Online Supplementary Materials: Detailed mathematical proofs and simulation results for "Estimation of Ordinary Differential Equation Parameters Using Constrained Local Polynomial Regression," by A. Adam Ding and Hulin Wu

1 Model notations and the main theorem

A general nonlinear ordinary differential equation (ODE) model

$$\frac{dX(t)}{dt} = F\{X(t);\theta\} \tag{1}$$

is measured with noise at time points $t_1, t_2, ..., t_n$ with observations

$$Y_i = Y(t_i) = X(t_i) + e(t_i), \quad i = 1, ..., n. \tag{2}$$

Using differential equation constraints, we can calculate the higher-order derivatives

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

Hence we propose incorporating differential equations into the local polynomial regression on a grid of time points $t = t_1^*, t_2^*, ..., t_m^*$ by minimizing the objective function

$$\sum_{k=1}^{m}\sum_{i=1}^{n}[Y_i - \{\alpha_k + \sum_{j=1}^{p}\frac{F^{(j-1)}(\alpha_k;\theta)}{j!}(t_i - t_k^*)^j\}]^2 K_h(t_i - t_k^*)\omega(t_k^*), \tag{3}$$

with respect to $\xi = (\alpha_1, ..., \alpha_m, \theta)^T$, where $\omega(t_k^*)$ are nonnegative weights over the time grid. This provides estimates $\hat{\alpha}_k = \hat{X}(t_k^*)$ and $\hat{\theta}$ simultaneously.

For a general nonlinear function F of the differential equation model, the optimization of (3) becomes a nonlinear minimization problem, thus we may lose the computational efficiency of the original local polynomial fitting. To solve this problem, we consider a linear estimator that results from one iteration of the Gauss-Newton optimization of (3) at a previous estimate $\xi = (\alpha_1^*, ..., \alpha_m^*, \theta^*)^T$. In matrix notation, the objective function (3) is $\{Y - G(\xi)\}^T W \{Y - G(\xi)\}$, where $Y = (Y_1, ..., Y_n, ..., Y_1, ..., Y_n)^T$ is a (nm)-dimensional vector with the observations Y_i's repeated m times, $G = (G_{1,1}, ..., G_{n,1}, ..., G_{1,m}, ..., G_{n,m})^T$ 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 a $nm \times nm$ diagonal weight matrix, that is,

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

Let $J = (\partial G/\partial \alpha_1, ..., \partial G/\partial \alpha_m, \partial G/\partial \theta_1, ..., \partial G/\partial \theta_q)_{\xi=\xi^*}$ denote the $nm \times (m+q)$ Jacobian matrix evaluated at $\xi = \xi^*$. Then a Gauss-Newton iteration minimizes (3) with $G(\xi)$ replaced by its linear approximation $G(\xi^*) + J(\xi - \xi^*)$. This results in the weighted linear least squares estimator

$$\hat{\xi} = (J^T W J)^{-1} J^T W \tilde{Y}, \tag{4}$$

where $\tilde{Y} = Y - G(\xi^*) + J\xi^*$.

Theorem 1 *We assume the following technical conditions*

(1)The differential equation (1) holds over a time interval $[a_0, b_0]$ *and have a bounded solution* $X(t)$. *We observe* $Y_i(t)$ *from model (2) at* $t = t_i \in [a_0, b_0]$, $i = 1, ..., n$. *The differential equation parameters* θ *are jointly estimated with* $\alpha_i = X(t_i^*)$ *over a time grid* $t_i^* \in [a_0, b_0]$, $i = 1, ..., m$. *The resulting estimator* $\hat{\xi}$ *is given by (4) with the linearization at a starting value* $\xi^* = (\alpha_1^*, ..., \alpha_m^*, \theta^*)^T$.

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

(3) The function $F(x)$ *in differential equation (1) has bounded p-th 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. Denote the moments*

of K by $\mu_j(K) = \int K(u)u^j du$. Then $\mu_0(K) = \int K(u)du = 1$, and all odd-order moments $\mu_j(K) = 0$ vanish.

(6) The observation time points $t_1, ..., t_n$ and fitted time points $t_1^*, ..., t_m^*$ comes from distribution with densities $f(t)$ and $f_g(t)$, $t \in [a_0, b_0]$. 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 $t \in [a_0, b_0]$.

Then conditional on the observation time points $t_1, ..., t_n$, fitted time points $t_1^*, ..., t_m^*$ and ξ^*, the differential equation parameter estimator $\hat{\theta}$ has conditional bias

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

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

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

$$var(\hat{\theta}) = \frac{\sigma^2}{n} A_F^{-1}[B_F - (C_F + C_F^T)]A_F^{-1}, \tag{5}$$

with $A_F = \int [F_\theta \circ F_{\theta^T} \circ \omega \circ f \circ f_g](t)dt$, $B_F = \int [(\omega \circ f_g \circ F_\theta)' \circ (\omega \circ f_g \circ F_{\theta^T})' \circ f](t)dt$ and $C_F = \int [(f' + f \circ F_X) \circ \omega \circ f_g \circ F_\theta \circ \{\omega \circ f_g \circ F_{\theta^T}\}'](t)dt$. Here and in the following we use the shorthand notations $[f \circ g](t) = f(t)g(t)$, $F_X(t) = [\frac{\partial}{\partial X} F(X;\theta)](t) = \frac{\partial}{\partial X} F(X;\theta)|_{X=X(t)}$, $F_\theta(t) = [\frac{\partial}{\partial \theta} F(X;\theta)](t) = \frac{\partial}{\partial \theta} F(X;\theta)|_{X=X(t)}$ and $F_{\theta^T}(t) = [F_\theta(t)]^T$.

Remark: With h small enough $o(n^{-1/(2p)})$, the bias in $\hat{\theta}$ is of $o_p(n^{-1/2})$, so the variance dominates. With $\omega(a_0) = \omega(b_0) = 0$ and choosing m big enough so that $mh^3 \to \infty$, then the Theorem states that $\hat{\theta}$ converges at the parametric rate $n^{-1/2}$. Furthermore, we can see how does this constrained local polynomial estimator improves upon the pseudo-least square estimator with unconstrained local polynomial estimator for X and X' using the same bandwidth. Since the bandwidth is selected so that bias is $o_p(n^{-1/2})$, we just need to compare their variances. For simplicity, let use consider the case of uniformly distributed t_is and t_k^*s on time interval $[0,1]$. Then $f(t) = f_g(t) = 1$, and the variance $var(\hat{\theta})$ becomes

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

where A_F is now $\int [F_\theta \circ F_{\theta^T} \circ \omega](t)dt$. Compare to the variance of Liang-Wu estimator (Liang and Wu 2010), our $var(\hat{\theta})$ have one less term $\frac{\sigma^2}{n} A_F^{-1} \int [(\omega \circ F_X \circ F_\theta)' \circ (\omega \circ F_X \circ F_{\theta^T})'](t)dt A_F^{-1}$. This is a positive semi-definite matrix, meaning our variance is smaller. This extra term in Liang-Wu's estimator corresponding to the error propogated from first stage estimator $\hat{X}(t)$. Since we restrict our $\hat{X}'(t) = F(\hat{X}(t); \hat{\theta})$, our estimators $\hat{X}'(t)$ and $\hat{X}(t)$ are related and their errors enter into variance of $\hat{\theta}$ only once through the term $\int F_X \circ \omega \circ \{F_\theta \circ (\omega \circ F_{\theta^T})' + (\omega \circ F_\theta)' \circ F_{\theta^T}\}](t)dt$.

The detailed proof of Theorem 1 is given as follows.

A 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.) The order of some common quantities would be useful. Let $S_{k,j} = \sum_{i=1}^n K_h(t_i - t_k^*)(t_i - t_k^*)^j$. Then

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

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

To consider properties of the estimator $\hat{\xi} = (J^T W J)^{-1} J^T W \tilde{Y}$ in (4), we first study the matrix $(J^T W J)^{-1}$ and $J^T W$. 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 $\widetilde{DX}_{i,k} = 1 + \sum_{j=1}^{p} \frac{(t_i - t_k^*)^j}{j!} D_{X,k}^{(j-1)}$ and $\widetilde{D\theta}_{i,k} = \sum_{j=1}^{p} \frac{(t_i - t_k^*)^j}{j!} D_{\theta^T,k}^{(j-1)}$ with

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

Since p is fixed, $\widetilde{DX}_{i,k}$ and $\widetilde{D\theta}_{i,k}$ are sums of fixed number 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 later often only need to focus on the lowest power term in $\widetilde{DX}_{i,k}$ and $\widetilde{D\theta}_{i,k}$. That is, 1 and $(t_i - t_k^*) D_{\theta^T,k}^{(0)}$ respectively.

Hence,

$$WJ = \begin{pmatrix} \omega(t_1^*)K_h(t_1 - t_1^*)\widetilde{DX}_{1,1} & \dots & 0 & \omega(t_1^*)K_h(t_1 - t_1^*)\widetilde{D\theta}_{1,1} \\ \vdots & \dots & \vdots & \vdots \\ \omega(t_1^*)K_h(t_n - t_1^*)\widetilde{DX}_{n,1} & \dots & 0 & \omega(t_1^*)K_h(t_n - t_1^*)\widetilde{D\theta}_{n,1} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & \omega(t_m^*)K_h(t_1 - t_m^*)\widetilde{DX}_{1,m} & \omega(t_m^*)K_h(t_1 - t_m^*)\widetilde{D\theta}_{1,m} \\ \vdots & \dots & \vdots & \vdots \\ 0 & \dots & \omega(t_m^*)K_h(t_n - t_m^*)\widetilde{DX}_{n,m} & \omega(t_m^*)K_h(t_n - t_m^*)\widetilde{D\theta}_{n,m} \end{pmatrix}, \tag{A.2}$$

and

$$J^TWJ = \begin{pmatrix} D_{m\times m} & L_{m\times q} \\ L^T_{q\times m} & C_{q\times q} \end{pmatrix}, \tag{A.3}$$

where the subscripts of the four sub-matrices denotes their dimensions. The matrix D is a $m \times m$ diagonal matrix with entries

$$D_k = \sum_{i=1}^{n} K_h(t_i - t_k^*)\omega(t_k^*)(\widetilde{DX}_{i,k})^2, \quad k = 1, ..., m. \tag{A.4}$$

The k-th row of the L matrix is

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

and

$$C = \sum_{k=1}^{m}\sum_{i=1}^{n} K_h(t_i - t_k^*)\omega(t_k^*)\widetilde{D\theta}_{i,k}^T\widetilde{D\theta}_{i,k}. \tag{A.6}$$

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 was given under (5). We give the proof of Lemma 1 in the subsection A.3. It is easy to check by block matrix algebra that

$$(J^TWJ)^{-1} = \begin{pmatrix} D_{m\times m} & L_{m\times q} \\ L^T_{q\times m} & C_{q\times q} \end{pmatrix}^{-1} = \begin{pmatrix} D^{-1} + D^{-1}LV^{-1}L^TD^{-1} & -D^{-1}LV^{-1} \\ -V^{-1}L^TD^{-1} & V^{-1} \end{pmatrix} \tag{A.7}$$

with $V = C - L^TD^{-1}L$. The order of quantities in (A.7) is described in the following lemma whose proof is in subsection A.3.

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

Using the results in Lemma 1 and 2,

$$(J^TWJ)^{-1} = \begin{pmatrix} D^{-1}_{m\times m} + o_p(\frac{1}{n}) & O_p(\frac{1}{mn})_{m\times q} \\ O_p(\frac{1}{mn})_{q\times m} & C^{-1}_{q\times q} + o_p(\frac{1}{mnh^2}) \end{pmatrix}. \tag{A.8}$$

Remark: For a d-dimensional X, the order analysis of the matrices would all remain the same. The $D_{m\times m}$ matrix would become $D_{md\times md}$ with diagonal block matrices D_k of size $d\times d$. And L_k would be matrices of size $d\times q$. As d is fixed, the multiplying of matrices with dimension d instead of 1 does not change the order. So the whole proof can be extended to d-dimensional X straightforwardly.

A.1 Bias

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

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

Denote $J = (J_{1,1}^T, J_{2,1}^T, ..., J_{n,1}^T, J_{1,2}^T, ..., J_{n,m}^T)^T$. Hence 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 Taylor expansion of $E(Y_i) = X(t_i)$ at time point $t = t_k^*$, we have

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

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

$$E(Y_i) - G_{i,k}(\xi^*) - J_{i,k}(\xi_0 - \xi^*) = (t_i - t_k^*)^{p+1}\frac{X^{(p+1)}(\tilde{t}_{i,k})}{(p+1)!} + O_p(n^{-2\delta}). \tag{A.9}$$

Denote $T_j = ((t_1 - t_1^*)^j, ..., (t_n - t_1^*)^j, (t_1 - t_2^*)^j, ..., (t_n - t_m^*)^j)^T$. Similar to the analysis in proof of Lemma 1, we analyze the order of J^TWT_j by focusing on the term with lowest power of $(t_i - t_k^*)$ as higher power leads to smaller order kernel sum. From (A.2), the first m elements in J^TWT_j are of the form $\sum_{i=1}^n K_h(t_i - t_k^*)(t_i - t_k^*)^j\omega(t_k^*)\widetilde{DX}_{i,k}$, $k = 1, ..., m$. The lowest power term in $\widetilde{DX}_{i,k}$ is 1 (i.e., $(t_i - t_k^*)^0$) so that those 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 those last q elements are of the same order as $\sum_{k=1}^m [\sum_{i=1}^n K_h(t_i - t_k^*)(t_i - t_k^*)^{j+1}] = \sum_{k=1}^m S_{k,j+1}$. That is, they are of order $O_p(mnh^{j+2})$ for p even, and $O_p(mnh^{j+1})$ for p odd by (A.1). In summation,

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

From (A.9), $E(Y) - G(\xi^*) - J(\xi_0 - \xi^*) = T_{p+1}O_p(1) + T_0 O_p(n^{-2\delta})$. Plug-in the orders of $J^T W T_{p+1}$ and $J^T W T_0$, we have that $J^T W\{E(Y) - G(\xi^*) - J(\xi_0 - \xi^*)\}$ is

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

Combining this with (A.7) and Lemma 2, the bias $Bias(\hat{\theta})$, when p is odd, is

$$\begin{aligned} & -V^{-1}L^T D^{-1} O_p(n(h^{p+1} + n^{-2\delta}))_{m\times 1} + V^{-1} O_p(mnh^2(h^{p+1} + n^{-2\delta}))_{q\times 1} \\ = \; & O_p(\tfrac{1}{mn})_{q\times m} O_p(n(h^{p+1} + n^{-2\delta}))_{m\times 1} + O_p(\tfrac{1}{mnh^2})_{q\times q} O_p(mnh^2(h^{p+1} + n^{-2\delta}))_{q\times 1} \\ = \; & O_p(h^{p+1} + n^{-2\delta}), \end{aligned}$$

where in the last equality an extra m factor in the first term comes from product of the matrices of sizes $q \times m$ and $m \times 1$ while the second term need no extra factor as q is fixed.

When p is even, $Bias(\hat{\theta})$ becomes

$$-V^{-1}L^T D^{-1} O_p(n(h^{p+2} + n^{-2\delta}))_{m\times 1} + V^{-1} O_p(mnh^2(h^{p} + n^{-2\delta}))_{q\times 1} = O_p(h^p + n^{-2\delta}).$$

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

A.2 Variance

Given $t_1, ..., t_n, t_1^*, ..., t_m^*, \xi^*$, the variance $var(\hat{\xi}) = (J^T W J)^{-1} J^T W var(\tilde{Y}) W J (J^T W J)^{-1}$. Since $(J^T W J)^{-1}$ is given in (A.7), we now calculate $J^T W var(\tilde{Y}) W J$. Denote $\Sigma = var((Y_1, ..., Y_n)^T) =$

$diag\{\underbrace{\sigma^2, ..., \sigma^2}_{n}\}$. So $var(\tilde{Y})$ are simply $m \times m$ blocks of Σ,

$$\begin{pmatrix} \overbrace{\Sigma, ..., \Sigma}^{m} \\ \vdots, \ddots, \vdots \\ \Sigma, ..., \Sigma \end{pmatrix}.$$

Thus direct calculation gives that

$$J^T W var(\tilde{Y}) W J = \begin{pmatrix} D^*_{m\times m} & L^*_{m\times q} \\ (L^*)^T_{q\times m} & C^*_{q\times q} \end{pmatrix}, \tag{A.10}$$

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

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

the k-th row in L^* is

$$L^*_k = \sigma^2 \omega(t^*_k)[\sum_{j=1}^{m} \omega(t^*_j) \sum_{i=1}^{n} K_h(t_i - t^*_k)K_h(t_i - t^*_j)\widetilde{DX}_{i,k}\widetilde{D\theta}_{i,j}], \text{ for } k = 1, ..., m, \tag{A.12}$$

and

$$C^* = \sum_{k=1}^{m}\sum_{j=1}^{m} \sigma^2 \omega(t^*_k)\omega(t^*_j)[\sum_{i=1}^{n} K_h(t_i - t^*_k)K_h(t_i - t^*_j)\widetilde{D\theta}^T_{i,k}\widetilde{D\theta}_{i,j}]. \tag{A.13}$$

Lemma 3

$$D^*_{k,k} = O_p(nh^{-1}), \qquad D^*_{k,j} = o_p(n) \qquad for\ k \neq j. \tag{A.14}$$

$$L^*_k = nmh^2\sigma^2\mu_2(K)[\omega \circ f \circ \{\omega \circ f_g \circ F_{\theta^T}\}'](t^*_k) + o_p(nmh^2). \tag{A.15}$$

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

$$C^* = nm^2h^4\sigma^2[\mu_2(K)]^2 B_F + o_p(nm^2h^4). \tag{A.16}$$

The definition of B_F is given under (5). The proof of Lemma 3 is given in subsection A.3.

$$var(\hat{\xi}) = \begin{pmatrix} D & L \\ L^T & C \end{pmatrix}^{-1} \begin{pmatrix} D^* & L^* \\ (L^*)^T & C^* \end{pmatrix} \begin{pmatrix} D & L \\ L^T & C \end{pmatrix}^{-1} = \begin{pmatrix} D^{**} & L^{**} \\ (L^{**})^T & C^{**} \end{pmatrix}$$

Using (A.7), we directly calculate $var(\hat{\theta}) = C^{**}$ as

$$V^{-1}L^TD^{-1}D^*D^{-1}LV^{-1} - V^{-1}(L^*)^TD^{-1}LV^{-1} - V^{-1}L^TD^{-1}L^*V^{-1} + V^{-1}C^*V^{-1}. \quad \text{(A.17)}$$

We first focus on the case of $\omega(a_0) = \omega(b_0) = 0$. Use Lemma 2 and Lemma 3, The first term is of order

$$O_p(\frac{1}{nm})_{q\times m}O_p(\frac{n}{h})O_p(\frac{1}{nm})_{m\times q} = O(\frac{1}{nmh})$$

where an extra factor m was added from the product of the matrices of sizes $q \times m$ and $m \times q$ (the diagonal matrices in the middle does not introduce any extra factor). The second term and the third term is of order

$$O_p(\frac{1}{nmh^2})_{q\times q}O_p(nmh^2)_{q\times m}O_p(\frac{1}{nm})_{m\times q} = O_p(\frac{1}{n}).$$

Again the extra factor m comes from the product of the matrices of sizes $q \times m$ and $m \times q$. The last term is of order

$$O_p(\frac{1}{nmh^2})O_p(nmh + nm^2h^4)O_p(\frac{1}{nmh^2}) = O_p(\frac{1}{nmh^3} + \frac{1}{n}).$$

So the first term is of smaller order, and the sum of all four terms is of order $O_p(\frac{1}{nmh^3} + \frac{1}{n})$.

For the second case of when $\omega(a_0) \neq 0$ or $\omega(b_0) \neq 0$, using Lemma 3 shows that the last term in (A.17) is now $O_p(\frac{1}{nmh^3} + \frac{1}{nh})$. The first three terms order remain the same and are now of smaller order. Hence the variance of $\hat{\theta}$ is of order $O_p(\frac{1}{nmh^3} + \frac{1}{nh})$.

We now derive the explicit variance formula when $var(\hat{\theta}) = O_p(\frac{1}{n})$. That is, when $\omega(a_0) = \omega(b_0) = 0$ and $mh^3 \to \infty$. By Lemma 1, $C = nmh^2\mu_2(K)A_F + o_p(nmh^2)$; by Lemma 2, $V^{-1} = C^{-1}[1 + o(h^2)]$; and $C^* = nm^2h^4\sigma^2[\mu_2(K)]^2B_F + o_p(nm^2h^4)$ as in (A.16). Hence the last term in (A.17) becomes $V^{-1}C^*V^{-1} = \frac{\sigma^2}{n}A_F^{-1}B_FA_F^{-1} + o(\frac{1}{n})$.

Now consider the third term, $-V^{-1}L^T D^{-1} L^* V^{-1}$ in (A.17). The matrix in the middle is

$$L^T D^{-1} L^* = \sum_{k=1}^{m} L_k^T D_k{}^{-1} L_k^*.$$

By Lemma 1, $D_k = n\omega(t_k^*)f(t_k^*) + o_p(n)$, $L_k = nh^2\mu_2(K)[f'(t_k^*) + f(t_k^*)D_{X,k}^{(0)}]\omega(t_k^*)D_{\theta^T,k}^{(0)} + o_p(nh^2)$. From (A.15) $L_k^* = nmh^2\sigma^2\mu_2(K)[\omega \circ f \circ \{\omega \circ f_g \circ F_{\theta^T}\}'](t_k^*) + o_p(nmh^2)$. Hence

$$\begin{aligned} L^T D^{-1} L^* &= nmh^4\sigma^2[\mu_2(K)]^2 \textstyle\sum_{k=1}^{m} [(f' + f \circ F_X) \circ \omega \circ F_\theta \circ \{\omega \circ f_g \circ F_{\theta^T}\}'](t_k^*)[1 + o_p(1)] \\ &= nm^2h^4\sigma^2[\mu_2(K)]^2 C_F + o_p(nm^2h^4) \end{aligned}$$

where $C_F = \int [(f' + f \circ F_X) \circ \omega \circ F_\theta \circ \{\omega \circ f_g \circ F_{\theta^T}\}' \circ f_g](t)dt$ as defined under (5). Then using $C = nmh^2\mu_2(K)A_F + o(nmh^2)$ from Lemma 1, $-V^{-1}L^T D^{-1} L^* V^{-1} = -\frac{\sigma^2}{n} A_F^{-1} C_F A_F^{-1} + o_p(\frac{1}{n})$.

The second term in (A.17) is the transpose of the third term. Combine them together, we have

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

A.3 Proofs of Lemmas

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

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

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

<u>Proof of Lemma 1</u>:

The proof comes from direct calculations using $\widetilde{DX}_{i,k} = 1 + \sum_{j=1}^{p}(t_i - t_k^*)^j \frac{D_{X,k}^{(j-1)}}{j!}$, $\widetilde{D\theta}_{i,k} = \sum_{j=1}^{p}(t_i - t_k^*)^j \frac{D_{\theta^T,k}^{(j-1)}}{j!}$ and (A.1). Notice that that $\widetilde{DX}_{i,k}$ has $p+1$ terms that each is of the form of powers $(t_i - t_k^*)^j$ times a bounded quantity. So D_k by (A.4) is sum of $(p+1)^2$ terms each of the form $S_{k,j} = \sum_{i=1}^{n} K_h(t_i - t_k^*)(t_i - t_k^*)^j$ times a bounded quantity. Specifically,

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

Since the asymptotic is done for fixed p, $m \to \infty$ and $n \to \infty$, asymptotically D_k corresponds to the term with biggest 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 $O_p(nh^{j+1})$. Either way, they are 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 by (A.5),

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

Here the first term in the sum is $S_{k,1}\omega(t_k^*)D^{(0)}_{\theta^T,k} = nh^2\mu_2(K)f'(t_k^*)\omega(t_k^*)D^{(0)}_{\theta^T,k} + o_p(nh^2)$, and the first term in the double sum is $S_{k,2}D^{(0)}_{X,k}D^{(0)}_{\theta^T,k}\omega(t_k^*) = nh^2\mu_2(K)f(t_k^*)\omega(t_k^*)D^{(0)}_{X,k}D^{(0)}_{\theta^T,k} + o_p(nh^2)$. The rest of terms are at most of order $O_p(nh^4) = o_p(nh^2)$. So

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

Now from (A.6) we consider $C = \sum_{k=1}^{m}[\sum_{l=1}^{p}\sum_{j=1}^{p} S_{k,l+j}(\frac{D^{(l-1)}_{\theta^T,k}}{l!})^T(\frac{D^{(j-1)}_{\theta^T,k}}{j!})]\omega(t_k^*)$. Inside the summation over k, for each k there are p^2 terms. The first term is $S_{k,2}\omega(t_k^*)D^{(0)}_{\theta,k}D^{(0)}_{\theta^T,k} = nh^2\mu_2(K)f(t_k^*)\omega(t_k^*)D^{(0)}_{\theta,k}D^{(0)}_{\theta^T,k} + o_p(nh^2)$. The rest of the terms are at most of order $O_p(nh^4) = o_p(nh^2)$. So after the sum over k, we have

$$C = \sum_{k=1}^{m} nh^2\mu_2(K)f(t_k^*)\omega(t_k^*)D^{(0)}_{\theta,k}D^{(0)}_{\theta^T,k} + o_p(mnh^2) = nmh^2\mu_2(K)\int[\omega\circ f\circ f_g\circ F_\theta\circ F_{\theta^T}](t)dt + o(nmh^2).$$

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

Proof of Lemma 2:

Since D is diagonal, by Lemma 1, D^{-1} diagonal with entries of $O_p(1/n)$. Hence

$$L^TD^{-1}L = O_p(nh^2)_{q\times m}O_p(\frac{1}{n})O_p(nh^2)_{m\times q} = O_p(nh^2\frac{1}{n}nh^2)m = O_p(nmh^4).$$

Note that we count an extra factor of m when multiplying the $q \times m$ matrix with a $m \times q$ matrix. Since D^{-1} is diagonal, multiplying by D^{-1} does not introduce the extra factor m. The rest of the order calculations of matrices products are all like this: direct multiplications of the order of each matrix and add an extra factor m for multiplication over the m dimension, but no such extra factor for diagonal matrix nor for multiplication over the fixed dimension q.

Now $V = C + O_p(nmh^4) = C[1 + O_p(h^2)]$ by Lemma 1. With the fixed $q \times q$ dimensions, we have $V^{-1} = C^{-1}[1 + O_p(h^2)]$. Then $D^{-1}LV^{-1} = O_p(\frac{1}{n}nh^2\frac{1}{nmh^2}) = O_p(\frac{1}{mn})$. And $D^{-1}LV^{-1}L^TD^{-1} = O_p(\frac{1}{mn}nh^2\frac{1}{n})m = O_p(\frac{h^2}{n})$.

Lemma 3

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

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

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

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

Proof of Lemma 3:

The analysis of the order of $D^*_{k,j}$ is similar to the analysis of D_k in proof of Lemma 1. There are $(p+1) \times (p+1)$ terms in $\widetilde{DX}_{i,k}\widetilde{DX}_{i,j}$ with the lowest power term being 1. So the first term in $D^*_{k,j}$ his $\sigma^2\omega(t^*_k)\omega(t^*_j)\sum_{i=1}^n K_h(t_i - t^*_k)K_h(t_i - t^*_j)$. For $k = j$, this becomes $\sum_{i=1}^n [K_h(t_i - t^*_k)]^2 = nh^{-1}\mu_0(K^2)f(t^*_k) + o_p(nh^{-1})$. Easy to check the rest of terms in $D^*_{k,j}$ is at most of order $O_p(nh)$. So

$$D^*_{k,k} = nh^{-1}\sigma^2[\omega(t^*_k)]^2\mu_0(K^2)[1 + o_p(1)] = O_p(nh^{-1}).$$

When $k \neq j$, $\sum_{i=1}^n K_h(t_i - t^*_k)K_h(t_i - t^*_j) \to n\int K(u)K_h(t^*_k - t^*_j + hu)f(t^*_k + hu)du = o_p(n)$. Hence we have $D^*_{k,j} = o_p(n)$ for $k \neq j$.

For L_k^*, by (A.12), we only need to consider the lowest power terms in $\widetilde{DX}_{i,k}\widetilde{D\theta}_{i,j}$, that is, $1(t_i - t_j^*)D^{(0)}_{\theta^T,j}$. So (A.12) becomes

$$\begin{aligned} L_k^* =& \ \sigma^2\omega(t_k^*)[\textstyle\sum_{i=1}^n K_h(t_i - t_k^*)\sum_{j=1}^m \omega(t_j^*)K_h(t_i - t_j^*)(t_i - t_j^*)D^{(0)}_{\theta^T,j}][1 + o_p(1)] \\ =& \ \sigma^2\omega(t_k^*)[\textstyle\sum_{i=1}^n K_h(t_i - t_k^*)mh^2\mu_2(K)[-\omega \circ f_g \circ F_{\theta^T}]'(t_i)][1 + o_p(1)] \\ =& \ -nmh^2\sigma^2\mu_2(K)[\omega \circ f \circ \{\omega \circ f_g \circ F_{\theta^T}\}'](t_k^*) + o_p(nmh^2). \end{aligned}$$

We now evaluate the order of C^* in more details. Again, for $\widetilde{D\theta}_{i,j}$ the leading term is $(t_i - t_{j^*})D^{(0)}_{\theta^T}(\alpha_j;\theta)$. So from (A.13), we have C^* becomes

$$\begin{aligned} & C^* \\ =& \ \sigma^2 \textstyle\sum_{i=1}^n \sum_{k=1}^m \sum_{j=1}^m K_h(t_i - t_k^*)K_h(t_i - t_{j^*})(t_i - t_k^*)(t_i - t_{j^*})\omega(t_k^*)\omega(t_j^*) \\ & \qquad D_\theta(\alpha_k;\theta)D_{\theta^T}(\alpha_j;\theta)[1 + o_p(1)] \\ =& \ \sigma^2[\textstyle\sum_{i=1}^n \sum_{k=1}^m \{K_h(t_i - t_k^*)(t_i - t_k^*)\omega(t_k^*)\}^2 D_\theta(\alpha_k;\theta)D_{\theta^T}(\alpha_k;\theta) \\ & \quad + \textstyle\sum_{i=1}^n \sum_{(k\neq j)=1}^m K_h(t_i - t_k^*)K_h(t_i - t_{j^*})(t_i - t_k^*)(t_i - t_{j^*})\omega(t_k^*)\omega(t_j^*) \\ & \qquad D_\theta(\alpha_k;\theta)D_{\theta^T}(\alpha_j;\theta)\ ]\ [1 + o_p(1)] \\ =& \ \sigma^2[nm\textstyle\int\int \frac{1}{h^2}\{K(\frac{x-z}{h})\}^2(x-z)^2[\omega(x)]^2 F_\theta(x)F_{\theta^T}(x)f_g(x)f(z)dxdz \\ & \quad + nm^2 \textstyle\int\int\int \frac{1}{h^2}K(\frac{x-z}{h})K(\frac{y-z}{h})(x-z)(y-z)\omega(x)\omega(y)F_\theta(x)F_{\theta^T}(y) \\ & \qquad f_g(x)f_g(y)f(z)dxdydz\ ]\ [1 + o_p(1)]. \end{aligned} \tag{A.18}$$

We further simplify the triple integral in the last expression using integral by parts: Let $K^*(t)$ be the anti-derivative of $K(t)$ and $K^{**}(t)$ the anti-derivative of $K^*(t)$. That is, $\frac{d^2}{dt^2}K^{**}(t) = \frac{d}{dt}K^*(t) = K(t)$. Denote $\kappa^*(x) = K^*(x)x - K^{**}(x)$ so that $\frac{d}{dx}\kappa^*(x) = K(x)x$. Hence

$$\begin{aligned} & \textstyle\int \frac{1}{h}K(\frac{x-z}{h})(x-z)\omega(x)F_\theta(x)f_g(x)dx \\ =& \ h\kappa^*(\frac{x-z}{h})\omega(x)F_\theta(x)f_g(x)|_{a_0}^{b_0} - \textstyle\int h\kappa^*(\frac{x-z}{h})[\omega \circ f_g \circ F_\theta]'(x)dx. \end{aligned} \tag{A.19}$$

Notice that the first term is different for two cases: (a) the weight function is zero at boundary points $\omega(a_0) = \omega(b_0) = 0$; and (b) the weight function is nonzero at boundary points $\omega(a_0) \neq 0$ or $\omega(b_0) \neq 0$. For the first case of $\omega(a_0) = \omega(b_0) = 0$, the first term in (A.19) equals zero. For the second case, the first term in (A.19) is non-zero.

We now consider the case where the first term in (A.19) equals zero. Then only the second term in (A.19) remains, which is (with a change of variable $u = \frac{x-z}{h}$)

$$\begin{aligned} & -\int h\kappa^*(\tfrac{x-z}{h})[\omega \circ f_g \circ F_\theta]'(x)dx \\ = & -h^2 \int \kappa^*(u)[\omega \circ f_g \circ F_\theta]'(z+hu)du \\ = & -h^2[\omega \circ f_g \circ F_\theta]'(z) \int \kappa^*(u)du[1+o_p(1)] \\ = & \; h^2\mu_2(K)[\omega \circ f_g \circ F_\theta]'(z)[1+o_p(1)]. \end{aligned}$$

This is because of $-\int \kappa^*(u)du = -\kappa^*(u)u|_{-1}^{1} + \int [K(u)u]udu = \mu_2(K)$ where without loss of generality, we assume that $K(u)$ has a compact support on $[-1,1]$, $\kappa^*(-1) = 0$ since $K^*(-1) = K^{**}(-1) = 0$. $\kappa^*(u)|_{-1}^{1} = \int [K(u)u]du = 0$ so $\kappa^*(1) = 0 + \kappa^*(-1) = 0$. Therefore $\kappa^*(u)u|_{-1}^{1} = 0$. Hence, plugging this into (A.18) we have

$$\begin{aligned} & C^* \\ = & \; \sigma^2\{nmh\mu_2(K^2)\int [\omega(x)]^2 F_\theta(x)F_{\theta^T}(x)f_g(x)f(x)dx \\ & +nm^2 \int \{h^2\mu_2(K)[\omega \circ f_g \circ F_\theta]'(z)\}\{h^2\mu_2(K)[\omega \circ f_g \circ F_{\theta^T}]'(z)\}f(z)dz\}[1+o_p(1)] \\ = & \; \sigma^2\{nmh\mu_2(K^2)\int [\omega^2 \circ f_g \circ f \circ F_\theta \circ F_{\theta^T}](x)dx \\ & +nm^2h^4[\mu_2(K)]^2 \int [(\omega \circ f_g \circ F_\theta)' \circ (\omega \circ f_g \circ F_{\theta^T})' \circ f](z)dz\}[1+o_p(1)] \\ = & \; O_p(nmh + nm^2h^4) \end{aligned}$$

For the second case where the first term in (A.19) cannot be dropped out, there is an extra term in the triple integral of (A.18) ,

$$nm^2 \int [h\kappa^*(\frac{b_0 - z}{h})]^2[\omega \circ f_g \circ F_\theta](z)[\omega \circ f_g \circ F_{\theta^T}](z)f(z)dz = O_p(nm^2h^3)$$

Therefore $C^* = O_p(nmh + nm^2h^3)$ instead.

This finishes the proof of Lemma 3.

B Numerical Studies

In this section, we compare the performance of the proposed method with the Liang-Wu's method (2008), the method of Ramsay et al. (2007) and the NLS estimator by Monte Carlo simulations. We evaluate the performance of the estimators by the average relative error (ARE) defined as

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

with $\hat{\theta}_i$ as the estimate for true θ in the ith simulation run with $i = 1, 2, ..., r$. The computational cost and convergence are also considered in evaluating different estimation methods, but the determination of the smoothing parameters is not considered in computational cost. However, the same bandwidth was used for the Liang-Wu's PsLS estimator and our new estimator for fair comparisons.

Since the Liang-Wu's PsLS estimator and the proposed new estimator in this paper are computationally efficient, they can be used as the starting point for the NLS estimator. So that this hybrid strategy may enjoy both computational efficiency of the PsLS estimator or the new estimator and high estimation accuracy of the NLS estimator. We will also evaluate the hybrid approaches in our simulation studies.

Example 1. Chen et al. (1999) proposed a system of differential equations to model protein and gene interactions. We simulated the data from a modified gene-protein interaction model,

$$\begin{aligned} \tfrac{d}{dt}X_1 &= \tfrac{a}{1+e^{-X_2}} - bX_1 \\ \tfrac{d}{dt}X_2 &= 2X_1 - cX_2 \end{aligned} \qquad \text{(A.20)}$$

with true parameter values $\theta = (a, b, c) = (1.5, 1, 2)$. Assume that X_1 and X_2 were measured over a grid of n equally-spaced time points, every time interval of 0.4 in the range of $t = [0, 20]$, with measurement errors as in equation (2) with (σ_1, σ_2) taking as $(0.1, 0.1)$, $(0.1, 0.3)$, $(0.3, 0.1)$ 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 simulated data set, we apply the proposed estimation method and other existing methods to obtain the following estimates: the NLS estimator $\hat{\theta}^{NLS}$ and Ramsay et al.'s collocation estimator $\hat{\theta}^{col}$, and the Liang-Wu's PsLS estimators $\hat{\theta}_n^{PLS}$ and $\hat{\theta}_m^{PLS}$ with $m = n$ and $m = [n^{4/3}]^+$ ($[z]^+$ denotes the largest integer that does not exceed z) respectively using the starting values generated randomly from 0 to twice the true parameter value, the NLS estimator $\hat{\theta}_{PLS}^{NLS}$ using the Liang-Wu's PsLS estimator $\hat{\theta}_n^{PLS}$ as the starting point, the proposed new estimator $\hat{\theta}_n^{new}$ and $\hat{\theta}_m^{new}$ using the Liang-Wu's PsLS estimators as the starting point, the NLS estimators $\hat{\theta}_{new,n}^{NLS}$ and $\hat{\theta}_{new,m}^{NLS}$ using the proposed estimators $\hat{\theta}_n^{PLS}$ and $\hat{\theta}_m^{PLS}$ respectively as the starting point. For the Liang-Wu PsLS estimator and the proposed estimator, the local quadratic polynomial smoothing was used and the piecewise linear weighted function suggested in Brunel (2008) was used: $w(t) = 1$ for $1 \leq t \leq 19$; $w(t) = t$ for $0 \leq t \leq 1$; $w(t) = 20 - t$ for $19 \leq t \leq 20$. Only one iteration was used for the proposed estimator starting at the Liang-Wu's PsLS estimator since very little accuracy improvement was obtained by more iterations. The Ramsay et al. (2007)'s collocation estimator $\hat{\theta}^{col}$ was implemented using the R package *CollocInfer* (available from http://www.bscb.cornell.edu/∼hooker) with 51 equally-spaced knots between $t = 0$ and $t = 20$.

Table 1 summarizes the AREs and computing times of the various estimators based on $r = 400$ simulation runs. From these simulation results, we observe the following patterns, in particular for the estimates of parameters a and b: 1) The Liang-Wu's PsLS estimator is always most computationally efficient, but the second worst in estimation accuracy (in AREs) except that the NLS estimator with twice the true parameter values as the stating point was the worst in estimation accuracy probably due to the local convergence. 2) The improvement of estimation accuracy by the data augmentation approach, $m > n$, is limited. 3) Ramsay et al. (2007)'s collocation estimator performs similarly in estimation accuracy to the NLS estimator when the PsLS estimator or our proposed new estimator was used as the starting point, but the computational cost of the collocation estimator (implemented using

(σ_1, σ_2)			Estimators								
			$\hat{\theta}^{NLS}$	$\hat{\theta}^{col}$	$\hat{\theta}^{NLS}_{PLS}$	$\hat{\theta}^{PLS}_{n}$	$\hat{\theta}^{new}_{n}$	$\hat{\theta}^{NLS}_{new,n}$	$\hat{\theta}^{PLS}_{m}$	$\hat{\theta}^{new}_{m}$	$\hat{\theta}^{NLS}_{new,m}$
(0.1,0.1)	AREs	a	44.36	13.74	8.58	18.20	11.81	8.58	18.25	11.64	8.58
		b	18.26	14.42	8.84	18.70	12.07	8.84	18.76	11.90	8.84
		c	35.40	2.09	1.55	1.55	1.53	1.55	1.55	1.53	1.55
	diverge		3.25	2.25	0	0	0	0	0	0	0
	time		1.36	26.82	0.91	0.19	0.27	0.99	0.63	1.05	1.74
(0.1,0.3)	AREs	a	49.85	22.34	17.64	23.89	18.85	17.64	23.71	18.75	17.64
		b	216.0	23.31	18.22	24.67	19.03	18.22	24.47	18.93	18.22
		c	32.15	4.05	3.58	3.58	3.55	3.58	3.59	3.55	3.58
	diverge		3.75	2.0	0	0	0	0	0	0	0
	time		1.36	25.28	0.99	0.18	0.25	1.08	0.58	0.97	1.80
(0.3,0.1)	AREs	a	81.39	28.32	21.25	53.56	33.07	20.96	53.20	32.85	20.96
		b	216.3	29.26	21.87	55.29	33.51	21.46	54.91	33.34	21.46
		c	39.6	3.60	3.30	3.33	3.29	3.30	3.32	3.28	3.30
	diverge		5.25	4.25	0	0	0	0	0	0	0
	time		1.46	24.27	1.25	0.18	0.25	1.24	0.58	0.98	1.97
(0.3,0.3)	AREs	a	87.09	35.02	33.08	54.81	41.69	30.13	54.42	41.63	29.32
		b	2.28	36.26	36.14	56.63	42.26	32.44	56.24	42.19	30.87
		c	34.65	4.93	4.64	4.66	4.60	4.64	4.66	4.59	4.65
	diverge		5.75	3.25	0	0	0	0	0	0	0
	time		1.51	25.50	1.33	0.18	0.25	1.33	0.58	.96	2.05

Table 1: Performance of different estimators for Example 1 with $n = 51$ observations: $\hat{\theta}^{NLS}$=NLS estimate starting randomly from 0 to the twice true values; $\hat{\theta}^{col}$=Ramsay et al's collocation estimate; $\hat{\theta}^{PLS}_{n}$=PsLS estimate on a grid of n times points; $\hat{\theta}^{PLS}_{m}$=PsLS estimate on a grid of $m = [n^{4/3}]^{+}$ times points; $\hat{\theta}^{NLS}_{PLS}$=NLS estimate starting at $\hat{\theta}^{PLS}_{n}$; $\hat{\theta}^{new}_{n}$=the proposed new estimate on a grid of n times points; $\hat{\theta}^{new}_{m}$=the proposed new estimate on a grid of $m = [n^{4/3}]^{+}$ times points; $\hat{\theta}^{NLS}_{new,n}$=NLS estimate starting at $\hat{\theta}^{new}_{n}$; $\hat{\theta}^{NLS}_{new,m}$=NLS estimate starting at $\hat{\theta}^{new}_{m}$

the Hooker's R package) is highest. However, it is possible to reduce the computational cost of the collocation estimator if a more efficient algorithm is used and the smoothing parameter or tuning parameter is more appropriately adjusted. 4) If the starting point of the NLS estimator is far from the true values, the NLS may converge to the local minima with a high computational cost and poor estimation accuracy. However, when the Liang-Wu's PsLS estimator or the proposed new estimator was used as the starting point, the NLS estimator is significantly improved to become the best among all the estimators in estimation accuracy with a reasonable price of computational cost. 5) The proposed new estimator is clearly better than the Liang-Wu's PsLS estimator in estimation accuracy with a small price of computational cost as we expected.

Example 2. In this second simulation example, 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 et al. (1962). This model was also used for simulation studies by Ramsay et al. (2007) and Liang & Wu (2008). We use this model to further investigate the finite-sample behavior of the proposed method and other existing methods for a different nonlinear differential equation model. The FitzHugh-Nagumo system can be written as

$$\begin{aligned} \tfrac{d}{dt}X_1 &= (X_1 + X_2 - \tfrac{X_1^3}{3})c, \\ \tfrac{d}{dt}X_2 &= -\tfrac{X_1 - a + bX_2}{c}, \end{aligned} \tag{A.21}$$

with true parameter values $\theta = (a, b, c) = (0.34, 0.2, 3)$ in our simulations. We similarly assume that X_1 and X_2 are measured over a grid of 51 equally-spaced time points, every 0.4 time interval in the range of $t = [0, 20]$ with measurement errors as in equation (2) with (σ_1, σ_2) taking as $(0.1, 0.1)$, $(0.1, 0.3)$, $(0.3, 0.1)$ or $(0.3, 0.3)$ for the measurement standard errors for X_1 and X_2 respectively. For each simulated data set, we apply the proposed estimation method and other existing methods to obtain all the estimates as in Example 1.

Table 2 summarizes the AREs and computing times of the various estimators based on

(σ_1, σ_2)			Estimators								
			$\hat{\theta}^{NLS}$	$\hat{\theta}^{col}$	$\hat{\theta}^{NLS}_{PLS}$	$\hat{\theta}^{PLS}_n$	$\hat{\theta}^{new}_n$	$\hat{\theta}^{NLS}_{new,n}$	$\hat{\theta}^{PLS}_m$	$\hat{\theta}^{new}_m$	$\hat{\theta}^{NLS}_{new,m}$
(0.1,0.1)	ARE	a	2.58	7.70	1.96	4.26	5.42	1.75	4.11	4.25	1.75
		b	14.1	58.2	12.5	19.64	20.71	11.95	19.7	19.3	12.0
		c	2.32	6.07	0.69	26.82	21.16	0.37	24.2	12.3	0.37
	diverge		36.75	6.00	4.25	0	0	0.50	0	0	0
	time		11.59	12.07	8.17	0.17	0.24	7.01	0.58	0.89	7.07
(0.1,0.3)	ARE	a	6.73	10.34	3.72	6.85	8.06	2.49	6.90	7.62	2.49
		b	42.1	69.7	33.6	52.78	49.28	28.9	52.4	50.4	28.9
		c	9.08	7.73	2.44	33.95	22.31	0.55	31.1	14.9	0.56
	diverge		39.75	5.25	14.75	0	0	1.25	0	0	1.25
	time		10.93	12.89	9.11	0.18	0.25	8.14	0.57	0.87	7.90
(0.3,0.1)	ARE	a	5.21	13.61	5.26	10.30	8.60	4.97	10.2	8.23	5.15
		b	22.9	73.0	24.4	29.49	27.43	23.18	29.3	27.6	23.9
		c	1.91	7.71	1.42	34.08	21.70	1.05	31.1	19.9	1.39
	diverge		34.50	5.75	8.75	0	0	2.00	0	0	2.00
	time		13.08	14.00	12.62	0.18	0.25	11.66	0.58	0.87	11.75
(0.3,0.3)	ARE	a	6.12	13.58	5.68	11.44	10.73	5.44	11.5	10.6	5.19
		b	35.0	82.6	35.8	55.02	53.28	36.1	54.9	53.6	35.2
		c	4.21	9.04	2.44	43.33	24.44	1.41	40.0	22.3	1.12
	diverge		39.75	6.25	20.00	0	0	7.00	0	0	7.75
	time		12.75	14.73	12.06	0.18	0.25	10.99	0.58	0.87	10.93

Table 2: Performance of different estimators for Example 2 with $n = 51$ observations: $\hat{\theta}^{NLS}$=NLS estimate starting randomly from 0 to the twice true values; $\hat{\theta}^{col}$=Ramsay et al's collocation estimate; $\hat{\theta}^{PLS}_n$=PsLS estimate on a grid of n times points; $\hat{\theta}^{PLS}_m$=PsLS estimate on a grid of $m = [n^{4/3}]^+$ times points; $\hat{\theta}^{NLS}_{PLS}$=NLS estimate starting at $\hat{\theta}^{PLS}_n$; $\hat{\theta}^{new}_n$=the proposed new estimate on a grid of n times points; $\hat{\theta}^{new}_m$=the proposed new estimate on a grid of $m = [n^{4/3}]^+$ times points; $\hat{\theta}^{NLS}_{new,n}$=NLS estimate starting at $\hat{\theta}^{new}_n$; $\hat{\theta}^{NLS}_{new,m}$=NLS estimate starting at $\hat{\theta}^{new}_m$.

$r = 400$ simulation runs. From this simulation study, additional interesting behaviors of different estimators were observed although most results still hold as in Example 1: 1) The NLS estimator is still the worst in both computational cost and convergence when the starting point is randomly generated from 0 to twice the true parameter values. The non-convergence frequency of the NLS estimator is much higher for the model in this example compared to that of Example 1. In this case, it is more important to get a good starting value for the NLS estimator. However we notice that the improvement using the Liang-Wu's PsLS estimator as the starting point is limited while using the proposed new estimator as the starting point, the NLS estimator can be significantly improved to become the best among all the estimators in estimation accuracy with a reasonable price of the computational cost. 2) In this case, Ramsay et al. (2007)'s collocation estimator performs worse than the best NLS estimator (the NLS estimator with our new estimator as the starting point) and also has a higher computational cost. 3) The proposed new estimator, as the starting point, can better improve the NLS estimator in the sense of AREs and convergence stability, compared to that of the PsLS estimator as the starting point. 4) In Table 3, we also reported the standard deviation (STD) of the estimators in Table 2 here. We can see that the trend and conclusions for the STD are similar to those for the AREs, i.e., comparing to the Liang-Wu's pseudo-least squares (PLS) estimator, our new estimator has a lower STD in all three parameters a, b and c for the cases $(\sigma_1, \sigma_2) = (0.3, 0.1)$ and $(0.3, 0.3)$; while for the other two cases, the two methods produce mixed performance in terms of both STD and ARE (for some parameters, the PLS estimator is better and for some other parameters, our new estimate is better). However, the NLS estimator using our new estimator as the initial value always performs better for all the cases in both STD and ARE, compared to those using the PLS estimator as the initial value. In summary, considering the computational cost, convergence rate, variance and ARE, the best estimator is the NLS estimator θ_{new}^{NLS} using the proposed new estimator as the starting point.

(σ_1, σ_2)			Estimators								
			$\hat{\theta}^{NLS}$	$\hat{\theta}^{col}$	$\hat{\theta}^{NLS}_{PLS}$	$\hat{\theta}^{PLS}_{n}$	$\hat{\theta}^{new}_{n}$	$\hat{\theta}^{NLS}_{new,n}$	$\hat{\theta}^{PLS}_{m}$	$\hat{\theta}^{new}_{m}$	$\hat{\theta}^{NLS}_{new,m}$
(0.1,0.1)	STD	a	7.73	14.2	4.59	5.71	6.93	2.23	5.19	5.24	2.22
		b	24.6	82.0	18.3	25.0	24.8	15.2	25.1	24.0	15.2
		c	13.63	15.5	6.31	2.95	3.64	0.47	3.08	3.62	0.47
	diverge		36.75	6.00	4.25	0	0	0.50	0	0	0
	time		11.59	12.07	8.17	0.17	0.24	7.01	0.58	0.89	7.07
(0.1,0.3)	STD	a	6.24	15.2	16.6	8.65	10.1	3.11	8.88	9.68	3.10
		b	39.6	87.0	70.4	66.7	61.9	36.8	66.2	63.2	36.7
		c	15.15	16.3	14.7	4.04	5.83	0.73	4.49	6.37	0.73
	diverge		39.75	5.25	14.75	0	0	1.25	0	0	1.25
	time		10.93	12.89	9.11	0.18	0.25	8.14	0.57	0.87	7.90
(0.3,0.1)	STD	a	6.75	19.1	7.17	13.0	10.6	6.34	12.8	10.0	6.34
		b	29.1	85.8	33.0	37.4	34.4	29.7	36.9	33.7	29.9
		c	9.60	15.8	6.09	5.96	4.39	1.32	5.30	5.56	1.33
	diverge		34.50	5.75	8.75	0	0	2.00	0	0	2.00
	time		13.08	14.00	12.62	0.18	0.25	11.66	0.58	0.87	11.75
(0.3,0.3)	STD	a	10.24	18.0	10.1	14.1	13.3	8.70	14.3	13.1	6.61
		b	47.1	95.1	45.7	70.3	66.8	45.7	69.9	67.3	44.6
		c	18.81	16.7	12.6	6.38	5.75	5.58	5.90	6.52	1.42
	diverge		39.75	6.25	20.00	0	0	7.00	0	0	7.75
	time		12.75	14.73	12.06	0.18	0.25	10.99	0.58	0.87	10.93

Table 3: Performance (standard deviation as percentage) of different estimators for Example 1 with $n = 51$ observations: $\hat{\theta}^{NLS}$=nonlinear least squares estimate using a random starting point; $\hat{\theta}^{col}$=Ramsay et al's collocation estimate using the same starting point; $\hat{\theta}^{PLS}$=pseudo-least squares estimate using the same starting point; $\hat{\theta}^{new}$=the proposed new estimate started from $\hat{\theta}^{PLS}$; $\hat{\theta}^{NLS}_{PLS}$=nonlinear least squares estimate started from $\hat{\theta}^{PLS}$; θ^{NLS}_{new}=nonlinear least squares estimate started from $\hat{\theta}^{new}$; θ^{NLS}_{col}=nonlinear least squares estimate started from $\hat{\theta}^{col}$.

References

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

Chen, J. & Wu, H. (2008). Efficient Local Estimation for Time-varying Coefficients in Deterministic Dynamic Models with Applications to HIV-1 Dynamics. *Journal of the American Statistical Association* **103**, 369-384.

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

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

Fang, Y., Wu, H. & Zhu, L. (2011). A Two-Stage Estimation Method for Random Coefficient Differential Equation Models with Application to Longitudinal HIV Dynamic Data, *Statistica Sinica*, **21**, 1145-1170.

FitzHugh, R. (1961). Impulses and Physiological States in Models of Nerve Membrane. *Biophysical Journal*, **1**, 445-466.

Liang, H. & Wu, H. (2008). Parameter Estimation of Differential Equation Models Using a Framework of Measurement Error in Regression Models. *Journal of the American Statistical Association*, **103**, 1570-1583.

Liang, H. & Wu, H. (2010). Correction on "Parameter Estimation of Differential Equation Models Using a Framework of Measurement Error in Regression Models (JASA, 103, 1570-1583)". *Journal of the American Statistical Association*, **105**, 1636.

Lu, T., Liang, H., Li, H. & Wu, H. (2011). High Dimensional ODEs Coupled with Mixed-Effects Modeling Techniques for Dynamic Gene Regulatory Network Identification, *Journal of the American Statistical Association*, **106**, 1242-1258.

Nagumo, J. S., Arimoto, S. & Yoshizawa, S. (1962). An Active Pulse Transmission Line Simulating a Nerve Axon. *Proceedings of the IRE*, **50**, 2061-2070.

Ramsay, J. O., Hooker, G., Campbell, D. & Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach (with discussion). *J. R. Statist. Soc. 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. & 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. *Journal of Immunology*. **187**, 4474-4482.