Statistica Sinica **24** (2014), 1613-1631
doi:http://dx.doi.org/10.5705/ss.2012.304

ESTIMATION OF ORDINARY DIFFERENTIAL EQUATION PARAMETERS USING CONSTRAINED LOCAL POLYNOMIAL REGRESSION

A. Adam Ding and Hulin Wu

Northeastern University and University of Rochester

Abstract: We propose to use a constrained local polynomial regression to estimate the unknown parameters in ordinary differential equation models with a goal of improving the smoothing-based two-stage pseudo-least squares estimate. The equation constraints are derived from the differential equation model and are incorporated into the local polynomial regression in order to estimate the unknown parameters in the differential equation model. We also derive the asymptotic bias and variance of the proposed estimator. Our simulation studies show that our estimator is clearly better than the pseudo-least squares estimator in estimation accuracy with a small price of computational cost. An application to immune cell kinetics and trafficking for influenza infection further illustrates the benefits of the proposed method.

Key words and phrases: Constrained optimization, local polynomial smoothing, ordinary differential equation.

1. Introduction

Differential equations are widely used to describe and quantify dynamic systems in many scientific fields. The so-called inverse problem of differential equation models, i.e., the estimation of unknown parameters based on experimental data of state variables, is quite challenging because the standard nonlinear least squares method may fail due to convergence problems, local minima, and high computational cost. Recently, alternative methods based on nonparametric smoothing have been proposed and investigated by Poyton et al. (2006), Ramsay et al. (2007), Chen and Wu (2008), Liang and Wu (2008), Brunel (2008). They intend to improve the computational efficiency and stability of the nonlinear least squares method at a cost of reduced estimation accuracy.

A general nonlinear ordinary differential equation model can be written as

$$\frac{dX(t)}{dt} = F\{X(t);\theta\}, \tag{1.1}$$

where $X(t)=\{X_1(t),\ldots,X_d(t)\}^T$ is a d-dimensional state vector, $\theta=(\theta_1,\ldots,\theta_q)^T$ is a q-dimensional vector of unknown parameters, and $F(\cdot)=\{F_1(\cdot),\ldots,F_d(\cdot)\}^T$

is a known nonlinear function vector. The proposed methodology with minor modifications is also applicable to more general differential equations with input variables. The process $X(t)$ is usually measured with noise, say

$$Y(t) = X(t) + e(t), \tag{1.2}$$

where the measurement error $e(t)$ is independent of $X(t)$ with mean zero and a covariance matrix Σ_e.

Denote the solution to the differential equation (1.1) as $X(t;\theta)$. Generally $X(t;\theta)$ does not have an analytic solution and needs to be obtained by solving the differential equations numerically. This results in computationally intensive and often numerically unstable estimation for θ. To avoid numerically solving the differential equations, nonparametric smoothing techniques were applied to the observed process to estimate the parameters θ via multiple-stage procedures in Poyton et al. (2006), Ramsay et al. (2007), Chen and Wu (2008), Liang and Wu (2008) and Brunel (2008). Particularly, Liang and Wu (2008) proposed using local polynomial estimation as the smoothing technique in the first stage, and obtained the pseudo-least square estimator for θ in the second estimation stage. There, (1.1) was only used in the second estimation stage, which results in a significant reduction of estimation efficiency of the pseudo-least squares estimator compared to the nonlinear least squares estimator. We propose a new approach to improve the Liang and Wu's pseudo-least squares estimator by combining the local polynomial smoothing and differential equation information. We expect the new method to gain more in estimation accuracy at a small computational cost.

2. Differential Equation-Constrained Local Polynomial Regression

2.1. Notation and method

Suppose the process $Y(t)$ is observed at time points $t_1, t_2, \ldots, t_n$, so (1.2) becomes

$$Y_i = Y(t_i) = X(t_i) + e(t_i), \quad i = 1, \ldots, n. \tag{2.1}$$

For notational simplicity, we present our model and method in the univariate case. However, the proposed methodologies and theoretical results can be easily extended to $d > 1$. In particular, we illustrate this point in our simulation studies and data analysis by applying the proposed method to multivariate cases in Section 3.

We can estimate $X(t)$ and its derivative at any time point t by nonparametric local polynomial smoothing of observed Y_is, $i = 1, \ldots, n$. Thus, we obtain the smoothing estimates for the process $\hat{X}(t_k^*)$ and its derivative $\hat{X}'(t_k^*)$ over a grid of time points $t = t_1^*, t_2^*, \ldots, t_m^*$. Liang and Wu (2008) proposed to: (1) use the local polynomial smoothing over the grid of observed time points t_i to yield estimates

for $\hat{X}(t_i)$ and $\hat{X}'(t_i)$, $i = 1, 2, \ldots, n$, and (2) estimate θ using the pseudo-least squares estimator

$$\hat{\theta}_{PLS} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} [\hat{X}'(t_i) - F\{\hat{X}(t_i); \theta\}]^2 \omega(t_i),$$

with $\omega(t_i)$ an appropriate weight function. In general, we can extend Liang and Wu's procedure over a time grid of size m which can be larger than the number of original measurements n, with

$$\hat{\theta}_{PLS} = \underset{\theta}{\operatorname{argmin}} \sum_{k=1}^{m} [\hat{X}'(t_k^*) - F\{\hat{X}(t_k^*); \theta\}]^2 \omega(t_k^*).$$

This modified pseudo-least squares estimator converges at the $n^{-1/2}$ rate (Liang and Wu (2008, 2010); Fang, Wu and Zhu (2011)).

Notice that Liang and Wu (2008)'s first stage smoothing was done without using the differential equation information. The differential equation was only used in the second stage to estimate θ based on the first stage smoothing results. The separation of these two stages results in a significant reduction in estimation accuracy of the differential equation parameters. Poyton et al. (2006) and Ramsay et al. (2007) used a spline approach to combine the smoothing stage with the differential equation information, which produced a more accurate and stable estimate. They minimize a criterion combining the residuals in smoothing fits to observations Y_i's and deviation of the smoothing fits from the differential equation model. Motivated by these ideas, we propose to estimate the differential equation parameters θ jointly with the state variable $\hat{X}(t_k^*)$ and $\hat{X}'(t_k^*)$, $k = 1, 2, \ldots, m$, and expect to improve on the Liang-Wu's pseudo-least squares estimator in estimation accuracy at a small computational cost.

The standard local pth-order polynomial regression estimates $X(t)$ and its derivative up to order p at time t can be obtained by minimizing

$$\sum_{i=1}^{n} \{Y_i - (\alpha + \sum_{j=1}^{p} \beta_j (t_i - t)^j)\}^2 K_h(t_i - t), \tag{2.2}$$

where $K(\cdot)$ is a symmetric kernel function, $K_h(\cdot) = K(\cdot/h)/h$, and h is the bandwidth. Then $X(t)$ can be estimated with that of α and the derivatives $X^{(j)}(t)/j!$ can be estimated with those of β_j, $j = 1, 2, \ldots, p$ (Fan and Gijbels (1996)). In (1.1), (α, β_1) in (2.2) should satisfy $\beta_1 = F(\alpha; \theta)$. The higher derivatives $X^{(j)}(t)$ can similarly be expressed as functions of $X(t)$ through (1.1). For example, if $D_X(X; \theta) = \partial F(X; \theta)/\partial X$, then

$$\frac{d^2 X(t)}{dt^2} = \frac{\partial}{\partial X} F\{X(t); \theta\} \frac{dX(t)}{dt} = D_X\{X(t); \theta\} F\{X(t); \theta\}.$$

With $F^{(1)}(X;\theta) = D_X(X;\theta)F(X;\theta)$, (1.1) implies that we should restrict $\beta_2 = F^{(1)}(\alpha;\theta)/2$. Similarly, let $D_X^{(j)}(X;\theta) = \partial F^{(j)}(X;\theta)/\partial X$, and $F^{(j)}(X;\theta) = D_X^{(j-1)}(X;\theta)F(X;\theta)$ with $F^{(0)}(X;\theta) = F(X;\theta)$ and $D_X^{(0)}(X;\theta) = D_X(X;\theta)$. Then we have

$$X^{(j)}(t) = \frac{d^j X(t)}{dt^j} = F^{(j-1)}\{X(t);\theta\}.$$

Thus, we have the general differential equation constraints

$$\beta_j = \frac{X^{(j)}(t)}{j!} = \frac{F^{(j-1)}\{X(t);\theta\}}{j!} = \frac{F^{(j-1)}\{\alpha;\theta\}}{j!}, \quad j = 1,\dots,p. \tag{2.3}$$

After plugging in the constraints, the objective function of the local polynomial regression (2.2) can be reformulated as

$$\sum_{i=1}^{n}[Y_i - \{\alpha + \sum_{j=1}^{p}\frac{F^{(j-1)}(\alpha;\theta)}{j!}(t_i - t)^j\}]^2 K_h(t_i - t). \tag{2.4}$$

The optimization of (2.4) jointly over α and θ provides estimates $\hat{\alpha} = \hat{X}(t)$ and $\hat{\theta}$ simultaneously.

The optimization of (2.4) is unlikely to provide a good estimate for $\hat{\theta}$ since it only uses the differential equation constraint of $X(t)$ at one time point t. Following Liang and Wu (2008) and Brunel (2008), we could estimate θ by integrating the objective function over the grid of time points $t = t_1^*, t_2^*, \dots, t_m^*$, to minimize

$$\sum_{k=1}^{m}\sum_{i=1}^{n}[Y_i - \{\alpha_k + \sum_{j=1}^{p}\frac{F^{(j-1)}(\alpha_k;\theta)}{j!}(t_i - t_k^*)^j\}]^2 K_h(t_i - t_k^*)\omega(t_k^*), \tag{2.5}$$

with respect to $\xi = (\alpha_1,\dots,\alpha_m,\theta)^T$, where $\omega(t_k^*)$ are nonnegative weights over the time grid, as suggested by Brunel (2008), and the bandwidth h can be determined by cross-validation approach or the plug-in method suggested by Liang and Wu (2008). The $\hat{\xi}$ that minimizes (2.5) is called the differential equation constrained local polynomial estimator.

For a general nonlinear function F of the differential equation model, the optimization of (2.5) becomes a nonlinear minimization problem, and we may lose the computational efficiency of the original local polynomial fitting. For this, we consider a linear estimator that results from one iteration of the Gauss-Newton optimization of (2.5) at a previous estimate $\xi^* = (\alpha_1^*,\dots,\alpha_m^*,\theta^*)^T$. In matrix notation, the objective function (2.5) is $[Y - G(\xi)]^T W[Y - G(\xi)]$, where $Y = (Y_1,\dots,Y_n,\dots,Y_1,\dots,Y_n)^T$ is a (nm)-dimensional vector with the observations Y_i's repeated m times, $G = (G_{1,1},\dots,G_{n,1},\dots,G_{1,m},\dots,G_{n,m})^T$ with

$$G_{i,k}(\xi) = G_{i,k}(\alpha_k,\theta) = \{\alpha_k + \sum_{j=1}^{p}\frac{F^{(j-1)}(\alpha_k;\theta)}{j!}(t_i - t_k^*)^j\}$$

and W is the $nm \times nm$ diagonal weight matrix

$$\text{Diag}\{\omega(t_1^*)K_h(t_1 - t_1^*), \ldots, \omega(t_1^*)K_h(t_n - t_1^*), \ldots, \omega(t_m^*)K_h(t_1 - t_m^*), \ldots, \omega(t_m^*)K_h(t_n - t_m^*)\}.$$

Let $J = (\partial G/\partial\alpha_1, \ldots, \partial G/\partial\alpha_m, \partial G/\partial\theta_1, \ldots, \partial G/\partial\theta_q)_{\xi=\xi^*}$ denote the $nm \times (m+q)$ Jacobian matrix evaluated at $\xi = \xi^*$. Then a Gauss-Newton iteration minimizes (2.5) with $G(\xi)$ replaced by its linear approximation $G(\xi^*) + J(\xi - \xi^*)$. This results in the weighted linear least squares estimator

$$\hat{\xi} = (J^T W J)^{-1} J^T W \tilde{Y}, \tag{2.6}$$

where $\tilde{Y} = Y - G(\xi^*) + J\xi^*$ is a (nm)-dimensional vector.

The selection of bandwidth h and m is an important issue for practice. We suggest selecting the bandwidth h using the plug-in method according to the recommendations of Liang and Wu (2008). This works very well in our numerical simulations and data analysis in Section 3. Selection of m for data-augmentation is less critical based on our simulation results. In theory, larger m is better if the computational cost does not increase too much. Thus, we can select m as large as can be afforded. Our method can also be adapted to handle the case with partially observed state variables or observed functions of state variables in principle, but it may be difficult to find initial values for the unobserved state variables.

2.2. Asymptotic property for $\hat{\theta}$

We need the following technical conditions.

(1) (1.1) holds over a time interval $[a_0, b_0]$ and has a bounded solution $X(t)$. We observe $Y_i(t)$ as (2.1) for $t = t_i \in [a_0, b_0]$, $i = 1, \ldots, n$. The differential equation parameters θ are jointly estimated with $\alpha_i = X(t_i^*)$ over a time grid $t_i^* \in [a_0, b_0]$, $i = 1, \ldots, m$. The resulting estimator $\hat{\xi}$ is given by (2.6) with the linearization at a starting value $\xi^* = (\alpha_1^*, \ldots, \alpha_m^*, \theta^*)^T$.

(2) The starting value is an estimator ξ^* such that $|\xi^* - \xi| = O_p(n^{-\delta})$ for some $\delta > 1/4$, $|\cdot|$ the L_∞ norm.

(3) The function $F(x)$ as (1.1) has a bounded pth order derivative.

(4) $n \to \infty$, $h \to 0$, $nh \to \infty$ and $m \to \infty$.

(5) The kernel function $K \geq 0$ is compactly supported and bounded, with $\mu_j(K) = \int K(u)u^j du$, $\mu_0(K) = \int K(u)du = 1$, and all odd-order moments $\mu_j(K) = 0$.

(6) The observation time points $t_1, \ldots, t_n$ and fitted time points $t_1^*, \ldots, t_m^*$ follow distributions with densities $f(t)$ and $f_g(t)$, $t \in [a_0, b_0]$, respectively. Over the time interval $t \in [a_0, b_0]$, $f(t) > 0$ and $f_g(t) > 0$ are bounded with continuous derivatives $f'(t)$ and $f_g'(t)$.

(7) The weight function $\omega(t) \geq 0$ is bounded over the time interval $t \in [a_0, b_0]$.

Theorem 1. *Under* (1)−(7)*, conditional on the observation time points* $t_1, \ldots, t_n$*, fitted time points* $t_1^*, \ldots, t_m^*$ *and* ξ^**, the differential equation parameter estimator* $\hat{\theta}$ *has conditional bias*

$$Bias(\hat{\theta}) = o_p(n^{-1/2}) + O_p(h^{p+1}) \ p \text{ odd}, \quad Bias(\hat{\theta}) = o_p(n^{-1/2}) + O_p(h^p) \ p \text{ even},$$

and conditional variance $var(\hat{\theta}) = O_p((nmh^3)^{-1} + (nh)^{-1})$ *if* $\omega(a_0) \neq 0$ *or* $\omega(b_0) \neq 0$, $var(\hat{\theta}) = O_p((nmh^3)^{-1} + n^{-1})$ *if* $\omega(a_0) = \omega(b_0) = 0$.

If $\omega(a_0) = \omega(b_0) = 0$ *and* $mh^3 \to \infty$,

$$var(\hat{\theta}) = \frac{\sigma^2}{n} A_F^{-1}[B_F - (C_F + C_F^T)]A_F^{-1}, \tag{2.7}$$

with $A_F = \int [F_\theta * F_{\theta^T} * \omega * f * f_g](t)dt$, $B_F = \int [(\omega * f_g * F_\theta)' * (\omega * f_g * F_{\theta^T})' * f](t)dt$ *and* $C_F = \int [(f' + f * F_X) * \omega * f_g * F_\theta * \{\omega * f_g * F_{\theta^T}\}'](t)dt$.

We use the shorthand notations $[f*g](t) = f(t)g(t)$, $F_X(t) = [\frac{\partial}{\partial X} F(X;\theta)](t) = \frac{\partial}{\partial X} F(X;\theta)|_{X=X(t)} = D_X(X;\theta)|_{X=X(t)}$, $F_\theta(t) = [\frac{\partial}{\partial \theta} F(X;\theta)](t) = \frac{\partial}{\partial \theta} F(X;\theta)|_{X=X(t)}$, and $F_{\theta^T}(t) = [F_\theta(t)]^T$. The proof outline of Theorem 1 is given in the Appendix and details are provided in the online supplementary materials.

We have used the random design of time points $t_1, \ldots, t_n$ and $t_1^*, \ldots, t_m^*$ in Theorem 1. We can also consider a fixed design with $\int_{t_0}^{t_i} f(t)dt = (i-1)/(n-1)$ and $\int_{t_0}^{t_k^*} f_g(t)dt = (k-1)/(m-1)$ for $i = 1, \ldots, n$ and $k = 1, \ldots, m$. The proof for the fixed design case is similar, but more tedious. For local polynomial regression for p odd, the asymptotic bias and variance are the same under random design and fixed design (Fan and Gijbels (1996, p.68)). We expect this is true under our model setting. But, in any case, $\hat{\theta}$ converges at the parametric $n^{-1/2}$ rate when $h = o(n^{-1/2p})$ and $m^{-1}h^{-3} = o(1)$. The function $\hat{\alpha}_k$ is still estimated at a nonparametric rate which is slower than $n^{-1/2}$, a result similar to those in Brunel (2008), Liang and Wu (2008, 2010), and Fang, Wu and Zhu (2011).

The result of this theorem is also similar to the one-step maximum likelihood approximation (Theorem 4.3 of Lehmann and Casella (1998)) in some sense. The one-step Newton-Raphson iteration of the likelihood equation starting at a $n^{-1/2}$-consistent estimator results in a more efficient estimator. Here our one-step Gauss-Newton iteration for maximizing (2.5) starting at $n^{-\delta}$ rate estimator ($\delta > 1/4$) result in a new $n^{-1/2}$ rate estimator for θ. Our one-iteration estimator (2.6) is a linear estimator that improves Liang-Wu's pseudo-least squares (PsLS) estimator (Liang and Wu (2008, 2010)). The initial estimator for the α_k's can be taken as the smoothing estimator without using the differential equation information. This one iteration starting from the Liang-Wu's PsLS estimator $\hat{\theta}$ and a local polynomial estimator for the α_k's results in a linear estimator with a

$n^{-1/2}$ rate for θ. The linearity may also be useful for extension of our method to mixed-effects differential equation models for longitudinal data (Fang, Wu and Zhu (2011)).

Remark 1. From the theorem, for a small enough $h = o(n^{-1/(2p)})$, the bias for the PsLS estimator and the proposed estimator are of order $o_p(n^{-1/2})$, so the variance dominates the mean squared error of the two estimators. For $\omega(a_0) = \omega(b_0) = 0$, and if m is chosen to be large enough so that $mh^3 \to \infty$, then we have an explicit expression (2.7) for the variance of our proposed estimator. Consider the case of uniformly distributed t_is and t_k^*s on $[0, 1]$. Here $f(t) = f_g(t) = 1$, and $var(\hat{\theta})$ is

$$\frac{\sigma^2}{n} A_F^{-1} \Big(\int [(\omega * F_\theta)' * (\omega * F_{\theta^T})' - F_X * \omega * \{F_\theta * (\omega * F_{\theta^T})' + (\omega * F_\theta)' * F_{\theta^T}\}](t)dt \Big) A_F^{-1},$$

where A_F is now $\int [F_\theta * F_{\theta^T} * \omega](t)dt$. The variance of Liang-Wu PsLS estimator has the extra term $(\sigma^2/n)A_F^{-1} \int [(\omega * F_X * F_\theta)' * (\omega * F_X * F_{\theta^T})'](t)dt A_F^{-1}$, a positive semi-definite matrix.

Remark 2. As long as h^p is of smaller order than $n^{-1/2}$, the order p of the polynomial is not important. As p increases, there are more terms in (2.5) and the computational burden increases. A value of $p = 1$ or $p = 2$ would be preferred in practice. In numerical studies we used $p = 2$, the same as that in Liang-Wu's method (2008) for the sake of comparison.

3. Numerical Studies

We compared the performance of the proposed method with Liang-Wu's method (2008), the method of Ramsay et al. (2007), and the nonlinear least squares estimator via Monte Carlo simulations. In addition, we applied the proposed method to a data set on immune cell trafficking for influenza infection to illustrate the usefulness of the proposed method. We measure performance of estimators by their average relative error

$$ARE = \sum_{i=1}^{r} |\frac{\hat{\theta}_i}{\theta} - 1|,$$

with $\hat{\theta}_i$ as the estimate for θ in the ith simulation runs with $i = 1, 2, \ldots, r$. Computational cost and convergence were also considered in evaluating the methods.

Since Liang-Wu's pseudo-least squares estimator and our estimator are computationally efficient, they can be used as the starting point for the nonlinear least squares estimator. This hybrid strategy may enjoy both the computational efficiency of the former and the high estimation accuracy of the latter. We also evaluated the performance of the hybrid approaches in our simulation studies.

Example 1. We simulated the data from the FitzHugh-Nagumo system of differential equations that were originally used to model the behavior of spike potentials in the giant axon of squid neurons, in FitzHugh (1961) and Nagumo, Arimoto, and Yoshizawa (1962). This model was also used for simulation studies by Ramsay et al. (2007) and Liang and Wu (2008). The FitzHugh-Nagumo system can be written as

$$\begin{aligned} \frac{d}{dt}X_1 &= (X_1 + X_2 - \frac{X_1^3}{3})c, \\ \frac{d}{dt}X_2 &= -\frac{X_1 - a + bX_2}{c}, \end{aligned} \tag{3.1}$$

with true parameter values $\theta = (a, b, c) = (0.34, 0.2, 3)$ in our simulations. We assumed that X_1 and X_2 were measured over a grid of $n = 51$ equally-spaced time points in $[0, 20]$ with measurement errors as in (2.1), with (σ_1, σ_2) being $(0.1, 0.1)$, $(0.1, 0.3)$, $(0.3, 0.1)$ or $(0.3, 0.3)$ for the measurement standard errors for X_1 and X_2 respectively. Thus, we obtained $n = 51$ data points. For each data set, we applied the proposed method and other existing methods to obtain: the nonlinear least squares estimator $\hat{\theta}^{NLS}$, Ramsay et al. (2007)'s collocation estimator $\hat{\theta}^{col}$, Liang-Wu's pseudo-least squares estimators $\hat{\theta}^{PLS}$ with $m = n$ grid points. The estimators used a common starting value of θ that was uniformly distributed on a cube centered at true value $(0.34, 0.2, 3)$ with one corner at $(0, 0, 0)$. The proposed new estimator $\hat{\theta}^{new}$ with $m = n$ was calculated by (2.6) starting at $\hat{\theta}^{PLS}$. For the hybrid approach, we found the nonlinear least squares estimator $\hat{\theta}^{NLS}_{PLS}$, $\hat{\theta}^{NLS}_{new}$, and $\hat{\theta}^{NLS}_{col}$ using respectively, $\hat{\theta}^{PLS}$, $\hat{\theta}^{new}$ and $\hat{\theta}^{col}$ as the starting points. For the Liang-Wu pseudo-least squares estimator and the proposed estimator, local quadratic polynomial smoothing was used and the piecewise linear weight function suggested in Brunel (2008) was used: $w(t) = 1$ for $1 \le t \le 19$; $w(t) = t$ for $0 \le t \le 1$; $w(t) = 20 - t$ for $19 \le t \le 20$. The collocation estimator $\hat{\theta}^{col}$ was implemented using the R package *CollocInfer* (Hooker, Xiao and Ramsay (2010)) with 51 equally-spaced knots between $t = 0$ and $t = 20$. The smoothing parameter for the collocation estimator was chosen as in the FitzHugh-Nagumo system demo example in the package. Otherwise we used the bandwidth recommended by Liang and Wu (2008): $\hat{h}_{opt} \times n^{-3/35}(\log n)^{-1/16}$. Here $\hat{h}_{opt}$ is the optimal bandwidth for the local polynomial fitting without the differential equation constraints, and we calculated it from the R package 'lokern' (Maechler (2010)). The methods were coded in R and run on the same computer.

Table 1 summarizes the average relative errors and computing times of the various estimators based on $r = 400$ simulation runs. Liang-Wu's pseudo-least squares estimator is always most computationally efficient and always converges, but its estimation accuracy in the sense of average relative errors is relatively

Table 1. Performance (ARE) of the estimators for Example 1 with $n = 51$ observations: $\hat{\theta}^{NLS}$=nonlinear least squares estimate using a random starting point; The collocation estimate $\hat{\theta}^{col}$ using the same starting point; The pseudo-least squares estimate $\hat{\theta}^{PLS}$ using the same starting point; Our estimate $\hat{\theta}^{new}$ started from $\hat{\theta}^{PLS}$; The nonlinear least squares estimate $\hat{\theta}^{NLS}_{PLS}$ started from $\hat{\theta}^{PLS}$; The nonlinear least squares estimate θ^{NLS}_{new} started from $\hat{\theta}^{new}$; The nonlinear least squares estimate θ^{NLS}_{col} started from $\hat{\theta}^{col}$.

(σ_1, σ_2)		parameter	Estimators						
			$\hat{\theta}^{NLS}$	$\hat{\theta}^{col}$	$\hat{\theta}^{PLS}$	$\hat{\theta}^{new}$	$\hat{\theta}^{NLS}_{PLS}$	$\hat{\theta}^{NLS}_{new}$	$\hat{\theta}^{NLS}_{col}$
(0.1,0.1)	ARE	a	2.58	7.70	4.26	5.42	1.96	1.75	1.75
		b	14.1	58.2	19.64	20.71	12.5	11.95	11.89
		c	2.32	6.07	26.82	21.16	0.69	0.37	0.37
	diverge		36.75	6.00	0	0	4.25	0.50	1.75
	time		11.59	12.07	0.17	0.24	8.17	7.01	20.12
(0.1,0.3)	ARE	a	6.73	10.34	6.85	8.06	3.72	2.49	2.49
		b	42.1	69.7	52.78	49.28	33.6	28.9	29.0
		c	9.08	7.73	33.95	22.31	2.44	0.55	0.56
	diverge		39.75	5.25	0	0	14.75	1.25	3.75
	time		10.93	12.89	0.18	0.25	9.11	8.14	21.01
(0.3,0.1)	ARE	a	5.21	13.61	10.30	8.60	5.26	4.97	5.03
		b	22.9	73.0	29.49	27.43	24.4	23.18	23.53
		c	1.91	7.71	34.08	21.70	1.42	1.05	1.06
	diverge		34.50	5.75	0	0	8.75	2.00	3.75
	time		13.08	14.00	0.18	0.25	12.62	11.66	25.11
(0.3,0.3)	ARE	a	6.12	13.58	11.44	10.73	5.68	5.44	5.49
		b	35.0	82.6	55.02	53.28	35.8	36.1	35.8
		c	4.21	9.04	43.33	24.44	2.44	1.41	1.49
	diverge		39.75	6.25	0	0	20.00	7.00	6.50
	time		12.75	14.73	0.18	0.25	12.06	10.99	26.18

poor. Our estimator, in comparison, improves the average relative errors for most cases at a small cost of computation. When used to initiate the nonlinear least squares estimate, it produces the best estimate in terms of the average relative errors in all the cases. The standard nonlinear least squares estimator with random starting points within the twice of the magnitudes of true parameter values is unstable, and has convergence problems. It can be significantly improved in computational cost and estimation accuracy if our estimator is used as the initial estimate. The collocation estimator has poor estimation accuracy and its computational cost is highest, in most cases, among all the methods. The comparison of computational cost here may need to be taken with a grain of salt as it is affected by the actual implementation procedure, in particular the selection of the penalty parameter. Additional simulation results from another nonlinear differential equation model are included in the online supplementary

materials; they show similar results. In the supplementary Table 3, we also reported the standard deviation (STD) of the estimators in Table 1. The trend and conclusions for the STD are similar to those for the AREs. Also included in the online supplementary materials are simulation results for evaluating the data augmentation size m. While our estimator's performance remains similar for larger m in most cases, increasing m does lead to improvement of the AREs in a few cases. However, for the NLS estimator θ_{new}^{NLS} using the proposed estimator as a starting point, the performance improvement is not significant. We recommend using $m = n$ when the proposed estimator is the starting point for the NLS estimator.

Example 2. We applied the method to a differential equation model for the growth and migration of influenza virus-specific effector CD8+ T cells among lymph node (T_E^m), spleen (T_E^s), and lung (T_E^l) of mice. The mechanistic differential equation model can be written as (Wu et al. (2011)),

$$\begin{aligned}
\frac{d}{dt}T_E^m &= [\rho_m D^m(t-\tau) - \delta_m]T_E^m - (\gamma_{ms} + \gamma_{ml})T_E^m, \\
\frac{d}{dt}T_E^s &= [\rho_s D^s(t-\tau) - \delta_s]T_E^s - \gamma_{sl}T_E^s + \gamma_{ms}T_E^m, \\
\frac{d}{dt}T_E^l &= \gamma_{ml}T_E^m + \gamma_{sl}T_E^s - \delta_l T_E^l,
\end{aligned} \tag{3.2}$$

where D^m denotes the number of mature dendritic cells in the mediastinal lymph node (MLN), D^s the number of mature dendritic cells in spleen; τ is the time delay of the effects of dendritic cells on CD8+ T cell proliferation; ρ_m and ρ_s are the proliferation rates of CD8+ T cells stimulated by per dendritic cell in MLN and spleen, respectively; δ_m , δ_s, and δ_l are the disappearance rates in MLN, spleen, and lung, respectively; γ_{ms} is the migration rate from MLN to spleen, γ_{ml} the migration rate from MLN to lung, and γ_{sl} the migration rate from spleen to lung. For this system, a total of $n = 77$ data points at 9 distinct time points for each of the three state variables, (T_E^m, T_E^s, T_E^l), are available (see Figure 1). The data for D^m are also available. In the analysis, the data of D^s are not available and are assumed to follow a similar pattern as D^m, see Wu et al. (2011) in this connection. The smoothed estimates of D^m were used in the analysis. More details can be found in Wu et al. (2011).

To stabilize the measurement error variance, a logarithm transformation is applied. If $X = (X_1, X_2, X_3)^\tau = (\log(T_E^m), \log(T_E^s), \log(T_E^l))^\tau$, the differential equations can be re-expressed as

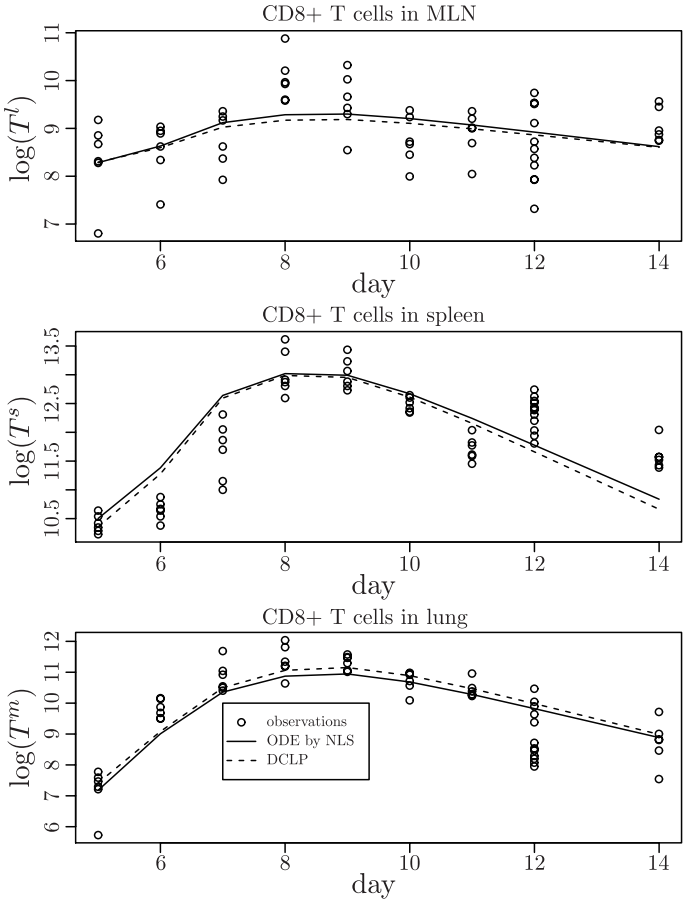

Figure 1. Data of influenza-specific CD8+ T cells in MLN, spleen, and lung and the corresponding fitted curves.

$$\begin{aligned}
\frac{d}{dt}X_1 &= \rho_m D^m(t-\tau) - \delta_m - \gamma_{ms} - \gamma_{ml},\\
\frac{d}{dt}X_2 &= \rho_s D^s(t-\tau) - \delta_s - \gamma_{sl} + \gamma_{ms}\exp(X_1 - X_2),\\
\frac{d}{dt}X_3 &= \gamma_{ml}\exp(X_1 - X_3) + \gamma_{sl}\exp(X_2 - X_3) - \delta_l.
\end{aligned} \tag{3.3}$$

Model (3.3) was fitted to data from day 5 to day 14 since the influenza-specific CD8+ T cells are not yet produced in Days 0-5. The time delay was set to $\tau = 3.08$ days, and parameters δ_m, δ_s and γ_{ml} were set to zero in Wu et al. (2011).

We applied our estimation method with a piece-wise linear weight function: $w(t) = 1$ for $6 \le t \le 13$; $w(t) = t - 5$ for $5 \le t \le 6$; $w(t) = 14 - t$ for $13 \le t \le 14$, as suggested by Brunel (2008). For comparisons, we also obtained Liang-Wu's

Table 2. The estimated parameter values by different procedures for CD8+ T cells data analysis. PsLS denotes the pseudo-least squares estimate; DCLP denotes the proposed differential equation constrained local polynomial estimate; NLS denotes the nonlinear least squares estimate; DCLP-NLS denotes the nonlinear least squares estimate started from the proposed estimate.

Parameters	**Estimation Methods**			
	PsLS	DCLP	NLS	DCLP-NLS
$T_E^m(5)$	$4.23E+3$	$4.23E+3$	$3.96E+3$	$3.96E+3$
$T_E^s(5)$	$3.33E+3$	$3.33E+3$	$3.64E+3$	$3.66E+3$
$T_E^l(5)$	$13.1E+3$	$13.1E+3$	$13.1E+3$	$13.1E+3$
ρ_m	$1.95E-5$	$1.46E-5$	$1.66E-5$	$1.66E-5$
ρ_s	$2.18E-5$	$4.78E-5$	$4.48E-5$	$4.47E-5$
δ_l	$1.53E-29$	3.96	3.96	3.97
γ_{ms}	$1.41E-1$	$1.38E-1$	$1.57E-1$	$1.57E-1$
γ_{ms}	$4.17E-5$	$6.11E-1$	$4.95E-1$	$4.96E-1$
residual sum of squares	112.4	18.48	15.77	15.77
average time	0.32	0.78	10.81	5.52

pseudo-least squares estimates and the nonlinear least squares estimates. A grid search was used to obtain the proposed differential equation constrained local polynomial estimates and the nonlinear least squares estimates. We report the results of parameter estimates for these estimation methods in Table 2 and fitted curves in Figure 1.

Our differential equation constrained local polynomial estimates of kinetic parameters are much closer to the nonlinear least squares estimates than the Liang-Wu's pseudo-least squares estimates, while both save computation. With our estimate as the starting point for the nonlinear least squares estimate, we achieved the convergence in approximately half the time that the original nonlinear least squares algorithm took, another benefit of the proposed differential equation constrained local polynomial approach.

4. Concluding Remarks

We propose a new estimation method for differential equation parameters based on the differential equation constraint local polynomial regression with a goal for improving the Liang-Wu's pseudo-least squares estimate. We investigated the asymptotic properties and finite-sample behaviors of the proposed method. Our simulation studies and data analysis showed that the proposed new estimator is clearly better than the Liang-Wu's pseudo-least squares estimator in estimation accuracy at a small cost of computation. Due to their computational efficiency, the pseudo-least squares estimator and the new estimator can be used as the starting point for the more refined nonlinear least squares estimate, and the new estimator is better for this purpose. Our simulation results also

demonstrate that the Ramsay et al. (2007)'s collocation method is more stable and can improve the estimation accuracy of the nonlinear least squares estimate significantly, but it cannot achieve the estimation accuracy of the nonlinear least squares estimate without using our proposed estimator as the starting point, and its computational cost is highest among all the methods in our simulation studies.

Lu, Liang, Li and Wu (2011) show that computationally efficient methods such as the pseudo-least squares estimate are useful in high-dimensional differential equation models where nonlinear least squares often fails. We expect that our approach can also improve the performance of the pseudo-least squares estimates in high-dimensional cases. This is a future research topic.

The accuracy of our estimate is not up to that of the nonlinear least squares estimate, and it requires the measurement of all state variables. It can be adapted to deal with latent state variables at the cost of computation; Careful investigations are needed to evaluate the trade-off between additional cost and benefits. The selection of the optimal bandwidth (h) and the data augmentation size (m) remains an open question. We followed the recommendations in Liang and Wu (2008) for bandwidth selection and this worked well in our numerical studies. The selection of m for data-augmentation is apparently not very critical. In theory, we can select m as large as possible, subject to the additional computational cost.

Ours is a linearized estimator in contrast to a nonlinear estimator such as Liang-Wu's pseudo-least squares estimator. Hence it is possible to extend the approach to population differential equation models (ODE). Longitudinal dynamic (random coefficient) ODE models have been suggested by Putter et al. (2002), Huang and Wu (2006), and Huang, Liu, and Wu (2006), in which the hierarchical Bayesian approach is used to estimate population dynamic parameters in HIV dynamic models from longitudinal clinical data. Li et al. (2002) proposed a spline-enhanced population model to study pharmacokinetics using a random time-varying coefficient ODE model. Guedj, Thiébaut, and Commenges (2007) used the maximum likelihood approach to directly estimate unknown parameters in random coefficient ODE models. Fang, Wu and Zhu (2011) extended the two-stage estimation method to random coefficient ODE models for longitudinal data. However, the extension of the differential equation constrained local polynomial estimator to the population mixed-effects ODE model is not trivial and remains an open research topic.

Acknowledgement

We appreciate helpful discussions with Dr. Hongqi Xue. This work is partially supported by NIH grants, HHSN272201000055C, RO1 AI087135, and the University of Rochester CTSI(RR024160) Pilot Award.

Appendix

Proof of Theorem 1. We analyze the order of estimation errors similar to the usual derivations of local polynomial regression; For example, see section 3.7 in Fan and Gijbels (1996). Let $S_{k,j} = \sum_{i=1}^{n} K_h(t_i - t_k^*)(t_i - t_k^*)^j$. Then

$$\begin{aligned} S_{k,j} &= nh^j f(t_k^*)\mu_j(K)[1 + o_p(1)] \quad j \text{ even}, \\ S_{k,j} &= nh^{j+1} f'(t_k^*)\mu_{j+1}(K)[1 + o_p(1)] \quad j \text{ odd}, \end{aligned} \tag{A.1}$$

where $f(t)$ is the density at t and $\mu_j(K) = \int K(u)u^j du$.

We consider properties of the estimator $\hat{\xi} = (J^T W J)^{-1} J^T W \tilde{Y}$ in (2.6). Since $G_{i,k}(\xi)$ only depends on (α_k, θ), the Jacobian matrix J is sparse with many zero elements:

$$J = \begin{pmatrix} \widetilde{DX}_{1,1} & \dots & 0 & \widetilde{D\theta}_{1,1} \\ \vdots & \dots & \vdots & \vdots \\ \widetilde{DX}_{n,1} & \dots & 0 & \widetilde{D\theta}_{n,1} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & \widetilde{DX}_{1,m} & \widetilde{D\theta}_{1,m} \\ \vdots & \dots & \vdots & \vdots \\ 0 & \dots & \widetilde{DX}_{n,m} & \widetilde{D\theta}_{n,m} \end{pmatrix},$$

where $\widetilde{DX}_{i,k} = 1 + \sum_{j=1}^{p}((t_i - t_k^*)^j/j!)D_{X,k}^{(j-1)}$ and $\widetilde{D\theta}_{i,k} = \sum_{j=1}^{p}((t_i - t_k^*)^j/j!)D_{\theta^T,k}^{(j-1)}$ with

$$\begin{aligned} D_{X,k}^{(j)} &= D_X^{(j)}(\alpha_k^*; \theta^*), \\ D_{\theta^T,k}^{(j)} &= D_{\theta^T}^{(j)}(\alpha_k^*; \theta^*) = \left(\frac{\partial}{\partial \theta_1} F^{(j)}(\alpha; \theta), \dots, \frac{\partial}{\partial \theta_q} F^{(j)}(\alpha; \theta) \right)_{\alpha = \alpha_k^*, \theta = \theta^*}. \end{aligned}$$

Since p is fixed, $\widetilde{DX}_{i,k}$ and $\widetilde{D\theta}_{i,k}$ are sums of fixed numbers of terms. Since by (A.1), the kernel sums of $(t_i - t_k^*)^j$ is at most of order $O_p(nh^j)$, the error analysis often focuses only on the lowest power term in $\widetilde{DX}_{i,k}$ and $\widetilde{D\theta}_{i,k}$, 1 and $(t_i - t_k^*)D_{\theta^T,k}^{(0)}$ respectively.

Direct calculation shows that

$$J^T W J = \begin{pmatrix} D_{m \times m} & L_{m \times q} \\ L_{q \times m}^T & C_{q \times q} \end{pmatrix}. \tag{A.2}$$

The matrix D is a $m \times m$ diagonal matrix with entries

$$D_k = \sum_{i=1}^{n} K_h(t_i - t_k^*)\omega(t_k^*)(\widetilde{DX}_{i,k})^2, \quad k = 1, \dots, m. \tag{A.3}$$

The kth row of the L matrix is

$$L_k = \sum_{i=1}^{n} K_h(t_i - t_k^*)\omega(t_k^*)\widetilde{DX}_{i,k}\widetilde{D\theta}_{i,k}, \tag{A.4}$$

and

$$C = \sum_{k=1}^{m}\sum_{i=1}^{n} K_h(t_i - t_k^*)\omega(t_k^*)\widetilde{D\theta}_{i,k}^T\widetilde{D\theta}_{i,k}. \tag{A.5}$$

Lemma 1. $D_k = n\omega(t_k^*)f(t_k^*) + o_p(n)$,

$$L_k = nh^2\mu_2(K)\omega(t_k^*)[f'(t_k^*)D^{(0)}_{\theta^T,k} + f(t_k^*)D^{(0)}_{X,k}D^{(0)}_{\theta^T,k}] + o_p(nh^2),$$

and $C = nmh^2\mu_2(K)A_F + o_p(nmh^2)$.

The definition of A_F is given under (2.7).

Proof of Lemma 1. The proof follows direct calculations using $\widetilde{DX}_{i,k} = 1 + \sum_{j=1}^{p}(t_i - t_k^*)^j D^{(j-1)}_{X,k}/j!$, $\widetilde{D\theta}_{i,k} = \sum_{j=1}^{p}(t_i - t_k^*)^j D^{(j-1)}_{\theta^T,k}/j!$, and (A.1). Here $\widetilde{DX}_{i,k}$ has $p+1$ terms, each of the form of powers $(t_i - t_k^*)^j$, multiplied by a bounded quantity. So D_k by (A.3) is the sum of $(p+1)^2$ terms, each of the form $S_{k,j} = \sum_{i=1}^{n} K_h(t_i - t_k^*)(t_i - t_k^*)^j$, multiplied by a bounded quantity. Thus

$$D_k = [S_{k,0} + \sum_{j=1}^{p} S_{k,j}(\frac{D^{(j-1)}_{X,k}}{j!}) + \sum_{l=1}^{p}(\frac{D^{(l-1)}_{X,k}}{l!})(S_{k,l} + \sum_{j=1}^{p} S_{k,l+j}\frac{D^{(j-1)}_{X,k}}{j!})]\omega(t_k^*).$$

For fixed p, $m \to \infty$ and $n \to \infty$, asymptotically D_k is the term with highest order among the $(p+1)^2$ terms. The leading term is $S_{k,0}\omega(t_k^*) = n\omega(t_k^*)f(t_k^*) + o_p(n)$ by (A.1). The rest of terms are of order $S_{k,j}$ for some $j \geq 1$, so are of order $O_p(nh^j)$ or $O_p(nh^{j+1})$, and at most of order $O_p(nh^2) = o_p(n)$. Hence the sum $D_k = n\omega(t_k^*)f(t_k^*) + o_p(n)$ is of order $O_p(n)$.

Similarly, the leading terms in $\widetilde{DX}_{i,k}\widetilde{D\theta}_{i,k}$ and $\widetilde{D\theta}_{i,k}^T\widetilde{D\theta}_{i,k}$ give the results for L_k and C. More detailed analysis can be found in the online supplemental materials. This finishes the proof of Lemma 1.

It is easy to check that

$$(J^TWJ)^{-1} = \begin{pmatrix} D_{m\times m} & L_{m\times q} \\ L^T_{q\times m} & C_{q\times q} \end{pmatrix}^{-1} = \begin{pmatrix} D^{-1} + D^{-1}LV^{-1}L^TD^{-1} & -D^{-1}LV^{-1} \\ -V^{-1}L^TD^{-1} & V^{-1} \end{pmatrix} \tag{A.6}$$

with $V = C - L^TD^{-1}L$. The order of quantities in (A.6) is given in a lemma whose proof is provided in the online supplemental materials.

Lemma 2. $L^TD^{-1}L = O_p(mnh^4)$, $V^{-1} = C^{-1}[1 + O_p(h^2)] = O_p(1/nmh^2)$, $D^{-1}LV^{-1} = O_p(1/mn)$ *and* $D^{-1}LV^{-1}L^TD^{-1} = O_p(h^2/n)$.

Using the results in Lemma 1 and 2,

$$(J^TWJ)^{-1} = \begin{pmatrix} D^{-1}_{m\times m} + o_p(\frac{1}{n}) & O_p(\frac{1}{mn})_{m\times q} \\ O_p(\frac{1}{mn})_{q\times m} & C^{-1}_{q\times q} + o_p(\frac{1}{mnh^2}) \end{pmatrix}, \tag{A.7}$$

where a matrix is of an order, such as $O((mn)^{-1})_{m\times q}$, when all its elements are of that order.

Remark 3. For a d-dimensional X, the order analysis of the matrices remains the same. The $D_{m\times m}$ matrix is then $D_{md\times md}$ with diagonal block matrices D_k of size $d\times d$, and the L_k are matrices of size $d\times q$. As d is fixed, the multiplication of matrices with dimension d instead of 1 does not change the order, and the proof extends to d-dimensional X.

A.1. Asymptotic bias

The bias of $\hat{\xi}$ given $t_1,\ldots,t_n,t_1^*,\ldots,t_m^*,\xi^*$ is

$$\begin{aligned} Bias(\hat{\xi}) &= (J^TWJ)^{-1}J^TWE(\tilde{Y}) - \xi_0 \\ &= (J^TWJ)^{-1}J^TW\{E(Y - G(\xi^*) + J\xi^*) - J\xi_0\} \\ &= (J^TWJ)^{-1}J^TW\{E(Y) - G(\xi^*) - J(\xi_0 - \xi^*)\}. \end{aligned}$$

Let $J = (J_{1,1}^T, J_{2,1}^T, \ldots, J_{n,1}^T, J_{1,2}^T, \ldots, J_{n,m}^T)^T$. The elements in $E(Y) - G(\xi^*) - J(\xi_0 - \xi^*)$ are those $E(Y_i) - G_{i,k}(\xi^*) - J_{i,k}(\xi_0 - \xi^*)$'s. With a Taylor expansion of $E(Y_i) = X(t_i)$ at time point $t = t_k^*$, we have

$$\begin{aligned} X(t_i) &= X(t_k^*) + \sum_{j=1}^{p} \frac{(t_i - t_k^*)^j}{j!} X^{(j)}(t_k^*) + (t_i - t_k^*)^{p+1} \frac{X^{(p+1)}(\tilde{t}_{i,k})}{(p+1)!} \\ &= G_{i,k}(\xi_0) + (t_i - t_k^*)^{p+1} \frac{X^{(p+1)}(\tilde{t}_{i,k})}{(p+1)!}, \end{aligned}$$

where $\tilde{t}_{i,k}$ is a point between t_k^* and t_i. Since $G_{i,k}(\xi_0) - G_{i,k}(\xi^*) - J_{i,k}(\xi_0 - \xi^*) = O_p(|\xi_0 - \xi^*|^2) = O_p(n^{-2\delta})$, we have

$$E(Y_i) - G_{i,k}(\xi^*) - J_{i,k}(\xi_0 - \xi^*) = (t_i - t_k^*)^{p+1} \frac{X^{(p+1)}(\tilde{t}_{i,k})}{(p+1)!} + O_p(n^{-2\delta}). \tag{A.8}$$

Let $T_j = ((t_1 - t_1^*)^j, \ldots, (t_n - t_1^*)^j, (t_1 - t_2^*)^j, \ldots, (t_n - t_m^*)^j)^T$. As in the proof of Lemma 1, we evaluate the order of J^TWT_j by focusing on the term with the lowest power of $(t_i - t_k^*)$. The first m elements in J^TWT_j are of the form $\sum_{i=1}^n K_h(t_i - t_k^*)(t_i - t_k^*)^j\omega(t_k^*)\widetilde{DX}_{i,k}$, $k = 1,\ldots,m$. The lowest power term in $\widetilde{DX}_{i,k}$ is 1 so that the m elements are of the same order as $S_{k,j} = \sum_{i=1}^n K_h(t_i - t_k^*)(t_i - t_k^*)^j$, which is $O_p(nh^j)$ for p even, and $O_p(nh^{j+1})$ for p odd, by (A.1).

The last q elements in J^TWT_j are $\sum_{k=1}^m[\sum_{i=1}^n K_h(t_i-t_k^*)(t_i-t_k^*)^j\omega(t_k^*)\widetilde{D\theta}_{i,k}]$. Again, the lowest power term in $\widetilde{D\theta}_{i,k}$ is $(t_i-t_k^*)$ so that the last q elements are of the same order as $\sum_{k=1}^m[\sum_{i=1}^n K_h(t_i-t_k^*)(t_i-t_k^*)^{j+1}]=\sum_{k=1}^m S_{k,j+1}$: of order $O_p(mnh^{j+2})$ for p even, and $O_p(mnh^{j+1})$ for p odd by (A.1). We have

$$J^TWT_j=\begin{pmatrix}O_p(nh^j)_{m\times 1}\\ O_p(mnh^{j+2})_{q\times 1}\end{pmatrix}\text{ for } j \text{ even},\quad \begin{pmatrix}O_p(nh^{j+1})_{m\times 1}\\ O_p(mnh^{j+1})_{q\times 1}\end{pmatrix}\text{ for } j \text{ odd}.$$

From (A.8), $E(Y)-G(\xi^*)-J(\xi_0-\xi^*)=T_{p+1}O_p(1)+T_0O_p(n^{-2\delta})$. Plug-in the orders of J^TWT_{p+1} and J^TWT_0, we have that $J^TW\{E(Y)-G(\xi^*)-J(\xi_0-\xi^*)\}$ is

$$\begin{pmatrix}O_p(n(h^{p+1}+n^{-2\delta}))_{m\times 1}\\ O_p(mnh^2(h^{p+1}+n^{-2\delta}))_{q\times 1}\end{pmatrix}\text{ for } p \text{ odd};\ \begin{pmatrix}O_p(n(h^{p+2}+n^{-2\delta}))_{m\times 1}\\ O_p(mnh^2(h^{p}+n^{-2\delta}))_{q\times 1}\end{pmatrix}\text{ for } p \text{ even}.$$

Combining this with (A.6) and Lemma 2, the bias in estimating θ is

$$\begin{aligned}Bias(\hat\theta)&=O_p(n^{-2\delta})+O_p(h^{p+1})\quad p \text{ odd},\\ Bias(\hat\theta)&=O_p(n^{-2\delta})+O_p(h^{p})\quad p \text{ even}.\end{aligned}$$

Since $\delta>-1/4$, $Bias(\hat\theta)=o_p(n^{-1/2})$ for $h=o(n^{-1/2p})$.

A.2. Asymptotic variance

For the variance of $\hat\xi$ given $t_1,\dots,t_n,t_1^*,\dots,t_m^*,\xi^*$, notice that $var(\hat\xi)=(J^TWJ)^{-1}J^TWvar(\tilde Y)WJ(J^TWJ)^{-1}$. Let $\Sigma=var((Y_1,\dots,Y_n)^T)=diag\{\underbrace{\sigma^2,\dots,\sigma^2}_{n}\}$, $\Sigma_Y=var(\tilde Y)=var(Y)$ are simply $m\times m$ blocks of Σ. Thus,

$$J^TWvar(\tilde Y)WJ=\begin{pmatrix}D^*_{m\times m} & L^*_{m\times q}\\ (L^*)^T_{q\times m} & C^*_{q\times q}\end{pmatrix},\tag{A.9}$$

where the (k,j)th element in D^* is

$$D^*_{k,j}=\sigma^2\omega(t_k^*)\omega(t_j^*)[\sum_{i=1}^n K_h(t_i-t_k^*)K_h(t_i-t_j^*)\widetilde{DX}_{i,k}\widetilde{DX}_{i,j}],\ \text{for } k,j=1,\dots,m,\tag{A.10}$$

the kth row in L^* is

$$L^*_k=\sigma^2\omega(t_k^*)[\sum_{j=1}^m\omega(t_j^*)\sum_{i=1}^n K_h(t_i-t_k^*)K_h(t_i-t_j^*)\widetilde{DX}_{i,k}\widetilde{D\theta}_{i,j}],\ \text{for } k=1,\dots,m,\tag{A.11}$$

and

$$C^*=\sum_{k=1}^m\sum_{j=1}^m\sigma^2\omega(t_k^*)\omega(t_j^*)[\sum_{i=1}^n K_h(t_i-t_k^*)K_h(t_i-t_j^*)\widetilde{D\theta}_{i,k}^T\widetilde{D\theta}_{i,j}].\tag{A.12}$$

Lemma 3.

$$D^*_{k,k} = O_p(nh^{-1}), \qquad D^*_{k,j} = o_p(n) \qquad for\ k \neq j. \tag{A.13}$$

$$L^*_k = nmh^2\sigma^2\mu_2(K)[\omega * f * \{\omega * f_g * F_{\theta^T}\}'](t^*_k) + o_p(nmh^2). \tag{A.14}$$

When $\omega(a_0) \neq 0$ *or* $\omega(b_0) \neq 0$, $C^* = O_p(nmh+nm^2h^3)$, *when* $\omega(a_0) = \omega(b_0) = 0$, $C^* = O_p(nmh + nm^2h^4)$, *and when* $\omega(a_0) = \omega(b_0) = 0$ *and* $mh^3 \to \infty$,

$$C^* = nm^2h^4\sigma^2[\mu_2(K)]^2B_F + o_p(nm^2h^4). \tag{A.15}$$

The proof of Lemma 3 is given in the online supplemental materials. Using (A.6), we find $var(\hat{\theta})$ as

$$V^{-1}L^TD^{-1}D^*D^{-1}LV^{-1} - V^{-1}(L^*)^TD^{-1}LV^{-1} - V^{-1}L^TD^{-1}L^*V^{-1} + V^{-1}C^*V^{-1}. \tag{A.16}$$

The order can be calculated using the results in Lemmas 1, 2, and 3. When $\omega(a_0) = \omega(b_0) = 0$, the first term in (A.16) is of smaller order $O(1/nmh)$ and ignored. When $mh^3 \to \infty$, we see from the other three terms that $var(\hat{\theta}) = O_p(1/n)$. Using Lemma 3, the last term in (A.16) is $V^{-1}C^*V^{-1} = (\sigma^2/n)$ $A_F^{-1}B_FA_F^{-1} + o(1/n)$, and using Lemmas 1 and 2, the third term is $-V^{-1}$ $L^TD^{-1}L^*V^{-1} = -(\sigma^2/n)A_F^{-1}C_FA_F^{-1} + o_p(1/n)$. The second term in (A.16) is the transpose of the third term. Combining, we have

$$var(\hat{\theta}) = \frac{\sigma^2}{n}A_F^{-1}[B_F - (C_F + C_F^T)]A_F^{-1} + o_p(\frac{1}{n}).$$

If $\omega(a_0) \neq 0$ or $\omega(b_0) \neq 0$, similar calculation using Lemma 3 shows that the variance of $\hat{\theta}$ is of order $O_p(1/nmh^3 + 1/nh)$.

More details are provided in the online supplemental materials.

References

Brunel, N. J-B. (2008). Parameter estimation of ODEs via nonparametric estimators. *Electronic J. Statist.* **2**, 124-1267.

Chen, J. and Wu, H. (2008). Efficient local estimation for time-varying coefficients in deterministic dynamic models with applications to HIV-1 dynamics. *J. Amer. Statist. Assoc.* **103**, 369-384.

Chen, T., He, H. L. and Church, G. M. (1999). Modeling gene expression with differential equations. *Pacific Symposium on Biocomputing* **4**, 29-40.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall/CRC.

Fang, Y., Wu, H. and Zhu, L. (2011). A two-stage estimation method for random coefficient differential equation models with application to longitudinal HIV dynamic data. *Statist. Sinica* **21**, 1145-1170.

FitzHugh, R. (1961). Impulses and physiological states in models of nerve membrane. *Biophysical J.* **1**, 445-466.

Guedj, J., Thiébaut, R. and Commenges, D. (2007). Maximum likelihood estimation in dynamical models of HIV. *Biometrics* **63**, 1198-1206.

Hooker, G., Xiao, L. and Ramsay, J. O. (2010). CollocInfer: Collocation Inference for Dynamic Systems. R package version 0.1.2. `http://CRAN.R-project.org/package=CollocInfer`.

Huang, Y., Liu, D. and Wu, H. (2006). Hierarchical Bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. *Biometrics* **62**, 413-423.

Huang, Y. and Wu, H. (2006). A Bayesian approach for estimating antiviral efficacy in HIV dynamic models. *J. Appl. Statist.* **33**, 155-174.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer.

Li, L., Brown, M. B., Lee, K. H. and Gupta, S. (2002). Estimation and inference for a spline-enhanced population pharmacokinetic model. *Biometrics* **58**, 601-611.

Liang, H. and Wu, H. (2008). Parameter estimation of differential equation models using a framework of measurement error in regression models. *J. Amer. Statist. Assoc.* **103**, 1570-1583.

Liang, H. and Wu, H. (2010). Correction on "Parameter estimation of differential equation models using a framework of measurement error in regression models (JASA, 103, 1570-1583)". *J. Amer. Statist. Assoc.* **105**, 1636.

Lu, T., Liang, H., Li, H. and Wu, H. (2011). High dimensional ODEs coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification. *J. Amer. Statist. Assoc.* **106**, 1242-1258.

Maechler, M. (2010). lokern: Kernel Regression Smoothing with Local or Global Plug-in Bandwidth. R package version 1.1-1. `http://CRAN.R-project.org/package=lokern`.

Nagumo, J. S., Arimoto, S. and Yoshizawa, S. (1962). An active pulse transmission line simulating a nerve axon. *Proc. the IRE* **50**, 2061-2070.

Poyton, A. A., Varziri, M. S., McAuley, K. B., McLellan, P. J. and Ramsay, J. O. (2006). Parameter estimation in continuous-time dynamic models using principal differential analysis. *Computers and Chemical Engineering* **30**, 698-708.

Putter, H., Heisterkamp, S. H., Lange, J. M. and De Wolf, F. (2002). A Bayesian approach to parameter estimation in HIV dynamical models. *Stat. Med.* **21**, 2199-2214.

Ramsay, J. O., Hooker, G., Campbell, D. and Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach (with discussion). *J. Roy. Statist. Soc. Ser. B* **69**, 741-796.

Wu, H., Kumar, A., Miao, H., Holden-Wiltse, J., Mosmann, T.R., Livingstone, A. M., Belz, G.T., Perelson, A. S., Zand, M. and Topham, D. J. (2011). Modeling of influenza-specific CD8+ T cells during the primary response indicates that the spleen is a major source of effectors. *J. Immunology* **187**, 4474-4482.

Department of Mathematics, Northeastern University, 360 Huntington Ave., Boston, MA 02115, U.S.A.

E-mail: a.ding@neu.edu

Department of Biostatistics and Computational Biology, University of Rochester School of Medicine and Dentistry, 601 Elmwood Avenue, Box 630, Rochester, New York 14642, U.S.A.

E-mail: Hulin_Wu@urmc.rochester.edu

(Received October 2012; accepted December 2013)