ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2014)
76, *Part* 4, *pp.* 735–748

Quick and easy one-step parameter estimation in differential equations

Peter Hall

University of Melbourne, Australia, and University of California, Davis, USA

and Yanyuan Ma

Texas A&M University, College Station, USA

[Received November 2012. Revised June 2013]

Summary. Differential equations are customarily used to describe dynamic systems. Existing methods for estimating unknown parameters in those systems include parameter cascade, which is a spline-based technique, and pseudo-least-squares, which is a local-polynomial-based two-step method. Parameter cascade is often referred to as a 'one-step method', although it in fact involves at least two stages: one to choose the tuning parameter and another to select model parameters. We propose a class of fast, easy-to-use, genuinely one-step procedures for estimating unknown parameters in dynamic system models. This approach does not need extraneous estimation of the tuning parameter; it selects that quantity, as well as all the model parameters, in a single explicit step, and it produces root-n-consistent estimators of all the model parameters. Although it is of course not as accurate as more complex methods, its speed and ease of use make it particularly attractive for exploratory data analysis.

Keywords: Criterion function; Differential equations; Dynamic systems; Kernel estimation; Non-parametric function estimator; One-step procedure; Smoothing parameter; Tuning parameter

1. Introduction

Dynamic system models have wide application, in areas including the biomedical sciences, ecology and physics. In consequence they are often studied by statisticians, who are interested particularly in estimating model parameters and in determining any smoothing, or tuning, parameters that are needed to estimate the model. Conceptually this is a simple parametric regression problem, with the inconvenience that the regression function is given implicitly as the solution of a differential equation. However, in practice that aspect has a non-trivial consequence: if we use a classical approach, such as least squares, we do not obtain a closed form solution and instead must resort to iterative numerical methods. Unfortunately, each iteration involves solving a differential equation at a candidate parameter value, as well as calculating the derivative of the solution function with respect to the parameter. All of this makes statistical inference, and particularly computation, quite challenging.

Indeed, the high computational cost of repeatedly solving differential equations has prompted statisticians to seek alternative approaches to this problem, using non-parametric methods. The best known technique is parameter cascade (Ramsay *et al.*, 2007), where a linear combination

Address for correspondence: Yanyuan Ma, Department of Statistics, Texas A&M University, College Station, TX 77843, USA.
E-mail: ma@stat.tamu.edu

 1369–7412/14/76735

of spline basis functions is employed to approximate the implicitly defined regression function, and model parameters are estimated so that the approximation function satisfies the differential equation, and fits the data, as well as possible. However, since the data are typically noisy then these two tasks do not necessarily lead to the same parameter values. The parameter cascade method balances these two conflicting requirements by minimizing the weighted average of two respective criterion functions.

If the spline order, the number of knots, the knot positions and the balancing weights are all predecided, estimation of the model parameters can be obtained together with the spline coefficients in a single optimization procedure. Thus, parameter cascade is sometimes referred to as a one-step estimation procedure, although of course the aforementioned tuning parameters, particularly the number of knots, must be chosen in a second operation. Theoretical properties of parameter cascade were not well appreciated until the work of Qi and Zhao (2010). So far, parameter cascade is probably the most popular statistical method for parameter estimation in dynamic systems and has enjoyed success in various applications. See, for example, Cao *et al.* (2008), Hooker (2009) and Hooker *et al.* (2011).

Encouraged by the success of spline-based non-parametric methods in these problems, kernel-based non-parametric methods were explored by Liang and Wu (2008). Their basic idea was, first, to estimate the regression function and its derivative by using kernel-based non-parametric methods, e.g. techniques founded on local polynomials, and then to adjust the parameter values so that the non-parametric estimators satisfied most closely the differential equation requirement. This idea was later adapted to estimate time varying parameters in differential equations (Chen and Wu, 2008a, b). Compared with the parameter cascade approach, these kernel-based methods perform non-parametric estimation while completely ignoring the model, and then derive the model parameters by minimizing the model discrepancy expressed in terms of the estimated functions. Since the non-parametric regression and parameter estimation stages are performed separately, these methods are referred to as two-step estimation procedures.

Compared with the 'one-step' parameter cascade procedure, the two-step kernel method is somewhat less popular. This is partly because it does not take into account the model structure while performing the non-parametric estimation step. One might think that, since the data were generated from the model, it would still be appropriate to smooth them first and then to treat model aspects by using the smoothed results. However, several issues conspire to make this approach problematic. For example, choosing the tuning parameter to optimize for non-parametric smoothing does not necessarily result in root-n-consistent parameter estimation. (Here n denotes sample size.) In fact, in the context of kernel methods there is not yet a quick, easy and attractive approach to tuning parameter choice that enjoys root n consistency.

Motivated by these drawbacks of existing one-step and two-step methods, in this paper we introduce a competitive approach where model parameters and the tuning parameter are estimated together, and the model parameter estimators are root n consistent. The method has substantial advantages of speed and ease of use over alternative approaches. It is based on minimizing $S(\boldsymbol{\beta}, h)$, defined at expression (9), where $\boldsymbol{\beta}$ is the vector of model parameters and h is the tuning parameter. It is completely automatic, involving only a single step; it is currently the only available method of that type. There are no extraneous tuning parameters to select, and the minimization step is performed with respect to only $\boldsymbol{\beta}$ and h, a fixed number of parameters, regardless of sample size, even though the methodology involves non-parametric function estimation. This is in contrast with methods based on classical non-parametric maximum likelihood estimation or profile maximum likelihood estimation, where the optimization is typically with respect to a number of parameters that diverges with sample size. The method has very good performance and, although not matching the performance of other techniques when their tuning

parameters are chosen optimally, it is much simpler. Its speed and ease of use make it particularly attractive for exploratory data analysis.

2. A class of one-step kernel estimators

2.1. General differential equation models

We begin by considering a problem that is substantially more general than those that have been considered previously, and then we specialize it to a more conventional setting so that our methodology and its properties can be compared with those of other researchers. Let $\mathbf{m}$ denote an l-dimensional function of a scalar variable x, and suppose that it satisfies

$$\mathbf{g}\{\mathbf{m}(x), \boldsymbol{\beta}\} = \sum_{j=1}^{q} \gamma_j \mathbf{m}^{(j)}(x) \tag{1}$$

for all x in an interval, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^{\mathrm{T}}$ are vectors, $\mathbf{g}(\cdot, \boldsymbol{\beta})$ is a known l-dimensional functional of $\boldsymbol{\beta}$, $\mathbf{m}^{(j)}$ denotes the jth derivative of $\mathbf{m}$ and it is assumed that $\boldsymbol{\gamma}$ is constrained in such a way that $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are identifiable from condition (1). (For example, we might insist that $\gamma_1 = 1$, arguing that the size of γ_1 is accommodated by altering the scale on which $\mathbf{g}$ is measured.) It is desired to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. We anticipate that the function $\mathbf{g}(\cdot, \boldsymbol{\beta})$ will be smooth, and so the left-hand side of equation (1) represents a smooth transformation of $\mathbf{m}$, whereas the transformation on the right-hand side of equation (1) reduces the smoothness of $\mathbf{m}$.

In this problem we observe information about $\mathbf{m}$ coming from independent and identically distributed data pairs $(X_i, \mathbf{Y}_i)$, for $i = 1, \ldots, n$, generated by a regression model,

$$\mathbf{Y} = \mathbf{m}(X) + \boldsymbol{\varepsilon}, \tag{2}$$

where $E(\boldsymbol{\varepsilon}|X) = \mathbf{0}$ and $\mathrm{cov}(\boldsymbol{\varepsilon}|X) = \boldsymbol{\Sigma}(X)$. Of course, there are many ways of estimating $\mathbf{m}$ from these data; we shall discuss some of them shortly. Let $\hat{\mathbf{m}}$ denote one such non-parametric estimator. A simple approach to estimating $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is to choose them to minimize an integrated squared norm,

$$\int \left\| \mathbf{g}\{\hat{\mathbf{m}}(x), \boldsymbol{\beta}\} - \sum_{j=1}^{q} \gamma_j \hat{\mathbf{m}}^{(j)}(x) \right\|^2 w(x)\,\mathrm{d}x, \tag{3}$$

where w is an appropriate weight function. (Here and throughout the text, we use $\|\cdot\|$ to denote the l_2-norm of a vector or the Frobenius norm of a matrix.) However, this approach will not always lead to root-n-consistent estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, because in general none of the estimators $\hat{\mathbf{m}}^{(j)}$, for $j \geqslant 0$, is root n consistent.

To avoid this problem we suggest modifying expression (3) so that the quantity inside the norm in the integrand can be estimated root n consistently. For this, for each x and each bounded function $\mathbf{b}$ we let $\mathbf{T}_x(\mathbf{b})$ denote a linear transformation of $\mathbf{b}$ that is tantamount to q-fold integration of that function, and consider the version of expression (3) that is obtained when this transformation is applied to the quantity in expression (3):

$$\int \left\| \mathbf{T}_x \left\{ \mathbf{g}(\hat{\mathbf{m}}, \boldsymbol{\beta}) - \sum_{j=1}^{q} \gamma_j \hat{\mathbf{m}}^{(j)} \right\} \right\|^2 w(x)\,\mathrm{d}x. \tag{4}$$

We choose $\mathbf{T}_x$ so that

$$\mathbf{T}_x(\mathbf{m}) \equiv \mathbf{0} \tag{5}$$

if condition (1) holds, and $\mathbf{T}_x(\hat{\mathbf{m}}^{(j)})$ can be estimated root n consistently for $j = 0, \ldots, q$. Since $\mathbf{g}(\mathbf{m}, \boldsymbol{\beta})$ is a smooth functional, this will generally mean that $\mathbf{T}_x\{\mathbf{g}(\hat{\mathbf{m}}, \boldsymbol{\beta})\}$ can also be estimated root n consistently.

2.2. Differential equations of degree 1

The particular version of condition (1) that is commonly investigated in practice, and is motivated by real data sets, has $q = 1$. In this case equation (1) has the form

$$\mathbf{m}'(x) = \mathbf{g}\{\mathbf{m}(x), \boldsymbol{\beta}\}, \tag{6}$$

where $\mathbf{m}' = \mathbf{m}^{(1)}$. In the remainder of this paper we shall confine attention to this setting. Note first that if the function $\mathbf{m}$ is supported on the interval $[0, 1]$ then it is appropriate to take

$$\mathbf{T}_x(\mathbf{m}) = \int_c^x [\phi'(u, x, c)\, \mathbf{m}(u) + \phi(u, x, c)\, \mathbf{g}\{\mathbf{m}(u), \boldsymbol{\beta}\}]\, \mathrm{d}u, \tag{7}$$

for any constant $c \in (0, 1)$ and for a function ϕ satisfying

$$\phi(x, x, c) = \phi(c, x, c) = 0, \qquad \text{for all } c, x \in (0, 1). \tag{8}$$

Observe also that $\mathbf{T}_x$ is a linear transformation and that, for any bounded function ϕ satisfying condition (8), property (5) holds.

If $\mathbf{T}_x$ is given by equation (7) then expression (4) has the form

$$S(\boldsymbol{\beta}, h) \equiv \int \left\| \int_c^x [\phi'(u, x, c)\, \hat{\mathbf{m}}(u, h) + \phi(u, x, c)\, \mathbf{g}\{\hat{\mathbf{m}}(u, h), \boldsymbol{\beta}\}]\, \mathrm{d}u \right\|^2 w(x)\, \mathrm{d}x, \tag{9}$$

where $\phi(u, x, c) = (x - u)(u - c)$ and, in practice, we take $w \equiv 1$. The notation $\hat{\mathbf{m}}(u, h)$ in expression (9) reflects the fact that the estimator $\hat{\mathbf{m}}$ depends on a bandwidth h. For a given h we choose $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_h$ to minimize $S(\boldsymbol{\beta}, h)$ in equation (9). As we shall show, this approach, in company with property (8), ensures that the bandwidth that is selected together with $\hat{\boldsymbol{\beta}}$, by minimizing $S(\boldsymbol{\beta}, h)$, is one that enables $\hat{\boldsymbol{\beta}}$ to be root n consistent for $\boldsymbol{\beta}$.

The quantity $\hat{\mathbf{m}}(u, h)$ can denote any standard kernel estimator of $\mathbf{m}$ with bandwidth h. For example, we might use $\hat{\mathbf{m}}(u, h) = \Sigma_{1 \leqslant i \leqslant n}\, \omega_i(x, h) \mathbf{Y}_i$, where

$$\omega_i(x, h) = \int_{s_{i-1}}^{s_i} K_h(t - x)\, \mathrm{d}t,$$

$$\omega_i(x, h) = \frac{K_h(X_i - x)}{\sum\limits_{1 \leqslant i \leqslant n} K_h(X_i - x)}$$

in the case of Gasser–Müller or Nadaraya–Watson estimators respectively, and

$$\omega_i(x) = \frac{K_h(X_i - x)\left\{ \sum\limits_j K_h(X_j - x)(X_j - x)^2 - (X_i - x) \sum\limits_j K_h(X_j - x)(X_j - x) \right\}}{\sum\limits_k K_h(X_k - x)\left\{ \sum\limits_j K_h(X_j - x)(X_j - x)^2 - (X_k - x) \sum\limits_j K_h(X_j - x)(X_j - x) \right\}}$$

if $\hat{\mathbf{m}}$ is a local linear estimator. Here, K is a kernel function, $K_h(u) = h^{-1} K(u/h)$ and $s_i = (X_i + X_{i+1})/2$ for $i = 1, \ldots, n - 1$, $s_0 = 0$ and $s_n = 1$. For simplicity, we have used a common

bandwidth h for estimating different components in $\mathbf{m}$. This simplification does not affect first-order asymptotic properties of the estimator of $\boldsymbol{\beta}$. When using either Gasser–Müller or Nadaraya–Watson estimators we suggest that c is taken to be a small constant approaching 0, but much larger than h. This has the effect of guarding against edge effects, and taking $c = h^{1/2}$ is usually sufficient. However, using $c = 0$ causes no problems if the estimator $\hat{\mathbf{m}}$ is robust against boundary effects, e.g. if it is a local linear estimator.

2.3. Further generalizations

In addition to generalizations to more complex models, such as that at expression (1), we can treat criteria that are intrinsically more general than $S(\boldsymbol{\beta}, h)$ at expression (9). This diversity can reduce the variance of estimators $\hat{\boldsymbol{\beta}}$. For simplicity we shall confine attention again to model (6).

Assume that $0 < c < \frac{1}{2}$, let $\psi_1, \psi_2, \ldots$ be a sequence of functions supported on $[c, 1-c]$ and satisfying $\psi_j(c) = \psi_j(1-c) = 0$ for each j, and note that if model (6) holds then

$$\mathbf{0} = \int_c^{1-c} \psi_j(x)[\mathbf{m}'(x) - \mathbf{g}\{\mathbf{m}(x), \boldsymbol{\beta}\}]\,\mathrm{d}x = -\int_c^{1-c} [\psi_j'(x)\,\mathbf{m}(x) + \psi_j(x)\,\mathbf{g}\{\mathbf{m}(x), \boldsymbol{\beta}\}]\,\mathrm{d}x$$

for each j. This motivates choosing $(\hat{\boldsymbol{\beta}}, h)$ to minimize

$$S_1(\boldsymbol{\beta}, h) = \sum_j w_j \left\| \int_c^{1-c} [\psi_j'(x)\,\hat{\mathbf{m}}(x,h) + \psi_j(x)\,\mathbf{g}\{\hat{\mathbf{m}}(x,h), \boldsymbol{\beta}\}]\,\mathrm{d}x \right\|^2, \tag{10}$$

where $w_1, w_2, \ldots$ are non-negative weights; compare with expression (9). Alternatively, if there are just J weights ψ_j then we can use either of the criteria

$$\begin{aligned} S_2(\boldsymbol{\beta}, h) = \sum_{j=1}^J \sum_{k=1}^J w_{jk} &\left(\int_c^{1-c} [\psi_j'(x)\,\hat{\mathbf{m}}(x,h) + \psi_j(x)\,\mathbf{g}\{\hat{\mathbf{m}}(x,h), \boldsymbol{\beta}\}]\,\mathrm{d}x \right)^{\mathrm{T}} \\ &\times \left(\int_c^{1-c} [\psi_k'(x)\,\hat{\mathbf{m}}(x,h) + \psi_k(x)\,\mathbf{g}\{\hat{\mathbf{m}}(x,h), \boldsymbol{\beta}\}]\,\mathrm{d}x \right), \end{aligned} \tag{11}$$

$$\begin{aligned} S_3(\boldsymbol{\beta}, h) = \sum_{j=1}^J \sum_{k=1}^J \int &\left(\int_c^x [\phi_j'(u,x,c)\,\hat{\mathbf{m}}(x,h) + \phi_j(u,x,c)\,\mathbf{g}\{\hat{\mathbf{m}}(x,h), \boldsymbol{\beta}\}]\,\mathrm{d}u \right)^{\mathrm{T}} \\ &\times \left(\int_c^x [\phi_k'(u,x,c)\,\hat{\mathbf{m}}(u,h) + \phi_k(u,x,c)\,\mathbf{g}\{\hat{\mathbf{m}}(u,h), \boldsymbol{\beta}\}]\,\mathrm{d}u \right) w_{jk}(x)\,\mathrm{d}x, \end{aligned} \tag{12}$$

where in equation (11) the weights w_{jk} are chosen so that the $J \times J$ matrix $\mathbf{W} = (w_{jk})$ is non-negative definite, and in equation (12) the weight functions $w_{jk}(x)$ have the property that $\mathbf{W}(x) = (w_{jk}(x))$ is non-negative definite for all except at most a finite number of values of x.

3. Theoretical properties

Let $\mathbf{g}'_\beta(\mathbf{m}, \boldsymbol{\beta})$ denote the first partial derivatives of $\mathbf{g}(\mathbf{m}, \boldsymbol{\beta})$ with respect to the p-vector $\boldsymbol{\beta}$, an $l \times p$ matrix. Let $\mathbf{g}''_\beta(\mathbf{m}, \boldsymbol{\beta})$ denote the second partial derivatives of $\mathbf{g}(\mathbf{m}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, written into a length l block row vector, with each block size $p \times p$. Similarly, write $\mathbf{g}'(\mathbf{m}, \boldsymbol{\beta})$, an $l \times l$ matrix, and $\mathbf{g}''(\mathbf{m}, \boldsymbol{\beta})$, an $l \times l^2$ matrix, for the first and second partial derivatives of $\mathbf{g}(\mathbf{m}, \boldsymbol{\beta})$ with respect to $\mathbf{m}$. Let $\phi(u, x, c)$, and its derivatives $\phi^{(j)}(u, x, c) = (\partial/\partial u)^j \phi(u, x, c)$ for $j = 1, 2$, satisfy condition (8). With $\boldsymbol{\beta}_0$ representing the true value of $\boldsymbol{\beta}$, define

$$\mathbf{b}(u,x,c) = \phi'(u,x,c)\,\mathbf{I}_l + \phi(u,x,c)\,\mathbf{g}'\{\mathbf{m}(u), \boldsymbol{\beta}_0\}, \tag{13}$$

$$\boldsymbol{\lambda}(x) = \int_0^x \phi(u,x,0)\,\mathbf{g}'_\beta\{\mathbf{m}(u), \boldsymbol{\beta}_0\}\,\mathrm{d}u, \tag{14}$$

denoting respectively an $l \times l$ matrix and an $l \times p$ matrix, where $\mathbf{I}_l$ is the size l identity matrix. If $\mathbf{v}$ is a vector or a matrix, put $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^{\mathrm{T}}$.

We take $\hat{\mathbf{m}}(u,h)$ to be a Gasser–Müller, Nadaraya–Watson or local linear estimator; see Section 2.2 for discussion. In the theorem below, $(\hat{\boldsymbol{\beta}}_h, h)$ denotes the minimizer of $S(\boldsymbol{\beta}, h)$, in expression (9), for any given $h \in \mathcal{H}_n$, where $\mathcal{H}_n$ is a large class of bandwidths, defined in condition 7; the regularity conditions 1–8 are those discussed in Appendix A; and the $p \times p$ matrices $\mathbf{A}$ and $\mathbf{B}$ are given by

$$\mathbf{A} = \int \boldsymbol{\lambda}^{\mathrm{T}}(x)^{\otimes 2}\, w(x)\,\mathrm{d}x,$$

$$\mathbf{B} = E\left\{ \left(\int_X^1 \mathbf{b}(X,x,0)\,\boldsymbol{\lambda}(x)\,w(x)\,\mathrm{d}x \right)^{\mathrm{T}} \frac{\boldsymbol{\Sigma}(X)}{f_X^2(X)} \left(\int_X^1 \mathbf{b}(X,x,0)\,\boldsymbol{\lambda}(x)\,w(x)\,\mathrm{d}x \right) \right\}. \tag{15}$$

If we substitute $(\mathbf{m}, \boldsymbol{\beta})$, satisfying condition (6), into expression (9) then expression (9) will reach the minimum possible value 0. It is not difficult to verify the converse, i.e., if expression (9) is 0 at a certain $(\mathbf{m}, \boldsymbol{\beta})$ then that $(\mathbf{m}, \boldsymbol{\beta})$ will automatically satisfy condition (6). This ensures that minimizing expression (9) does not lead us to a problem which is different from that of solving expression (6). In fact, under condition 7, $\hat{\mathbf{m}}(x,h) \to \mathbf{m}(x)$ when $n \to \infty$; hence the above observation also ensures that the estimator that minimizes expression (9) yields a consistent estimator of $\boldsymbol{\beta}$.

The following theorem, which is derived in the on-line appendix, shows not only that the limiting distribution of $\hat{\boldsymbol{\beta}}_h$ is normal for each $h \in \mathcal{H}_n$ but also that, in a well-defined sense, asymptotic normality can be asserted uniformly in $h \in \mathcal{H}_n$. In Appendix A we shall discuss properties when $h \notin \mathcal{H}_n$.

Theorem 1. If conditions 1–8 hold then, for each $h \in \mathcal{H}_n$, $n^{1/2}(\hat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}_0)$ is asymptotically normally distributed with zero mean and covariance matrix $\boldsymbol{\Sigma} = \mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^{\mathrm{T}}$. Indeed, $n^{1/2}(\hat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}_0) = \mathbf{N}_n + o_p(1)$, uniformly in $h \in \mathcal{H}_n$, where the random p-vectors $\mathbf{N}_n$, for $n \geqslant 1$, all have the $N(\mathbf{0}, \boldsymbol{\Sigma})$ distribution and do not depend on h.

Theorem 1 is readily generalized to cases where, for given h, $\hat{\boldsymbol{\beta}}_h$ minimizes one of the criteria $S_1(\boldsymbol{\beta}, h)$, $S_2(\boldsymbol{\beta}, h)$ and $S_3(\boldsymbol{\beta}, h)$, defined at expressions (10), (11) and (12) respectively. For example, consider $l = 1$. If the criterion is S_3 then, for $l = 1$, the theorem holds with $\mathbf{A}$ and $\mathbf{B}$ redefined by

$$\mathbf{A} = \int \boldsymbol{\Lambda}(x)\,\mathbf{W}(x)\,\boldsymbol{\Lambda}^{\mathrm{T}}(x)\,\mathrm{d}x,$$

$$\mathbf{B} = E\left[\left\{ \int_X^1 \boldsymbol{\Lambda}(x)\,\mathbf{W}(x)\,\mathbf{b}(X,x,0)\ \mathrm{d}x \right\}^{\otimes 2} \frac{\sigma^2(X)}{f_X^2(X)} \right],$$

where

$$\mathbf{b}(u,x,c) = \phi'(u,x,c) + \phi(u,x,c)g'\{m(u), \boldsymbol{\beta}_0\}, \tag{16}$$

$$\boldsymbol{\Lambda}(x) = \int_c^x \mathbf{g}'_\beta\{m(u), \boldsymbol{\beta}_0\}\,\boldsymbol{\phi}^{\mathrm{T}}(u,x,c)\,\mathrm{d}u. \tag{17}$$

If the criterion is S_2, at expression (11), then theorem 1 is again obtained, this time with

$$\mathbf{A} = \boldsymbol{\Lambda}(1-c)\mathbf{W}\boldsymbol{\Lambda}^{\mathrm{T}}(1-c),$$

$$\mathbf{B} = E\left[\left\{\boldsymbol{\Lambda}(1-c)\mathbf{W}\,\mathbf{b}(X, 1-c, 0)\frac{\sigma(X)}{f_X(X)}\right\}^{\otimes 2}\right],$$

where $\mathbf{b}(u, x, c)$ and $\boldsymbol{\Lambda}(x)$ are again as at equations (16) and (17). Still in the context of S_2, the minimum variance is $\{\boldsymbol{\Lambda}(1-c)\mathbf{W}\boldsymbol{\Lambda}^{\mathrm{T}}(1-c)\}^{-1}$ and is obtained when

$$\mathbf{W}^{-1} = E\left\{\mathbf{b}(X, 1-c, 0)\frac{\sigma(X)}{f_X(X)}\right\}^{\otimes 2}.$$

An estimator with these weights has minimum variance in the class of estimators defined by minimizing S_2, although it is apparently not optimal more widely for estimating $\boldsymbol{\beta}$. There appears to be no estimator that retains the simplicity of our one-step approach and, at the same time, has minimum variance in a wide sense.

4. Numerical properties

Since $S(\boldsymbol{\beta}, h)$ is twice differentiable with respect to $\boldsymbol{\beta}$ and h, minimization of expression (9) can be performed by using the Newton–Raphson algorithm. In practice we restrict h to be at least the maximum distance between two neighbouring observations, and at most $n^{-1/3}$. This choice works well in practice and also ensures condition 7 in Appendix A. The algorithm could stop at a local minimum, but this problem is readily alleviated by starting with several different values and evaluating $S(\boldsymbol{\beta}, h)$ at a set of grid points, to gain global knowledge of the function.

4.1. Simulated examples

We performed a series of simulation studies to investigate the finite sample performance of our methodology. In our first set of simulations the true underlying model was a simple linear ordinary differential equation of the form $m'(x) = \beta m(x)$. This equation has explicit solution $m(x, \beta) = \exp(\beta x) + c$, for any constant c. Of course, during the estimation procedure we did not take advantage of the solution. We generated n observations from the model $Y = \exp(\beta X) + \varepsilon$. We took $\beta = 1$ and assumed that ε was normal $N(0, \sigma^2)$, and we experimented with sample sizes $n = 500$ and $n = 1000$, in combination with standard deviations $\sigma = 0.1, 0.2, 0.3$. Here and below, each experiment was repeated 1000 times.

The results are given in Table 1 and show that our methodology yields estimators with very small biases and variances. Although the sample size $n = 250$ seems too small to corroborate our theoretical results, the conclusions of theorem 1 are reflected well for sample sizes $n = 500$ and $n = 1000$, and in particular the averages of the empirical standard errors of the estimators are close to their theoretical values. Indeed, when the sample size is small, a bootstrap procedure can be performed in lieu of theorem 1 to assess the variability of the estimator. Specifically, one can randomly sample $(X_{b1}, \mathbf{Y}_{b1}), \ldots, (X_{bn}, \mathbf{Y}_{bn})$ from the original observations and obtain $(\hat{\boldsymbol{\beta}}_b, \hat{h}_b)$ by minimizing expression (9) constructed by using the bootstrap data. Repeating this procedure for $b = 1, \ldots, B$, and calculating the sample variance–covariance matrix from $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_B$, we obtain a bootstrap estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$.

In our second set of simulations we considered a second-order differential equation, $m''(x) = -\beta^2 m(x)$, which equivalently can be written as a set of two first-order ordinary differential equations, $m_1'(x) = \beta m_2(x)$ and $m_2'(x) = -\beta m_1(x)$. We generated data from the model $Y_1 = \cos(\beta X) + \varepsilon_1$ and $Y_2 = -\sin(\beta X) + \varepsilon_2$, which satisfies the set of equations for $\beta = 1$. The errors

Table 1. Simulation 1: average values of estimators of β, $\bar{\hat{\beta}}$, standard deviations of estimators of β, $\mathrm{sd}(\hat{\beta})$, and average values of estimators of standard deviations of β, $\widehat{\mathrm{sd}}(\hat{\beta})$, for $n = 250, 500, 1000$ and $\sigma = 0.1$, $0.2, 0.3$†

	Values for $n = 250$			*Values for* $n = 500$			*Values for* $n = 1000$		
	$\sigma = 0.1$	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.1$	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.1$	$\sigma = 0.2$	$\sigma = 0.3$
$\bar{\hat{\beta}}$	0.9999	0.9998	1.0001	0.9996	0.9994	0.9989	1.0001	1.0002	1.0004
$\mathrm{sd}(\hat{\beta})$	0.0259	0.0487	0.0717	0.0166	0.0321	0.0478	0.0108	0.0213	0.0320
$\widehat{\mathrm{sd}}(\hat{\beta})$	0.0283	0.0566	0.0849	0.0176	0.0351	0.0526	0.0113	0.0226	0.0338

†The true value of β was 1.

Table 2. Simulation 2: average values of estimators of β, $\bar{\hat{\beta}}$, standard deviations of estimators of β, $\mathrm{sd}(\hat{\beta})$, and average values of estimators of standard deviations of β, $\widehat{\mathrm{sd}}(\hat{\beta})$, for $n = 250, 500, 1000, 1500$ and $\sigma = 0.1, 0.2, 0.3$†

	Values for $n = 250$			*Values for* $n = 500$			*Values for* $n = 1000$			*Values for* $n = 1500$		
	$\sigma = 0.1$	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.1$	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.1$	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.1$	$\sigma = 0.2$	$\sigma = 0.3$
$\bar{\hat{\beta}}$	1.0002	0.9989	0.9993	0.9995	0.9987	0.9991	0.9997	0.9994	0.9995	1.0001	1.0002	0.9999
$\mathrm{sd}(\hat{\beta})$	0.0405	0.0781	0.1172	0.0264	0.0527	0.0786	0.0168	0.0336	0.0501	0.0134	0.0269	0.0405
$\widehat{\mathrm{sd}}(\hat{\beta})$	0.0460	0.0921	0.1381	0.0285	0.0570	0.0855	0.0183	0.0366	0.0549	0.0143	0.0286	0.0429

†The true value of β was 1.

ε_1 and ε_2 were taken to be normal $N(0, \sigma^2)$. Results for various sample sizes and error variances are summarized in Table 2. Again the procedure proposed produces estimators with small biases and variances. However, the average estimated standard deviation does not match its theoretical counterpart closely until the sample size becomes relatively large, reflecting the fact that, as we shall see in subsequent simulation studies, for models of greater complexity the asymptotic results in theorem 1 require larger sample sizes before they are fully apparent in practice.

In the third set of simulations we experimented with estimating a bivariate parameter, $\boldsymbol{\beta} = (\beta_1, \beta_2)^{\mathrm{T}}$, while keeping the number of equations, and their order, equal to 1. Specifically, the differential equation was taken to be $m'(x) = \beta_1 m(x) + \beta_2$, and we generated data from the model $Y = \beta_1^{-1} \exp(\beta_1 X + \beta_2) + \varepsilon$, where the mean function was a solution of the differential equation, and ε was normal $N(0, \sigma^2)$. As can be seen from the results that are reported in Table 3, the methodology again performs well, and in this relatively complex model the asymptotic theory in theorem 1 requires larger sample sizes before it is evident empirically.

In our fourth set of simulations we extended the previous three studies to cases where both the number of differential equations and the number of parameters exceeded 1. Specifically, we took the equations to be $m_1'(x) = \beta_1 \beta_2 m_2(x)$ and $m_2'(x) = -\beta_2^{-1} \beta_1 m_1(x)$, and we generated data from the models $Y_1 = \cos(\beta_1 x) + \varepsilon_1$ and $Y_2 = \beta_2^{-1} \sin(\beta_1 x) + \varepsilon_2$. In Table 4 we report results for various sample sizes and reach the same conclusions as before.

We also experimented with data drawn from the well-known FitzHugh–Nagumo equations (FitzHugh, 1961; Nagumo *et al.*, 1962). This set of equations involves three parameters and has the form

Table 3. Simulation 3: average values of estimators of β, $\tilde{\hat{\beta}}$, standard deviations of estimators of β, $\mathrm{sd}(\hat{\beta})$, and average values of estimators of standard deviations of β, $\widehat{\mathrm{sd}}(\hat{\beta})$, for $n = 250, 500, 1000, 1500, 2000$ and $\sigma = 0.1, 0.2, 0.3$†

	Values for $\sigma = 0.1$		*Values for* $\sigma = 0.2$		*Values for* $\sigma = 0.3$	
	β_1	β_2	β_1	β_2	β_1	β_2
$n = 250$						
$\tilde{\hat{\beta}}$	0.9958	1.0031	0.9926	1.0069	0.9918	1.0087
$\mathrm{sd}(\hat{\beta})$	0.2675	0.1731	0.3781	0.2539	0.4520	0.3114
$\widehat{\mathrm{sd}}(\hat{\beta})$	0.3187	0.2078	0.6383	0.4170	0.9611	0.6285
$n = 500$						
$\tilde{\hat{\beta}}$	1.0055	0.9968	1.0081	0.9959	1.0061	0.9977
$\mathrm{sd}(\hat{\beta})$	0.1610	0.1042	0.2556	0.1684	0.3169	0.2135
$\widehat{\mathrm{sd}}(\hat{\beta})$	0.1794	0.1174	0.3590	0.2352	0.5395	0.3536
$n = 1000$						
$\tilde{\hat{\beta}}$	1.0051	0.9971	1.0120	0.9938	1.0147	0.9928
$\mathrm{sd}(\hat{\beta})$	0.0954	0.0622	0.1834	0.1206	0.2484	0.1654
$\widehat{\mathrm{sd}}(\hat{\beta})$	0.1073	0.0705	0.2147	0.1411	0.3224	0.2118
$n = 1500$						
$\tilde{\hat{\beta}}$	1.0010	0.9995	1.0033	0.9985	1.0062	0.9973
$\mathrm{sd}(\hat{\beta})$	0.0756	0.0491	0.1481	0.0967	0.2103	0.1375
$\widehat{\mathrm{sd}}(\hat{\beta})$	0.0811	0.0534	0.1622	0.1068	0.2435	0.1604
$n = 2000$						
$\tilde{\hat{\beta}}$	1.0026	0.9983	1.0054	0.9965	1.0063	0.9961
$\mathrm{sd}(\hat{\beta})$	0.0618	0.0408	0.1221	0.0806	0.1785	0.1180
$\widehat{\mathrm{sd}}(\hat{\beta})$	0.0670	0.0442	0.1340	0.0884	0.2011	0.1326

†The true value of β was $(1, 1)^{\mathrm{T}}$.

$$m_1'(x) = \beta_3\{m_1(x) - \tfrac{1}{3}m_1(x)^3 + m_2(x)\},$$
$$m_2'(x) = -\beta_3^{-1}\{m_1(x) - \beta_1 + \beta_2 m_2(x)\}.$$

We generated data from the noisy equations $Y_1 = m_1(x) + \varepsilon_1$ and $Y_2 = m_2(x) + \varepsilon_2$, and in Table 5 we report results for various sample sizes and error variances. Again the method performs well, and the greater complexity of this setting means that the asymptotic properties require large sample sizes to be fully visible.

4.2. Data example

We applied our method to the famous lynx and hare data (Odum and Barrett (2004), page 191). The data set contains the recorded numbers of Canadian lynx and snowshoe hares between 1845 and 1935, based on data collected by the Hudson Bay company. Because the two species have a typical predator–prey relationship, we used the classical Lotka–Volterra model (Borrelli and Coleman, 1996) to describe changes in the two population sizes. Specifically, if $m_1(x)$ and $m_2(x)$ represent the numbers of snowshoe hares and Canadian lynxes respectively, at time x, they satisfy

$$m_1'(x) = \exp(\beta_1)\, m_1(x) - \exp(\beta_2)\, m_1(x)\, m_2(x),$$
$$m_2'(x) = -\exp(\beta_3)\, m_2(x) + \exp(\beta_4)\, m_1(x)\, m_2(x).$$

Table 4. Simulation 4: average values of estimators of β, $\tilde{\hat{\beta}}$, standard deviations of estimators of β, $\mathrm{sd}(\hat{\beta})$, and average values of estimators of standard deviations of β, $\widehat{\mathrm{sd}}(\hat{\beta})$, for $n = 250, 500, 1000, 1500, 2000$ and $\sigma = 0.1, 0.2, 0.3$†

	Values for $\sigma = 0.1$		*Values for* $\sigma = 0.2$		*Values for* $\sigma = 0.3$	
	β_1	β_2	β_1	β_2	β_1	β_2
n = 250						
$\tilde{\hat{\beta}}$	1.0004	1.0019	0.9981	1.0030	0.9943	1.0055
$\mathrm{sd}(\hat{\beta})$	0.0496	0.0501	0.0982	0.1011	0.1454	0.1496
$\widehat{\mathrm{sd}}(\hat{\beta})$	0.0578	0.0576	0.1168	0.1169	0.1778	0.1794
n = 500						
$\tilde{\hat{\beta}}$	1.0013	1.0009	1.0021	1.0020	1.0018	1.0031
$\mathrm{sd}(\hat{\beta})$	0.0324	0.0320	0.0638	0.0639	0.0954	0.0965
$\widehat{\mathrm{sd}}(\hat{\beta})$	0.0360	0.0357	0.0721	0.0717	0.1088	0.1084
n = 1000						
$\tilde{\hat{\beta}}$	1.0001	1.0002	0.9995	0.9998	0.9994	1.0003
$\mathrm{sd}(\hat{\beta})$	0.0218	0.0216	0.0437	0.0432	0.0656	0.0652
$\widehat{\mathrm{sd}}(\hat{\beta})$	0.0232	0.0230	0.0465	0.0462	0.0699	0.0695
n = 1500						
$\tilde{\hat{\beta}}$	0.9996	1.0005	0.9986	1.0006	0.9977	1.0011
$\mathrm{sd}(\hat{\beta})$	0.0164	0.0176	0.0328	0.0349	0.0491	0.0526
$\widehat{\mathrm{sd}}(\hat{\beta})$	0.0181	0.0180	0.0363	0.0361	0.0545	0.0543
n = 2000						
$\tilde{\hat{\beta}}$	1.0002	1.0004	1.0001	1.0006	1.0002	1.0009
$\mathrm{sd}(\hat{\beta})$	0.0149	0.0149	0.0297	0.0300	0.0447	0.0451
$\widehat{\mathrm{sd}}(\hat{\beta})$	0.0153	0.0152	0.0306	0.0304	0.0460	0.0457

†The true value of β was $(1, 1)^{\mathrm{T}}$.

We parameterized the coefficients on the right-hand side by using an exponential form, to ensure that these coefficients are positive. Using the method proposed, the estimated parameter values are $\hat{\beta} = (4.5273, 1.4418, 4.0914, 0.4172)^{\mathrm{T}}$, with associated standard errors $(0.4452, 0.3365, 3.5567, 3.0131)^{\mathrm{T}}$. On the basis of these estimated values we plotted the solutions of the differential equations and graphed the estimated curves against the observations; Fig. 1. Here, either an initial condition or a boundary condition needs to be given to determine uniquely the solution of the differential equation system and the curves. Because our method does not involve these aspects, we simply used a least squares criterion to pick an initial value that minimizes the total distance between the differential equation predictions and the observations. From Fig. 1 we can see that the method captures population fluctuations reasonably well, even though the fit is not precise. This imprecision is due partly to the nature of the data, which are known to be inaccurate because of the difficulty of determining empirically the size of a population, and partly to limitations of the rather crude Lotka–Volterra model of the relationship between predator and prey.

4.3. Comparison with competing methods

We applied our estimator in the setting of Ramsay *et al.* (2007) and found that our estimation standard errors for the three parameters were 0.0390, 0.1573 and 0.5522. Relative to their results

Table 5. Simulation 5: average values of estimators of β, $\tilde{\hat{\beta}}$, standard deviations of estimators of β, sd($\hat{\beta}$), and average values of estimators of standard deviations of β, $\widehat{\text{sd}}(\hat{\beta})$, for $n = 250, 500, 1000, 1500, 2000$ and $\sigma = 0.1, 0.2, 0.3$†

	Values for $\sigma = 0.1$			*Values for* $\sigma = 0.2$			*Values for* $\sigma = 0.3$		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
$n = 250$									
$\tilde{\hat{\beta}}$	4.9363	0.9851	0.4949	4.9222	0.9772	0.4903	4.7183	0.9392	0.4868
sd($\hat{\beta}$)	2.2596	0.5110	0.0218	3.3960	0.7688	0.0283	3.8731	0.8806	0.0323
$\widehat{\text{sd}}(\hat{\beta})$	3.9317	0.8933	0.0116	7.8449	1.7930	0.0233	11.7125	2.6881	0.0351
$n = 500$									
$\tilde{\hat{\beta}}$	5.0006	1.0034	0.4984	4.9666	0.9946	0.4953	4.9124	0.9816	0.4928
sd($\hat{\beta}$)	0.9943	0.2258	0.0076	1.6072	0.3665	0.0125	1.9398	0.4437	0.0163
$\widehat{\text{sd}}(\hat{\beta})$	1.1844	0.2687	0.0059	2.2795	0.5215	0.0119	3.4008	0.7796	0.0179
$n = 1000$									
$\tilde{\hat{\beta}}$	5.0050	1.0020	0.4991	4.9993	1.0006	0.4970	4.9886	0.9971	0.4956
sd($\hat{\beta}$)	0.4435	0.1016	0.0038	0.7755	0.1785	0.0071	1.0266	0.2371	0.0097
$\widehat{\text{sd}}(\hat{\beta})$	0.4322	0.0991	0.0035	0.8698	0.1998	0.0070	1.3179	0.3033	0.0105
$n = 1500$									
$\tilde{\hat{\beta}}$	5.0005	1.0006	0.4991	4.9975	0.9992	0.4977	4.9974	0.9981	0.4963
sd($\hat{\beta}$)	0.2709	0.0624	0.0028	0.5033	0.1164	0.0052	0.7101	0.1646	0.0071
$\widehat{\text{sd}}(\hat{\beta})$	0.2726	0.0627	0.0026	0.5549	0.1279	0.0053	0.8477	0.1955	0.0080
$n = 2000$									
$\tilde{\hat{\beta}}$	4.9868	0.9962	0.4993	4.9485	0.9887	0.4981	4.9284	0.9837	0.4971
sd($\hat{\beta}$)	0.2104	0.0487	0.0022	0.3916	0.0909	0.0042	0.5699	0.1326	0.0058
$\widehat{\text{sd}}(\hat{\beta})$	0.2084	0.0480	0.0022	0.4254	0.0982	0.0044	0.6471	0.1496	0.0066

†The true value of β was $(5, 1, 0.5)^{\text{T}}$.

of 0.0149, 0.0643 and 0.0264, ours are obviously worse. However, this is hardly surprising, since the competing method is substantially more difficult to implement than ours and uses a smoothing parameter that requires significantly more skill to compute. We also compared our approach with the method of Liang and Wu (2008), in the same setting as theirs, and obtained the estimation standard deviations 0.0541, 0.1446 and 0.4168. In comparison with the corresponding results of 0.08, 0.12 and 0.17 in Liang and Wu (2008), our results are better for some parameters but worse for others, and the method of Liang and Wu (2008) also requires a delicate choice of tuning parameter.

5. Discussion

We have proposed a quick and easy-to-use one-step kernel method for parameter estimation in differential-equation-based dynamic models. The method avoids bandwidth choice via cross-validation or plug-in techniques. Instead, the bandwidth is treated as one of the parameters, and they are chosen together by using a single optimization procedure. It is the only genuinely one-step procedure available in the literature. Our proposal shows that it is possible to avoid solving the differential equations and tuning the smoothing parameters simultaneously. Intuitively, this results from the fact that the dynamic system provides much more information than a pure non-parametric model.

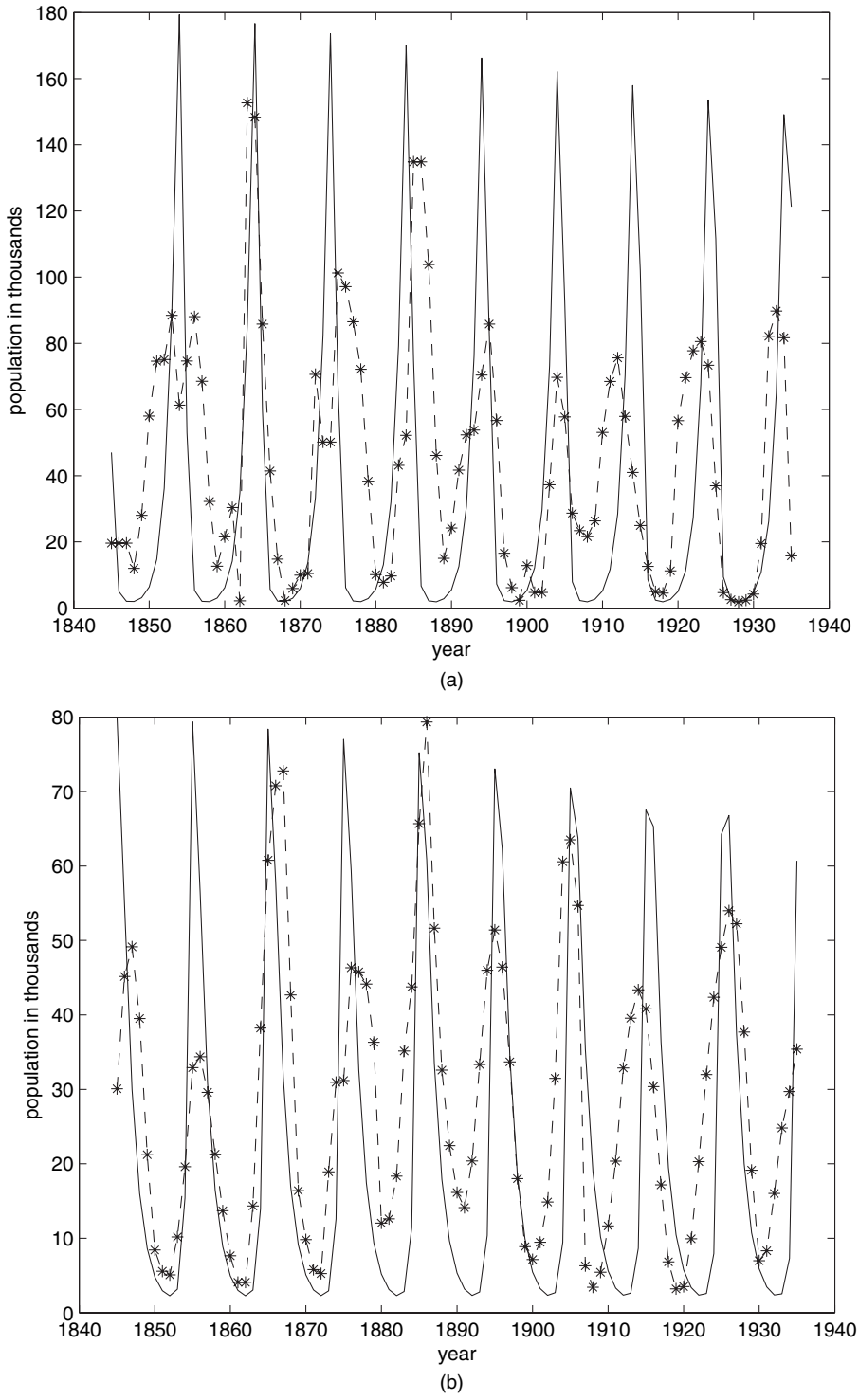

Fig. 1. Estimated (———) and observed (– – –) fluctuation of (a) the hare and (b) the lynx populations

The straightforwardness of our one-step kernel method comes also with a cost, however. Although our parameter estimators are root n consistent, they generally are not as accurate as the optimal spline-based method of Ramsay *et al.* (2007).

Appendix A: Regularity conditions and discussion

The following regularity conditions were imposed in theorem 1.

Condition 1. The true regression function $\mathbf{m}$ and its first two derivatives are bounded on the interval $[0, 1]$.

Condition 2. The function $\mathbf{g}(\mathbf{m}, \boldsymbol{\beta})$ has two bounded derivatives with respect to $\mathbf{m}$ and two bounded, Hölder continuous derivatives with respect to $\boldsymbol{\beta}$, and $\mathbf{g}$ and $\mathbf{m}$ satisfy condition (6).

Condition 3. The data pairs $(X_i, \mathbf{Y}_i)$ are independent and identically distributed as $(X, \mathbf{Y})$, generated by the model (2), and the regression error $\boldsymbol{\varepsilon}(x) = \mathbf{Y} - E(\mathbf{Y}|X = x)$ satisfies $E\{\boldsymbol{\varepsilon}(x)\} = 0$ for each x, and $\sup_{x \in [0,\,1]} E\|\boldsymbol{\varepsilon}(x)\|^C < \infty$ for each $C > 0$.

Condition 4. The support of the distribution of X equals $[0, 1]$, the density f_X of X is bounded away from 0 and ∞ on that interval, and the first two derivatives of f_X, when that function is viewed as restricted to $[0, 1]$, are bounded there.

Condition 5. The kernel function K is a bounded, symmetric, compactly supported and Hölder continuous density function, and the weight function w is positive and uniformly bounded.

Condition 6. The function ϕ of three variables satisfies $\phi(x, x, c) = \phi(c, x, c) = 0$ for each c and x, and $\phi(u, x, c)$, $\partial\phi(u, x, c)/\partial u$ and $\partial^2\phi(u, x, c)/\partial u^2$ are bounded functions of (u, x, c).

Condition 7. To define $(\hat{\boldsymbol{\beta}}_h, h)$ we minimize $S(\boldsymbol{\beta}, h)$, at expression (9), over $\boldsymbol{\beta} \in \mathcal{B}$ for given $h \in \mathcal{H}_n$, where $\mathcal{H}_n$ is the set of h such that $n^{\eta_1 - 1/2} \leqslant h \leqslant n^{-1/4 - \eta_1}$, $\eta_1 \in (0, \frac{1}{8})$ is otherwise arbitrary, and $\mathcal{B}$ is the class of p-vectors $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leqslant \eta_2$, with $\eta_2 > 0$ chosen sufficiently small but not depending on n, and $\boldsymbol{\beta}_0$ denoting the unique value of $\boldsymbol{\beta}$ for which condition (6) holds.

Condition 8. In expression (9) we take the unqualified outer integral on the right-hand side to be over $c \leqslant x \leqslant 1 - c$, where $c = c(n)$ has the following properties. If local polynomial methods are used, $c \to 0$ (in this case $c \equiv 0$ is permitted) and, if Gasser–Müller or Nadaraya–Watson techniques are employed, $c \to 0$ and nc^4 is bounded away from 0.

In reference to condition 8, if we take $c \in (0, \frac{1}{2})$ to be fixed then the variance of $\hat{\boldsymbol{\beta}}_h$ is relatively complex. In particular the assumption $c \to 0$ in condition 8 simplifies the expression for the variance of $\hat{\boldsymbol{\beta}}_h$.

The class $\mathcal{H}_n$ of bandwidths, introduced in condition 7, has the property that $\hat{\boldsymbol{\beta}}_h$ is asymptotically normally distributed, with zero mean, for each $h \in \mathcal{H}_n$. However, if that were all that was desired then the definition $\mathcal{H}_n = \{h : n^{\eta_1 - 1/2} \leqslant h \leqslant n^{-1/4 - \eta_1}\}$ could be generalized to $\mathcal{H}_n = [h_1, h_2]$, where $h_j = h_j(n)$, for $j = 1$ and $j = 2$, should be chosen so that $h_1 < h_2$, $n^{1/2} h_1 \to \infty$ and $n^{1/4} h_2 \to 0$ as $n \to \infty$. The extra restriction in the definition of $\mathcal{H}_n$ in condition 7 is imposed to obtain the 'uniform asymptotic normality' result that is asserted in theorem 1.

Outside the range $[h_1, h_2]$, if h converges to 0 at the same rate as $n^{-1/4}$ or $n^{-1/2}$, or either more slowly than $n^{-1/4}$ or more quickly than $n^{-1/2}$, then $\hat{\boldsymbol{\beta}}_h$ is not necessarily normally distributed with zero mean and asymptotic variance of order n^{-1}. Using a lengthy argument it can be shown that for relatively small or large h the value of $\inf_\beta S(\boldsymbol{\beta}, h)$ is relatively large, and that such values of h are not selected by the algorithm that was suggested in Section 2.2.

References

Borrelli, R. L. and Coleman, C. S. (1996) *Differential Equations: a Modeling Perspective*. New York: Wiley.
Cao, J., Fussmann, G. F. and Ramsay, J. O. (2008) Estimating a predator-prey dynamical model with the parameter cascades method. *Biometrics*, **64**, 959–967.
Chen, J. and Wu, H. (2008a) Efficient local estimation for time-varying coefficients in deterministic dynamic models with applications to HIV-1 dynamics. *J. Am. Statist. Ass.*, **103**, 369–384.
Chen, J. and Wu, H. (2008b) Estimation of time-varying parameters in deterministic dynamic models. *Statist. Sin.*, **18**, 987–1006.
FitzHugh, R. (1961) Impulses and physiological states in theoretical models of nerve membrane. *Biophys. J.*, **1**, 445–466.
Hooker, G. (2009) Forcing function diagnostics for nonlinear dynamics. *Biometrics*, **65**, 928–936.
Hooker, G., Ellner, S. P., de Vargas Roditi, L. and Earn, D. J. D. (2011) Parameterizing state space models for infectious disease dynamics by generalized profiling measles in Ontario. *Interface*, **8**, 961–974.
Liang, H. and Wu, H. (2008) Parameter estimation for differential equation models using a framework of measurement error in regression models. *J. Am. Statist. Ass.*, **103**, 1570–1583.
Nagumo, J. S., Arimoto, S. and Yoshizawa, S. (1962) An active pulse transmission line simulating a nerve axon. *Proc. Inst. Radio Engrs*, **50**, 2061–2070.
Odum, E. P. and Barrett, G. W. (2004) *Fundamentals of Ecology*. Belmont: Brooks Cole.
Qi, X. and Zhao, H. (2010) Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations. *Ann. Statist.*, **38**, 435–481.
Ramsay, J. O., Hooker, G., Campbell, D. and Cao, J. (2007) Parameter estimation for differential equations: a generalized smoothing approach (with discussion). *J. R. Statist. Soc.* B, **69**, 741–796.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Appendix B: Proof of Theorem 1'.