $n = 100$ (observations times are uniformly sampled between $t = 0$ and $t = 100$) and two noise levels $\sigma = 3$ and $\sigma = 6$. For the computation of the regression splines \hat{Y} , we select manually the knots location instead of using the GCV driven selection (to avoid overfitting). We have placed four equispaced knots respectively at time $t = 0, 33, 66$ and 100 . The true parameter is $\theta^* = (k_1^*, k_2^*) =$

(0.0593, 0.0296) and the initial condition x_0^* equals (0, 0, 100). For the Kalman estimator, we select λ_n by cross-validation among the values $\lambda^v = \{10^k\}_{k \in [5 \ 16]}$.

(n, σ)		MSE (10^{-6})	ARE (10^{-2})	$E_P \left(X_{\hat{\theta}, \hat{x}_0} \right)$	$\Delta \left(X_{\hat{\theta}, \hat{x}_0} \right)$
(200, 3)	$\hat{\theta}^{NLS}$	4.20	5.16	43.56	4.18
	$\hat{\theta}^K$	3.97	4.77	42.79	4.13
	$\hat{\theta}^{GS}$	12.87	12.23	46.80	9.28
(200, 6)	$\hat{\theta}^{NLS}$	17.09	9.92	87.76	8.43
	$\hat{\theta}^K$	16.49	9.43	85.45	8.28
	$\hat{\theta}^{GS}$	77.87	23.28	93.69	17.77
(100, 3)	$\hat{\theta}^{NLS}$	8.21	7.43	44.95	6.04
	$\hat{\theta}^K$	8.78	7.37	43.03	6.15
	$\hat{\theta}^{GS}$	22.32	12.60	48.01	9.45
(100, 6)	$\hat{\theta}^{NLS}$	36.89	15.27	90.76	12.24
	$\hat{\theta}^K$	34.98	14.91	86.19	12.36
	$\hat{\theta}^{GS}$	86.74	24.39	94.91	18.63

Tab. 1: Results for the Toy Model 1 ; partially observed.

The results are presented in table 1. The GS estimator is outperformed by the Kalman and NLS estimators, moreover our approach improves the parameter estimation accuracy in terms of MSE and ARE in almost every cases comparing to the NLS and also minimizes prediction error. Regarding the missing state reconstruction both methods gives similar results.

Misspecified model (Toy Model 2) In our simulation, we give also some insight in the case of misspecified models. Indeed, our perturbed 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 $v \in L^2([0, T], \mathbb{R}^d)$ an unknown function. We do not provide any theoretical analysis for this kind of model misspecification. The Kalman estimator gives more accurate estimation than the NLS estimator in that case, as we consider perturbations of the initial model. Moreover, the optimal control \bar{u} obtained along the parameter estimation can be used as a correction term to add to the initial model to counter-balance misspecification. This implies potentially a better prediction power. The true model is nearly the same model as above

$$\dot{X} = A_\theta X + v(t) \quad (27)$$

with $\theta^* = (k_1^*, k_2^*) = (0.0593, 0.0296)$ and $x_0^* = (0, 0, 100)$, but we add a perturbation $v : [0, T] \mapsto \mathbb{R}^3$ with entries equal to $0.4 \times \sin(\frac{t}{5})$. Nevertheless for parameter estimation, we still use the unperturbed model $\dot{X} = A_\theta X$.

In the case of the Kalman estimator $\hat{\theta}^K$, the optimal control \bar{u} can be used for correcting the model and for defining a new model

$$\dot{X} = A_{\hat{\theta}} X + \bar{u}. \quad (28)$$

We are then interested in evaluating the prediction error of $X_{\hat{\theta}, \hat{x}_0, \bar{u}}$, defined as $E_P(X_{\hat{\theta}, \hat{x}_0, \bar{u}})$. We also estimate the error between the true first state value and the obtained reconstruction with the corrected model. As shown in the introduction, Generalized Smoothing can also evaluate a correction term for $\hat{\theta}^{GS}$, defined as $\bar{u}(t) = \hat{X}(t, \hat{\theta}^{GS}) - A_{\hat{\theta}} \hat{X}(t, \hat{\theta}^{GS})$ (where $\hat{X}(t, \hat{\theta}^{GS})$ is the spline corresponding to the estimated parameter $\hat{\theta}^{GS}$ with adaptive $\hat{\lambda}$). In the case of NLS, we cannot compute a correction \bar{u} , as the estimated trajectories are exactly solution of the ODE for $\hat{\theta}^{NLS}$. In the case of Generalized Smoothing, we have $\Delta(X_1^*; X_{\hat{\theta}, \hat{x}_0}) \approx \Delta(X_1^*; X_{\hat{\theta}, \hat{x}_0, \bar{u}})$ because the hidden parts are (almost) exactly trajectories of the ODE with parameter $\hat{\theta}^{GS}$. The estimates that change is the Kalman-based one.

(n, σ)		MSE (10^{-5})	ARE (10^{-2})	$E_P(X_{\hat{\theta}, \hat{x}_0})$	$E_P(X_{\hat{\theta}, \hat{x}_0, \bar{u}})$	$\Delta(X_{\hat{\theta}, \hat{x}_0}^{unobs})$	$\Delta(X_{\hat{\theta}, \hat{x}_0, \bar{u}}^{unobs})$
(200, 3)	$\hat{\theta}^{NLS}$	4.14	19.13	52.24	52.24	19.08	19.08
	$\hat{\theta}^K$	3.66	17.46	47.78	47.75	18.82	18.87
	$\hat{\theta}^{GS}$	7.56	27.34	55.13	50.99	22.27	22.26
(200, 6)	$\hat{\theta}^{NLS}$	4.99	18.65	92.95	92.95	19.90	19.90
	$\hat{\theta}^K$	4.68	18.14	88.30	88.02	20.66	19.79
	$\hat{\theta}^{GS}$	13.21	29.66	97.25	94.64	26.09	26.08
(100, 3)	$\hat{\theta}^{NLS}$	4.88	19.56	52.66	52.66	19.48	19.48
	$\hat{\theta}^K$	4.56	18.53	48.07	47.86	19.76	19.23
	$\hat{\theta}^{GS}$	10.04	29.32	56.33	55.06	23.71	23.66
(100, 6)	$\hat{\theta}^{NLS}$	7.96	23.32	96.63	96.63	21.69	21.69
	$\hat{\theta}^K$	7.59	22.36	89.19	88.65	23.88	21.56
	$\hat{\theta}^{GS}$	15.63	32.77	101.41	98.81	26.14	26.15

Tab. 2: Results for Toy Model 2, partially observed model; misspecified case

The GS parameter estimator is outperformed by the Kalman and the NLS estimator. Our approach improves the estimation accuracy for θ (lower MSE and ARE in every cases) on the NLS estimator. This difference is bigger than in the well specified case (Toy Model 1), as we are more robust to the presence of a perturbation than the NLS. The Kalman estimator gives also better prediction error in every cases and the correction \bar{u} slightly improves the prediction errors. Nevertheless, the NLS estimator provides the smallest $\Delta(X_1^*; X_{\hat{\theta}, \hat{x}_0})$ among all estimation methods in every cases but the first one. Nonetheless using \bar{u} minimizes in most of case the error for X_1 estimation for our approach and allows us to obtain slightly better result than the NLS estimator.

The correction term \bar{u} is related (correlated) to the perturbation $t \mapsto v(t)$ as we can in figure 1, where we plot the mean of each component of \bar{u} , when $(n, \sigma) = (200, 3)$. Even though the scale is not the same (we need to rescale by 10^{-5} for easing comparisons), the correction \bar{u} exhibits some important features of the true one, such as oscillations with a period close to the period of v . The analysis of \bar{u} is beyond the scope of that paper, but the presence of strong patterns in u can be used to detect misspecification, in the same way that the analysis of residuals permits to detect lack of fit in regression models.

6.3 Real case example: Methanation reaction

We consider an ODE model introduced in [18] for describing the dynamics of carbon monoxide and hydrogen methanation over a supported nickel catalyst by transient isotopic tracer in a gradientless circulating reactor. This “Methanation reaction” model is a linear autonomous equation in \mathbb{R}^4 , with a forcing term. A important difference w.r.t the previous is the nonlinearity in parameters as we have

$$A_\theta = \begin{pmatrix} -\frac{V+V'+F_0^{C0}/W}{\beta C^{CO}/W+C^{CO}} & 0 & 0 & 0 \\ \frac{V+V'}{\beta C^{H_2O}/W} & -\frac{V+V'+v_5}{\beta C^{H_2O}/W} & 0 & \frac{v_5}{\beta C^{H_2O}/W} \\ \frac{V'}{\beta C^{CO_2}/W} & 0 & -\frac{V'+v_6}{\beta C^{CO_2}/W} & \frac{v_6}{\beta C^{CO_2}/W} \\ 0 & \frac{v_5}{C^{O_s}} & \frac{v_6}{C^{O_s}} & -\frac{v_5+v_6}{C^{O_s}} \end{pmatrix}$$

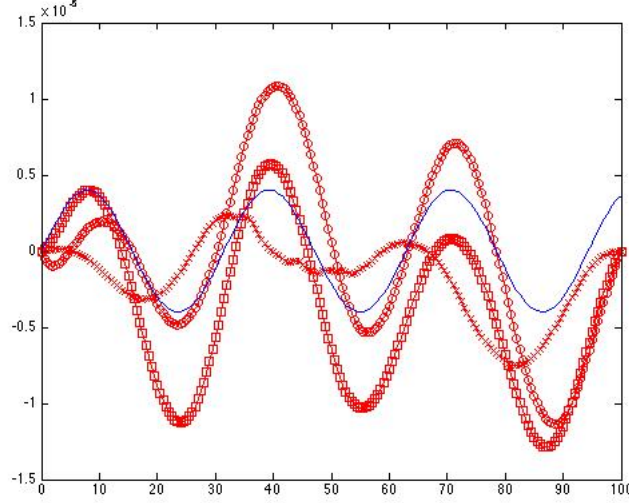


Fig. 1: Toy Model 2, $(n, \sigma) = (200, 3)$: Mean correction \bar{u} (red curve); rescaled true perturbation $10^{-5} \times v$ (blue curve)

and $r_\theta = \left(\frac{F_i^{CO} z_i^{CO}}{\beta C^{CO} / W + C^{CO} \tau}, 0, 0, 0 \right)^\top$. The state X is defined as $X^\top = (X^{CO}, X^{H_2O}, X^{CO_2}, X^{O_s})$, 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 . The total amount of oxygen X^{O_s} cannot be measured. Some of the parameter are already known:

- F_i^{CO} / F_0^{CO} : inlet/outlet flow rates of CO (0.59/0.45)
- z_i^{CO} : the constant fraction of ^{18}O present in oxygen atoms of the CO inlet flow rate (0.132)
- V/V' : rates of production (0.124/0.01)
- C^j : concentrations of gas phases in the reaction system ($j = CO, H_2O, CO_2$)
- W total weight of catalyst within system (0.744)
- β volume of dead space (206.1)

Our aim is to estimate the parameter $\theta = (C^{Col}, C^{Os}, v_5, v_6)$.

For simulating the datasets, we use two sample sizes $n = 100$ and $n = 50$ (observations are uniformly sampled the time interval $[0, 40]$), with 2 noise levels $\sigma = 0.002$ and $\sigma = 0.004$. The true parameter value is the estimate provided in [18], i.e $\theta^* = (0.1, 11.1, 0.35, 0.008)$ and with initial condition equals to $x_0^* = (0, 0, 0, 0)$. For the computation of the Kalman estimator, we select λ among 1, 5, 20, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000. Finally, the nonparametric estimate \hat{Y} is a regression splines, with knots selected manually (instead of GCV selection, because of overfitting): we use three equispaced knots at times $t = 0, 20, 40$.

(n, σ)		MSE	ARE	$E_P(X_{\hat{\theta}, \hat{x}_0})$	$E_P(X_{\hat{\theta}, \hat{x}_0, \hat{u}})$	$\Delta(X_{\hat{\theta}, \hat{x}_0}^{unobs})$	$\Delta(X_{\hat{\theta}, \hat{x}_0, \hat{u}}^{unobs})$
(100, 0.002)	$\hat{\theta}^{NLS}$	17.28	1.09	19.43	19.43	6.15	6.15
	$\hat{\theta}^K$	3.60	1.06	8.22	1.11	2.16	0.70
	$\hat{\theta}^{GS}$	21.54	1.14	19.45	19.45	6.41	6.41
(50, 0.002)	$\hat{\theta}^{NLS}$	57.23	2.19	31.49	31.49	12.62	12.62
	$\hat{\theta}^K$	21.38	2.05	9.26	3.49	2.58	1.38
	$\hat{\theta}^{GS}$	58.05	2.11	44.40	44.39	12.40	12.40
(100, 0.004)	$\hat{\theta}^{NLS}$	50.98	1.55	41.60	41.60	12.00	12.00
	$\hat{\theta}^K$	26.76	1.44	7.54	2.11	2.50	1.35
	$\hat{\theta}^{GS}$	55.61	1.59	33.66	33.66	15.96	15.96
(50, 0.004)	$\hat{\theta}^{NLS}$	80.03	2.25	43.13	43.13	14.75	14.75
	$\hat{\theta}^K$	35.87	2.16	28.69	3.06	7.79	1.57
	$\hat{\theta}^{GS}$	94.30	2.29	44.59	44.59	17.86	17.86

Tab. 3: Methanation Model

The results are presented in table 3, that gathers the statistics about the parameter estimation accuracy, and the prediction of the complete state, and in particular the estimation of the hidden variable X^{Os} . The Kalman estimator gives more accurate parameter estimates than Nonlinear Least Squares or Generalized Smoothing. The dramatic difference for the MSE comes from the estimation of C^{Os} that is of greater magnitude than the other parameters, thus ARE seems more relevant for comparisons. However, the MSE enlighten the difficulty for NLS and GS estimator to correctly estimate C^{Os} ; moreover, a great number of outliers for C^{Os} estimates have been removed for the NLS estimation before computing ARE and MSE. Additionally, state estimation improves

dramatically, as the prediction error E_P and missing state reconstruction Δ of the Kalman estimator outperforms the two others. This improvement is even more significative when the correction \bar{u} is used.

The difference can be partly explained by the nonlinearity in parameters that makes their estimation more difficult. We can have estimates that are far from the true parameter value, but in the case of the Kalman estimator, the important errors for the parameter are balanced by a more important correction term \bar{u} that ameliorates significantly state estimation and prediction.

7 Discussions

We have considered the statistical problem of parameter and state estimation of a linear Ordinary Differential Equations as an Optimal Control problem. By doing this, we follow the lines drawn in [33] or in the two-step approaches, that consist in defining a statistical criterion more adapted to ordinary differential equation than the likelihood. A new theory was needed in order to assess the statistical efficiency of this new estimator, that heavily relies on the Linear-Quadratic Theory. Indeed, the linear structure of the model gives a closed-form for the criterion S which permits to establish the needed regularity properties for statistical analysis. An important question is to determine the conditions under which we can apply the same methodology for nonlinear ODEs. It is probably more involved but the characterization used here is directly generalized by the Pontryagin's Maximum Principle, that gives also a tractable way to solve the optimal control problem.

An important feature of our approach is that we can cope with model misspecification, and the estimation process gives a way to evaluate the lack-of-fit thanks to the analysis of the control \bar{u} . Thanks to that, we are able to estimate properly the parameters, but also to do prediction and state estimation. Our experiments show that we can have better performance than the classical NLS and Generalized Smoothing and that it is beneficial to account for possible perturbation. A good choice for the trade-off hyperparameter λ is then necessary, and our selection methodology is satisfying in practice but needs more insight to explain its influence for the selection of good predictors, in particular for hidden states. The penalty term $\|u\|_{L^2}$ is an energy related to the degrees of freedom of the predictor $X_{\theta, x_0, u}$, but it is not related to the usual criterion of model complexity for smoothing.

In our analysis, we assume that the observability $O_\theta(T)$ is nonsingular, which

avoids the use of the quadratic form $x_0^\top Q x_0$ in the criterion. If this regularization term is used, the mechanics of the proof would be the same, but with $Q = Q_n$ that should tend to 0, as n tends to infinity. Nevertheless, it can have consequences on the asymptotics of the estimators, as it corresponds to cases where the loss of information is too big and needs additional information. Quite interestingly, our criterion about identifiability remains tractable, and can be relatively easy to check in practice.

Appendix: State and Parameter Inference for Partially Observed ODE

A Derivation of deterministic Kalman filter estimator using Linear-Quadratic Theory

In this section we describe more precisely how the deterministic Kalman Filter is constructed (see [35] for an introduction), it involves two steps:

1. For a given initial condition x_0 we determine the minimum cost expression thanks to theorem A.1 (subsection A.1).
2. inimal cost is a quadratic form w.r.t final condition and hence it exists an unique final condition (and hence a unique initial condition by ODE solution uniqueness) minimizing this minimal cost (subsection A.2).

A.1 x_0 fixed, minimal cost expression

To derive a closed form for the minimal cost for a given x_0 . For that we define the reverse time functions:

$$\begin{aligned}\tilde{X}_{\theta,x_0,u}(t) &= X_{\theta,x_0,u}(T-t), \quad \tilde{A}_\theta(t) = -A_\theta(T-t) \\ \tilde{r}_\theta(t) &= -r_\theta(T-t), \quad \tilde{B}(t) = -B(T-t), \quad \tilde{Y}(t) = \hat{Y}(T-t)\end{aligned}\tag{29}$$

And by denoting

$$\tilde{W}_1 = \begin{pmatrix} C^T C & -C^T \tilde{Y} \\ -\tilde{Y}^T C & \tilde{Y}^T \tilde{Y} \end{pmatrix}, \quad Q_1 = \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix}, \quad Z_{\theta,x_0,u} = \begin{pmatrix} \tilde{X}_{\theta,x_0,u} \\ 1 \end{pmatrix}\tag{30}$$

we can rewrite our cost under the form:

$$\begin{aligned}C(\hat{Y}; x_0, u, \theta, \lambda) &= \tilde{C}(\hat{Y}; Z_{\theta,x_0,u}(T), u, \theta, \lambda) \\ &:= Z_{\theta,x_0,u}(T)^T Q_1 Z_{\theta,x_0,u}(T) + \int_0^T \|Z_{\theta,x_0,u}(t)\|_{\tilde{W}_1}^2 dt + \lambda \int_0^T \|u(t)\|_2^2 dt\end{aligned}\tag{31}$$

The issue here is to minimize (31) in a non-finite dimensional space but thanks to results coming from Optimal control and Riccati theory we know that for a given θ and a given $Z(T)$ it exists a unique control \bar{u} such that

$$\tilde{C}(\hat{Y}; Z(T), \bar{u}, \theta, \lambda) = \min_{u \in L^2} \tilde{C}(\hat{Y}; Z(T), u, \theta, \lambda)$$

It is the main point of the following theorem for a given θ and $Z(T)$ it ensures the existence, the uniqueness of this control \bar{u} and gives a closed form for both \bar{u} and $\tilde{C}(\hat{Y}; Z(T), \bar{u}, \theta, \lambda)$.

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:*

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

and the cost:

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

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:

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

For a given t_0 we want to minimize the cost $C(t_0, u, U)$ on $L^2([0, T], \mathbb{R}^d)$.

We know it exists an unique control \bar{u} , called optimal control, associated to the trajectory $z_{\bar{u}}$, called optimal trajectory, minimizing this cost. Moreover \bar{u} is under the closed-feedback loop form $\bar{u}(t) = U^{-1}(t)E(t)B(t)z_{\bar{u}}(t)$ where E is the matricial solution of the ODE:

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

this ODE its called Ricatti equation associated to LQ problem composed of the cost $C(t_0, u, U)$ and the ODE $\dot{z}_u(t) = A(t)z_u(t) + B(t)u(t)$, $z(t_0) = z_0$. Moreover $E(t)$ is symetric and the minimal cost is equal to: $C(t_0, \bar{u}, U) = -z_0^T E(t_0) z_0$.

By identifying in the last theorem A with $\begin{pmatrix} \tilde{A}_\theta(t) & \tilde{r}_\theta(t) \\ 0 & 0 \end{pmatrix}$, Q with Q_1 , W with W_1 and U with λI_d we obtain the corresponding minimal cost reached for the optimal cost \bar{u} for a given initial condition x_0

$$\tilde{C}(\hat{Y}; Z_{\theta, x_0, \bar{u}}(T), \bar{u}, \theta, \lambda) = -\tilde{X}_{\theta, x_0, \bar{u}}(0)^T \tilde{E}_\theta(0) \tilde{X}_{\theta, x_0, \bar{u}}(0) - 2\tilde{X}_{\theta, x_0, \bar{u}}(0)^T \tilde{h}_\theta(0) - \tilde{\alpha}_\theta(0). \quad (32)$$

with the associated ODE:

$$\begin{aligned}\dot{\widetilde{E}}_\theta(t) &= C^T C - \widetilde{A}_\theta \widetilde{E}_\theta - \widetilde{E}_\theta \widetilde{A}_\theta - \frac{1}{\lambda} \widetilde{E}_\theta^2, \quad \widetilde{E}_\theta(T) = -Q \\ \dot{\widetilde{h}}_\theta(t) &= -C^T \widetilde{Y} - \widetilde{A}_\theta^T \widetilde{h}_\theta - \widetilde{E}_\theta \widetilde{r}_\theta - \frac{1}{\lambda} \widetilde{E}_\theta \widetilde{h}_\theta, \quad \widetilde{h}_\theta(T) = 0 \\ \dot{\widetilde{\alpha}}_\theta &= \widetilde{Y}^T \widetilde{Y} - 2\widetilde{r}_\theta^T \widetilde{h}_\theta - \frac{1}{\lambda} \widetilde{h}_\theta^T \widetilde{h}_\theta, \quad \widetilde{\alpha}_\theta(T) = 0\end{aligned}$$

To be able to apply Theorem A.1 we need W_1 to belong to $L^\infty([0, T], \mathbb{R}^{d \times d})$, that is why we require $\widehat{Y} \in L^\infty([0, T], \mathbb{R}^{d'})$.

A.2 Optimal x_0 selection

For a given x_0 we have obtained the minimal cost expression w.r.t control. How can we choose x_0 in order to minimize this minimal cost?

We recall that $\widetilde{X}_{\theta, x_0, \bar{u}}(0) = X_{\theta, x_0, \bar{u}}(T)$ so $\widetilde{C}(\widehat{Y}; Z_{\theta, x_0, \bar{u}}(T), \bar{u}, \theta, \lambda)$ defined by (32) is a quadratic form w.r.t the final condition $(\widetilde{\alpha}_\theta(0))$ do not depend on $\widetilde{X}_{\theta, x_0, \bar{u}}(0)$. Since it makes no difference to minimize $\widetilde{C}(\widehat{Y}; Z_{\theta, x_0, \bar{u}}(T), \bar{u}, \theta, \lambda)$ w.r.t the final condition instead of x_0 because of unicity of ODE solution we look for the final condition minimizing (32). Hence if $\widetilde{E}_\theta(0)$ is invertible the minimum is reached for

$$-\widetilde{E}_\theta(0)^{-1} \widetilde{h}_\theta(0) \tag{33}$$

we denote \widehat{x}_0 the unique initial condition such that $X_{\theta, \widehat{x}_0, \bar{u}}(T) = -\widetilde{E}_\theta(0)^{-1} \widetilde{h}_\theta(0)$.

In that case the minimal cost is equal to:

$$\widetilde{C}(\widehat{Y}; Z_{\theta, \widehat{x}_0, \bar{u}}(T), \bar{u}, \theta, \lambda) = \widetilde{h}_\theta(0)^T \widetilde{E}_\theta(0)^{-1} \widetilde{h}_\theta(0) - \widetilde{\alpha}_\theta(0)$$

and for a given parameter θ we have:

$$\begin{aligned}S(\widehat{Y}; \theta, \lambda) &= \min_{x_0 \in \mathbb{R}^d} \left(\min_{u \in L^2} C(\widehat{Y}; x_0, u, \theta, \lambda) \right) \\ &= \widetilde{h}_\theta(0)^T \widetilde{E}_\theta(0)^{-1} \widetilde{h}_\theta(0) - \widetilde{\alpha}_\theta(0) \\ &= \widetilde{h}_\theta(0)^T \widetilde{E}_\theta(0)^{-1} \widetilde{h}_\theta(0) + \int_0^T \left(\widetilde{Y}(t)^T \widetilde{Y}(t) - 2\widetilde{r}_\theta(t)^T \widetilde{h}_\theta(t) - \frac{1}{\lambda} \widetilde{h}_\theta(t)^T \widetilde{h}_\theta(t) \right) dt\end{aligned}$$

A.3 Minimal cost expression

By posing $E_\theta(t) = -\widetilde{E}_\theta(T-t)$, $\widehat{h}_\theta(t) = -\widetilde{h}_\theta(T-t)$ we define our estimator as:

$$\widehat{\theta} = \arg \min_{\theta \in \Theta} S(\widehat{Y}; \theta, \lambda)$$

with the functional criteria:

$$\begin{aligned} S(Y; \theta, \lambda) &= -h_\theta(T, Y)^T E_\theta(T)^{-1} h_\theta(T, Y) \\ &+ \int_0^T (Y(t)^T Y(t) - 2r_\theta(t)^T h_\theta(t, Y) - \frac{1}{\lambda} h_\theta(t, Y)^T h_\theta(t, Y)) dt \end{aligned} \quad (34)$$

the associated ODE:

$$\begin{aligned} \dot{E}_\theta(t) &= C^T C - A_\theta^T E_\theta - E_\theta A_\theta - \frac{1}{\lambda} E_\theta^2, \\ \dot{h}_\theta(t, Y) &= -\alpha_\theta(t) h_\theta(t, Y) - \beta_\theta(t, Y) \\ (E_\theta(0), h_\theta(0, Y)) &= (Q, 0) \end{aligned} \quad (35)$$

and the functions α and β defined by:

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

Hence we have obtained the expression for the optimal control, the minimal cost and the final state value presented in Theorem 2.2.

B State Estimation: Controls of the variations of the adjoint variables

Lemma B.1. *We have $\|E_\theta(t) - E_{\theta'}(t)\|_2 \leq K_1 \overline{E_\lambda} e^{L_1 \frac{\overline{E_\lambda}}{\lambda}} \|\theta - \theta'\|$ by denoting $\overline{E_\lambda} = \sup_{t, \theta \in [0, T] \times \Theta} \|E_\theta(t)\|_2$ and $\overline{E_\lambda} \leq K_2 e^{\frac{L_1}{\lambda}}$*

Proof. Thanks to condition 3 $\forall \theta \in \Theta$ $t \mapsto A_\theta(t)$ is continuous on $[0, T]$ and $\forall \theta \in \Theta$ $t \mapsto E_\theta(t)$ is defined on $[0, T]$ and obviously continuous on the same interval as an ODE solution.

$\forall (\theta, \theta') \in \Theta^2$ we have:

$$\begin{aligned} \dot{E}_\theta(t) - \dot{E}_{\theta'}(t) &= A_{\theta'}(t)^T E_{\theta'}(t) - A_\theta(t)^T E_\theta(t) \\ &+ E_{\theta'}(t) A_{\theta'}(t) - E_\theta(t) A_\theta(t) \\ &+ \frac{1}{\lambda} (E_{\theta'}^2(t) - E_\theta^2(t)) \end{aligned}$$

and by integrating between 0 and t , taking the norm gives us:

$$\begin{aligned} \|E_\theta(t) - E_{\theta'}(t)\|_2 &\leq \int_0^t \|A_{\theta'}(s)^T E_{\theta'}(s) - A_\theta(s)^T E_\theta(s)\|_2 ds \\ &+ \int_0^t \|E_{\theta'}(s) A_{\theta'}(s) - E_\theta(s) A_\theta(s)\|_2 ds \\ &+ \frac{1}{\lambda} \int_0^t \|E_{\theta'}^2(s) - E_\theta^2(s)\|_2 ds \end{aligned}$$

and:

$$\begin{aligned} \int_0^t \|E_\theta^2(s) - E_{\theta'}^2(s)\|_2 ds &\leq \sqrt{d} \int_0^t \|E_\theta(s)\|_2 \|E_\theta(s) - E_{\theta'}(s)\|_2 ds \\ &+ \sqrt{d} \int_0^t \|E_{\theta'}(s)\|_2 \|E_\theta(s) - E_{\theta'}(s)\|_2 ds \\ &\leq 2\sqrt{dE_\lambda} \int_0^t \|E_\theta(s) - E_{\theta'}(s)\|_2 ds \end{aligned}$$

by denoting $\overline{E_\lambda} = \sup_{t, \theta \in [0, T] \times \Theta} \|E_\theta(t)\|_2$.

Now we bound the remaining term:

$$\begin{aligned} \int_0^t \|A_\theta(s)^T E_\theta(s) - A_{\theta'}(s)^T E_{\theta'}(s)\|_2 ds &\leq \sqrt{d} \int_0^t \|A_\theta(s)\|_2 \|E_\theta(s) - E_{\theta'}(s)\|_2 ds \\ &+ \sqrt{d} \int_0^t \|E_{\theta'}(s)\|_2 \|A_\theta(s) - A_{\theta'}(s)\|_2 ds \\ &\leq \sqrt{d\overline{A}} \int_0^t \|E_\theta(s) - E_{\theta'}(s)\|_2 ds \\ &+ \sqrt{d\overline{E_\lambda}} \int_0^t \|A_\theta(s) - A_{\theta'}(s)\|_2 ds \end{aligned}$$

Using these bounds in the main inequality drive us to the following inequality:

$$\begin{aligned} \|E_\theta(t) - E_{\theta'}(t)\|_2 &\leq 2\sqrt{d}\left(\frac{\overline{E_\lambda}}{\lambda} + \overline{A}\right) \int_0^t \|E_\theta(s) - E_{\theta'}(s)\|_2 ds \\ &+ 2\sqrt{d\overline{E_\lambda}} \int_0^t \|A_\theta(s) - A_{\theta'}(s)\|_2 ds \end{aligned}$$

then Gronwall's lemma gives us

$$\|E_\theta(t) - E_{\theta'}(t)\|_2 \leq 2\sqrt{d\overline{E_\lambda}} \int_0^t \|A_\theta(s) - A_{\theta'}(s)\|_2 dt \cdot e^{\int_0^t 2\sqrt{d}\left(\frac{\overline{E_\lambda}}{\lambda} + \overline{A}\right) dt}$$

and we obtain thanks to Cauchy-Schwarz inequality:

$$\|E_\theta(t) - E_{\theta'}(t)\|_2 \leq 2\sqrt{dT\overline{E_\lambda}} \|A_\theta - A_{\theta'}\|_{L^2} e^{2\sqrt{d}\left(\frac{\overline{E_\lambda}}{\lambda} + \overline{A}\right)T}$$

which gives us the proper results using $\theta \mapsto A_\theta$ continuity.

The bound for $\overline{E_\lambda}$ is obtained by a direct application of Gronwall's lemma:

$$\begin{aligned} \|E_\theta(t)\|_2 &\leq |C|^2 + \int_0^t \|A_{\theta'}(s)^T E_{\theta'}(s) + A_\theta(s)^T E_\theta(s) + \frac{1}{\lambda} E_\theta(s)^T E_\theta(s)\|_2 ds \\ &\leq |C|^2 + 2\sqrt{d} \int_0^t \left(\|A_{\theta'}(s)\|_2 + \frac{\overline{E_\lambda}}{\lambda} \right) \|E_{\theta'}(s)\|_2 ds \\ &\leq |C|^2 e^{2\sqrt{d}\left(\overline{A} + \frac{\overline{E_\lambda}}{\lambda}\right)T} \end{aligned}$$

hence

$$\overline{E_\lambda} \leq |C|^2 e^{2\sqrt{d}AT} e^{2\sqrt{d}\frac{\overline{E_\lambda}}{\lambda}}$$

□

Lemma B.2. *Assuming condition C3 and C4 we know it exists constants K_i such that:*

$$\begin{aligned} \|h_\theta(t, Y) - h_{\theta'}(t, Y')\|_2 &\leq K_6 e^{L_1 \frac{\overline{E_\lambda}}{\lambda}} \|Y - Y'\|_{L^2} \\ &+ (K_7 + K_8 \overline{E_\lambda}) \left(K_4 + \frac{K_5}{\lambda} \overline{E_\lambda} e^{L_1 \frac{\overline{E_\lambda}}{\lambda}} \right) e^{2L_1 \frac{\overline{E_\lambda}}{\lambda}} \|\theta - \theta'\| \\ &+ \left(K_9 e^{2L_1 \frac{\overline{E_\lambda}}{\lambda}} + K_{10} e^{L_1 \frac{\overline{E_\lambda}}{\lambda}} \right) \overline{E_\lambda} \|\theta - \theta'\| \end{aligned}$$

and

$$\|h_\theta(t, Y)\|_2 \leq \sqrt{T} d^2 \|C\|_2 e^{\sqrt{d}(\overline{A} + \frac{\overline{E_\lambda}}{\lambda})T} \|Y\|_{L^2} + \sqrt{d \overline{E_\lambda} \overline{r_\theta}}$$

Proof. We know that $h_\theta(t, Y) = -\int_0^t R_\theta(t, s) C^T Y(s) ds - \int_0^t R_\theta(t, s) E_\theta(s) r_\theta(s) ds$ hence $\forall (Y, Y') \in L^\infty([0, T], \mathbb{R}^{d'})$ we have:

$$\begin{aligned} \|h_\theta(t, Y) - h_{\theta'}(t, Y')\|_2 &\leq \sqrt{d} \|C\|_2 \int_0^t \|R_\theta(t, s)\|_2 \|Y(s) - Y'(s)\|_2 ds \\ &+ \sqrt{d} \|C\|_2 \int_0^t \|R_\theta(t, s) - R_{\theta'}(t, s)\|_2 \|Y'(s)\|_2 ds \\ &+ d\overline{r} \int_0^t \|R_\theta(t, s) - R_{\theta'}(t, s)\|_2 \|E_\theta(s)\|_2 ds \\ &+ d\overline{r} \int_0^t \|R_{\theta'}(t, s)\|_2 \|E_\theta(s) - E_{\theta'}(s)\|_2 ds \\ &+ d\overline{r_\theta} \|\theta - \theta'\|_2 \int_0^t \|R_{\theta'}(t, s)\|_2 \|E_{\theta'}(s)\|_2 ds \end{aligned}$$

Cauchy Schwarz inequality gives us:

$$\begin{aligned} \|h_\theta(t, Y) - h_{\theta'}(t, Y')\|_2 &\leq \sqrt{d} \|C\|_2 \|R_\theta(\cdot, s)\|_{L^2} \|Y - Y'\|_{L^2} \\ &+ \left(\sqrt{d} \|C\|_2 \|Y'\|_{L^2} + d\sqrt{T\overline{r}\overline{E_\lambda}} \right) \|R_\theta(\cdot, s) - R_{\theta'}(\cdot, s)\|_{L^2} \\ &+ d\overline{r} \|R_\theta(\cdot, s)\|_{L^2} \|E_\theta - E_{\theta'}\|_{L^2} \\ &+ d\sqrt{T\overline{r_\theta}} \|\theta - \theta'\|_2 \|R_\theta(\cdot, s)\|_{L^2} \overline{E_\lambda} \end{aligned} \tag{36}$$

We straightforwardly bound $\|R_\theta(t, s)\|_2$ by application of Gronwall's lemma:

$$\begin{aligned} \|R_\theta(t, s)\|_2 &\leq \sqrt{d} + \sqrt{d} \left(\overline{A} + \frac{\overline{E_\lambda}}{\lambda} \right) \int_s^t \|R_\theta(u, s)\|_2 du \\ &\leq \sqrt{d} e^{\sqrt{d}(\overline{A} + \frac{\overline{E_\lambda}}{\lambda})T} := K_3 e^{L_1 \frac{\overline{E_\lambda}}{\lambda}} \end{aligned}$$

Using successively norm inequalities and Gronwall's lemma we obtain:

$$\begin{aligned}
\|R_{\theta'}(t, s) - R_{\theta}(t, s)\|_2 &\leq \sqrt{d} \|\alpha_{\theta}(t) - \alpha_{\theta'}(t)\|_2 \|R_{\theta}(t, s)\|_2 \\
&+ \sqrt{d} \|R_{\theta}(t, s) - R_{\theta'}(t, s)\|_2 \|\alpha_{\theta'}(t)\|_2 \\
&\leq \sqrt{d} \left(\overline{A_{\theta}} \|\theta - \theta'\|_2 + \frac{1}{\lambda} \|E_{\theta}(t) - E_{\theta'}(t)\|_2 \right) \|R_{\theta}(t, s)\|_2 \\
&+ \sqrt{d} \|R_{\theta}(t, s) - R_{\theta'}(t, s)\|_2 \left(\overline{A} + \frac{\overline{E_{\lambda}}}{\lambda} \right) \\
&\leq \sqrt{d} \left(\overline{A_{\theta}} + \frac{1}{\lambda} K_1 \overline{E_{\lambda}} e^{L_1 \frac{\overline{E_{\lambda}}}{\lambda}} \right) K_3 e^{L_1 \frac{\overline{E_{\lambda}}}{\lambda}} \|\theta - \theta'\| \\
&+ \sqrt{d} \left(\overline{A} + \frac{\overline{E_{\lambda}}}{\lambda} \right) \|R_{\theta}(t, s) - R_{\theta'}(t, s)\|_2 \\
&\leq \sqrt{d} e^{\sqrt{d} A T} K_3 \left(\overline{A_{\theta}} + \frac{1}{\lambda} K_1 \overline{E_{\lambda}} e^{L_1 \frac{\overline{E_{\lambda}}}{\lambda}} \right) e^{2L_1 \frac{\overline{E_{\lambda}}}{\lambda}} \|\theta - \theta'\| \\
&:= \left(K_4 + \frac{K_5}{\lambda} \overline{E_{\lambda}} e^{L_1 \frac{\overline{E_{\lambda}}}{\lambda}} \right) e^{2L_1 \frac{\overline{E_{\lambda}}}{\lambda}} \|\theta - \theta'\|
\end{aligned}$$

and applying this bound in 36 gives the following inequality:

$$\begin{aligned}
\|h_{\theta}(t, Y) - h_{\theta'}(t, Y')\|_2 &\leq \sqrt{d} \|C\|_2 K_3 e^{L_1 \frac{\overline{E_{\lambda}}}{\lambda}} \|Y - Y'\|_{L^2} \\
&+ \left(\sqrt{d} \|C\|_2 \|Y'\|_{L^2} + d\sqrt{T\overline{r}E_{\lambda}} \right) \left(K_4 + \frac{K_5}{\lambda} \overline{E_{\lambda}} e^{L_1 \frac{\overline{E_{\lambda}}}{\lambda}} \right) e^{2L_1 \frac{\overline{E_{\lambda}}}{\lambda}} \|\theta - \theta'\| \\
&+ d\overline{r} K_3 e^{L_1 \frac{\overline{E_{\lambda}}}{\lambda}} K_1 \overline{E_{\lambda}} e^{L_1 \frac{\overline{E_{\lambda}}}{\lambda}} \|\theta - \theta'\| \\
&+ d\sqrt{T\overline{r}\overline{\theta}} K_3 e^{L_1 \frac{\overline{E_{\lambda}}}{\lambda}} \overline{E_{\lambda}} \|\theta - \theta'\|_2 \\
&\leq K_6 e^{L_1 \frac{\overline{E_{\lambda}}}{\lambda}} \|Y - Y'\|_{L^2} \\
&+ (K_7 + K_8 \overline{E_{\lambda}}) \left(K_4 + \frac{K_5}{\lambda} \overline{E_{\lambda}} e^{L_1 \frac{\overline{E_{\lambda}}}{\lambda}} \right) e^{2L_1 \frac{\overline{E_{\lambda}}}{\lambda}} \|\theta - \theta'\| \\
&+ \left(K_9 e^{2L_1 \frac{\overline{E_{\lambda}}}{\lambda}} + K_{10} e^{L_1 \frac{\overline{E_{\lambda}}}{\lambda}} \right) \overline{E_{\lambda}} \|\theta - \theta'\|
\end{aligned}$$

By a similar computation we obtain:

$$\|h_{\theta}(t, Y)\|_2 \leq \sqrt{T} d^2 \|C\|_2 e^{\sqrt{d}(\overline{A} + \frac{\overline{E_{\lambda}}}{\lambda})T} \|Y\|_{L^2} + \sqrt{d\overline{E_{\lambda}}\overline{r}\overline{\theta}}$$

□

Lemma B.3. *Assuming condition C3 and C4 we know it exists constants K_i such that: $\|E_{\theta}^{-1}(T) - E_{\theta'}^{-1}(T)\|_2 \leq (\frac{K_{12}}{\lambda} + K_{11}) e^{K_{13} + \frac{K_{14}}{\lambda}} \|\theta - \theta'\|$ and $\|E_{\theta}^{-1}(T)\|_2 \leq \frac{K_{15}}{\lambda}$*

Proof. We have already shown $\forall \theta \in \Theta \ t \mapsto E_{\theta}(t)$ is defined on $[0, T]$ and obviously continuous on the same interval as an ODE solution.

When $E_{\theta}^{-1}(t)$ is defined we know it follows the ODE:

$$\begin{aligned}
\frac{d}{dt} (E_\theta^{-1}(t)) &= -E_\theta^{-1}(t) \dot{E}_\theta(t) E_\theta^{-1}(t) \\
&= -E_\theta^{-1}(t) (C^T C - A_\theta(t)^T E_\theta(t) - E_\theta(t) A_\theta(t) - \frac{1}{\lambda} E_\theta(t)^T E_\theta(t)) E_\theta^{-1}(t) \\
&= \frac{1}{\lambda} + E_\theta^{-1}(t) A_\theta^T(t) + A_\theta(t) E_\theta^{-1}(t) - E_\theta^{-1}(t) C^T C E_\theta^{-1}(t)
\end{aligned}$$

By hypothesis $\forall \theta \in \Theta$ $E_\theta^{-1}(T)$ is defined and by continuity of $(\theta, t) \mapsto \text{Det}(E_\theta(t))$ using chain rule we know for each θ it exists an interval $[T - \varepsilon, T]$ and a open ball $B_\varrho(\theta)$ where $(t, \theta) \mapsto E_\theta(t)$ is non-singular. Because of $\theta \mapsto E_\theta(T - \varepsilon)$ continuity and differentiability it exists a constant K_{11} such that: $\|E_\theta^{-1}(T - \varepsilon) - E_{\theta'}^{-1}(T - \varepsilon)\|_2 \leq K_{11} \|\theta - \theta'\|$.

By defining $\overline{E_{\varepsilon, \varrho, \theta, \lambda}^{-1}} = \sup_{[T - \varepsilon, T] \times B_\varrho(\theta)} \|E_\theta^{-1}(t)\|_2$ and using successively norm inequalities and Gronwall's lemma we obtain:

$$\begin{aligned}
\|E_\theta^{-1}(t) - E_{\theta'}^{-1}(t)\|_2 &\leq \int_{T - \varepsilon}^t \left\| E_\theta^{-1}(s) A_\theta^T(s) + A_\theta(s) E_\theta^{-1}(s) - E_{\theta'}^{-1}(s) A_{\theta'}^T(s) - A_{\theta'}(s) E_{\theta'}^{-1}(s) \right\|_2 ds \\
&+ \int_{T - \varepsilon}^t \left\| E_{\theta'}^{-1}(s) C^T C E_{\theta'}^{-1}(s) - E_\theta^{-1}(s) C^T C E_\theta^{-1}(s) \right\|_2 ds \\
&+ K_{11} \|\theta - \theta'\| \\
&\leq 2\sqrt{dA} \int_{T - \varepsilon}^t \|E_\theta^{-1}(s) - E_{\theta'}^{-1}(s)\|_2 ds \\
&+ 2\sqrt{dE_{\theta, \lambda}^{-1}} \int_{T - \varepsilon}^t \|A_\theta(s) - A_{\theta'}(s)\|_2 ds \\
&+ 2d^{\frac{3}{2}} \overline{E_{\theta, \lambda}^{-1}} \|C\|_2^2 \int_{T - \varepsilon}^t \|E_\theta^{-1}(s) - E_{\theta'}^{-1}(s)\|_2 ds \\
&+ K_{11} \|\theta - \theta'\| \\
&\leq \left(2\sqrt{dA} + 2d^{\frac{3}{2}} \overline{E_{\varepsilon, \varrho, \theta, \lambda}^{-1}} \|C\|_2^2 \right) \int_{T - \varepsilon}^t \|E_\theta^{-1}(s) - E_{\theta'}^{-1}(s)\|_2 ds \\
&+ \left(2\sqrt{dT} \overline{E_{\varepsilon, \varrho, \theta, \lambda}^{-1}} \overline{A_\theta} + K_{11} \right) \|\theta - \theta'\| \\
&\leq \left(2\sqrt{dT} \overline{E_{\varepsilon, \varrho, \theta, \lambda}^{-1}} \overline{A_\theta} + K_{11} \right) e^{\left(2\sqrt{dA} + 2d^{\frac{3}{2}} \overline{E_{\varepsilon, \varrho, \theta, \lambda}^{-1}} \|C\|_2^2 \right) T} \|\theta - \theta'\|
\end{aligned}$$

For $\|E_\theta^{-1}(T)\|_2$ we can obtain a uniform bound w.r.t θ using Gronwall's lemma:

$$\begin{aligned}
\|E_\theta^{-1}(t)\|_2 &\leq \int_{T - \varepsilon}^t \left\| \frac{1}{\lambda} + E_\theta^{-1}(s) A_\theta^T(s) + A_\theta(s) E_\theta^{-1}(s) - E_\theta^{-1}(t) C^T C E_\theta^{-1}(t) \right\|_2 ds \\
&\leq \frac{1}{\lambda} + \int_{T - \varepsilon}^t \sqrt{d} \left(2\overline{A} + d \overline{E_{\varepsilon, \varrho, \theta, \lambda}^{-1}} \|C\|_2^2 \right) \|E_\theta^{-1}(t)\|_2^2 ds \\
&\leq \frac{1}{\lambda} e^{\sqrt{d} \left(2\overline{A} + d \overline{E_{\varepsilon, \varrho, \theta, \lambda}^{-1}} \|C\|_2^2 \right) T} \\
&:= \frac{K_{15}}{\lambda}
\end{aligned}$$

by using this upper bound in the previous inequality we obtain the desired result. \square

C Gradient Computation

For optimization purpose we need to compute the gradient of $S(\widehat{Y}; \theta, \lambda)$.

C.1 Notation in row vector for the adjoint ODE vector field

We will define the solution of the adjoint ODE in row formulation, we introduce

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

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

$$\begin{aligned} \dot{Q}_\theta &= F(Q_\theta, \theta, t) \\ Q_\theta(0) &= \begin{pmatrix} 0_{1,d} & Q^r \end{pmatrix}^T \end{aligned}$$

by introducing $Q^r := (Q_1^T, \dots, Q_d^T)^T$ the row formulation of Q and 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:

$$\begin{aligned} G(Q_\theta, \theta, t) &:= -\left(A_\theta^T + \frac{E_\theta}{\lambda}\right) \widehat{h}_\theta - (C^T Y + E_\theta r_\theta) \\ H_{(j-1)d+i}(Q_\theta, \theta) &:= \delta_{i,j} - (A_{\theta,i}^T E_j + A_{\theta,j}^T E_{\theta,i} + \frac{1}{\lambda} E_{\theta,i}^T E_{\theta,j}) \end{aligned}$$

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

For the next subsections we will drop dependence in θ for A_θ , r_θ , E_θ , \widehat{h}_θ

C.2 Gradient computation by sensitivity equation

Straightforward computation gives us :

$$\begin{aligned} \nabla_\theta S(\widehat{Y}; \theta, \lambda) &= -2 \int_0^T \left(\frac{\partial r(t)}{\partial \theta}^T \widehat{h}(t) + \frac{\partial \widehat{h}(t)}{\partial \theta}^T r(t) + \frac{1}{\lambda} \frac{\partial \widehat{h}(t)}{\partial \theta}^T \widehat{h}(t) \right) dt \\ &\quad - \frac{\partial \widehat{h}(T)}{\partial \theta}^T E(T)^{-1} \widehat{h}(T) \\ &\quad - \left(\widehat{h}(T)^T \frac{\partial (E(T)_i^{-1})}{\partial \theta} + (E(T)_i^{-1})^T \frac{\partial \widehat{h}(T)}{\partial \theta} \right)_{1 \leq i \leq d}^T \widehat{h}(T) \end{aligned}$$

with:

$$\frac{\partial (E(T)^{-1})}{\partial \theta_j} = -E(T)^{-1} \left(\frac{\partial E(T)}{\partial \theta_j} \right) E(T)^{-1}$$

thus we need to compute $\frac{\partial Q_\theta(t)}{\partial \theta}$ solution of the sensitivity equation:

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

and we know that $R_\theta(0) = (0, Q^r)$ so $\frac{\partial Q_\theta(0)}{\partial \theta} = 0$, hence we can obtain $\frac{\partial Q_\theta(t)}{\partial \theta}$ by solving the Cauchy problem:

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

In order to compute sensitivity equation we need to compute $\frac{\partial F}{\partial Q}(Q_\theta, \theta, t)$ and $\frac{\partial F}{\partial \theta}(Q_\theta, \theta, t)$, for $\frac{\partial F}{\partial Q}(Q_\theta, \theta, t)$ and $\frac{\partial F}{\partial \theta}(Q_\theta, \theta, t)$ we obtain:

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

with:

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

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

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

because:

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

- $\frac{\partial}{\partial \theta} (A_j^t E_i + A_i^t E_j) = E_i^t \frac{\partial A_j}{\partial \theta} + E_j^t \frac{\partial A_i}{\partial \theta}$ where $\frac{\partial A_i}{\partial \theta} = \left(\frac{\partial A_i}{\partial \theta_1} \dots \frac{\partial A_i}{\partial \theta_p} \right)$ a $d \times p$ matrix

References

- [1] M. Bardi and I. Capuzzo-Dolcetta. *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*. Modern Birkhauser Classic. Birkhauser, 2008.
- [2] R. Bellman and K.J. Astrom. On structural identifiability. *Mathematical Biosciences*, 7:329–339, 1970.
- [3] A. Bensoussan. *Stochastic Control of Partially Observable Systems*. Cambridge University Press, 2004.
- [4] P.J. Bickel and Y. Ritov. Nonparametric estimators which can be plugged-in. *Annals of Statistics*, 31(4):4, 2003.
- [5] N. J-B. Brunel. Parameter estimation of ode's via nonparametric estimators. *Electronic Journal of Statistics*, 2:1242–1267, 2008.
- [6] N. J-B. Brunel and Q. Clairon F. D'Alche-Buc. Parameter estimation of ordinary differential equations with orthogonality conditions. *JASA*, 109:173–185, 2014.
- [7] Nicolas J-B. Brunel and Quentin Clairon. A tracking approach to parameter estimation in linear ordinary differential equations. Technical report, 2014. submitted.
- [8] M. Enqvist C. Lyzell, T. Glad and L. Ljung. Difference algebra and system identification. *Automatica*, 47:1896–1904, 2011.
- [9] B. Calderhead, M. Girolami, and N.D. Lawrence. Accelerating bayesian inference over nonlinear differential equations with gaussian processes. In *Advances in Neural Information Processing Systems 21 - Proceedings of the 2008 Conference*, 2009.
- [10] D.A. Campbell and O. Chkrebtii. Maximum profile likelihood estimation of differential equation parameters through model based smoothing state estimates. *Mathematical Biosciences*, 2013.
- [11] I. Chernovena, B. Freydin, B. Hipszner, and Apanasovich T. V. Estimation of nonlinear differential equation for glucose-insulin dynamics in type i diabetic patients using generalized smoothing. *Annals of Applied Statistics*, 8(2):886–904, 2014.
- [12] 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.
- [13] S. P. Ellner and J. Guckenheimer. *Dynamic Models in Biology, Number vol. 13 in Princeton Paperbacks*. Princeton, NJ: Princeton University Press, 2006.
- [14] Hein W Engl, Christoph Flamm, Philipp Kügler, James Lu, Stefan Müller, and Peter Schuster. Inverse problems in systems biology. *Inverse Problems*, 25(12), 2009.

- [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] 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.
- [18] J. Happel, I. Suzuki, P. Kokayeff, and V. Fthenakis. Multiple isotope tracing of methanation over nickel catalyst. *Journal of Catalysis*, 65:59–77, 1980.
- [19] G. Hooker. Forcing function diagnostics for nonlinear dynamics. *Biometrics*, 65:928–936, 2009.
- [20] G. Hooker and S. P. Ellner. Goodness of fit in nonlinear dynamics: Mis-specified rates or mis-specified states? arxiv:1312.0294., arXiv preprint, 2013.
- [21] Y. Huang and H. Wu. A bayesian approach for estimating antiviral efficacy in hiv dynamic models. *Journal of Applied Statistics*, 33:155–174, 2006.
- [22] E. Hubert. Essential components of an algebraic differential equation. *J. Symbolic Computation*, 28:657–680, 1999.
- [23] C. Noiret L. Denis-Vidal, G. Joly-Blanchart. Some effective approaches to check the identifiability of uncontrolled nonlinear systems. *Mathematics and Computers in Simulation*, 57:35–44, 2000.
- [24] Z. Li, M.R. Osborne, and T. Prvan. Parameter estimation of ordinary differential equations. *IMA Journal of Numerical Analysis*, 25:264–285, 2005.
- [25] 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.
- [26] Costello J. C.-Küffner R. Vega N. M. Prill R. J. Camacho D. M. ... & DREAM5 Consortium. Marbach, D. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804., 2012.
- [27] 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.
- [28] W. K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79:147–168, 1997.
- [29] M.A. Nowak and R.M. May. *Virus Dynamics: Mathematical Principles of Immunology and Virology*. Oxford University Press, 2000.
- [30] H. Pohjanpalo. System identifiability based on the power series expansion of the solution. *Mathematical Biosciences*, 41:21–33, 1978.
- [31] L. Pronzato. Optimal experimental design and some related control problems. *Automatica*, 44:303–325, 2008.
- [32] Xin Qi and Hongyu Zhao. Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations. *The Annals of Statistics*, 1:435–481, 2010.

- [33] 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.
- [34] D. Ruppert, M.P. Wand, and R.J. Carroll. *Semiparametric regression*. Cambridge series on statistical and probabilistic mathematics. Cambridge University Press, 2003.
- [35] E. Sontag. *Mathematical Control Theory: Deterministic finite-dimensional systems*. Springer-Verlag (New-York), 1998.
- [36] A.M. Stuart. Inverse problems: A bayesian perspective. *Acta Numerica*, pages 451–559, 2010.
- [37] A.W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilities Mathematics. Cambridge University Press, 1998.
- [38] J. M. Varah. A spline least squares method for numerical parameter estimation in differential equations. *SIAM J.sci. Stat. Comput.*, 3(1):28–46, 1982.
- [39] Eric Walter and Luc Pronzato. *Identification of parametric models*. Communications and Control Engineering. Springer Verlag New-York, 1997.
- [40] H. Wu, T. Lu, H. Xue, and H. Liang. Sparse additive odes for dynamic gene regulatory network modeling. *Journal of the American Statistical Association*, 109(506):700–716, 2014.