SIEVE ESTIMATION OF CONSTANT AND TIME-VARYING COEFFICIENTS IN NONLINEAR ORDINARY DIFFERENTIAL EQUATION MODELS BY CONSIDERING BOTH NUMERICAL ERROR AND MEASUREMENT ERROR

BY HONGQI XUE, HONGYU MIAO AND HULIN WU[1]

University of Rochester

This article considers estimation of constant and time-varying coefficients in nonlinear ordinary differential equation (ODE) models where analytic closed-form solutions are not available. The numerical solution-based nonlinear least squares (NLS) estimator is investigated in this study. A numerical algorithm such as the Runge–Kutta method is used to approximate the ODE solution. The asymptotic properties are established for the proposed estimators considering both numerical error and measurement error. The B-spline is used to approximate the time-varying coefficients, and the corresponding asymptotic theories in this case are investigated under the framework of the sieve approach. Our results show that if the maximum step size of the p-order numerical algorithm goes to zero at a rate faster than $n^{-1/(p \wedge 4)}$, the numerical error is negligible compared to the measurement error. This result provides a theoretical guidance in selection of the step size for numerical evaluations of ODEs. Moreover, we have shown that the numerical solution-based NLS estimator and the sieve NLS estimator are strongly consistent. The sieve estimator of constant parameters is asymptotically normal with the same asymptotic co-variance as that of the case where the true ODE solution is exactly known, while the estimator of the time-varying parameter has the optimal convergence rate under some regularity conditions. The theoretical results are also developed for the case when the step size of the ODE numerical solver does not go to zero fast enough or the numerical error is comparable to the measurement error. We illustrate our approach with both simulation studies and clinical data on HIV viral dynamics.

1. Introduction. Ordinary differential equations (ODE) are widely used to model dynamic processes in many scientific fields such as engineering, physics, econometrics and biomedical sciences. In particular, new biotechnologies allow scientists to use ODE models to more accurately describe biological processes such as genetic regulatory networks, tumor cell kinetics, epidemics and viral dynamics of infectious diseases [Chen, He and Church (1999), Jansson and Revesz

(1975), Michelson and Leith (1997), Daley and Gani (1999), Anderson and May (1991), Brookmeyer and Gail (1994), Nowak and May (2000)]. The mathematical modeling approach has made a great impact on these scientific fields over the past decades. For instance, ODE models have been used to quantify HIV viral dynamics which resulted in important scientific findings [Ho et al. (1995), Wei et al. (1995), Perelson et al. (1996, 1997)]. Comprehensive reviews of the application of ODE models in HIV dynamics can be found in Perelson and Nelson (1999), Nowak and May (2000), Tan and Wu (2005) and Wu (2005).

Although differential equation models have been widely used in scientific research, very little statistical research has been dedicated to parameter estimation and inference for differential equation models. The existing statistical methods for ODE models include the nonlinear least squares method [Bard (1974), van Domselaar and Hemker (1975), Benson (1979), Li, Osborne and Pravan (2005)], the smoothing-based techniques [Swartz and Bremermann (1975), Varah (1982), Chen and Wu (2008), Liang and Wu (2008), Brunel (2008)], the principal differential analysis (PDA) [Ramsay (1996), Heckman and Ramsay (2000), Poyton et al. (2006), Ramsay et al. (2007), Varziri et al. (2008)] and the Bayesian approaches [Putter et al. (2002), Huang, Liu and Wu (2006), Donnet and Samson (2007)]. However, very few of these publications rigorously address the theoretical issues and study the asymptotic properties of the proposed estimators when both measurement error and numerical error are significant. In this paper, we intend to investigate statistical estimation methods for both constant and time-varying parameters in ODE models and study the asymptotic properties of the proposed estimators under the framework of the sieve approach.

Denote a general set of ODE models containing only constant parameters as

$$(1.1) \quad \begin{cases} \dfrac{d\mathbf{X}(t)}{dt} = \mathbf{F}\{t, \mathbf{X}(t), \boldsymbol{\beta}\}, & \forall t \in [t_0, T], \\ \mathbf{X}(t_0) = \mathbf{X}_0, \end{cases}$$

and denote a general set of ODE models with both constant and time-varying parameters as

$$(1.2) \quad \begin{cases} \dfrac{d\mathbf{X}(t)}{dt} = \mathbf{F}\{t, \mathbf{X}(t), \boldsymbol{\beta}, \eta(t)\}, & \forall t \in [t_0, T], \\ \mathbf{X}(t_0) = \mathbf{X}_0, \end{cases}$$

where $\mathbf{X}(t) = \{X_1(t), \ldots, X_K(t)\}^T$ is a K-dimensional state variable vector, $\boldsymbol{\beta}$ is a d-dimensional vector of unknown constant parameters with true value $\boldsymbol{\beta}_0$, $\eta(t)$ is an unknown time-varying parameter with true value $\eta_0(t)$ (here we only consider a single time-varying parameter, the proposed methodology can be extended to include multiple time-varying parameters although it is tedious and cumbersome in notation), $\mathbf{F}(\cdot) = \{F_1(\cdot), \ldots, F_K(\cdot)\}^T$ is a vector of differentiable functions whose forms are known and $\mathbf{X}(t_0) = \mathbf{X}_0$ is the initial value. Equations (1.1) and (1.2) are called state equations. Obviously, equation (1.1) is a special case of

(1.2). The function $\mathbf{F}(t, \mathbf{X}, \boldsymbol{\beta})$ in (1.1) or $\mathbf{F}(t, \mathbf{X}, \boldsymbol{\beta}, \eta)$ in (1.2) is assumed to fulfil the Lipschitz assumption to $\mathbf{X}$ [with the Lipschitz constant independent of the unknown parameters $\boldsymbol{\beta}$ and $\eta(\cdot)$] ensuring existence and uniqueness of the solutions to (1.1) and (1.2) [see Hairer, Nørsett and Wanner (1993) and Mattheij and Molenaar (2002)]. Let $\mathbf{X}(t, \boldsymbol{\beta})$ and $\mathbf{X}(t, \boldsymbol{\beta}, \eta(t))$ denote the true solutions to (1.1) and (1.2) for given $\boldsymbol{\beta}$ and $\eta(\cdot)$, respectively. We usually use notation $\mathbf{X}(t)$ to denote $\mathbf{X}(t, \boldsymbol{\beta}_0)$ or $\mathbf{X}(t, \boldsymbol{\beta}_0, \eta_0(t))$ in this article. Our objective is to estimate the unknown parameters $\boldsymbol{\beta}$ and $\eta(\cdot)$ based on the measurements of the state variables, $\mathbf{X}(t)$ or their functions.

If a closed-form solution to (1.1) or (1.2) is available, the standard statistical approaches for nonlinear regression or time-varying coefficient regression models can be used to estimate unknown parameters. In practice, (1.1) and (1.2) usually do not have closed-form solutions for a nonlinear $\mathbf{F}$. In this case, numerical methods such as the Runge–Kutta algorithm [Runge (1895), Kutta (1901)] have to be used to approximate the solution of the ODEs for a given set of parameter values and initial conditions. Consequently, the nonlinear least squares (NLS) principle (minimizing the residual sum of squares of the differences between the experimental observations and numerical solutions) can be used to obtain the estimates of the unknown parameters. The NLS method for (1.1) was first described by mathematicians in 1970s [Bard (1974), van Domselaar and Hemker (1975), Benson (1979)]. The NLS method was also widely used to estimate the unknown parameters in ODE models in the fields of mathematics, computer science and control engineering. In the 1990s, the NLS method was extended to estimate time-varying parameters in (1.2). For example, the NLS method with spline approximation to time-varying parameters has been successfully applied to pharmacokinetic [Li et al. (2002)], physiologic [Thomaseth et al. (1996)] and HIV studies [Adams (2005)].

Though the NLS was the earliest and the most popular method developed for estimating the parameters in ODE models, so far the proposed NLS estimators and their asymptotic properties for ODE models have not been systematically studied, in particular, for time-varying parameter estimates. The influence of the numerical approximation error of ODEs on the asymptotic properties has not been analyzed. All existing studies took the numerical solution as the true solution and did not consider the difference between them. The difficulty is due to the co-existence of both measurement error and numerical error, and the standard theories of the NLS method [Jennerich (1969), Malinvaud (1970), Wu (1981), Delgado (1992)] cannot be directly applied. In this article, we intend to fill this gap.

The rest of the article is organized as follows. In Section 2, we discuss the identifiability problem of ODE models. Then we introduce the numerical solution-based NLS estimators for (1.1) and (1.2), and study their asymptotic properties in Sections 3 and 4, respectively. The asymptotic properties of the proposed estimators, including strong consistency, rate of convergence and asymptotic normalities, are established using the tools of empirical processes [Pollard (1984, 1990), Pakes and Pollard (1989), van der Vaart and Wellner (1996), Ma and Kosorok (2005),

Wellner and Zhang (2007)] and the sieve methods [Grenander (1981), Shen and Wong (1994), Huang (1996) and Shen (1997)]. We perform simulation studies to investigate the finite-sample performance of the proposed estimation methods in Section 5. In this section, we also apply the proposed approaches to a set of ODE models for HIV dynamics. We provide a summary and discussion for the proposed methods in Section 6. Finally, the proofs of all the theoretical results are given in the Appendix.

2. Identifiability of ODE models. Identifiability of ODE models is a critical question to answer before parameter estimation. To verify the uniqueness of parameter estimates for given system inputs and outputs, both analytical and numerical techniques have been developed for ODE models since 1950s. Before jumping into technical details, two commonly used definitions of identifiability are given as follows [Bellman and Åström (1970), Cobelli, Lepschy and Jacur (1979), Walter (1987), Ljung and Glad (1994), Audoly et al. (2001), Jeffrey and Xia (2005)].

DEFINITION 1.   Globally identifiable: a system structure is said to be globally identifiable if for any two parameter vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ in the parameter space $\mathcal{B}$, $\mathbf{X}(t, \boldsymbol{\beta}_1) = \mathbf{X}(t, \boldsymbol{\beta}_2)$ can be satisfied for all t if and only if $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$.

However, global identifiability is a strong condition to satisfy and usually difficult to verify in practice. Therefore, the definition of at-a-point identifiability was introduced by Ljung and Glad (1994) and Quaiser and Mönnigmann (2009) as follows.

DEFINITION 2.   At-a-point identifiable: a system structure is said to be locally (or globally) identifiable at a point $\boldsymbol{\beta}_*$ if for any $\boldsymbol{\beta}$ within an open neighborhood of $\boldsymbol{\beta}_*$ (or within the entire parameter space), $\mathbf{X}(t, \boldsymbol{\beta}) = \mathbf{X}(t, \boldsymbol{\beta}_*)$ can be satisfied for all t if and only if $\boldsymbol{\beta} = \boldsymbol{\beta}_*$.

A number of methods have been proposed for identifiability analysis of ODE models, including *structural* [Bellman and Åström (1970), Ljung and Glad (1994), Xia and Moog (2003)], *practical* [e.g., Rodriguez-Fernandez, Egea and Banga (2006), Miao et al. (2008)] and *sensitivity-based* [e.g., Jolliffe (1972), Quaiser and Mönnigmann (2009)] approaches. Due to the limited space, we may not be able to provide an exhaustive list of publications on identifiability of ODE models. In this article, the structural identifiability analysis techniques are of particular interest mainly due to the theoretical completeness.

Various structural identifiability approaches have been proposed, such as power series expansion [Pohjanpalo (1978)], similarity transformation [Vajda et al. (1989), Chappel and Godfrey (1992)] and implicit function theorem method [Xia (2003), Xia and Moog (2003), Miao et al. (2008), Wu et al. (2008)]. Particularly, Ollivier (1990) and Ljung and Glad (1994) introduced another approach in the

framework of differential algebra [Ritt (1950), Kolchin (1973)]. The differential algebra approach is suitable to general nonlinear dynamic systems, and it has been successfully applied to nonlinear differential equation models, including models with time-varying parameters [Audoly et al. (2001)]. In this article, the differential algebra approach is employed to verify the identifiability of ODE models with both constant and time-varying parameters.

For most structural identifiability analysis techniques such as the implicit function theorem method and the differential algebra approach, a key step is the elimination of latent variables via taking derivatives and algebraic operations, which makes such techniques suitable for multivariate ODE models with partially observed state variables. After all unobserved state variables are eliminated, equations involving only given inputs, measured outputs and unknown parameters can be obtained. If we consider the parameters as unknowns, it is easy to verify that the identifiability of unknown parameters is determined by the number of roots of these equations.

For illustration purposes, we consider a classical HIV dynamic model with both constant and time-varying parameters [Nowak and May (2000), Huang, Rosenkranz and Wu (2003), Wu et al. (2005)] as an example:

$$(2.1) \quad \begin{cases} \dfrac{d}{dt}T_U(t) = \lambda - \rho T_U(t) - \eta(t)T_U(t)V(t), \\ \dfrac{d}{dt}T_I(t) = \eta(t)T_U(t)V(t) - \delta T_I(t), \\ \dfrac{d}{dt}V(t) = N\delta T_I(t) - cV(t), \end{cases}$$

where T_U is the concentration of uninfected target CD4+ T cells, T_I the concentration of infected CD4+ T cells, $V(t)$ the viral load, λ the proliferation rate of uninfected CD4+ T cells, ρ the death rate of uninfected CD4+ T cells, $\eta(t)$ the time-varying infection rate depending on antiviral drug efficacy, δ the death rate of infected cells, c the clearance rate of free virions, N the number of virions produced by a single infected cell on average. This model will also be used in our numerical studies in Section 5. For notational simplicity, let x_1, x_2 and x_3 denote T_U, T_I and V, and let $y_1 = T_U + T_I = x_1 + x_2$ and $y_2 = V = x_3$ denote the measurable outputs, respectively. Then (2.1) can be re-written as

$$(2.2) \quad \begin{cases} x_1' = \lambda - \rho x_1 - \eta(t)x_1 x_3, \\ x_2' = \eta(t)x_1 x_3 - \delta x_2, \\ x_3' = N\delta x_2 - cx_3, \end{cases}$$

where x_1', x_2' and x_3' denote the derivatives of x_1, x_2 and x_3, respectively. We adopt the following ranking for variable elimination [Ljung and Glad (1994)],

$$(2.3) \quad \eta \prec y_2 \prec y_1 \prec \boldsymbol{\beta} \prec x_3 \prec x_2 \prec x_1,$$

where $\boldsymbol{\beta} = (\lambda, \rho, N, \delta, c)^T$ is the vector of constant unknown parameters. By taking the higher order derivatives on both sides of (2.2) and using some algebra elimination techniques, we can eliminate x_1, x_2 and x_3 from (2.2) using the ranking (2.3) to obtain

$$(2.4) \qquad y_1^{(2)} + (\rho + \delta)y_1' + \delta\rho y_1 - \delta\lambda + \eta(t)y_2(y_1' + \delta y_1 - \lambda) = 0,$$

$$(2.5) \qquad y_2^{(2)} + (\delta + c)y_2' + \delta c y_2 - \eta(t)y_2(N\delta y_1 - y_2' - c y_2) = 0,$$

where $y_1^{(2)}$ and $y_2^{(2)}$ denote the second-order derivative of $y_1(t)$ and $y_2(t)$, respectively. Therefore, $\eta(t)$ can be expressed in terms of measurable system outputs and other constant unknown parameters either from (2.4) as

$$(2.6) \qquad \eta(t) = \frac{y_1^{(2)} + (\rho + \delta)y_1' + \delta\rho y_1 - \delta\lambda}{-y_2(y_1' + \delta y_1 - \lambda)}$$

or from (2.5) as

$$(2.7) \qquad \eta(t) = \frac{y_2^{(2)} + (\delta + c)y_2' + \delta c y_2}{y_2(N\delta y_1 - y_2' - c y_2)}.$$

Thus, $\eta(t)$ is identifiable if all the constant unknown parameters are identifiable. To verify the identifiability of all unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \eta)^T$, equations (2.6) and (2.7) can be combined to obtain

$$
\begin{aligned}
(2.8) \quad & y_1^{(2)} y_2 y_2' - y_1' y_2 y_2^{(2)} - \delta y_1 y_2 y_2^{(2)} + \lambda y_2 y_2^{(2)} - (\delta + c)y_1' y_2 y_2' \\
& + (\rho\delta + \rho + \delta - \delta^2 - \delta c)y_1 y_2 y_2' + c y_2 y_2' \\
& + \rho c y_1' y_2{}^2 + (\rho\delta c - \delta^2 c)y_1 y_2{}^2 - N\delta y_1 y_1^{(2)} y_2 \\
& + c y_1^{(2)} y_2{}^2 - N\delta(\rho + \delta)y_1 y_1' y_2 - N\delta^2 \rho y_1{}^2 y_2 + N\delta^2 \lambda y_1 y_2 = 0.
\end{aligned}
$$

The equation above only involves measurable system outputs $[(T_U + T_I)$, V and their derivatives] and constant unknown parameters. We assume that the derivatives of $(T_U + T_I)$ and V exist and are continuous up to order 2. Although the derivatives of $(T_U + T_I)$ and V are usually not directly measured in experiments, for theoretical identifiability analysis, they are known once $(T_U + T_I)$ and V are measured (e.g., via numerical evaluation). Finally, it can be verified that (2.8) is of order 0 and of degree > 1 in $\boldsymbol{\theta}$, so (2.8) satisfies the sufficient conditions given in Ljung and Glad (1994) and $\boldsymbol{\beta} = (\lambda, \rho, N, \delta, c)^T$ is thus at-a-point identifiable at the true parameter point. Therefore, $\eta(t)$ is also at-a-point identifiable at the true parameter point. For more detailed techniques for identifiability analysis of ODE models, we refer readers to the references listed above.

3. ODE models with constant parameters. Throughout this article, we let $\|\mathbf{a}\|$ be the Euclidean norm (or L_2 norm) of a vector (or a matrix) $\mathbf{a}$; $\|\mathbf{A}\|_\infty = \max_{1 \le i \le m} \sum_{j=1}^n |a_{ij}|$ be the supremum norm of an $m \times n$ matrix $\mathbf{A}$, where a_{ij} is the (i, j)th element of $\mathbf{A}$; $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}^T$ for a matrix $\mathbf{A}$; $C^r[a, b]$ be the class of functions with r-order continuous derivative on the interval $[a, b]$; $\|f\|_\infty = \sup_t |f(t)|$ be the supremum norm of a function f; and $x \wedge y$ denotes $\min(x, y)$. Moreover, for a random vector $\mathbf{Z} \sim P$, where P is a probability measure, we let $\|f(\mathbf{Z})\|_2 = \|f\|_{P,2} = (\int f^2 \, dP)^{1/2}$ be the $L_2(P)$-norm of a function f.

3.1. *Measurement model and estimator.* In this section, we consider ODE models with constant parameters, that is, equation (1.1), over the time range of interest $I = [t_0, T]$ $(-\infty < t_0 < T < +\infty)$, where the initial value $\mathbf{X}_0 = \mathbf{X}(t_0)$ is assumed to be known in this article. In reality, $\mathbf{X}(t)$ cannot be measured exactly and directly; instead, its surrogate $\mathbf{Y}(t)$ can be measured. For simplicity, here we assume an additive measurement error model to describe the relationship between $\mathbf{X}(t_i)$ and the surrogate $\mathbf{Y}(t_i)$,

$$(3.1) \qquad \mathbf{Y}(t_i) = \mathbf{X}(t_i) + \boldsymbol{\varepsilon}(t_i),$$

at random or fixed design time points $t_1, \ldots, t_n$, where the measurement errors $(\boldsymbol{\varepsilon}(t_1), \ldots, \boldsymbol{\varepsilon}(t_n))$ are independent with mean zero and a diagonal variance–covariance matrix Σ. Moreover, in the case of random design, assume that the measurement errors are independent of $\mathbf{X}(t)$. Equation (3.1) is called the observation or measurement equation.

If (1.1) does not have a closed-form solution, we need to resort to numerical techniques to obtain numerical solutions at discrete time points. In this article, we consider a general one-step numerical method. Let $t_0 = s_0 < s_1 < \cdots < s_{m-1} = T$ be grid points on the interval I, $h_j = s_j - s_{j-1}$ be the step size and $h = \max_{1 \le j \le m-1} h_j$ be the maximum step size, and $\mathbf{X}_j^h$ and $\mathbf{X}_{j+1}^h$ be the numerical approximations to the true solutions $\mathbf{X}(s_j)$ and $\mathbf{X}(s_{j+1})$, respectively, which can be typically written as

$$(3.2) \qquad \mathbf{X}_{j+1}^h = \mathbf{X}_j^h + h \boldsymbol{\Phi}(s_j, \mathbf{X}_j^h, \mathbf{X}_{j+1}^h, h),$$

where the specific form of $\boldsymbol{\Phi}$ depends on the numerical method. The common numerical methods include the Euler backward method, the trapezoidal rule, the r-stage Runge–Kutta algorithm (r is usually between 2 and 5), and so on. Among these algorithms, the 4-stage Runge–Kutta algorithm [Mattheij and Molenaar (2002), page 53, Hairer, Nørsett and Wanner (1993), page 134] has been well developed and widely used in practice. Therefore, we employ the 4-stage Runge–Kutta algorithm as an example in our numerical studies.

Define $e^h = \max_{0 \le j \le m-1} \|\mathbf{X}(s_j) - \mathbf{X}_j^h\|$, which is called the numerical error or the global discretization error [Hairer, Nørsett and Wanner (1993), page 159, Mattheij and Molenaar (2002), page 57]. If $e^h = O(h^p)$, p is called the order

of the numerical method. It is necessary to establish a relationship between the number of grid points m (or the maximum step size h) and the sample size of measurements n since the asymptotic properties of the proposed estimators are related to both numerical error and measurement error. To our best knowledge, this is the first attempt to establish such as a relationship.

Following Mattheij and Molenaar [(2002), page 58] the interpolation technique is commonly used if the measurement points $(t_i, i = 1, 2, \ldots, n)$ are not coincident with the grid points $(s_j, j = 1, 2, \ldots, m - 1)$ of the numerical method, and the cubic Hermite interpolation is often adopted. Let $\tilde{\mathbf{X}}(t, \boldsymbol{\beta})$ denote the interpolated numerical solution of $\mathbf{X}(t, \boldsymbol{\beta})$ obtained from the numerical method for given $\boldsymbol{\beta}$, and then (3.1) can be approximately rewritten as $\mathbf{Y}(t) \approx \tilde{\mathbf{X}}(t, \boldsymbol{\beta}_0) + \boldsymbol{\varepsilon}(t)$. The simple numerical solution-based NLS estimator $\hat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}_0$ minimizes

$$(3.3) \qquad \Xi_1(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{K} [Y_j(t_i) - \tilde{\mathbf{X}}_j(t_i, \boldsymbol{\beta})]^2.$$

Note that if the data are correlated or the measurement variances are heterogeneous, the weighted NLS can be used. The theoretical results can be extended to the weighted NLS. Also note that we can easily obtain the estimator $\hat{\mathbf{X}}(t) = \tilde{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_n)$ for $\mathbf{X}(t)$.

To minimize the NLS objective function (3.3), the standard gradient optimization methods may fail due to the complicated nonlinear ODE model and the NLS objective function may have multiple local minima or may be ill-behaved [Englezos and Kalogerakis (2001)]. Fortunately, various global optimization methods are available to more reliably solve the parameter estimation problem for ODE models, although the global optimization methods are very computationally intensive. Moles, Banga and Keller (2004) compared the performance and computational cost of seven global optimization methods, including the differential evolution method [Storn and Price (1997)]. Their results suggest that the differential evolution method outperforms the other six methods with a reasonable computational cost. Improved performance can be achieved using a hybrid method combining gradient methods and global optimization methods. A hybrid method based on the scatter search and sequential quadratic programming (SQP) has been proposed by Rodriguez-Fernandez, Egea and Banga (2006), who showed that the hybrid scatter search method is much faster than the differential evolution method for a simple HIV ODE model. In addition, Miao et al. (2008) also suggested that global optimization methods should be used for general nonlinear ODE models. Here we combine the differential evolution, the scatter search method and the SQP local optimization technique to implement our NLS minimization.

3.2. *Asymptotic properties.* In this section, we study the asymptotic properties of the proposed numerical solution-based NLS estimator when both measurement error and numerical error are considered. First we make the following assumptions:

A1. $\boldsymbol{\beta} \in \mathcal{B}$, where $\mathcal{B}$ is a compact subset of $\mathcal{R}^d$ with a finite diameter $R_{\boldsymbol{\beta}}$.

A2. $\Omega = \{\mathbf{X}(t, \boldsymbol{\beta}) : t \in I, \boldsymbol{\beta} \in \mathcal{B}\}$ is a closed and bounded convex subset of $\mathcal{R}^K$.

A3. There exist two constants $-\infty < \mathbf{c}_1 < \mathbf{c}_2 < +\infty$ such that $c_1 \le \mathbf{Y}(t) \le c_2$ for all $t \in I$.

A4. All partial derivatives of $\mathbf{F}(t, \mathbf{X}, \boldsymbol{\beta})$ up to order p with respect to t and $\mathbf{X}$ exist and are continuous.

A5. The numerical method for solving ODEs is of order p.

A6. For any $\boldsymbol{\beta} \in \mathcal{B}$, $E_t[\mathbf{X}(t, \boldsymbol{\beta}) - \mathbf{X}(t, \boldsymbol{\beta}_0)]^2 = 0$ if and only if $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

A7. The first and second partial derivatives, $\frac{\partial \mathbf{X}(t, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ and $\frac{\partial^2 \mathbf{X}(t, \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \, \partial \boldsymbol{\beta}^T}$, exist and are continuous and uniformly bounded for all $t \in I$ and $\boldsymbol{\beta} \in \mathcal{B}$.

A8. For the ODE numerical solution $\tilde{\mathbf{X}}(t, \boldsymbol{\beta})$, the first and second partial derivatives, $\frac{\partial \tilde{\mathbf{X}}(t, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ and $\frac{\partial^2 \tilde{\mathbf{X}}(t, \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \, \partial \boldsymbol{\beta}^T}$, exist and are continuous and uniformly bounded for all $t \in I$ and $\boldsymbol{\beta} \in \mathcal{B}$.

A9. Let $0 < c_3 < c_4 < \infty$ be two constants. For random design points, $t_1, \dots, t_n$ are i.i.d. The joint density function $\phi(t, \mathbf{y})$ of $(t, \mathbf{Y})$ satisfies $c_3 \le \phi(t, \mathbf{y}) \le c_4$ for all $(t, \mathbf{y}) \in [t_0, T] \times [\mathbf{c}_1, \mathbf{c}_2]$.

A10. The true parameter $\boldsymbol{\beta}_0$ is an interior point of $\mathcal{B}$.

A11. $\tilde{\boldsymbol{\beta}}$ is an interior point of $\mathcal{B}$, where $\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathcal{B}} E_0[\mathbf{Y}(t) - \tilde{\mathbf{X}}(t, \boldsymbol{\beta})]^T \times [\mathbf{Y}(t) - \tilde{\mathbf{X}}(t, \boldsymbol{\beta})]$ and E_0 is the expectation with respect to $P_{\boldsymbol{\beta}_0}$, the joint probability distribution of $(t, \mathbf{Y}(t))$ at true value $\boldsymbol{\beta}_0$.

A12. $\mathbf{V}_1 = \{E_t(\frac{\partial \mathbf{X}}{\partial \boldsymbol{\beta}_0} \frac{\partial \mathbf{X}}{\partial \boldsymbol{\beta}_0^T})\}^{-1} E_t(\frac{\partial \mathbf{X}}{\partial \boldsymbol{\beta}_0} \Sigma \frac{\partial \mathbf{X}}{\partial \boldsymbol{\beta}_0^T}) \{E_t(\frac{\partial \mathbf{X}}{\partial \boldsymbol{\beta}_0} \frac{\partial \mathbf{X}}{\partial \boldsymbol{\beta}_0^T})\}^{-1}$ is positive definite, where $E_t[g(t)]$ is expectation of function $g(t)$ with respect to t.

A13. $\tilde{\mathbf{V}}_1 = \{E_t(\frac{\partial \tilde{\mathbf{X}}}{\partial \tilde{\boldsymbol{\beta}}} \frac{\partial \tilde{\mathbf{X}}}{\partial \tilde{\boldsymbol{\beta}}^T})\}^{-1} E_0(\frac{\partial \tilde{\mathbf{X}}}{\partial \tilde{\boldsymbol{\beta}}} [\mathbf{Y}(t) - \tilde{\mathbf{X}}(t, \tilde{\boldsymbol{\beta}})]^{\otimes 2} \frac{\partial \tilde{\mathbf{X}}}{\partial \tilde{\boldsymbol{\beta}}^T}) \{E_t(\frac{\partial \tilde{\mathbf{X}}}{\partial \tilde{\boldsymbol{\beta}}} \frac{\partial \tilde{\mathbf{X}}}{\partial \tilde{\boldsymbol{\beta}}^T})\}^{-1}$ is positive definite.

Assumptions A1–A4 are general requirements for existence of numerical solutions of ODE models. Assumption A5 from Mattheij and Molenaar (2002, pages 55 and 56) defines the precision of the numerical algorithm. For example, the Euler backward method, the trapezoidal rule, the 4-stage and 5-stage Runge–Kutta are of order 1, 2, 4 and 5, respectively. Theorem 2.13 in Hairer, Nørsett and Wanner [(1993), page 153] provides sufficient and necessary conditions for the numerical method to be of order p. Theorems 3.1 and 3.4 in Hairer, Nørsett and Wanner [(1993), pages 157 and 160] give the magnitude of the numerical error of the numerical algorithms. Assumption A6 is required for identifiability and imposed for consistency. From Section 2, we know that the HIV model (2.1) is at-a-point identifiable at the true value $\boldsymbol{\beta}_0$. This result and assumption A9 are sufficient conditions for assumption A6 to be satisfied. Assumptions A7–A9 are needed for consistency. Assumptions A10–A13 are needed for the proof of asymptotic normality in Theorem 3.2.

THEOREM 3.1. *Assume that there exists a $\lambda > 0$ such that $h = O(n^{-\lambda})$, then under assumptions* A1–A10, *we have $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 \to 0$, almost surely under $P_{\boldsymbol{\beta}_0}$.*

THEOREM 3.2. (i) *For $h = O(n^{-\lambda})$ with $\lambda > 1/(p \wedge 4)$ where p is the order of the numerical method* (3.2), *under assumptions* A1–A10 *and* A12, *we have that $n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, \mathbf{V}_1)$.*

(ii) *For $h = O(n^{-\lambda})$ with $0 < \lambda \leq 1/(p \wedge 4)$, under assumptions* A1–A9, A11 *and* A13, *we have that $n^{1/2}(\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}) \xrightarrow{d} N(0, \tilde{\mathbf{V}}_1)$ with $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O(h^{(p \wedge 4)/2}) = O(n^{-\lambda(p \wedge 4)/2})$ and $\|\tilde{\mathbf{V}}_1 - V_1\| = O(h^{(p \wedge 4)/2}) = O(n^{-\lambda(p \wedge 4)/2})$.*

The detailed proofs of Theorems 3.1 and 3.2 are provided in the Appendix. The basic idea for the proofs is motivated by Pakes and Pollard (1989) in which a general central limit theorem is proved for a broad class of simulation estimators, that is, the objective function of the estimator is too complicated to evaluate directly, and instead the Monte Carlo simulation is used to approximate the objective function to obtain the estimator. The asymptotic properties of the simulation-based estimator are established using a general central limit theorem under nonstandard conditions given in Huber (1967) and Pollard (1985), which are often called the Huber–Pollard Z-theorem [see Theorem 3.3.1 in van der Vaart and Wellner (1996)]. In this article, we use the same theorem to prove the asymptotic normality of the numerical solution-based NLS estimator for ODEs. Similarly, our objective function $\Xi_1(\beta)$ in (3.3) cannot be directly evaluated; instead we have to approximate it by solving (1.1) numerically. Thus, similar ideas in Pakes and Pollard (1989) can be borrowed to establish the asymptotic results of our estimator in Theorems 3.1 and 3.2.

REMARK 1. Theorems 3.1 and 3.2 can be extended to fixed design points $t_i \in [t_0, T]$ $(i = 1, \ldots, n)$. Assume that there exists a distribution function $Q(t)$ with corresponding density $\varphi(t)$ such that, with $Q_n(t)$, the empirical distribution of $(t_1, \ldots, t_n)$, $\sup_{t \in [t_0, T]} |Q_n(t) - Q(t)| = O_p(n^{-1/2})$ and $\varphi(t)$ is bounded away from zero and has continuous second-order derivative on $[t_0, T]$. Define $E_t[g(t)]$ be the integral $\int_{t_0}^{T} g(t) \, dQ(t)$ for function $g(t)$. Similarly we can prove Theorems 3.1 and 3.2 for the fixed design if we replace assumption A9 by above assumption.

REMARK 2. From the proof of Theorem 3.2 in the Appendix, we still have $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O(h^{(p \wedge 4)/2})$ and $\|\tilde{V}_1 - V_1\| = O(h^{(p \wedge 4)/2})$ for a fixed constant h, which is independent of the sample size n. This suggests that, if the maximum step size h of the numerical algorithm for solving ODEs is a fixed constant, the numerical solution-based NLS estimator is not consistent. Instead the asymptotic bias is in the order of $h^{(p \wedge 4)/2}$.

Notice that our asymptotic results provide a theoretical foundation for the relationship between the numerical step size and sample size, that control numerical error and measurement error, respectively, for the widely-used NLS estimator

based on the numerical solutions of the ODEs. Intuitively, the smaller the numerical step size is, better the estimator is. However, a smaller step size will increase the computational cost and this may become a serious problem when the ODE system is large and the computational cost is high. It is important to study the trade-off between the numerical error and measurement error when the computational cost needs to be taken into consideration. Our theoretical results show that, only when the numerical step size, which controls the numerical error and computational cost, goes to zero with a rate faster than a particular rate $n^{-1/(p \wedge 4)}$, the numerical solution-based NLS estimator converges to the true value of the parameters with the rate of root-n. In addition, the asymptotic variance of the NLS estimator is the one as if the true solution $\mathbf{X}(t)$ is exactly known.

The asymptotic variance–covariance matrix needs to be estimated in order to perform statistical inference for unknown parameters $\boldsymbol{\beta}$. There are some standard methods that can be used. The first approach is to use the observed pseudo-information matrix based on the NLS objective function (3.3). The observed pseudo-information matrix is defined as $\mathcal{I}_1(\boldsymbol{\beta}) = -\frac{\partial^2 \Xi_1}{\partial \boldsymbol{\beta}^2}$. The standard error of $\hat{\boldsymbol{\beta}}_n$ can then be approximated by $\mathcal{I}_1^{-1/2}(\hat{\boldsymbol{\beta}}_n)/\sqrt{n}$. In practice, we have noted that the inverse of the observed pseudo-information matrix provides a reasonable approximation to the asymptotic variance–covariance matrix V_1. Rodriguez-Fernandez, Egea and Banga (2006) also proposed this approach for parameter inference in ODE models.

The second approach is the weighted bootstrap method [Ma and Kosorok (2005)]. Let W_i, $i = 1, \ldots, n$, denote n i.i.d. positive random weights with mean one [$E(W) = 1$] and variance one [$\text{Var}(W) = 1$]. The weights, W_i are independent of $\{\boldsymbol{\beta}, t, \mathbf{Y}(t)\}$. For (1.1), the weighted M-estimator $\hat{\boldsymbol{\beta}}_n^0$ satisfies

$$\hat{\boldsymbol{\beta}}_n^0 = \arg\min \sum_{i=1}^{n} \sum_{j=1}^{K} W_i [Y_j(t_i) - \tilde{\mathbf{X}}_j(t_i, \boldsymbol{\beta})]^2.$$

From Corollary 2 and Theorem 2 in Ma and Kosorok (2005), given $\{t_i, \mathbf{Y}(t_i)\}$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^0 - \hat{\boldsymbol{\beta}}_n)$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ have the same limiting distribution, then the weighted M-estimator $\hat{\boldsymbol{\beta}}_n^0$ can be used for inference on $\hat{\boldsymbol{\beta}}_n$.

Note that the empirical bootstrap has been used for statistical inference for ODE models [Joshi, Seidel-Morgenstern and Kremling (2006)]. However, the asymptotic properties of the empirical bootstrap estimators are quite difficult to derive. This is why we propose to use the weighted bootstrap method instead of the empirical bootstrap approach.

4. ODE models with both constant and time-varying parameters.

4.1. *Measurement model and estimator.* In this section, we consider (1.2) with both constant and time-varying parameters, where the initial value $\mathbf{X}_0 = \mathbf{X}(t_0)$ is

assumed to be known. Again, $\mathbf{X}(t)$ is not observed directly in practice; instead, we observe its surrogate $\mathbf{Y}(t)$ through (3.1).

Let $\mathcal{A}$ be the following class of functions,

$$(4.1) \qquad \mathcal{A} = \{\eta \in C^{\mu}[t_0, T] : |\eta^{(\mu)}(z_1) - \eta^{(\mu)}(z_2)| \le L|z_1 - z_2|^{\gamma}\},$$

where μ is a nonnegative integer, $\gamma \in (0, 1]$, $\varrho = \mu + \gamma > 0.5$, and L an unknown constant. The smoothness assumption is often used in nonparametric curve estimation. Usually, either $\varrho = 1$ (i.e., $\mu = 0$ and $\gamma = 1$) or $\varrho = 2$ (i.e., $\mu = 1$ and $\gamma = 1$) should be satisfied in various situations. Denote $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \eta)^T$. Then the parameter space is denoted by $\Theta = \{\boldsymbol{\theta} : \boldsymbol{\beta} \in \mathcal{B}, \eta \in \mathcal{A}\} = \mathcal{B} \times \mathcal{A}$.

In this article, we use the method of sieves to approximate $\eta_0(t)$ on the support interval $[t_0, T]$ of t. The basic idea of the sieve approach is to approximate an infinite-dimensional parameter space Θ by a series of finite-dimensional parameter spaces Θ_n, which depend on the sample size n, and then to estimate the parameter on the finite-dimensional spaces Θ_n instead of Θ. The concept of sieve was first proposed by Grenander (1981). Since then, the sieve method has been a powerful tool in the area of nonparametric and semiparametric statistics [Shen and Wong (1994), Huang (1996), van der Vaart and Wellner (1996), Section 3.4, Shen (1997), Huang and Rossini (1997), Huang (1999), He, Fung and Zhu (2002), Xue, Lam and Li (2004) and Huang, Zhang and Zhou (2007)].

Here we apply the sieve estimation method to (1.2) with a time-varying parameter. First, we approximate $\eta(t)$ by B-spline functions on the support interval I of t. Let $t_0 = u_0 < u_1 < \cdots < u_q = T$ be a partition of the interval I, where $q = O(n^v)$ $(0 < v < 0.5)$ is a positive integer such that $\max_{1 \le j \le q} |u_j - u_{j-1}| = O(n^{-v})$. Then we have $N = q + l$ normalized B-spline basis functions of order $l + 1 \ge \varrho$ [see Huang (2003), page 1618] that form a basis for the linear spline space. We denote these basis functions in the forms of a vector $\boldsymbol{\pi}(t) = (B_1(t), \ldots, B_N(t))^T$ with which $\eta(t)$ can be approximated by $\boldsymbol{\pi}(t)^T \boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)^T \in \mathcal{R}^N$ is the spline coefficient vector with $\boldsymbol{\alpha}_0$ corresponding to $\eta_0(t)$. Such approximation is extensively used in nonparametric and semiparametric problems [Stone (1985), Shen and Wong (1994), Shen (1997), Huang (1999) and Huang (2003)]. The readers are referred to Schumaker [(1981), page 118] for more details about the construction of the basis functions. Regression spline approximation to a nonparametric function can always be expressed as a linear function of basis functions so that the problem of time-varying coefficients can be transformed into an estimation problem for a number of constant parameters. Thus the estimation methods and computational algorithms developed for (1.1) with constant coefficients in Section 3 can be employed for (1.2) with both constant and time-varying parameters.

For any $\boldsymbol{\theta}_i \in \mathcal{B} \times \mathcal{A}$ $(i = 1, 2)$, we define a distance

$$(4.2) \qquad d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| + \|\eta_1 - \eta_2\|_2.$$

Denote set

$$\mathcal{A}_n = \left\{ \eta(t) = \sum_{i=1}^{N} B_i(t)\alpha_i : \max_{1 \leq i \leq N} |\alpha_i| \leq \ell_n \right\},$$

where $\ell_n \leq n^{(2l-1)/[2l'(2l+1)]}$ with a constant l' arbitrarily close to l [see Shen (1997), page 2560], then $\Theta_n = \{\theta : \beta \in \mathcal{B}, \eta \in \mathcal{A}_n\} = \mathcal{B} \times \mathcal{A}_n$ can be used as a sieve of Θ. In fact, for any $\theta = (\beta^T, \eta)^T \in \Theta$, by Corollary 6.21 in Schumaker (1981), there exists $\eta_n \in \mathcal{A}_n$ such that $\|\eta_n - \eta\|_\infty = O_p(n^{-\nu\varrho})$. Denote $\theta_n = (\beta^T, \eta_n)^T \in \Theta_n$, then $d(\theta, \theta_n) = O_p(n^{-\nu\varrho})$. Equation (1.2) now becomes

$$\frac{d\mathbf{X}(t)}{dt} \approx \mathbf{F}\{t, \mathbf{X}(t), \beta, \pi(t)^T \alpha\}.$$

For this approximation model, let $\tilde{\mathbf{X}}(t, \beta, \pi(t)^T \alpha)$ be the numerical approximation of $\mathbf{X}(t, \beta, \eta(t))$ that can be obtained from the same numerical algorithm as described in Section 3. Equation (3.1) can be approximated by $\mathbf{Y}(t) \approx \tilde{\mathbf{X}}(t, \beta, \pi(t)^T \alpha_0) + \varepsilon(t)$. The numerical solution-based sieve NLS estimator $\hat{\theta}_n = (\hat{\beta}_n^T, \hat{\eta}_n)^T$ is defined as

$$(4.3) \qquad \hat{\theta}_n = \underset{\theta \in \Theta_n}{\arg\inf} \, \Xi_2(\theta) = \underset{\theta \in \Theta_n}{\arg\inf} \sum_{i=1}^{n} \sum_{j=1}^{K} [Y_j(t_i) - \tilde{X}_j(t_i, \beta, \eta(t))]^2.$$

When we substitute the sieve NLS estimators $\hat{\theta}_n$ into the numerical approximation, we can obtain the estimator $\hat{\mathbf{X}}(t) = \tilde{\mathbf{X}}(t, \hat{\beta}_n, \hat{\eta}_n(t))$.

4.2. *Asymptotic properties.* The empirical objective function for the sieve NLS method proposed in Section 4.1 is a second-order loss function which is not a likelihood function. We cannot use the standard information calculation of the maximum likelihood estimator (MLE) based on orthogonal projections in semiparametric models [Bickel et al. (1993)], and the asymptotic normality theory for semiparametric MLEs obtained by Huang (1996, Theorem 6.1) does not apply to our case. Fortunately, Ma and Kosorok (2005) and Wellner and Zhang (2007) extended the Huang's asymptotic normality results to more general semiparametric M-estimators by using a so-called pseudo-information calculation. We are able to employ these new asymptotic results to asymptotic properties of the proposed sieve NLS estimator, and the following additional assumptions are needed:

B1. The true time-varying parameter $\eta_0(\cdot) \in \mathcal{A}$, where $\mathcal{A}$ is denoted in (4.1).
B2. All partial derivatives of $\mathbf{F}$ up to order p with respect to $t, \mathbf{X}$, and η, respectively, exist and are continuous.
B3. For any $\beta \in \mathcal{B}$ and $\eta \in \mathcal{A}$, $E_t[\mathbf{X}(t, \beta, \eta(t)) - \mathbf{X}(t, \beta_0, \eta_0(t))]^2 = 0$ if and only if $\beta = \beta_0$ and $P\{t : \eta(t) = \eta_0(t)\} = 1$.

B4. The first and second partial Fréchet-derivatives [van der Vaart and Wellner (1996), page 373] in the norm d defined in (4.2), $\frac{\partial \mathbf{X}(t,\boldsymbol{\beta},\eta)}{\partial \boldsymbol{\beta}}$, $\frac{\partial \mathbf{X}(t,\boldsymbol{\beta},\eta)}{\partial \eta}$, $\frac{\partial^2 \mathbf{X}(t,\boldsymbol{\beta},\eta)}{\partial \boldsymbol{\beta} \, \partial \boldsymbol{\beta}^T}$, $\frac{\partial^2 \mathbf{X}(t,\boldsymbol{\beta},\eta)}{\partial \boldsymbol{\beta} \, \partial \eta}$ and $\frac{\partial^2 \mathbf{X}(t,\boldsymbol{\beta},\eta)}{\partial \eta^2}$ exist and are continuous and uniformly bounded for all $t \in I$, $\boldsymbol{\beta} \in \mathcal{B}$ and $\eta \in \mathcal{A}$.

B5. For the ODE numerical solution $\tilde{\mathbf{X}}(t,\boldsymbol{\beta},\eta)$, the first and second partial Fréchet-derivatives in the norm d, $\frac{\partial \tilde{\mathbf{X}}(t,\boldsymbol{\beta},\eta)}{\partial \boldsymbol{\beta}}$, $\frac{\partial \tilde{\mathbf{X}}(t,\boldsymbol{\beta},\eta)}{\partial \eta}$, $\frac{\partial^2 \tilde{\mathbf{X}}(t,\boldsymbol{\beta},\eta)}{\partial \boldsymbol{\beta} \, \partial \boldsymbol{\beta}^T}$, $\frac{\partial^2 \tilde{\mathbf{X}}(t,\boldsymbol{\beta},\eta)}{\partial \boldsymbol{\beta} \, \partial \eta}$ and $\frac{\partial^2 \tilde{\mathbf{X}}(t,\boldsymbol{\beta},\eta)}{\partial \eta^2}$ exist and are continuous and uniformly bounded for all $t \in I$, $\boldsymbol{\beta} \in \mathcal{B}$ and $\eta \in \mathcal{A}$.

B6. For $K \geq 2$, $\mathbf{V}_2 = \mathbf{S}_1^{-1} \mathbf{S}_2 (\mathbf{S}_1^{-1})^T$ is positive definite, where $\mathbf{S}_1$ and $\mathbf{S}_2$ are defined in (A.5) and (A.6) in the Appendix, respectively.

B7. v satisfies the restrictions $0.25/\varrho < v < 0.5$ and $v(2 + \varrho) > 0.5$, where ϱ is the measure of smoothness of $\eta(t)$ defined in assumption (B1).

THEOREM 4.1. *Assume that there exists a $\lambda > 0$ such that $h = O(n^{-\lambda})$ and under assumptions* A1–A4, A9, A10 *and* B1–B5, *we have $d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) \to 0$, almost surely under $P_{\boldsymbol{\theta}_0}$.*

THEOREM 4.2. *Assume that there exists a $\lambda > 1/[2(p \wedge 4)]$ such that $h = O(n^{-\lambda})$ where p is the order of the numerical algorithm (3.2), and under assumptions* A1–A4, A9, A10 *and* B1–B5, *we have $d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = O_p(n^{-v\varrho} + n^{-(1-v)/2})$.*

From Theorem 4.2, we know that $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_p(n^{-v\varrho} + n^{-(1-v)/2})$ and $\|\hat{\eta}_n(t) - \eta_0(t)\|_2 = O_p(n^{-v\varrho} + n^{-(1-v)/2})$. If $v = 1/(1 + 2\varrho)$, the rate of convergence of $\hat{\eta}_n$ is $n^{-\varrho/(1+2\varrho)}$, which is the same as the optimal rate of the standard nonparametric function estimation [Stone (1982)]. Theorem 4.3 below states that the rate of weak convergence of $\hat{\boldsymbol{\beta}}_n$ achieves $n^{-1/2}$ under some additional assumptions.

THEOREM 4.3. *For the maximum step size $h = O(n^{-\lambda})$ with $\lambda > 1/(p \wedge 4)$, under assumptions* A1–A4, A9, A10 *and* B1–B7, *and $K \geq 2$, we have $n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, \mathbf{V}_2)$.*

REMARK 3. For the case $h = O(n^{-\lambda})$ with $1/[2(p \wedge 4)] < \lambda \leq 1/(p \wedge 4)$, similar results to case (ii) in Theorem 3.2 can be obtained.

For $K = 1$, Theorem 4.3 does not hold, since in this case the special perturbation direction $a^*(t)$ given in (A.4) is $\frac{\partial X}{\partial \beta_0} / \frac{\partial X}{\partial \eta_0}$, which leads to both $\mathbf{S}_1$ in (A.5) and $\mathbf{S}_2$ in (A.6) to be zero (see the proof of Theorem 4.3 in the Appendix). In this article, we consider one special case that we assume there exists an additive relationship

between $\boldsymbol{\beta}$ and $\eta(\cdot)$ as follows:

$$(4.4) \qquad \frac{dX(t)}{dt} = F\{t, X(t), \beta + \eta(t)\},$$

which is a special case of (1.2), then the function $X(t)$ has the form of $X(t, \beta + \eta(t))$. In this case, we are able to establish similar asymptotic normality results under the identifiability constraint $E_t \eta(t) = 0$. Note that Schick (1986) studied a similar problem under a semiparametric regression model and used the same identifiability constraint for the unknown function $\eta(t)$ to establish the asymptotic normality for the constant parameters. We follow a similar idea and use B-spline approximation for $\eta(t)$. We center the B-spline estimator of $\eta(t)$ as follows:

$$\hat{\eta}_n(t_i) \approx \sum_{i=1}^{N} B_j(t_i)\hat{\alpha}_j - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N} B_j(t_i)\hat{\alpha}_j = \sum_{j=1}^{N} \hat{\alpha}_j \left[B_j(t_i) - \frac{1}{n}\sum_{i=1}^{n} B_j(t_i) \right],$$

which is subject to the constraints $\sum_{i=1}^{n} \hat{\eta}_n(t_i) = 0$. Under similar assumptions, the strong consistency and the rate of weak convergence of the estimators, similar to those of Theorems 4.1 and 4.2, can be obtained. In particular, the asymptotic normality can be established as follows:

PROPOSITION 1. *For (4.4) with $K = 1$, when the maximum step size $h = O(n^{-\lambda})$ with $\lambda > 1/(p \wedge 4)$, under assumptions A1–A4, A9, A10, B1–B5, B7 and in addition $E_t[\eta(t)] = 0$, we have $n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \overset{d}{\to} N(0, V_3)$, where $V_3 = \sigma_0^2 \{E_t(\frac{\partial X}{\partial \xi})^2\}^{-1}$ with $\xi = \beta_0 + \eta_0(t)$.*

The proof of this proposition is different from that of Theorem 4.3 and is given in the Appendix.

REMARK 4.   By combining Theorem 4.3 and Proposition 1, we can see that the proposed sieve NLS estimator is asymptotically normal with a convergence rate of $\sqrt{n}$ for $K \geq 2$ under assumption B6, but we are only able to prove the result for a special ODE model (4.4) for $K = 1$. This is because the asymptotic covariance V_2 defined in B6 is always singular in the case of $K = 1$, and is only possibly nonsingular in the case of $K \geq 2$. Since V_2 is always singular for $K = 1$, we derive the asymptotic distribution for the special ODE model (4.4) using a different approach which results in Proposition 1.

Similar approaches proposed in Section 3 can be used to estimate the asymptotic variance–covariance matrix for $(\hat{\boldsymbol{\beta}}_n, \hat{\eta}_n(t))$. For the first approach, the observed pseudo-information matrix can be evaluated by replacing $\eta(t)$ with the spline approximation $\pi^T(t)\alpha$, that is, to rewrite the objective function $\Xi_2(\boldsymbol{\theta})$ in the expression (4.3) as $\Xi_2(\boldsymbol{\beta}, \boldsymbol{\alpha})$. Then the observed pseudo-information matrix $\mathcal{I}_2(\boldsymbol{\beta}, \boldsymbol{\alpha})$

can be defined as

$$
\mathcal{I}_2(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \begin{pmatrix} -\dfrac{\partial^2 \Xi_2}{\partial \boldsymbol{\beta}^2} & -\dfrac{\partial^2 \Xi_2}{\partial \boldsymbol{\beta} \, \partial \boldsymbol{\alpha}} \\[2ex] -\dfrac{\partial^2 \Xi_2}{\partial \boldsymbol{\alpha} \, \partial \boldsymbol{\beta}} & -\dfrac{\partial^2 \Xi_2}{\partial \boldsymbol{\alpha}^2} \end{pmatrix}.
$$

The standard error of $(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\alpha}}_n)$ is approximately $\mathcal{I}_2^{-1/2}(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\alpha}}_n)/\sqrt{n}$ from which the standard error of $\hat{\boldsymbol{\beta}}_n$ can be obtained. We also find that the inverse of the observed pseudo-information matrix provides a reasonable approximation to $\mathbf{V}_2$ via our simulation studies in the next section.

Similarly the weighted bootstrap method can also be used. For (1.2), the weighted M-estimators $(\hat{\boldsymbol{\beta}}_n^0, \hat{\boldsymbol{\alpha}}_n^0)$ satisfy

$$
(\hat{\boldsymbol{\beta}}_n^0, \hat{\boldsymbol{\alpha}}_n^0) = \arg\min \sum_{i=1}^n \sum_{j=1}^K W_i [Y_j(t_i) - \tilde{\mathbf{X}}_j(t_i, \boldsymbol{\beta}, \boldsymbol{\pi}(t)^T \boldsymbol{\alpha})]^2.
$$

Based on Corollary 2 and Theorem 2 in Ma and Kosorok (2005), given $\{t_i, \mathbf{Y}(t_i)\}$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^0 - \hat{\boldsymbol{\beta}}_n)$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ have the same limiting distribution which can be used to justify the weighted bootstrap for inference on $\hat{\boldsymbol{\beta}}_n$ and $\hat{\eta}_n(t)$.

5. Numerical studies. In this section, we consider the HIV dynamic model described in Section 2. Recall that in this system, $T_U(t)$, $T_I(t)$ and $V(t)$ are state variables and $(\lambda, \rho, \delta, N, c, \eta(t))^T$ are kinetic parameters. By introducing the time-varying infection rate $\eta(t)$ in this HIV dynamic model, the model can flexibly describe the long-term viral dynamics. In clinical studies, only viral load, $V(t)$ and total CD4+ T cell count, $T(t) = T_U(t) + T_I(t)$, are closely monitored and measured over time. For easy illustration and computational simplicity, we fix the parameters ρ and δ in our numerical studies, and our objective is to estimate three constant parameters and one time-varying parameter, $(\lambda, N, c, \eta(t))^T$ based on measurements of viral load and total CD4+ T cell count.

5.1. *Monte Carlo simulation study.* The following parameter values and initial conditions were used to simulate observation data for (2.1): $T_U(0) = 600$, $T_I(0) = 30$, $V(0) = 10^5$, $\lambda = 36$, $\rho = 0.108$, $N = 1000$, $\delta = 0.5$, $c = 3$. For comparison purpose, we generated the measurement data of $V(t)$ and $T(t)$ for four scenarios in our simulation studies: (i) $\eta(t) = \eta$ is a small constant, $\eta = 9.5 \times 10^{-6}$; (ii) $\eta(t)$ is time-varying but with a smaller (10%) variation, $\eta(t) = 9 \times 10^{-5} \times \{1 - 0.9\cos(\pi t/400)\}$; (iii) $\eta(t) = \eta$ is a larger constant, $\eta = 3.84 \times 10^{-5}$; and (iv) $\eta(t)$ is time-varying but with a large (10-fold) variation, $\eta(t) = 9 \times 10^{-5} \times \{1 - 0.9\cos(\pi t/40)\}$. Note that for cases (i) and (iii), the values of constant η were chosen to be approximately the average of $\eta(t)$ over the period of time interval for cases (ii) and (iv), respectively.

Let $y_1 = T = T_U + T_I$ denote the total number of infected and uninfected CD4+ T cells and $y_2 = V$ denote the viral load, the measurement models are given as follows:

$$y_{1i} = T(t_i) + \varepsilon_{1i},$$

$$y_{2i} = V(t_i) + \varepsilon_{2i},$$

where ε_{1i} and ε_{2i} are independent and follow normal distributions with mean zero and variances σ_{1i}^2 and σ_{2i}^2, respectively. The HIV dynamic model was numerically solved within the time range $[0, 20]$ to generate the simulated data at each time interval of 0.5 using the 4-stage Runge–Kutta algorithm. Consequently, the corresponding sample size is 40. The 20% measurement errors were added to the numerical results of the ODE model according to the observation equations above. We applied the proposed estimation methods in Sections 3 and 4 to the simulated data for the 4 cases to evaluate the performance of the proposed estimators and the effect of the model misspecification. To stabilize the computational algorithm, we log-transformed the data. We also fixed parameters ρ and δ as their true values.

For evaluating the performance of the estimation methods, we define the average relative estimation error (ARE) as

$$\text{ARE} = \frac{1}{M} \sum_{j=1}^{M} \frac{|\hat{\theta}_j - \theta|}{|\theta|} \times 100\%,$$

where $\hat{\theta}_j$ is the estimate of the parameter vector θ from the jth simulation data set, and $M = 500$ is the total number of simulation runs.

In Table 1, the AREs of the constant parameters (λ, N, c) are listed. In addition, we also report σ_{ODE}^2 as the average of the estimated variance by the observed pseudo-information matrix and σ_{emp}^2 as the empirical variance based on simulation runs. Based on these results, we can see that, when the change of $\eta(t)$ is small as a function of time t or η is a small constant, the estimation of parameters is always good by fitting a constant η model as observed by the low ARE values. However, when the change of $\eta(t)$ is large or η is a large constant, misspecification of $\eta(t)$ may produce large AREs for all parameter estimates. In particular, when $\eta(t)$ is time-varying with a large variation, using a constant η model may result in very poor estimates for all constant parameters. The variance estimates based on the pseudo-information agree well with the empirical estimates based on simulations, which shows that the pseudo-information-based variance estimate is reasonably good. The evaluation of the bootstrap variance estimation is prohibited in our simulation study due to high computational cost.

In Figure 1, the average trajectories of estimated $\eta(t)$ are compared to the true trajectories of $\eta(t)$ for four different scenarios. From this figure, we observed a similar trend as the constant parameter estimates. The misspecification of $\eta(t)$ produces estimation error, in particular for the cases with a large variation of $\eta(t)$

TABLE 1

Simulation results for constant η and the time-varying $\eta(t)$ models. The ARE is calculated based on 500 simulation runs for the HIV dynamic model. In addition, σ^2_{ODE} is the average of the estimated variance by the observed pseudo-information, and σ^2_{emp} is the empirical variance based on simulations. The sample size is n = 40 and the noise level is about 20%

| Change of $\eta(t)$	True $\eta(t)$ model	Fitted $\eta(t)$ model	λ			N			c		
			ARE(%)	σ^2_{ODE}	σ^2_{emp}	ARE(%)	σ^2_{ODE}	σ^2_{emp}	ARE(%)	σ^2_{ODE}	σ^2_{emp}
Small	Constant	Constant	3.19	2.49	1.91	17.7	3.23e+04	4.94e+04	17.4	0.313	0.425
		Time-varying	6.45	9.82	8.77	22.9	7.14e+04	8.63e+04	20.5	0.593	0.635
	Time-varying	Constant	3.77	2.38	2.08	17.9	3.36e+04	4.71e+04	19.8	0.331	0.432
		Time-varying	6.40	9.16	8.55	22.6	6.53e+04	7.93e+04	20.9	0.543	0.637
Large	Constant	Constant	6.29	8.19	12.1	72.5	1.13e+06	8.53e+05	67.3	9.22	6.75
		Time-varying	7.34	9.19	15.6	88.8	3.25e+06	1.13e+06	82.5	26.2	8.96
	Time-varying	Constant	94.2	13.7	7.02	994	5.86e+07	1.25e+08	1899	1660	3780
		Time-varying	15.6	31.5	48.2	29.5	1.67e+05	1.67e+05	25.1	1.81	1.37
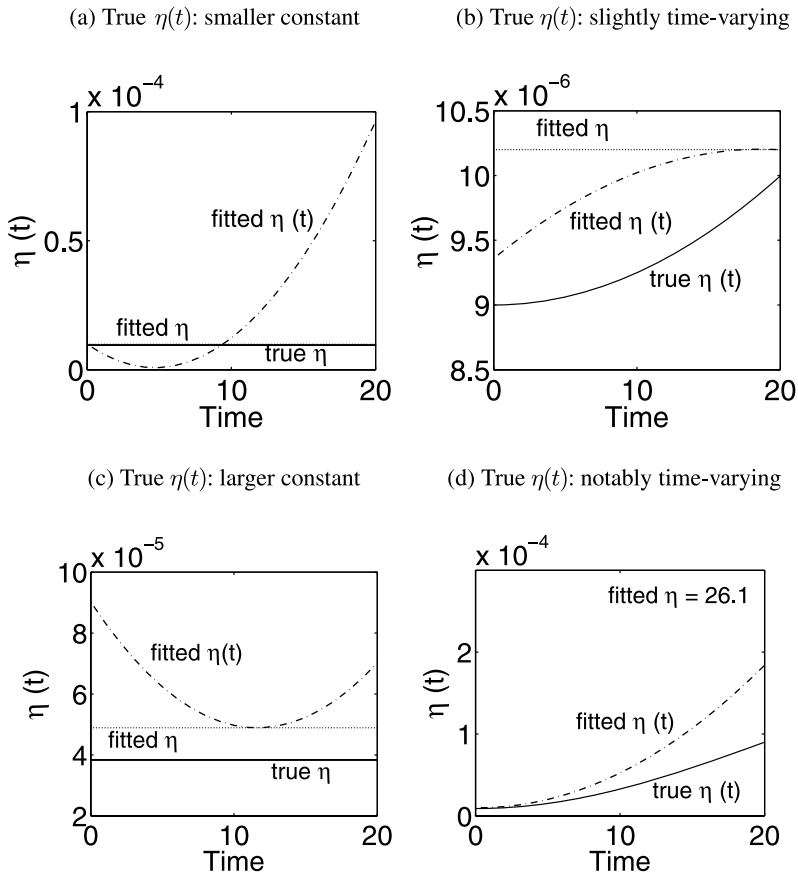
FIG. 1. *Simulation results for constant η and the time-varying $\eta(t)$ models. In each figure, the true model of η (solid), the constant η model (dotted) and the time-varying $\eta(t)$ model (dash-dotted) are plotted and compared.*

or a large constant η. When the model of $\eta(t)$ is correctly specified, the estimates based on the proposed methods are reasonably good. In order to evaluate the robustness of the proposed approach, we also performed further simulation studies for a complex function $\eta(t) = 9.0 \times 10^{-6} + 9.0 \times 10^{-7} \times t \times \{1 - 0.5\sin(\pi t/5.8)\}$ under the same simulation settings (i.e., 40 time points, 20% error, 500 simulation runs). The results suggest that the sieve estimator can still capture the essential pattern of the complex $\eta(t)$ reasonably well (plots not shown).

5.2. *Application to AIDS clinical data.*   To illustrate applicability and feasibility of our proposed methods and theories, we also applied the proposed estimation methods to fit the HIV dynamic model to a clinical data set obtained from an HIV-1 infected patient who was treated with an antiretroviral therapy. Very frequent viral load measurements were collected from this patient after initiating the antiretro-

viral regimen: 13 measurements during the first day, 14 measurements from day 2 to week 2, and then one measurement at weeks 4, 8, 12, 14, 20, 24, 28, 32, 36, 40, 44, 48, 52, 56, 64, 74 and 76, respectively. In addition, the measurements of total CD4+ T cell counts were also taken at Day 1, weeks 2 and 4, and monthly thereafter. Equation (2.1) was used to estimate HIV kinetic parameters using the viral load and total CD4+ T cell data.

For simplicity of illustration and computation, we fixed the initial conditions of the state variables in (2.1) as $T_U(0) = 1$, $T_I(0) = 551$, $V(0) = 6.38 \times 10^4$, which were derived from the baseline measurements. We also fixed two parameters, as in the simulation study, $\rho = 0.10$ and $\delta = 0.434$, which were taken from the estimates in literature. Our objective is to estimate the three constant parameters (λ, N, c) and the time-varying parameter $\eta(t)$ as in the simulation study. As we proposed in Section 4, we employed B-splines to approximate $\eta(t)$. We positioned the spline knots at equally-spaced time points (the log-time scale was used since the distribution of observation time points is highly-skewed). We selected the order of splines and the number of spline knots using the model selection criterion AICc given by

$$\text{AICc} = n \ln\left(\frac{\text{RSS}}{n}\right) + \frac{2nk}{n - k - 1},$$

where RSS is the residual of the sum of squares obtained from the NLS model fitting, n is the total number of observations and k is the number of unknown parameters [including the coefficients in the spline representation of $\eta(t)$]. Note that as a practical guideline, if the number of unknown parameters exceeds $n/40$ (where n is the sample size), the AICc instead of AIC should be used. For our clinical data, the sample size n is equal to 65, and the number of unknown parameters varies between 6 and 13 for different scenarios, which is much larger than $n/40 = 65/40 = 1.6$. Thus the AICc is more appropriate for our applications. In general, the AICc converges to the AIC as the sample size gets larger, thus the AICc is often suggested to be employed regardless of the sample size [Burnham and Anderson (2004)]. For our application, we used AICc and compared the models with the splines of order 3 and 4, and the number of knots from 3 to 10. In Table 2, the AICc values for these different models are reported, from which the best model was selected as the spline with order 3 and 5 knots for $\eta(t)$ approximation.

We used the weighted bootstrap method to calculate both the confidence intervals for the constant parameters and the confidence bands for the time-varying parameter. The basic idea of the weighted bootstrap method is provided in Sections 3.2 and 4.2. For the computational implementation, we first generated a positive random weight for each data point in the raw data set from the exponential distribution with mean one and variance one. By repeating this step, a large number of (say, 1000) sets of weights can be generated. Second, for each set of weights, the ODE model is fitted to the data to obtain parameter estimates by minimizing

TABLE 2
*Model selection results for B-spline approximation of the
time-varying parameter $\eta(t)$*

Model	Spline order	Number of knots	AICc
1		3	−222.3
2		4	−242.8
3		5	−252.8
4		6	−243.8
5	3	7	−250.6
6		8	−246.0
7		9	−246.8
8		10	−244.4
9		3	−
10		4	−233.2
11		5	−230.3
12		6	−242.6
13	4	7	−249.1
14		8	−245.5
15		9	−244.9
16		10	−240.5

the weighted residual sum of squares (see Sections 3 and 4). Recall that the time-varying parameter in the model has been approximated by B-splines, then both the constant parameters and the constant B-spline coefficients are actually estimated. Once the estimates of the B-spline coefficients are obtained, we construct the B-splines which approximate the time-varying parameter. Thus, we eventually obtain 1000 estimates for each constant parameter and 1000 B-splines for each time-varying parameter. Third, for each constant parameter, we select the 2.5% and 97.5% quantiles of the 1000 estimates to form the 95% confidence intervals for this parameter. For the time-varying parameter, at a single time point, the 1000 B-splines have 1000 values. We also select the 2.5% and 97.5% quantiles of the 1000 values at this time point to eventually form the 95% pointwise confidence bands for the time-varying parameter.

Model fitting results are given in Figure 2 and Table 3. From Figures 2(a) and (b), we can see that the fitting is reasonably good for both CD4+ T cell counts and viral load data. The estimates of constant parameters (λ, N, c) are listed in Table 3, and the 95% bootstrap confidence intervals of the estimates are also provided. The uninfected cell proliferation rate (λ) was estimated as 46.52 cells per day, the average number of virions produced by one infected cell (N) was estimated as 1300 per day and the clearance rate of free virions was 4.35 per day which corresponds to a half-life of 3.8 hours. All these estimates are in the ballpark of similar estimates from other methods [Perelson et al. (1996, 1997)]. In
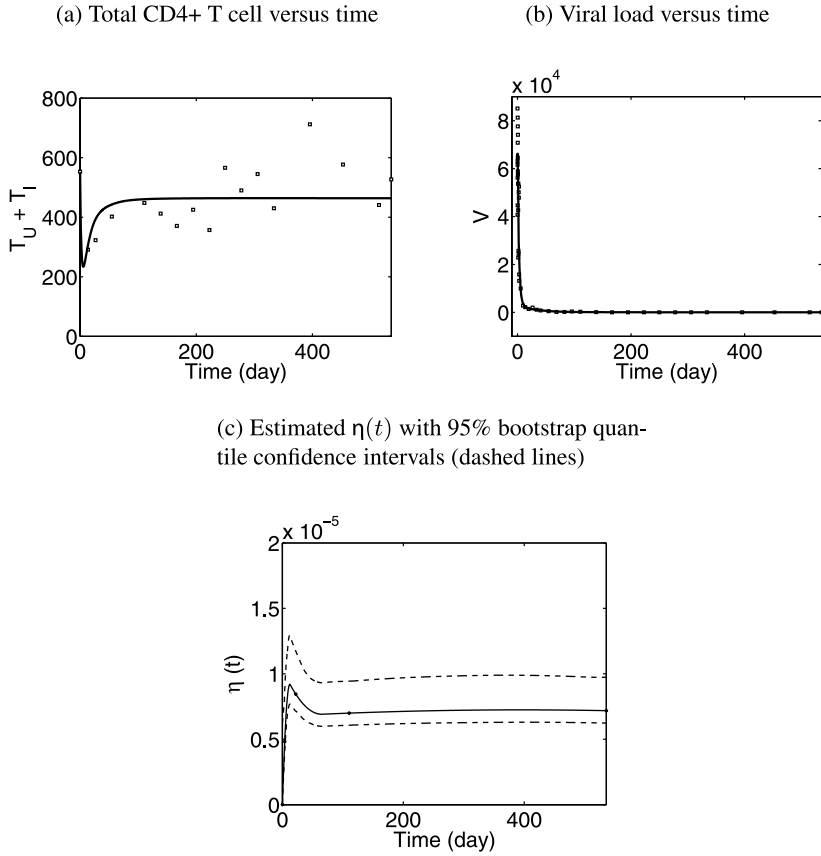
(a) Total CD4+ T cell versus time   (b) Viral load versus time

(c) Estimated $\eta(t)$ with 95% bootstrap quan-
tile confidence intervals (dashed lines)

FIG. 2.   *Model fitting results with $\eta(t)$ approximated by B-splines of order 3 and 5 knots.*

Figure 2(c), the estimated trajectory of the time-varying parameter $\eta(t)$ (the viral infection rate), is plotted with 95% bootstrap quantile confidence intervals, which shows an initial fluctuation but converges to a constant after 2 to 3 months.

6. Discussion. In this paper, we have systematically studied numerical solution-based NLS estimators for general nonlinear ODE models which the

TABLE 3
The constant parameter estimation results

Parameter	Estimate	95% confidence interval
λ	46.52	[43.20, 51.04]
N	1300.39	[251.93, 4628.26]
c	4.35	[0.98, 14.83]

closed-form solutions are not available. Both constant and time-varying parameters are considered. For the model involved time-varying parameters, we formulated the estimator under the framework of sieve approach. Our main contribution is the establishment of the asymptotic properties for the proposed numerical solution-based NLS estimators (including the sieve NLS estimator for the time-varying parameter) with consideration of both numerical error and measurement error. Our results show that if the maximum step size of the p-order numerical algorithm goes to zero at a rate faster than $n^{-1/(p \wedge 4)}$, the numerical error is negligible compared with the measurement error. This provides guidance in selecting the step size for numerical evaluations of ODEs. Moreover, we have shown that the numerical solution-based NLS estimator and the sieve NLS estimator for the model with a time-varying parameter are strongly consistent. The sieve estimator of constant parameters is asymptotically normal with the same asymptotic co-variance as that of the case where the true solution is exactly known, while the estimator of the time-varying parameter has an optimal convergence rate under some regularity conditions. We also obtained the theoretical results for the case when the step size of the ODE numerical solver does not go to zero fast enough or the numerical error is comparable to the measurement error [see case (ii) of Theorem 3.2 and Remark 3]. To our best knowledge, this is the first time that the sieve method has been extended to the case of ODE models which have no closed-form solutions, and the sieve-based theories were used to establish the asymptotic results and construct confidence intervals (bands) for both constant and time-varying parameters. Note that we only considered a single time-varying parameter in the model, but the methodologies can be extended to multiple time-varying parameters although it is more tedious to implement.

Note that the NLS estimators have good properties under some assumptions and are more accurate compared to other estimates such as those proposed in Ramsay et al. (2007), Chen and Wu (2008) and Liang and Wu (2008). But the price that we have to pay is the high computational cost to obtain the NLS estimates. To reduce the computational burden, we may use the rough estimates from other methods [Ramsay et al. (2007), Chen and Wu (2008), Liang and Wu (2008)] to narrow down the search range for the NLS optimization algorithm. More efficient optimization algorithms may also be employed to speed up the computation. We are also considering to parallel our global optimization algorithms on high-performance computers. Hopefully these efforts can help us to handle a reasonable size of ODE models.

This article only considered the initial value problem (IVP), that is, the initial conditions are assumed to be given. In practice, the initial conditions can be estimated from the data. However, the generalizations of the theoretical results to the cases of estimated initial conditions and other boundary value problems as well as constraints on parameters are not trivial. Also note that, if there is more than one time-varying parameter in the model, similar identifiability techniques in Section 2

may be applied to these parameters one by one, sequentially. Spline approximation to these multiple time-varying parameters can be used for estimation. But the computation and theoretical results are more complicated in this case. However, these generalizations are worth further investigations in future.

APPENDIX: PROOFS

LEMMA 1. *Under conditions* A1–A5, $\sup_{t \in I} \|\tilde{\mathbf{X}}(t, \boldsymbol{\beta}) - \mathbf{X}(t, \beta)\|_\infty = O(h^{p \wedge 4})$ *for any given* $\boldsymbol{\beta} \in \mathcal{B}$ *in* (1.1).

PROOF.   By Theorem 3.4 in Hairer, Nørsett and Wanner [(1993), page 160] under conditions A1–A5, for the pth order numerical algorithm (3.2) for (1.1), its global discretization error satisfies

$$\max_{0 \leq i \leq m-1} \|\tilde{\mathbf{X}}(s_i, \boldsymbol{\beta}) - \mathbf{X}(s_i, \boldsymbol{\beta})\|_\infty = O(h^p) \qquad \text{for given } \boldsymbol{\beta} \in \mathcal{B}.$$

When t is not coincident with the grid points of the numerical algorithm, the cubic Hermite interpolation [de Boor (1978), page 51] will be used to obtain the solution at time t. In this case,

$$\sup_{t \in I \setminus \{s_i : 0 \leq i \leq m-1\}} \|\tilde{\mathbf{X}}(t, \boldsymbol{\beta}) - \mathbf{X}(t, \boldsymbol{\beta})\|_\infty = O(h^4).$$

Then it follows that

$$\sup_{t \in I} \|\tilde{\mathbf{X}}(t, \boldsymbol{\beta}) - \mathbf{X}(t, \boldsymbol{\beta})\|_\infty$$

$$\leq \sup_{t \in I \setminus \{s_i : 0 \leq i \leq m-1\}} \|\tilde{\mathbf{X}}(t, \boldsymbol{\beta}) - \mathbf{X}(t, \boldsymbol{\beta})\|_\infty$$

$$+ \max_{t \in \{s_i : 0 \leq i \leq m-1\}} \|\tilde{\mathbf{X}}(t, \boldsymbol{\beta}) - \mathbf{X}(t, \boldsymbol{\beta})\|_\infty$$

$$= O(h^4) + O(h^p).$$

In general, h is less than 1, $O(h^4) + O(h^p) = O(h^{p \wedge 4})$, which completes the proof.   □

Moreover, Lemma 1 can be extended to the ODE model (1.2) with both constant and time-varying parameters, since for this model, it can be verified that the result of Theorem 3.1 in Hairer, Nørsett and Wanner [(1993), page 157] is still valid for any given $\boldsymbol{\beta} \in \mathcal{B}$ and $\eta \in \mathcal{A}$ under condition B2 (it can be derived using the Taylor expansion and the Chain rule), which leads to the same conclusion as Theorem 3.4 in Hairer, Nørsett and Wanner [(1993), page 160]. For Theorems 3.1 and 3.2, the proofs for the univariate and multivariate cases are the same. For presentation and notation simplicity, we only outline the proof for the univariate case below.

PROOF OF THEOREM 3.1.   Denote $\tilde{M}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n[Y(t_i) - \tilde{X}(t_i,\boldsymbol{\beta})]^2$, $M_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n[Y(t_i) - X(t_i,\boldsymbol{\beta})]^2$ and $M(\boldsymbol{\beta}) = [Y(t) - X(t,\boldsymbol{\beta})]^2$.

First, we claim that $E_0[M(\boldsymbol{\beta})]$ reaches its unique minimum at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. In fact,

$$
\begin{aligned}
E_0[M(\boldsymbol{\beta})] &= E_0[Y(t) - X(t,\boldsymbol{\beta})]^2 \\
&= E_0[Y(t) - X(t,\boldsymbol{\beta}_0) + X(t,\boldsymbol{\beta}_0) - X(t,\boldsymbol{\beta})]^2 \\
&= E_0[Y(t) - X(t,\boldsymbol{\beta}_0)]^2 + E_t[X(t,\boldsymbol{\beta}_0) - X(t,\boldsymbol{\beta})]^2 \\
&= E_0[\varepsilon(t)]^2 + E_t[X(t,\boldsymbol{\beta}_0) - X(t,\boldsymbol{\beta})]^2 \\
&\geq E_0[\varepsilon(t)]^2 = E_0[M(\boldsymbol{\beta}_0)],
\end{aligned}
$$

where the third equality holds because the intersection term equals zero according to the following calculation:

$$
\begin{aligned}
E_0[\varepsilon(t)]&[X(t,\boldsymbol{\beta}_0) - X(t,\boldsymbol{\beta})] \\
&= E_t E_0\{[\varepsilon(t)][X(t,\boldsymbol{\beta}_0) - X(t,\boldsymbol{\beta})]|t\} \\
&= E_t\{[X(t,\boldsymbol{\beta}_0) - X(t,\boldsymbol{\beta})]E_0[\varepsilon(t)]\} \\
&= 0,
\end{aligned}
$$

because of $E_0[\varepsilon(t)] = 0$. Moreover, $E_t[X(t,\boldsymbol{\beta}) - X(t,\boldsymbol{\beta}_0)]^2 = 0$ if and only if $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ from assumption A6. Thus the above claim holds. Under assumption A10, it follows that the first-order derivative $\frac{\partial E_0[M(\boldsymbol{\beta})]}{\partial \boldsymbol{\beta}}$ of $E_0[M(\boldsymbol{\beta})]$ at $\boldsymbol{\beta}_0$ equals to zero and the second-order derivative $\frac{\partial^2 E_0[M(\boldsymbol{\beta})]}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}^T}$ of $E_0[M(\boldsymbol{\beta})]$ at $\boldsymbol{\beta}_0$ is positive definite. By assumptions A7 and A9, the second-order derivative of $E_0[M(\boldsymbol{\beta})]$ in a small neighborhood of $\boldsymbol{\beta}_0$ is bounded away from 0 and ∞. Then the second-order Taylor expansion of $E_0[M(\boldsymbol{\beta})]$ gives that there exists a constant $0 < C < \infty$ such that

$$
E_0[M(\hat{\boldsymbol{\beta}}_n) - M(\boldsymbol{\beta}_0)] \geq C\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|^2.
$$

Thus it is sufficient to prove $E_0[M(\boldsymbol{\beta}_0)] - E_0[M(\hat{\boldsymbol{\beta}}_n)] \to 0$, a.s.

Let $N_1(\varepsilon, \mathcal{Q}, \mathcal{F})$ be the covering number of the class $\mathcal{F}$ in the probability measure $\mathcal{Q}$, as given in Pollard (1984, page 25). From Lemma 4.1 in Pollard (1990), we have that $N_1(\varepsilon, L_2, \mathcal{B}) \leq (\frac{3R_\beta}{\varepsilon})^d$. Let $\mathcal{F}_n$ be the set $\{M_n(\boldsymbol{\beta}): \boldsymbol{\beta} \in \mathcal{B}\}$. With the Taylor expansion, for any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{B}$, we can easily obtain

$$
|M_n(\boldsymbol{\beta}_1) - M_n(\boldsymbol{\beta}_2)| \leq C\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|,
$$

where C is some constant. Then for any probability measure Q, we have

$$
\sup_Q N_1(\varepsilon, Q, \mathcal{F}_n) \leq N_1(\varepsilon/C, L_2, \mathcal{B}) \leq C\left(\frac{1}{\varepsilon}\right)^d \qquad \text{for } 0 < \varepsilon < 1.
$$

Then by Theorem II.37 in Pollard (1984), $\sup_{\boldsymbol{\beta}} |M_n(\boldsymbol{\beta}) - E_0 M(\boldsymbol{\beta})| \to 0$, a.s., under $P_{\boldsymbol{\beta}_0}$. Then we have $M_n(\hat{\boldsymbol{\beta}}_n) - E_0[M(\hat{\boldsymbol{\beta}}_n)] \to 0$ and $M_n(\boldsymbol{\beta}_0) - E_0[M(\boldsymbol{\beta}_0)] \to 0$, a.s.

Next, by Lemma 1,

$$\tilde{M}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}[Y(t_i) - \tilde{X}(t_i, \boldsymbol{\beta})]^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}[Y(t_i) - X(t_i, \boldsymbol{\beta}) + O(n^{-\lambda(p\wedge 4)})]^2$$

(A.1)

$$= \frac{1}{n}\sum_{i=1}^{n}[Y(t_i) - X(t_i, \boldsymbol{\beta})]^2 + O(n^{-\lambda(p\wedge 4)})$$

$$= M_n(\boldsymbol{\beta}) + O(n^{-\lambda(p\wedge 4)}).$$

Then

$$\tilde{M}_n(\hat{\boldsymbol{\beta}}_n) - E_0[M(\boldsymbol{\beta}_0)]$$

$$\geq \tilde{M}_n(\hat{\boldsymbol{\beta}}_n) - E_0[M(\hat{\boldsymbol{\beta}}_n)]$$

$$= M_n(\hat{\boldsymbol{\beta}}_n) + O(n^{-\lambda(p\wedge 4)}) - E_0[M(\hat{\boldsymbol{\beta}}_n)]$$

and

$$\tilde{M}_n(\hat{\boldsymbol{\beta}}_n) - E_0[M(\boldsymbol{\beta}_0)]$$

$$\leq \tilde{M}_n(\boldsymbol{\beta}_0) - E_0[M(\boldsymbol{\beta}_0)]$$

$$= M_n(\boldsymbol{\beta}_0) + O(n^{-\lambda(p\wedge 4)}) - E_0 M(\boldsymbol{\beta}_0).$$

Hence $\tilde{M}_n(\hat{\boldsymbol{\beta}}_n) - E_0[M(\boldsymbol{\beta}_0)] \to 0$, a.s. Thus

$$|E_0[M(\hat{\boldsymbol{\beta}}_n)] - E_0[M(\boldsymbol{\beta}_0)]|$$

$$\leq |\tilde{M}_n(\hat{\boldsymbol{\beta}}_n) - E_0[M(\hat{\boldsymbol{\beta}}_n)]| + |\tilde{M}_n(\hat{\boldsymbol{\beta}}_n) - E_0[M(\boldsymbol{\beta}_0)]| \to 0 \qquad \text{a.s.}$$

Since $\boldsymbol{\beta}_0$ is the unique minimum point for $E_0[M(\boldsymbol{\beta})]$, $\hat{\boldsymbol{\beta}}_n$ is almost surely consistent with respect to $P_{\boldsymbol{\beta}_0}$. □

PROOF OF THEOREM 3.2. For the proof of part (i), it suffices to verify conditions of Theorem 2 in Pollard (1985). Denote $\tilde{G}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}[Y(t_i) - \tilde{X}(t_i, \boldsymbol{\beta})]\frac{\partial \tilde{X}(t_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$, $G_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}[Y(t_i) - X(t_i, \boldsymbol{\beta})]\frac{\partial X(t_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ and $G(\boldsymbol{\beta}) = E_0[Y(t) - X(t, \boldsymbol{\beta})]\frac{\partial X(t, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$. Obviously, $\tilde{G}_n(\hat{\boldsymbol{\beta}}_n) = 0$ and $G(\boldsymbol{\beta}_0) = E_t E_0(\{[Y(t) - X(t, \boldsymbol{\beta}_0)]\frac{\partial X(t, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0}\}|t) = 0$ from $E_0[Y(t)|t] = X(t, \boldsymbol{\beta}_0)$.

First, we verify the following result: $\sqrt{n}[\tilde{G}_n(\boldsymbol{\beta}_0) - G(\boldsymbol{\beta}_0)] \xrightarrow{d} N(0, H_1)$. For fixed t, according to the multivariate inequality of Kolmogorov type for L_2-norms of derivatives [Babenko, Kofanov and Pichugov (1996), page 9], we

have $\|\frac{\partial \tilde{\mathbf{X}}(t,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} - \frac{\partial \mathbf{X}(t,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\| \leq C\|\frac{\partial^2 \tilde{\mathbf{X}}(t,\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} - \frac{\partial^2 \mathbf{X}(t,\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\|_{\infty}^{1/2}\|\tilde{\mathbf{X}}(t,\boldsymbol{\beta}) - \mathbf{X}(t,\boldsymbol{\beta})\|_{\infty}^{1/2} \leq$
$C'\|\tilde{\mathbf{X}}(t,\boldsymbol{\beta}) - \mathbf{X}(t,\boldsymbol{\beta})\|_{\infty}^{1/2}$ for two constants C and C', where the second inequality
holds because of the uniform boundedness of both $\frac{\partial^2 \mathbf{X}(t,\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$ and $\frac{\partial^2 \tilde{\mathbf{X}}(t,\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$ under con-
ditions A7 and A8. Based on $\sup_{t \in I} \|\tilde{\mathbf{X}}(t,\boldsymbol{\beta}) - \mathbf{X}(t,\beta)\|_{\infty} = O(n^{-\lambda(p \wedge 4)})$ from
Lemma 1, it follows that $\|\frac{\partial \tilde{\mathbf{X}}(t,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} - \frac{\partial \mathbf{X}(t,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\| = O(n^{-\lambda(p \wedge 4)/2})$. Considering that
$Y(t_i) - X(t_i, \boldsymbol{\beta}_0)$ and $\frac{\partial \tilde{X}(t_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0}$ are bounded, we have

$$
\begin{aligned}
\sqrt{n}&[\tilde{G}_n(\boldsymbol{\beta}_0) - G(\boldsymbol{\beta}_0)] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [Y(t_i) - \tilde{X}(t_i, \boldsymbol{\beta}_0)] \frac{\partial \tilde{X}(t_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0} \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [Y(t_i) - X(t_i, \boldsymbol{\beta}_0) + O(n^{-\lambda(p \wedge 4)})] \left[\frac{\partial X(t_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0} + O(n^{-\lambda(p \wedge 4)/2}) \right] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [Y(t_i) - X(t_i, \boldsymbol{\beta}_0)] \frac{\partial X(t_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0} + O(n^{-\lambda(p \wedge 4)/2 + 1/2}).
\end{aligned}
$$

When $\lambda > 1/(p \wedge 4)$, $O(n^{-\lambda(p \wedge 4)/2 + 1/2}) = o(1)$. So for the above expression, we
have

$$
\begin{aligned}
\sqrt{n}&[\tilde{G}_n(\boldsymbol{\beta}_0) - G(\boldsymbol{\beta}_0)] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [Y(t_i) - X(t_i, \boldsymbol{\beta}_0)] \frac{\partial X(t_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0} + o(1) \\
&= \sqrt{n}[G_n(\boldsymbol{\beta}_0) - G(\boldsymbol{\beta}_0)] + o(1).
\end{aligned}
$$

Based on the general central limit theorem, $\sqrt{n}[G_n(\boldsymbol{\beta}_0) - G(\boldsymbol{\beta}_0)] \to N(0, \mathbf{H}_1)$
with

$$
\mathbf{H}_1 = E_0[Y(t) - X(t, \boldsymbol{\beta}_0)]^2 \left[\frac{\partial X(t, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0} \right]^{\otimes 2} = \sigma_0^2 E_t \left[\frac{\partial X(t, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0} \right]^{\otimes 2}.
$$

Second, let $\delta_n \downarrow 0$. For $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq \delta_n$, we want to show that

$$
\sqrt{n}[\tilde{G}_n(\boldsymbol{\beta}) - G(\boldsymbol{\beta})] - \sqrt{n}[\tilde{G}_n(\boldsymbol{\beta}_0) - G(\boldsymbol{\beta}_0)] = o_p(1).
$$

In fact, from the first step above, for any $\boldsymbol{\beta} \in \mathcal{B}$, we have that $\sqrt{n}[\tilde{G}_n(\boldsymbol{\beta}) - G_n(\boldsymbol{\beta})] = o_p(1)$. Then

$$
\begin{aligned}
\sqrt{n}&[\tilde{G}_n(\boldsymbol{\beta}) - G(\boldsymbol{\beta})] - \sqrt{n}[\tilde{G}_n(\boldsymbol{\beta}_0) - G(\boldsymbol{\beta}_0)] \\
&= \sqrt{n}[G_n(\boldsymbol{\beta}) - G(\boldsymbol{\beta})] - \sqrt{n}[G_n(\boldsymbol{\beta}_0) - G(\boldsymbol{\beta}_0)] + o_p(1).
\end{aligned}
$$

From Lemma 4.1 in Pollard (1990), we have that $N_1(\varepsilon, L_2, \mathcal{B}) \leq (\frac{3R}{\varepsilon})^d$. Let Λ_n be the set $\{G_n(\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathcal{B}\}$ for any $X \in \mathcal{X}$. Using a Taylor series expansion, for any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{B}$, we can easily obtain

$$|G_n(\boldsymbol{\beta}_1) - G_n(\boldsymbol{\beta}_2)| \leq C \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|,$$

where C is some constant. Then for any probability measure Q, we have

$$N_1(\varepsilon, L_2(Q), \Lambda_n) \leq N_1(\varepsilon/C, L_2, \mathcal{B}) \leq C \left(\frac{1}{\varepsilon}\right)^d,$$

and thus

$$\log N_1(\varepsilon, L_2(Q), \Lambda_n) \leq d \log \frac{1}{\varepsilon}.$$

Since $\int_0^1 \log(1/\varepsilon) \, d\varepsilon < \infty$, Λ_n is a P-Donsker class by Theorem 2.5.2 in van der Vaart and Wellner (1996). Hence $\sqrt{n}[G_n(\boldsymbol{\beta}) - G(\boldsymbol{\beta})] - \sqrt{n}[G_n(\boldsymbol{\beta}_0) - G(\boldsymbol{\beta}_0)] = o_p(1)$.

Third, with some simple calculations, we have $G(\boldsymbol{\beta}) = E_t[X(t, \boldsymbol{\beta}_0) - X(t, \boldsymbol{\beta})] \frac{\partial X(t, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$, then

$$\frac{\partial G(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\int \left\{\frac{\partial X(t, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\}^{\otimes 2} d\Phi(t) + \int [X(t, \boldsymbol{\beta}_0) - X(t, \boldsymbol{\beta})] \frac{\partial^2 X(t, \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \, \partial \boldsymbol{\beta}^T} d\Phi(t)$$

and $\frac{\partial G(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta} = \boldsymbol{\beta}_0} = -E_t\{\frac{\partial X(t, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0}\}^{\otimes 2}$. Denote $\mathbf{H}_2 = E_t\{\frac{\partial X(t, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0}\}^{\otimes 2}$. Then by using the Taylor series expansion again, the function $G(\boldsymbol{\beta})$ is Fréchet-differentiable at $\boldsymbol{\beta}_0$ with nonsingular derivative $\mathbf{H}_2$.

Thus all conditions of Theorem 2 in Pollard (1985) are satisfied, then Theorem 3.2(i) holds with $\mathbf{V}_1 = \mathbf{H}_2^{-1} \mathbf{H}_1 (\mathbf{H}_2^{-1})^T = \sigma_0^2 \{E_t[\frac{\partial X(t, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0}]^{\otimes 2}\}^{-1}$.

For the proof of case (ii) of Theorem 3.2, it is easy to verify the conditions of Theorem 2 in Pollard (1985) for the asymptotic normality. Now we just need to show $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + O(h^{(p \wedge 4)/2})$ and $\tilde{V}_1 = V_1 + O(h^{(p \wedge 4)/2})$. Denote $\tilde{M}(\boldsymbol{\beta}) = [Y(t) - \tilde{X}(t, \boldsymbol{\beta})]^2$ and $\tilde{G}(\boldsymbol{\beta}) = E_0[Y(t) - \tilde{X}(t, \boldsymbol{\beta})] \frac{\partial \tilde{X}(t, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$. Since $E_0[\tilde{M}(\boldsymbol{\beta})]$ reaches its minimum at $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$, then the first-order derivative of $E_0[\tilde{M}(\boldsymbol{\beta})]$ at $\tilde{\boldsymbol{\beta}}$ equals 0, that is, $\tilde{G}(\tilde{\boldsymbol{\beta}}) = 0$. Then similar to the proof of case (i) above, we have

$$\tilde{G}(\boldsymbol{\beta}) = E_0[Y(t) - \tilde{X}(t, \boldsymbol{\beta})] \frac{\partial \tilde{X}(t, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

$$= E_0[Y(t) - X(t, \boldsymbol{\beta}) + O(h^{p \wedge 4})] \left[\frac{\partial X(t, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} + O(h^{(p \wedge 4)/2})\right]$$

$$= G(\boldsymbol{\beta}) + O(h^{(p \wedge 4)/2}).$$

It follows that $\tilde{G}(\tilde{\boldsymbol{\beta}}) = G(\tilde{\boldsymbol{\beta}}) + O(h^{(p\wedge 4)/2})$, then $G(\tilde{\boldsymbol{\beta}}) = O(h^{(p\wedge 4)/2})$ from $\tilde{G}(\tilde{\boldsymbol{\beta}}) = 0$. The Taylor series expansion yields that there exist constants $0 < c_1, c_2 < \infty$ such that

$$c_1 \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \le |G(\tilde{\boldsymbol{\beta}}) - G(\boldsymbol{\beta}_0)| \le c_2 \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|.$$

Thus $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O(h^{(p\wedge 4)/2})$ from $G(\boldsymbol{\beta}_0) = 0$. Similarly we can show that $\|\tilde{\mathbf{V}}_1 - \mathbf{V}_1\| = O(h^{(p\wedge 4)/2})$. $\square$

Some definitions and notation are necessary in order to prove Theorems 4.1–4.3. Denote $\tilde{M}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^K [Y_j(t_i) - \tilde{\mathbf{X}}_j(t_i, \boldsymbol{\beta}, \eta(t_i))]^2$, $M_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^K [Y_j(t_i) - \mathbf{X}_j(t_i, \boldsymbol{\beta}, \eta(t_i))]^2$ and $M(\boldsymbol{\theta}) = \sum_{j=1}^K [Y_j - \mathbf{X}_j(t, \boldsymbol{\beta}, \eta(t))]^2$. We define a semidistance ρ on Θ as

$$\rho^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = E_0\{(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \dot{M}_1(\boldsymbol{\theta}) + \dot{M}_2(\boldsymbol{\theta})[\eta - \eta_0]\}^2,$$

where $\dot{M}_1$ is the score function of M for $\boldsymbol{\beta}$, and $\dot{M}_2$ is the score operator of M for η, both evaluated at the true parameter value $\boldsymbol{\theta}_0$. Similarly to the proof in Huang and Rossini [(1997), page 966] when $\mathbf{V}_2(\boldsymbol{\theta}_0)$, defined in assumption B6, is positive definite, and $\dot{M}_1$ and $\dot{M}_2$ are bounded away from $+\infty$ and $-\infty$, if $\rho(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = O_p(r_n)$, then $d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = O_p(r_n)$; and if $\rho(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) \to 0$ almost surely under $P_{\boldsymbol{\theta}_0}$, then $d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) \to 0$ almost surely under $P_{\boldsymbol{\theta}_0}$.

PROOF OF THEOREM 4.1.   Similarly to the proof of Theorem 3.1, we have that $E_0[M(\boldsymbol{\theta}_0)]$ reaches its unique minimum at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. It follows that

$$E_0[M(\boldsymbol{\theta}_0) - M(\hat{\boldsymbol{\theta}}_n)] \ge C\rho^2(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0),$$

where C is some constant. Thus if $E_0[M(\boldsymbol{\theta}_0)] - E_0[M(\hat{\boldsymbol{\theta}}_n)] \to 0$, almost surely under $P_{\boldsymbol{\theta}_0}$, then $d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) \to 0$, almost surely under $P_{\boldsymbol{\theta}_0}$.

Let $\mathcal{A}_n^\delta$ be the set $\{\eta \in \mathcal{A}_n, \|\eta - \eta_{n0}\|_2 \le \delta\}$ and $N_2(\varepsilon, L_\infty, \mathcal{A}_n^\delta)$ be its bracketing number with respect to L_∞ [see Definition 2.1.6, van der Vaart and Wellner (1996)], where η_{n0} is the map point of η_0 in the sieve $\mathcal{A}_n$. By the calculation of Shen and Wong [(1994), page 597] for any $\varepsilon \le \delta$, we have

$$N_2(\varepsilon, L_\infty, \mathcal{A}_n^\delta) \le C(\delta/\varepsilon)^N,$$

where $N = q + l$ is the number of B-splines basis functions. Let $\mathcal{F}_n$ be the set $\{M_n(\boldsymbol{\theta}) : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le \delta, \eta \in \mathcal{A}_n, \|\eta - \eta_{n0}\|_2 \le \delta\}$. For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta_n$, we can easily obtain

$$|M_n(\boldsymbol{\theta}_1) - M_n(\boldsymbol{\theta}_2)| \le C(\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| + \|\eta_1 - \eta_2\|_\infty)$$

using Taylor's expansion. Hence

$$\begin{aligned}
N_2(\varepsilon, L_\infty, \mathcal{F}_n) &\le N_1(\varepsilon/2, L_2, \mathcal{B}) \times N_2(\varepsilon/2, L_\infty, \mathcal{A}_n^\delta) \\
&\le C(3R_d/\varepsilon)^d (\delta/\varepsilon)^N \\
&\le C'(1/\varepsilon)^{N+d}.
\end{aligned}$$

Note that, since $N_2(\varepsilon, L_\infty, \mathcal{F}_n)$ depends on n in the above expression, we cannot directly use Theorem II.37 in Pollard (1984) to obtain $\sup_{\mathcal{F}_n} |M_n(\boldsymbol{\theta}) - E_0[M(\boldsymbol{\theta})]| \to 0$, a.s., under $P_{\boldsymbol{\theta}_0}$. Fortunately, we can still get this result based on (A.2) in Xue, Lam and Li (2004). Thus we have $M_n(\hat{\boldsymbol{\theta}}_n) - E_0[M(\hat{\boldsymbol{\theta}}_n)] \to 0$ and $M_n(\boldsymbol{\theta}_{n0}) - E_0[M(\boldsymbol{\theta}_{n0})] \to 0$, a.s., where $\boldsymbol{\theta}_{n0}$ is the map point of $\boldsymbol{\theta}_0$ in the sieve Θ_n.

From the extension of Lemma 1 for any given $\boldsymbol{\beta} \in \mathcal{B}$ and $\eta(t) \in \mathcal{A}$ in (1.2), similarly to (A.1), we have

$$(A.2) \qquad \tilde{M}_n(\boldsymbol{\theta}) = M_n(\boldsymbol{\theta}) + O\big(n^{-\lambda(p\wedge4)}\big).$$

Then the remaining steps are similar to those in the proof of Theorem 3.1. $\quad\square$

PROOF OF THEOREM 4.2. We apply Theorem 3.4.1 in van der Vaart and Wellner (1996) to obtain the rate of convergence.

For $\boldsymbol{\theta}_{n0}$ in the proof of Theorem 4.1, define $\boldsymbol{\theta}_{n0} \mapsto \rho_1(\boldsymbol{\theta}, \boldsymbol{\theta}_{n0})$ be a map from Θ_n to $[0, \infty)$ as $\rho_1^2(\boldsymbol{\theta}, \boldsymbol{\theta}_{n0}) = E_0[M(\boldsymbol{\theta})] - E_0[M(\boldsymbol{\theta}_{n0})]$. Choose $\delta_n = \rho(\boldsymbol{\theta}_0, \boldsymbol{\theta}_{n0})$. For $\delta_n < \delta < \infty$, denote $\Omega = \{\boldsymbol{\theta} : \boldsymbol{\theta} \in \Theta_n, \delta/2 < \rho(\boldsymbol{\theta}, \boldsymbol{\theta}_{n0}) \le \delta\}$. From the definition of ρ_1, we have $\sup_\Omega E_0[M(\boldsymbol{\theta}_{n0})] - E_0[M(\boldsymbol{\theta})] \le -\frac{\delta^2}{4}$.

Let Ξ_n be the set $\{M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta}_{n0}) : \boldsymbol{\theta} \in \Theta_n\}$ and $\tilde{J}(\delta, L_2(P), \Xi_n)$ be the $L_2(P)$-norm bracketing integral of the sieve Θ_n. From the proof of Theorem 4.1, we have

$$\tilde{J}(\delta, L_2(P), \Xi_n) = \int_0^\delta \sqrt{1 + \log N_2(\varepsilon, L_2(P), \Xi_n)} \, d\varepsilon$$

$$\le \int_0^\delta \sqrt{1 + \log N_2(\varepsilon, L_\infty, \Xi_n)} \, d\varepsilon$$

$$\le C N^{1/2} \delta.$$

Let

$$\phi_n(\delta) = \tilde{J}(\delta, L_2(P), \Xi_n)\left(1 + \frac{\tilde{J}(\delta, L_2(P), \Xi_n)}{\delta^2 \sqrt{n}}\right) = N^{1/2}\delta + \frac{N}{\sqrt{n}}.$$

Obviously, $\phi_n(\delta)/\delta^{1+\tau}$ is a decreasing function in δ for $0 < \tau < 1$. Then by Lemma 3.4.2 in van der Vaart and Wellner (1996), we have

$$E_0\Big[\sup_\Omega \sqrt{n}(M_n - M)(\boldsymbol{\theta} - \boldsymbol{\theta}_{n0})\Big] \preceq \phi_n(\delta).$$

For $\lambda > 1/[2(p \wedge 4)]$, from (A.2), it follows that

$$\sqrt{n}[\tilde{M}_n(\boldsymbol{\theta}) - M_n(\boldsymbol{\theta})] = O\big(n^{1/2-\lambda(p\wedge4)}\big) = o(1).$$

Then we have that

$$E_0\Big[\sup_\Omega \sqrt{n}(\tilde{M}_n - M)(\boldsymbol{\theta} - \boldsymbol{\theta}_{n0})\Big] \preceq \phi_n(\delta).$$

Then the conditions of Theorem 3.4.1 in van der Vaart and Wellner (1996) are satisfied for the δ_n, ρ_1 and $\phi_n(\delta)$ above. Therefore we have $-r_n^2 \rho_1(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_{n0}) = O_p(1)$, where r_n satisfies $r_n^2 \phi_n(\frac{1}{r_n}) \leq \sqrt{n}$. It follows that $r_n = N^{-1/2} n^{1/2} = n^{(1-v)/2}$. Thus $\rho_1(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_{n0}) = O_p(n^{-(1-v)/2})$.

Now, we define a distance ρ_2 as

$$\rho_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| + \|\eta_1 - \eta_2\|_\infty.$$

Let ς be a positive constant. Similarly to the proof of Theorem 3.2 in Huang (1999), it is easy to follow that for any $\boldsymbol{\theta}$ with $\rho_2(\boldsymbol{\theta}, \boldsymbol{\theta}_{n0}) \leq \varsigma$, there exist constants $0 < c_1, c_2 < \infty$ such that

$$-c_1 d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_{n0}) + O_p(n^{-2v\varrho}) \leq -\rho_1^2(\boldsymbol{\theta}, \boldsymbol{\theta}_{n0}) \leq -c_2 d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_{n0}) + O_p(n^{-2v\varrho}).$$

Therefore, for a constant $c_2 > 0$,

$$c_2 d^2(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_{n0}) \leq O_p(n^{-2v\varrho} + n^{-(1-v)}).$$

Because $d(\boldsymbol{\theta}_{n0}, \boldsymbol{\theta}_0) \leq \rho_2(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) = O_p(n^{-v\varrho})$, we have $d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = O_p(n^{-v\varrho} + n^{-(1-v)/2})$. □

PROOF OF THEOREM 4.3.   We prove this theorem using Theorem 6.1 in Wellner and Zhang (2007). It suffices to validate conditions A1–A6 of Theorem 6.1 in Wellner and Zhang (2007). From the proof of Theorems 4.1 and 4.2 above, it is easy to see that condition A1 regarding consistency and rate of convergence and condition A2 for Theorem 6.1 in Wellner and Zhang (2007) hold.

For condition A3, we need to calculate the pseudo-information matrix. For any fixed $\eta \in \mathcal{A}$, let $\mathcal{A}_0 = \{\eta_\omega(\cdot) : \omega$ in a neighborhood of $0 \in \mathcal{R}\}$ be a smooth curve in $\mathcal{A}$ running through η_0 at $\omega = 0$, that is, $\eta_{\omega=0}(t) = \eta_0(t)$. Denote $\frac{\partial}{\partial \omega} \eta_\omega(t)|_{\omega=0} = a(t)$ and the space generated by such $a(t)$ as Υ. The score functions of $\boldsymbol{\beta}$ and η are

$$\dot{M}_1 = \frac{\partial M}{\partial \boldsymbol{\beta}} = -2 \sum_{j=1}^{K} (Y_j - \mathbf{X}_j) \frac{\partial \mathbf{X}_j}{\partial \boldsymbol{\beta}_0},$$

$$\dot{M}_2[a] = \frac{\partial M}{\partial \eta_0} = -2 \sum_{j=1}^{K} (Y_j - \mathbf{X}_j) \frac{\partial \mathbf{X}_j}{\partial \eta_0} a(t).$$

We also set

$$\dot{M}_{11} = \frac{\partial^2 M}{\partial \boldsymbol{\beta}_0 \, \partial \boldsymbol{\beta}_0^T} = 2 \sum_{j=1}^{K} \left[\frac{\partial \mathbf{X}_j}{\partial \boldsymbol{\beta}_0} \frac{\partial \mathbf{X}_j}{\partial \boldsymbol{\beta}_0^T} - (Y_j - \mathbf{X}_j) \frac{\partial^2 \mathbf{X}_j}{\partial \boldsymbol{\beta}_0 \, \partial \boldsymbol{\beta}_0^T} \right],$$

$$\dot{M}_{12}[a] = \dot{M}_{21}^T[a] = \frac{\partial^2 M}{\partial \boldsymbol{\beta}_0 \, \partial \boldsymbol{\eta}_0} = 2 \sum_{j=1}^{K} \left[\frac{\partial \mathbf{X}_j}{\partial \boldsymbol{\beta}_0} \frac{\partial \mathbf{X}_j}{\partial \boldsymbol{\eta}_0} - (Y_j - \mathbf{X}_j) \frac{\partial^2 \mathbf{X}_j}{\partial \boldsymbol{\beta}_0 \, \partial \boldsymbol{\eta}_0} \right] a(t)$$

and

$$\dot{M}_{22}[a_1, a_2] = \frac{\partial^2 M}{\partial \eta_0^2} = 2 \sum_{j=1}^{K} \left[\left(\frac{\partial \mathbf{X}_j}{\partial \eta_0} \right)^2 - (Y_j - \mathbf{X}_j) \frac{\partial^2 \mathbf{X}_j}{\partial \eta_0^2} \right] a_1(t) a_2(t),$$

where $a_1(t), a_2(t) \in \Upsilon$. Following the idea from the proofs of the asymptotic results for semiparametric M-estimator in Ma and Kosorok (2005) and Wellner and Zhang (2007), we assume that the special perturbation direction $\mathbf{a}^*(t) = (a_1^*(t), \ldots, a_d^*(t))^T$ with $a_i^*(t) \in \Upsilon$ for $1 \leq i \leq d$, satisfies $E_0\{\dot{M}_{12}[a] - \dot{M}_{22}[\mathbf{a}^*, a]\} = 0$ for any $a \in \Upsilon$. Some calculations yield

$$E_0\{\dot{M}_{12}[a] - \dot{M}_{22}[\mathbf{a}^*, a]\}$$

$$= 2 \sum_{j=1}^{K} E_0\left[\left\{ \frac{\partial \mathbf{X}_j}{\partial \boldsymbol{\beta}_0} \frac{\partial \mathbf{X}_j}{\partial \eta_0} - (Y_j - \mathbf{X}_j) \frac{\partial^2 \mathbf{X}_j}{\partial \boldsymbol{\beta}_* \partial \eta_0} \right\} a(t) \right.$$

$$\left. - \left\{ \left(\frac{\partial \mathbf{X}_j}{\partial \eta_0} \right)^2 - (Y_j - \mathbf{X}_j) \frac{\partial^2 \mathbf{X}_j}{\partial \eta_0^2} \right\} a(t) \mathbf{a}^*(t) \right]$$

$$= 2 \sum_{j=1}^{K} E_t E_0\left(\left[\left\{ \frac{\partial \mathbf{X}_j}{\partial \boldsymbol{\beta}_0} \frac{\partial \mathbf{X}_j}{\partial \eta_0} - (Y_j - \mathbf{X}_j) \frac{\partial^2 \mathbf{X}_j}{\partial \boldsymbol{\beta}_* \partial \eta_0} \right\} a(t) \right. \right.$$

$$\left. \left. - \left\{ \left(\frac{\partial \mathbf{X}_j}{\partial \eta_0} \right)^2 - (Y_j - \mathbf{X}_j) \frac{\partial^2 \mathbf{X}_j}{\partial \eta_0^2} \right\} a(t) \mathbf{a}^*(t) \right] \middle| t \right).$$

It follows that

$$(A.3) \quad \mathbf{a}^*(t) = \frac{\sum_{j=1}^{K} E_*[\{\partial \mathbf{X}_j/\partial \boldsymbol{\beta}_0 \, \partial \mathbf{X}_j/\partial \eta_0 - (Y_j - \mathbf{X}_j)\partial^2 \mathbf{X}_j/\partial \boldsymbol{\beta}_0 \, \partial \eta_0\}|t]}{\sum_{j=1}^{K} E_0[\{(\partial \mathbf{X}_j/\partial \eta_0)^2 - (Y_j - \mathbf{X}_j)\partial^2 \mathbf{X}_j/\partial \eta_0^2\}|t]}.$$

Since $E_0[\mathbf{Y}(t)|t] = \mathbf{X}(t)$, $\mathbf{a}^*(t)$ in (A.3) can be simplified as

$$(A.4) \quad \mathbf{a}^*(t) = \frac{\sum_{j=1}^{K} \partial \mathbf{X}_j/\partial \boldsymbol{\beta}_0 \, \partial \mathbf{X}_j/\partial \eta_0}{\sum_{j=1}^{K} (\partial \mathbf{X}_j/\partial \eta_0)^2}.$$

For $K \geq 2$, both

$$(A.5) \quad \mathbf{S}_1 = E_0(\dot{M}_{11} - \dot{M}_{12}[\mathbf{a}^*]) = 2 \sum_{j=1}^{K} E_t\left[\frac{\partial \mathbf{X}_j}{\partial \boldsymbol{\beta}_0} \frac{\partial \mathbf{X}_j}{\partial \boldsymbol{\beta}_0^T} - \frac{\partial \mathbf{X}_j}{\partial \boldsymbol{\beta}_0} \frac{\partial \mathbf{X}_j}{\partial \eta_0} \mathbf{a}^*(t) \right]$$

and

$$(A.6) \quad \mathbf{S}_2 = E_0(\dot{M}_1 - \dot{M}_2[\mathbf{a}^*])^{\otimes 2} = 4 \sum_{j=1}^{K} \sigma_j^2 E_t\left\{ \frac{\partial \mathbf{X}_j}{\partial \boldsymbol{\beta}_0} - \frac{\partial \mathbf{X}_j}{\partial \eta_0} \mathbf{a}^*(t) \right\}^{\otimes 2}$$

are nonsingular. Let $\mathbf{V}_2 = \mathbf{S}_1^{-1} \mathbf{S}_2 (\mathbf{S}_1^{-1})^T$. Thus condition A3 of finite variance for Theorem 6.1 in Wellner and Zhang (2007) is satisfied.

Conditions A4 and A5 for Theorem 6.1 in Wellner and Zhang (2007) can be verified by similar arguments as condition (i) and C3 in the proof of Theorem 4 in Xue, Lam and Li (2004), respectively. Condition A6 of smoothness of the model can be easily verified using a straightforward Taylor expansion where n^{-c_1} is just the rate of convergence in Theorem 4.2 and faster than $n^{-1/4}$, and $c_2 = 2$, which completes the proof.   □

PROOF OF PROPOSITION 1. Let $\mathcal{G}$ be the set of a real valued functions g on $[a, b]$ which are absolutely continuous and satisfy $\int_a^b g^2(t)\,dt < \infty$ and $E_t g(t) = 0$. Similarly to the proof of Theorem 4.3, for any fixed $\eta \in \mathcal{A}$, let $\mathcal{A}_0 = \{\eta_\omega(\cdot) : \omega$ in a neighborhood of $0 \in \mathcal{R}\}$ be a smooth curve in $\mathcal{A}$ running through η_0 at $\omega = 0$, that is, $\eta_{\omega=0}(t) = \eta_0(t)$. Denote $\frac{\partial}{\partial \omega}\eta_\omega(t)|_{\omega=0} = a(t)$ and restrict $a(t) \in \mathcal{G}$. Denote the space generated by such $a(t)$ as Υ. The score functions of β and η are

$$\dot{M}_1 = \frac{\partial M}{\partial \beta} = -2(Y - X)\frac{\partial X}{\partial \xi},$$

$$\dot{M}_2[a] = \frac{\partial M}{\partial \eta} = -2(Y - X)\frac{\partial X}{\partial \xi}a(t)$$

with $\xi = \beta_0 + \eta_0(t)$. Let $\dot{P}$ be the linear span of $\dot{M}_2[a]$. Since $E_0\{\dot{M}_1\dot{M}_2[a]\} = 0$ for any $a(t) \in \Upsilon$, it follows that $\dot{M}_1$ is orthogonal to $\dot{P}$. Thus the efficient score function of β is just $\dot{M}_1$. Then the pseudo-information is $E_0[\dot{M}_1^2]$. The rest of the proof is similar to that of Theorem 4.3, where the efficient score function and the pseudo-information are updated as discussed before, and the least favorable direction can be selected by any $a \in \Upsilon$.   □

REFERENCES

ADAMS, B. M. (2005). Non-parametric parameter estimation and clinical data fitting with a model of HIV infection. Ph.D. thesis, North Carolina State Univ. MR2623319

ANDERSON, R. M. and MAY, R. M. (1991). *Infectious Diseases of Humans: Dynamics and Control.* Oxford Univ. Press, Oxford.

AUDOLY, S., BELLU, G., D'ANGIO, L., SACCOMANI, M. P. and COBELLI, C. (2001). Global identifiability of nonlinear models of biological systems. *IEEE Trans. Biomed. Eng.* **48** 55–65.

BABENKO, V. F., KOFANOV, V. A. and PICHUGOV, S. A. (1996). Multivariate inequalities of Kolmogorov type and their applications. In *Multivariate Approximation and Splines* (G. Nuraberger, J. W. Schmidt and G. Walz, eds.) 1–12. Birkhauser, Basel. MR1484990

BARD, Y. (1974). *Nonlinear Parameter Estimation.* Academic Press, New York. MR0326870

BELLMAN, R. and ÅSTRÖM, K. J. (1970). On structural identifiability. *Math. Biosci.* **7** 329–339.

BENSON, M. (1979). Parameter fitting in dynamic models. *Ecol. Mod.* **6** 97–115.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press, Baltimore. MR1245941

BROOKMEYER, R. and GAIL, M. H. (1994). *AIDS Epidemiology: A Quantitative Approach. Monographs in Epidemiology and Biostatistics* **23**. Oxford Univ. Press, New York.

BRUNEL, N. (2008). Parameter estimation of ODE's via nonparametric estimators. *Electron. J. Statist.* **2** 1242–1267. MR2471285

BURNHAM, K. P. and ANDERSON, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociol. Methods Res.* **33** 261. MR2086350

CHAPPEL, M. J. and GODFREY, K. R. (1992). Structural identifiability of the parameters of a nonlinear batch reactor model. *Math. Biosci.* **108** 245–251. MR1154720

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. MR2420240

CHEN, T., HE, H. L. and CHURCH, G. M. (1999). Modeling gene expression with differential equations. *Pac. Symp. Biocomput.* 29–40.

COBELLI, C., LEPSCHY, A. and JACUR, R. (1979). Identifiability of compartmental systems and related structural proerties. *Math. Biosci.* **44** 1–18.

DALEY, D. J. and GANI, J. (1999). *Epidemic Modeling*. Cambridge Univ. Press, Cambridge.

DE BOOR, C. (1978). *A Practical Guide to Splines*. Springer, New York. MR0507062

DELGADO, M. A. (1992). Semiparametric generalized least squares in the multivariate nonlinear regression model. *Econometric Theory* **8** 203–222. MR1179510

DONNET, S. and SAMSON, A. (2007). Estimation of parameters in incomplete data models defined by dynamical systems. *J. Statist. Plann. Inference* **137** 2815–2831. MR2323793

ENGLEZOS, P. and KALOGERAKIS, N. (2001). *Applied Parameter Estimation for Chemical Engineers*. Dekker, New York.

GRENANDER, U. (1981). *Abstract Inference*. Wiley, New York. MR0599175

HAIRER, E., NØRSETT, S. P. and WANNER, G. (1993). *Solving Ordinary Differential Equations I: Nonstiff Problems*, 2nd ed. Springer, Berlin. MR1227985

HE, X., FUNG, W. K. and ZHU, Z. Y. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* **89** 579–590. MR1929164

HECKMAN, N. E. and RAMSAY, J. O. (2000). Penalized regression with model-based penalties. *Canad. J. Statist.* **28** 241–258. MR1792049

HO, D. D., NEUMANN, A. U., PERELSON, A. S. ET AL. (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **373** 123–126.

HUANG, J. (1996). Efficient estimation for the proportinal hazards model with interval censoring. *Ann. Statist.* **24** 540–568. MR1394975

HUANG, J. (1999). Efficient estimation of the partly linear additive Cox model. *Ann. Statist.* **27** 1536–1563. MR1742499

HUANG, J. and ROSSINI, A. J. (1997). Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *J. Amer. Statist. Assoc.* **92** 960–967. MR1482126

HUANG, J. Z. (2003). Local asymptotics for polynomial spline regression. *Ann. Statist.* **31** 1600–1635. MR2012827

HUANG, J. Z., ZHANG, L. and ZHOU, L. (2007). Efficient estimation in marginal partially linear models for longitudinal/clustered data using plines. *Scand. J. Statist.* **34** 451–477. MR2368793

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. MR2227489

HUANG, Y., ROSENKRANZ, S. L. and WU, H. (2003). Modeling HIV dynamics and antiviral response with consideration of time-varying drug exposures, adherence and phenotypic sensitivity. *Math. Biosci.* **184** 165–186.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* 221–233. Univ. California Press, Berkeley. MR0216620

JANSSON, B. and REVESZ, L. (1975). Analysis of the growth of tumor cell populations. *Math. Biosci.* **19** 131–154.

JEFFREY, A. M. and XIA, X. (2005). Identifiability of HIV/AIDS model. In *Deterministic and Stochastic Models of AIDS Epidemics and HIV Infections with Intervention* (W. Y. Tan and H. Wu, eds.) 255–286. World Scientific, Singapore.

JENNERICH, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.* **40** 633–643. MR0238419

JOLLIFFE, I. T. (1972). Discarding variables in a principal component analyssis. I: Artificial data. *J. Roy. Statist. Soc. Ser. C Appl. Statist.* **21** 160–172. MR0311034

JOSHI, M., SEIDEL-MORGENSTERN, A. and KREMLING, A. (2006). Exploiting the boostrap method for quantifying parameter confidence intervals in dynamic systems. *Metabolic Engineering* **8** 447–455.

KOLCHIN, E. (1973). *Differential Algebra and Algebraic Groups*. Academic Press, New York. MR0568864

KUTTA, W. (1901). Beitrag zur näherungsweisen itegration totaler differentialgleichungen. *Zeitschr. Math. Phys.* **46** 435–453.

LI, L., BROWN, M. B., LEE, K. H. and GUPTA, S. (2002). Estimation and inference for a spline-enhanced pupulation pharmacokinetic model. *Biometrics* **58** 601–611. MR1933534

LI, Z., OSBORNE, M. R. and PRAVAN, T. (2005). Parameter estimation of ordinary differential equations. *IMA J. Numer. Anal.* **25** 264–285. MR2126204

LIANG, H. and WU, H. (2008). Parameter estimation for differential equation models using a framework of measurement error in regression models. *J. Amer. Statist. Assoc.* **103** 1570–1583. MR2504205

LJUNG, L. and GLAD, T. (1994). On global identifiability for arbitrary model parametrizations. *Automatica* **30** 265–276. MR1261705

MA, S. and KOSOROK, M. R. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *J. Multivariate Anal.* **96** 190–217. MR2202406

MALINVAUD, E. (1970). The consistancy of nonlinear regressions. *Ann. Math. Statist.* **41** 956–969. MR0261754

MATTHEIJ, R. and MOLENAAR, J. (2002). *Ordinary Differential Equations in Theory and Practice*. SIAM, Philadelphia. MR1946758

MIAO, H., DYKES, C., DEMETER, L. M., CAVENAUGH, J., PARK, S. Y., PERELSON, A. S. and WU, H. (2008). Modeling and estimation of kinetic parameters and replicative fitness of HIV-1 from flow-cytometry-based growth competition experiments. *Bull. Math. Biol.* **70** 1749–1771. MR2430325

MICHELSON, S. and LEITH, J. T. (1997). Tumor heterogeneity and growth control. In *Tumor Heterogeneity and Growth Control* (J. A. Adam and N. Bellomo, eds.) 295–326. Birkhäuser, Boston.

MOLES, C. G., BANGA, J. R. and KELLER, K. (2004). Solving nonconvex climate control problems: Pitfalls and algorithm performances. *Appl. Soft. Comput.* **5** 35–44.

NOWAK, M. A. and MAY, R. M. (2000). *Virus Dynamics: Mathematical Principles of Immunology and Virology*. Oxford Univ. Press, Oxford. MR2009143

OLLIVIER, F. (1990). Le problème de l'identifiabilité globale: Étude thé orique, méthodes effectives et bornes de complexité. Ph.D. thesis, École Polytechnique, Paris, France.

PAKES, A. and POLLARD, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica* **57** 1027–1057. MR1014540

PERELSON, A. S., ESSUNGER, P., CAO, Y., VESANEN, M., HURLEY, A., SAKSELA, K., MARKOWITZ, M. and HO, D. D. (1997). Decay characteristics of HIV-1-infected compartments during combination therapy. *Nature* **387** 188–191.

PERELSON, A. S. and NELSON, P. W. (1999). Mathematical analysis of HIV-1 dynamics in vivo. *SIAM Rev.* **41** 3–44. MR1669741

PERELSON, A. S., NEUMANN, A. U., MARKOWITZ, M., LEONARD, J. M. and HO, D. D. (1996). HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. *Science* **271** 1582–1586.

POHJANPALO, H. (1978). System identifiability based on the power series expansion of the solution. *Math. Biosci.* **41** 21–33. MR0507373

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York. MR0762984

POLLARD, D. (1985). New ways to prove central limit theorems. *Econometric Theory* **1** 295–314.

POLLARD, D. (1990). *Empirical Processes Theory and Applications*. IMS, Hayward, CA. MR1089429

POYTON, A. A., VARZIRI, M. S., MCAULEY, K. B., MCLELLEN, 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.

QUAISER, T. and MÖNNIGMANN, M. (2009). Systematic identifiability testing for unambiguous mechanistic modeling—application to JAK-STAT, MAP kinase, and NF-κB signaling pathway models. *BMC Sys. Bio.* **3** 50.

RAMSAY, J. O. (1996). Principal Differential Analysis: Data Reduction by Differential Operators. *J. Roy. Statist. Soc. Ser. B* **58** 495–508. MR1394362

RAMSAY, J. O., HOOKER, G., CAMPBELL, D. and CAO, J. (2007). Parameter estimation for differential equations: A generalized smoothing approach (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 741–796. MR2368570

RITT, J. F. (1950). *Differential Algebra*. Amer. Math. Soc., Providence, RI. MR0035763

RODRIGUEZ-FERNANDEZ, M., EGEA, J. A. and BANGA, J. R. (2006). Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. *BMC Bioinformatics* **7** 1–18.

RUNGE, C. (1895). Ueber die numerische Auflösung von Differentialgleichungen. *Math. Ann.* **46** 167–178. MR1510879

SCHICK, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* **14** 1139–1151. MR0856811

SCHUMAKER, L. L. (1981). *Spline Functions*. Wiley, New York. MR0606200

SHEN, X. (1997). On methods of sieves and penalization. *Ann. Statist.* **25** 2555–2591. MR1604416

SHEN, X. and WONG, W. H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22** 580–615. MR1292531

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR0673642

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705. MR0790566

STORN, R. and PRICE, K. (1997). Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *J. Global Optim.* **11** 341–359. MR1479553

SWARTZ, J. and BREMERMANN, H. (1975). Discussion of parameter estimation in biological modeling: Algorithms for estimation and evaluation of the estimates. *J. Math. Biol.* **1** 241–275.

TAN, W. Y. and WU, H. (2005). *Deterministic and Stochastic Models of AIDS Epidemics and HIV Infections With Intervention*. World Scientific, Singapore. MR2169300

THOMASETH, K., ALEXANDRA, K. W., BERNHARD, L. ET AL. (1996). Integrated mathematical model to assess β-cell activity during the oral glucose test. *Amer. J. Phisiol.* **270** E522–E531.

VAJDA, S., RABITZ, H., WALTER, E. and LECOURTIER, Y. (1989). Qualitative and quantitative identifiability analysis of nonlinear chemical kinetiv-models. *Chem. Eng. Commun.* **83** 191–219.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. MR1385671

VAN DOMSELAAR, B. and HEMKER, P. W. (1975). Nonlinear parameter estimation in initial value problmes. Report NW18/75, Math. Centrum, Amsterdam.

VARAH, J. M. (1982). A spline least squares method for numerical parameter estimation in differential equations. *SIAM J. Sci. Comput.* **3** 28–46. MR0651865

VARZIRI, M. S., POYTON, A. A., MCAULEY, K. B., MCLELLEN, P. J. and RAMSAY, J. O. (2008). Selecting optimal weighting factors in iPDA for parameter estimation in continuous-time dynamic models. *Comp. Chem. Eng.* **32** 3011–3022.

WALTER, E. (1987). *Identifiability of Parameteric Models*. Pergamon Press, Oxford.

WEI, X., GHOSH, S. K., TAYLOR, M. E. ET AL. (1995). Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* **373** 117–122.

WELLNER, J. A. and ZHANG, Y. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *Ann. Statist.* **35** 2106–2142. MR2363965

WU, C. F. (1981). Asymptotic theory of nonlinear least squares estiamtion. *Ann. Statist.* **9** 501–513. MR0615427

WU, H. (2005). Statistical methods for HIV dynamic studies in AIDS clinical trials. *Stat. Methods Med. Res.* **14** 1–22. MR2135921

WU, H., HUANG, Y., ACOSTA, E. P. ET AL. (2005). Modeling long-term HIV dynamics and antiretroviral response: Effects of drug potency, pharmacokinetics, adherence, and drug resistance. *JAIDS* **39** 272–283.

WU, H., ZHU, H., MIAO, H. and PERELSON, A. S. (2008). Identifiability and statistical estimation of dynamic parameters in HIV/AIDS dynamic models. *Bull. Math. Biol.* **70** 785–799. MR2393024

XIA, X. (2003). Estimation of HIV/AIDS parameters. *Automatica J. IFAC* **39** 1983–1988. MR2142834

XIA, X. and MOOG, C. H. (2003). Identifiability of nonlinear systems with applications to HIV/AIDS models. *IEEE Trans. Automat. Control* **48** 330–336. MR1957979

XUE, H., LAM, K. F. and LI, G. (2004). Sieve maximum likelihood estimator for semiparametric regression models with current status data. *J. Amer. Statist. Assoc.* **99** 346–356. MR2062821

DEPARTMENT OF BIOSTATISTICS
AND COMPUTATIONAL BIOLOGY
UNIVERSITY OF ROCHESTER SCHOOL
OF MEDICINE AND DENTISTRY
601 ELMWOOD AVENUE, BOX 630
ROCHESTER, NEW YORK 14642
USA
E-MAIL: Hongqi_Xue@urmc.rochester.edu
        Hongyu_Miao@urmc.rochester.edu
        hwu@bst.rochester.edu