

Another look at estimating parameters in systems of ordinary differential equations via regularization

Ivan Vujačić*

Seyed Mahdi Mahmoudi**, Ernst Wit**

*Department of Mathematics, Vrije Universiteit Amsterdam, The Netherlands

** Department of Statistics and Probability, University of Groningen, The Netherlands

Van Dantzig seminar, March 6, 2014

Introduction

- System of ordinary differential equations (ODEs) in the standard form

$$\begin{cases} x'(t) = f(x(t), t; \theta), & t \in [0, T], \\ x(0) = \xi, \end{cases} \quad (1)$$

where $x(t), \xi \in \mathbb{R}^d$ and $\theta \in \mathbb{R}^p$.

- $x(t; \theta, \xi)$ denotes the solution of (1) for given ξ, θ .
- Many processes in science and engineering are modelled by (1).

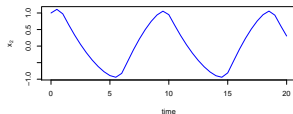
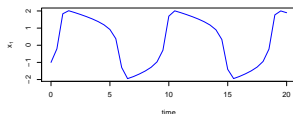
Example: The FitzHugh-Nagumo neural spike potential equations

$$\begin{cases} x_1'(t) = c\{x_1(t) - x_1(t)^3/3 + x_2(t)\}, \\ x_2'(t) = -\frac{1}{c}\{x_1(t) - a + bx_2(t)\}. \end{cases}$$

- x_1 represents the voltage across an axon membrane.
- x_2 summarizes outward currents.

Example:

- $\xi_1 = -1, \xi_2 = 1.$
- $a = 0.2, b = 0.2, c = 3.$



The problem

Noisy observations of $x(t; \theta_0, \xi_0)$ of some states of the system are available:

$$y_i(t_j) = x_i(t_j; \theta_0, \xi_0) + \varepsilon_i(t_j), \quad i = 1, \dots, d_1; j = 1, \dots, n.$$

where $0 \leq t_1 \leq \dots \leq t_n \leq T$.

For simplicity, we consider Gaussian errors.

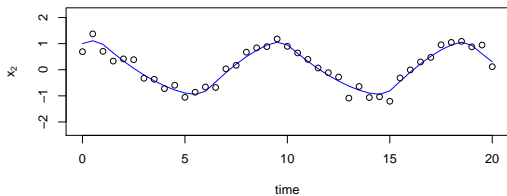
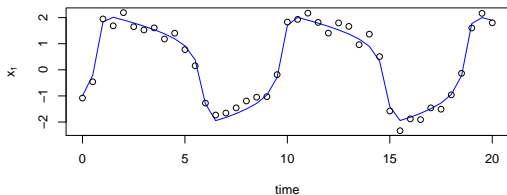
Goal

Estimate θ_0 from the data Y , where $Y = (y_i(t_i))_{ij}$.

This is inverse problem for the coefficients in a system of ODEs.

If ξ_0 is not known it is considered as parameter and estimated as well.

FhNdata from R package 'CollocInfer'



Some existing approaches

- 1 Non-linear least squares (MLE)
- 2 Smooth and match estimators
- 3 Generalized profiling procedure

Non-linear least squares

- 1 Numerical solution $\hat{x}(t; \theta, \xi)$ of the ODE system.
- 2 Criterion $M_n(\theta, \xi)$.

$$M_n(\theta, \xi) = - \sum_{i=1}^{d_1} \sum_{j=1}^n \log p(y_i(t_j) | \hat{x}_i(t_j; \theta, \xi)),$$

where $p(y_i(t_j) | \hat{x}_i(t_j; \theta, \xi))$ is the probability density function of the data.

- NLS estimator is \sqrt{n} -consistent and asymptotically efficient.
- Assumption: the maximum step size of the numerical solver goes to zero.
- Otherwise NLS is not consistent. [Xue et al., 2010]



Xue, H., Miao, H. and Wu, Hulin (2010).

Sieve estimation of constant and time-varying coefficients in nonlinear ordinary differential equation models by considering both numerical error and measurement error.

Annals of statistics, 38:2351–2387.

Smooth and match estimator


- 1 Smoother $\hat{x}(t)$
- 2 Criterion $M_n(\theta)$


$$M_n(\theta) = \int_0^T \|\hat{x}'(t) - f(\hat{x}(t), \theta)\|^q w(t) dt.$$

The \sqrt{n} -consistency was shown for:

- regression splines for $0 < q \leq \infty$. [Brunel et al., 2008]
- kernel estimator for $q = 2$. [Gugushvili and Klaassen, 2012]

Asymptotic normality was shown for regression splines for $q = 2$.
[Brunel et al., 2008]

 Brunel, N. J. et al. (2008).
Parameter estimation of ode's via nonparametric estimators.
Electronic Journal of Statistics, 2:1242–1267.

 Gugushvili, S. and Klaassen, C. A. J. (2012).
 \sqrt{n} -consistent parameter estimation for systems of ordinary differential equations:
bypassing numerical integration via smoothing.
Bernoulli, 18:1061–1098.

Smooth and match estimator: integral criterion

- 1 Smoother $\hat{x}(t)$
- 2 Criterion $M_n(\theta, \xi)$

$$M_n(\theta, \xi) = \int_0^T \|\hat{x}(t) - \xi - \int_0^t f(x(s), \theta) ds\|^2 dt.$$

For $f(x(t), \theta) = g(x(t))\theta$, $g: \mathbb{R}^d \rightarrow \mathbb{R}^{d \times p}$ \sqrt{n} -consistency was shown for:

- local polynomials [Dattner and Klaassen(2013)].
- certain step function estimator in [Vujacic et al.(2014)].



Dattner, I., Klaassen, C.A.:

Estimation in systems of ordinary differential equations linear in the parameters.
arXiv preprint arXiv:1305.4126, (2013)



Vujačić, I., Dattner, I., González, J., Wit, E. :

Time-course window estimator for ordinary differential equations linear in the parameters.

Statistics and Computing, (2014) (To appear in Statistics and Computing. Published online.)

Generalized profiling procedure

- 1 Model based smoother $\hat{x}(t; \theta, \xi)$, where $\hat{x} = \operatorname{argmin}_{x \in \mathcal{X}_m} J(x)$.
- 2 Criterion $M_n(\theta, \xi)$

Inner criterion

$$J(x) = - \sum_{i=1}^{d_1} \sum_{j=1}^n \log p(y_i(t_j) | x_i(t_j; \theta, \xi)) + \lambda \sum_{i=1}^d w_i \int_0^T \{x'_i(t) - f_i(x(t), t, \theta)\}^2 dt,$$

Outer criterion

$$M_n(\theta, \xi) = - \sum_{i=1}^{d_1} \sum_{j=1}^n \log p(y_i(t_j) | \hat{x}_i(t_j; \theta, \xi)).$$

- The estimator is consistent and asymptotically efficient.
[Ramsay et al.(2007)]
- The only frequentist approach that can handle partially observed systems.



Ramsay, J.O., Hooker, G., Campbell, D., Cao, J.:

Parameter estimation for differential equations: a generalized smoothing approach.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), **69**(5):
741–796, (2007)

The framework:

- Stochastic or deterministic approximation \hat{x} of the solution.
- Criterion function M_n .

This talk

For simplicity let ξ_0 be known.

Otherwise, define augmented vector $\theta^* = (\theta, \xi)$.

The framework:

1. $\hat{x}(\theta) = \operatorname{argmin}_{x \in \mathcal{X}_m} \mathcal{J}_{\alpha, \gamma}(x | \theta),$
2. $\hat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} M_n(\theta | \hat{x}(\theta), Y).$

We consider log-likelihood criterion M_n .

Aim

Define $\mathcal{J}_{\alpha, \gamma}$ such that:




- It yields asymptotically efficient estimator.
- It can handle partially observed systems.

Structure of the rest of the presentation

- 1 Background on regularization theory.
- 2 Applying the regularization theory to ODE problem.
- 3 Asymptotic results.
- 4 Conceptual comparison with the generalized profiling procedure.

Only theory in this talk; no simulation studies.

1. Background on regularization theory.

-  Vasin, V. V. and Ageev, A. L. (1995).
Ill-posed problems with a priori information, volume 3.
Walter de Gruyter.
-  Engl, H. W., Hanke, M., and Neubauer, A. (1996).
Regularization of inverse problems, volume 375.
Springer.
-  Pöschl, C. (2008).
Tikhonov regularization with general residual term.
University Innsbruck.

Well-posedness in the sense of Hadamard

Let $F : \mathcal{X} \rightarrow \mathcal{Y}$ where \mathcal{X}, \mathcal{Y} are linear normed spaces and consider the equation

$$F(x) = y, \quad (2)$$

$x \in \mathcal{X}, y \in \mathcal{Y}$.

The problem (2) is *well-posed in the sense of Hadamard* on $(\mathcal{X}, \mathcal{Y})$ if:

- 1 The solution of (2) exists.
- 2 It is unique.
- 3 It is continuous with respect to y .

The problem (2) is *ill-posed* on $(\mathcal{X}, \mathcal{Y})$ if it is not well-posed.

Equation

$$F(x) = y, \quad (3)$$

can be solved on a set $S \subset \mathcal{X}$ by minimizing *objective functional*

$$\mathcal{J}(x) = \|F(x) - y\|^2,$$

on S .

Quasisolution of equation (3) on $S \subset \mathcal{X}$ is any minimizer of \mathcal{J} on S .

It is also called pseudo solution or least squares solution.

Remark:

This idea dates back to the beginning of the 19th century (Gauss, Legendre).

Stabilizing functional and Tikhonov regularization

- Ω - *stabilizing functional*
- Ω incorporates a priori information on the smoothness of the solution x .
- Ω is usually given by a norm or a semi-norm on \mathcal{X} .

Tikhonov regularization involves minimization of the *Tikhonov functional*

$$\mathcal{T}_\alpha(x) = \mathcal{J}(x) + \alpha\Omega(x - x_0),$$

where

- x_0 is *trial solution*
- $\alpha \geq 0$ is *regularization parameter*

Similarity functional and generalized Tikhonov regularization

- *Similarity functional* \mathcal{S} incorporates a priori information on values of x .
- \mathcal{S} measures the closeness of the solution to this a priori information.

Generalized Tikhonov regularization involves minimization of

$$\mathcal{T}_{\alpha,\gamma}(x) = \mathcal{J}(x) + \alpha\Omega(x - x_0) + \gamma\mathcal{S}(x),$$

where $\gamma \geq 0$ is the *penalty parameter*.

- We will call $\mathcal{T}_{\alpha,\gamma}$ *generalized Tikhonov functional*.
- We will call any minimizer of $\mathcal{T}_{\alpha,\gamma}$ *generalized Tikhonov regularizer*.

Finite-dimensional approximation

Numerical minimization - on some finite-dimensional subspace $\mathcal{X}_m \subset \mathcal{X}$.

Minimal assumptions:

- 1 $\mathcal{X}_1 \subset \mathcal{X}_2 \subset \dots$
- 2 $\cup_{m=1}^{\infty} \mathcal{X}_m$ is dense in \mathcal{X} .

Remarks:

- In statistics literature \mathcal{X}_m s are called *sieves*.
- Finite-dimensional approximation is a form of regularization.
- It is called *self regularization* or *regularization by projection*.

Generalized Tikhonov functional

$$\mathcal{T}_{\alpha,\gamma}(x) = \mathcal{J}(x) + \alpha\Omega(x - x_0) + \gamma\mathcal{S}(x).$$

- 1 Objective functional \mathcal{J} .
- 2 Stabilizing functional Ω .
- 3 Similarity functional \mathcal{S} .
- 4 Finite-dimensional approximation.

2. Applying the regularization theory to ODE problem.

Is the problem

$$\begin{cases} x'(t) = f(x(t), t; \theta), & t \in [0, T], \\ x(0) = \xi, \end{cases}$$

ill-posed?

NO.

Is the problem

$$x'(t) = f(x(t), t; \theta), \quad t \in [0, T],$$

ill-posed?

YES.

Even if the initial conditions are known, non-uniqueness can still be introduced through finite dimensional approximation.

Finite-dimensional approximation

- The construction is for fixed θ .
- We suppress dependence on θ for notational simplicity.
- Solution of the system belongs to $(C^1[0, T])^d$.
- $\mathcal{X}_m \subset C^1[0, T]$ linear subspace of dimension m with basis $\{h_1, \dots, h_m\}$.
- Each component of x is approximated by an element of \mathcal{X}_m .

$$x_i(t) = \sum_{k=1}^m \beta_{ik} h_k(t) = \beta_i^\top h(t),$$

where

- $\beta_i = (\beta_{i1}, \dots, \beta_{im})^\top$
- $h(t) = (h_1(t), \dots, h_m(t))^\top$

\mathcal{J} - objective functional

Consider

$$x'(t) = f(x(t), t; \theta), \quad t \in [0, T],$$

for fixed θ .

- Define $F(x(\cdot)) = x'(\cdot) - f(x(\cdot), \cdot, \theta)$,
- ODE system is equivalent to the equation $F(x) = 0_d$.

The corresponding objective functional is

$$\mathcal{J}(x) = \|x' - f(x, \cdot, \theta)\|_{2,w}^2.$$

where

- $w = (w_1, \dots, w_d)$, $w_i > 0$ for $i = 1, \dots, d$,
- $\|x\|_{2,w} = \sqrt{\sum_{i=1}^d w_i \int_0^T x_i^2(t) dt}$.

Ω - stabilizing functional

Here we list two options common in the literature.

Norm in $(L_2[0, T])^d$

$$\Omega(x) = \|x\|_{2,w}^2 = \sum_{i=1}^d w_i \int_0^T x_i^2(t) dt.$$

Norm in Sobolev space $(H^2[0, T])^d$

$$\Omega(x) = \sum_{i=1}^d v_i \int_0^T \{x_i''(t)\}^2 dt.$$

\mathcal{S} - similarity functional

The observations Y represent:

- the data for the problem of the estimation of θ_0 .
- a priori information for the problem of finding the solution $x(t; \theta_0, \xi_0)$.

We have:

- The true distribution of the data g .
- Postulated, a priori distribution of the solution $p(\cdot|x(\cdot; \theta, \xi))$.
- "Distance" between g and $p(\cdot|x(\cdot; \theta, \xi))$ should be small.

Taking KL divergence yields:

$$\mathcal{S}(x) = KL(g(\cdot); p(\cdot|x)) \approx - \sum_{i=1}^{d_1} \sum_{j=1}^n \log p(y_i(t_j)|x_i(t_j)).$$

$\mathcal{T}_{\alpha,\gamma}$ - generalized Tikhonov functional

For fixed θ the generalized Tikhonov functional is

$$\mathcal{T}_{\alpha,\gamma}(x(\beta)) = \mathcal{J}(x(\beta)) + \alpha\Omega(x(\beta) - x_0) + \gamma\mathcal{S}(x(\beta)), \quad (4)$$

where the functionals \mathcal{J} , Ω and \mathcal{S} are defined in previous slides.

The regularized solution is found by optimizing (4) over \mathcal{X}_m^d .

This can be achieved by optimizing (4) with respect to β over \mathbb{R}^{dm} :

$$\hat{\beta} = \operatorname{argmin}_{\beta \in \mathbb{R}^{dm}} \mathcal{T}_{\alpha,\gamma}(x(\beta)),$$

and applying basis expansion $\hat{x}_i(t) = \sum_{k=1}^m \hat{\beta}_{ik} h_k(t) = \hat{\beta}_i^\top h(t)$.

Artificial example: smooth and match estimators fit into the proposed framework

$$\mathcal{T}_{\alpha,\gamma}(x) = \mathcal{J}(x) + \alpha\Omega(x - x_0) + \gamma\mathcal{S}(x).$$

- Take trial solution x_0 to be some smoother of the data.
- $\hat{x} = \operatorname{argmin}_{x \in \mathcal{X}_m^d} T_{\infty,0}(x) = x_0$.

$$M_n(\theta) = \int_0^T \|\hat{x}'(t) - f(\hat{x}(t), \theta)\|^q w(t) dt,$$

Remark:

Similarly, taking trial solution x_0 to be numerical solution yields NLS.

3. Asymptotics

1. $\hat{x}(\theta) = \operatorname{argmin}_{x \in \mathcal{X}_m^d} \mathcal{J}_{\alpha, \gamma}(x | \theta),$
2. $\hat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} M_n(\theta | \hat{x}(\theta), Y).$

We consider log-likelihood criterion M_n and

$$\Omega(x) = \sum_{i=1}^d v_i \int_0^T \{x_i''(t)\}^2 dt.$$

Result for

$$\Omega(x) = \|x\|_{2,w}^2$$

carries over without any modification.



Qi, X. and Zhao, H. (2010).

Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations.

The Annals of Statistics, 38(1):435–481.

Union of sieves is dense in $(C^1[0, T])^d$

$$A_n(\theta, \xi) = \|x^o(\theta, \xi, \cdot) - w\|_\infty \vee \left\| \frac{dx^o}{dt}(\theta, \xi, \cdot) - \frac{dw}{dt} \right\|_\infty \vee \left\| \frac{d^2x^o}{dt^2}(\theta, \xi, \cdot) - \frac{d^2w}{dt^2} \right\|_\infty$$

$$B_n(\theta, \xi) = \|x^u(\theta, \xi, \cdot) - v\|_\infty \vee \left\| \frac{dx^u}{dt}(\theta, \xi, \cdot) - \frac{dv}{dt} \right\|_\infty \vee \left\| \frac{d^2x^u}{dt^2}(\theta, \xi, \cdot) - \frac{d^2v}{dt^2} \right\|_\infty.$$

Lemma

Under Assumption 2 of [Qi and Zhao, 2010], there exist a sequence of finite-dimensional subspaces \mathcal{X}_n of $C^1[0, T]$ such that for any compact subset Θ_0 of Θ and any compact subset Ξ_0 of Ξ , it holds

$$\lim_{n \rightarrow \infty} r_n = 0,$$

where

$$r_n = \max \left\{ \sup_{(\theta, \xi) \in \Theta_0 \times \Xi_0} \inf_{w \in \mathcal{X}_n, w(0) = \xi_0^o} A_n(\theta, \xi), \sup_{(\theta, \xi) \in \Theta_0 \times \Xi_0} \inf_{v \in \mathcal{X}_n, v(0) = \xi_0^u} B_n(\theta, \xi) \right\}.$$

Theorem (Consistency)

Let Assumptions 1-5 from [Qi and Zhao, 2010] hold. If as $n \rightarrow \infty$

① $r_n \rightarrow 0$

② $\alpha_n \rightarrow 0$

③ $\gamma_n \rightarrow 0$

then $\hat{\theta}_n - \theta_0 = o_P(1)$.

$$\mathcal{J}_{\alpha,\gamma}(x) = \mathcal{J}(x) + \alpha\Omega(x - x_0) + \gamma\mathcal{S}(x).$$

Theorem (Asymptotic efficiency)

Let Assumptions 1-6 from [Qi and Zhao, 2010] hold. If $r_n = o(n^{-1})$, $\alpha_n = o(n^{-2})$ and $\gamma_n = o(n^{-2})$ as $n \rightarrow \infty$ then $\hat{\theta}_n$ is asymptotically normal with the same asymptotic covariance matrix as that of the maximum likelihood estimation.

4. Conceptual comparison with the generalized profiling procedure.

Generalized profiling fits into the proposed framework

Inner criterion of the generalized profiling procedure

$$J(x) = - \sum_{i=1}^{d_1} \sum_{j=1}^n \log p(y_i(t_j) | x_i(t_j; \theta)) + \lambda \sum_{i=1}^d w_i \int_0^T \{x'_i(t) - f_i(x(t), t, \theta)\}^2 dt$$

can be written as

$$J(x) = \lambda \left\{ \frac{1}{\lambda} \mathcal{S}(x) + \mathcal{J}(x) \right\} = \lambda \mathcal{T}_{0,1/\lambda}(x).$$

Thus, model based smoother \hat{x} is

$$\hat{x} = \operatorname{argmin}_{x \in \mathcal{X}_m^d} \mathcal{T}_{0,1/\lambda}(x).$$

Smoothing VS Generalized Tikhonov regularization

”For solutions to the dynamic systems, however, the roles of goodness of fit and ‘roughness penalty’ seems more likely reversed, with fidelity to the ODE the major concern and the ‘error distribution’ of the data an afterthought (Chong Gu - in the discussion section of [Ramsay et al.(2007)]).

In the generalized profiling:

- Fidelity to the ODE term is the penalty.
- λ must approach ∞ : leads to ill conditioning in the optimization.

In the regularization formulation

- Fidelity to the ODE term is the main term— objective functional.
- γ must approach 0: no ill conditioning in the optimization.

Generalized Tikhonov regularizer and its special cases

Parameters	$\mathcal{T}_{\alpha,\gamma}(x)$	$\hat{x} = \operatorname{argmin}_{x \in \mathcal{X}_m^d} \mathcal{T}_{\alpha,\gamma}(x)$
$\alpha > 0, \gamma > 0$	$\mathcal{J}(x) + \alpha\Omega(x - x_0) + \gamma\mathcal{S}(x)$	Gen. Tikhonov's regularizer
$\alpha = 0, \gamma = 0$	$\mathcal{J}(x)$	Ivanov's quasi solution
$\alpha > 0, \gamma = 0$	$\mathcal{J}(x) + \alpha\Omega(x - x_0)$	Tikhonov's regularizer
$\alpha = 0, \gamma > 0$	$\mathcal{J}(x) + \gamma\mathcal{S}(x)$	model based smoother
$\alpha = \infty, \gamma = 0$	$\mathcal{J}(x_0)/\delta(x - x_0)$	trial solution x_0

Table: The last row should be interpreted as $\mathcal{T}_{\alpha,0}(x) \rightarrow \mathcal{J}(x_0)/\delta(x - x_0)$ as $\alpha \rightarrow +\infty$, where δ is the Dirac's delta function.

- Regularization provides a coherent and principled framework for defining an approximation of the solution of ODE.
- ODE system is solved in the least square sense.

Acknowledgments

- Bartek Knapik
Department of mathematics, Vrije Universiteit Amsterdam, The Netherlands
- Itai Dattner
Department of statistics, University of Haifa, Israel

Questions, comments,...