Differential equations and empirical likelihood

Debashis Paul

Department of Statistics
University of California, Davis

with Sanjay Chaudhuri (NUS) and Xiaoyan Liu (UCD)

June 23, 2016

Workshop on New Applications of Empirical Likelihood

Institute for Mathematical Sciences, National University of Singapore

General formulation

- Observed data:

$$\mathbf{Y}_i = \mathbf{X}(t_i) + \varepsilon_i, \qquad i = 1, \ldots, n; \tag{1}$$

  where $t_i \in [T_0, T_1]$.

- Parametric ODE model:

$$\frac{d}{dt}\mathbf{X}(t) = G(\mathbf{X}(t), t; \theta) \tag{2}$$

  subject to an initial condition $\mathbf{X}(T_0) = \mathbf{X}_0$.

- If $G(\mathbf{x}, t; \theta) \equiv G(\mathbf{x}, \theta)$, then the system is *autonomous*.

Estimating θ by regression

- Treat $\mathbf{X}(t)$ as a function of (θ, X_0).
- If X_0 is known, the trajectory for $\mathbf{X}(t) = \mathbf{X}(t, \theta)$ can be determined for all $t \in [T_0, T_1]$, *up to a given level of precision* by numerically solving (2), for example by Runge-Kutta method.
- One obvious approach is to estimate θ by solving the least squares problem:

$$\widehat{\theta}^{LS} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} w_i \|\mathbf{Y}_i - \mathbf{X}(t_i; \theta)\|^2 \tag{3}$$

based on some *known weights* $\{w_i\}_{i=1}^{n}$.

Computational bottleneck

- Typically, (3) is solved by iteratively solving a sequence of linearized regression problem (e.g., Levenberg-Marquardt method). This involves, at each step of the iterations,
  1. Solve for $\mathbf{X}(t;\theta)$ on $[T_0, T_1]$ given the current value of θ.
  2. Solve for $\frac{d}{dt}\mathbf{X}(t;\theta)$ on the same interval, by solving the corresponding ODE.
- Both these steps need knowledge of X_0, and evaluation of the trajectory on a fine grid.

Alternative approaches : Two-stage estimation

- The first stage involves estimating $X(t)$ and $X'(t)$ separately using local polynomial method.
- The second stage involves nonlinear regression based on the following description:

$$\widehat{\mathbf{X}}'(t_i) = G(\widehat{\mathbf{X}}(t_i), t_i; \theta) + \mathbf{e}_i, \qquad i = 1, \ldots, n.$$

- The method has been analyzed in detail by Chen and Wu (2008a, 2008b).
- Choice of weights in the nonlinear regression can impact the quality of estimation. Brunel (2008) (who used splines rather than local polynomial regression for smoothing $\mathbf{X}(t)$), established $\sqrt{n}$-consistency of $\widehat{\theta}$ under a careful choice of weights that vanish at the boundaries.

Alternative approaches : Parameter cascading

- A penalized least squares method whereby the trajectory X is represented in terms of splines and the ODE itself acts as a penalty on the trajectory. The parameters of the ODE are fitted using a parameter cascading method (Ramsay, Hooker, Campbell and Cao, 2007).
- Specifically, after expressing $\widehat{X}(t) = \sum_{j=1}^{M} \beta_j \Phi_j(t)$, they minimize

$$\sum_{i=1}^{n} w_i \|\mathbf{Y}_i - \widehat{\mathbf{X}}(t_i)\|^2 + \lambda \int \|\widehat{\mathbf{X}}'(t) - G(\widehat{\mathbf{X}}(t), t; \theta)\|^2 dt$$

  for some $\lambda > 0$.
- Qi and Zhao (2010) provided theoretical analysis of the parameter cascading method.

Alternative approaches : Linear functionals

- Hall and Ma (2014) proposed another smoothing-based estimator by making use of a class of linear *test functionals*.
- Specifically, if ψ_j, $j = 1, \ldots, J$ are C^1 functions supported on $[c, 1-c]$ for some $c \in (0, 1/2)$, and satisfying $\psi_j(c) = \psi_j(1-c) = 0$, then estimate θ by minimizing

$$\sum_{j=1}^{J} W_j \| \int_c^{1-c} \left(\psi_j(t)\widehat{\mathbf{X}}(t) + \psi_j(t) G(\widehat{\mathbf{X}}(t), t; \theta) \right) dt \|^2$$

  where $\widehat{\mathbf{X}}(t)$ is obtained by local polynomial regression.
- They show that the estimator is $\sqrt{n}$-consistent under appropriate range of the bandwidth and appropriately chosen $\{\psi_j\}_{j=1}^{J}$.

Empirical likelihood framework

- The main idea is to combine the nonparametric smoothing with the parametric regression of $\widehat{\mathbf{X}}'_{\mathbf{h}}(t)$ on $\widehat{\mathbf{X}}_{\mathbf{h}}(t)$, so that the bandwidths of smoothers can be chosen simultaneously.
- *EL step:* Use the following normal equations as the "core estimating equations":

$$\sum_{i=1}^{n} w_i \frac{\partial}{\partial\theta} \|\widehat{\mathbf{X}}'_{\mathbf{h}}(t_i) - G(\widehat{\mathbf{X}}_{\mathbf{h}}(t_i), t_i; \theta)\|^2 = 0 \qquad (4)$$

  where $\mathbf{w} := (w_i)_{i=1}^{n} \in \mathcal{S}_n$ (n-dimensional simplex).
- Need additional estimating equations to control the bandwidths $\mathbf{h}$ and improve stability.

Parametric integral constraint

- We impose the restriction

$$\sum_{i=1}^{n} w_i \left(\widehat{\mathbf{X}}''(t_i) - \frac{\partial}{\partial t} G(\widehat{\mathbf{X}}(t), t; \theta) \,|_{t=t_i} \right) = 0. \qquad (5)$$

- Maximize $\prod_{i=1}^{n} w_i$ subject to (4) and (5), given θ and $\mathbf{h}$, to obtain

$$\ell_P(\theta, \mathbf{h}) := \sum_{i=1}^{n} \log \hat{w}_i(\theta, \mathbf{h}).$$

Nonparametric integral constraints

- Given a set of weights $\boldsymbol{\nu} := (\nu_i)_{i=1}^n \in \mathcal{S}_n$, we require:

$$\sum_{i=1}^{n} \nu_i(\mathbf{Y}_i - \widehat{\mathbf{X}}(t_i)) = 0. \tag{6}$$

- Based on bias-variance trade-off used to select the optimal global bandwidth, we also impose:

$$\sum_{i=1}^{n} \nu_i \left((Y_{j,i} - \widehat{X}_{j,h_j}(t_i))^2 - c(K_j) n h_j^5 (\widehat{X}''_{j,h_j}(t_i))^2 \right) = 0, \tag{7}$$

for $j = 1, ldots, d$, where $c(K_j)$ is a constant determined by the kernel K_j used in obtaining the smoother $\widehat{X}_{j,h_j}$.

- Maximize $\prod_{i=1}^n \nu_i$ subject to (4) and (5), given $\mathbf{h}$, to obtain

$$\ell_{NP}(\mathbf{h}) := \sum_{i=1}^{n} \log \hat{\nu}_i(\mathbf{h}).$$

Final penalized loss function

- We minimize the following loss function to obtain the final estimates of θ and $\mathbf{h}$.

$$L(\theta, \mathbf{h}) = -\ell_P(\theta, \mathbf{h}) - \ell_{NP}(\mathbf{h}) + \eta \sum_{j=1}^{d} \mathrm{trace}(\mathbf{S}_{j,h_j}^{(1)}) \qquad (8)$$

  where $\eta \geq 0$ is a specified constant, and $\mathbf{S}_{j,h_j}^{(k)}$ denotes the smoother matrix used to obtain $\widehat{X}_{j,h_j}^{(k)}(t)$ for $k = 0, 1, 2$.
- We choose either $\eta = 0$ (no trace-based penalty), or $\eta = 2$ (motivated by AIC).

Simulation study

We use the following *Lotka-Volterra* system as a testbed for our procedure. The same system has been studied by Brunel (2008) and Ramsay et al. (2007).
$\mathbf{X}(t) = (x(t), y(t))^T$, where

$$\begin{aligned} x'(t) &= \theta_1 x(t) - \theta_3 x(t) y(t) \\ y'(t) &= -\theta_2 y(t) + \theta_4 x(t) y(t). \end{aligned}$$

Observational errors:

$$\boldsymbol{\varepsilon}_i = (\varepsilon_i^x, \varepsilon_i^y)^T \overset{i.i.d.}{\sim} N(0, \sigma^2 I_2).$$

Detailed settings

- We compare with the results obtained by Brunel (2008). So, we choose an interval $[T_0, T_1]$ such that $T_1 - T_0 = 20$. Also, we choose $(x(T_0), y(T_0)) = (1, 3)$, $(\theta_1, \theta_2, \theta_3, \theta_4) = (1, 1.5, 1.5, 2)$ and $\sigma = 0.2$.
- We report the results when we have $n = 200$ equally spaced time points in the interval.
- The proposed procedure requires a decent initial estimator for fast convergence. We use a two-step procedure where the first step involves estimation of $(x(t), y(t))$ and their derivatives by local quadratic regression. However, we use cross-validated bandwidths (h_1^{opt}, h_2^{opt}) for obtaining $(\widehat{x}(t), \widehat{y}(t))$, and $(h_1^{opt}/2, h_2^{opt}/2)$ for obtaining $(\widehat{x}'(t), \widehat{y}'(t))$.

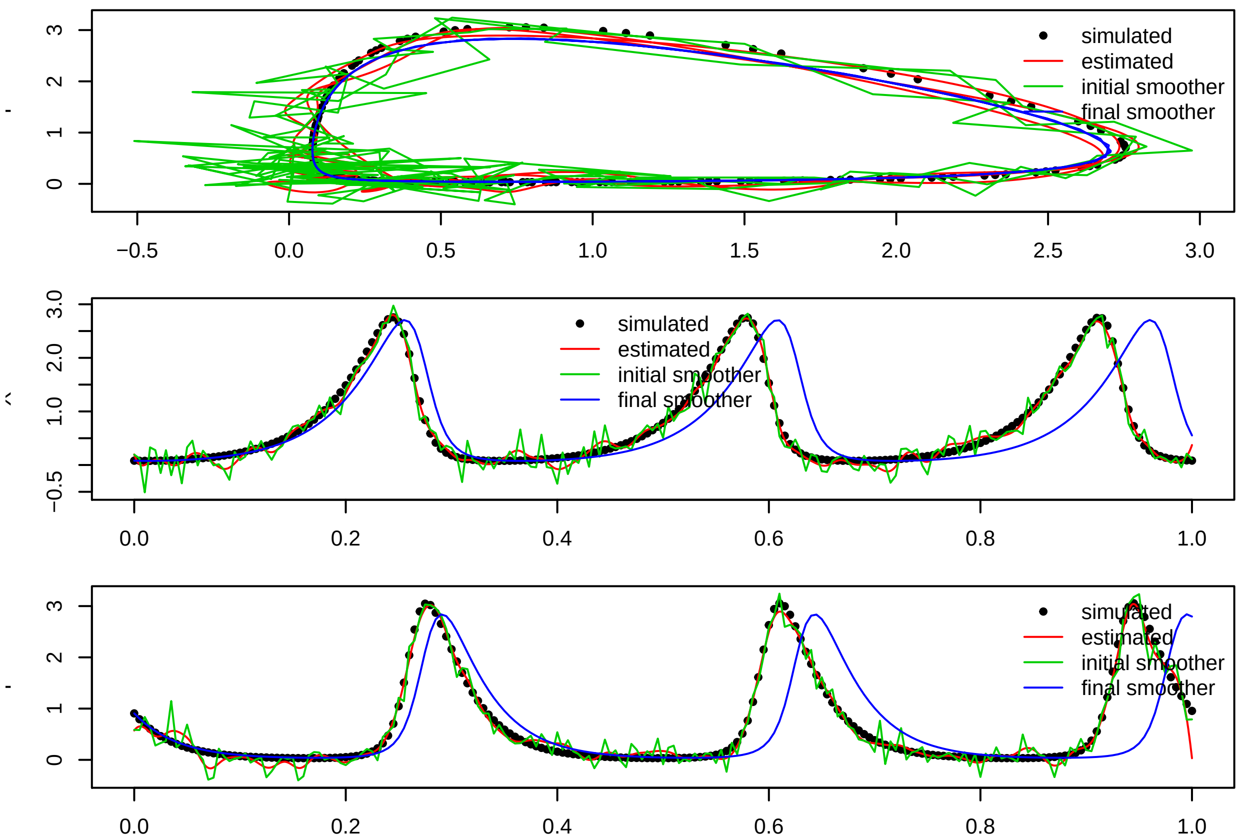
Parameter	Criterion	Brunel	Initial	Proposed $(\eta = 0)$	Proposed$(\eta = 2)$
θ_1	Bias	-.08	-.0550	-.0693	-.0345
	(SD)	(.07)	(.0677)	(.0806)	(.0709)
θ_2	Bias	-0.10	-.1422	-.0950	-.0716
	(SD)	(.10)	(.1358)	(.1266)	(.1110)
θ_3	Bias	-.13	-.1047	-.1117	-.0475
	(SD)	(.09)	(.0718)	(.1040)	(.0813)
θ_4	Bias	-.17	-.1845	-.1410	-.0831
	(SD)	(.12)	(.1130)	(.1524)	(.1073)
σ_x^2	Bias		-.0135	-.0204	-.0185
	(SD)		(.0035)	(.0028)	(.0051)
σ_y^2	Bias		-.0132	-0.0215	-.0193
	(SD)		(.0036)	(.0028)	(.0048)
$\hat{h}_x$	Mean		.0106	.0071	.0079
$\hat{h}_y$	Mean		.0106	.0068	.0076

Summary and future work

- Empirical likelihood method provides a flexible approach for estimating ODEs without having to evaluate the trajectories explicitly.
- Appropriate penalization scheme is needed to adjust for bandwidth, which can also be incorporated by combining the parametric and nonparametric aspects.
- The method enjoys a certain robustness to noise characteristics.
- Asymptotic theory needs to developed.

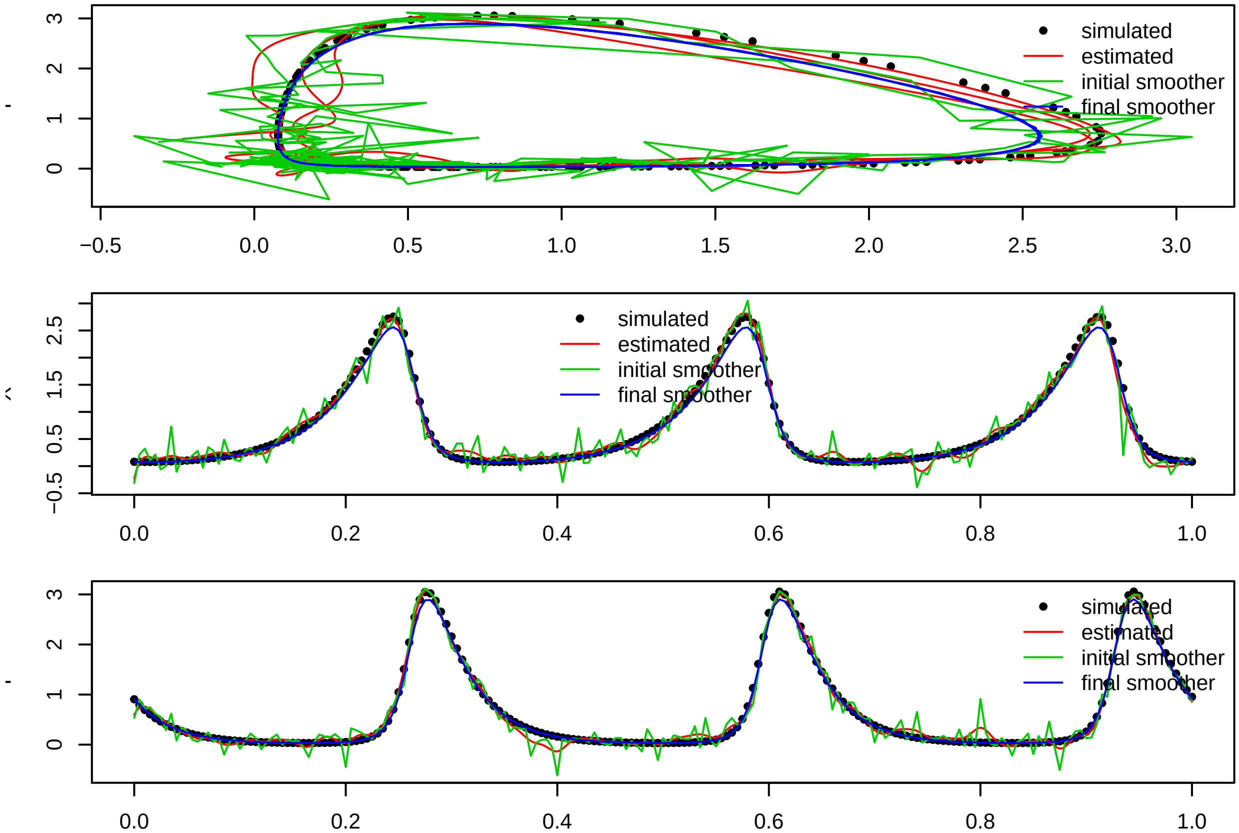
References

- Brunel, N. J.-B. (2008). Parameter estimation of ODEs via nonparametric estimators. *Electronic Journal of Statistics*, **2**, 1242–1267.
- Chen, J. and Wu, H. (2008). Estimation of time-varying parameters in deterministic dynamic models with application to HIV infections. *Statistica Sinica*, **18**, 987–1006.
- Hall, P. and Ma, Y. (2014). Quick and easy one-step parameter estimation in differential equations. *Journal of Royal Statistical Society, Series B*, **76**, 735–748.
- Ramsay, J. O., Hooker, G., Campbell, D. and Cao. J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society, Series B*, **69**, 741–796.



simulated
estimated
initial smoother
final smoother
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
simulated
estimated
initial smoother
final smoother
0.0
0.2
0.4
0.6
0.8
1.0
−0.5
1.0
2.0
3.0
simulated
estimated
initial smoother
final smoother
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3


```
[1] "run: 27"
```


simulated
estimated
initial smoother
final smoother
simulated
estimated
initial smoother
final smoother
simulated
estimated
initial smoother
final smoother

```
[1] "run: 28"
```