Parameter estimation in Sloppy Differential Equations with Optimal Control

Nicolas J-B. Brunel & Quentin Clairon
ENSIIE & Laboratoire de Mathématiques et Modélisation d'Evry,
UMR CNRS 8071, Universite d'Evry, France

April 15, 2016
GT Contrôle - UPMC

Outline

1. ODE MODELS AND STATISTICAL INVERSE PROBLEMS

ODE Models

- Ordinary Differential Equations are derived commonly in applications (engineering/biology/...)

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

  Mechanistic Model derived from precise description of underlying phenomenon

- Some recurrent features:
  - f nonlinear in $(\mathbf{x}, \theta)$
  - θ often high-dimensional
  - x_0 is a "nuisance parameter"

ODE: sharp and uncertain models

- Several way to derive a wrong model for real-life
  1. Apply physical / chemical rules for describing mechanisms of inner interactions (e.g. $R_k = \theta_{ij}^k x_i^{\alpha_{ijk}} x_j^{\beta_{ijk}}, \ldots ?$)

$$F(t, \mathbf{x}, \theta) = \sum_k R_k(t, \mathbf{x}, \theta) + \ldots$$
$$= f(t, \mathbf{x}, \theta) + \varepsilon g(t, \mathbf{x})$$

  2. Interactions with "outside" that acts as a forcing function u (deterministic or random)

$$F(t, \mathbf{x}, \theta) = f(t, \mathbf{x}, \theta) + u(t)$$

  3. Approximation of a stochastic processes (e.g continuous time Markov processes)

$$\text{"}F(t, \mathbf{x}, \theta)dt = f(t, \mathbf{x}, \theta)dt + \sigma(t, \mathbf{x})dW_t\text{"}$$

  4. Preliminary expert and simulation-based calibration

$$F(t, \mathbf{x}, \theta) = f(t, \mathbf{x}, \theta^{Prior}, \theta)$$

Parameter identification from noisy data

- f smooth vector field defined on $t \in [0, T]$, $\mathbf{x} \in \Omega \subset \mathbb{R}^d$, $\theta \in \Theta \subset \mathbb{R}^p$, with unique solution $t \mapsto X_{\theta,x_0}(t)$, C^1 w.r.t θ, x_0.
- Assume noisy and discrete observations of the trajectories

$$y_i = h(X_{\theta,x_0}(t_i)) + \varepsilon_i$$

i.e we have a deterministic "state space model":

$$\begin{cases} \dot{x}(t) & = f(t, x(t), \theta) \\ y(t_i) & = h(x(t_i)) + \varepsilon_{t_i} \end{cases}$$

- Nonlinear Least Squares (regression) is the classical solution

$$\left(\hat{x}_0, \hat{\theta}\right) \in \arg\min_{x_0,\theta} \sum_{i=1}^{n} \left| y_i - h(X_{\theta,x_0}(t_i)) \right|^2$$

- Important limitations:
  - Computational load: nonlinear optimization, multiple ODE integration, ODE stability
  - NLS criterion very wriggly and flat regions (identifiability problems)

FitzHugh-Nagumo model: simple and ill-posed model

- Classical Benchmark borrowed from Ramsay et al., JRSS(B) 2007:

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

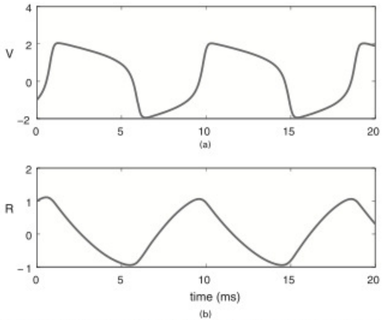
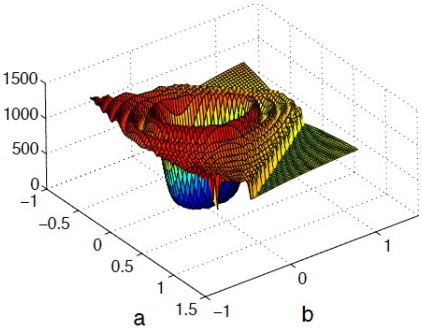


Fig. 1. Limiting behaviour of (a) voltage V and (b) recovery R variables defined by the FitzHugh–Nagumo equations (2) with parameter values $a=0.2$, $b=0.2$ and $c=3.0$ and initial conditions $(V_0, R_0)=(-1, 1)$

NLS asymptotic criterion for FitzHugh-Nagumo



$$RSS(\theta) = \sum_{i=1}^{n} \left| X_{\theta, x_0^*}(t_i) - X_{\theta,^* x_0^*}(t_i) \right|^2$$

Initial conditions are known $(V_0, R_0) = (-1, 1)$ and $c^* = 3$ is fixed. True values : $(a^*, b^*) = (0.2, 0, 2)$.

Sloppy Models and propagation of uncertainty

- Gauss-Newton (or Levenberg-Marquardt) solves sequentially the quadratic program: $\theta^{(k+1)} = \theta^{(k)} + \delta$ with

$$\begin{cases} \min_{\delta \in \mathbb{R}^p} \sum_{i=1}^{n} \left\| \left(y_i - X_{\theta^{(k)},x_0}(t_i) \right) - D_\theta X_{\theta^{(k)},x_0}(t_i)\delta \right\|^2 \\ s.t. \ \|\delta\|^2 \leq \gamma_k \end{cases} \tag{1}$$

- Fisher Information Matrix (Gaussian case)

$$\mathscr{I}(\theta) \propto \mathscr{S}(\theta) = \sum_{i=1}^{n} D_\theta X_{\theta,x_0}(t_i)^\top D_{\theta,x_0} X_\theta(t_i)$$

  where $D_\theta X_{\theta,x_0}(t)$ is the sensitivity matrix.

- Sloppy model (Sethna, 2007) means that $D_\theta X_{\theta,x_0}$ is badly conditionned: $\mathscr{S}(\theta)$ s.p.d. with $\lambda_1 > \cdots > \lambda_p$ and $\frac{\lambda_1}{\lambda_p} \geq 10^4$

Sloppy Models can be disappointing and misleading as model and data errors can produce spurious estimators and predictions.

2. REGULARIZATION AND APPROXIMATE SOLUTIONS OF ODEs

1. GENERALIZED PROFILING

Problem to solve

- We assume $y_i = X^*(t_i) + \varepsilon_i$, and $\dot{X}^* = f(t, X^*(t), \theta^*) + u^*(t)$.
  - $u^*(t)$ some regular function (forcing function/input function)
  - $u^*(t) = \varepsilon g(X^*(t))$
  - $u^*(t) = f(t, X^*(t), \theta^{**}) - f(t, X^*(t), \theta^*)$

The big aim: Can we perform correct estimation of θ^* with the wrong (but useful) model $\{f(x,\theta), \theta\}$?
This work is an intermediate step: robustify and show it is ok for $u^* = 0$....

Nonparametric estimation and model uncertainty

- Build a function $\hat{X}$ close to the data $(t_i, y_i), i = 1, \ldots, n$ that respects approximately the ODE
- Easy to do with smoothing splines that solves

$$\hat{X} = \arg\min_X \sum_{i=1}^{n} |y_i - X(t_i)|^2 + \lambda \int_0^T \|D^2 X\|_2^2 \, dt$$

where D is the differential operator.

- General P-splines smoothing defined with a linear differential operator $P_\theta = \sum_{j=0}^{J} \theta_j D^j$ (Ramsay and Silverman, 2005) by solving

$$\hat{X}_\theta = \arg\min_X \sum_{i=1}^{n} |y_i - X(t_i)|^2 + \lambda \int_0^T \|P_\theta X\|_2^2 \, dt$$

Nonlinear Smoothing and Generalized Profiling (Ramsay, 2007)

- $\hat{X}_\theta$ has a series expansion (in a B-splines basis), that depends on the coefficients θ: $\hat{X}_\theta(t) = \sum_k c_k(\theta) B_k(t)$.
- In the linear case: for unknown ODE model, it is possible to solve

$$\min_{\mathbf{c},\theta} \sum_{i=1}^{n} |y_i - B(t_i)\mathbf{c}|^2 + \lambda \int_0^T \|P_\theta B(t)\mathbf{c}\|_2^2 \, dt$$

- In general: we can look for a proxy $\hat{X}(t,\theta,x_0) = \sum_{k=1}^{K} c_k B_k(t)$ and replace linear operator by nonlinear operator

$$\min_{\mathbf{c},\theta} \sum_{i=1}^{n} \left| y_i - \hat{X}(t_i) \right|^2 + \lambda \int_0^T \left\| \dot{\hat{X}} - f\left(t, \hat{X}, \theta\right) \right\|_2^2 dt$$

Generalized Profiling

- The global optimization in $(\mathbf{c}, \theta)$ is hard to do and turned into *Parameter Cascading*: For each θ, profile on $\mathbf{c}$ by solving

$$\mathbf{c}_\lambda(\theta) \hat{=} \arg\min_{\mathbf{c}} \sum_{i=1}^{n} |y_i - B(t_i)\mathbf{c}|^2 + \lambda \int_0^T \left\| \dot{B}(t)\mathbf{c} - f(t, B(t)\mathbf{c}, \theta) \right\|_2^2 dt$$

- The approximate solution is $\hat{X}_\lambda(t,\theta) = B(t)\mathbf{c}_\lambda(\theta)$ and we solve

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

- Parameter Cascading: Iterate until convergence
  1. Optimize on $\mathbf{c}_\lambda(\theta)$ (inner optimization)
  2. Optimize on $\hat{\theta}_\lambda$ (middle optimization)
  3. Optimize on λ (outer optimization)

Generalized Profiling and Forcing function

- Function $\hat{X}_\lambda(t,\theta)$ is not solution of ODE: $\hat{u}(t) = \frac{d}{dt}\hat{X}_\lambda(t,\theta) - f(t, \hat{X}_\lambda(t,\theta), \theta) \neq 0$ and $\hat{u}$ is a "forcing function"; represents approximation or model error (used by Hooker et al, 2009, 2015 for lack-of-fit testing)
- Qi et al (2010) showed that under regularity assumptions +
  - $K = K_n \to \infty$ for error approximation of B-splines expansion
  - $\lambda = \lambda_n \to \infty$ but carefully selected for avoiding overfitting by $\hat{X}_\lambda(t,\theta)$ (discrepancy with solutions X_{θ,x_0})

  then Generalized Profiling is consistent and asymptotically efficient; but in practice choice of λ can be important, and influence of K_n hard to assess.

2. REGULARIZATION AND PERTURBED MODELS

Another look at Generalized Profiling

- GP: $\hat{X}(t,\theta,x_0)$ is solution of the approximate model

$$\dot{x} = f(t,x,\theta) + \hat{u}$$

  with $\hat{u}$ the forcing function.
- We introduce the solution $X_{\theta,x_0,u}$ of the perturbed model

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

- Any nonparametric estimate $\hat{X}$ obtained from $(y_1,\ldots,y_n)$ is close to X^* and we have $\hat{X} = X_{\theta,x_0,u_\theta}$ with

$$u_\theta(t) = \frac{d}{dt}\hat{X}(t) - f(t,\hat{X}(t),\theta)$$

Perturbed Model and Control Theory

- Any reasonable nonparametric estimate $\hat{X}$ should be close to the true solution X^* and we minimize $\left\|\hat{X} - X_{\theta,x_0}\right\|_{L^2}^2$.
  In order to take into account model uncertainty and sloppiness, we consider the minimization problem

$$\begin{cases} \min_{\theta,u,x_0} \left\|\hat{X} - X_{\theta,x_0,u}\right\|_{L^2}^2 \\ s.t \ \|u\|_{L^2}^2 \leq c \end{cases}$$

- For the estimation of θ, we introduce the criterion

$$C_\lambda(\hat{X};\theta,u,x_0) = \int_0^T \left\|\hat{X}(t) - X_{\theta,x_0,u}(t)\right\|^2 dt + \lambda \int_0^T \|u(t)\|^2 dt$$

  but we profile on the nuisance parameter u

$$\mathscr{S}_\lambda(\hat{X};\theta,x_0) = \min_{u\in L^2} C_\lambda(\hat{X};\theta,u,x_0)$$

- Tracking estimator is defined as

$$\hat{\theta}_\lambda^T = \arg\min_\theta \mathscr{S}_\lambda\left(\hat{X};\theta,x_0\right)$$

State Estimation and Partially Observed Systems

- When partial observations are available $y_i = CX^*(t_i) + \varepsilon_i$, we smooth the observed variable $\hat{Y}(t) \approx Y^*(t)$ (with splines, local polynomial...)
- The criterion is then replaced by

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

and $\mathscr{S}_\lambda(\hat{Y};\theta,x_0) = \min_{u\in L^2} C_\lambda(\hat{Y};\theta,u,x_0)$.

- Initial state unknown: profile again

$$\mathscr{S}_\lambda(\hat{Y};\theta) = \min_{x_0} \min_{u\in L^2} C_\lambda(\hat{Y};\theta,u,x_0).$$

3. LINEAR CASE AND TRACKING ESTIMATOR

Linear Case: known IC, all states observed

- Observations are $y_i = X^*(t_i) + \varepsilon_i$ and X^* is solution of

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

on $[0, T]$, with $X(0) = x_0^*$ known.

- If $\hat{X}$ is nonparametric proxy for X^* then we have an explicit expression for the profiled cost:

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

where $t \mapsto h(t)$ is defined through the Final Value Problem (adjoint problem)

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

and $E(T) = 0,\ h(T) = 0$.

3. ASYMPTOTIC ANALYSIS

Tracking Estimator: M-estimator

- If the model is correct i.e $\dot{X}^* = f(t, X^*, \theta^*)$
- Consistency, rate and asymptotic normality of Tracking estimator based
  - Identifiability of the model $(X_{\theta^*} = X_\theta \Rightarrow \theta = \theta^*)$
  - regularity of criterion $S(X, \theta, \lambda)$ w.r.t $\theta \in \Theta$ and $X \in H^1 = \left\{ X | \dot{X} \in L^2([0,T]) \right\}$ $((t,\theta) \mapsto A_\theta(t), r_\theta(t)$ are C^1 on $[0,T] \times \Theta)$.
  - Consistency and plugin properties of nonparametric estimate used (here $\hat{X}$ is a consistent regression spline)
- Some preliminary results:
  - For all θ, for all $X \in H^1$, for all $\lambda > 0$, $S(X, \theta, \lambda) < \infty$
  - For all $X \in H^1$, $\theta \mapsto S(X, \theta, \lambda)$ is C^1
  - $S(X, \theta, \lambda) = \int_0^T G(X(t), \theta)\, dt$
- CONSISTENCY : For any $\lambda > 0$, if $\left\| \hat{X} - X^* \right\|_{L^2} \longrightarrow 0$ in probability then $\hat{\theta}^T \longrightarrow \theta^*$ in Probability.

Root-n rate and asymptotic normality

- Use linearization of S around θ^* such that :

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

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

$$\Gamma(X) = \int_0^T \left(\frac{\partial\left(A_{\theta*}(t).X^*\right)}{\partial\theta} + \frac{1}{\lambda}\frac{\partial h_{\theta^*}(t, X^*)}{\partial\theta}\right)^T \left(\int_t^T R_{\theta^*}(T-t, T-s)X(s)ds\right) d \tag{3}$$

R_{θ^*} is the resolvant of the ODE.

- NORMALITY: For any $\lambda > 0$, if $\hat{X}$ is a regression spline then $\Gamma(\widehat{X}) - \Gamma(X^*) = O_P(n^{-1/2})$ and is asymptotically normal and

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

4. A SLOPPY MODEL

Example: chemical engineering model

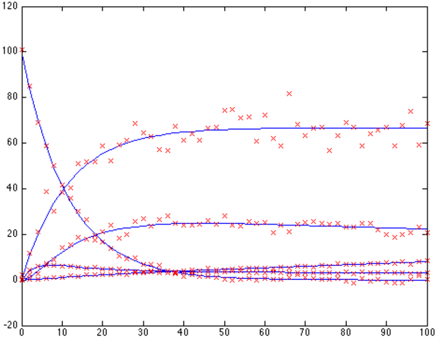
- Linear ODE $\dot{X} = A_\theta X$

$$A_\theta = \begin{pmatrix} -(\theta_1+\theta_2) & 0 & 0 & 0 & 0 \\ \theta_1 & 0 & 0 & 0 & 0 \\ \theta_2 & 0 & -(\theta_3+\theta_4) & 0 & \theta_5 \\ 0 & 0 & \theta_3 & 0 & 0 \\ 0 & 0 & \theta_4 & 0 & -\theta_5 \end{pmatrix} \quad (4)$$

  The initial condition is known and equal to $x_0^* = (100, 0, 0, 0, 0)$ and the true parameter value is $\theta^* = (5.93, 2.96, 2.05, 27.5, 4) \times 10^{-4}$.

- Sloppy model: Fisher Information Matrix has spectrum with ratio $\frac{\lambda_5}{\lambda_1} \simeq 1.5 \times 10^4$

Convergence to equilibrium



Comparison of NLS, GS and Tracking: well-specified model

(n, σ)		$Bias(\widehat{\theta}) \times 10^{-2}$	$Tr\left(V(\widehat{\theta})\right) \times 10^{-4}$	$\text{MSE} \times 10^{-4}$
(100, 4)	$\widehat{\theta}^{T}$	0.05	0.3	0.31
	$\widehat{\theta}^{NLS}$	0.03	0.85	0.86
	$\widehat{\theta}^{GP}$	0.05	0.38	0.38
(100, 8)	$\widehat{\theta}^{T}$	0.46	1.15	1.28
	$\widehat{\theta}^{NLS}$	0.08	2.12	2.13
	$\widehat{\theta}^{GP}$	0.21	1.93	1.95
(50, 4)	$\widehat{\theta}^{T}$	0.45	0.59	0.74
	$\widehat{\theta}^{NLS}$	0.05	1.44	1.44
	$\widehat{\theta}^{GP}$	0.26	1.01	1.03
(50, 8)	$\widehat{\theta}^{T}$	0.28	3.02	3.05
	$\widehat{\theta}^{NLS}$	0.04	6.96	6.96
	$\widehat{\theta}^{GP}$	0.41	3.54	3.61

Table:

Misspecified model for α-pinene

- Real model is $\dot{X} = A_\theta X + v$ with $v(t) = 0.1\sin(\frac{\pi}{50}t) \times (1\,1\,1\,1\,1)^\top$.
- Estimation is performed by assuming homogeneous model.

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

Table:

5. General Case and Pontryagin

Nonlinear and partially observed systems

- The initial model is

$$\begin{cases} \dot{x} & = f(t, x(t), \theta) \\ Y_i & = CX_{\theta, x_0}(t_i) + \varepsilon_i \end{cases}$$

  with unknown initial conditions. The perturbed model $\dot{x} = f(t, x(t), \theta) + Bu$ has solution $X_{\theta, x_0, u}$.
- The criterion is

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

$$\mathscr{S}_\lambda(\hat{Y}; \theta, x_0) = \min_{u \in L_a} \mathscr{C}_\lambda(\hat{Y}; \theta, u, x_0)$$

  where $L_a \subseteq L^2$ is the set of admissible control.
- Tracking estimator is defined as

$$\left(\hat{\theta}_\lambda^T, \widehat{x_0} \right) = \arg\min_\theta \mathscr{S}_\lambda \left(\hat{Y}; \theta, x_0 \right)$$

  For simplicity, we denote $\theta = (\theta, x_0)$.

Extension to semiparametric models

- Straightforward extension to estimation of time-varying parameters $\vartheta : [0, T] \longrightarrow \mathbb{R}^{d_f}$, with the ODE

$$\dot{x} = f(t, x, \theta, \vartheta)$$

- Criterion is

$$\mathscr{C}'_\lambda(\hat{Y}; \theta, u, x_0) = \int_0^T \left\| \hat{Y}(t) - CY_{\theta,\vartheta}(t) \right\|^2 dt + \lambda \int_0^T \left\| \ddot{\vartheta}(t) \right\|^2 dt$$

- Tracking estimator is defined as

$$\hat{\theta}_\lambda^T = \arg\min_\theta \mathscr{S}_\lambda \left(\hat{Y}; \theta, x_0 \right)$$

- The perturbed model is

$$\begin{cases} \dot{x}_1 & = f(t, x(t), \theta, z_1(t)) \\ \dot{z}_1(t) & = z_2(t) \\ \dot{z}_2(t) & = u(t) \end{cases}$$

Necessary and sufficient conditions

- The problem to solve is

$$\begin{aligned} &\min_u \int_0^T \|CX(t) - Y(t)\|_2^2 \, dt + \lambda \int_0^T \|u(t)\|_2^2 \, dt \\ &\text{s.t.} \begin{cases} \dot{X}(t) = f(t, X(t), \theta) + Bu(t) \\ X(0) = x_0 \\ u \in L^2\left([0, T], \mathbb{R}^{d_u}\right) \end{cases} \end{aligned} \tag{5}$$

  for varying $\lambda > 0$, and for all $\theta \in \Theta$, x_0 in $\mathbb{R}^d$ and the smooth function $\hat{Y}$ (close to Y^*).
  - Need to show existence (and uniqueness) of an optimal solution $u_{Y,\theta,\lambda}$ for "all" Y, θ and λ.
  - Find a tractable expression for $(\theta, Y) \mapsto \mathscr{S}(Y, \theta, \lambda)$

Solution with Pontryagin

- If f has compact support Q, and C^1 w.r.t (t,x) on $[0,T]\times Q$, Y is C^0 ; for all θ, $\lambda > 0$, there exists an optimal process $(X_{\theta,\bar{u}}, \bar{u})$ such that

$$\mathscr{S}(Y,\theta,\lambda) = \mathscr{C}_\lambda(\hat{X};\theta,u,x_0)$$

- Solved by Pontryagin Maximum Principle

$$\mathscr{S}(Y,\theta,\lambda) = \|CX_{\theta,\bar{u}} - Y\|^2_{L^2} + \frac{1}{4\lambda}\|Bp_\theta\|^2_{L^2}$$

where p_θ is the adjoint vector solution of

$$\begin{cases} \dot{X}_{\theta,\overline{u}}(t) = f(t, X_{\theta,\overline{u}}(t),\theta) + \frac{1}{2\lambda}BB^T p_\theta(t) \\ \dot{p}_\theta(t) = -\frac{\partial f}{\partial x}(t, X_{\theta,\overline{u}}(t),\theta)^T p_\theta(t) + 2C^T(CX_{\theta,\overline{u}}(t) - Y(t)) \\ (X_{\theta,\overline{u}}(0), p_\theta(T)) = (x_0, 0) \end{cases} \tag{6}$$

and the optimal control is

$$\overline{u}_\theta(t) = \frac{1}{2\lambda}B^T p_\theta(t). \tag{7}$$

Existence and uniqueness of the solution

- If f is C^2 on Q and f_{xx} is bounded on $[0, T] \times Q \times \Theta$, and Y is C^2 , then
  - $\exists \lambda_1$ s.t. for all $\lambda \geq \lambda_1$, $\left(X_{\theta,\bar{u}}, \bar{u}\right)$ is a local minimum
  - $\exists \lambda_2$ s.t. for all $\lambda \geq \lambda_2$, $\left(X_{\theta,\bar{u}}, \bar{u}\right)$ is a global minimum
- Based on quadratic conditions for strong minimum (Milyutin and Osmolovskii).
- For $\lambda \geq \lambda_2$, the solution $(X_{\theta,\bar{u}}, Bp_\theta)$ is unique

$$
\begin{aligned}
&\dot{X}_{\theta,\overline{u}}(t) = f(t, X_{\theta,\overline{u}}(t), \theta) + \frac{1}{2\lambda} BB^T p_\theta(t) \\
&\dot{p}_\theta(t) = -\frac{\partial f}{\partial x}(t, X_{\theta,\overline{u}}(t), \theta)^T p_\theta(t) + 2C^T (CX_{\theta,\overline{u}}(t) - Y(t)) \\
&(X_{\theta,\overline{u}}(0), p_\theta(T)) = (x_0, 0)
\end{aligned}
\tag{8}
$$

Consistency

- Additional assumptions:
  - $\hat{Y} = \sum_{k=1}^{K} \beta_{kK} p_{kK}(t) = \beta_K^T p_K(t)$ is a regression spline,
  - Structural identifiability
    $\forall(\theta, x_0) \in \Theta \times Q \,;\, CX_{\theta,x_0} = CX_{\theta^*,x_0^*} \Longrightarrow \begin{cases} \theta = \theta^* \\ x_0 = x_0^* \end{cases}$
  - the function $(t,x,\theta) \mapsto f(t,x,\theta)$ is C^3 on $[0,T] \times Q \times \Theta$
  - conditions for K_n and regularity for consistency of $\hat{Y}$ in L^2
- For λ big enough,

$$\left|\mathscr{S}(Y^*;\theta,\lambda) - \mathscr{S}(\widehat{Y};\theta,\lambda)\right| \leq K_\lambda \left\|Y^* - \widehat{Y}\right\|_{L^2}$$

Theorem

Under regulariy conditions and $\widehat{Y} \longrightarrow Y^*$ in probability (L^2), then for any $\lambda > F(Y^*)$, we have:

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

Asymptotics: Normality and Rate

Proposition

Let ζ such that $\widehat{Y} \in B(Y^*, \zeta)$. Under regularity and λ big enough:

$$\widehat{\theta}_\lambda^T - \theta^* = -2 \frac{\partial^2 \mathscr{S}(Y^*; \theta^*, \lambda)}{\partial \theta^T \partial \theta}^{-1} \nabla_\theta \mathscr{S}(\widehat{Y}; \theta^*, \lambda) + o_P(1)$$

Proposition

Under regularity and λ big enough, then $Y \longmapsto \nabla_\theta \mathscr{S}(Y; \theta, \lambda)$ is differentiable on $B(Y^*, \zeta)$ for all θ and $(Y_1, Y_2) \longmapsto D(\nabla_\theta \mathscr{S}(.; \theta, \lambda))(Y_1).Y_2$ is s.t.

$$D(\nabla_\theta \mathscr{S}(.; \theta, \lambda))(Y_1).Y_2 = < V(Y_1, \theta), Y_2 > \text{, with}$$

$V(Y_1, \theta) \in C^1\left([0, T], \mathbb{R}^{p \times d'}\right)$.

Theorem

If $\widehat{Y}$ is a regression spline and under regularity conditions , then for any λ big enough, $\widehat{\theta}_\lambda^T - \theta^*$ is asymptotically normal and

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

FitzHughNagumo with perturbations

The true parameters are $a^* = b^* = 0.2$ and $c^* = 3$ and $x_0^* = (V_0^*, R_0^*) = (-1, 1)$. Original model is altered by a step function Z defined by:

$$
\begin{aligned}
Z(t) &= 0 && \text{if } 0 \leq t \leq 5 \\
&= 0.3 && \text{if } 5 < t \leq 10 \\
&= 0 && \text{if } 10 < t \leq 15 \\
&= 0.3 && \text{if } 15 < t \leq 20
\end{aligned}
$$

This function is originally present in the model proposed by [?] to picture an exogenous stimuli. Hence, the true model is in fact

$$
\begin{cases}
\dot{V} = c\left(V - \frac{V^3}{3} + R + Z\right) \\
\dot{R} = -\frac{1}{c}(V - a + bR)
\end{cases}
\tag{9}
$$

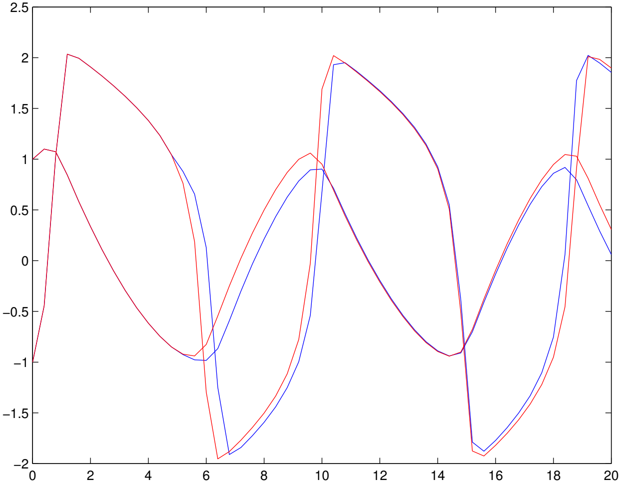


Figure: Solution of exact ODE in blue, solution of perturbed ODE in red.

Comparison of estimators

(n,σ)		$Bias(\widehat{\theta})$	$Tr\left(V(\widehat{\theta})\right)\times 10^{-4}$	MSE $(\times 10^{-2})$
$(100, 0.1)$	θ^*			
	$\widehat{\theta}^T$	0.30	6.36	3.41
	$\widehat{\theta}^{NLS}$	0.43	13	7.13
	$\widehat{\theta}^{GP}$	0.30	7.26	3.48
$(50, 0.1)$				
	$\widehat{\theta}^T$	0.28	9.31	3.45
	$\widehat{\theta}^{NLS}$	0.42	17	6.52
	$\widehat{\theta}^{GP}$	0.36	11	5.47

Table: Bias, Variance, MSE, for Parameter estimation for ill specified FitzHugh-Nagumo model

5. CONCLUSION

Conclusion

- For Sloppy Models: biased estimator for reducing variance and MSE
- Parametric and semiparametric in a seamless framework in the general case
- Dealing with model uncertainty: a good profiled statistical criterion is the value function of a good control problem
- Current directions:
  - Impact of wrong models and selection of λ
  - Approximate ODE model (e.g linear or 2nd order), and identification of networks with sparse model
  - Use of Dynamic Programming for computing the criterion and u as $u(t) = g(X(t))$
  - Stochastic Differential Equations

References and Related Works

- Q. Clairon, N. Brunel . *Parameter Estimation with Pontryagin Maximum Principle*, submitted.
- Q. Clairon, N. Brunel . *State and Parameter Estimation of Partially Observed Linear ODE models*, submitted.
- N. Brunel, Q. Clairon, *A Tracking Approach to Parameter estimation in Linear ODE Models*, Electronic Journal of Statistics, 2015.
- J. O. Ramsay, G. Hooker, J. Cao, D. Campbell, *Parameter estimation for Differential Equations: A generalized Smoothing Approach*, JRSS(B) 2007.
- H. Engl, C. Flamm, P. Kluger, J. Lu, S. Muller, P. Schuster, *Inverse problems in Systems Biology*, Inverse Problems, 2009.
- J. Brynjarsdottir et A. O'Hagan. *Learning about physical parameters: The importance of model discrepancy*. Inverse Problems, 30:24, 2014.