SCIENCE CHINA
Mathematics

• ARTICLES •

September 2011 Vol. 54 No. 9: 1815–1828
doi: 10.1007/s11425-011-4224-7

# A nonparametric estimation of the infection curve

LIN HuaZhen[1,*], YIP Paul S. F.[2] & HUGGINS Richard M.[3]

[1]*School of Statistics, Southwestern University of Finance and Economics, Chengdu* 611130*, China;*
[2]*Social Work and Social Administration, University of Hong Kong, Hong Kong, China;*
[3]*Department of Mathematics and Statistics, The University of Melbourne, Melbourne, Australia*
*Email: linhz@swufe.edu.cn, sfpyip@hku.hk, huggins@ms.unimelb.edu.au*

Received October 3, 2010; accepted January 25, 2011; published online August 1, 2011

**Abstract** Predicting the future course of an epidemic depends on being able to estimate the current numbers of infected individuals. However, while back-projection techniques allow reliable estimation of the numbers of infected individuals in the more distant past, they are less reliable in the recent past. We propose two new nonparametric methods to estimate the unobserved numbers of infected individuals in the recent past in an epidemic. The proposed methods are noniterative, easily computed and asymptotically normal with simple variance formulas. Simulations show that the proposed methods are much more robust and accurate than the existing back projection method, especially for the recent past, which is our primary interest. We apply the proposed methods to the 2003 Severe Acute Respiratory Syndorme (SARS) epidemic in Hong Kong.

**Keywords** epidemic, back-projection, nonparametric method, infection curve

**MSC(2000):** 62G05, 62P10

**Citation:** Lin H Z, Yip P S F, Huggins R M. A nonparametric estimation of the infection curve. Sci China Math, 2011, 54(9): 1815–1828, doi: 10.1007/s11425-011-4224-7

## 1 Introduction

An important public health issue that arises over the course of an epidemic determines how many individuals are infected at a given time. This quantity is of concern to policy-makers and managers of health care systems as well as epidemiologists. For example, the 2003 SARS epidemic in Hong Kong, which killed 298 persons and infected about 1800, presented one of the most serious global health threats since the HIV/AIDS epidemic. One of the reasons leading to the epidemic was the considerable uncertainty about the current epidemic state during the course of this epidemic.

A key feature of infectious disease data is that infected individuals are only observed when they are diagnosed so the exact infection times are unknown and hence full information on the current epidemic state is not available. Limited methods are available for analysing epidemic data. One is mathematically convenient curves, for example, exponential or polynomial. Predictions based on these models might not be reliable because these parametric curves are not data-based (see [13, 25]). On the other hand, transmission models (see [2, 19, 20, 6]) try to explore the latent nature of the spread of the disease. Unfortunately, data are usually inadequate to estimate key model parameters, and these transmission models have not been extensively studied and used.

Back projection (see [8, 9]) has become one of the most popular methods of reconstructing the past pattern of infections and it is also widely used to predict future numbers of cases with the disease

*Corresponding author

 

(see [5, 7, 27]). It makes elementary assumptions about the way the data are generated and the only additional information required is knowledge of the distribution of the time from infection to clinical diagnosis (see [7]). However, the back-projection method is not without problems. One problem with the back projection method is ill-posed inverse (see [26]). To avoid the problem, it is necessary to impose some kind of structure on the infection curve. Some implementations of back projection used a smooth parametric model (see [14]) or a parametric step function (see [8, 9, 27]) for the incidence curve. However, as the epidemic is only partially observed, it is not easy to correctly specify the incidence curve. Other investigators have allowed a nonparametric form for the incidence curve. Brookmeyer [10], Bacchetti et al. [5] and Liao and Brookmeyer [22] used a smoothing spline method based on a penalized likelihood. Becker, Watson, and Carlin [7] applied a smoothed EM algorithm developed by Silverman et al. [29]. The smoother used in [7] is easy to compute, but does not perform well at boundaries. For example, Figures 1(b) and (c) show simulation results of the average of the estimated incidence curve over 500 replications of an epidemic, and the smoothed EM, denoted EMS, of Becker and Watson [7], is biased, and is noticeably unsatisfactory for times close to the end of the observation period. Another common problem with the parametric and nonparametric back-projection methods is that the theoretical properties of the methods are largely unknown.

To overcome the disadvantages of the existing approaches, in this paper two alternate approaches are proposed. Firstly, via the independent incubation process among the infected individuals, we propose the one-step estimate of the infective number at each day $j$ based on the number of observed cases. The one-step estimator has closed-form expression and is obtained without any effort on the program and the computation. Simulations in Section 4 indicate that the one-step estimator performs much better than the unsmoothed and the smoothed back projection methods in terms of the mean square error. Considering that the one-step estimator may have slightly large variance for the number of infections in the recent past due to the very little accurate information available, and because that the recent past is our interest, we make an effort to improve the one-step estimate in the recent past by "borrowing" or making use of information of the neighbouring time points. That is, smoothing the one-step estimates over time using one of the existing smoothing techniques. Since the one-step estimator has closed-form, any existing smoothing techniques can be applied without any extra effort to the programming. Compared with the existing methods, the one-step and the smoothing one-step methods are noniterative, easy to compute and are shown to be asymptotically normal both with simple variance formulas.

The paper is organized as follows. In Section 2, we give the one-step estimator and the smoothed one-step estimator of the incidence. The asymptotic normality of the proposed estimators are established in Section 3. Section 4 conducts simulation studies to compare the behavior of the one-step estimator, the smoothed one-step estimator and the back-projection methods. Finally, in Section 5, the approach is applied to the Hong Kong SARS data. The simulations and the application indicate that our proposed methods are robust and efficient. Some discussions are given in Section 6.

## 2 Notation and estimation

### 2.1 Notation

Typical data from epidemics are interval censored so that cases are reported in batches on a daily or weekly basis. To reflect this, we divide the time axis $[0, \tau]$ over which data have been collected into intervals of equal length, that may be thought of as "months", "weeks" or "days". These are indexed by the nonnegative integers $j, j = 1, \ldots, n$, where $n$ is the most recent interval beyond which no detected cases are available. We call $\tau$ the current time. Let the observed data $d_j$ be the number of cases detected on day $j$ for $j = 1, 2, \ldots, n$, $z_j$ be the unobserved number of individuals who were infected on day $j$, and $z_{ju}$ be the number of individuals infected on day $j$ with incubation period $u, u = 0, \ldots, k$, where $k$ denotes the longest incubation time. Then $z_j = \sum_{u=0}^{k} z_{ju}$. Let $p_u$ be the probability that an infected individual is detected on day $u$, $u = 0, \ldots, k$ after infection. We suppose that $p_u, u = 0, \ldots, k$ are known and assume that the incubation processes for different infected individuals are independent. This assumption is also

made by the existing methods including back-projection. For infectious disease with a short incubation time such as SARS, reliable information on the incubation times is available and accurate estimates of the $p_u$ are readily obtained [15, 1]. For more detail on the estimate of the incubation distribution, see [3–5].

### 2.2 The one-step estimation

Our objective is to estimate the expected number of infective individuals $Ez_j$ at each day $j$ using the observed number of cases $d_j, j = 1, \ldots, n$. A natural estimator of $Ez_j$ is the conditional mean of $z_j = \sum_{u=0}^{k} z_{ju}$ given the observed data. For that, we consider the conditional distribution of $z_j$ given the observation data $\{d_j, j = 1, 2, \ldots, n\}$. A natural way to obtain the conditional distribution is computing the probability distributions of the data (e.g., the likelihood) and the probability distributions of $z_j$. However, in our case, both the distributions are difficult to obtain due to the feature of partial observation of the epidemic data.

Some observations form the estimator proposed in the paper.

(1) In all of the observed data, only $d_{j+u}$ which is related to $z_{ju}$ and thus the conditional distribution of $z_{ju}$ given the observed data $\{d_j, j = 1, 2, \ldots, n\}$ is equal to the conditional distribution of $z_{ju}$ given $d_{j+u}$.

(2) The incubation process among infected individuals is independent.

(3) $d_r = \sum_{u=0}^{k} z_{r-u,u}$.

The observations (2) and (3) imply that given $d_r$, the conditional distributions of $\{z_{r-u,u}, u = 0, \ldots, k\}$ are independent multinomial distributions based upon $d_r$ trials with probabilities $\{p_u, u = 0, \ldots, k\}$. Hence,

$$E\{z_{r-u,u}|d_r\} = p_u d_r. \tag{2.1}$$

Let $\mathcal{D} = \{d_j, j = 1, 2, \ldots, n\}$, the observation (1) implies that $E\{z_{ju}|\mathcal{D}\} = E\{z_{ju}|d_{j+u}\}$. This, coupling with (2.1), for $j \leqslant n - k$, we get

$$E\{z_{ju}|\mathcal{D}\} = E\{z_{ju}|d_{j+u}\} = p_u d_{j+u}. \tag{2.2}$$

However, (2.2) cannot be used when $j > n - k$, because $d_{j+u}$ is unobservable. Since

$$Ez_j = \frac{\sum_{u=0}^{n-j} Ez_{ju}}{\sum_{u=0}^{n-j} p_u},$$

a simple estimator, that we call one-step estimator (termed OS), of $Ez_j$ is,

$$\hat{z}_j = \frac{\sum_{u=0}^{n-j} p_u d_{j+u}}{\sum_{u=0}^{n-j} p_u}, \tag{2.3}$$

where $p_u = 0$ when $u > k$. It is easy to show that $E\{\hat{z}_j\} = Ez_j$, hence $\hat{z}_j$ is an unbiased estimator of $Ez_j$.

It is interesting to make a comparison of the one-step estimator and the back-projection estimator. To motivate the back-projection estimator note that

$$E\{d_j|z_1, \ldots, z_n\} = \sum_{i=1}^{j} z_i p_{j-i}, \tag{2.4}$$

so that the back-projection estimator is based on the conditional distribution of $d_j$ given $z_1, \ldots, z_n$. Contrarily, the one-step estimator is based on the conditional distribution of $z_j$ given the observed data $d_1, \ldots, d_n$. Let $\lambda_j = E[z_j]$, then the back-projection estimator of $\lambda_j$ is obtained iteratively from

$$\lambda_j^{(m)} = \frac{\lambda_j^{(m-1)}}{\sum_{u=0}^{n-j} p_u} \sum_{u=0}^{n-j} \frac{d_{j+u} p_u}{\sum_{i=1}^{j+u} \lambda_i^{(m-1)} p_{j+u-i}}, \tag{2.5}$$

which is the combination of the E step and the M step of the EM algorithm (see [7]). The EM algorithm generally converges very slowly and can be time-consuming, for example, averaging around 10 minutes for each repetition of simulation 1 in Section 4. Also, because the complication of the computation of the back-projection estimator, the used nonparametric techniques to smooth the back-projection are restrictive, for example, the smoother used in [7], which is easy to compute, but does not perform well at boundaries. However, very little programming effort and time are needed to compute the one-step estimator. Furthermore, it is difficult to establish the asymptotical properties of the back-projection estimators and these are largely unknown. In contrast, as the one-step estimator has a closed form, under the regular conditions given in Appendix A.1, it can be proved that the one-step estimator is asymptotically normal (see Theorem 1 in Section 3).

### 2.3 The smoothed one-step estimation

Theorem 1 in Section 3 and the results in Tables 1, 2 and Figure 1 of Section 4 show that the variance of the one-step estimator $\hat{z}_j$ increases for $j$ close to $n$. Since the infection number near the current time is our primarily interest, it is worthy to make an effort to reduce the variance of the one-step estimator and hence reduce the mean square error for the recent past. A natural method to reduce the variance is applying nonparametric techniques to smooth the one-step estimator over time. As a smoothing method, we choose the local linear model (see [16]). This method has many good statistical properties. For example, it adapts automatically to the boundary of design points, which is especially important for our problem because our interest is on the boundary. We also note that it may be possible to improve the EMS estimator by using a more reasonable smoother than that which has been used in the EMS algorithm. However, a more reasonable smoother generally means greater computational complexity. Specifically, for the EMS estimator, we need to apply the smooth technique to each iterative EM step of the back projection method, which is a huge computational burden. However, since the one-step estimator has a closed-form, any existing nonparametric smoothing techniques can be used without any extra programming and computational efforts.

Write $t_j = j\delta_n$, where $\delta_n = \tau/n$, so that $t_j$ is the absolute time at the end of the $j$th interval. Now, $z_j$ is the number of new infectives arising in the $j$th interval so that $z_1, \ldots, z_n$ arises from a discretization of an underlying continuous time infection process. Let $\lambda(t)$ be the intensity of this continuous time process over the interval $[0, \tau]$ and $\eta$ be the size of the underlying population so that we can take $\Lambda_j = \eta \int_{t_{j-1}}^{t_j} \lambda(s)ds = \eta \int_{t_j-\delta_n}^{t_j} \lambda(s)ds = \Lambda(t_j)$, where $\Lambda(t) = \eta \int_{t-\delta_n}^{t} \lambda(s)ds$ is a differentiable function. Since $\Lambda(t)$ is differentiable, for any fixed $t_0 \in [0, \tau]$ and each $t$ close to $t_0$, a Taylor expansion gives,

$$\Lambda(t) \approx \Lambda(t_0) + \Lambda'(t_0)(t - t_0) \equiv \beta_1 + \beta_2(t - t_0), \tag{2.6}$$

where $\beta_1$ and $\beta_2$ depend on $t_0$. This, coupling with $E\hat{z}_j = Ez_j = \Lambda_j$, motivates a local linear model fitted using a locally weighted linear regression. We estimate $\beta = (\beta_1, \beta_2)$ by minimizing

$$\ell(\beta) = \sum_{j=1}^{n} \{\hat{z}_j - \beta_1 - \beta_2(t_j - t_0)\}^2 K_h(t_j - t_0), \tag{2.7}$$

where $K_h(\cdot) = K(\cdot/h)/h$, in which $K(\cdot)$ denotes a kernel function and $h$ is a bandwidth. The kernel is introduced so that the local model (2.6) is only applied to the data close to $t_0$. Denote the minimizer of (2.7) by $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)'$. From (2.7), for fixed $t_0$ we obtain the closed form estimator,

$$\hat{\beta} = \left( \sum_{j=1}^{n} x_{t_j} x'_{t_j} K_h(t_j - t_0) \right)^{-1} \sum_{j=1}^{n} x_{t_j} K_h(t_j - t_0) \hat{z}_j, \tag{2.8}$$

where $x_t = (1, t - t_0)'$. Then for $t_0 \in [0, \tau]$, $\Lambda(t_0)$ is estimated by $\hat{\Lambda}(t_0) = \hat{\beta}_1$ and the $\Lambda_j$ are estimated by $\hat{\Lambda}_j = \hat{\Lambda}(t_j)$, $j = 1, \ldots, n$. We refer to these as the smoothed one-step estimates (SOS).

## 3 Large sample properties

In this section, we investigate the asymptotic properties of the one-step and smoothed one-step estimators. Firstly, we consider the one-step estimator. By [21, p. 98], and Conditions (iv) and (v), it is straightforward to get

**Theorem 1.** *For any fixed $j$, $\sqrt{\nu_{n-j}}\,(\hat{z}_j - Ez_j) \to N(0, \sigma_j^2)$, where $\nu_{n-j} = \sum_{u=0}^{n-j} p_u$ and $\sigma_j^2 = \lim_{n\to\infty} \frac{\sum_{u=0}^{n-j} p_u^2 \mathrm{Var}(d_{j+u})}{\nu_{n-j}}$.*

Theorem 1 implies that the convergent rate of $\hat{z}_j$ depends on $j$, the convergent rate decreases with $j$ increasing. As a result, the variance of the estimator for the number of infection will increase when $j$ is close to $n$. The conclusion is confirmed by the simulation studies in Section 4.

Denote $u_2 = \int_{-\infty}^{+\infty} x^2 K(x)dx$ and $v_0 = \int_{-\infty}^{+\infty} K^2(x)dx$, we have the following theorem for the smoothed one-step estimator.

**Theorem 2.** *As $n \to \infty$, $h \to 0$, under the regularity conditions given in the Appendix.*

1. *If $\tau - t_0 = O(h)$ and $nh \to +\infty$, we have*

$$(nh)^{1/2}\{\hat{\Lambda}(t_0) - \Lambda(t_0) - h^2\Lambda''(t_0)u_2/2\} \to N(0, b(t_0, t_0)), \tag{3.1}$$

*where $b(t,t)$ is a continuous function and defined by* (A.2) *in the Appendix. Hence if $\tau - t_j = O(h)$, then*

$$(nh)^{1/2}(\hat{\Lambda}_j - \Lambda_j - h^2\Lambda''(t_j)u_2/2) \to N(0, b(t_j, t_j)). \tag{3.2}$$

2. *If $\tau - t_0 = O(1)$, we have*

$$n^{1/2}\{\hat{\Lambda}(t_0) - \Lambda(t_0) - h^2\Lambda''(t_0)u_2/2\} \to N(0, \tilde{b}(t_0, t_0)), \tag{3.3}$$

*where $\tilde{b}(t,t)$ is a continuous function and also defined by* (A.2) *in the Appendix. Hence if $\tau - t_j = O(1)$, then*

$$n^{1/2}(\hat{\Lambda}_j - \Lambda_j - h^2\Lambda''(t_j)u_2/2) \to N(0, \tilde{b}(t_j, t_j)). \tag{3.4}$$

Therefore, the asymptotic bias of $\hat{\Lambda}_j$ is $\mathrm{bias}(\hat{\Lambda}_j) = h^2\Lambda''(t_j)u_2/2$, and the asymptotic variance of $\hat{\Lambda}_j$ is

$$\mathrm{var}(\hat{\Lambda}_j) = \begin{cases} b(t_j, t_j)/(nh), & \tau - t_j = O(h), \\ \tilde{b}(t_j, t_j)/n, & \tau - t_j = O(1). \end{cases}$$

If $\tau - t_j = O(1)$, so that $t_j$ is not close to the current time $\tau$, increasing $h$ cannot decrease the variance, but does increase the bias and the optimal bandwidth to estimate $\Lambda_j$ is $h = 0$. By (2.7), $\hat{\Lambda}_j = \hat{z}_j$ when $h = 0$. These results suggest that when $t_j$ is far away from the current time $\tau$, the smoothing step cannot improve the one-step estimator. This is confirmed by the simulations in Section 4.

For $\tau - t_j = O(h)$ and $t_j$ is close to the current time, we need to select the bandwidth $h$. Theoretically, an optimal local bandwidth is obtained by minimizing the integrated mean squared error given by $\sum_{j=r}^{n}[\mathrm{Bias}^2\{\hat{\Lambda}_j\} + \mathrm{Var}\{\hat{\Lambda}_j\}]$, where $r$ is the time point from which we smooth the one-step estimator. The estimation of the bias can be obtained by the empirical bias approach proposed by Ruppert [28], which has been proved to work well in related studies (see [23, 24]). The proof of the theorem shows that the variance-covariance matrix of $(\hat{\beta}_1(t), h\hat{\beta}_2(t))$ can be estimated by $\mathrm{V} = (nh)^{-1}\hat{A}^{-1}\widehat{\Sigma}\hat{A}^{-1}$, where $\hat{A} = n^{-1}H^{-1}\sum_{j=1}^{n} x_{t_j}x_{t_j}' K_h(t_j - t)H^{-1}$,

$$\widehat{\Sigma} = \frac{h}{n}H^{-1}\sum_{s=1}^{n}\left(\sum_{j=1}^{s} x_{t_j}K_h(t_j - t)\frac{p_{s-j}}{\sum_{u=0}^{n-j} p_u}\right)\left(\sum_{j=1}^{s} x_{t_j}K_h(t_j - t)\frac{p_{s-j}}{\sum_{u=0}^{n-j} p_u}\right)' \hat{\mathrm{V}}\mathrm{ar}(d_s)H^{-1},$$

and

$$H = \mathrm{diag}(1, h).$$

The variance of $\hat{\Lambda}_j$ is estimated by the $(1,1)$-entry of the matrix V with $t$ replaced by $t_j = \tau j/n$.

When $h = 0$, the variance of $\hat{\Lambda}_j$ can be estimated by

$$\sum_{s=j}^{n} \left( \frac{p_{s-j}}{\sum_{u=0}^{n-j} p_u} \right)^2 \hat{\mathrm{Var}}(d_s),$$

which is exactly the empirical version of the variance of the one-step estimator $\hat{z}_j$ (see Theorem 1).

In the example concerning the SARS epidemic in Hong Kong, we will give a method to determine the point from which we smooth the one-step estimators. In practice, we are interested in the number of the infected individuals in the recent past, that is, the target time always is close to the current time. Hence, generally, we need to smooth the one-step estimator.

## 4 Simulations studies

### 4.1 Comparison of the one-step estimator and the back projection estimators

Since the properties of the back-projection are unknown, we cannot compare the one-step estimator with the back projection methods via theoretical results and instead, we conduct a numerical study. Two models are considered. The first concerns an infection processe without intervention, and $z_j$ depends on the size of the infective population just before $j$. Following traditional infection models, we simulate an epidemic process with hazard function $h(t) = 0.05y(t-)$, where $y(t-)$ is the total number of infectives in the population just before time $t$. We use a Weibull distribution with shape 1.5 and scale 8 to model the distribution of the incubation time (see Figure 1(a)). The epidemic commences with 15 infective individuals. We conducted 500 simulations and the average of the total number of infected individuals was 959.08 (sd= 220.19). We obtained the estimates of the incidence curve using the one-step estimator (termed OS), the back projection estimator (termed BP) and the back projection method with a smoothed EM (termed EMS, see [7]). Figure 1(b) shows the average of the estimated incidence curve over the 500 replications for the OS estimator, the BP estimator and the EMS estimator. Table 1 gives bias, SD and RMSE (root mean squared error) of the estimators of the number of infectives at $j = 15, 30, 45, 60, 75, 90$

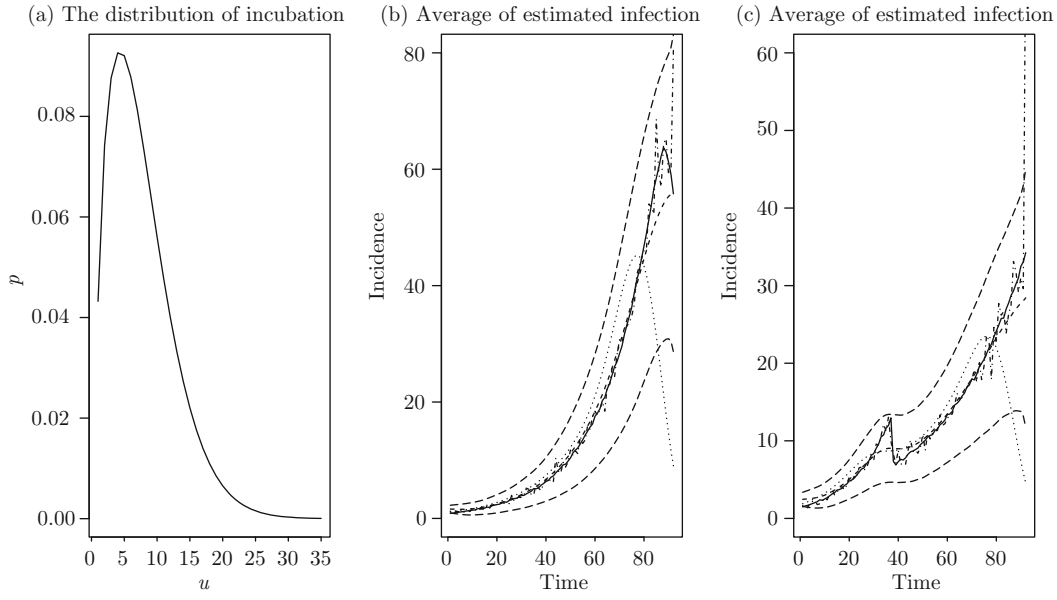


**Figure 1** (a) The distribution of the incubation time; (b) and (c) The mean estimated infection curve based on 500 simulations using the proposed one-step method (dashed), the EMS (dotted) and the BP (dotted-linear), the empirical pointwise 95% confidential limits of the proposed one-step method (dashed), as well as the true infection curve (solid). (b) Simulation 1; (c) Simulation 2.

**Table 1** Comparisons of the performance of the OS, the BP and the EMS estimators for the number of infected when applied to simulation 1

| | | days ($j$) | | | | | |
|---|---|---|---|---|---|---|---|
| | | 15 | 30 | 45 | 60 | 75 | 90 |
| OS | Bias | 0.123 | 0.130 | 0.387 | 0.862 | 1.065 | −9.054 |
| | SD | 0.580 | 1.054 | 2.124 | 4.370 | 8.630 | 11.876 |
| | RMSE | 0.592 | 1.062 | 2.159 | 4.455 | 8.695 | 14.934 |
| EMS | Bias | 0.217 | 0.481 | 1.136 | 2.694 | 8.979 | −37.607 |
| | SD | 0.612 | 1.149 | 2.326 | 4.988 | 10.219 | 5.420 |
| | RMSE | 0.650 | 1.245 | 2.589 | 5.669 | 13.603 | 37.995 |
| BP | Bias | 0.134 | −0.379 | −0.817 | 0.916 | 0.485 | −5.294 |
| | SD | 4.501 | 6.999 | 13.767 | 27.963 | 47.480 | 88.524 |
| | RMSE | 4.503 | 7.009 | 13.791 | 27.978 | 47.483 | 88.682 |

for each method. The bias is defined by the difference of the estimator from the mean number of cases $E[z_j]$ as generated by the simulation model. From Figure 1(b) and Table 1, we see that the BP method has the largest variance and is considerably inefficient. The EMS is biased, particularly for times close to $n = 95$. In contrast, the proposed one-step estimator yielded a estimator that has much less bias than the EMS and has much less variance than the BP, as a result, has consistently smaller RMSE than the EMS and the BP estimators, and the improvement of the one-step estimate over the EMS and the BP increases as $j$ becomes closer to $n$. Hence, the one-step estimator is much better than the EMS estimator and the BP estimator. The considerable inefficience of the BP estimator is caused by the well-known ill-posedness of the inverse problem, which can be appreciated by observing the following equation obtained by (2.4):

$$d_j = \sum_{i=1}^{j} E(z_i)p_{j-i} + \varepsilon_j.$$

Since $p_{j-i}$ smoothly change over $i$, as a result, relatively large perturbations of $E(z_i), i = 1, \ldots, n$ can give rise to very slight perturbations in the data $d_j, j = 1, \ldots, n$ and conversely. It follows from this that least squares, minimum $\chi^2$, or maximum likelihood solutions will be very sensitive to slight changes in the data.

Our second simulation considers an infection process with a control factor. The infection process was time dependent with hazard function: $h(t) = \beta(t)y(t)$, where $\beta(t) = 0.06$ for $t \leqslant 40$ and $\beta(t) = 0.03$ for $t > 40$, so the hazard drops at $t = 40$. The epidemic commenced with 20 infective individuals. The results displayed in Figure 1(c) and Table 2 yield similar conclusions to our first set of simulations.

**Table 2** Comparisons of the performance of the OS, the BP and the EMS estimators for the number of infected when applied to simulated data 2

| | | days ($j$) | | | | | |
|---|---|---|---|---|---|---|---|
| | | 15 | 30 | 45 | 60 | 75 | 90 |
| OS | Bias | 0.380 | 0.088 | 1.475 | 0.255 | 0.428 | −2.750 |
| | SD | 0.890 | 1.782 | 2.219 | 2.943 | 4.547 | 6.512 |
| | RMSE | 0.967 | 1.784 | 2.664 | 2.954 | 4.567 | 7.069 |
| EMS | Bias | 0.778 | 0.870 | 1.505 | 1.170 | 3.674 | −16.932 |
| | SD | 0.977 | 1.950 | 2.221 | 3.127 | 5.302 | 3.130 |
| | RMSE | 1.249 | 2.135 | 2.683 | 3.338 | 6.451 | 17.219 |
| BP | Bias | 0.222 | −0.585 | −0.331 | 1.258 | −1.023 | 3.787 |
| | SD | 7.429 | 13.076 | 15.411 | 23.209 | 31.344 | 50.292 |
| | RMSE | 7.432 | 13.089 | 15.415 | 23.243 | 31.361 | 50.435 |

### 4.2 Comparison of the one-step, smoothed one-step and smoothed back-projection estimators

We conducted simulations to compare the performance of the smoothed one-step estimator (SOS) with the one-step estimator (OS) and the smoothed back-projection estimators (EMS). Table 3 gives the bias, SD and RMSE of the resulting estimators for the number of infectives at $j = 66, 70, 74, 78, 82, 86, 90, 94$ using the SOS estimator with $h = 10$, the OS and the EMS estimators using the first simulation. From Table 3 we see that the SOS estimator has slightly less variance and less MSE than the OS estimator at time $j \geqslant 74$, while the SOS estimator has larger bias and larger MSE than the OS estimator when time $j \leqslant 74$. Hence, the SOS estimator is better than the OS estimator when time is close to the present, and the OS estimator is better than the SOS estimator when time is far away from the present. There results are consistent with Theorem 2 in Section 3. Simulations according to the second simulation scenario lead to the same but more confirmative conclusions and are reported in Table 4.

### 4.3 Testing the accuracy of standard error formula

We now test the accuracy of our standard error formula given in Section 3. We provide the results of simulations with $n = 100$ and data $z_j = \sum_{u=0}^{m}\{z_{ju}\}$, where $z_{ju}$ are independent and generated according to Poisson distribution with mean $\Lambda_j p_u$ for each $j = 1, \ldots, n$ and $u = 0, 1, \ldots, m$. Let $X \sim N(\mu, \sigma^2)$, $\mu = 10$, $\sigma = 4$ and define $p_u = \Pr(\min(\max([X], 0), n) = u)$, where $[X]$ denotes the integer part of $X$. We assume $\Lambda_j = 80 + 10(j - 50)^2 + 2j$. We generated 500 simulations.

For each simulated dataset, we obtained estimates of the incidence curve using the proposed approach with bandwidths $h = 0, 0.5, 1$ and 2 to test the accuracy of our standard error formulas, where $h = 0$

**Table 3** Simulation 1 to compare the performance of the smoothing one-step, one-step and the smoothing back projection estimators for the number of infections

| | | days | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 66 | 70 | 74 | 78 | 82 | 86 | 90 | 94 |
| SOS | Bias | 2.997 | 2.895 | 2.710 | 1.018 | −2.204 | −6.264 | −9.747 | −1.932 |
| | SD | 6.002 | 6.927 | 7.842 | 8.711 | 9.535 | 10.374 | 11.356 | 12.668 |
| | RMSE | 6.709 | 7.508 | 8.297 | 8.770 | 9.787 | 12.119 | 14.965 | 12.814 |
| OS | Bias | 1.049 | 1.225 | 1.643 | 0.763 | −1.532 | −5.071 | −9.054 | −3.416 |
| | SD | 5.758 | 6.947 | 8.299 | 9.581 | 10.574 | 11.223 | 11.876 | 12.869 |
| | RMSE | 5.853 | 7.054 | 8.460 | 9.611 | 10.684 | 12.316 | 14.934 | 13.315 |
| EMS | Bias | 4.868 | 7.093 | 9.337 | 7.988 | −0.234 | −16.557 | −37.607 | −47.441 |
| | SD | 7.067 | 8.690 | 10.007 | 10.481 | 9.853 | 8.105 | 5.420 | 2.584 |
| | RMSE | 8.581 | 11.217 | 13.686 | 13.178 | 9.856 | 18.435 | 37.995 | 47.511 |

**Table 4** Simulation 2 to compare the performance of the smoothing one-step, one-step and the smoothing back projection estimators for the number of infections

| | | days | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 70 | 72 | 74 | 76 | 88 | 90 | 92 | 94 |
| SOS | Bias | 1.422 | 1.309 | 1.627 | 1.013 | −0.289 | −0.600 | −0.942 | −1.721 |
| | SD | 4.092 | 4.309 | 4.537 | 4.775 | 6.420 | 6.732 | 7.058 | 7.399 |
| | RMSE | 4.332 | 4.503 | 4.819 | 4.881 | 6.426 | 6.759 | 7.121 | 7.596 |
| OS | Bias | 0.131 | 0.071 | 0.457 | −0.089 | 0.002 | −0.003 | −0.282 | −0.968 |
| | SD | 4.133 | 4.320 | 4.581 | 4.910 | 6.868 | 7.085 | 7.323 | 8.215 |
| | RMSE | 4.135 | 4.321 | 4.604 | 4.911 | 6.868 | 7.085 | 7.329 | 8.272 |
| EMS | Bias | 1.536 | 1.614 | 2.161 | 1.810 | −6.180 | −11.532 | −17.866 | −24.585 |
| | SD | 4.215 | 4.484 | 4.784 | 5.092 | 5.057 | 4.182 | 3.112 | 2.033 |
| | RMSE | 4.486 | 4.765 | 5.250 | 5.404 | 7.985 | 12.267 | 18.135 | 24.669 |

**Table 5** True and estimated standard errors for Simulation 3

| | | time $j$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | 10 | 20 | 40 | 60 | 80 | 90 |
| $h=0$ | SD | 34.094 | 25.805 | 9.813 | 10.107 | 26.574 | 43.645 |
| | $\text{SE}_{ave}$ | 30.956 | 25.992 | 11.025 | 11.112 | 26.005 | 42.106 |
| $h=0.5$ | SD | 33.178 | 25.442 | 10.328 | 10.056 | 25.879 | 40.714 |
| | $\text{SE}_{ave}$ | 30.764 | 25.858 | 10.984 | 11.098 | 25.921 | 41.902 |
| $h=1$ | SD | 32.638 | 26.093 | 10.119 | 9.794 | 25.787 | 41.689 |
| | $\text{SE}_{ave}$ | 30.392 | 25.588 | 10.889 | 11.000 | 25.653 | 41.819 |
| $h=2$ | SD | 31.165 | 25.050 | 9.756 | 9.456 | 24.715 | 41.130 |
| | $\text{SE}_{ave}$ | 28.925 | 24.574 | 10.592 | 10.696 | 24.719 | 41.279 |

corresponds to the OS estimator and $h = 0.5, 1$ and 2 correspond to the SOS estimator. The standard deviations, denoted by SD in Table 5, of 500 estimated $\hat{\Lambda}_j$, based on 500 simulations, can be regarded as the true standard errors. The average and standard deviations of 500 estimated standard errors, denoted by $\text{SE}_{ave}$ and $\text{SE}_{sd}$, summarize the overall performance of the standard error formula. Table 5 presents the results at the points at $j = 10, 20, 40, 60, 80, 90$, which correspond to the 10th, 20th, 40th, 60th, 80th and 90th percentiles of the distribution of time. The performance of the standard error formula is quite satisfactory.

## 5 Reconstructing the infection curve for the 2003 SARS epidemic in Hong Kong

The SARS epidemic poses one of the most serious global health threats since the AIDS epidemic. Here we use the proposed method to estimate the number of infected cases based on the reported cases over the duration of the epidemic. The daily number of reported cases of severe acute respiratory syndrome is obtained from the Department of Health of the Hong Kong Administrative Region. The first observed case occurred on 11th March 2003, which is set to be $j = 0$. There were 1150 cases up to 13th April 2003. On 10th April 2003 and 11th April 2003, the trend of the severe acute respiratory syndrome showed an abnormal pattern with 28 and 61 reported cases, respectively. It is suggested that a reporting delay occurred in the previous day, and some of the cases released on 11th April 2003 should be counted as the cases on the 10th April (see [12]). Averages for the two days, that is 44 and 45 cases, are used in the analysis.

There were no infection times reported. But some information exists on the incubation. Tsang et al. [30] suggest that the incubation period varies from 2 days to 11 days; whereas the Department of Health in Hong Kong reports that the incubation period varies from 2 days to 7 days. In view of these statements, Chau and Yip [12] suggested that the parameters of the distribution are chosen to satisfy the followings:

i. the minimum incubation time is 2 days;

ii. more than 90% of the infections are reported within 7 days of their infections;

iii. more than 99% of the infections are reported within 11 days of their infections.

Furthermore, Chau and Yip [12] suggested using the Weibull family to model the incubation time. Let $U$ be a continuous random variable representing the incubation time. The Weibull densities have the form: $f(u, \zeta, \eta, \theta) = \zeta^\eta \eta (u-\theta)^{\eta-1} \exp(-\zeta^\eta (u-\theta)^\eta)$, $u > \theta$, where $\zeta > 0$ and $\eta > 0$. The parameter $\theta$ represents the minimum incubation time, and $\zeta$ and $\eta$ are the parameters that together determine the shape of the curves. Following Tsang et al. [30], the Department of Health and the latest two conditions of Chau and Yip [12], we choose $\theta = 2$, $\zeta = 0.4057$ and $\eta = 1.1793$, so, $u_0 = u_1 = 0$, $u_2 = 0.2936, u_3 = 0.2516, u_4 = 0.1763, u_5 = 0.1126, u_6 = 0.0721, u_7 = 0.0382, u_8 = 0.0248, u_9 = 0.0132, u_{10} = 0.0075, u_{11} = 0.01$, where $u_j = \Pr\{j < U \leqslant j+1\}, j = 2, \ldots, 10$, $u_{11} = \Pr\{U \geqslant 11\}$.

Figure 2 gives the OS estimator for the incidence and the associated 95% pointwise confidence interval. Since the day 19/5 is the end of the SARS epidemic, the information of the infection even on the day 19/5 has already been provided by the data, the OS and EMS estimators are similar, and it is not necessary to smooth the OS estimator based on the whole data. The pattern of the infection curve is in line with the outbreak occuring (see [12]). The first infection wave, which started around 16th March 2003 in Amoy Garden, a large residential estate made up of many individual blocks. This was initiated by a patient who was treated for chronic renal failure but had been infected by SARS at Prince of Wales Hospital. He visited Amoy Garden on 14 and 19 March 2003 and used the toilet of his brother's flat. After the first wave, the epidemic had spread throughout Hong Kong. In the second wave, there were cluster infections in various hospitals. Two regional hospitals, the United Christian Hospital and the Princess Margaret Hospital, which started admitting SARS patients resulting from the first outbreak around 26th March 2003, both reported local outbreaks in the hospitals. 386 of 1755 infections were medical and healthcare workers. On 10th April 2003, home quarantine was implemented for all households with contacts of confirmed SARS patients. This preventive measure was implemented at the third wave. It seems that this preventive measure was very effective in preventing the spread in the community.

Note that the EMS estimator performs quite well retrospectively. However, in general, the current time $\tau$ will not be the end of the epidemic but may be some intermediate time when the epidemic is still running its course. To appreciate the performance of the OS estimator, the SOS estimator and EMS estimator under the case when the information is not complete, we use the observed data on and before the day 17th April. The result estimators are displayed in Figure 3. The result estimators show that the

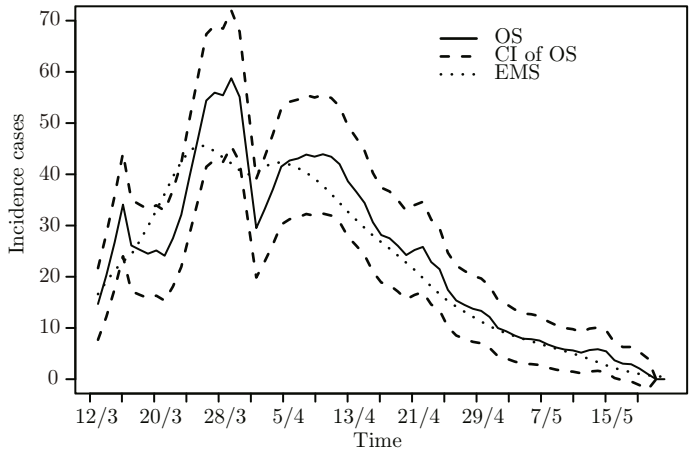


**Figure 2** The estimated incidence of SARS and their corresponding pointwise 95% confidence intervals in Hong Kong using the whole data.

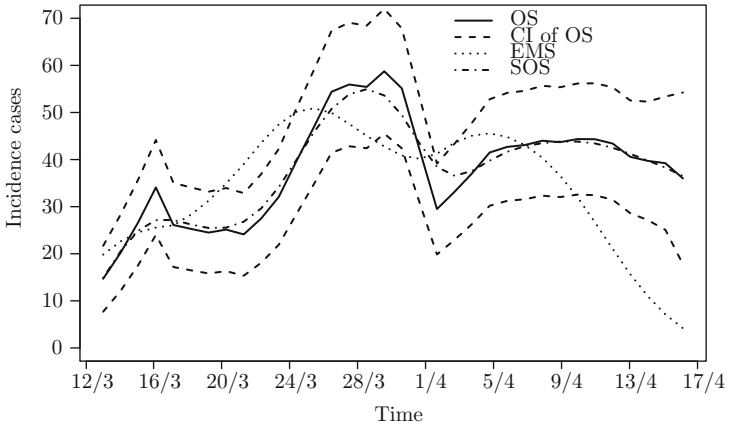


**Figure 3** The estimated incidence of SARS and their corresponding pointwise 95% confidence intervals in Hong Kong using the observed data on and before the day 17th April.

performance of the EMS method changes rapidly on 5th April from close to the one-step estimator to far away from the OS estimator. That suggests that it may be necessary to smooth the OS estimator after 5th April. Considering that the data before the day 19th May have already provided all the information on the epidemic, we can regard the OS estimator based on the whole data as the true incidence. Therefore, we can approximate the mean squared error (MSE) of $\hat{\Lambda}_j$ by $\sum_{j=1}^{n}(\hat{\Lambda}_j - \tilde{\Lambda}_j)^2$, where $\tilde{\Lambda}_j$ is the OS estimator based on the whole data. With the definition, the MSEs of the OS, SOS and EMS estimators based on the data before 17th April are 68.79, 66.59 and 5643.51, respectively, suggesting that the SOS is a little bit better than the OS estimator, the EMS estimator performs poorly when the information of epidemic is not complete, where SOS is obtained by smoothing the OS estimator after the 5th of April with the bandwidth $h = 1.5$. We choose $h$ using the method described in Section 4.

## 6 Discussion

We propose a new nonparametric method to estimate the unobserved infection numbers. The key idea is that we try to estimate the infective number based on the incubation process, which is independent among the infected individuals, rather than directly modelling the infectious process, which is difficult and may be impossible. We develop a simple closed-form expression to estimate the number of infections with the assumption of independent incubation process, which is easy to be satisfied in a real epidemic. Our method is noniterative. The simulations of Section 4 indicate that our method is more powerful, robust, accurate as well as much easier to compute than the back-projection method.

As the case counts provide very little information about recent infections, the variance of the nonparametric one-step estimator is large for the recent past. We reduce this by borrowing strength from the estimate of earlier time and although may introduce some limited bias, the resultant estimator has smaller mean square error for the recent past by choosing an adaptive bandwidth. The simulations, the SARS data and the theoretical results show that the smoothing step can improve the estimator for the recent past considerably.

The new method performs best if the estimates are only smoothed for times near the current time. If $t_j$ is far away from the current time, most of the infected on $t_j$ have been diagnosed as cases and hence, "borrowing" the information near $t_j$ gives only a marginal increase in the amount of information but can introduce bias, and as a result, increase the mean squared error. On the other hand, if $t_j$ is close to the present, the information on the numbers of infected at $t_j$ is limited, hence, the "borrowing" the information near $t_j$ can increase significantly the amount of information, even in the same time can introduce the bias, but by choosing the suitable bandwidth, the reduction in the variance may be larger than the increase in the bias, resulting in a reduction of the mean squared error.

The SOS estimator requires the numbers of infected smoothly change over time. This may not be true when some intervention is implemented. If the numbers of infected do not smoothly change over time, the estimators of the numbers of infected around the time, at which a intervention is implemented, may have a little biased (see Figure 1(c) for the second simulation in Section 4.1). A varying bandwidth with small value at implementing time point may be helpful to handle with the problem. In addition, the incubation process is estimated from exogenous data, which may add some uncertainty in the proposed estimators. The uncertainty depends on the model and the data from which the incubation process is estimated. These problems, including the model and estimation of the incubation process and their effect on the estimator of infection curve, will be considered in our future work.

**Acknowledgements** Lin's research was supported in part by National Natural Science Foundation of China (Grant Nos. 10771148, 11071197). Yip's research was supported by an RGC grant, the Chief Executive Community Project and Hong Kong Jockey Club Charities Trust.

### References

1 Anderson R M, Fraserl C, Ghanil A C, et al. Epidemiology, transmission dynamics and control of SARS: the 2002–2003 epidemic. Phil Trans R Soc Lond B, 2004, 359: 1091–1105

2 Anderson R M, Medley G F, May R M, et al. A preliminary study of the transmission dynamics of the Human Immunodeficiency Virus (HIV), the causative agent of AIDS. IMA J Math Appl Med Biol, 1986, 3: 229–263
3 Bacchetti P, Jewell N P. Nonparametric-estimation of the incubation period of aids based on a prevalent cohort with unknown infection times. Biometrics, 1991, 47: 947–960
4 Bacchetti P. Estimating the incubation period of aids by comparing population infection and diagnosis patterns. J Amer Statist Assoc, 1990, 85: 1002–1008
5 Bacchetti P, Segal M R, Hessol N A, et al. Different AIDS incubation periods and their impacts on reconstructing human immunodeficiency virus epidemics and projecting AIDS incidence. Proc Natl Acad Sci USA, 1993, 90: 2194–2196
6 Becker N G, Britton T. Statistical studies of infectious disease incidence. J Roy Stati Soc, Ser B, 1999, 61: 287–307
7 Becker N G, Watson L F, Carlin J B. A method of nonparametric back-projection and its application to aids data. Statistics in Medicine, 1991, 10: 1527–1542
8 Brookmeyer R, Gail M H. A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. J Amer Statist Assoc, 1988, 83: 301–308
9 Brookmeyer R. Minimum size of the acquired-immunodeficiency-syndrome (AIDS) epidemic in the united-states. Lancet, 1986, 2 : 1320–1322
10 Brookmeyer R. Reconstruction and future-trends of the AIDS epidemic in the united-states. Science, 1991, 253: 37–42
11 Brookmeyer R. Discussion of "Backcalculation of HIV infection rates" by Bacchetti, Segal, and Jewell. Statist Sci, 1993, 8: 102–104
12 Chau P H, Yip P S F. Monitoring the severe acute respiratory syndrome epidemic and assessing effectiveness of interventions in Hong Kong Special Administrative Region. J Epidemiology Community Health, 2003, 57: 766–769
13 Curran J W, Morgan M W, Hardy A M, et al. The epidemiology of AIDS: current status and future prospects. Science, 1985, 229: 1352–1357
14 Day N E, Gore S M, McGee M A, et al. Predictions of the AIDS epidemic in the U.K. — The use of the back projection method. Phil Trans R Soc Lond B, 1989, 325: 123–134
15 Donnelly C A, Fisher M C, Fraser C, et al. Epidemiological and genetic analysis of severe acute respiratory syndrome. Lancet, 2004, 4: 672–683
16 Fan J, Gijbels I. Local Polynomial Modelling and Its Applications. London: Chapman & Hall, 1996
17 Gail M H, Brookmeyer R. Methods for projecting course of acquired immunodeficiency syndrome epidemic. J National Cancer Institute, 1988, 80: 900–911
18 Gasser T, Muller H G, Mammitzsch V. Kernels for nonparametric curve estimation. J Roy Statist Soc B, 1985, 47: 238–252
19 Isham V. Mathematical-modeling of the transmission dynamics of HIV infection and AIDS — a review. J Roy Statist Soc Ser A-Statistics in Society, 1988, 151: 5–30
20 Isham V, Medley G. Models for infectious human diseases: Their structure and relation to data. Cambridge: Cambridge University Press, 1989
21 Lehmann E L. Elements of Large-Sample Theory. New York: Springer, 1998
22 Liao J, Brookmeyer R. An empirical Bayes approach to smoothing in backcalculation of HIV infection rates. Biometrics, 1995, 51: 579–588
23 Lin X, Carroll R J. Nonparametric function estimation for clustered data when the predictor is measured without/with error. J Amer Statist Assoc, 2000, 95: 520–534
24 Lin H Z, Yip P S F, Huggins R M. A double-nonparametric procedure for estimating the number of delay-reported cases. Stat Med, 2008, 27: 3325–3339
25 Mcevoy M, Tillett H E. Some problems in the prediction of future numbers of cases of the acquired immunodeficiency syndrome in the U.K. Lancet, 1985, ii: 541–542
26 O' Sullivan F. A statistical perspective on ill-posed inverse problems. Statist Sci, 1986, 1: 502–527
27 Rosenberg P S, Gail M H. Backcalculation of flexible linear-models of the human-immunodeficiency-virus infection curve. J Roy Statist Soc Ser C, 1991, 40: 269–182
28 Ruppert D. Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. J Amer Statist Assoc, 1997, 92: 1049–1062
29 Silverman B W, Jones M C, Wilson J D, et al. A smoothed em approach to indirect estimation problems, with particular reference to stereology and emission tomography. J Roy Statist Soc Ser B-Methodological, 1990, 52: 271–324
30 Tsang K W T, Ho P L, Ooi C G C, et al. A cluster of cases of severe acute respiratory syndrome in Hong Kong. New England J Medicine, 2003, 348: 1977–1985

## Appendix

Let $U$ denote the incubation time and $f(\cdot)$ be the density function of $U$. To determine the properties of the estimator, we impose the following regularity conditions on $\Lambda(\cdot)$, $f(\cdot)$ and the kernel function:

i. $f(\cdot)$ is a continuous function with bounded support $[0, \tau_0]$;

ii. $\Lambda(\cdot)$ is bounded and continuous function on $[0, \tau]$ and $\Lambda''(\cdot)$ is continuous at the point $t_0$;

iii. The kernel $K$ is a symmetric density function with bounded support;

iv. $\sup_s d_s < \infty$. For any fixed $j$,

$$\sigma_j^2 = \lim_{n\to\infty} \frac{\sum_{u=0}^{n-j} p_u^2 \mathrm{Var}(d_{j+u})}{\sum_{u=0}^{n-j} p_u};$$

v. $d_1, \ldots, d_n$ are independent random variables.

Condition (iv) require that the variance of the number of observed cases in unity interval is bounded.

Let $a_{ju} = p_u / \sum_{u=0}^{n-j} p_u$ for $u = 0, 1, \ldots, k, j = 1, \ldots, n$ and rewrite $\hat{z}_j$ as $\hat{z}_j = \sum_{u=0}^{n-j} a_{ju} d_{j+u}$. Here, $p_u = 0$ for $u > k$. Since

$$a_{ju} = \frac{\Pr\{u\delta_n \leqslant U < (u+1)\delta_n\}}{\Pr\{U < (n-j+1)\delta_n\}} = \frac{\int_{t_u}^{t_u+\delta_n} f(t)dt}{\int_0^{\tau-t_j+\delta_n} f(t)dt},$$

and

$$a_{j,s-j} = \frac{\int_{t_s-t_j}^{t_s-t_j+\delta_n} f(t)dt}{\int_0^{\tau-t_j+\delta_n} f(t)dt}, \tag{A.1}$$

where $t_j = j\delta_n$, $a_{j,s-j}$ is a continuous function of $t_j$ and it follows that

$$b_n(t_j, t_m) = \sum_{s=\max(j,m)}^{n} a_{j,s-j} a_{m,s-m} \mathrm{Var}(d_s)$$

is a continuous function of $t_j$ and $t_m$. From (A.1), we see that $a_{j,s-j} = O(\delta_n/(\min(\tau - t_j + \delta_n, \tau_0)))$. Then using condition (iv) and noting $\sum_{s=j}^{n} a_{j,s-j} = 1$, we have $b_n(t_j, t_j) = O(\delta_n) = O(1/n)$ if $\tau - t_j = O(1)$ and $b_n(t_j, t_j) = O(1/(nh))$ if $\tau - t_j = O(h)$, where $h \to 0$ and $nh \to \infty$. Denote

$$b(t,t) = \lim_{n\to\infty} nhb_n(t,t), \quad \text{if} \quad \tau - t = O(h),$$

and

$$\tilde{b}(t,t) = \lim_{n\to\infty} nb_n(t,t), \qquad \text{if} \quad \tau - t = O(1). \tag{A.2}$$

*Proof of Theorem* 2. Let $c_n = (nh)^{-1/2}$, $H = \mathrm{diag}(1, h)$, $\beta = (\beta_1, \beta_2)' = (\Lambda(t_0), \Lambda'(t_0))'$ and $\overline{\Lambda}(t) = \Lambda(t_0) + \Lambda'(t_0)(t - t_0)$,

$$\begin{aligned}
c_n^{-1}H(\hat{\beta} - \beta) &= c_n^{-1}H\bigg(\sum_{j=1}^{n} x_{t_j} x_{t_j}' K_h(t_j - t_0)\bigg)^{-1} \sum_{j=1}^{n} x_{t_j} K_h(t_j - t_0)\hat{z}_j \\
&\quad - c_n^{-1}H\bigg(\sum_{j=1}^{n} x_{t_j} x_{t_j}' K_h(t_j - t_0)\bigg)^{-1} \sum_{j=1}^{n} x_{t_j} x_{t_j}' K_h(t_j - t_0)\beta \\
&= c_n^{-1}H\bigg(\sum_{j=1}^{n} x_{t_j} x_{t_j}' K_h(t_j - t_0)\bigg)^{-1} \sum_{j=1}^{n} x_{t_j} K_h(t_j - t_0)\{\hat{z}_j - Ez_j\} \\
&\quad + c_n^{-1}H\bigg(\sum_{j=1}^{n} x_{t_j} x_{t_j}' K_h(t_j - t_0)\bigg)^{-1} \sum_{j=1}^{n} x_{t_j} K_h(t_j - t_0)\{\Lambda(t_j) - \overline{\Lambda}(t_j)\} \\
&\equiv A_n^{-1}B_n + A_n^{-1}C_n,
\end{aligned} \tag{A.3}$$

where $x_t = (1, t-t_0)'$, $A_n = \frac{1}{n}H^{-1}\sum_{j=1}^n x_{t_j}x'_{t_j}K_h(t_j-t_0)H^{-1}$, $B_n = \frac{1}{c_n n}H^{-1}\sum_{j=1}^n x_{t_j}K_h(t_j-t_0)\{\hat{z}_j - Ez_j\}$, $C_n = \frac{1}{c_n n}H^{-1}\sum_{j=1}^n x_{t_j}K_h(t_j-t_0)\{\Lambda(t_j)-\overline{\Lambda}(t_j)\}$. Following Fan and Gijbels [16], the conditions on $K(\cdot)$ and $t_0 \in [0,\tau]$, we have

$$A_n = \frac{1}{\tau}\begin{pmatrix} 1 & 0 \\ 0 & u_2 \end{pmatrix}(1+o_p(1)) = A + o_p(1), \tag{A.4}$$

$$C_n = \frac{h^2\Lambda''(t_0)}{2\tau c_n}\begin{pmatrix} u_2 \\ u_3 \end{pmatrix}(1+o_p(1)), \tag{A.5}$$

where $u_r = \int_{-\infty}^{+\infty} x^r K(x)dx$. Now we consider $B_n$.

$$\begin{aligned} B_n &= \frac{1}{c_n n}H^{-1}\sum_{j=1}^n x_{t_j}K_h(t_j-t_0)\sum_{u=0}^{n-j} a_{ju}\{d_{j+u}-Ed_{j+u}\} \\ &= \frac{1}{c_n n}H^{-1}\sum_{j=1}^n\sum_{s=j}^n x_{t_j}K_h(t_j-t_0)a_{j,s-j}\{d_s - Ed_s\}, \end{aligned}$$

exchanging the summation, we have

$$B_n = \frac{1}{c_n n}H^{-1}\sum_{s=1}^n\sum_{j=1}^s x_{t_j}K_h(t_j-t_0)a_{j,s-j}\{d_s-Ed_s\}.$$

Hence by condition (v), we get

$$\begin{aligned} \mathrm{Var}(B_n) &= \frac{1}{c_n^2 n^2}H^{-1}\sum_{s=1}^n\left(\sum_{j=1}^s x_{t_j}K_h(t_j-t_0)a_{j,s-j}\right)\left(\sum_{j=1}^s x_{t_j}K_h(t_j-t_0)a_{j,s-j}\right)\mathrm{Var}(d_s)H^{-1} \\ &= \frac{1}{c_n^2 n^2}H^{-1}\sum_{s=1}^n\sum_{j=1}^s x_{t_j}x'_{t_j}K_h^2(t_j-t_0)a_{j,s-j}^2\mathrm{Var}(d_s)H^{-1} \\ &\quad + \frac{1}{c_n^2 n^2}H^{-1}\sum_{s=1}^n\sum_{j\neq m,j,m=1}^s x_{t_j}x'_{t_m}K_h(t_j-t_0)K_h(t_m-t_0)a_{j,s-j}a_{m,s-m}\mathrm{Var}(d_s)H^{-1} \\ &= \frac{1}{c_n^2 n^2}H^{-1}\sum_{j=1}^n x_{t_j}x'_{t_j}K_h^2(t_j-t_0)b_n(t_j,t_j)H^{-1} \\ &\quad + \frac{1}{c_n^2 n^2}H^{-1}\sum_{j\neq m,j,m=1}^n x_{t_j}x'_{t_m}K_h(t_j-t_0)K_h(t_m-t_0)b_n(t_j,t_m)H^{-1} \\ &= hE\begin{pmatrix} 1 & (t_j-t_0)/h \\ (t_j-t_0)/h & (t_j-t_0)^2/h^2 \end{pmatrix}K_h^2(t_j-t_0)b_n(t_j,t_j)(1+o_p(1)) \\ &\quad + (n-1)hE\begin{pmatrix} 1 & (t_j-t_0)/h \\ (t_m-t_0)/h & (t_j-t_0)(t_m-t_0)/h^2 \end{pmatrix} \\ &\quad \times K_h(t_j-t_0)K_h(t_m-t_0)b_n(t_j,t_m)(1+o_p(1)). \end{aligned}$$

Following Fan and Gijbels [16], the conditions on $K(\cdot)$ and $t_0 \in (0,\tau)$, if $\tau - t_0 = O(h)$, we have

$$\mathrm{Var}(B_n) = \begin{pmatrix} A_1 & hA_2 \\ hA_2 & h^2A_3 \end{pmatrix}(1+o_p(1)), \tag{A.6}$$

where $A_1 = b(t_0,t_0)/\tau^2$, $A_2 = b^{(10)}(t_0,t_0)u_2/\tau^2$, $A_3 = b^{(11)}(t_0,t_0)u_2^2/\tau^2$, and

$$b^{(k_1,k_2)}(x_1,x_2) = \frac{\partial^{(k_1+k_2)}b(x_1,x_2)}{\partial x_1^{k_1}\partial x_2^{k_2}}, \quad k_1,k_2 = 0,1.$$

The first part of Theorem 2 follows from (A.3)–(A.6). The second part of Theorem 2 can be proved in the same way described above.