LOCAL LINEAR REGRESSION VERSUS BACKCALCULATION IN FORECASTING

by

Xiaochun Li

B.Sc., Tsinghua University, Beijing, China, 1986

M.Sc., Tsinghua University, Beijing, China, 1988

M.Sc., University of Saskatchewan, 1991

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

Department of Statistics

We accept this thesis as conforming

to the required standard

....

....

....

....

THE UNIVERSITY OF BRITISH COLUMBIA

August 1996

Abstract

The local linear forecasting estimator is proposed in this thesis as an alternative technique to either parametric regression or the backcalculation approach in the context of forecasting for independent data.

The asymptotic bias and variance of the local linear forecasting estimator are derived and used to develop procedures for the estimation of the optimal bandwidth for forecasting. Both the theoretical and the computational aspects of these procedures are explored. Simulation study shows that a cross-validation procedure has the best performance in forecasting among four bandwidth estimation procedures under study.

Simulations and statistical analyses show that the backcalculation approach is very vulnerable to violations of the assumptions underlying this approach and that its application to AIDS data fails to achieve its two primary goals, to forecast the numbers of new AIDS cases and to estimate the historical HIV infection curve.

To test the proposed forecasting estimator over parametric regression, both techniques are applied to the Canadian AIDS data and the UK AIDS data. The results of the two examples expose the weakness of parametric regression and show that the proposed technique does better than parametric regression in forecasting.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my thesis supervisor Nancy Heckman for her guidance in the development of my thesis, for her excellent advice and support, and for being an influential role model throughout my Ph.D. program. I am also very grateful to the members of my thesis committee, Jean Meloche, Steve Marion, and in particular, John Petkau for their careful reading of the manuscript and valuable comments.

In addition, I am very much indebted to Jim Ramsay for making my stay at McGill possible and for his inspiration in statistics. I would also like to acknowledge the help from Ping Yan from the Laboratory Centre for Disease Control, who provided me with the Canadian AIDS data, and the friendship and support of fellow graduate students and staff members in the Department of Statistics.

Finally, I would like to thank the University and the Department of Statistics for the indispensible financial support I received throughout my Ph.D program.

Chapter 1

Introduction

Suppose bivariate data $\{(t_i, Y_i), i = 1, \ldots, n\}$ are observed at times $0 \leq t_1 \leq t_2 \leq \ldots \leq t_n \leq 1$. Here the t_is are in $[0, 1]$, which may be normalized physical times. Thus any time beyond 1 is the "future". What can be said about the future path of this stream of data by studying the information available up to the present? Specifically, what is the "expected" value of Y at some $t > 1$?

Consider the AIDS example. A new disease, acquired immunodeficiency syndrome (AIDS), was defined by the Center for Disease Control in the United States in 1982. AIDS is believed to be caused by the human immunodeficiency virus, or HIV. This disease spread rapidly in the United States: reported cases of AIDS increased from 295 in 1981 to nearly $42,000$ in 1991 with reported deaths increasing from 126 to more than $30,000$ in the same period. AIDS is now one of the major causes of death in the United States [6]. AIDS has spread and caused serious concern around the world. Even though the current data suggest a slowing down in the growth of the AIDS epidemic in the United States, Northern Europe, Canada and Australia, the spread of HIV infection is rampant in Africa. AIDS in South America and Asia seems to be at the onset of explosive spread [6].

Uncertainty about AIDS has generated research involving epidemiologists and statisticians about the relationship between HIV and AIDS, the transmission of the disease and the future course of AIDS incidence. Mechanisms for monitoring the epidemic have been set up around the world. In the United States, the number of new AIDS cases each month has been available from the CDC (Centers for Disease Control) since 1982 (earlier data from 1975 and prior to 1982 being combined to assure confidentiality [1]) while in Canada quarterly numbers have been recorded by the FCA (the Federal Centre for AIDS) since 1979 [23] and the LCDC (the Laboratory Centre for Disease Control).

Figure 1.1 depicts the quarterly numbers of AIDS cases reported within six years of diagnosis in Canada. Those data are considered to be complete in reporting [23]. This thesis will focus on the problem of forecasting with complete or adjusted AIDS data, since adjusting AIDS data for under-reporting and reporting-delays would pose a separate research problem on its own. Note that the data shown here are not the cumulated numbers of AIDS cases. Each data point represents the number of new AIDS cases in the corresponding quarter. In this case, t_i corresponds to the end of the ith quarter and Y_i observed at t_i, the number of new AIDS cases in the ith quarter $(t_{i-1}, t_i]$. One goal in AIDS research is forecasting the number of AIDS cases that will occur in future time periods. This number is important to health organizations and governments for estimating the future demand in health care and allocating funds accordingly.

To summarize data and make forecasts based on observed data of the form $\{(t_i, Y_i), i = 1, \ldots, n\}$, various approaches exist in different theoretical frameworks:

1. the linear operator approach;

2. the regression approach;

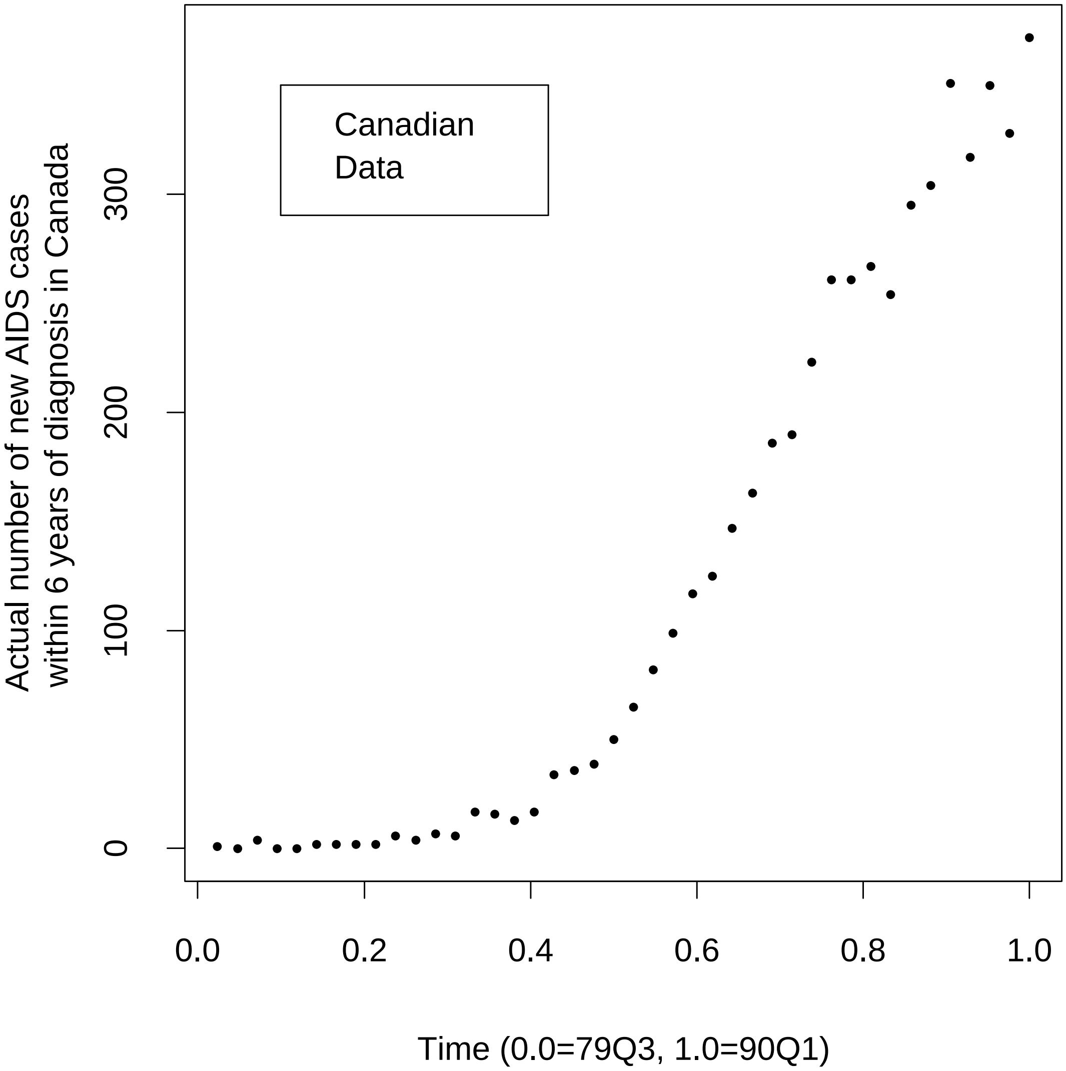

Figure 1.1: The quarterly number of AIDS cases reported within six years of diagnosis from the fourth quarter of 1979 (79Q4) to the first quarter of 1990 (90Q1). One unit in Time is 1/42 with 42 being the total number of quarters from 79Q4 to 90Q1 inclusive.

3. the time series approach.

This thesis focuses on the first two approaches, its main goal being the development of methodology for regression. Its application to AIDS data will be used as an example.

1.1 The linear operator approach

Assume the t_is are in $[0, 1]$. In the special case of equally spaced data, $t_i = i/n$, $t_n = 1$ being the "present" time at which the last observation was made. Suppose (t, Y) has a certain probabilistic structure and $E\{g(Y_i)|t_i\} = m(t_i)$, g being a transformation function and m some smooth function. For a fixed design where the t_is are not random, $E\{g(Y_i)|t_i\}$ means $E\{g(Y_i)\}$ with Y_i observed at $t = t_i$.

The linear operator approach assumes that data can be expressed in a functional form, $E\{g(Y_i)|t_i\} = L_i(I) = m(t_i)$, where m is a smooth function and L_i is a known linear operator on a certain function space where the unknown I resides. In this thesis, the case where L_i is the evaluation operator, i.e., $L_iI = I(t_i)$, is considered as a problem in regression (approach 2) in which no additional information other than that in the data is available. The function I will have a specific meaning in applications. In the AIDS example, I is the rate of HIV infection and the Y_is are the numbers of AIDS cases. The linear operator L_i for the AIDS data will be described later. The linear operator approach summarizes the data and enables forecasting by estimating I first. Then the operator L_j is applied to this estimated I to give an estimate of $m(t_j)$. This approach has the potential benefit of embedding prior knowledge of the structure of the data into the linear operator L_i. The technique called backcalculation [3] which exemplifies this approach is illustrated below through its application to the AIDS data.

The prevalent theory in AIDS research postulates that an AIDS patient was infected with HIV at a certain time s in the past which took some time to develop to the advanced stage of HIV infection and then to the time point of AIDS diagnosis, t. The length of time between the point of HIV infection at s and the point of AIDS diagnosis at t is called the incubation time, whose distribution can be modeled by $F(\cdot|s)$, a time dependent distribution function. Estimates of F have been found from cohort studies. So F is usually assumed known.

Based on the assumed relationship between HIV infection and AIDS diagnosis, the expected number of AIDS cases before t can be calculated by the formula,

$$\begin{aligned} a(t) &= E(\text{number of AIDS cases diagnosed before time } t) \\ &= \int_0^t I(s)F(t-s|s)ds, \end{aligned} \tag{1.1}$$

where $I(s)$ is the expected number of new HIV infections at time instant s, $s \geq 0$. $I(\cdot)$ is unknown and to be estimated. The convolution formula (1.1) uses an integral operator to link the HIV infections and the AIDS diagnoses.

From (1.1), for equally spaced t_is

$$\begin{aligned} m(t_i) &= E(Y_i) \\ &= E(\text{number of AIDS cases in } (t_{i-1}, t_i]) = a(t_i) - a(t_{i-1}) \\ &= \int_0^{t_i} I(s)F(t_i - s|s)ds - \int_0^{t_{i-1}} I(s)F(t_{i-1} - s|s)ds \\ &= \int_0^{t_i} I(s)\left(F(t_i - s|s) - F(t_i - \frac{1}{n} - s|s)\right) ds && (1.2) \\ &\equiv L_i(I), && (1.3) \end{aligned}$$

where $F(u|s) = 0$, if $u \leq s$. The linear operators, i.e., the L_is so defined, model the dynamics of the two processes: HIV infection and the AIDS diagnosis [3].

In general, define

$$m(t) = \int_0^t I(s)\left(F(t-s|s) - F(t - \frac{1}{n} - s|s)\right) ds. \tag{1.4}$$

Thus to forecast the number of AIDS cases in a future quarter, say $(1+\Delta_n - 1/n, 1+\Delta_n]$, one must estimate $m(t)$ at $t = 1 + \Delta_n$.

The first step to forecasting is to estimate the HIV infection curve I from the Y_is, the numbers of AIDS cases. This is done by confining $I(\cdot)$ to an appropriate function space $\mathcal{H}$ and minimizing a fitting criterion, e.g.,

$$\hat{I}(\cdot) = argmin_{I \in \mathcal{H}} \left[\sum_{i=1}^{n} (Y_i - m(t_i))^2 + \lambda \int_0^1 I''^2(t)dt \right]. \tag{1.5}$$

The parameter λ in (1.5) is the smoothing parameter which can be chosen by cross-validation. More detailed discussion of cross-validation is provided in Chapter 2 and Chapter 4. The above procedure for estimating I from the Y_is is the so-called backcalculation [3]. In an appropriately chosen $\mathcal{H}$, the minimizer $\hat{I}$ can be written as a linear combination of $1, t$ and the Riesz representors of L_is [31]. For details, see Chapter 2.

After $\hat{I}$ is obtained, $m(1+\Delta_n)$ can be predicted by using $\hat{I}$ in formula (1.4). However, for s close to 1, $\hat{I}(s)$ is usually not reliable. The information about $I(s)$ is contained in the Ys observed after time s. Thus the data contain no information about $I(s)$ for $s > 1$. Further, for $s < 1$, the closer a time point s is to 1 the less reliable the backcalculated $\hat{I}(s)$ is because there are fewer available data. A conservative approach is to be less ambitious and get a lower bound for $m(1 + \Delta_n)$ instead. A useful lower bound can be found if the disease has a long incubation. For a short term forecast, a great proportion of new HIV infections arising in time period $(1, 1 + \Delta_n]$ are still latent and thus will not contribute to the number of new AIDS cases occurring in $(1 + \Delta_n - 1/n, 1 + \Delta_n]$. Therefore, the percentage of new AIDS cases contributed by the HIV infections that occur in $(1, 1 + \Delta_n]$ may be negligible.

From (1.4), we have

$$m(1 + \Delta_n) \geq \int_0^1 I(s) \left(F(1 + \Delta_n - s|s) - F(1 + \Delta_n - \frac{1}{n} - s|s) \right) ds \equiv LB_n, \tag{1.6}$$

which in effect takes I to be zero over $(1, 1 + \Delta_n]$. An estimate of LB_n can be obtained by plugging $\hat{I}$ into the above formula for LB_n,

$$\widehat{LB}_n \equiv \int_0^1 \hat{I}(s) \left(F(1 + \Delta_n - s|s) - F(1 + \Delta_n - \frac{1}{n} - s|s) \right) ds. \tag{1.7}$$

More examples of the linear operator approach in science and engineering can be found in [29] and [30].

1.2 The regression approach

The regression approach assumes an unknown underlying regression function m for the data or the transformed data, i.e., $E\{g(Y_i)|t_i\} = m(t_i)$. It will be assumed hereafter that the data are properly transformed and $E(Y_i|t_i) = m(t_i)$ in the regression approach. The regression approach uses the data $\{Y_i\}_1^n$ to estimate m directly. This general approach includes parametric regression in which m depends on a global parametric form and local regression in which m is only assumed to be smooth. The next two sections contain details of both types of regression in the context of forecasting.

1.2.1 Parametric extrapolation

This approach assumes that the trend of the data continues over a period of time in a postulated parametric form, for example, $log(E(Y_i)) = \beta_0 + \beta_1 t_i$. One fits this parametric model to the entire data set and then uses this fit to extrapolate to obtain the forecast. In 1986 the Public Health Service in the United States gave a forecast of $270,000$ for the cumulative number of AIDS cases in the United States by the end of 1991 by extrapolating a polynomial model. The actual number of reported AIDS cases turned out to be 206,000 [6].

As noted by a few authors ([9],[20]), parametric extrapolation can provide useful short-term forecasts. However, there are a few criticisms of this approach:

1. Parametric extrapolation does not make use of any available information on the progression of the HIV infections to AIDS disease and thus may be less efficient

than backcalculation.

2. The parametric assumption is not verifiable and therefore is a considerable source of uncertainty. Thus parametric extrapolation is unable to give a long-term forecast.

As for the first criticism, it is not necessarily the case that a method that uses less data is less efficient than a method that uses more data. The discussion in Chapter 6 will show that the local linear forecasting estimator that uses only $\{(t_i, Y_i)\}_1^n$, suggests better forecasts than backcalculation that uses $\{(t_i, Y_i)\}_1^n$ and additional data.

As for the second criticism, so far as forecasting is concerned, future predictions of medium to long term pose a challenge for statisticians. If a parametric model is "correct" it will be able to give a long-term forecast. However, no models are "correct" since they are simplified representations of the real world. When a set of data is fitted to a specific parametric form, the fitted curve may be compromised to fit all the data well at the expense of local structure in the data. Especially of interest is the local structure near the present, $t = 1$. Forecasts based on the extrapolation from such a fitted curve are unlikely to be as good as forecasts based an extrapolation from a fitted curve of more recent data.

As Healy and Tillet [20] write, "It doesn't seem entirely reasonable to give the early data ... a high degree of influence in forecasting the future." When fitting the data to a log-linear model, they superimposed subjectively-chosen weights decreasing into the past. This method should be expected to be an improvement over simple parametric fitting when "appropriate" weights are chosen. It might be desirable to do away with the parametric assumption and also to couple the fitting with a data-driven method for the choice of weights. The local regression method proposed in the next section does exactly this.

1.2.2 A local linear forecasting estimator

In general, to estimate a function m at a certain point t, the approach of local linear regression assumes that m has a linear structure in a neighbourhood of t: $m(x) = \beta_0 + \beta_1(x-t)$, and estimates β_0, β_1 by weighted regression. The weights assigned to the observations close to t are much bigger than the weights assigned to the rest of observations.

That is, let the estimator of $m(t)$ be defined as follows:

$$\hat{m}_{h_n}(t) \equiv \hat{\beta}_{0,t} + \hat{\beta}_{1,t}(t-t) = \hat{\beta}_{0,t}, \tag{1.8}$$

where

$$\hat{\beta}_t = \begin{pmatrix} \hat{\beta}_{0,t} \\ \\ \hat{\beta}_{1,t} \end{pmatrix} = argmin \sum_{1}^{n} (Y_i - \beta_0 - \beta_1(t_i - t))^2 K(\frac{t_i - t}{h_n}), \tag{1.9}$$

with $K(\cdot)$ a kernel function that gives more weight to the observations close to t (defined by h_n) than the rest of the observations, and h_n the so-called bandwidth that determines the size of the neighbourhood of t. For example, if the kernel function is the indicator function over $[-1,1]$, then $K((t_i - t)/h_n) = 1$ for those t_is with $|t_i - t| \leq h_n$ and 0 otherwise.

Since $\hat{m}_{h_n}(t)$ is calculated from the Y_is, it is a random variable with statistical properties which are affected by the bandwidth h_n and the kernel K. In practice the choice of K has little effect on $\hat{m}_{h_n}(t)$ while the choice of h_n has a large effect. Therefore we assume K is specified in Chapter 4 and describe data driven choice of h_n.

For now assume h_n is given. Unless m is a straight line, $\hat{m}_{h_n}(t)$ is in general a biased estimator of $m(t)$, that is, $Bias\{\hat{m}_{h_n}(t)\} \equiv E\{\hat{m}_{h_n}(t)|t_1, \ldots, t_n\} - m(t) \neq 0$. Note that if the kernel function is the indicator function over $[-1,1]$, the bandwidth h_n reflects the number of Y_is used in the regression. A large h_n means that a large number of Y_is is used in the regression and a small h_n means otherwise. Therefore a large bandwidth causes $\hat{m}_{h_n}(t)$ to have a small variance and a small bandwidth causes $\hat{m}_{h_n}(t)$ to have a large variance. For commonly used kernels, when h_n is small, $\hat{m}_{h_n}(t)$ will have a small bias but a large variance; when h_n is large, $\hat{m}_{h_n}(t)$ will have a large bias but a small variance. If h_n is small enough, $\hat{m}_{h_n}(t)$ will behave like an interpolant while if h_n is large enough, $\hat{m}_{h_n}(t)$ will become the familiar least squares fit. A balance between the bias and the variance is desirable and this balance is usually achieved by minimizing the mean squared error of $\hat{m}_{h_n}(t)$, i.e., $E\{(\hat{m}_{h_n}(t) - m(t))^2 \,|t_1, \ldots, t_n\}$.

To predict m at $1 + \Delta_n$, let $t = 1 + \Delta_n$ in (1.9). Therefore, $\hat{m}_{h_n}(1 + \Delta_n) = \hat{\beta}_{0,1+\Delta_n}$. To avoid cumbersome calculations of the asymptotics later in Chapter 3, an algebraic substitution is employed to re-center the data at 1 instead of $1 + \Delta_n$. Note that

$$\begin{aligned}
\hat{\beta}_{1+\Delta_n} &= \begin{pmatrix} \hat{\beta}_{0,1+\Delta_n} \\ \hat{\beta}_{1,1+\Delta_n} \end{pmatrix} \\
&= argmin \sum_1^n (Y_i - \beta_0 - \beta_1(t_i - 1 - \Delta_n))^2 K(\frac{t_i - 1 - \Delta_n}{h_n}) \\
&= argmin \sum_1^n (Y_i - (\beta_0 - \beta_1\Delta_n) - \beta_1(t_i - 1))^2 K(\frac{t_i - 1 - \Delta_n}{h_n}). \qquad (1.10)
\end{aligned}$$

Let

$$\beta_0^* = \beta_0 - \beta_1\Delta_n, \quad \beta_1^* = \beta_1, \quad K^*(\frac{t_i - 1}{h_n}) = K(\frac{t_i - 1 - \Delta_n}{h_n}). \qquad (1.11)$$

Then

$$\hat{\beta}_1^* = \begin{pmatrix} \hat{\beta}^*_{0,1} \\ \hat{\beta}^*_{1,1} \end{pmatrix} = argmin \sum_1^n (Y_i - \beta_0^* - \beta_1^*(t_i - 1))^2 K^*(\frac{t_i - 1}{h_n}), \qquad (1.12)$$

and $\hat{m}_{h_n}(1 + \Delta_n) = \hat{\beta}^*_{0,1} + \hat{\beta}^*_{1,1}\Delta_n$.

Formulation (1.12) will be used hereafter with the superscript $*$ dropped. The notation $\hat{m}_{h_n,1}(1 + \Delta_n)$ instead of $\hat{m}_{h_n}(1 + \Delta_n)$ will be used to denote the forecasting estimator that forecasts Δ_n ahead with data centered at $t = 1$ and with bandwidth h_n. To summarize:

Definition 1.1 *The estimator* $\hat{m}_{h_n,1}(1+\Delta_n)$ *of* $m(1+\Delta_n)$ *is defined as*

$$\hat{m}_{h_n,1}(1+\Delta_n) = \hat{\beta}_{0,1} + \hat{\beta}_{1,1}\Delta_n,$$

where

$$\hat{\beta}_1 = \begin{pmatrix} \hat{\beta}_{0,1} \\ \\ \hat{\beta}_{1,1} \end{pmatrix} = argmin \sum_{1}^{n} (Y_i - \beta_0 - \beta_1(t_i - 1))^2 K(\frac{t_i - 1}{h_n}). \tag{1.13}$$

The bandwidth h_n can be chosen by plug-in or cross-validation approaches. The development of those approaches for forecasting in Chapter 4 is the main contribution of this thesis. Results of the statistical properties derived in Chapter 3 serve as the cornerstone for all bandwidth estimation procedures.

In the next chapter, theoretical results for the linear operator approach will be presented and applied to the AIDS data.

Chapter 2

The linear operator approach

The linear operator approach will be discussed under the theoretical framework of a Hilbert space. Standard results will be presented without proof. Proofs and other details can be found in [31]. Examples of the linear operator approach, details of computation and some asymptotic results can be found in [28]-[31].

The following theorem is central to the linear operator approach.

Theorem 2.1 *Assume the following conditions:*

1. $\mathcal{H}$ *is a Hilbert space over the reals,* $\mathcal{R}$, *with inner product* $<,>$ *and norm* $\|\cdot\|$;

2. *For each* $i \in \{1, \ldots, n\}$, $L_i : \mathcal{H} \to \mathcal{R}$, *is a bounded linear functional;*

3. $P : \mathcal{H} \to \mathcal{H}_1$, *is a projection operator with* $\mathcal{H}_1$ *a subspace of* $\mathcal{H}$;

4. *The null space of* P *is* $\mathcal{H}_0 = span\{\phi_1, \ldots, \phi_{n_0}\}$, *where the* $\phi_j s$ *are linearly independent and* $n_0 < \infty$;

5. $\mathcal{F}(\boldsymbol{Y}, L_1(I), \ldots, L_n(I)) : \mathcal{R}^n \times \mathcal{R}^n \to \mathcal{R}$ *is a real function, where* $\boldsymbol{Y} = (y_1, \ldots, y_n)^t$.

For given $\lambda \geq 0$, if the minimizer $\hat{I}$ of

$$\mathcal{F}(\boldsymbol{Y}, L_1(I), \ldots, L_n(I)) + \lambda \|P(I)\|^2 \tag{2.1}$$

exists in $\mathcal{H}$, it is of the form:

$$\hat{I} = \sum_{j=1}^{n_0} d_j \phi_j + \sum_{j=1}^{n} c_j \xi_j, \tag{2.2}$$

where the d_js and c_js are real numbers, $\xi_j = P(\eta_j)$ and η_j is the Riesz representor of L_j, that is, $L_j(I) =< \eta_j, I >$ for all $I \in \mathcal{H}$.

Corollary 2.1 *Assume the conditions in Theorem 2.1. In addition, assume the matrices $T = (T_{ij})_{n \times n_0}$, $T_{ij} = L_i(\phi_j)$ and $\Sigma = (\Sigma_{ij})_{n \times n}$, $\Sigma_{ij} =< \xi_i, \xi_j >$ are of full column rank, and that the w_js are positive.*

Then

$$\frac{1}{n} \sum_{j=1}^{n} (Y_j - L_j(I))^2 w_j + \lambda \|P(I)\|^2 \tag{2.3}$$

has a unique minimizer $\hat{I}$ in $\mathcal{H}$ and is as in (2.2) with

$$\boldsymbol{d} = (d_1, \ldots, d_{n_0})^t = (T^t A^{-1} T)^{-1} T^t A^{-1} \boldsymbol{Y}, \tag{2.4}$$

$$\boldsymbol{c} = (c_1, \ldots, c_n)^t = A^{-1}[I - T(T^t A^{-1} T)^{-1} T^t A^{-1}] \boldsymbol{Y}, \tag{2.5}$$

where $A = \Sigma + n\lambda I$.

Remarks:

1. Theorem 2.1 greatly reduces the complexity of the task of finding the minimizer $\hat{I}$ in $\mathcal{H}$ (often of infinite dimension) to finding $\hat{I}$ in the finite dimensional subspace spanned by $\phi_1, \ldots, \phi_{n_0}, \xi_1, \ldots, \xi_n$. Roughly speaking, a parametric model of span $\phi_1, \ldots, \phi_{n_0}$ determined by P, and the additional functions, $\xi_1, \ldots, \xi_n$, allow us a more flexible fit to our n data points.

2. Theorem 2.1 is a very general result in that $\mathcal{F}$ can be a general function. However, we are only interested in $\mathcal{F}$ being $-n^{-1}log(likelihood)$, in which case (2.1) is called a penalized likelihood. When the Y_js are independent and normally distributed with mean $L_j(I)$ and variance $(2w_j)^{-1}$, $\mathcal{F}$ is $-n^{-1}log$(normal likelihood) and is the first term as in (2.3). We will assume (2.1) to be a penalized likelihood hereafter. The number λ, known as the smoothing parameter, governs the relative importance of the goodness of fit of the data to that of the penalty P. If λ is zero, the fit of the data is of the utmost importance and the penalty is ignored, with the result that $\hat{I}$ is the maximum likelihood estimator. On the other hand, a large value of λ forces $\hat{I}$ close to the null space of P and results in an estimate close to a parametric fit. Any λ between these two extremes reflects a compromise between the goodness of fit to the data and the form of I emphasized by the penalty.

3. An automatic, i.e., data-driven choice of the smoothing parameter λ might be desired. Cross-validation is a technique commonly used for the estimation of an optimal λ. Details of this method will be presented in Chapter 4.

Though the Riesz representation theorem assures the existence of the η_js, the representors of the L_js, it does not give the analytical forms of the η_js. The theory of reproducing kernel Hilbert spaces (r.k.h.s.) makes it possible to calculate those ξ_js analytically. This is important in applications.

This chapter will present the relevant results of r.k.h.s. and then apply them to the case of backcalculation.

2.1 Some results on reproducing kernel Hilbert spaces

The theory of r.k.h.s. is very powerful and very useful in the linear operator approach in general.

Definition 2.1 *Let $\mathcal{H}$ be a Hilbert space of real-valued functions over* $[0,1]$,

$\mathcal{H} \equiv \{I: \ [0,1] \to \mathcal{R}\}$.

Then $\mathcal{H}$ is a r.k.h.s. if and only if for all $t \in [0,1]$,

$$L_t: \ \mathcal{H} \to \mathcal{R} \ \ \textit{with} \ \ L_t(I) = I(t)$$

is a bounded linear functional.

Definition 2.2 *If $\mathcal{H}$ is a r.k.h.s., then by the Riesz representation theorem, for any t, there exists a function $K_t \in \mathcal{H}$ such that*

$$< I, K_t >= I(t), \quad \textit{for all } I \in \mathcal{H}.$$

$$\textit{Let} \ \ K(s,t) \equiv K_t(s): \ [0,1] \times [0,1] \to \mathcal{R}. \tag{2.6}$$

Then K is called the reproducing kernel of $\mathcal{H}$.

As shown in the following lemma, in a r.k.h.s. it is very easy to use the reproducing kernel to calculate the Riesz representor of a bounded linear operator.

Lemma 2.1 *Assume that L is a bounded linear operator on a r.k.h.s., $\mathcal{H}$. Its Riesz representor is η where $\eta(t) = L(K_t)$.*

Proof: Let η be the Riesz representor of L in $\mathcal{H}$. So for all $I \in \mathcal{H}$, $L(I) =< \eta, I >$. In particular, for $I = K_t$, $L(K_t) =< \eta, K_t >$. By Definition 2.2, $< \eta, K_t >$ is also equal to $\eta(t)$. So $\eta(t) = L(K_t)$. □

A commonly chosen r.k.h.s. is the class of smooth functions with square-integrable rth derivatives.

Definition 2.3 $W_2^r[0,1] \equiv \{I : [0,1] \to \mathcal{R} : \ I, \, I', \ldots, I^{(r-1)}$ *are absolutely continuous in* $[0,1]$ *with* $\int_0^1 I^{(r)}(u)^2 du < \infty\}$. *The inner product is defined as:*

$$< f, g >= \sum_{j=0}^{r-1} f^{(j)}(0) g^{(j)}(0) + \int_0^1 f^{(r)}(u) g^{(r)}(u) du.$$

Standard results say that this space is a r.k.h.s. and has a kernel of a known form.

Lemma 2.2 $W_2^r[0,1]$ *is a r.k.h.s. with the reproducing kernel,*

$$K(s,t) = \sum_{j=0}^{r-1} \frac{s^j}{j!}\frac{t^j}{j!} + \int_0^1 \left(\frac{1}{(r-1)!}\right)^2 (s-u)_+^{r-1}(t-u)_+^{r-1} du, \tag{2.7}$$

where $(s-u)_+ = s-u$, *if* $s > u$ *and* 0 *otherwise.*

2.2 Application to the AIDS data: backcalculation

Recall that in the linear operator approach in the AIDS example in Section 1.1, $E(Y_j)$, the expected number of new AIDS cases in the jth quarter, is assumed to be linked to an unknown function I by a linear operator L_j,

$$L_j(I) = \int_0^{t_j} I(s)(F(t_j - s|s) - F(t_j - \frac{1}{n} - s|s))ds, \tag{2.8}$$

where I is the HIV infection function and $F(\cdot|s)$ is the distribution function of the incubation time from HIV infection to AIDS diagnosis at the instant s. F is assumed known so the operator L_j is known. Obviously L_j is a linear operator.

In this application, the HIV infection function I will be assumed to be in $\mathcal{H} = W_2^r[0,1]$ with $r = 2$. This is a very minimal assumption on I because $W_2^2[0,1]$ is a very general class of functions. No specific assumptions are made on those functions other than smoothness and integrability: I and I' are absolutely continuous with $\int_0^1 I''(u)^2du < \infty$.

Recovering the historical HIV infection pattern I is an important goal in AIDS research. It is believed that I is a smooth curve. Therefore rough Is should be penalized. The quantity $\int_0^1 I''(u)^2du$ is a measure of the roughness of I so it can be used as the penalty term.

To put it statistically, finding a reasonably smooth I in $W_2^2[0,1]$ that fits the observed AIDS incidence data well is achieved by finding the $\hat{I}$ that minimizes

$$\mathcal{F}(\boldsymbol{Y}, L_1(I), \ldots, L_n(I)) + \lambda \int_0^1 I''(u)^2 du, \tag{2.9}$$

where $\mathcal{F}$ is the negative of a loglikelihood function divided by n, $I \in W_2^2[0,1]$ and $\lambda \geq 0$.

The smoothing parameter λ places a control on the roughness of I relative to the goodness of fit of the $L_j(I)$s to the Y_js. Minimizing the penalized likelihood means choosing an I in $W_2^2[0,1]$ with a tradeoff between the goodness of fit to the data and the roughness of I. This tradeoff is determined by the value of the smoothing parameter λ. An optimal λ for this tradeoff minimizes the prediction error,

$$PE(\lambda) = n^{-1} \sum_{i=1}^{n} \{Y_i^* - \hat{m}_\lambda(t_i)\}^2,$$

where the Y_i^*s are hypothetical new observations at t_is, $\hat{m}_\lambda(t_i) = L_i(\hat{I}_\lambda)$ and $\hat{I}_\lambda$ minimizes (2.9). This optimal λ can be estimated by a standard technique called cross-validation. Further details on cross-validation will follow in Chapter 4.

The minimizer $\hat{I}$ of (2.9) can be determined by using Theorem 2.1 and the theory of r.k.h.s. in Section 2.1.

Theorem 2.2 *If the minimizer $\hat{I}$ of (2.9) exists in $W_2^2[0,1]$, it takes the form:*

$$\hat{I}(t) = \beta_0 + \beta_1 t + \sum_{j=1}^{n} \beta_{j+1} \xi_j(t), \tag{2.10}$$

where

$$\begin{aligned} \xi_j(t) \quad = \quad & \int_0^{t_j} \left[st(s \wedge t) - (s+t)\frac{(s \wedge t)^2}{2} + \frac{(s \wedge t)^3}{3} \right] \times \\ & \qquad (F(t_j - s|s) - F(t_j - \frac{1}{n} - s|s))ds, \end{aligned} \tag{2.11}$$

and $s \wedge t = s$ if $s < t$ and t otherwise.

Theorem 2.2 enables one to find the minimizer in a finite dimensional subspace instead of in $W_2^2[0,1]$, a space of infinite dimension.

The succeeding lemmas verify the conditions of Theorem 2.1 for its application to the AIDS data. So the ensemble of the following lemmas serves as the proof of Theorem 2.2.

Lemma 2.2 in the previous section assures that Condition 1 of Theorem 2.1 is satisfied. $W_2^2[0,1]$ is a Hilbert space with the inner product

$$< f, g >= f(0)g(0) + f'(0)g'(0) + \int_0^1 f''(u)g''(u)du, \quad f, g \in W_2^2[0,1],$$

and the norm induced by the above inner product,

$$\|I\| = \left[I(0)^2 + I'(0)^2 + \int_0^1 I''(u)^2 du \right]^{1/2}.$$

The lemma below checks Condition 2 of Theorem 2.1.

Lemma 2.3 *Suppose that for any $s > 0$, $F(\cdot|s)$ is a distribution function with $F(u|s) = 0$ if $u \leq 0$ and that $F(u|s)$ is continuous in both u and s.*

For each $j \in \{1, \ldots, n\}$, the operator

$$L_j : \ L_j(I) = \int_0^{t_j} I(s)(F(t_j - s|s) - F(t_j - \frac{1}{n} - s|s))ds, \ I \in W_2^2[0,1]$$

is a bounded linear operator on $W_2^2[0,1]$.

Proof: L_j is well-defined since $I(s)(F(t_j - s|s) - F(t_j - \frac{1}{n} - s|s))$ is continuous in s and thus is integrable. The linearity of L_j is obvious.

By the definition of the boundedness of a linear operator, L_j is bounded if there exists C, $0 < C < \infty$, such that

$$|L_j(I)| \leq C\|I\|, \tag{2.12}$$

for all $I \in W_2^2[0,1]$.

Since $0 \leq F(t_j - s|s) - F(t_j - \frac{1}{n} - s|s) \leq 1$, it follows that

$$\begin{aligned} |L_j(I)| &= |\int_0^{t_j} I(s)(F(t_j - s|s) - F(t_j - \frac{1}{n} - s|s))ds| \\ &\leq \int_0^{t_j} |I(s)|ds \leq \int_0^1 |I(s)|ds \leq \Big(\int_0^1 I(s)^2 ds\Big)^{1/2}. \end{aligned} \tag{2.13}$$

To bound $\int_0^1 |I(s)|ds$, first consider

$$I(s) = I(0) + sI'(0) + \int_0^s \int_0^u I''(v)dvdu. \tag{2.14}$$

Applying the inequality

$$(a+b+c)^2 \leq 3(a^2+b^2+c^2),$$

to (2.14) gives

$$I(s)^2 \leq 3(I(0)^2 + s^2 I'(0)^2 + (\int_0^s \int_0^u I''(v)dvdu)^2). \tag{2.15}$$

By Schwarz's inequality,

$$(\int_0^s \int_0^u I''(v)dvdu)^2 \leq (\int_0^s \int_0^u I''(v)^2 dvdu)(\int_0^s \int_0^u 1^2 dvdu),$$

which is bounded by $\int_0^1 I''(v)^2 dv$ for $0 \leq s \leq 1$.

For $0 \leq s \leq 1$, applying the relationship in (2.15) yields:

$$\begin{aligned} I(s)^2 &\leq 3(I(0)^2 + s^2 I'(0)^2 + \int_0^1 I''(v)^2 dv) \\ &\leq 3(I(0)^2 + I'(0)^2 + \int_0^1 I''(v)^2 dv) \\ &= 3\|I\|^2. \end{aligned} \tag{2.16}$$

Using (2.16) in (2.13) gives $|L_j(I)| \leq 3^{1/2}\|I\|$. Therefore L_j is a bounded linear operator on $W_2^2[0,1]$. □

For Condition 3 of Theorem 2.1, the following lemma gives the projection operator P corresponding to the penalty term $\int_0^1 I''(u)^2 du$ and the basis functions ϕ_js that span the null space of P.

Lemma 2.4 *Let*

$$P : W_2^2[0,1] \to W_2^2[0,1] : \ P(I)(t) = I(t) - I(0) - I'(0)t. \tag{2.17}$$

Then P is a projection operator onto $\mathcal{H}_1$ with $\mathcal{H}_1 \oplus \mathcal{H}_0 = W_2^2[0,1]$, where $\mathcal{H}_0 = span\{1, t\}$ is the null space of P and $\|P(I)\|^2 = \int_0^1 I''(u)^2 du$.

The next lemma uses the results of the r.k.h.s. to calculate the Riesz representor η_j of L_j in $W_2^2[0,1]$ and its projected image $\xi_j = P(\eta_j)$ for L_j as in (2.8) and P as in (2.17).

Lemma 2.5 *In $W_2^2[0,1]$, ξ_j, the projection of the Riesz representor η_j of L_j is:*

$$\begin{aligned} \xi_j(t) &= \int_0^{t_j} \left[st(s \wedge t) - (s+t)\frac{(s \wedge t)^2}{2} + \frac{(s \wedge t)^3}{3} \right] \times \\ &\quad (F(t_j - s|s) - F(t_j - \frac{1}{n} - s|s))ds. \end{aligned} \tag{2.18}$$

Remark: The ξ_js are not splines (cubic polynomials) as in the trivial case when $L_j(I) = I(t_j)$. One can show that ξ_j defined by (2.18) is an increasing function with $\xi_j(0) = 0$, concave in $[0, t_j]$ and a straight line beyond t_j.

Proof: For $r = 2$, (2.7) gives the reproducing kernel as

$$K(s,t) = 1 + st + st(s \wedge t) - (s+t)\frac{(s \wedge t)^2}{2} + \frac{(s \wedge t)^3}{3}. \tag{2.19}$$

Recall that $K_t(\cdot) = K(\cdot, t)$. By Lemma 2.1, the Riesz representor η_j of L_j evaluated at t is:

$$\eta_j(t) =< \eta_j, K_t >= L_j(K_t) = \int_0^{t_j} K_t(s)(F(t_j - s|s) - F(t_j - \frac{1}{n} - s|s))ds.$$

So

$$\begin{aligned} \xi_j(t) &= P(\eta_j)(t) = \eta_j(t) - \eta_j(0) - \eta_j'(0)t \\ &= L_j(K_t) - L_j(K_0) - \frac{d}{dt}(L_j(K_t)) \mid_{t=0} t \ . \end{aligned} \tag{2.20}$$

One can show that

$$\frac{d}{dt}(L_j(K_t))\,|_{t=0} = \int_0^{t_j} s(F(t_j - s|s) - F(t_j - \frac{1}{n} - s|s))ds.$$

Therefore

$$\xi_j(t) = \int_0^{t_j} (K_t(s) - 1 - st)(F(t_j - s|s) - F(t_j - \frac{1}{n} - s|s))ds. \qquad \square$$

The results in this chapter will be used in Chapter 6.

The next chapter will contain the derivation of the statistical properties of the local linear forecasting estimator $\hat{m}_{h_n,1}(1 + \Delta_n)$ for the regression approach.

Chapter 3

Asymptotics for the local linear regression estimator

This chapter will cover the asymptotic properties of $\hat{m}_{h_n,t}(t+\Delta_n)$, as defined below based on data $\{(t_i, Y_i)\}_1^n$, for t_is random and t_is nonrandom respectively. These asymptotic results will help us understand the local linear forecasting estimator and choose an optimal bandwidth for the estimator. The results and conditions are different from those in non-forecasting regression problems (see, e.g. [8]). However, the proofs are similar.

In non-forecasting regression problems, e.g., estimating $m(t)$ by local regression for t in the domain of design points t_is, data on both sides of t are used in regression. Therefore usually a kernel function K symmetric about 0 is used. Often in non-forecasting regression problems, K is assumed to be a density function. However, in forecasting data are centered at the boundary $t = 1$, so only the left part of the kernel function K is actually used because there are no data beyond $t = 1$. For this reason, the local linear forecasting estimator is defined by a kernel function K with a negative support. Moreover, the local linear forecasting estimator does not require K to be a density function but a general function that satisfies Condition 3.1 below.

Definition 3.1 *Let $\hat{m}_{h_n,t}(t') = \hat{\beta}_{0,t} + \hat{\beta}_{1,t}(t' - t)$, where $\hat{\beta}_{0,t}$ and $\hat{\beta}_{1,t}$ minimize $\sum_1^n (Y_i - \beta_0 - \beta_1(t_i - t))^2 K((t_i - t)/h_n)$ with K a kernel function having a negative support.*

Note that the data are centered at t in Definition 3.1. The estimator $\hat{m}_{h_n,t}(t')$ forecasts $t' - t$ ahead of t using data prior to t since the kernel function K has a negative support. The main application of the asymptotic results for $\hat{m}_{h_n,t}(t')$ will be to the case $t = 1$ and $t' - t = \Delta_n$ (see (1.13)). However, the more general asymptotic results for t and t' in a neighbourhood of 1 will be used in Chapter 4.

From Definition 3.1, the asymptotic properties of $\hat{m}_{h_n,t}(t + \Delta_n)$ are completely determined by those of $\hat{\beta}_t \equiv (\hat{\beta}_{0,t}, \hat{\beta}_{1,t})^t$. Note that the superscript "t" of a vector or a matrix indicates the action of taking the transpose of a vector or a matrix and should not be confused with the scalar t as either an argument of a function or a subscript of a variable. The asymptotic bias and variance of $\hat{\beta}_t$ will be presented first.

For the case of random t_is, assume the following conditions:

1. $Y_i = m(t_i) + \epsilon_i$, where the ϵ_is are independent with mean 0 and variance σ^2;

2. t_is are iid with density $f(\cdot)$ and are independent of the ϵ_is;

3. the density function f is known and satisfies: f is bounded away from 0 and f is continuous on $[0, 1]$;

4. K is square integrable on its compact support $[-1, 0]$ with

$$u_0 > 0, \ u_0 u_2 - u_1^2 > 0, \tag{3.1}$$

where $u_i = \int_{-1}^{0} u^i K(u) du$;

let $u_i^* = \int_{-1}^{0} u^i K(u)^2 du$;

5. m'' is continuous over $[0, 1+\Delta_n]$;

6. $h_n \to 0$ and $nh_n \to \infty$.

Theorem 3.1 *Let $\boldsymbol{e} = (1,1)^t$. Under the above assumptions, as $n \to \infty$,*

$$\frac{2}{h_n^2}\begin{pmatrix} E(\hat{\beta}_{0,t}|\boldsymbol{t}) - m(t) \\ h_n\left(E(\hat{\beta}_{1,t}|\boldsymbol{t}) - m'(t)\right) \end{pmatrix} - m''(t)\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1}\begin{pmatrix} u_2 \\ u_3 \end{pmatrix} = o_p(1)\boldsymbol{e}, \tag{3.2}$$

and

$$nh_nVar\left(\begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}\begin{pmatrix} \hat{\beta}_{0,t} \\ \hat{\beta}_{1,t} \end{pmatrix}|\boldsymbol{t}\right) -$$

$$\frac{\sigma^2}{f(t)}\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1}\begin{pmatrix} u_0^* & u_1^* \\ u_1^* & u_2^* \end{pmatrix}\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1} = o_p(1)\boldsymbol{e}, \tag{3.3}$$

uniformly in $t \in [a_n, 1]$ with $\liminf a_n/h_n \geq 1$.

Remarks:

1. Note that here "$\theta_n(t) = O_p((nh)^{-1/2})$ uniformly for $t \in [a_n, 1]$" means that

$$\lim_{C\to+\infty}\lim_{n\to\infty}\sup_{t\in[a_n,1]} P((nh)^{1/2}|\theta_n(t)| > C) = 0,$$

which is weaker than

$$\lim_{C\to+\infty}\lim_{n\to\infty} P(\sup_{t\in[a_n,1]}(nh)^{1/2}|\theta_n(t)| > C) = 0.$$

Thus, for a fixed sample size, some values of $\theta_n(t)$ could be far from 0.

2. It is not necessary to require that K have an one-sided support if a bandwidth is given because there are no data for $t > 1$ and thus any kernel will have an one-sided support automatically. However, some bandwidth selection procedures (e.g. FCV in Chapter 4) need this requirement to have certain asymptotic properties (see Chapter 4).

The proof of Theorem 3.1 will be postponed until its corollaries are presented and proven. The results in Theorem 3.1 make calculations of the asymptotic bias and variance of $\hat{m}_{h_n,t}(t+\Delta_n)$ very easy. The corollary below follows easily from Theorem 3.1.

Corollary 3.1 *Assume Conditions 1, 2, 3, 4 and 5. In addition assume*

6′. $\Delta_n = Dn^{-1/5}$ and $h_n = Hn^{-1/5}$ for some D, $H > 0$. Let $\delta = D/H$.

Let

$$Bias(\hat{m}_{h_n,t}(t+\Delta_n)|\boldsymbol{t}) \equiv E(\hat{m}_{h_n,t}(t+\Delta_n)|\boldsymbol{t}) - m(t+\Delta_n). \tag{3.4}$$

As $n \to \infty$,

$$\frac{2}{\Delta_n^2}Bias(\hat{m}_{h_n,t}(t+\Delta_n)|\boldsymbol{t}) - m''(t)\left((\frac{u_2^2 - u_1u_3}{\delta^2} + \frac{u_0u_3 - u_1u_2}{\delta})/(u_0u_2 - u_1^2) - 1 \right) = o_p(1) \tag{3.5}$$

and

$$nf(t)\Delta_n Var(\hat{m}_{h_n,t}(t+\Delta_n)|\boldsymbol{t}) - \sigma^2\delta(1,\delta)\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1}\begin{pmatrix} u_0^* & u_1^* \\ u_1^* & u_2^* \end{pmatrix}\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1}\begin{pmatrix} 1 \\ \delta \end{pmatrix} = o_p(1) \tag{3.6}$$

uniformly in $t \in [a_n, 1]$ with $\liminf a_n/h_n \geq 1$.

Proof of Corollary 3.1:

Taylor expansion of $m(t+\Delta_n)$ at t yields:

$$m(t+\Delta_n) = m(t) + \Delta_n m'(t) + \frac{\Delta_n^2}{2}m''(t) + o(\Delta_n^2), \tag{3.7}$$

uniformly for $t \in [0,1]$.

$$\begin{aligned} \text{So} \quad & \frac{2}{\Delta_n^2}Bias(\hat{m}_{h_n,t}(t+\Delta_n)|\boldsymbol{t}) \\ = \quad & \frac{2}{\Delta_n^2}\left[E(\hat{m}_{h_n,t}(t+\Delta_n)|\boldsymbol{t}) - m(t+\Delta_n)\right] \end{aligned}$$

$$
\begin{aligned}
&= \frac{2}{\Delta_n^2}\left[E(\hat{m}_{h_n,t}(t+\Delta_n)|\boldsymbol{t}) - m(t) - \Delta_n m'(t) - \frac{\Delta_n^2}{2} m''(t) + o(\Delta_n^2)\right] \\
&= \frac{2}{\delta^2 h_n^2}\left[(1,\delta)\begin{pmatrix} E(\hat{\beta}_{0,t}|\boldsymbol{t}) - m(t) \\ h_n\left(E(\hat{\beta}_{1,t}|\boldsymbol{t}) - m'(t)\right)\end{pmatrix}\right] - m''(t) + o(1) \\
&= (\frac{1}{\delta^2},\frac{1}{\delta})\frac{2}{h_n^2}\begin{pmatrix} E(\hat{\beta}_{0,t}|\boldsymbol{t}) - m(t) \\ h_n\left(E(\hat{\beta}_{1,t}|\boldsymbol{t}) - m'(t)\right)\end{pmatrix} - m''(t) + o(1).
\end{aligned}
\tag{3.8}
$$

By result (3.2) of Theorem 3.1, as $n \to \infty$ this expression converges in probability uniformly to

$$
\begin{aligned}
&(\frac{1}{\delta^2},\frac{1}{\delta})m''(t)\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1}\begin{pmatrix} u_2 \\ u_3 \end{pmatrix} - m''(t) \\
&= m''(t)\left((\frac{u_2^2 - u_1 u_3}{\delta^2} + \frac{u_0 u_3 - u_1 u_2}{\delta})/(u_0 u_2 - u_1^2) - 1\right).
\end{aligned}
\tag{3.9}
$$

So (3.5) is proved. Now consider the variance.

$$
\begin{aligned}
&nf(t)\Delta_n Var(\hat{m}_{h_n,t}(t+\Delta_n)|\boldsymbol{t}) \\
&= nf(t)\Delta_n Var(\hat{\beta}_{0,t} + \hat{\beta}_{1,t}\Delta_n|\boldsymbol{t}) \\
&= nf(t)\Delta_n Var\left\{(1,\delta)\begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}\hat{\beta}_t|\boldsymbol{t}\right\} \\
&= nf(t)\Delta_n (1,\delta) Var\left\{\begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}\hat{\beta}_t|\boldsymbol{t}\right\}\begin{pmatrix} 1 \\ \delta \end{pmatrix} \\
&= \delta(1,\delta)f(t)\left(nh_n Var\left\{\begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}\hat{\beta}_t|\boldsymbol{t}\right\}\right)\begin{pmatrix} 1 \\ \delta \end{pmatrix}.
\end{aligned}
\tag{3.10}
$$

By result (3.3) of Theorem 3.1, as $n \to \infty$ the last expression converges in probability to:

$$\delta(1,\delta)f(t)\frac{\sigma^2}{f(t)}\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1}\begin{pmatrix} u_0^* & u_1^* \\ u_1^* & u_2^* \end{pmatrix}\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1}\begin{pmatrix} 1 \\ \delta \end{pmatrix}$$

$$= \sigma^2\delta(1,\delta)\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1}\begin{pmatrix} u_0^* & u_1^* \\ u_1^* & u_2^* \end{pmatrix}\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1}\begin{pmatrix} 1 \\ \delta \end{pmatrix}$$

uniformly in t. □

Since f and m'' are uniformly continuous on $[0,1]$, Corollary 3.2 follows immediately from Corollary 3.1.

Corollary 3.2 *Assume the same conditions as in Corollary 3.1. As $n \to \infty$,*

$$\frac{2}{\Delta_n^2}Bias(\hat{m}_{h_n,t}(t+\Delta_n)|\boldsymbol{t}) - m''(1)\left((\frac{u_2^2 - u_1u_3}{\delta^2} + \frac{u_0u_3 - u_1u_2}{\delta})/(u_0u_2 - u_1^2) - 1 \right) = o_p(1), \tag{3.11}$$

$$nf(1)\Delta_n Var(\hat{m}_{h_n,t}(t+\Delta_n)|\boldsymbol{t}) - \sigma^2\delta(1,\delta)\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1}\begin{pmatrix} u_0^* & u_1^* \\ u_1^* & u_2^* \end{pmatrix}\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1}\begin{pmatrix} 1 \\ \delta \end{pmatrix} = o_p(1) \tag{3.12}$$

uniformly in $t \in [1-\rho_n, 1]$ with any sequence $\rho_n \geq 0$, $\rho_n \to 0$.

Remarks:

1. As is usually done in regression problems, we consider the conditional bias and variance of our estimates.

2. Corollary 3.2 says that the dominant terms in both the asymptotic conditional bias and the variance of $\hat{m}_{h_n,t}(t+\Delta_n)$ are equal to those of $\hat{m}_{h_n,1}(1+\Delta_n)$ for $t \in [1-\rho_n, 1]$.

3. For a finite sample, Δ_n is given and h_n will be chosen to minimize the asymptotic conditional mean squared error ($AMSE$) of $\hat{m}_{h_n,1}(1+\Delta_n)$. Because of Condition $6'$, choosing h_n is equivalent to choosing $\delta = \Delta_n/h_n$. Although minimizing the actual finite sample conditional MSE would be best, its form is far too complicated.

4. Based on the results in Corollary 3.1 one can write out the $AMSE$ of $\hat{m}_{h_n,1}(1+\Delta_n)$ as the square of the asymptotic bias plus the asymptotic variance. Condition $6'$ sets a guideline on the magnitude of Δ_n, that is, on how far ahead one can forecast if the rate $n^{-4/5}$ is to be achieved for the $AMSE$ of $\hat{m}_{h_n,1}(1+\Delta_n)$. Stone [27] has shown that under appropriate regularity conditions, the optimal rate of $AMSE$ for a nonparametric estimator of a twice differentiable function is $n^{-4/5}$. If Δ_n grows faster (but still slowly enough so that $\Delta_n \to 0$), for example, $\Delta_n/h_n \to \infty$, a rate slower than $n^{-4/5}$ can be achieved under appropriate conditions. This can be seen by the reasoning below.

   For $\Delta_n \to 0$, the bias and variance of $\hat{m}_{h_n,1}(1+\Delta_n)$ are

$$\begin{aligned} Bias(\hat{m}_{h_n,1}(1+\Delta_n)|\boldsymbol{t}) &= E\left(\hat{m}_{h_n,1}(1+\Delta_n) - m(1+\Delta_n)|\boldsymbol{t}\right) \\ &= E\left(\hat{\beta}_{0,1} - m(1) + \Delta_n[\hat{\beta}_{1,1} - m'(1)]|\boldsymbol{t}\right) + O(\Delta_n^2), \\ Var(\hat{m}_{h_n,1}(1+\Delta_n)|\boldsymbol{t}) &= Var(\hat{\beta}_{0,1} + \hat{\beta}_{1,1}\Delta_n|\boldsymbol{t}) \\ &= Var(\hat{\beta}_{0,1}|\boldsymbol{t}) + 2\Delta_n Cov(\hat{\beta}_{0,1}, \hat{\beta}_{1,1}|\boldsymbol{t}) + \Delta_n^2 Var(\hat{\beta}_{1,1}|\boldsymbol{t}). \end{aligned}$$

   By the results in Theorem 3.1, for $t = 1$ and any $\Delta_n > 0$ with $\Delta_n \to 0$

$$\begin{aligned} Bias(\hat{m}_{h_n,1}(1+\Delta_n)|\boldsymbol{t}) &\sim O_p(h_n^2) + O_p(h_n\Delta_n) + O(\Delta_n^2), \\ Var(\hat{m}_{h_n,1}(1+\Delta_n)|\boldsymbol{t}) &\sim O_p(1/(nh_n)) + O_p(\Delta_n/(nh_n^2)) + O_p(\Delta_n^2/(nh_n^3)) \\ &= O_p(1/(nh_n))\left(1 + O_p(\Delta_n/h_n) + O_p(\Delta_n^2/h_n^2)\right). \end{aligned}$$

If $\Delta_n/h_n \to \infty$, we have

$$\begin{aligned} Bias(\hat{m}_{h_n,1}(1+\Delta_n)|\boldsymbol{t}) &\sim O_p(\Delta_n^2), \\ Var(\hat{m}_{h_n,1}(1+\Delta_n)|\boldsymbol{t}) &\sim O_p(\Delta_n^2/(nh_n^3)). \end{aligned}$$

To minimize the $AMSE$, we need to make $Bias^2$ and Var of the same magnitude, i.e., $\Delta_n^4 \propto \Delta_n^2/(nh_n^3)$, or $\Delta_n^2 h_n^3 \propto 1/n$. If $\Delta_n/h_n \to \infty$, $\Delta_n^2 h_n^3 \propto 1/n$ implies $nh_n^5 \to 0$ and $n{\Delta_n}^5 \to \infty$. As a result, if $\Delta_n/h_n \to \infty$ under the conditions $nh_n^3 \to \infty$ and $\Delta_n^2 h_n^3 \propto 1/n$, the $AMSE$ of $\hat{m}_{h_n,1}(1+\Delta_n)$ achieves the rate $O_p(\Delta_n^4)$ (equivalently $O_p(\Delta_n^2/(nh_n^3))$, which is slower than $n^{-4/5}$ since $n\Delta_n^5 \to \infty$.

5. Recall that in Chapter 1.2.2, the algebraic substitution (1.11) is used to re-center the data. This recentering yields a simple dependence of $AMSE$ on δ, thus making it easier to find the optimal bandwidth by finding the optimal δ. If the original formulation (1.9) is used with the data centered at $1+\Delta_n$, then the estimate of $m(1+\Delta_n)$ is $\hat{\beta}_{0,1+\Delta_n}$. The formulae of the asymptotic bias and variance of $\hat{\beta}_{0,1+\Delta_n}$ will show clearly the advantage of centering data at 1 over centering at $1+\Delta_n$. Under the same assumptions as in Corollary 3.1 but with the additional assumption that $D < H$, one can show that the asymptotic bias and variance satisfy

$$\begin{aligned} &\frac{2}{h_n^2}\left[Bias(\hat{m}_{h_n,1}(1+\Delta_n)|\boldsymbol{t}) - m''(1+\Delta_n)\frac{v_2^2 - v_1 v_3}{v_0 v_2 - v_1^2}\right] \\ &\equiv \frac{2}{h_n^2}\left[E(\hat{\beta}_{0,1+\Delta_n} - m(1+\Delta_n)|\boldsymbol{t}) - m''(1+\Delta_n)\frac{v_2^2 - v_1 v_3}{v_0 v_2 - v_1^2}\right] \\ &= o_p(1), \end{aligned} \tag{3.13}$$

$$\begin{aligned} &nh_n f(1+\Delta_n) Var(\hat{m}_{h_n,1}(1+\Delta_n)|\boldsymbol{t}) - \\ &\begin{pmatrix} v_0 & v_1 \\ v_1 & v_2 \end{pmatrix}^{-1} \begin{pmatrix} v_0^* & v_1^* \\ v_1^* & v_2^* \end{pmatrix} \begin{pmatrix} v_0 & v_1 \\ v_1 & v_2 \end{pmatrix}^{-1} = o_p(1) \end{aligned} \tag{3.14}$$

where $v_i = \int_{-1}^{-\delta} v^i K(v)dv$, $v_i^* = \int_{-1}^{-\delta} v^i K(v)^2 dv$. Minimization of the resulting $AMSE$ of $\hat{\beta}_{0,1+\Delta_n}$ is computationally intensive since δ in the upper limit of the integrals depends on h_n. Therefore if $AMSE(h_n)$ is to be minimized over a grid of values of h_n, all the integrals have to be computed for each h_n. If K is, say, the indicator function on $[-1,0]$, these integrals can be found in closed form. If, however, K is a truncated normal density, then the intregrals must be evaluated numerically.

There is another advantage of re-centering the data: after re-centering the data at 1, h_n is the "real" bandwidth used in forecasting. If the data were instead centered at $1+\Delta_n$, the effective bandwidth would be $h_n - \Delta_n$ since there are no data beyond 1.

Lemma 3.1 below is used in the proof of Theorem 3.1.

Lemma 3.1 *Suppose that $g(\cdot)$ and $f(\cdot)$ are bounded over $[0,1]$ and $W(\cdot)$ is square integrable over $\mathcal{R}$ and that the t_is are iid with density $f(\cdot)$. If $nh_n \equiv nh \to \infty$, then as $n \to \infty$,*

$$\frac{1}{nh}\sum_{i=1}^{n} W(\frac{t_i - t}{h})g(t_i) - \frac{1}{h}\int_0^1 W(\frac{u-t}{h})g(u)f(u)du = O_p((nh)^{-1/2}) \tag{3.15}$$

uniformly for $t \in [0,1]$.

Proof: Let $Z = (nh)^{-1}\sum_{i=1}^{n} W((t_i - t)/h)g(t_i)$.

By Chebyshev's inequality, $Z = E(Z) + O_p((Var(Z))^{1/2})$.

$$\begin{aligned} E(Z) &= \frac{1}{h}E\left(W(\frac{t_1 - t}{h})g(t_1)\right) \\ &= \frac{1}{h}\int_0^1 W(\frac{u-t}{h})g(u)f(u)du, \\ Var(Z) &= \frac{1}{nh^2}Var\left(W(\frac{t_1 - t}{h})g(t_1)\right) \\ &\leq \frac{1}{nh^2}E\left(W^2(\frac{t_1 - t}{h})g^2(t_1)\right) \\ &= \frac{1}{nh}\int_0^1 W^2(\frac{u-t}{h})g^2(u)f(u)du/h. \end{aligned} \tag{3.16}$$

Since g and f are bounded, and $\int_0^1 W^2((u-t)/h)du/h = \int_{-t/h}^{(1-t)/h} W^2(s)ds < \infty$,

$$Var(Z) \quad = \quad O(\frac{1}{nh}), \tag{3.17}$$

which does not depend on t. □

Proof of Theorem 3.1: Write

$$\sum_1^n (Y_i - \beta_0 - \beta_1(t_i - t))^2 K(\frac{t_i - t}{h_n}) = (\boldsymbol{Y} - \Lambda\beta)^t \boldsymbol{W}(\boldsymbol{Y} - \Lambda\beta), \tag{3.18}$$

where

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \Lambda = \begin{pmatrix} 1 & t_1 - t \\ \vdots & \vdots \\ 1 & t_n - t \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \text{ and}$$

$$\boldsymbol{W} = diag(K(\frac{t_i - t}{h_n})).$$

The minimizer of (3.18) is:

$$\hat{\beta}_t = (\Lambda^t \boldsymbol{W} \Lambda)^{-1} \Lambda^t \boldsymbol{W} \boldsymbol{Y}, \tag{3.19}$$

provided that $\Lambda^t \boldsymbol{W} \Lambda$ is invertible and positive-definite. So for $\boldsymbol{t}$ with $\Lambda^t \boldsymbol{W} \Lambda$ invertible,

$$E(\hat{\beta}_t | \boldsymbol{t}) = (\Lambda^t \boldsymbol{W} \Lambda)^{-1} \Lambda^t \boldsymbol{W} \boldsymbol{m} \tag{3.20}$$

where $\boldsymbol{m} = (m(t_1), \ldots, m(t_n))^t$, and

$$Var(\hat{\beta}_t | \boldsymbol{t}) = \sigma^2 (\Lambda^t \boldsymbol{W} \Lambda)^{-1} \Lambda^t \boldsymbol{W}^2 \Lambda (\Lambda^t \boldsymbol{W} \Lambda)^{-1}. \tag{3.21}$$

Result (3.28) below shows that, when suitably normalized, $\Lambda^t \boldsymbol{W} \Lambda$ converges in probability to a positive definite matrix. Since we are interested in convergence in probability of $E(\hat{\beta}_t | \boldsymbol{t})$ and $Var(\hat{\beta}_t | \boldsymbol{t})$, it suffices to analyze (3.20) and (3.21).

The asymptotic bias will be derived first.

The Taylor expansion of m at t yields:

$$m(t_i) = m(t) + m'(t)(t_i - t) + \frac{m''(t)}{2}(t_i - t)^2 + o((t_i - t)^2)$$

uniformly in t_i with $|t_i - t| \leq h_n$. Only these t_is are considered because the support of K is $[-1, 0]$.

Therefore, for $|t_i - t| \leq h_n$

$$\begin{aligned} m(t_i) &= (1, t_i - t)\begin{pmatrix} m(t) \\ m'(t) \end{pmatrix} + \frac{m''(t)}{2} q_i {h_n}^2 + h_n^2 o(q_i) \\ &= \Lambda_i \begin{pmatrix} m(t) \\ m'(t) \end{pmatrix} + q_i h_n^2 \left(m''(t)/2 + o(1)\right) \qquad (3.22) \end{aligned}$$

where Λ_i is the ith row of Λ and

$$\boldsymbol{q} = (q_1, q_2, \ldots, q_n)^t \equiv \left(\left(\frac{t_1 - t}{h_n}\right)^2, \ldots, \left(\frac{t_n - t}{h_n}\right)^2\right)^t.$$

Since $w_i \equiv K((t_i - t)/h_n) = 0$ for $|t_i - t| > h_n$, so

$$\begin{aligned} &\begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}\left(E(\hat{\beta}_t|\boldsymbol{t}) - \begin{pmatrix} m(t) \\ m'(t) \end{pmatrix}\right) \\ &= \begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}(\Lambda^t \boldsymbol{W}\Lambda)^{-1}\Lambda^t \boldsymbol{W}\boldsymbol{q}{h_n}^2(m''(t)/2 + o(1)) \\ &= {h_n}^2 \begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}\left(\frac{1}{nh_n}\Lambda^t \boldsymbol{W}\Lambda\right)^{-1}\begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix} \times \\ &\qquad \begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}^{-1}\left(\frac{1}{nh_n}\Lambda^t \boldsymbol{W}\boldsymbol{q}\right)(m''(t)/2 + o(1)). \qquad (3.23) \end{aligned}$$

For $i, j = 1, 2$

$$\begin{aligned}
&\left(\begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}^{-1} \frac{1}{nh_n}\Lambda^t \boldsymbol{W} \Lambda \begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}^{-1}\right)_{ij} \\
&= \frac{1}{h_n^{i+j-2}} \frac{1}{nh_n} (\Lambda^t \boldsymbol{W} \Lambda)_{ij} \\
&= \frac{1}{h_n^{i+j-2}} \frac{1}{nh_n} \sum_{s=1}^{n} K\left(\frac{t_s - t}{h_n}\right)(t_s - t)^{i+j-2} \\
&= \frac{1}{nh_n} \sum_{s=1}^{n} K\left(\frac{t_s - t}{h_n}\right)\left(\frac{t_s - t}{h_n}\right)^{i+j-2} \qquad (3.24)
\end{aligned}$$

which by Lemma 3.1, is equal to

$$\frac{1}{h_n}\int_0^1 K\left(\frac{u-t}{h_n}\right)\left(\frac{u-t}{h_n}\right)^{i+j-2} f(u)du + O_p((nh_n)^{-1/2}).$$

Substituting $s = (u - t)/h_n$ yields

$$\int_{-t/h_n}^{(1-t)/h_n} K(s)s^{i+j-2} f(t + sh_n)ds + O_p((nh_n)^{-1/2}). \qquad (3.25)$$

For $t \in [a_n, 1]$ with $\liminf a_n/h_n > 1$, $[-t/h_n, (1 - t)/h_n] \cap [-1, 0] = [-1, 0]$ for n sufficiently large. So for $n \to \infty$ and $t \in [a_n, 1]$, expression (3.25) is equal to

$$\int_{-1}^{0} s^{i+j-2} K(s) f(t + sh_n)ds + O_p((nh_n)^{-1/2}), \qquad (3.26)$$

where $O_p((nh_n)^{-1/2})$ holds uniformly in $t \in [a_n, 1]$. The last expression converges (not only in probability) to

$$\int_{-1}^{0} s^{i+j-2} K(s) f(t)ds = f(t)u_{i+j-2}$$

uniformly in $t \in [a_n, 1]$ because

$$\begin{aligned}
&|\int_{-1}^{0} s^{i+j-2} K(s)\{f(t + sh_n) - f(t)\}ds| \\
&\leq \sup_{t\in[a_n,1]} |f(t + sh_n) - f(t)| \int_{-1}^{0} |s^{i+j-2} K(s)|ds \\
&\leq \sup_{t\in[a_n,1]} |f(t + sh_n) - f(t)| (\int_{-1}^{0} |s^{2(i+j-2)} ds)^{1/2} (\int_{-1}^{0} K(s)^2 ds)^{1/2} \\
&\to 0. \qquad (3.27)
\end{aligned}$$

So

$$\begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}^{-1} \frac{1}{nh_n}\Lambda^t \boldsymbol{W} \Lambda \begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}^{-1} \xrightarrow{p} f(t) \begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}, \tag{3.28}$$

uniformly in $t \in [a_n, 1]$, as $n \to \infty$.

Similarly, for $i = 1, 2$,

$$\begin{aligned} &\left(\begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}^{-1} \frac{1}{nh_n}\Lambda^t \boldsymbol{W} \boldsymbol{q} \right)_i \\ &= \frac{1}{nh_n} \sum_{s=1}^{n} K\left(\frac{t_s - t}{h_n}\right) \left(\frac{t_s - t}{h_n}\right)^{i+1} \\ &= \frac{1}{h_n} \int_0^1 K\left(\frac{u - t}{h_n}\right) \left(\frac{u - t}{h_n}\right)^{i+1} f(u)du + O_p((nh_n)^{-1/2}), \end{aligned}$$

uniformly in $t \in [a_n, 1]$. This expression converges to $f(t)u_{i+1}$ uniformly in $t \in [a_n, 1]$, as $n \to \infty$.

Thus

$$\begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}^{-1} \frac{1}{nh_n}\Lambda^t \boldsymbol{W} \boldsymbol{q} \xrightarrow{p} f(t) \begin{pmatrix} u_2 \\ u_3 \end{pmatrix}. \tag{3.29}$$

By (3.23), (3.28) and (3.29), it follows that

$$\begin{aligned} \frac{2}{h_n^2} \begin{pmatrix} E(\hat{\beta}_{0,t}|\boldsymbol{t}) - m(t) \\ h_n \left(E(\hat{\beta}_{1,t}|\boldsymbol{t}) - m'(t)\right) \end{pmatrix} &\xrightarrow{p} m''(t) \left(f(t) \begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix} \right)^{-1} \left(f(t) \begin{pmatrix} u_2 \\ u_3 \end{pmatrix} \right) \\ &= m''(t) \begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1} \begin{pmatrix} u_2 \\ u_3 \end{pmatrix} \end{aligned} \tag{3.30}$$

uniformly in $t \in [a_n, 1]$.

To calculate the asymptotic variance, first consider,

$$\begin{aligned}
& nh_n Var\left(\begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}\begin{pmatrix} \hat{\beta}_{0,t} \\ \hat{\beta}_{1,t} \end{pmatrix} |\boldsymbol{t}\right) \\
&= nh_n\sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix} (\Lambda^t\boldsymbol{W}\Lambda)^{-1}\Lambda^t\boldsymbol{W}^2\Lambda(\Lambda^t\boldsymbol{W}\Lambda)^{-1}\begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix} \\
&= \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}\left(\frac{\Lambda^t\boldsymbol{W}\Lambda}{nh_n}\right)^{-1}\begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix} \times \\
& \begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}^{-1} \frac{1}{nh_n}\Lambda^t\boldsymbol{W}^2\Lambda \begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}^{-1} \times \\
& \begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}\left(\frac{\Lambda^t\boldsymbol{W}\Lambda}{nh_n}\right)^{-1}\begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}.
\end{aligned} \tag{3.31}$$

By calculations similar to those in the proof of (3.28),

$$\begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}^{-1} \frac{1}{nh_n}\Lambda^t\boldsymbol{W}^2\Lambda \begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}^{-1} \xrightarrow{p} f(t)\begin{pmatrix} u_0^* & u_1^* \\ u_1^* & u_2^* \end{pmatrix}. \tag{3.32}$$

Using results (3.28) and (3.32) in (3.31) gives

$$\begin{aligned}
& (nh_n)Var\left(\begin{pmatrix} 1 & 0 \\ 0 & h_n \end{pmatrix}\begin{pmatrix} \hat{\beta}_{0,t} \\ \hat{\beta}_{1,t} \end{pmatrix} |\boldsymbol{t}\right) = \\
& \sigma^2 \left(f(t)\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}\right)^{-1}\left(f(t)\begin{pmatrix} u_0^* & u_1^* \\ u_1^* & u_2^* \end{pmatrix}\right)\left(f(t)\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}\right)^{-1} + o_p(1) \\
&= \frac{\sigma^2}{f(t)}\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1}\begin{pmatrix} u_0^* & u_1^* \\ u_1^* & u_2^* \end{pmatrix}\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1} + o_p(1),
\end{aligned} \tag{3.33}$$

uniformly for $t \in [a_n, 1]$. $\square$.

Under appropriate conditions, Theorem 3.1 and its corollaries hold for the fixed design where t_is are pre-specified design points rather than random variables. The analogous theorem and corollaries for the fixed design are presented below.

Theorem 3.2 *For the fixed design case, assume the same conditions on m, K, h_n: 1, 4, 5 and 6 as in the random design. In addition assume that K' is continuous, that $t_i \equiv t_i^n$ with $0 < t_1 < \ldots < t_n = 1$ and*

$$\exists\, f : \int_0^{t_i} f(t)dt = i/n, \ \inf_{t\in[0,1]} f(t) > 0 \tag{3.34}$$

with f continuous. Then as $n \to \infty$,

$$\frac{2}{h_n^2}\begin{pmatrix} E(\hat\beta_{0,t}) - m(t) \\ h_n\left(E(\hat\beta_{1,t}) - m'(t)\right)\end{pmatrix} - m''(t)\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2\end{pmatrix}^{-1}\begin{pmatrix} u_2 \\ u_3\end{pmatrix} = o(1)\boldsymbol{e}, \tag{3.35}$$

$$nh_nVar\left(\begin{pmatrix} 1 & 0 \\ 0 & h_n\end{pmatrix}\begin{pmatrix} \hat\beta_{0,t} \\ \hat\beta_{1,t}\end{pmatrix}\right) -$$

$$\frac{\sigma^2}{f(t)}\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2\end{pmatrix}^{-1}\begin{pmatrix} u_0^* & u_1^* \\ u_1^* & u_2^*\end{pmatrix}\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2\end{pmatrix}^{-1} = o(1)\boldsymbol{e}, \tag{3.36}$$

uniformly in $t \in [a_n, 1]$ with $\liminf a_n/h_n > 1$.

The condition on the design points t_is in the theorem above is analogous to Condition 3 for the random design case. This condition serves as a guideline to the selection of the design points in an experiment.

Corollary 3.3 *Assume the same conditions as those in Theorem 3.2 except with Condition 6 replaced by Condition 6′ of Corollary 3.1. Then as $n \to \infty$,*

$$\sup_{t\in[a_n,1]}\left|\frac{2}{\Delta_n^2}Bias(\hat m_{h_n,t}(t+\Delta_n)) - \right.$$

$$m''(t)\left((\frac{u_2^2 - u_1 u_3}{\delta^2} + \frac{u_0 u_3 - u_1 u_2}{\delta})/(u_0 u_2 - u_1^2) - 1 \right)\Bigg| = o(1), \qquad (3.37)$$

and

$$\sup_{t\in[a_n,1]} \Bigg| nf(t)\Delta_n Var(\hat{m}_{h_n,t}(t+\Delta_n)) - \sigma^2\delta(1,\delta)\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1}$$

$$\cdot \begin{pmatrix} u_0^* & u_1^* \\ u_1^* & u_2^* \end{pmatrix}\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1}\begin{pmatrix} 1 \\ \delta \end{pmatrix}\Bigg| = o(1). \quad (3.38)$$

Corollary 3.4 *Assume the same conditions as in Corollary 3.3. As $n \to \infty$,*

$$\sup_{t\in[1-\rho_n,1]} \Bigg| \frac{2}{\Delta_n^2} Bias(\hat{m}_{h_n,t}(t+\Delta_n)) -$$

$$m''(1)\left((\frac{u_2^2 - u_1 u_3}{\delta^2} + \frac{u_0 u_3 - u_1 u_2}{\delta})/(u_0 u_2 - u_1^2) - 1 \right)\Bigg| = o(1), \qquad (3.39)$$

$$\sup_{t\in[1-\rho_n,1]} \Bigg| nf(1)\Delta_n Var(\hat{m}_{h_n,t}(t+\Delta_n)) - \sigma^2\delta(1,\delta)\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1}$$

$$\cdot \begin{pmatrix} u_0^* & u_1^* \\ u_1^* & u_2^* \end{pmatrix}\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1}\begin{pmatrix} 1 \\ \delta \end{pmatrix}\Bigg| = o(1). \quad (3.40)$$

The proof of Theorem 3.2 for the fixed design case is along the same lines as for the random design case except that the following lemma is used instead of Lemma 3.1 when replacing a sum by an integral. The proofs of the corollaries of Theorem 3.2 are the same as those of the corollaries of Theorem 3.1.

Lemma 3.2 *Assume f, W' and g' continuous on $[0,1]$ and f satisfies (3.34). Then $\exists\ C > 0$, such that*

$$\sup_{t\in[0,1]} \left| \frac{1}{nh}\sum_{i=1}^{n} W(\frac{t_i - t}{h})g(t_i) - \frac{1}{h}\int_0^1 W(\frac{u-t}{h})g(u)f(u)du \right| \le \frac{C}{nh^2}. \qquad (3.41)$$

Remark: Note that the error term $O((nh^2)^{-1})$ in Lemma 3.2 is smaller than the error term $O((nh)^{-1/2})$ in Lemma 3.1 for h of order $n^{-1/5}$.

The proof of Lemma 3.2 requires the result of Lemma 3.3.

Lemma 3.3 *If* $\inf_{t\in[0,1]} f(t) > 0$*, and if* $0 \equiv t_0 < t_1 < \ldots < t_n = 1$*, then* $\exists\, C_2 > 0$ *such that* $\sup_{i,n} |t_i - t_{i-1}| \le C_2/n$.

Proof: Let $C_2 = 1/\inf_{t\in[0,1]} f(t)$. By the assumption on f,

$$\frac{1}{n} = \int_{t_{i-1}}^{t_i} f(u)du \ge \int_{t_{i-1}}^{t_i} \inf_{t\in[0,1]} f(u)du = C_2^{-1}(t_i - t_{i-1}).$$

Thus $t_i - t_{i-1} \le C_2/n,\ \forall i, n.$ □

Proof of Lemma 3.2:

$$\begin{aligned}
&\frac{1}{h}\int_0^1 W(\frac{u-t}{h})g(u)f(u)du = \frac{1}{h}\sum_{i=1}^{n}\int_{t_{i-1}}^{t_i} W(\frac{u-t}{h})(gf)(u)du \\
&= \frac{1}{h}\sum_{i=1}^{n}\int_{t_{i-1}}^{t_i} W(\frac{t_i-t}{h})g(t_i)f(u)du \\
&+ \frac{1}{h}\sum_{i=1}^{n}\int_{t_{i-1}}^{t_i}\left[W(\frac{u-t}{h})g(u) - W(\frac{t_i-t}{h})g(t_i)\right] f(u)du \\
&\equiv A + B. \qquad (3.42)
\end{aligned}$$

Consider the first term in (3.42).

$$\begin{aligned}
A &= \frac{1}{h}\sum_{i=1}^{n} W(\frac{t_i-t}{h})g(t_i)\int_{t_{i-1}}^{t_i} f(u)du \\
&= \frac{1}{nh}\sum_{1}^{n} W(\frac{t_i-t}{h})g(t_i) \qquad (3.43)
\end{aligned}$$

by assumption (3.34).

The lemma will be proved if $|B| \le C/(nh^2)$ for some $C > 0$.

Since f, W' and g' are continuous over $[0,1]$, there exists $C_1 > 0$, such that

$$\left|W(\frac{u-t}{h})g(u) - W(\frac{t_i-t}{h})g(t_i)\right| f(u) \le C_1\left|\frac{u-t_i}{h}\right|. \qquad (3.44)$$

Thus

$$\begin{aligned}
|B| &\leq \frac{1}{h}\sum_{1}^{n}\int_{t_{i-1}}^{t_i} C_1 \left|\frac{u-t_i}{h}\right| du \\
&\leq \frac{1}{h}\sum_{1}^{n}\int_{t_{i-1}}^{t_i} C_1 \left|\frac{t_{i-1}-t_i}{h}\right| du \\
&= \frac{C_1}{h^2}\sum_{1}^{n}(t_i-t_{i-1})^2 \\
&\leq \frac{C_1}{h^2}\sum_{1}^{n}(\frac{C_2}{n})^2 \\
&= \frac{C_1C_2^2}{nh^2} \equiv \frac{C}{nh^2}. \qquad \Box \qquad (3.45)
\end{aligned}$$

The results in this chapter will be used in the next chapter for the estimation of an optimal bandwidth for forecasting.

Chapter 4

The choice of a smoothing parameter

The plug-in approach and the cross-validation approach are commonly used in choosing a bandwidth for a local regression estimator in the non-forecasting setting. Each approach seeks to minimize some measure of the discrepancy between the estimated and the true function and tries to estimate the bandwidth at which the minimum of such a measure occurs. There is abundant literature on the merits and limitations of both methods. For a comprehensive review, see [7], [12], [13] and [19]. The ideas of choosing a bandwidth for forecasting by both approaches will be developed in this chapter. The applicability of these approaches relies upon the fact that the data are independent. Any correlation structure in the data will affect any automatic selection of bandwidth (see, e.g. [16]).

4.1 The plug-in approach

A plug-in bandwidth is an estimate of an "optimal bandwidth" in a certain sense. To estimate the function m at t by $\hat{m}_{h_n}(t)$ (defined in 1.8), an optimal bandwidth $h_{opt}(t)$ may be defined to be the one that minimizes the mean squared error (MSE) of $\hat{m}_{h_n}(t)$:

$$h_{opt}(t) \equiv argmin_{h_n} MSE(t, h_n),$$

where $MSE(t, h_n) \equiv E(\hat{m}_{h_n}(t) - m(t)|\boldsymbol{t})^2$ in the random design and $MSE(t, h_n) \equiv E(\hat{m}_{h_n}(t) - m(t))^2$ in the fixed design. Since the idea is the same for both designs, the selection of an optimal bandwidth will be described for the random design. The bandwidth $h_{opt}(t)$ is called a local bandwidth since $MSE(t, h_n)$ is a criterion of goodness of fit at a single point t. However, if m is believed to be reasonably smooth, a constant bandwidth for all t will suffice ([7]). In such a case, an optimal bandwidth h^G_{opt} may be defined to be the one that minimizes the sum of the mean squared errors:

$$h^G_{opt} \equiv argmin_{h_n} MSE_G(h_n)$$

where $MSE_G(h_n) \equiv n^{-1} \sum_{i=1}^n E\left((\hat{m}_{h_n}(t_i) - m(t_i))^2|\boldsymbol{t}\right)$. Since h^G_{opt} will be used to estimate $m(t)$ for all t, it is called a global bandwidth. Another commonly used criterion for a global bandwidth is the integrated mean squared error,

$$\begin{aligned} MISE(h_n) &\equiv \int_0^1 MSE(t, h_n)dt \\ &= \int_0^1 E\left((\hat{m}_{h_n}(t) - m(t))^2|\boldsymbol{t}\right) dt; \qquad (4.1) \end{aligned}$$

a global bandwidth can be obtained by minimizing $MISE(h_n)$. $MSE_G(h_n)$ may be viewed as a discretized version of $MISE(h_n)$ when the t_is are equally spaced over $[0, 1]$.

Note that all three criteria mentioned above involve the unknown function m and the data $\{(t_i, Y_i)\}_1^n$. So the optimal bandwidth depends on m and σ^2 and thus has to be estimated. The plug-in approach estimates the optimal bandwidth by minimizing an estimate of the asymptotic expression for MSE or $MISE$.

The criterion $MSE(t, h_n)$ will be used hereafter. The context of the plug-in approach that of non-forecasting and some relevant results will be presented first and then the generalization of the plug-in approach to the forecasting setting will be discussed.

4.1.1 Introduction

In this section suppose $Y_i = m(t_i) + \epsilon_i$, where the ϵ_is are independent with mean 0 and variance $\sigma^2(t_i)$. When the errors are homoscedastic, $\sigma^2(t_i) \equiv \sigma^2$. Results for heteroscedastic errors in the non-forecasting problem will be presented because they encompass the special case of homoscedastic errors. In particular, a technique by Fan and Gijbels [7] of estimating $m''(1)$ in the case of heteroscedastic errors will be described and used in forecasting.

In general, to estimate $m^{(\nu)}(t)/\nu! \equiv \beta_\nu$, $\nu = 0, 1, \ldots$, a polynomial of degree p ($p \geq \nu$) is fitted to the data with the following fitting criterion:

$$\hat{\beta}_t = argmin \sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p} \beta_j (t_i - t)^j)^2 K(\frac{t_i - t}{h_n}), \tag{4.2}$$

and ${\hat{m}_{h_n}}^{(\nu)}(t)$ is set to $\nu! \hat{\beta}_{\nu,t}$. In curve estimation, the kernel K in (4.2) is usually taken to be symmetric around 0 over either the reals or $[-1, 1]$. The performance of ${\hat{m}_{h_n}}^{(\nu)}(t)$ is assessed by its MSE: $E\left\{\left({\hat{m}_{h_n}}^{(\nu)}(t) - m^{(\nu)}(t)\right)^2 |\boldsymbol{t}\right\}$. According to the general results in [7], when $p - \nu$ is odd, the asymptotic MSE ($AMSE$) of ${\hat{m}_{h_n}}^{(\nu)}(t)$ is

$$\beta_{p+1}^2 b_\nu^2 h_n^{2(p+1-\nu)} + a_\nu \frac{\sigma^2(t)}{f(t) n h_n^{1+2\nu}}. \tag{4.3}$$

Here

$$\begin{aligned}
\boldsymbol{a} &\equiv (a_0, a_1, \ldots, a_p)^t = diag(S^{-1} S^* S^{-1}), & (4.4)\\
\boldsymbol{b} &\equiv (b_0, b_1, \ldots, b_p)^t = S^{-1}(s_{p+1}, \ldots, s_{2p+2})^t, & (4.5)\\
S &\equiv (s_{i,j}) \text{ with } s_{i,j} \equiv s_{i+j-2} = \int u^{i+j-2} K(u) du, & (4.6)\\
S^* &\equiv \left(s_{i,j}^*\right) \text{ with } s_{i,j}^* \equiv s_{i+j-2}^* = \int u^{i+j-2} K^2(u) du. & (4.7)
\end{aligned}$$

The first term in (4.3) is the square of the asymptotic bias which depends on $m^{(p+1)}(t) = (p+1)!\beta_{p+1}$, while the second term is the asymptotic variance which depends on $\sigma^2(t)$. For example, when $m(t)$ is estimated by a local linear estimator, i.e., $\nu = 0$ and $p = 1$, the square of the asymptotic bias of $\hat{m}_{h_n}(t)$ is $\beta_2^2 b_0^2 h_n^4$ (or $m''(t)^2 b_0^2 h_n^4/4$) and its asymptotic variance is $a_0\sigma^2(t)/(f(t)nh_n)$.

Formula (4.3) shows that if $m^{(p+1)}(t) \neq 0$, a large bandwidth h_n will create a large bias but a small variance, and a small bandwidth will yield a small bias but a large variance. Requiring both the bias and the variance to be small at the same time is setting conflicting goals for h_n. Therefore a tradeoff between the bias and the variance is needed and this is achieved by choosing h_n to minimize the asymptotic MSE of the estimator.

From (4.3), the bandwidth that minimizes the $AMSE$ of $\hat{m}_{h_n}{}^{(\nu)}(t)$ is:

$$h_{\nu,opt}(t) = \left(\frac{(2\nu+1)a_\nu\sigma^2(t)}{2(p+1-\nu)b_\nu^2\beta_{p+1}^2 nf(t)}\right)^{\frac{1}{2p+3}}. \tag{4.8}$$

For example, when m is to be estimated by a local line, i.e., $\nu = 0$ and $p = 1$, this locally optimal bandwidth is

$$h_{0,opt}(t) = \left(\frac{a_0\sigma^2(t)}{4b_0^2\beta_2^2 nf(t)}\right)^{\frac{1}{5}} = \left(\frac{a_0\sigma^2(t)}{b_0^2 m''(t)^2 nf(t)}\right)^{\frac{1}{5}} \propto n^{-1/5}. \tag{4.9}$$

Plugging this optimal bandwidth into the formula for the $AMSE$ of $\hat{m}_{h_n}(t)$, one can see that the rate of $n^{-4/5}$ is achieved for its $AMSE$.

Note that formula (4.8) depends on two unknowns,

$$\sigma^2(t) \text{ and } m^{(p+1)}(t) = \beta_{p+1}(p+1)!.$$

These unknowns need to be estimated and these estimates are then plugged into formula (4.8) to yield an estimate, $\hat{h}_{\nu,opt}(t)$, of the optimal bandwidth $h_{\nu,opt}(t)$. For example, if the curve m ($\nu = 0$) is being estimated at t, to get an optimal bandwidth one has to estimate $m''(t)$ and $\sigma^2(t)$ first.

4.1.2 Plug-in approach using complete asymptotics for forecasting ($CAMSE$)

We only consider the case of homoscedastic errors, i.e., $\sigma^2(t) \equiv \sigma^2$. In forecasting, the criterion that the plug-in approach considers is also the mean squared error:

$$MSE(h_n) \equiv MSE(1+\Delta_n, h_n) = E\left([\hat{m}_{h_n,1}(1+\Delta_n) - m(1+\Delta_n)]^2 | \boldsymbol{t}\right). \qquad (4.10)$$

Again, only the random design case will be presented since the ideas and the results are the same (under proper conditions) for both designs.

Under Condition 6′ of Corollary 3.3, for Δ_n given, the asymptotic $MSE(h_n)$ can be written in terms of $\delta = \Delta_n/h_n$, so the notation $AMSE(\delta)$ is used instead.

By Corollary 3.1 or 3.3, the $AMSE$ of $\hat{m}_{h_n,1}(1+\Delta_n)$ is

$$\begin{aligned} AMSE(\delta) &= \frac{m''(1)^2}{4}\Delta_n^4\left((\frac{u_2^2 - u_1u_3}{\delta^2} + \frac{u_0u_3 - u_1u_2}{\delta})/(u_0u_2 - u_1^2) - 1\right)^2 \\ &+ \frac{\sigma^2\delta}{nf(1)\Delta_n}(1,\delta)\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1}\begin{pmatrix} u_0^* & u_1^* \\ u_1^* & u_2^* \end{pmatrix}\begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1}\begin{pmatrix} 1 \\ \delta \end{pmatrix}. \end{aligned} \qquad (4.11)$$

If the first term of $AMSE(\delta)$ is a monotone and non-increasing function of δ, the above expression shows, in an asymptotic sense, that for a given Δ_n, a very large δ (corresponding to a very small h_n) will yield estimates with a small bias but a large variance and that a small δ (corresponding to a large h_n) will yield estimates with a small variance but a large bias. For a general kernel function K, the relationship between h_n and the asymptotic bias is not as obvious as it is in the non-forecasting setting. In the forecasting setting, a trade-off between the asymptotic bias and the asymptotic variance can be achieved by choosing the $h_n = \Delta_n/\delta$ with δ minimizing $AMSE(\delta)$.

Observe that as $\delta \to 0$ and $+\infty$, $AMSE(\delta) \to +\infty$. So a minimum of $AMSE(\delta)$

in $(0, +\infty)$ is guaranteed. The minimum can be found by setting the first derivative of $AMSE$ to zero and choosing the positive root which gives the smallest $AMSE$.

Whereas in curve estimation for the non-forecasting setting an explicit formula (4.8) for the optimal h_n exists, no such formula exists in forecasting. Finding the optimal h_n or equivalently δ for forecasting, calls for solving a seventh degree polynomial equation (the one resulting from setting the first derivative of $AMSE(\delta)$ to zero) for δ. An alternative method for minimizing $AMSE$ is by a grid search.

Note that like in curve estimation, estimating the optimal h_n in forecasting requires the estimation of m'' and σ^2. This issue will be discussed in Section 4.1.4.

4.1.3 The plug-in approach using the finite sample variance for forecasting ($HAMSE$)

Recall that the discrepancy measure (4.10), the MSE of the forecasting estimator, is comprised of two parts: a bias component and a variance component. The bias component involves the unknown function m but by (3.21) the variance component equals $\sigma^2(1, \Delta_n)(\Lambda^t \boldsymbol{W} \Lambda)^t \Lambda^t \boldsymbol{W}^2 \Lambda (\Lambda^t \boldsymbol{W} \Lambda)(1, \Delta_n)^t$, which does not involve m. The estimation of the variance parameter σ^2 is relatively easier than the estimation of m''. Therefore we can directly use the exact variance expression rather than using its asymptotic expression [7]. The asymptotic variance describes the variability when $n \to \infty$. Specifically, the following can be used in lieu of $AMSE$.

Definition 4.1 *The asymptotic MSE using the finite sample variance for the forecasting estimator* $\hat{m}_{h_n,1}(1 + \Delta_n)$ *is defined as:*

$$\begin{aligned} HAMSE(\delta) &= \frac{m''(1)^2}{4} \Delta_n^4 \left(\left(\frac{u_2^2 - u_1 u_3}{\delta^2} + \frac{u_0 u_3 - u_1 u_2}{\delta}\right) / (u_0 u_2 - u_1^2) - 1 \right)^2 \\ &\quad + \sigma^2 (1, \Delta_n)(\Lambda^t \boldsymbol{W} \Lambda)^t \Lambda^t \boldsymbol{W}^2 \Lambda (\Lambda^t \boldsymbol{W} \Lambda)(1, \Delta_n)^t. \qquad (4.12) \end{aligned}$$

Expression (4.12) uses only "half" of the asymptotic results of the bias and the variance of $\hat{m}_{h_n,1}(1+\Delta_n)$ in that $HAMSE$ is defined as the sum of the asymptotic bias and the finite sample variance of the forecasting estimator, hence the acronym "$HAMSE$". Recall that $\boldsymbol{W} = diag(K((t_i-1)/h_n)) = diag(K(\delta(t_i-1)/\Delta_n))$. The optimal δ, δ_{opt}, that minimizes $HAMSE(\cdot)$ can be found by a grid search and the optimal bandwidth is set to be: $h_{HAMSE,opt} = \Delta_n/\delta_{opt}$.

Note that both the plug-in approach using $AMSE$ and the plug-in approach using $HAMSE$ are based on the same discrepancy measure, the mean squared error of $\hat{m}_{h_n,1}(1+\Delta_n)$. The difference is that the latter approach uses the exact variance of $\hat{m}_{h_n,1}(1+\Delta_n)$ in the variance component instead of its asymptotic expression.

4.1.4 Estimation of m'' and σ^2 – an introduction

The plug-in approach in the non-forecasting setting calls for the estimation of the second derivative m'' of the unknown function and the variance function σ^2.

When homoscedastic errors are assumed, σ^2 can be estimated by the Rice estimator [25].

Definition 4.2 *The Rice estimator of the variance σ^2 is defined as:*

$$\hat{\sigma}^2_{Rice} = \sum_{i=1}^{n-1} \frac{(Y_{i+1}-Y_i)^2}{2(n-1)}. \tag{4.13}$$

Under certain conditions, the Rice estimator is $n^{1/2}$-consistent, which means $n^{1/2}(\hat{\sigma}^2_{Rice} - \sigma^2) \to 0$ in probability. Other estimators of σ^2 can be found in [11].

When heteroscedastic errors are assumed, Fan and Gijbels propose that $\sigma^2(t)$ can be estimated from residuals of the local polynomial regression [7]:

$$\hat{\sigma}^2(t) = \frac{1}{tr\{\boldsymbol{W}\} - tr\{(\Lambda_p^t \boldsymbol{W} \Lambda_p)^{-1} \Lambda_p^t \boldsymbol{W}^2 \Lambda_p\}} \sum_{i=1}^{n} (Y_i - \tilde{Y}_i)^2 K(\frac{t_i - t}{h_n}), \tag{4.14}$$

where

$$\Lambda_p = \begin{pmatrix} 1 & t_1 - t & \dots & (t_1 - t)^p \\ \vdots & \vdots & & \vdots \\ 1 & t_n - t & \dots & (t_n - t)^p \end{pmatrix}, \quad \boldsymbol{W} = diag(K((t_j - t)/h_n)), \tag{4.15}$$

p is the degree of the polynomial and $\tilde{\boldsymbol{Y}} = (\tilde{Y}_1, \dots, \tilde{Y}_n)^t = \Lambda_p \hat{\beta}_t$, $\hat{\beta}_t$ being the usual non-forecasting estimate as defined in (4.2). Clearly $\hat{\sigma}^2(t)$ can be used in the cases of either homoscedastic or heteroscedastic errors. Note that the estimation of $\sigma^2(t)$ requires again a choice of a bandwidth which may differ from the bandwidth for curve estimation.

The estimation of m'' requires a higher order method. Recall that $m^{(\nu)}$ is estimated by locally fitting a pth degree polynomial with $p \geq \nu$. Ruppert and Wand [26] recommend that $p - \nu$ should be taken to be odd to obtain an accurate estimate of $m^{(\nu)}(t)$. So $m''(t)$ should be estimated by locally fitting at least a cubic polynomial. Again the estimation of m'' requires the choice of an optimal bandwidth, which is different from (usually bigger than) the bandwidth that is used for estimating the function m itself.

To summarize the idea of the plug-in approach, to get an optimal h_n for curve estimation one needs to estimate σ^2 and m''. The estimation of σ^2 and m'' requires two additional bandwidths whose optimal values depend on higher derivatives of m. To avoid this spiralling argument, one has to come up with an initial bandwidth for estimating derivatives and σ^2. Therefore in the plug-in approach, the problem of choosing a good initial bandwidth has attracted special attention ([26],[12],[7]).

4.1.5 Estimation of m'' and σ^2 for forecasting

The discussion will be restricted to the homoscedastic errors case. Recall that the forecasting estimator is defined as $\hat{m}_{h_n,1}(1 + \Delta_n) = \hat{\beta}_{0,1} + \hat{\beta}_{1,1}\Delta_n$ where

$$\hat{\beta}_1 = argmin \sum_1^n (Y_i - \beta_0 - \beta_1(t_i - 1))^2 K((t_i - 1)/h_n).$$

The objective is to choose an $h_n = \Delta_n/\delta$ such that $AMSE(\delta)$ or $HAMSE(\delta)$ achieves a minimum.

Since both objective functions involve the unknowns σ^2 and $m''(1)$, these need to be estimated first. Existing techniques for estimating both in curve fitting can be applied directly to forecasting.

For estimating σ^2, $\hat{\sigma}^2_{Rice}$ suffices because of the assumption of homoscedastic errors.

Estimation of $m''(1)$ is a much more complicated matter. As discussed earlier, $m''(1)$ can be estimated by fitting a local cubic polynomial around $t = 1$. Let $\nu = 2$, $p = 3$ and $t = 1$ in (4.2), that is,

$$\text{let } \hat{\beta}_1 = argmin \sum_{i=1}^{n} (Y_i - \sum_{j=0}^{3} \beta_j (t_i - 1)^j)^2 K(\frac{t_i - 1}{h_n}), \tag{4.16}$$

and let $\hat{m}_{h_n,1}{}''(1) = 2\hat{\beta}_{1,1}$. Note that the kernel function K and the bandwidth h_n used to estimate $m''(1)$ can be different from those used for forecasting.

Formula (4.8) gives the optimal bandwidth for estimating $m''(1)$ as follows:

$$h_{2,opt}(1) = \left(\frac{5a_2\sigma^2}{4b_2^2\beta_4^2 n f(1)} \right)^{\frac{1}{9}}, \tag{4.17}$$

which depends on σ^2 and one higher unknown derivative $m^{(4)}(1)$ ($= 4!\beta_4$).

Ordinarily the fourth derivative of a function m is difficult to estimate, partly because the optimal bandwidth depends on higher order derivatives of m. However, an idea due to Fan and Gijbels [7] eliminates the need to estimate $m^{(4)}(1)$. To avoid estimating higher order derivatives when estimating $m^{(\nu)}$ by a pth degree polynomial, these authors have introduced a statistic RSC,

$$RSC(t, h_n) = \hat{\sigma}^2(t)\{1 + (p+1)V\}, \tag{4.18}$$

where $\hat{\sigma}^2(t)$ is as defined in (4.14) and V is the first diagonal element of the matrix $(\Lambda_p^t \boldsymbol{W} \Lambda_p)^{-1} \Lambda_p^t \boldsymbol{W}^2 \Lambda_p (\Lambda_p^t \boldsymbol{W} \Lambda_p)^{-1}$. The motivation for using RSC is that its minimum reflects to some extent a trade-off between the bias and variance of the fit.

Fan and Gijbels [7] have shown that the optimal bandwidth $h_o(t)$ that minimizes the asymptotic value of $E(RSC(t, h_n)|\boldsymbol{t})$ is:

$$h_o(t) = \left(\frac{a_0 \sigma^2(t)}{2C_p \beta_{p+1}^2 n f(t)}\right)^{\frac{1}{2p+3}}, \tag{4.19}$$

where

$$C_p = \frac{s_{2p+2} - (s_{p+1}, \ldots, s_{2p+1}) S^{-1} (s_{p+1}, \ldots, s_{2p+1})^t}{s_0}, \tag{4.20}$$

and s_j is as defined in (4.6). Comparing (4.8) and (4.19) yields:

$$h_{\nu,opt}(t) = \left(\frac{(2\nu+1)}{(p+1-\nu)} \frac{a_\nu}{a_0} \frac{C_p}{b_\nu^2}\right)^{\frac{1}{2p+3}} h_o(t). \tag{4.21}$$

Applying this relationship to the estimation of $m''(1)$, i.e., $t = 1$, $\nu = 2$, $p = 3$, leads to

$$h_{2,opt} \equiv h_{2,opt}(1) = \left(\frac{5a_2 C_3}{2a_0 b_2^2}\right)^{\frac{1}{9}} h_o(1). \tag{4.22}$$

The point is that $h_o(1)$ can be estimated from the sample by minimizing RSC, yielding an estimate of $h_{2,opt}$ since the scaling factor in (4.22) depends on the kernel function only. Thus the optimal bandwidth for estimating $m''(1)$ can be estimated from (4.22) using the information in the finite sample at hand.

The following algorithm summarizes the steps to get the optimal bandwidth for estimating $m''(1)$.

Algorithm 1 *:*

1. *Fit a local cubic polynomial centered at $t = 1$ as in (4.16) to the data for each value of h_n on a grid;*

2. *find $\hat{h}_o(1)$ that minimizes $RSC(1, h_n)$ on that grid of h_n;*

3. *get $\hat{h}_{2,opt}$ via (4.22);*

4. *estimate $m''(1)$ by fitting a local cubic polynomial to the data using (4.16) with bandwidth $\hat{h}_{2,opt}$.*

4.2 Cross-validation

4.2.1 Background to the non-forecasting setting

Cross-validation is a technique commonly used to choose a smoothing parameter. Most references in literature applies cross-validation in the context of curve fitting by splines or kernel estimators. However the idea of cross-validation is the same for both techniques and can be applied in other contexts, e.g., bandwidth selection in local linear regression. For a brief historical note on cross-validation, see [15] (pages 152-153).

This section will present the motivation and results on cross-validation in the setting of curve fitting. A generic symbol λ is used for the smoothing parameter. In the linear operator approach, λ will be the parameter used in the penalty term. In the local regression approach, λ will be h_n, the bandwidth of the kernel.

Let $\hat{m}_\lambda(\cdot)$ denote the fit to the data Y_is with smoothing parameter λ. Ideally, one would want to choose a λ such that the prediction error is minimized. A criterion reflecting this objective is

$$PE(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(Y_i^* - \hat{m}_\lambda(t_i))^2, \tag{4.23}$$

where the Y_i^*s are new observations made at the design points t_i. Thus the Y_i^*s are independent of the Y_is but have the same distribution as the Y_is.

To see the relationship to the criterion $MSE_G(\lambda)$ used in the plug-in approach, note that

$$\begin{aligned} E(PE(\lambda)|\boldsymbol{t}) &= \frac{1}{n}E\left(\sum_{i=1}^{n}(Y_i^* - m(t_i))^2 + 2\sum_{i=1}^{n}(Y_i^* - m(t_i))(m(t_i) - \hat{m}_\lambda(t_i)) \right. \\ &\qquad \left. + \sum_{i=1}^{n}(m(t_i) - \hat{m}_\lambda(t_i))^2|\boldsymbol{t}\right) \\ &= \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}E((m(t_i) - \hat{m}_\lambda(t_i))^2|\boldsymbol{t}) \\ &= \sigma^2 + MSE_G(\lambda). \end{aligned} \tag{4.24}$$

So minimizing the expected prediction error $E(PE(\lambda)|\boldsymbol{t})$ is equivalent to minimizing $MSE_G(\lambda)$.

Of course the Y_i^*s are unknown and so to implement this idea the Y_i^*s have to be replaced by observables. Simply minimizing (4.23) with the Y_i^*s replaced by the Y_is would usually result in choosing $\lambda = 0$, the λ that minimizes $\sum_1^n (Y_i - \hat{m}_\lambda(t_i))^2$ and yields an estimate of m that interpolates the data. Thus the fitted curve would be very bumpy. This is a direct consequence of underestimating the prediction error $PE(\lambda)$ because $\hat{m}_\lambda(\cdot)$ is "closer" to the Y_is than to the Y_i^*s since $\hat{m}_\lambda(\cdot)$ is fitted to the former.

The idea of cross-validation is to correct this downward bias in the estimation of $PE(\lambda)$. Note that Y_i^* is independent of $\hat{m}_\lambda(t_i)$ but Y_i is not. Cross-validation gets around this dependence by substituting $\hat{m}_\lambda^{(-i)}(t_i)$ for $\hat{m}_\lambda(t_i)$, where $\hat{m}_\lambda^{(-i)}(\cdot)$ is the curve fitted to the same data but with the ith data point (t_i, Y_i) removed. As a result Y_i and $\hat{m}_\lambda^{(-i)}(t_i)$ are independent.

Definition 4.3 *The cross-validation function is defined as*

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_\lambda^{(-i)}(t_i))^2, \tag{4.25}$$

where for each $i \in \{1, \ldots, n\}$, $\hat{m}_\lambda^{(-i)}(\cdot)$ is estimated by the same procedure as $\hat{m}_\lambda(\cdot)$ but with (t_i, Y_i) removed from the data set. The cross-validation choice of the smoothing parameter, λ_{CV}, minimizes CV.

By taking the expectation, we see that $CV(\lambda)$ is approximately unbiased for $PE(\lambda)$.

$$\begin{aligned} E(CV(\lambda)|\boldsymbol{t}) &= \frac{1}{n}E\left(\sum_{i=1}^{n}(Y_i - m(t_i))^2 + 2\sum_{i=1}^{n}(Y_i - m(t_i))(m(t_i) - \hat{m}_\lambda^{(-i)}(t_i))\right. \\ &\qquad\left. + \sum_{i=1}^{n}(m(t_i) - \hat{m}_\lambda^{(-i)}(t_i))^2|\boldsymbol{t}\right) \\ &= \sigma^2 + \frac{1}{n}E\left(\sum_{i=1}^{n}(m(t_i) - \hat{m}_\lambda^{(-i)}(t_i))^2|\boldsymbol{t}\right) \\ &\approx \sigma^2 + MSE_G(\lambda). \end{aligned} \tag{4.26}$$

Under suitable conditions, λ_{CV}, the smoothing parameter chosen by cross-validation, can be shown to converge almost surely to the optimal smoothing parameter λ_{PE}, the minimizer of the prediction error. Härdle and Marron [14] have shown such a result for the Nadaraya-Watson estimator.

On a practical level, cross-validation is a computationally intensive method. Since CV can rarely be minimized analytically, usually CV is minimized on a grid of λ in a specified interval $[\underline{\lambda}, \bar{\lambda}]$. Therefore for each combination of grid point λ_j and omitted design point t_i, the entire curve fitting procedure has to be repeated. This amounts to $n_\lambda \cdot n$ curve fits with n_λ being the number of grid points for λ.

In some cases, a short-cut formula is available which expresses $CV(\lambda)$ in terms of quantities which can be evaluated directly from applying the curve fitting procedure once to the entire data set. With this formula, only one regression curve is fitted for each value of λ. The following lemma gives the short-cut formula and conditions for its validity.

Lemma 4.1 *For each $i \in \{1, \ldots, n\}$, let $\hat{m}_\lambda^{(-i)}(\cdot)$ be the fit with (t_i, Y_i) removed from the data set $\{(t_j, Y_j)\}_1^n$; and let $\hat{m}_\lambda^{(-i)*}(\cdot)$ be the fit with (t_i, Y_i) replaced by $(t_i, \hat{m}_\lambda^{(-i)}(t_i))$ in the data set $\{(t_j, Y_j)\}_1^n$.*

Suppose that $\hat{\boldsymbol{m}}_\lambda \equiv (\hat{m}_\lambda(t_1), \ldots, \hat{m}_\lambda(t_n))^t = H_\lambda \boldsymbol{Y}$, with H_λ not dependent on $\boldsymbol{Y}$ and with its ith diagonal element $[H_\lambda]_{ii} \neq 1$. If for each i, $\hat{m}_\lambda^{(-i)*}(t_i) = \hat{m}_\lambda^{(-i)}(t_i)$ for any set of Y_is, then

$$\begin{aligned} Y_i - \hat{m}_\lambda^{(-i)}(t_i) &= \frac{Y_i - \hat{m}_\lambda(t_i)}{1 - [H_\lambda]_{ii}}, \text{ and so} \qquad (4.27) \\ CV(\lambda) &= \frac{1}{n} \sum_{i=1}^n \frac{(Y_i - \hat{m}_\lambda(t_i))^2}{(1 - [H_\lambda]_{ii})^2}. \qquad (4.28) \end{aligned}$$

Remark: The short-cut version of the CV function (4.28) casts insight on the idea of cross-validation. As discussed earlier, $n^{-1} \sum_i^n (Y_i - \hat{m}_\lambda(t_i))^2$ tends to underestimate the prediction error. By scaling the ith term by $1/(1-[H_\lambda]_{ii})^2$, CV function tries to eliminate this bias.

Proof: Since $\hat{\boldsymbol{m}}_\lambda = H_\lambda \boldsymbol{Y}$ and H_λ does not depend on $\boldsymbol{Y}$, the following holds:

$$\hat{m}_\lambda^{(-i)*}(t_i) = \sum_{j \neq i} [H_\lambda]_{ij} Y_j + [H_\lambda]_{ii} \hat{m}_\lambda^{(-i)}(t_i). \tag{4.29}$$

Using the assumption $\hat{m}_\lambda^{(-i)*}(t_i) = \hat{m}_\lambda^{(-i)}(t_i)$ yields:

$$\begin{aligned} Y_i - \hat{m}_\lambda^{(-i)}(t_i) &= Y_i - \hat{m}_\lambda^{(-i)*}(t_i) \\ &= Y_i - \sum_{j \neq i} [H_\lambda]_{ij} Y_j - [H_\lambda]_{ii} \hat{m}_\lambda^{(-i)}(t_i) \\ &= Y_i - \sum_1^n [H_\lambda]_{ij} Y_j + [H_\lambda]_{ii} (Y_i - \hat{m}_\lambda^{(-i)}(t_i)) \\ &= Y_i - \hat{m}_\lambda(t_i) + [H_\lambda]_{ii} (Y_i - \hat{m}_\lambda^{(-i)}(t_i)). \end{aligned} \tag{4.30}$$

Therefore, $Y_i - \hat{m}_\lambda^{(-i)}(t_i) = (Y_i - \hat{m}_\lambda(t_i))/(1-[H_\lambda]_{ii})$. Applying this relationship to the definition of $CV(\lambda)$ (4.25) gives the short-cut formula (4.28). □

The above applies to the scenario of curve fitting in the domain of the design variable when a global smoothing parameter λ is desired. If one wants to choose a local value of λ depending on t, localization of the CV function is straightforward by incorporating a weight function into CV. For example, a CV function for a local smoothing parameter can be defined as

$$CV(\lambda, t) = n^{-1} \sum_1^n (Y_i - \hat{m}_\lambda^{(-i)}(t_i))^2 w_i(t), \tag{4.31}$$

where $w_i(t)$ is a weight function that gives more weight to points near t than the rest. This weight function $w_i(t)$ can depend on another smoothing parameter λ^*. The $\lambda = \lambda(t)$ that minimizes $CV(\cdot, t)$ will determine the amount of smoothing around t.

4.2.2 CV for the linear operator approach

The minimizer of the penalized likelihood,

$$\mathcal{F}(\boldsymbol{Y}, L_1, \ldots, L_n) = n^{-1}\sum_1^n (Y_j - L_j(I))^2 + \lambda \int I''(u)^2 du,$$

for $I \in W_2^2[0,1]$, gives an estimator satisfying the conditions of Lemma 4.1. So the short-cut formula holds for this case. For proof and details, see [31].

4.2.3 CV for the local regression approach in curve estimation

The local polynomial regression estimator, $\hat{m}_{h_n}(t) = \hat{\beta}_{0,t}$, of $m(t)$ with $\hat{\beta}_t$ defined in (4.2), satisfies the conditions for the short-cut formula. The ith row of the H_λ matrix in this case is $[\Lambda_p]_i[(\Lambda_p^t \boldsymbol{W}\Lambda_p)^{-1}\Lambda_p^t\boldsymbol{W}]$, where $[\Lambda_p]_i$ is the ith row of Λ_p, Λ_p and $\boldsymbol{W}$ are as defined in (4.15) with $t = t_i$, $i = 1, \ldots, n$. Obviously, this H_λ matrix does not depend on Y. The lemma below verifies the other condition in Lemma 4.1 for the short-cut formula to hold.

Lemma 4.2 *For each $i \in \{1, \ldots, n\}$, let $\{w_{j,t_i}\}_1^n$ be a set of non-negative numbers with $w_{i,t_i} > 0$. Suppose that $\hat{\beta}_{t_i}^{(-i)}$ minimizes*

$$\sum_{j\neq i}\left(Y_j - \sum_{l=0}^{p}\beta_l(t_j - t_i)^l\right)^2 w_{j,t_i} \tag{4.32}$$

and that $\hat{\beta}_{t_i}^{(-i)*}$ minimizes

$$\begin{aligned}&\sum_{j\neq i}\left(Y_j - \sum_{l=0}^{p}\beta_l(t_j - t_i)^l\right)^2 w_{j,t_i} + \left(\hat{\beta}_{0,t_i}^{(-i)} - \sum_{l=0}^{p}\beta_l(t_i - t_i)^l\right)^2 w_{i,t_i}\\ &= \sum_{j\neq i}\left(Y_j - \sum_{l=0}^{p}\beta_l(t_j - t_i)^l\right)^2 w_{j,t_i} + \left(\hat{\beta}_{0,t_i}^{(-i)} - \beta_0\right)^2 w_{i,t_i}.\end{aligned} \tag{4.33}$$

Then $\hat{\beta}_{0,t_i}^{(-i)} = \hat{\beta}_{0,t_i}^{(-i)*}$ for $i \in \{1, \ldots, n\}$. Here $\hat{\beta}_{0,t_i}^{(-i)}$ and $\hat{\beta}_{0,t_i}^{(-i)*}$ are the first component of $\hat{\beta}_{t_i}^{(-i)}$ and $\hat{\beta}_{t_i}^{(-i)*}$ respectively.

Remark: The above lemma is the essence of the short-cut formula for regression with the fitting criterion

$$\sum_{1}^{n}\left(Y_j-\sum_{l=0}^{p}\beta_l(t_j-t)^l\right)^2 w_{j,t},$$

with $\hat{m}_{h_n}(t)=\hat{\beta}_0$. In particular the short-cut formula holds for local linear regression estimator ($p=1$) and local constant regression estimator ($p=0$, Nadaraya-Waston estimator), that is, when $w_{j,t}=K((t_j-t)/h_n)$ with K being a kernel function.

Proof: By the definition of $\hat{\beta}_{t_i}^{(-i)*}$,

$$\begin{aligned}
&\sum_{j\neq i}\left(Y_j-\sum_{l=0}^{p}\hat{\beta}_{l,t_i}^{(-i)*}(t_j-t_i)^l\right)^2 w_{j,t_i}+\left(\hat{\beta}_{0,t_i}^{(-i)}-\hat{\beta}_{0,t_i}^{(-i)*}\right)^2 w_{i,t_i}\\
&\leq\sum_{j\neq i}\left(Y_j-\sum_{l=0}^{p}\hat{\beta}_{l,t_i}^{(-i)}(t_j-t_i)^l\right)^2 w_{j,t_i}+\left(\hat{\beta}_{0,t_i}^{(-i)}-\hat{\beta}_{0,t_i}^{(-i)}\right)^2 w_{i,t_i}\\
&\leq\sum_{j\neq i}\left(Y_j-\sum_{l=0}^{p}\hat{\beta}_{l,t_i}^{(-i)*}(t_j-t_i)^l\right)^2 w_{j,t_i} \qquad (4.34)
\end{aligned}$$

since $\hat{\beta}_{t_i}^{(-i)}$ minimizes (4.32). Therefore $\hat{\beta}_{0,t_i}^{(-i)}=\hat{\beta}_{0,t_i}^{(-i)*}$. □

4.2.4 CV for forecasting: *FCV*

The following two sections contain discussion of ideas of cross-validation for forecasting based on different assumptions.

Unlike the case of curve fitting in the data range where the Y_is can be used to "cross-validate" the estimates $\hat{m}_{h_n}(t_i)$ computed with bandwidth h_n, the forecasting case has no "future" data beyond 1 to cross-validate $\hat{m}_{h_n,1}(1+\Delta_n)$ for h_n.

One natural idea would be to leave out a portion of the most recent data, use the forecasting estimator and the rest of the data to estimate the left-out data and minimize some measure of the discrepancy between the left-out Y_is and their estimated values to choose an optimal bandwidth h_n for forecasting. This idea has the flavour of conventional cross-validation in that a statistical model is built on part of the data and then "validated" by the rest of the data.

A formal definition of such a CV function for forecasting can be formed as below.

Definition 4.4 *The cross-validation function for forecasting is defined as:*

$$FCV(h_n) = \frac{1}{|\mathcal{S}|} \sum_{t_i \in \mathcal{S}} (\hat{m}_{h_n, t_i - \Delta_n}(t_i) - Y_i)^2, \tag{4.35}$$

where $\mathcal{S} \equiv \{t_i : \ t_i \in [1 - \rho\Delta_n, 1]\}$ *with* $\rho \geq 1$ *and* $|\mathcal{S}| = \#t_i s \in \mathcal{S}$.

Remarks:

1. Note that $\hat{m}_{h_n, t_i - \Delta_n}(t_i)$ as defined in Definition 3.1 centers the data at $t_i - \Delta_n$ and predicts Δ_n ahead. For each left-out t_i, the forecasting estimator $\hat{m}_{h_n, t_i - \Delta_n}(t_i)$ uses the rest of the data prior to the time point $t_i - \Delta_n$ to estimate $m(t_i)$. Then $\hat{m}_{h_n, t_i - \Delta_n}(t_i)$, the "forecast" at $t = t_i$, is compared to Y_i.

2. From Corollary 3.2 or 3.4, $FCV(h_n)$ could use more t_is, with $t_i \in [1 - \rho_n, 1]$ for $\rho_n \to 0$. But for computational convenience, $FCV(h_n)$ leaves out the part of the recent data with $t_i \in [1 - \rho\Delta_n, 1]$.

3. Independent work by Hart [17] uses a cross-validation idea, $TSCV$ (time series cross-validation), similar to FCV in the context of forecasting. He assumes an $AR(1)$ model for the ϵ_is and forecasts one step, $1/n$ ahead, using a locally constant regression estimator. In contrast, we assume independent ϵ_is and forecast $\Delta_n \propto n^{-1/5}$ ahead. Hart and Yi [18] recently modified $TSCV$ to $OSCV$ (one-sided cross-validation) for the bandwidth selection for curve estimation by local linear regression. They showed that the optimal bandwidth estimated by $OSCV$ is less variable than that estimated by the traditional CV as defined in (4.25). In both these papers ([17], [18]), the asymptotic mean squared error of the forecast has a simple form, $B^2 h_n^4 + V/(nh_n)$ where B and V are constants depending on K, m'',

σ^2 and f, since the estimate is forecasting $1/n$ ahead. But $AMSE(\delta)$ has a more complicated form than $B^2h_n^4 + V/(nh_n)$ since we are forecasting further ahead.

Recall that $AMSE(\delta)$, the asymptotic mean squared error of $\hat{m}_{h_n,1}(1+\Delta_n)$ is used in the plug-in approaches to estimate the optimal bandwidth for forecasting. The following corollary of Theorem 3.2 stated for non-random t_is sheds some light on the relationship between $FCV(h_n)$ and $AMSE(\delta)$.

Corollary 4.1 *Assume the same conditions as in Corollary 3.4. Then*

$$n^{4/5}\{E(FCV(h_n)) - \sigma^2 - AMSE(\delta)\} = o(1). \tag{4.36}$$

Proof: Since $\hat{m}_{h_n,t_i-\Delta_n}(t_i)$ is independent of Y_i,

$$\begin{aligned} & E(FCV(h_n)) \\ & = \frac{1}{|\mathcal{S}|}\sum_{t_i\in\mathcal{S}}\left\{E(\hat{m}_{h_n,t_i-\Delta_n}(t_i) - m(t_i))^2 + E(m(t_i) - Y_i)^2\right\} \\ & = \frac{1}{|\mathcal{S}|}\left\{\sum_{t_i\in\mathcal{S}} E(\hat{m}_{h_n,t_i-\Delta_n}(t_i) - m(t_i))^2 + \sigma^2\right\}. \end{aligned} \tag{4.37}$$

By Corollary 3.4,

$$\sup_{t_i\in\mathcal{S}} n^{4/5}\left(E\left[\hat{m}_{h_n,t_i-\Delta_n}(t_i) - m(t_i)\right]^2 - AMSE(\delta)\right) = o(1).$$

As a result,

$$n^{4/5}\{E(FCV(h_n)) - \sigma^2 - AMSE(\delta)\} = o(1). \qquad \Box.$$

The lemma below will ensure that the stronger uniform results in Theorem 3.2 and Corollaries 3.3 and 3.4 hold for t_is random. Then by the same proof, the above corollary is true for t_is random.

Lemma 4.3 *Suppose that $0 \equiv t_0 \leq t_1 \leq \ldots \leq t_n \leq t_{n+1} \equiv 1$ and that $t_1, \ldots, t_n$ are order statistics of a distribution with density function f. Suppose that W and W' are* *bounded in the support of* W*,* g' *and* f *are continuous over* $[0,1]$ *and* $\inf_{t\in[0,1]} f(t) = \alpha > 0$*. Then if* $nh_n^4 \to \infty$*,*

$$\sup_{t\in[0,1]} \left| \frac{1}{nh_n} \sum_1^n W(\frac{t_i - t}{h_n}) g(t_i) - \frac{1}{h_n} \int_0^1 W(\frac{u-t}{h_n})(gf)(u)du \right| \to 0 \tag{4.38}$$

in probability as $n \to \infty$*.*

The proof of the above lemma will use standard results on order statistics from the uniform distribution over $[0,1]$ (see, e.g. [21]). Those results are stated in the lemma below without proof.

Lemma 4.4 *Let* $U_{(i)}$*,* $i = 1, \ldots, n$ *be order statistics from the uniform distribution over* $[0,1]$*. Let* $U_{(0)} \equiv 0$*, and* $U_{(n+1)} \equiv 1$*.*

Then

1. $E(U_{(1)}) = \frac{1}{n+1}$*,* $E(U_{(1)}^2) = \frac{2}{(n+1)(n+2)}$*;*
2. $E(U_{(i)}) = \frac{i}{n+1}$*,* $Var(U_{(i)}^2) = \frac{i(n+1-i)}{(n+1)^2(n+2)}$*, for* $i = 1, \ldots, n$*;*
3. $E(U_{(i)} - U_{(i-1)}) = E(U_{(1)})$ *and* $E(U_{(i)} - U_{(i-1)})^2 = E(U_{(1)}^2)$*, for* $i = 1, \ldots, n+1$*.*

Proof of Lemma 4.3: Since

$$\begin{aligned} &\frac{1}{(n+1)h_n} \sum_1^{n+1} W(\frac{t_i - t}{h_n}) g(t_i) \\ &= \frac{n}{n+1} \frac{1}{nh_n} \sum_1^n W(\frac{t_i - t}{h_n}) g(t_i) + \frac{1}{(n+1)h_n} W(\frac{t_{n+1} - t}{h_n}) g(t_{n+1}), \end{aligned} \tag{4.39}$$

we have

$$\begin{aligned} &\sup_{t\in[0,1]} \left| \frac{1}{(n+1)h_n} \sum_1^{n+1} W(\frac{t_i - t}{h_n}) g(t_i) - \frac{1}{nh_n} \sum_1^n W(\frac{t_i - t}{h_n}) g(t_i) \right| \\ &= \sup_{t\in[0,1]} \left| \frac{1}{n+1} \left[\frac{1}{nh_n} \sum_1^n W(\frac{t_i - t}{h_n}) g(t_i) \right] + \frac{1}{(n+1)h_n} W(\frac{t_{n+1} - t}{h_n}) g(t_{n+1}) \right|. \end{aligned}$$

The assumptions on W, g and h_n ensure that the last expression will go to 0 as $n \to \infty$. Therefore this lemma is true if

$$\sup_{t\in[0,1]} \left| \frac{1}{(n+1)h_n} \sum_1^{n+1} W(\frac{t_i - t}{h_n}) g(t_i) - \frac{1}{h_n} \int_0^1 W(\frac{u-t}{h_n})(gf)(u)du \right| \to 0 \tag{4.40}$$

in probability as $n \to \infty$.

Consider the integral in (4.40):

$$
\begin{aligned}
&\frac{1}{h_n}\int_0^1 W(\frac{u-t}{h_n})(gf)(u)du \\
&= \frac{1}{h_n}\sum_{i=1}^{n+1}\int_{t_{i-1}}^{t_i} W(\frac{u-t}{h_n})(gf)(u)du \\
&= \frac{1}{h_n}\sum_{i=1}^{n+1}\int_{t_{i-1}}^{t_i} W(\frac{t_i-t}{h_n})g(t_i)f(u)du + \\
&\qquad \frac{1}{h_n}\sum_{i=1}^{n+1}\int_{t_{i-1}}^{t_i}\left[W(\frac{u-t}{h_n})g(u) - W(\frac{t_i-t}{h_n})g(t_i)\right] f(u)du. \qquad (4.41)
\end{aligned}
$$

Plugging (4.41) in the left hand side of (4.40) gives:

$$
\begin{aligned}
&\sup_{t\in[0,1]}\left|\left\{\frac{1}{(n+1)h_n}\sum_{1}^{n+1} W(\frac{t_i-t}{h_n})g(t_i) - \frac{1}{h_n}\sum_{i=1}^{n+1}\int_{t_{i-1}}^{t_i} W(\frac{t_i-t}{h_n})g(t_i)f(u)du\right\}\right. \\
&\qquad \left. - \left\{\frac{1}{h_n}\sum_{i=1}^{n+1}\int_{t_{i-1}}^{t_i}\left[W(\frac{u-t}{h_n})g(u) - W(\frac{t_i-t}{h_n})g(t_i)\right] f(u)du\right\}\right| \\
&\equiv \sup_{t\in[0,1]}\left|A(t) - B(t)\right|. \qquad (4.42)
\end{aligned}
$$

By Chebyshev's inequality, for any $\epsilon > 0$,

$$
\begin{aligned}
P\left(\sup_{t\in[0,1]}|A(t)-B(t)| > \epsilon\right) &\le \frac{E\left(\sup_{t\in[0,1]}|A(t)-B(t)|\right)}{\epsilon} \\
\le \quad & \frac{E\left(\sup_{t\in[0,1]}|A(t)|\right) + E\left(\sup_{t\in[0,1]}|B(t)|\right)}{\epsilon}. \qquad (4.43)
\end{aligned}
$$

Therefore the lemma is proved if the right hand side of (4.43) $\to 0$ as $n \to \infty$.

Consider $E\left(\sup_{t\in[0,1]}|A(t)|\right)$ first.

$$
\begin{aligned}
A(t) &= \frac{1}{(n+1)h_n}\sum_{1}^{n+1} W(\frac{t_i-t}{h_n})g(t_i) - \frac{1}{h_n}\sum_{i=1}^{n+1}\int_{t_{i-1}}^{t_i} W(\frac{t_i-t}{h_n})g(t_i)f(u)du \\
&= \frac{1}{h_n}\sum_{1}^{n+1} W(\frac{t_i-t}{h_n})g(t_i)\frac{1}{n+1} - \frac{1}{h_n}\sum_{i=1}^{n+1} W(\frac{t_i-t}{h_n})g(t_i)\int_{t_{i-1}}^{t_i} f(u)du, \qquad (4.44)
\end{aligned}
$$

which can be written in terms of the order statistics of a uniform distribution.

Let

$$
U_{(i)} = \int_0^{t_i} f(u)du, \;\; i = 0, \ldots, n+1.
$$

Then $U_{(1)},\ldots,U_{(n)}$ are order statistics of the uniform distribution over $[0,1]$ and

$$U_{(0)}=0,\quad U_{(n+1)}=1.$$

Using the facts

$$\begin{aligned}
\frac{1}{n+1} &= E(U_{(i)}-U_{(i-1)}),\\
\int_{t_{i-1}}^{t_i} f(u)du &= U_{(i)}-U_{(i-1)},\ \ i=1,\ldots,n+1
\end{aligned}$$

in $A(t)$ and then re-arranging the terms yield:

$$\begin{aligned}
A(t) &= \frac{1}{h_n}\sum_{i=1}^{n+1} W(\frac{t_i-t}{h_n})g(t_i)\left\{EU_{(i)}-U_{(i)}\right\}\\
&\qquad -\frac{1}{h_n}\sum_{i=1}^{n+1} W(\frac{t_i-t}{h_n})g(t_i)\left\{EU_{(i-1)}-U_{(i-1)}\right\}. \qquad (4.45)
\end{aligned}$$

Recall that $U_{(0)}=0$ and $U_{(n+1)}=1$, so

$$\begin{aligned}
A(t) &= \frac{1}{h_n}\sum_{i=1}^{n} W(\frac{t_i-t}{h_n})g(t_i)\left\{EU_{(i)}-U_{(i)}\right\}\\
&\qquad -\frac{1}{h_n}\sum_{i=2}^{n+1} W(\frac{t_i-t}{h_n})g(t_i)\left\{EU_{(i-1)}-U_{(i-1)}\right\}\\
&= \frac{1}{h_n}\sum_{i=2}^{n+1}\left\{W(\frac{t_{i-1}-t}{h_n})g(t_{i-1})-W(\frac{t_i-t}{h_n})g(t_i)\right\}\left\{EU_{(i-1)}-U_{(i-1)}\right\}. (4.46)
\end{aligned}$$

Using the condition $|(Wg)'|\leq C$ in (4.46) yields:

$$\sup_{t\in[0,1]}|A(t)| \leq \frac{C}{h_n}\sum_{i=2}^{n+1}\left|\frac{t_{i-1}-t_i}{h_n}\right|\left|EU_{(i-1)}-U_{(i-1)}\right|. \qquad (4.47)$$

Because of the fact that

$$\begin{aligned}
U_{(i)}-U_{(i-1)} &= \int_{t_{i-1}}^{t_i} f(u)du\\
&\geq \inf_{t\in[0,1]} f(t)(t_i-t_{i-1})=\alpha(t_i-t_{i-1}), \qquad (4.48)
\end{aligned}$$

$\sup_{t\in[0,1]}|A(t)|$ can be bounded by

$$\frac{C}{\alpha h_n^2}\sum_{i=2}^{n+1}\left|U_{(i)}-U_{(i-1)}\right|\left|EU_{(i-1)}-U_{(i-1)}\right|.$$

Schwarz' inequality can be used to bound the expected value of the above by

$$\frac{C}{\alpha h_n^2}\sum_{i=2}^{n+1}\left(E\left(U_{(i)}-U_{(i-1)}\right)^2\right)^{1/2}\left(Var(U_{(i-1)})\right)^{1/2}. \tag{4.49}$$

Using fact **3** of Lemma 4.4 then yields

$$\begin{aligned} E\left(\sup_{t\in[0,1]}|A(t)|\right) &\le \frac{C}{\alpha h_n^2}\left(E\left(U_{(1)}\right)^2\right)^{1/2}\sum_{i=2}^{n+1}\left(Var(U_{(i-1)})\right)^{1/2} \\ &= \frac{C}{\alpha h_n^2}\left(E\left(U_{(1)}\right)^2\right)^{1/2}\sum_{i=1}^{n}\left(Var(U_{(i)})\right)^{1/2}. \end{aligned}$$

Again, by facts **1** and **2** of Lemma 4.4, the last expression equals

$$\begin{aligned} &\frac{C}{\alpha h_n^2}\left(\frac{2}{(n+1)(n+2)}\right)^{1/2}\sum_{i=1}^{n}\left(\frac{i(n+1-i)}{(n+1)^2(n+2)}\right)^{1/2} \\ &= \frac{C}{\alpha h_n^2}\left(\frac{2}{(n+1)(n+2)}\right)^{1/2}\frac{n+1}{(n+2)^{1/2}}\left\{\frac{1}{n+1}\sum_{i=1}^{n}\left(\frac{i}{n+1}\left(1-\frac{i}{n+1}\right)\right)^{1/2}\right\}. \end{aligned}$$

Since

$$\frac{1}{n+1}\sum_{i=1}^{n}\left(\frac{i}{n+1}\left(1-\frac{i}{n+1}\right)\right)^{1/2} \to \int_0^1 x^{1/2}(1-x)^{1/2}dx,$$

as $n \to \infty$, we have

$$E\left(\sup_{t\in[0,1]}|A(t)|\right) = O\left(\frac{1}{n^{1/2}h_n^2}\right) \to 0 \tag{4.50}$$

under the condition that $nh_n^4 \to \infty$.

Now consider $E\left(\sup_{t\in[0,1]}|B(t)|\right)$, where

$$B(t) = \frac{1}{h_n}\sum_{i=1}^{n+1}\int_{t_{i-1}}^{t_i}\left[W(\frac{u-t}{h_n})g(u) - W(\frac{t_i-t}{h_n})g(t_i)\right]f(u)du. \tag{4.51}$$

By the assumptions of the lemma, W and W' are bounded in the support of W, and f and g' are continuous on $[0,1]$. So there exists $C_1 > 0$, such that

$$\left|W(\frac{u-t}{h_n})g(u) - W(\frac{t_i-t}{h_n})g(t_i)\right| f(u) \le C_1\left|\frac{u-t_i}{h_n}\right|. \tag{4.52}$$

Therefore,

$$\begin{aligned}
\sup_{t\in[0,1]} |B(t)| &\le \frac{1}{h_n}\sum_{i=1}^{n+1}\int_{t_{i-1}}^{t_i} C_1 \left|\frac{u-t_i}{h_n}\right| du \\
&\le \frac{1}{h_n}\sum_{i=1}^{n+1}\int_{t_{i-1}}^{t_i} C_1 \left|\frac{t_{i-1}-t_i}{h_n}\right| du \\
&= \frac{C_1}{h_n^2}\sum_{i=1}^{n+1}(t_{i-1}-t_i)^2 \\
&\le \frac{C_1}{h_n^2}\frac{1}{\alpha^2}\sum_{i=1}^{n+1}(U_{(i)}-U_{(i-1)})^2, \qquad (4.53)
\end{aligned}$$

by (4.48).

Finally,

$$\begin{aligned}
E\left(\sup_{t\in[0,1]}|B(t)|\right) &\le \frac{C_1}{\alpha^2 h_n^2}\sum_{i=1}^{n+1} E\left(U_{(i)}-U_{(i-1)}\right)^2 \\
&= \frac{C_1}{\alpha^2 h_n^2}(n+1)E\left(U_{(1)}^2\right) \\
&= \frac{C_1}{\alpha^2 h_n^2}(n+1)\frac{2}{(n+1)(n+2)} \\
&= O(\frac{1}{nh_n^2}) \to 0 \qquad (4.54)
\end{aligned}$$

under the condition that $nh_n^4 \to \infty$.

So by (4.50) and (4.54), the right hand side of (4.43) $\to 0$. □

Unfortunately, no short-cut formula for FCV has been identified, making the implementation of this criterion computationally intensive. The cross-validation criterion in the next section has a short-cut formula.

4.2.5 CV for forecasting, BCV

FCV uses the data up to $t_i - \Delta_n$ to forecast $m(t_i)$. In contrast the "backcast" cross-validation BCV in this section will use data to "forecast" the past. To understand the definition of BCV, recall that $\hat{m}_{h_n,1}(1+s) \equiv \hat{\beta}_{0,1} + \hat{\beta}_{1,1}s$ estimates $m(1+s)$ with data centered at $t = 1$. Suppose that m is really a line. Then if $\hat{m}_{h_n,1}(1+s)$ is a good estimate of $m(1+s)$ for $s \in [-\Delta_n, 0]$, $\hat{m}_{h_n,1}(1+s)$ is an equally good estimate of $m(1+s)$ as it is for $s \in [0, \Delta_n]$. If m is not a line, we know by a Taylor expansion that m is approximately linear in a neighbourhood of $t = 1$. So we can fit a line locally with the data centered at 1 and estimate $m(1+s)$ using $\hat{\beta}_{0,1}$ and $\hat{\beta}_{1,1}$ to calculate $\hat{m}_{h_n,1}(1+s)$. The accuracy of the fit is then assessed by the data in $[1-\Delta_n, 1]$. Since for $s \in [-\Delta_n, 0]$, $\hat{m}_{h_n,1}(1+s)$ gives a "backcast" instead of a forecast, we call the following cross-validation criterion BCV.

Definition 4.5 *Let $\mathcal{S}_1 = \{t_i : \ t_i \geq 1 - \Delta_n\}$ and $|\mathcal{S}_1| \equiv$ the number of $t_i s \in \mathcal{S}_1$.*

$$BCV(h_n) \equiv \frac{1}{|\mathcal{S}_1|} \sum_{t_i \in \mathcal{S}_1} c(1-t_i) \left(\hat{m}_{h_n,1}{}^{(-i)}(t_i) - Y_i \right)^2, \tag{4.55}$$

where $\hat{m}_{h_n,1}{}^{(-i)}(t_i) = \hat{\beta}_{0,1}^{(-i)} + \hat{\beta}_{1,1}^{(-i)}(t_i - 1)$, $\hat{\beta}_1^{(-i)}$ minimizes $\sum_{j \neq i}(Y_j - \beta_0 - \beta_1(t_j - 1))^2 K((t_j - 1)/h_n)$ and $c(\cdot)$ is a function that may depend on Δ_n but not on h_n.

In (4.55), we recommend $c(\cdot)$ to be chosen so that

$$E(BCV(h_n)) \approx AMSE(\delta) + constant, \tag{4.56}$$

where $AMSE(\delta)$ is the asymptotic MSE of $\hat{m}_{h_n,1}(1+\Delta_n)$ and $\delta = \Delta_n/h_n$. BCV always centers the data at $t = 1$ when estimating $m(t_i)$, thus using data in $[1-h_n, 1]$ but FCV in the last section centers the data at $t_i - \Delta_n$ when estimating $m(t_i)$, using data in $[t_i - \Delta_n - h_n, t_i - \Delta_n]$.

Suppose that a $c(\cdot)$ exists such that (4.56) holds. Then BCV behaves approximately like $AMSE(\delta)$. The hope is that the bandwidth that minimizes BCV also approximately minimizes $AMSE(\delta)$. The optimal bandwidth chosen by the BCV criterion is defined to be the one that minimizes BCV and will be sought by a grid search. Working with (4.55) directly is computationally intensive because for each h_n on the search grid and for each data point (t_i, Y_i) with $t_i \geq 1 - \Delta_n$, a regression needs to be computed. Fortunately, a short-cut formula holds for (4.55). This is shown in Lemma 4.6.

To choose a $c(\cdot)$ to satisfy (4.56), note that

$$\begin{aligned} E(BCV(h_n)) &= \frac{1}{|\mathcal{S}_1|}\sum_{t_i\in\mathcal{S}_1} c(1-t_i)E\left\{\left(\hat{m}_{h_n,1}{}^{(-i)}(t_i)-m(t_i)-[Y_i-m(t_i)]\right)^2\right\} \\ &= \frac{1}{|\mathcal{S}_1|}\sum_{t_i\in\mathcal{S}_1} c(1-t_i)E\left\{\left(\hat{m}_{h_n,1}{}^{(-i)}(t_i)-m(t_i)\right)^2\right\}+ \\ &\qquad\qquad \sigma^2\frac{1}{|\mathcal{S}_1|}\sum_{t_i\in\mathcal{S}_1} c(1-t_i) \\ &\approx \frac{1}{|\mathcal{S}_1|}\sum_{t_i\in\mathcal{S}_1} c(1-t_i)E\left\{(\hat{m}_{h_n,1}(t_i)-m(t_i))^2\right\}+ \\ &\qquad\qquad \sigma^2\frac{1}{|\mathcal{S}_1|}\sum_{t_i\in\mathcal{S}_1} c(1-t_i), \end{aligned} \tag{4.57}$$

if $|\mathcal{S}_1|^{-1}\sum_{t_i\in\mathcal{S}_1} c(1-t_i)E\left\{\left(\hat{m}_{h_n,1}{}^{(-i)}(t_i)-m(t_i)\right)^2\right\}$ can be replaced with negligible error by $|\mathcal{S}_1|^{-1}\sum_{t_i\in\mathcal{S}_1} c(1-t_i)E\left\{(\hat{m}_{h_n,1}(t_i)-m(t_i))^2\right\}$.

Thus to have (4.56), it suffices to find $c(1-t_i)$s such that

$$\frac{1}{|\mathcal{S}_1|}\sum_{t_i\in\mathcal{S}_1} c(1-t_i)E\left\{(\hat{m}_{h_n,1}(t_i)-m(t_i))^2\right\}\approx AMSE(\delta). \tag{4.58}$$

Since $AMSE(\delta)=O(n^{-4/5})$ by the asymptotic results in Chapter 3, we require the difference between the left and right sides of (4.58) to be $o(n^{-4/5})$.

The following Theorem is stated for fixed design with equally spaced t_is, $t_i=i/n$, for the choice of the c_is.

Theorem 4.1 *Assume the same conditions as in Corollary 3.3. Let* $k=[n-n\Delta_n]$, $\boldsymbol{e}=(1,\ldots,1)^t$ *(a* $k\times 1$ *vector),* $\boldsymbol{z}\equiv(z_k,\ldots,z_n)^t=(1-t_k,\ldots,1-t_n)^t$ *and* $\boldsymbol{c}\equiv(c_k,\ldots,c_n)^t=(c(z_k),\ldots,c(z_n))^t$. *Note that* $\mathcal{S}_1=\{t_k,\ldots,t_n\}$. *If* $c(\cdot)$ *satisfies:*

$$\begin{aligned} \boldsymbol{c}^t\boldsymbol{e} &= |\mathcal{S}_1| \\ \boldsymbol{c}^t\left(\frac{\boldsymbol{z}}{\Delta_n}\right) &= -|\mathcal{S}_1| \\ \boldsymbol{c}^t\left(\frac{\boldsymbol{z}}{\Delta_n}\right)^2 &= |\mathcal{S}_1| \\ \boldsymbol{c}^t\left(\frac{\boldsymbol{z}}{\Delta_n}\right)^3 &= -|\mathcal{S}_1| \\ \boldsymbol{c}^t\left(\frac{\boldsymbol{z}}{\Delta_n}\right)^4 &= |\mathcal{S}_1| \end{aligned} \tag{4.59}$$

where $(\boldsymbol{z}/\Delta_n)^i = ((z_k/\Delta_n)^i, \ldots, (z_n/\Delta_n)^i)^t$, *then*

$$\frac{1}{|\mathcal{S}_1|}\sum_{t_i \in \mathcal{S}_1} c_i E\{\,(\hat{m}_{h_n,1}(t_i) - m(t_i))^2\,\} - AMSE(\delta) = o(n^{-4/5}). \tag{4.60}$$

Proof: Since the t_is are equally spaced, we can write

$$\begin{aligned}\frac{1}{|\mathcal{S}_1|}\sum_{t_i \in \mathcal{S}_1} c_i E\left\{[\hat{m}_{h_n,1}(t_i) - m(t_i)]^2\right\} &= \frac{1}{|\mathcal{S}_1|}\sum_{i=k}^{n} c_i E\left\{[\hat{m}_{h_n,1}(t_i) - m(t_i)]^2\right\}\\ &\equiv B + V,\end{aligned}$$

where

$$\begin{aligned} B &\equiv \frac{1}{|\mathcal{S}_1|}\sum_{i=k}^{n} c_i Bias^2(\hat{m}_{h_n,1}(t_i)),\\ V &\equiv \frac{1}{|\mathcal{S}_1|}\sum_{i=k}^{n} c_i Var(\hat{m}_{h_n,1}(t_i)). \end{aligned}$$

Theorem 3.2 tells us that

$$\begin{aligned} E(\hat{\beta}_{0,1}) - m(1) &= \frac{m''(1)}{2} b_1 h_n^2 + o(h_n^2)\\ E(\hat{\beta}_{1,1}) - m'(1) &= \frac{m''(1)}{2} b_2 h_n + o(h_n)\\ Var(\hat{\beta}_{0,1}) &= \frac{1}{nh_n}\frac{\sigma^2}{f(1)}\Sigma_{11} + o(\frac{1}{nh_n})\\ Cov(\hat{\beta}_{0,1}, \hat{\beta}_{1,1}) &= \frac{1}{nh_n^2}\frac{\sigma^2}{f(1)}\Sigma_{12} + o(\frac{1}{nh_n^2})\\ Var(\hat{\beta}_{1,1}) &= \frac{1}{nh_n^3}\frac{\sigma^2}{f(1)}\Sigma_{22} + o(\frac{1}{nh_n^3}), \end{aligned}$$

where

$$\begin{aligned} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} &= \begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1} \begin{pmatrix} u_2 \\ u_3 \end{pmatrix}\\ \Sigma &= \begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1} \begin{pmatrix} u_0^* & u_1^* \\ u_1^* & u_2^* \end{pmatrix} \begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix}^{-1}, \end{aligned}$$

and u_i and u_i^* are as defined in Condition 4 in Chapter 3.

So, for $|t-1| \leq \Delta_n$,

$$\begin{aligned}
Bias(\hat{m}_{h_n,1}(t)) &= E(\hat{m}_{h_n,1}(t) - m(t)) \\
&= E\left(\hat{\beta}_{0,1} + \hat{\beta}_{1,1}(t-1) - m(t)\right) \\
&= E\left\{\hat{\beta}_{0,1} - m(1) + [\hat{\beta}_{1,1} - m'(1)](t-1)\right\} \\
&\quad -[m(t) - m(1) - m'(1)(t-1)] \\
&= \left\{b_1 \frac{h_n^2}{2} + b_2 \frac{h_n}{2}(t-1)\right\} m''(1) + o(h_n^2) \\
&\quad - \left\{m''(1)\frac{(t-1)^2}{2} + o((t-1)^2)\right\} \\
&= \frac{m''(1)}{2}\left\{b_1 h_n^2 - b_2 h_n(1-t) - (1-t)^2\right\} + o(\Delta_n^2), \qquad (4.61)
\end{aligned}$$

and

$$\begin{aligned}
Var(\hat{m}_{h_n,1}(t)) &= Var\left(\hat{\beta}_{0,1} + \hat{\beta}_{1,1}(t-1)\right) \\
&= Var(\hat{\beta}_{0,1}) + 2(t-1)Cov(\hat{\beta}_{0,1}, \hat{\beta}_{1,1}) + (t-1)^2 Var(\hat{\beta}_{1,1}) \\
&= \left(\frac{1}{nh_n}\Sigma_{11} + \frac{2(t-1)}{nh_n^2}\Sigma_{12} + \frac{(t-1)^2}{nh_n^3}\Sigma_{22}\right)\frac{\sigma^2}{f(1)} + o(\frac{1}{nh_n}),
\end{aligned}$$
$$(4.62)$$

where $o(\Delta_n^2)$ and $o(1/(nh_n))$ do not depend on t as long as $|t-1| \leq \Delta_n$. By (4.61) and (4.62) and the fact that $h_n = \Delta_n/\delta$,

$$\begin{aligned}
B &= \frac{1}{|\mathcal{S}_1|}\sum_{i=k}^{n} c_i Bias^2(\hat{m}_{h_n,1}(t_i)), \\
&= \left[\frac{m''(1)}{2}\right]^2 \frac{1}{|\mathcal{S}_1|}\sum_{i=k}^{n} c_i \left\{b_1 h_n^2 - b_2 h_n(1-t_i) - (1-t_i)^2 + o(\Delta_n^2)\right\}^2 \\
&= \left[\frac{m''(1)}{2}\right]^2 \frac{\Delta_n^4}{|\mathcal{S}_1|}\sum_{i=k}^{n} c_i \left\{\frac{b_1}{\delta^2} - \left(\frac{z_i}{\Delta_n}\right)\frac{b_2}{\delta} - \left(\frac{z_i}{\Delta_n}\right)^2 + o(1)\right\}^2 \\
&= \left[\frac{m''(1)}{2}\right]^2 \frac{\Delta_n^4}{|\mathcal{S}_1|}\sum_{i=k}^{n} c_i \left\{\frac{b_1^2}{\delta^4} - \left(\frac{z_i}{\Delta_n}\right)\frac{2b_1 b_2}{\delta^3} + \left(\frac{z_i}{\Delta_n}\right)^2 \frac{b_2^2 - 2b_1}{\delta^2} + \right. \\
&\qquad \left. + \left(\frac{z_i}{\Delta_n}\right)^3 \frac{2b_2}{\delta} + \left(\frac{z_i}{\Delta_n}\right)^4 + o(1)\right\}
\end{aligned}$$

$$\begin{aligned} &= \left[\frac{m''(1)}{2}\right]^2 \frac{\Delta_n^4}{|\mathcal{S}_1|} \left\{ \boldsymbol{c}^t \boldsymbol{e} \frac{b_1^2}{\delta^4} - \boldsymbol{c}^t \left(\frac{\boldsymbol{z}}{\Delta_n}\right) \frac{2b_1 b_2}{\delta^3} + \boldsymbol{c}^t \left(\frac{\boldsymbol{z}}{\Delta_n}\right)^2 \frac{b_2^2 - 2b_1}{\delta^2} \right. \\ &\qquad \left. + \boldsymbol{c}^t \left(\frac{\boldsymbol{z}}{\Delta_n}\right)^3 \frac{2b_2}{\delta} + \boldsymbol{c}^t \left(\frac{\boldsymbol{z}}{\Delta_n}\right)^4 + o(1) \right\}, \end{aligned} \tag{4.63}$$

and

$$\begin{aligned} V &= \frac{1}{|\mathcal{S}_1|} \sum_{i=k}^{n} c_i Var(\hat{m}_{h_n,1}(t_i)) \\ &= \frac{1}{|\mathcal{S}_1|} \sum_{i=k}^{n} c_i \left\{ \frac{1}{nh_n} \Sigma_{11} - \frac{2(1-t_i)}{nh_n^2} \Sigma_{12} \right. \\ &\qquad \left. + \frac{(1-t_i)^2}{nh_n^3} \Sigma_{22} + o(\frac{1}{nh_n}) \right\} \frac{\sigma^2}{f(1)} \\ &= \frac{1}{|\mathcal{S}_1|} \sum_{i=k}^{n} c_i \left\{ \frac{1}{nh_n} \Sigma_{11} - \frac{2\Delta_n}{nh_n^2} \Sigma_{12} \left(\frac{z_i}{\Delta_n}\right) \right. \\ &\qquad \left. + \frac{\Delta_n^2}{nh_n^3} \Sigma_{22} \left(\frac{z_i}{\Delta_n}\right)^2 + o(\frac{1}{nh_n}) \right\} \frac{\sigma^2}{f(1)} \\ &= \frac{1}{|\mathcal{S}_1|} \left\{ \boldsymbol{c}^t \boldsymbol{e} \frac{1}{nh_n} \Sigma_{11} - \boldsymbol{c}^t \left(\frac{\boldsymbol{z}}{\Delta_n}\right) \frac{2\Delta_n}{nh_n^2} \Sigma_{12} \right. \\ &\qquad \left. + \boldsymbol{c}^t \left(\frac{\boldsymbol{z}}{\Delta_n}\right)^2 \frac{\Delta_n^2}{nh_n^3} \Sigma_{22} + o(\frac{1}{nh_n}) \right\} \frac{\sigma^2}{f(1)}. \end{aligned} \tag{4.64}$$

By (4.61) and (4.62), for $t = 1 + \Delta_n$, the square of the bias and the variance of $\hat{m}_{h_n,1}(1+\Delta_n)$ are

$$\begin{aligned} Bias^2(\hat{m}_{h_n,1}(1+\Delta_n)) &= \left[\frac{m''(1)}{2}\right]^2 \left\{ b_1^2 h_n^4 + 2b_1 b_2 h_n^3 \Delta_n + (b_2^2 - 2b_1) h_n^2 \Delta_n^2 \right. \\ &\qquad \left. - 2b_2 h_n \Delta_n^3 + \Delta_n^4 + o(\Delta_n^4) \right\} \\ &= \left[\frac{m''(1)}{2}\right]^2 \Delta_n^4 \left\{ \frac{b_1^2}{\delta^4} + \frac{2b_1 b_2}{\delta^3} + \frac{b_2^2 - 2b_1}{\delta^2} \right. \\ &\qquad \left. - \frac{2b_2}{\delta} + 1 + o(1) \right\} \end{aligned} \tag{4.65}$$

and

$$Var(\hat{m}_{h_n,1}(1+\Delta_n)) = \left\{ \frac{1}{nh_n} \Sigma_{11} + \frac{2\Delta_n}{nh_n^2} \Sigma_{12} + \frac{\Delta_n^2}{nh_n^3} \Sigma_{22} + o(\frac{1}{nh_n}) \right\} \frac{\sigma^2}{f(1)}. \tag{4.66}$$

Conditions in (4.59) ensure that

$$B - Bias^2(\hat{m}_{h_n,1}(1+\Delta_n)) = o(\Delta_n^4)$$

and

$$V - Var(\hat{m}_{h_n,1}(1+\Delta_n)) = o(1/(nh_n)).$$

Since $h_n, \Delta_n \propto n^{-1/5}$, we have

$$B - Bias^2(\hat{m}_{h_n,1}(1+\Delta_n)) = o(n^{-4/5}),$$

and

$$V - Var(\hat{m}_{h_n,1}(1+\Delta_n)) = o(n^{-4/5}).$$

Therefore the theorem is proven. □

Remarks:

1. Since for each value of $[n\Delta_n]$, $\boldsymbol{c}$ needs to be calculated in the implementation of BCV, one might be interested in the limiting forms of conditions in (4.59) when $n \to \infty$ and $n\Delta_n \to \infty$. Let $c(t) = g(t/\Delta_n)$. For equally spaced t_is, one can show that there exists a function g, e.g., a fourth degree polynomial, such that

$$\begin{aligned} \int_0^1 g(u)du &= 1 \\ \int_0^1 ug(u)du &= -1 \\ \int_0^1 u^2g(u)du &= 1 \\ \int_0^1 u^3g(u)du &= -1 \\ \int_0^1 u^4g(u)du &= 1. \end{aligned} \tag{4.67}$$

2. The conditions in (4.59) are sufficient for (4.60) but may not be necessary. For example, if $m''(1) = 0$ the last two conditions in (4.59) are not needed for (4.60) to hold. However, if $m''(1) \neq 0$ and all the coefficients of powers of δ in (4.65) and (4.66) are non-zero, then the conditions in (4.59) are necessary for (4.60).

3. So far our discussion about BCV has been limited to simply providing guidelines for choosing $c(\cdot)$. It has not been shown that the h_n minimizing BCV is close to h_{opt}, the bandwidth that equals Δ_n/δ_{opt} with δ_{opt} minimizing $AMSE(\delta)$.

Note that when $|\mathcal{S}_1| > 5$, there may not be a unique vector $\boldsymbol{c}$ that satisfies (4.59). In the implementation of BCV, we will use the vector $\boldsymbol{c}$ that minimizes $\boldsymbol{c}^t\boldsymbol{c}$ subject to the conditions in (4.59).

Lemma 4.5 *Suppose that $|\mathcal{S}_1| \geq 5$ and that the $t_k, \ldots, t_n$ are distinct. Let*

$$\begin{aligned} X^t &\equiv \left[\boldsymbol{e}, \left(\frac{\boldsymbol{z}}{\Delta_n}\right), \left(\frac{\boldsymbol{z}}{\Delta_n}\right)^2, \left(\frac{\boldsymbol{z}}{\Delta_n}\right)^3, \left(\frac{\boldsymbol{z}}{\Delta_n}\right)^4\right] \quad \textit{and} \\ \boldsymbol{b}^t &\equiv (|\mathcal{S}_1|, -|\mathcal{S}_1|, |\mathcal{S}_1|, -|\mathcal{S}_1|, |\mathcal{S}_1|). \end{aligned} \tag{4.68}$$

Then the minimizer of $\boldsymbol{c}^t\boldsymbol{c}$ subject to (4.59) is $\boldsymbol{c} = X^t(XX^t)^{-1}\boldsymbol{b}$.

Proof: Let

$$L(\boldsymbol{c}, \boldsymbol{\lambda}) = \boldsymbol{c}^t\boldsymbol{c} - \boldsymbol{\lambda}^t(X\boldsymbol{c} - \boldsymbol{b}), \tag{4.69}$$

where $\boldsymbol{\lambda}$ is a 5 by 1 vector.

Setting the derivatives of L with respect to $\boldsymbol{c}$ and $\boldsymbol{\lambda}$ to zero yields:

$$\frac{\partial L}{\partial \boldsymbol{c}} = 2\boldsymbol{c} - X^t\boldsymbol{\lambda} = 0 \Rightarrow \boldsymbol{c} = \frac{1}{2}X^t\boldsymbol{\lambda}, \tag{4.70}$$

$$\frac{\partial L}{\partial \boldsymbol{\lambda}} = X\boldsymbol{c} - \boldsymbol{b} = 0 \Rightarrow X\boldsymbol{c} = \boldsymbol{b}. \tag{4.71}$$

Plugging $\boldsymbol{c}$ of (4.70) in to (4.71) gives

$$X\boldsymbol{c} = \frac{1}{2}XX^t\boldsymbol{\lambda} = \boldsymbol{b}. \tag{4.72}$$

Since $t_k, \ldots, t_n$ are distinct, X is of full row rank of 5. Therefore XX^t is invertible. Inverting XX^t in (4.72) results in $\boldsymbol{\lambda} = 2(XX^t)^{-1}\boldsymbol{b}$. Plugging this $\boldsymbol{\lambda}$ into (4.70) yields $\boldsymbol{c} = X^t(XX^t)^{-1}\boldsymbol{b}$, which may be either a minimizer or a maximizer. Since L is unbounded above, this $\boldsymbol{c}$ is a minimizer. □

Unlike FCV, BCV has the short-cut formula we established to enable one to calculate only one regression for each h_n on the search grid. Thus the process of estimating the optimal bandwidth h_n is fast. We turn to that problem now.

Lemma 4.6

$$BCV(h_n) = \frac{1}{|\mathcal{S}_1|} \sum_{t_i \in \mathcal{S}_1} c(1-t_i) \frac{(\hat{m}_{h_n,1}(t_i) - Y_i)^2}{(1-[H_{h_n}]_{ii})^2},$$

where

$$[H_{h_n}]_{ii} = (1, t_i - 1) \times \ \textit{the ith column of} \ [(\Lambda^t \boldsymbol{W} \Lambda)^{-1} \Lambda^t \boldsymbol{W}],$$
$$\Lambda^t = \begin{pmatrix} 1 & \dots & 1 \\ \\ t_1 - 1 & \dots & t_n - 1 \end{pmatrix}, \ \textit{and} \ \boldsymbol{W} = diag(K(\frac{t_j - 1}{h_n})) \tag{4.73}$$

with K a positive kernel over its support.

Proof: Recall that

$$\hat{\beta}_1 = argmin \sum_{j=1}^{n} (Y_j - \beta_0 - \beta_1(t_j - 1))^2 K(\frac{t_j - 1}{h_n}).$$

Fix i and let

$$\hat{\alpha}_{t_i} = argmin \sum_{j=1}^{n} (Y_j - \alpha_0 - \alpha_1(t_j - t_i))^2 K(\frac{t_j - 1}{h_n}).$$

Then

$$\begin{aligned} \hat{\alpha}_{0,t_i} &= \hat{\beta}_{0,1} + \hat{\beta}_{1,1}(t_i - 1) \\ \hat{\alpha}_{1,t_i} &= \hat{\beta}_{1,1}. \end{aligned} \tag{4.74}$$

The usual estimate of $m(t_i)$ using $\hat{\beta}_1$ is $\hat{\beta}_{0,1} + \hat{\beta}_{1,1}(t_i - 1)$ which is equal to $\hat{\alpha}_{0,t_i}$, the usual estimate of $m(t_i)$ based on $\hat{\alpha}_{t_i}$. Thus, both parameterizations yield the same estimate of $(m(t_1), \ldots, m(t_n))^t$, namely $H_{h_n}\boldsymbol{Y}$. For the same fixed i, also let

$$\begin{aligned}
\hat{\beta}_1^{(-i)} &= argmin \sum_{j \neq i}^{n} (Y_j - \beta_0 - \beta_1(t_j - 1))^2 K(\frac{t_j - 1}{h_n}) \\
\hat{\beta}_1^{(-i)*} &= argmin \sum_{j \neq i}^{n} (Y_j - \beta_0 - \beta_1(t_j - 1))^2 K(\frac{t_j - 1}{h_n}) \\
&\qquad + \left(\hat{\beta}_{0,1}^{(-i)} - \beta_0 - \beta_1(t_i - 1)\right)^2 K(\frac{t_i - 1}{h_n}), \\
\hat{\alpha}_{t_i}^{(-i)} &= argmin \sum_{j \neq i}^{n} (Y_j - \alpha_0 - \alpha_1(t_j - t_i))^2 K(\frac{t_j - 1}{h_n}) \\
\hat{\alpha}_{t_i}^{(-i)*} &= argmin \sum_{j \neq i}^{n} (Y_j - \alpha_0 - \alpha_1(t_j - t_i))^2 K(\frac{t_j - 1}{h_n}) \\
&\qquad + \left(\hat{\alpha}_{0,t_i}^{(-i)} - \alpha_0 - \alpha_1(t_i - t_i)\right)^2 K(\frac{t_i - 1}{h_n}),
\end{aligned}$$

then

$$\begin{aligned}
\hat{\alpha}_{0,t_i}^{(-i)} &= \hat{\beta}_{0,1}^{(-i)} + \hat{\beta}_{1,1}^{(-i)}(t_i - 1) \\
\hat{\alpha}_{1,t_i}^{(-i)} &= \hat{\beta}_{1,1}^{(-i)}, \\
\hat{\alpha}_{0,t_i}^{(-i)*} &= \hat{\beta}_{0,1}^{(-i)} + \hat{\beta}_{1,1}^{(-i)*}(t_i - 1) \\
\hat{\alpha}_{1,t_i}^{(-i)*} &= \hat{\beta}_{1,1}^{(-i)*}.
\end{aligned}$$

From Lemma 4.2, $\hat{\alpha}_{0,t_i}^{(-i)*} = \hat{\alpha}_{0,t_i}^{(-i)}$. Thus from Lemma 4.1, we have

$$\hat{\alpha}_{0,t_i}^{(-i)} - Y_i = \frac{\hat{\alpha}_{0,t_i} - Y_i}{1 - [H_{h_n}]_{ii}}.$$

By the relationship between the $\hat{\alpha}$s and $\hat{\beta}$s, the short-cut formula holds. □

In the next chapter, simulations will be carried out for the local linear forecasting estimator to compare the performance of the four methods of bandwidth estimation.

Chapter 5

Simulation: local linear forecasting estimator

5.1 Motivation and data

In this chapter we will study the local linear forecasting estimator of $m(1+\Delta_n)$ for $\Delta_n = 0.1$ and 0.2 on simulated data sets, $Y_i = m(t_i) + \epsilon_i$ with $\epsilon_i, i = 1, \ldots, n$ iid normal variables $\sim N(0, \sigma = 0.1)$, and equally spaced t_is, $t_i = i/n$. Data sets with sample sizes $n = 50$ and 100 will be generated from each of three ms, $m_1(t) = t$, $m_2(t) = t^2$ and

$$m_3(t) = \begin{cases} 2^{-7/2}(cos(4\pi t) + 1), & t \leq 1/2, \\ t^{5/2}, & t > 1/2. \end{cases} \tag{5.1}$$

Methods of estimating an optimal bandwidth by using the plug-in procedures $CAMSE$, $HAMSE$ from Sections 4.1.2 and 4.1.3 and the cross-validation procedures FCV and BCV from Sections 4.2.4 and 4.2.5 are applied to each data set. The estimators of $m''(1)$ and σ^2 needed for the plug-in procedures are those in Section 4.1.5. The kernel function K used in all four methods is the density function of the standard normal with support truncated to $[-1, 0]$. For this kernel function, the asymptotic mean square error by formula (4.11) is

$$\begin{aligned} AMSE(\delta) &= \frac{m''(1)^2}{4}\Delta_n^4(0.1532668/\delta^2 + 0.9663585/\delta + 1)^2 \\ &\quad + \frac{\sigma^2\delta}{n\Delta_n}(4.034252 + 12.1699\delta + 12.211216\delta^2). \end{aligned} \tag{5.2}$$

The optimal bandwidth for forecasting is: $h_{opt} = \Delta_n/\delta_{opt}$, where δ_{opt} minimizes $AMSE(\delta)$. The goal is to compare the estimates of $m(1 + \Delta_n)$ and the estimates of optimal bandwidths by all the bandwidth estimation procedures discussed in Chapter 4.

The functions t^p, $p = 1$ and 2, are chosen to study the sensitivity or the robustness of plug-in procedures of bandwidth estimation to the violation of assumptions on the regression functions in the asymptotic analysis in Chapter 4. Recall that plug-in procedures require the estimation of $m''(1)$ in order to estimate the optimal bandwidth for forecasting. In the estimation of $m''(1)$, the fourth derivative of m at 1, $m^{(4)}(1)$, is assumed to exist and be non-zero. Therefore, it is interesting to study the performance of the plug-in procedures when $m^{(4)}(1)$ is zero.

In contrast, the function m_3 is chosen because it has a non-zero fourth derivative at $t = 1$ and non-constant second derivative over $[0, 1]$. Although m_3'' has a discontinuity point at $t = 1/2$, Corollary 4.1 still holds when the FCV procedure is applied to m_3. Recall that Theorem 3.1 assumes that in Condition 5 m'' is continuous over $[0, 1 + \Delta_n]$ and that the results concerning the asymptotic bias and the variance of $\hat{m}_{h_n,t}(t + \Delta_n)$ hold uniformly for $t \in [a_n, 1]$ with $\liminf a_n/h_n \geq 1$. It can be easily shown that under the conditions of Theorem 3.1 but with Condition 5 replaced by the condition that m'' is continuous over $[a_0, 1 + \Delta_n]$ for some $0 < a_0 < 1$, the conclusions in Theorem 3.1 hold uniformly for $t \in [a_0, 1]$. Therefore Corollary 4.1 holds since FCV uses estimates of $m(t_i)$s with $t_i \geq 1 - \rho\Delta_n$ and these t_is are in $[a_0, 1]$ for n sufficiently large.

It is also interesting to see how the magnitude of $m''(1)$ will affect the estimated optimal bandwidth and the resulting forecasts. If $m''(1) = 0$, the square of the asymptotic bias of $\hat{m}_{h_n}(1+\Delta_n)$ is of a higher order than Δ_n^4 and thus the h_n that minimizes the $AMSE$ of $\hat{m}_{h_n}(1+\Delta_n)$ is of a different order from $n^{-1/5}$. In particular, if m is a line, the exact bias of $\hat{m}_{h_n}(1+\Delta_n)$ is zero. Thus the $AMSE$ in this case has only the variance term which equals the second term in (4.11), resulting in the asymptotically optimal bandwidth h_n being ∞. It can be proved from (5.2) that a larger absolute value of $m''(1)$ will result in a smaller asymptotically optimal bandwidth for the kernel function used in this chapter. The three ms under study have $m_1''(1) = 0$, $m_2''(1) = 2$ and $m_3''(1) = 3.75$.

5.2 Results and comments

Tables 5.1-5.3 display the summary statistics of the Δ_n-ahead forecasts ($\Delta_n = 0.1, 0.2$) along with the values of $\hat{h}_{opt}$, and the estimates of optimal bandwidth for the forecasts by the four procedures applied to 100 simulated data sets, each for sample sizes $n = 50$ or 100 and the functions $m = m_1$, m_2 and m_3. Since the estimates of the optimal bandwidth are quite variable, the median of the estimates of the optimal bandwidth for 100 simulations is a better indicator of the center of these estimates than the mean. The median of the estimates of the optimal bandwidth for 100 simulations is denoted as $\hat{h}_{opt}^{[M]}$. In the tables, the standard deviation (s.d.) and the mean square error (m.s.e.) of $\hat{m}_{h_n,1}(1+\Delta_n)$ and $\hat{h}_{opt}$ are estimated respectively as follows. Let $\hat{\theta}_b$ generically represent either $\hat{m}_{h_n,1}(1+\Delta_n)$ or $\hat{h}_{opt}$ for the bth simulation, $b = 1, \ldots, 100$, and θ the true value, i.e., $\theta = m(1+\Delta_n)$ or h_{opt}. Then $(\text{s.d.})^2$ is $\sum_{b=1}^{100} \left(\hat{\theta}_b - \bar{\hat{\theta}}\right)^2/100$ with $\bar{\hat{\theta}} = \sum_{b=1}^{100} \hat{\theta}_b/100$ and the m.s.e. is $(\theta - \bar{\hat{\theta}})^2 + (\text{s.d.})^2$. The last line of each table displays the true $m(1+\Delta_n)$ and h_{opt}.

Table 5.1: Summary of 100 forecasts ($\hat{m}$s) and estimates of optimal bandwidths by local linear regression for $m = m_1$ with sample size $n = 50$ and 100 respectively. The minimum of $AMSE$ is displayed in column (m.s.e) for **true value**.

$m = m_1$, $n = 50$						
method	$\boldsymbol{\Delta_n = 0.1}$			$\boldsymbol{\Delta_n = 0.2}$		
	$\bar{\hat{m}}$ (s.d.)	(m.s.e.)	$\hat{h}_{opt}^{[M]}$ (s.d.)	$\bar{\hat{m}}$ (s.d.)	(m.s.e.)	$\hat{h}_{opt}^{[M]}$ (s.d.)
$CAMSE$	1.08 (0.19)	(0.04)	0.30 (0.27)	1.17 (0.34)	(0.12)	0.27 (0.26)
$HAMSE$	1.08 (0.19)	(0.04)	0.27 (0.27)	1.18 (0.35)	(0.12)	0.24 (0.26)
FCV	1.11 (0.07)	(0.01)	0.51 (0.21)	1.20 (0.08)	(0.01)	0.51 (0.12)
BCV	1.10 (0.10)	(0.01)	0.31 (0.24)	1.23 (0.18)	(0.03)	0.19 (0.15)
true value	**1.10**	**0.00**	∞	**1.20**	**0.00**	∞
$m = m_1$, $n = 100$						
method	$\boldsymbol{\Delta_n = 0.1}$			$\boldsymbol{\Delta_n = 0.2}$		
	$\bar{\hat{m}}$ (s.d.)	(m.s.e.)	$\hat{h}_{opt}^{[M]}$ (s.d.)	$\bar{\hat{m}}$ (s.d.)	(m.s.e.)	$\hat{h}_{opt}^{[M]}$ (s.d.)
$CAMSE$	1.10 (0.08)	(0.01)	0.33 (0.28)	1.21 (0.14)	(0.02)	0.30 (0.25)
$HAMSE$	1.10 (0.08)	(0.01)	0.33 (0.24)	1.20 (0.14)	(0.02)	0.30 (0.24)
FCV	1.10 (0.05)	(0.00)	0.51 (0.21)	1.20 (0.06)	(0.00)	0.51 (0.10)
BCV	1.09 (0.10)	(0.01)	0.19 (0.22)	1.20 (0.18)	(0.03)	0.18 (0.14)
true value	**1.10**	**0.00**	∞	**1.20**	**0.00**	∞

Table 5.2: Summary of 100 forecasts ($\hat{m}$s) and estimates of optimal bandwidths by local linear regression for $m = m_2$ with sample size $n = 50$ and 100 respectively. The minimum of $AMSE$ is displayed in column (m.s.e) for **true value**.

$m = m_2$, $n = 50$						
method	$\boldsymbol{\Delta_n = 0.1}$			$\boldsymbol{\Delta_n = 0.2}$		
	$\bar{m}$ (s.d.)	(m.s.e.)	$\hat{h}_{opt}^{[M]}$ (s.d.)	$\bar{m}$ (s.d.)	(m.s.e.)	$\hat{h}_{opt}^{[M]}$ (s.d.)
$CAMSE$	1.16 (0.24)	(0.06)	0.24 (0.21)	1.33 (0.43)	(0.20)	0.22 (0.20)
$HAMSE$	1.16 (0.24)	(0.06)	0.22 (0.20)	1.33 (0.43)	(0.20)	0.21 (0.20)
FCV	1.16 (0.09)	(0.01)	0.29 (0.12)	1.33 (0.11)	(0.02)	0.29 (0.09)
BCV	1.13 (0.13)	(0.02)	0.30 (0.27)	1.37 (0.25)	(0.07)	0.20 (0.16)
true value	**1.21**	**0.01**	**0.34**	**1.44**	**0.02**	**0.32**

$m = m_2$, $n = 100$						
method	$\boldsymbol{\Delta_n = 0.1}$			$\boldsymbol{\Delta_n = 0.2}$		
	$\bar{m}$ (s.d.)	(m.s.e.)	$\hat{h}_{opt}^{[M]}$ (s.d.)	$\bar{m}$ (s.d.)	(m.s.e.)	$\hat{h}_{opt}^{[M]}$ (s.d.)
$CAMSE$	1.16 (0.12)	(0.02)	0.24 (0.26)	1.34 (0.20)	(0.05)	0.23 (0.22)
$HAMSE$	1.17 (0.12)	(0.02)	0.24 (0.21)	1.34 (0.21)	(0.05)	0.22 (0.20)
FCV	1.18 (0.08)	(0.01)	0.26 (0.09)	1.36 (0.11)	(0.02)	0.24 (0.08)
BCV	1.15 (0.14)	(0.02)	0.20 (0.28)	1.37 (0.20)	(0.04)	0.20 (0.18)
true value	**1.21**	**0.01**	**0.30**	**1.44**	**0.02**	**0.27**

Table 5.3: Summary of 100 forecasts ($\hat{m}$s) and estimates of optimal bandwidths by local linear regression for $m = m_3$ with sample size $n = 50$ and 100 respectively. The minimum of $AMSE$ is displayed in column (m.s.e) for **true value**.

$m = m_3$, $n = 50$						
method	$\boldsymbol{\Delta_n = 0.1}$			$\boldsymbol{\Delta_n = 0.2}$		
	$\bar{m}$ (s.d.)	(m.s.e.)	$\hat{h}_{opt}^{[M]}$ (s.d.)	$\bar{m}$ (s.d.)	(m.s.e.)	$\hat{h}_{opt}^{[M]}$ (s.d.)
$CAMSE$	1.21 (0.28)	(0.08)	0.24 (0.21)	1.45 (0.51)	(0.28)	0.23 (0.19)
$HAMSE$	1.21 (0.28)	(0.08)	0.22 (0.21)	1.45 (0.51)	(0.28)	0.21 (0.19)
FCV	1.19 (0.10)	(0.02)	0.26 (0.09)	1.37 (0.12)	(0.06)	0.25 (0.10)
BCV	1.13 (0.16)	(0.05)	0.30 (0.28)	1.39 (0.25)	(0.10)	0.19 (0.17)
true value	**1.27**	**0.02**	**0.26**	**1.58**	**0.05**	**0.24**

$m = m_3$, $n = 100$						
method	$\boldsymbol{\Delta_n = 0.1}$			$\boldsymbol{\Delta_n = 0.2}$		
	$\bar{m}$ (s.d.)	(m.s.e.)	$\hat{h}_{opt}^{[M]}$ (s.d.)	$\bar{m}$ (s.d.)	(m.s.e.)	$\hat{h}_{opt}^{[M]}$ (s.d.)
$CAMSE$	1.20 (0.20)	(0.04)	0.24 (0.26)	1.43 (0.33)	(0.13)	0.22 (0.22)
$HAMSE$	1.20 (0.20)	(0.04)	0.22 (0.24)	1.44 (0.33)	(0.13)	0.20 (0.22)
FCV	1.19 (0.09)	(0.01)	0.21 (0.08)	1.42 (0.15)	(0.05)	0.21 (0.09)
BCV	1.16 (0.16)	(0.04)	0.21 (0.29)	1.44 (0.21)	(0.06)	0.19 (0.17)
true value	**1.27**	**0.01**	**0.22**	**1.58**	**0.04**	**0.21**

Results in Tables 5.1-5.3 reveal that, among the four procedures examined, FCV produces the best forecasts in terms of the mean square error (m.s.e.) and that the m.s.e. of FCV is closest to the minimum of $AMSE$. The m.s.e. of forecasts given by the other cross-validation procedure, BCV, is smaller than the m.s.e. of forecasts given by the two plug-in procedures in most cases (eight cases out of twelve). However, BCV ties with the two plug-in procedures (in terms of the m.s.e.) in three cases and does worse in one case ($m = m_1$, $\Delta_n = 0.2$ and $n = 100$) where the m.s.e. of forecasts by BCV is 1% more than those of forecasts by plug-in procedures. Comparing the estimated mean and the estimated standard deviation of forecasts by all four procedures shows that forecasts by plug-in procedures have a bias comparable to the bias of forecasts by cross-validation procedures but tend to have a much larger variance, and that the two plug-in procedures produce almost identical results.

Tables 5.1-5.3 also show that the m.s.e.s of forecasts by all procedures increase as $|m''(1)|$ increases, which is confirmed by examining plots in Figure 5.1 where the minimum of $AMSE$ curve (solid line) increases as $|m''(1)|$ increases from $m_1''(1) = 0$ to $m_2''(1) = 2$ and $m_3''(1) = 3.75$. Moreover, the m.s.e.s of forecasts by all procedures are smaller for the larger sample size $n = 100$. This is expected since the convergence rate of $AMSE$ of $\hat{m}_{h_n,1}(1 + \Delta_n)$ is $n^{-4/5}$ and gets smaller when n gets larger.

From Tables 5.1-5.3, we see that among all four procedures, the estimates of the optimal bandwidths calculated by FCV are the least variable and have medians agreeing with the true optimal bandwidths reasonably well except for $m = m_1$. Other procedures may do well for some cases but badly in others.

Recall that all four procedures are devised to estimate the optimal bandwidth by simulating the $AMSE$ curve and that $E(FCV) - \sigma^2$ is approximately $AMSE$. To help us understand why FCV has the best performance, in the following figures, $AMSE(\delta)$ will be plotted as a function of h_n with $h_n = \Delta_n/\delta$. Moreover, the shifted FCV,

$FCV - \sigma^2$, will be plotted instead of FCV.

The two plug-in procedures, $CAMSE$ and $HAMSE$, produce almost identical simulated $AMSE$ curves. For example, Figure 5.2 shows that the $CAMSE$ and $HAMSE$ curves are in good agreement for each of 9 simulated data sets for function m_3 with $n = 50$ and $\Delta_n = 0.1$. The similarity of those curves explains why the $CAMSE$ and $HAMSE$ procedures produce very similar forecasts and estimates of optimal bandwidths. From now on only the $CAMSE$ curves are plotted for plug-in procedures.

Of the four procedures, only FCV on average simulates $AMSE$ curves reasonably well in all cases, which is not surprising in the light of Corollary 4.1. In contrast, as Figure 5.1 shows, the median of the $CAMSE$ curves (dashed) agrees with the $AMSE$ curve only for $m = m_3$ but the median of the FCV curves agrees with the $AMSE$ curve for all three functions. This observation suggests that the condition of $m^{(4)}(1) \neq 0$ is necessary for plug-in procedures to work.

Furthermore, in all cases the $AMSE$ curve simulated by FCV is less variable than the $AMSE$ curve simulated by plug-in procedures. For example, Figure 5.3 of quantiles of the FCV curves and $CAMSE$ curves shows that the $AMSE$ curve simulated by FCV is less variable than the $AMSE$ curve simulated by $CAMSE$ for $m = m_3$. This observation probably explains why forecasts obtained through FCV are less variable than forecasts by plug-in procedures.

The medians of simulated $AMSE$ curves by BCV are too rough to resemble anything like the $AMSE$ curve in all cases; see Figure 5.4 for example. This is probably due to the greatly oscillating values of the c_is, the weights used in BCV which minimize $\boldsymbol{c}^t\boldsymbol{c}$ subject to (4.59). For example, when $n = 50$ and $\Delta_n = 0.1$, we have

$\boldsymbol{c}^t = (220.75, -753.77, 632.54, 492.46, -996.23, 409.25)$.

As Figure 5.1 shows, FCV approximates the asymptotic mean squared error very well for $m = m_1$ and m_2 but not as well for $m = m_3$.

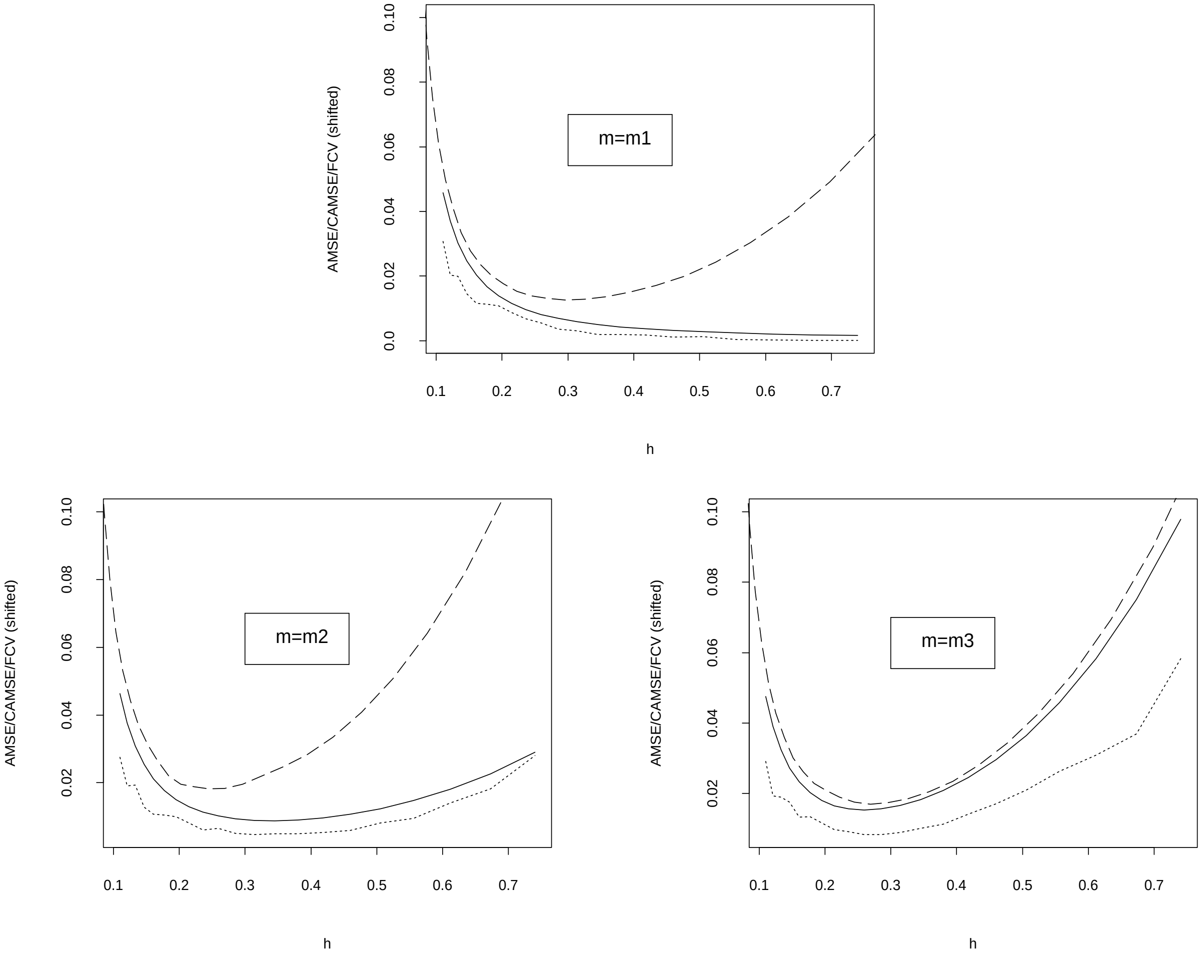

Figure 5.1: The median of 100 shifted FCV curves (dotted), the median of 100 $CAMSE$ curves (dashed) and the true $AMSE$ curve (solid) for $m = m_1, m_2$ and m_3 with $n = 50$ and $\Delta_n = 0.1$.

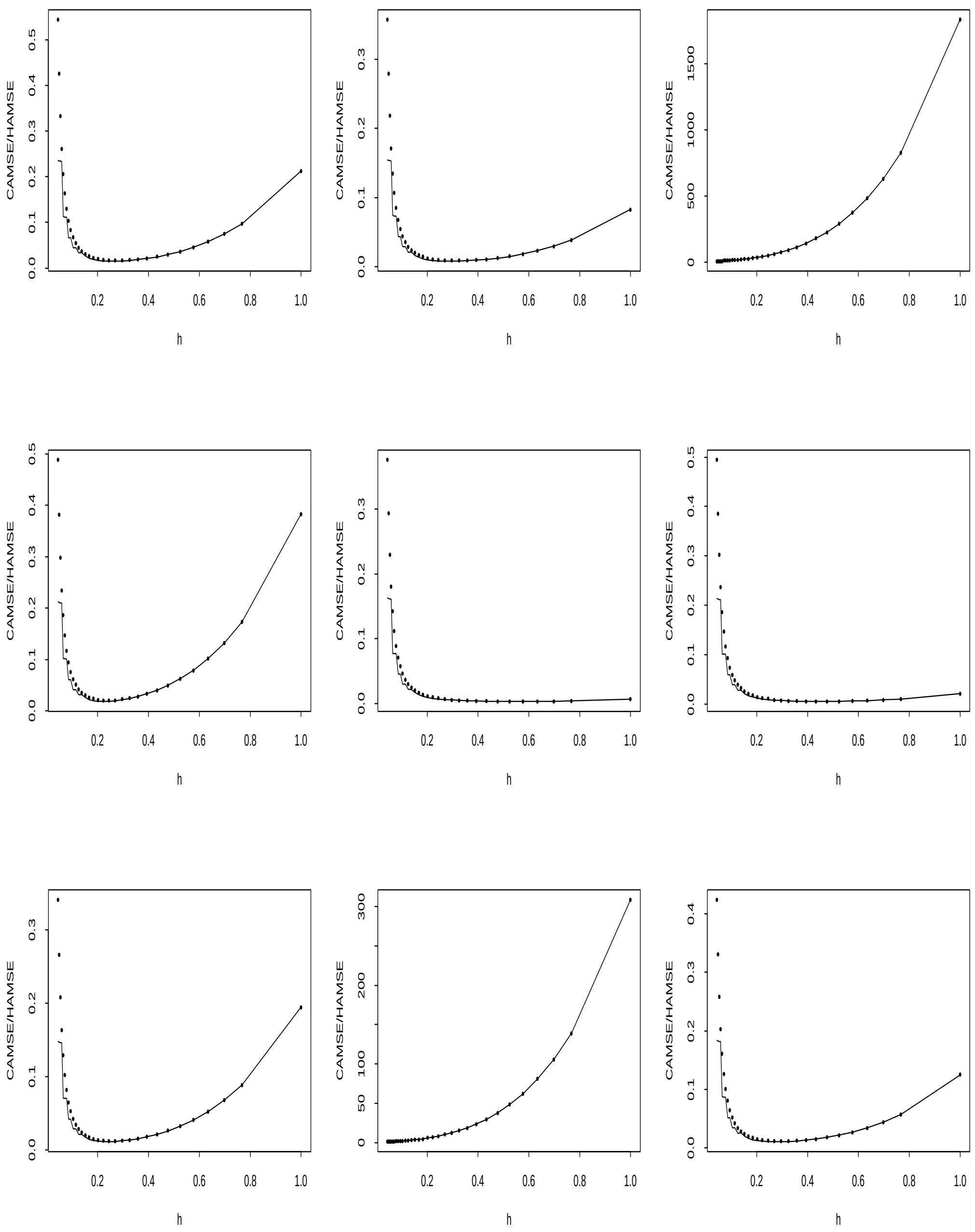

Figure 5.2: Pairs of $CAMSE$ (dotted) and $HAMSE$ (solid) curves for 9 simulated data sets for $m = m_3$, $n = 50$ and $\Delta_n = 0.1$.

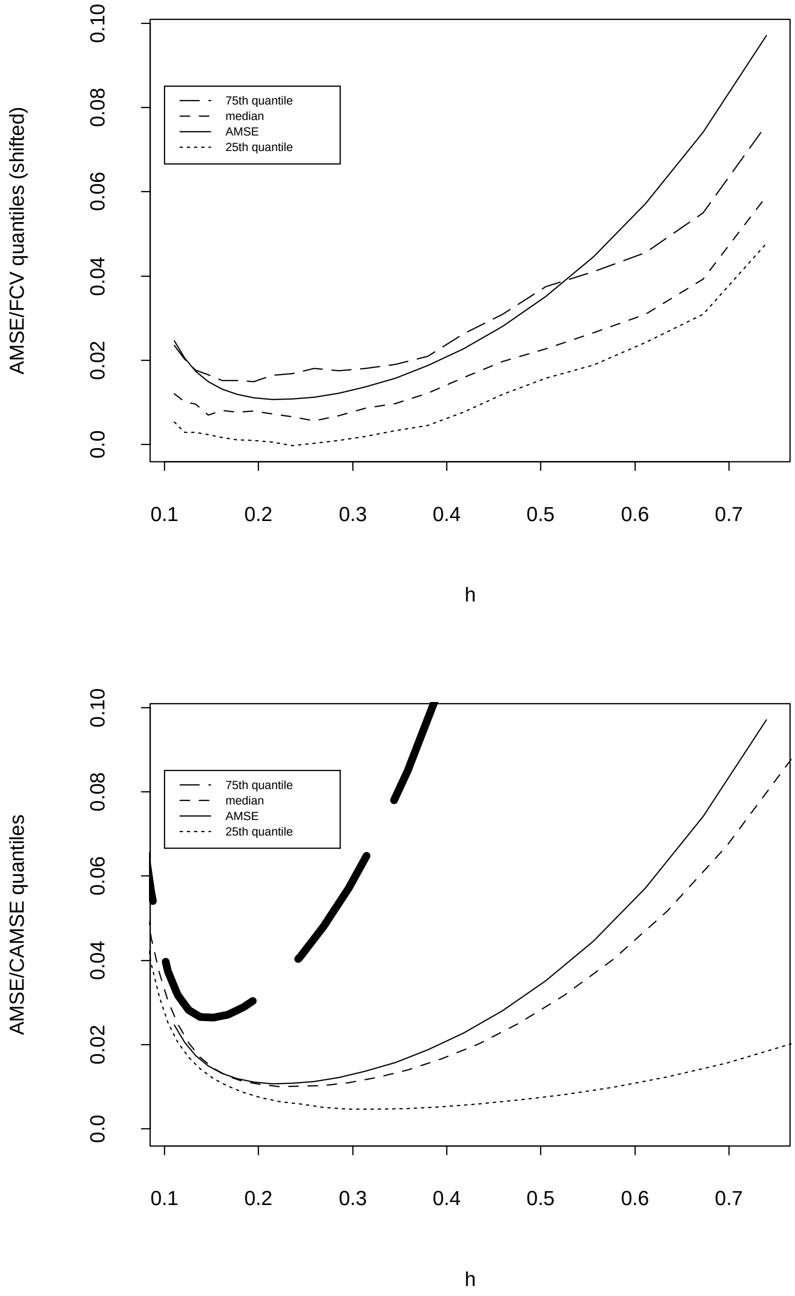

Figure 5.3: The $AMSE$ curve, the 25th and the 75th quantiles of shifted FCV and $CAMSE$ respectively, for $m = m_3$, $n = 100$ and $\Delta_n = 0.1$.

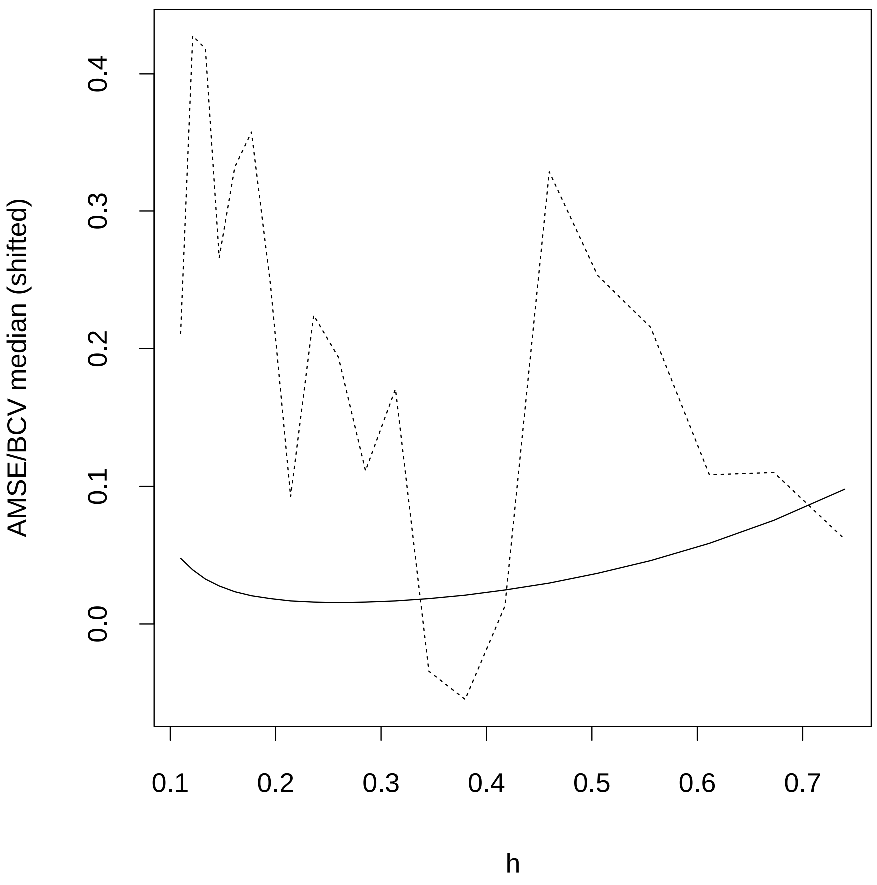

Figure 5.4: The median of 100 shifted BCV curves (dotted line) and $AMSE$ curve (solid line) for $m = m_3$, $n = 50$ and $\Delta_n = 0.1$.

It is easy to see why the shifted $FCV(h_n)$ under-estimates $AMSE(\delta)$, that is, $AMSE(\Delta_n/h_n)$, particularly for large values of h_n for $m = m_3$. Recall that the FCV score is an average of the estimates of the prediction errors for $m(t_i)$s with $t_i \in [1-\Delta_n, 1]$ and that by Corollary 3.1 or 3.3, $AMSE$ of $\hat{m}_{h_n,t_i-\Delta_n}(t_i)$ depends on $m''(t_i - \Delta_n)$, while $AMSE(\delta)$ depends on $m''(1)$. The second order derivatives of m_1 and m_2 are constants, with $m_1'' \equiv 0$ and $m_2'' \equiv 2$. So for $t_i \in [1 - \Delta_n, 1]$, $m''(t_i - \Delta_n)$ is equal to $m''(1)$. Thus, for m_1 and m_2, one sees an agreement in shape between the shifted FCV scores and the $AMSE$ curve. But for m_3, $m_3''(t) = 15t^{1/2}/4$ for $t > 1/2$, which is an increasing function of t. Therefore for $t_i \in [1-\Delta_n, 1]$, $m_3''(t_i - \Delta_n)$ is less than $m_3''(1)$. Even though as $n \to \infty$, the $m''(t_i - \Delta_n)$s converge to $m''(1)$, for a finite sample, one would expect the shifted FCV scores to under-estimate the $AMSE$ curve and that the discrepancy increases as h_n increases. However, Figure 5.5 shows that for fixed Δ_n the discrepancy is smaller for larger sample size n, as would be expected.

Based on our analysis and the above observations from our modest simulation study, we conclude that FCV has the best performance and thus is recommended.

The next chapter will contain the comparison of the backcalculation method and the local linear method applied to a fictitious AIDS example.

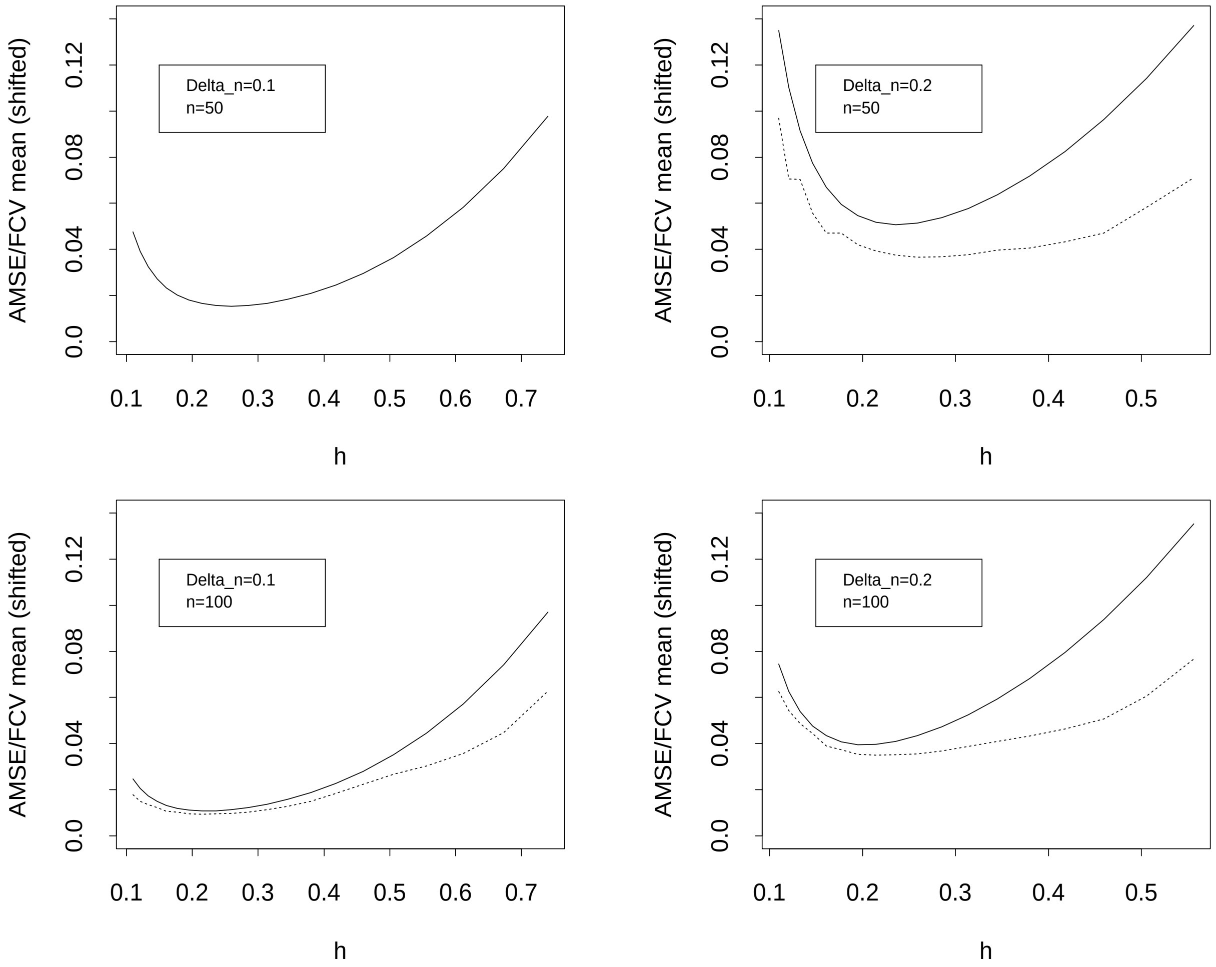

Figure 5.5: The $AMSE$ curve (solid) and the mean of 100 shifted FCV curves (dotted) for $m = m_3$ with $\Delta_n = 0.1, 0.2$ and $n = 50, 100$ respectively.

Chapter 6

The local method versus backcalculation

This chapter offers the arguments that the proposed local method should serve as an alternative to the backcalculation approach in forecasting.

6.1 Comments on the methods

Backcalculation is used to achieve two goals: to estimate the historical HIV infection curve up to the present and to forecast the number of new AIDS cases using the estimated HIV infection curve.

Backcalculation assumes the following:

1. the adjusted AIDS incidence data are accurate;

2. the incubation distribution, F, is accurately modelled and does not vary across subpopulations represented in the counts of new cases.

In reality, the raw AIDS incidence data are affected by under-reporting and reporting delays. In some developing countries, AIDS data are highly incomplete. Though adjusting the raw AIDS data for under-reporting and reporting delays improves the quality of the data significantly, the adjusted AIDS data still contain errors depending on the imputation methods, the quality of the raw data and the random nature of the data. In short, accurate AIDS counts are not available. This is more of a problem for backcalculation (Section 6.2) than for the local linear forecasting estimator.

Furthermore, knowledge about the "true" F is suspect. Knowing F is crucial to the backcalculation approach but F is very hard to estimate. Recall that F is the distribution of the incubation time from HIV infection to AIDS diagnosis. Usually external data other than the AIDS incidence data at hand are used in the procedure for estimating F. There are several problems with that procedure. First, exact infection times are known only for a few highly selected groups of individuals. So usually F is estimated from specific cohort(s) [3] and then plugged into the backcalculation model for a certain (or general) HIV infected population as the "true" incubation distribution. The applicability of F obtained from a cohort (a particular subpopulation) to a more general population is doubtful. In fact, Bacchetti et al [2] conclude from their analysis that the incubation distributions across the cohorts under study are quite different, resulting in very different estimates of HIV infection curves. Second, the changes of the definition of an AIDS case [1] (once in 1985 and then again in 1987 in the United States) further complicate the estimation of the incubation distribution. Third, under-reporting and reporting delays also affect the estimation of the incubation distribution.

In contrast, the local linear forecasting estimator has only one goal: to forecast the number of new AIDS cases. It assumes that the expectation of the adjusted AIDS incidence data reflects the true level of the underlying AIDS epidemic, i.e., $E(Y_i|t_i) = m(t_i)$ (with m the true level of the underlying AIDS epidemic) and uses the AIDS counts (adjusted when necessary) directly. The local linear forecasting estimator does not use any external data other than the AIDS data.

One might think that backcalculation should provide good estimates since it employs both external data and the AIDS data. However, as shown in the simulation study described in the next section, it does not produce better forecasts than the local linear forecasting estimator that uses the AIDS data only. In addition, it will be shown that backcalculation can not be used to produce a good estimate of the HIV infection curve up to present. The next two sections will contain discussions of the performance of the two methods.

6.2 A small simulation study

This simulation study has two objectives:

1. to compare the forecasts of the number of new AIDS cases by backcalculation and the local linear forecasting estimator,
2. to investigate how the violation of either of the two assumptions about the AIDS incidence data and the incubation distribution function F will affect the results of backcalculation.

To achieve the two objectives, the performance of both approaches is investigated in the simplest set-up: $Y_i = m(t_i) + \epsilon_i$ where the ϵ_is are independent normal random variables with standard deviation σ. Although the assumption of independent normal errors may not be realistic for the AIDS count data, it serves to show the statistical weakness of the backcalculation approach and simplifies the computation. This fictitious example will use the simplest model for F: an incubation time distribution function which is independent of the infection time, i.e., $F(t-s|s) \equiv F(t-s)$.

The simulated data are $Y_i = m(t_i) + \epsilon_i$, $i = 1, \ldots, 20$, where the ϵ_is are i.i.d. normal variables with mean zero and variance 0.25, $t_i = i/20$, $m(t_i) = \int_0^{t_i} I(s)(F(t_i - s) - F(t_{i-1} - s))ds$, $I(s) = -200(s+1)(s-2)$ and $F(t) = 1 - e^{-t}$. One hundred simulated data sets are generated from the above model. Figure 6.1 shows the incubation function F and the HIV infection rate I.

To attain the first objective, $\hat{m}(1 + \Delta_n)$, an estimate of $m(1 + \Delta_n)$, is calculated by both approaches for each of the simulation data sets for $\Delta_n = 0.1$ and $\Delta_n = 0.2$. As explained in the last section, the "true" F should be taken with a grain of salt. It will be interesting to see how the forecasts are affected by the assumed form of F. Therefore for the backcalculation approach, $\hat{m}(1 + \Delta_n)$ is estimated under a few assumed forms of F, the true F: $F(t) = 1 - e^{-t}$ and wrong Fs: $F(t) = 1 - e^{-t/\beta}$ with $\beta = 4, 1.1, 0.9, 0.5$. Since the true β is 1, $\beta = 4$ or 0.5 is grossly wrong while $\beta = 1.1$ or 0.9 is pretty close to the truth. The application of backcalculation with each F to the 100 simulated data sets yields 100 $\hat{m}(1 + \Delta_n)$s. In backcalculation, the smoothing parameter λ is chosen by cross-validation (Section 4.2.2). The application of the local linear forecasting estimator to the same data also yields 100 $\hat{m}(1 + \Delta_n)$s, but not depending on the assumed form of F. Method FCV is used to choose the bandwidth for the local linear forecasting estimator because simulations in the last chapter suggest that FCV has the best performance among the four competing bandwidth estimating procedures. For each set of 100 $\hat{m}(1 + \Delta_n)$s, the average, the standard deviation (s.d.) and the mean squared error (m.s.e.) are computed and displayed in Table 6.1. For this fictitious example, the mean squared errors can be calculated since the true value of $m(1 + \Delta_n)$ is known.

Results in Table 6.1 show that the local linear forecasting estimator always gives forecasts with smaller mean squared errors than backcalculation. Comparison of the averaged forecasts and standard deviations suggests that forecasts by backcalculation

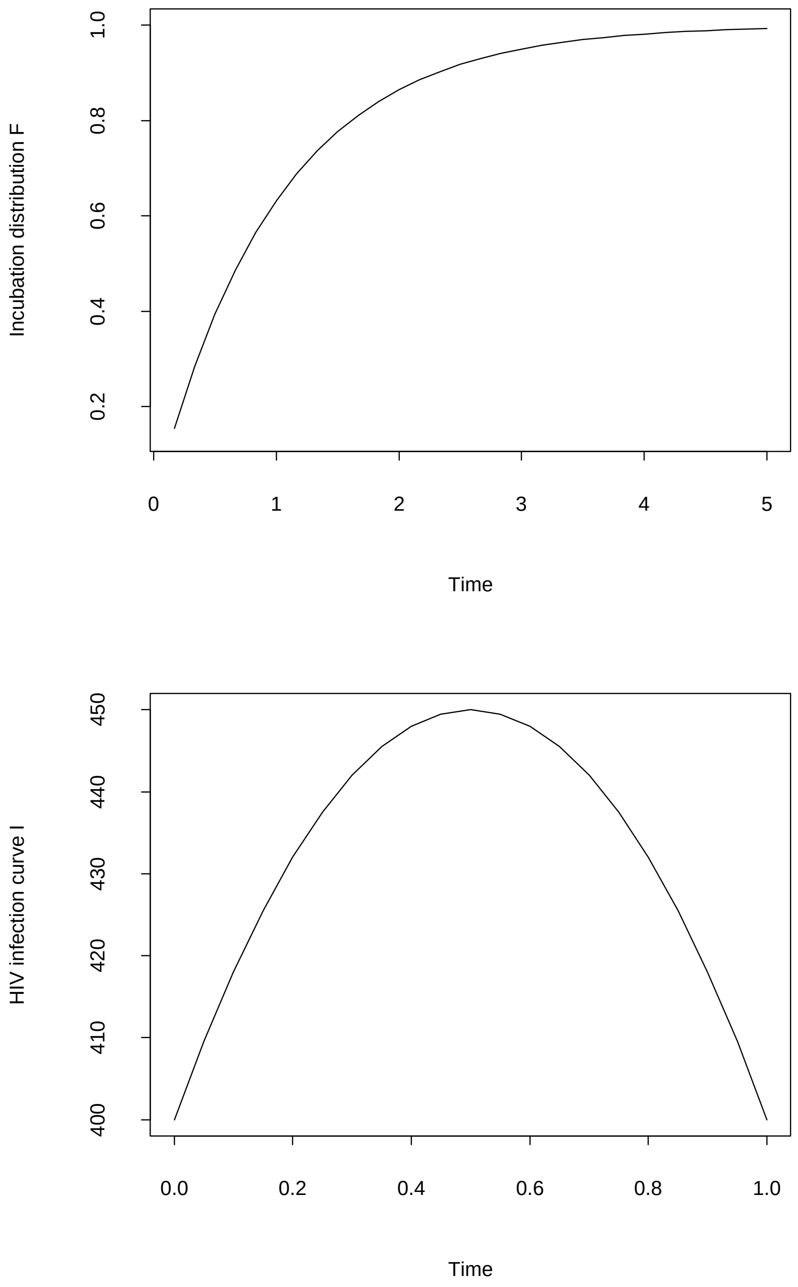

Figure 6.1: The HIV infection function I, $I(s) = -200(s+1)(s-2)$ and the incubation function F, $F(t) = 1 - e^{-t}$ used in the generation of the fictitious AIDS data.

Table 6.1: Summary of forecasts by backcalculation and local linear regression for 100 realizations of the fictitious AIDS data.

method	assumed	average of forecasts (s.d.) (m.s.e.)	
	incubation F	$\Delta_n = 0.1$	$\Delta_n = 0.2$
backcalculation	$1 - e^{-t/4}$	14.27 (1.58) (2.52)	14.85 (3.64) (13.37)
	$1 - e^{-x/1.1}$	14.10 (2.55) (6.50)	14.23 (7.42) (55.13)
	$1 - e^{-t}$, (**true**)	14.45 (2.94) (8.75)	15.52 (7.60) (58.80)
	$1 - e^{-t/0.9}$	14.31 (1.45) (2.14)	14.90 (3.24) (10.66)
	$1 - e^{-t/0.5}$	13.59 (8.10) (65.88)	12.49 (27.00) (733.04)
local linear			
FCV	NA	14.58 (0.62) (0.61)	15.51 (0.96) (1.95)
true value	$1 - e^{-t}$	**14.11**	**14.50**

are highly variable. For example, for $\Delta_n = 0.1$, even when the "true" F is used in the backcalculation, the standard deviation (s.d.) of 100 forecasts by backcalculation is 2.94, which is much higher than the corresponding standard deviation of 0.62 for the local linear forecasting estimator. Results of backcalculation when the "true" F is used suggest that forecasts should not be made based on the extrapolated $\hat{I}$.

Estimates of the conservative lower bounds for the numbers of new AIDS cases, LB_n of (1.6), are also calculated according to formula (1.7) for the backcalculation approach from the same simulated data. The average and the standard deviation of the 100 $\widehat{LB}_n$s

Table 6.2: Summary of lower bounds of forecasts by backcalculation for 100 realizations of the fictitious AIDS data.

method	assumed	average of $\widehat{LB}_n$ (s.d.)	
	incubation F	$\Delta_n = 0.1$	$\Delta_n = 0.2$
backcalculation	$1 - e^{-t/4}$	13.51 (0.89)	13.18 (0.87)
	$1 - e^{-t/1.1}$	12.89 (1.26)	11.78 (1.15)
	$1 - e^{-t}$, (**true**)	12.76 (1.21)	11.55 (1.09)
	$1 - e^{-t/0.9}$	12.71 (0.83)	11.38 (0.75)
	$1 - e^{-t/0.5}$	11.87 (0.54)	9.72 (0.44)
true lower bound	$1 - e^{-t}$, (**true**)	**12.69**	**11.48**
true value	$1 - e^{-t}$, (**true**)	**14.11**	**14.50**

for each F are summarized in Table 6.2. The true lower bounds given by formula (1.6) are calculated with F the "true" distribution used in the generation of the simulation data.

The s.d.'s of the lower bounds in Table 6.2 are much smaller than the s.d.'s of forecasts based on extrapolated $\hat{I}$s in Table 6.1. This confirms that the problem shown in Table 6.1 is due to the extrapolation of $\hat{I}$.

To attain the second objective of studying the sensitivity of backcalculation to the two parameters: the assumed form of the incubation function F and the presence of noise in the Y_is, each parameter is allowed to vary separately. First, estimates of I are backcalculated based on three similar Fs from the Y_is with no error. The three Fs are

$F(t) = 1 - e^{-t/\beta}$ with $\beta = 1.1, 1$ (true) and 0.9 (see Figure 6.2). A larger value of β means a longer incubation period. The resulting $\hat{I}$ based on Fs with $\beta = 1.1$ (long F), 1 (true F) and 0.9 (short F) are plotted together with the true I in Figure 6.2. Note that the true I is fully recovered when the true F is used in the backcalculation and when there are no errors in the Y_is.

To study the sensitivity of backcalculation to the presence of errors in the Y_is, estimates of I are backcalculated using the true F (the one used to generate the simulation data, $\beta = 1$) for different realizations of the errors in the Y_is with σ always equal to 0.5. The first plot of $\hat{I}$ in Figure 6.3 shows that I is fully recovered when there is no error in the Y_is. The other three $\hat{I}$s in Figure 6.3 are backcalculated from three realizations of the Y_is with errors. These three plots of $\hat{I}$ show that when there are errors in the Y_is, the backcalculated $\hat{I}$s are highly variable.

To get a better idea of the variability of the backcalculated $\hat{I}$s when the Y_is contain error, Figure 6.4 summarizes the 100 $\hat{I}$s by plotting the averaged $\hat{I}$, the 25th quantile, the 75th quantile and the true I.

The simulation results show that the results produced by the backcalculation approach are sensitive to the assumed form of the incubation distribution F and to the presence of errors in the Y_is and that the forecast based on the extrapolated $\hat{I}$ is highly variable (even when the true F is used). In comparison, the local linear estimator gives better forecasts by the criterion of MSE. Furthermore, the local linear estimator is much faster to compute and therefore is less expensive.

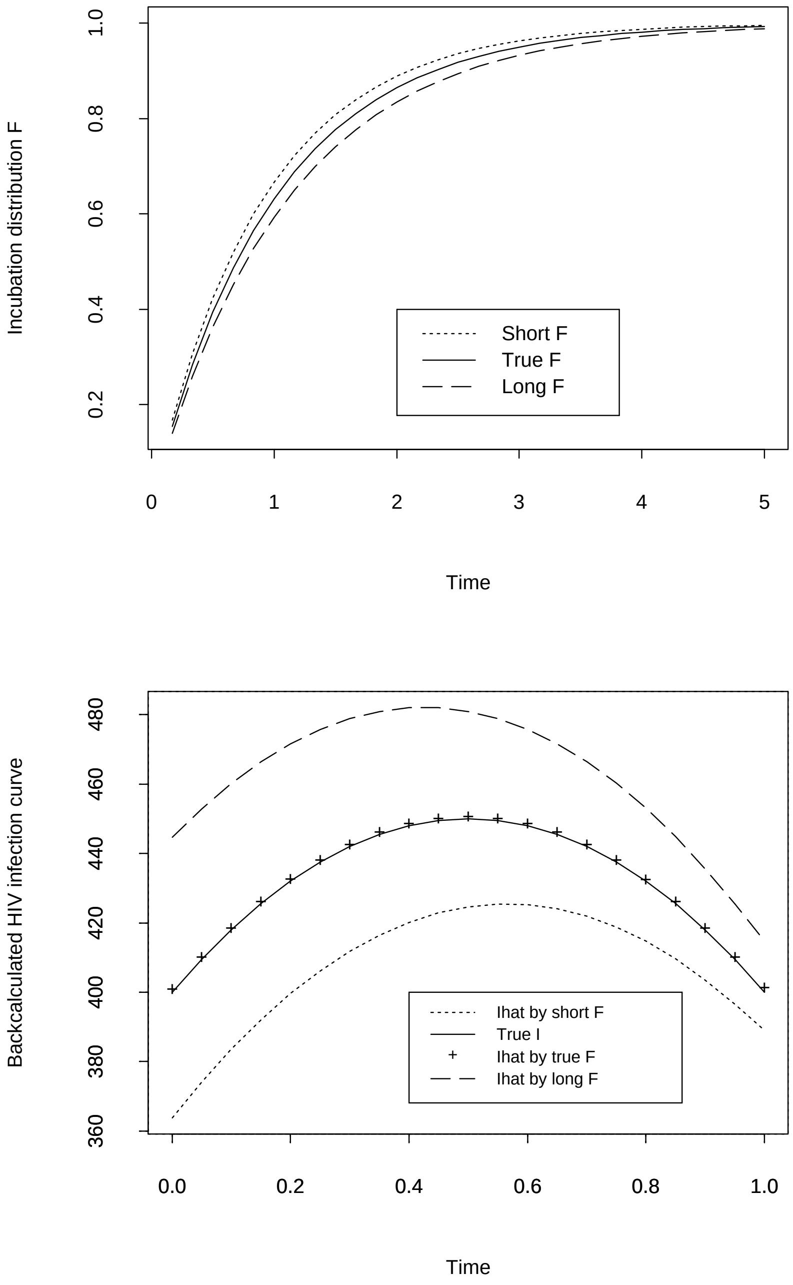

Figure 6.2: The sensitivity of $\hat{I}$ to the assumption of F in the fictitious AIDS data ($\sigma = 0$). True $F(t) = 1 - e^{-t}$, long $F(t) = 1 - e^{-t/1.1}$ and short $F(t) = 1 - e^{-t/0.9}$.

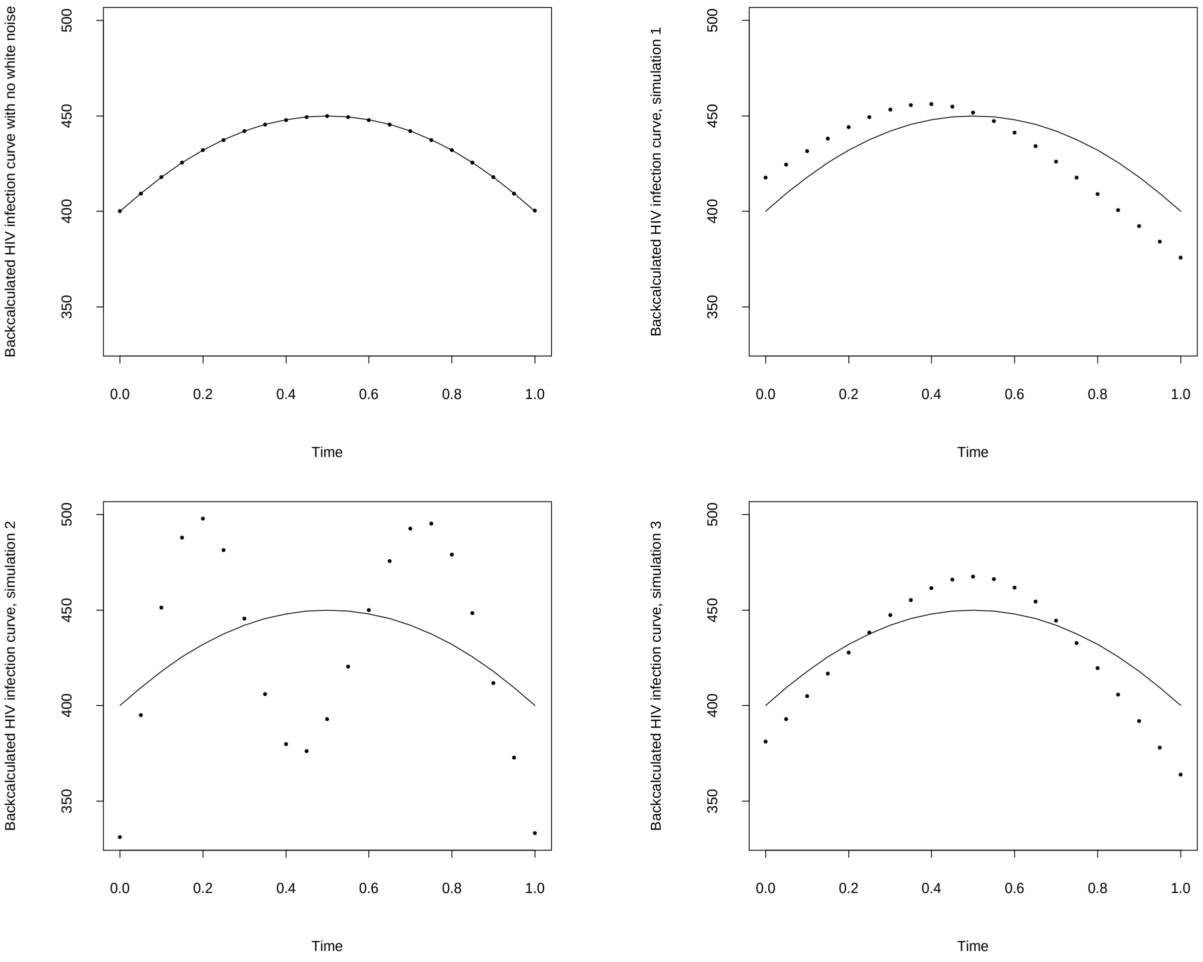

Figure 6.3: The sensitivity of $\hat{I}$ to the presence of the errors ($\sigma = 0.5$) in the fictitious AIDS data when the true F is used. The first plot shows that I is recovered from the error-free AIDS data while the rest show that $\hat{I}$ is highly distorted by the presence of errors in the AIDS incidence data. The solid curve (–) in each plot is the true I and the dotted line $(\ldots)$ the backcalculated $\hat{I}$.

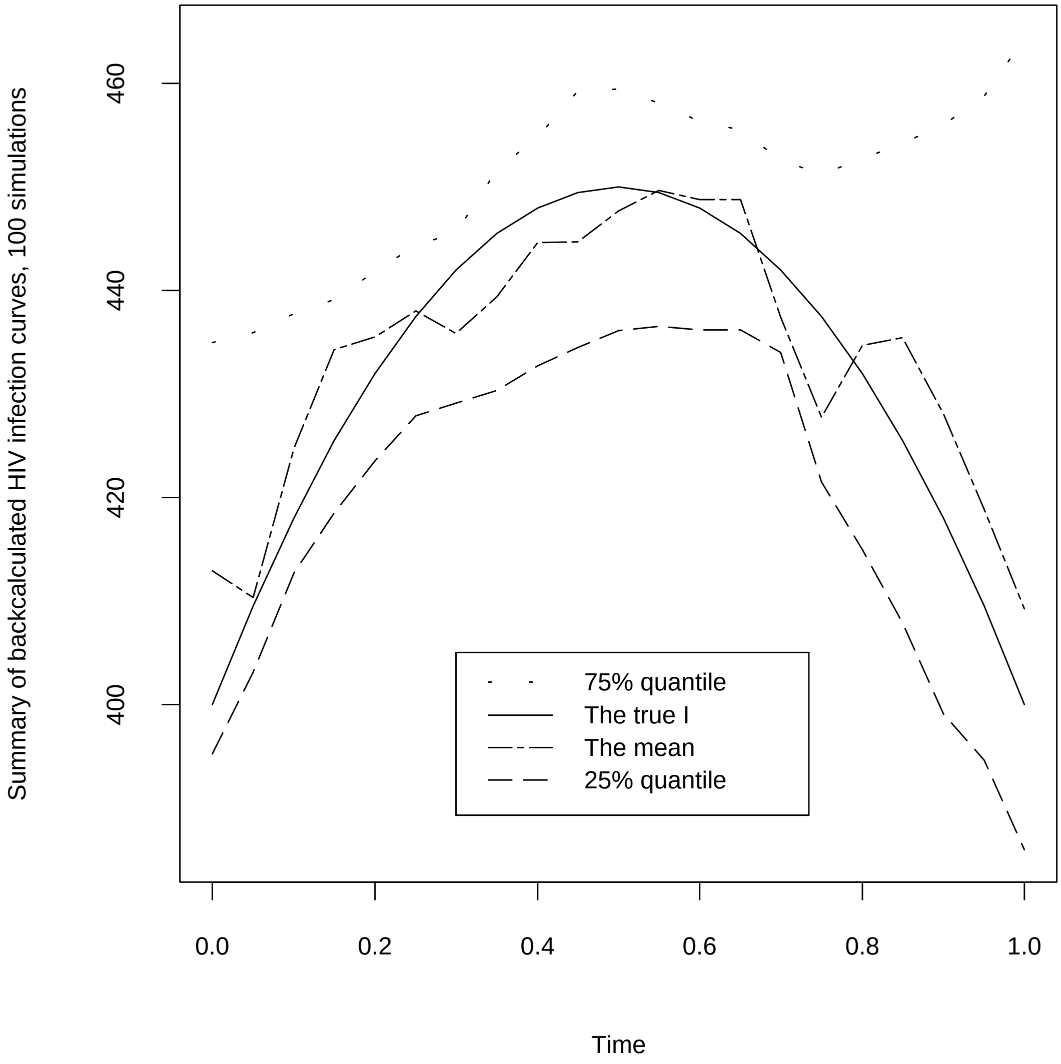

Figure 6.4: The variability of backcalculated $\hat{I}$.

6.3 Discussion of the methods and the simulation results

This section will offer theoretical insights on how the violation of either of the two assumptions on the AIDS incidence data and the incubation function F will make backcalculation perform poorly.

The reason that backcalculation is highly sensitive to inaccuracies in both the AIDS incidence data and the incubation function F is that backcalculation is a mathematically ill-posed problem. Equation (1.4)

$$m(t) = \int_0^t I(s)\left(F(t-s|s) - F(t-\frac{1}{n}-s|s)\right)ds.$$

is a Volterra equation of the first kind [22]. For details on Volterra equations, see [22].

According to [22], a linear Volterra equation of the first kind is defined as

$$\int_0^t K(t,s)I(s)ds \;=\; u(t),\;\; 0 \leq t \leq T \tag{6.1}$$

while a linear Volterra equation of the second kind is defined as

$$I(t) - \int_0^t \mathcal{K}(t,s)I(s)ds \;=\; g(t),\;\; 0 \leq t \leq T. \tag{6.2}$$

The purpose is to solve these equations for I.

An equation is well-posed if a small change in parameters, e.g. in K or u in (6.1), or in $\mathcal{K}$ or g in (6.2), causes only a small change in the solution I [22]. Equation (6.1) is not a well-posed problem.

Volterra equations of the second kind have been studied more extensively and are understood better than those of the first kind. Roughly speaking they are relatively well-posed in contrast to those of the first kind. A Volterra equation of the first kind may be converted to one of the second kind by differentiating (6.1):

$$K(t,t)I(t) \;+\; \int_0^t \frac{\partial K(t,s)}{\partial t}I(s)ds \;=\; u'(t). \tag{6.3}$$

If $K(t,t)$ does not vanish in $0 \leq t \leq T$, then dividing (6.3) by $K(t,t)$ yields a standard Volterra equation of the second kind and can be solved by existing algorithms. The exact knowledge of u (consequently the exact knowledge of u') makes the equation (6.3) possibly well-posed. Under regularity conditions, a continuous solution of equation (6.1) exists uniquely and is the same as the continuous solution of equation (6.3). Even though Volterra equations of the second kind are relatively more tractable in theory, solving them numerically is known to be hard.

However, in the backcalculation approach,

$$K(s,t) = F(t-s|s) - F(t-1/n-s|s), \tag{6.4}$$

with F being the incubation function, so $K(t,t) = 0$ for any t. Thus (6.3) is again an equation of the first kind. According to [22], a Volterra equation of the first kind can be converted into one of the second kind if K and u are sufficiently differentiable. This result is not very useful in practice because "u" (e.g. the functional form of m in backcalculation) is the very unknown that is of interest to us. As mentioned earlier only the *exact* knowledge of "u" may help convert equations of the first kind to those of the well-posed second kind.

Suppose that (1.4) can be converted into an equation of the second kind by differentiating. Then one might propose to solve the original equation by instead solving the resulting equation. Since in reality m is unknown, one needs to estimate derivatives of m from the raw data. It is well known that accurate estimation of derivatives of m from the raw data is hard, since errors inherited from the original data propagate (in terms of mean squared error). Those errors would be further amplified in $\hat{I}$ due to the ill-posed nature of the problem. Penalizing I, e.g., controlling the magnitude of $\int_0^1 I''(s)^2 ds$ when backcalculating I, does not change the ill-posed nature of the problem. This is well manifested by the simulations in the last section. A small amount of error added to m can distort the resulting estimate of I completely. See Figure 6.3. Of course, all of above discussions about Volterra equations assumes that u and K are known at all values of t, not just at the t_is.

Smoothing the data $\{Y_i\}_1^n$ prior to applying the backcalculation procedure does not help either. This can be explained by the ill-posed nature of the problem.

The ill-posed nature of backcalculation is also manifested in the sensitivity of backcalculation to the assumed form of the incubation function F. See Figure 6.2.

Figure 6.2 also exhibits a phenomenon which Brookmeyer [1] calls "nonidentifiability" of I and F: for the same set of "AIDS incidence" data if a long incubation period is assumed (e.g., $\beta = 1.1$), backcalculation produces an inflated estimate of the infection rate in order to match the observed incidence. On the other hand, if a short incubation period is assumed (e.g., $\beta = 0.9$) backcalculation underestimates the infection rates. Bacchetti [2] makes the same observation.

Recall that one major goal of the backcalculation approach is to reconstruct (estimate) the HIV infection curve $I(\cdot)$ from the AIDS incidence series. However, this nonidentifiability, accompanied by the unfortunate mathematical ill-posed nature of backcalculation, makes backcalculation unfit for the task of reconstructing the past to present HIV infection curve I.

It is also important to note that backcalculation is limited to estimating only a proportion of HIV infections. Even though there is a strong correlation between HIV infection and AIDS disease, not every HIV infected person will develop AIDS in his lifetime. The proportion of people who are infected with HIV virus and eventually develop AIDS is unknown. This is another reason that backcalculation can not recover the whole picture of the historical HIV infection pattern.

In short, one can not expect to get a reliable HIV infection curve from backcalculation unless the true F and accurate AIDS incidence data (after correction for under-reporting and reporting delays) are available. Other methods must be devised to estimate the HIV infections.

Brookmeyer [1] believes that greater improvements in reconstructing the HIV infection rates may come from empirical data on recent infection rates rather than from smoothing procedures or parametric models for I using backcalculation. As the AIDS epidemic continues, more and more data on HIV seroprevalence are available from cross-sectional surveys (since the mid-1980s) and longitudinal cohort studies. Thus obtaining infection rates directly has become possible. Brookmeyer [4][5] has proposed new approaches based on cross-sectional surveys and cohort studies in his recent research.

Though the proposed local linear estimator does not attempt to estimate the HIV infections, it does a better job at forecasting than backcalculation. The local linear estimator will be applied to two real data sets from Canada and the United Kingdom in the next chapter.

Chapter 7

Data analysis

In this chapter, forecasts using the local linear forecasting estimator and Healy and Tillett's method [20] will be compared to the corresponding true or corrected numbers of new AIDS cases. The two data sets at hand are the Canadian AIDS data, and the UK AIDS data used by Healy and Tillett [20]. For a given data set, a few recent observations will be left out and treated as "future", unknown observations. "Forecasts" by the two methods will be made using the rest of the data and compared to the left out data, actual values of these "future" observations.

7.1 Data, methods and results

The Canadian AIDS data are provided by the Laboratory Centre for Disease Control. Although the national quarterly numbers of new AIDS cases are available from the fourth quarter of 1979 to the first quarter of 1996, quarterly numbers prior to the second quarter of 1991 are assumed to be free of reporting delays. This agrees with previous work [23] which considered reporting to have ceased within six years. Since correcting for reporting delays and under-reporting is beyond the scope of this thesis, the subset of the Canadian AIDS cases that were diagnosed between the beginning of the fourth quarter of 1979 and the end of the first quarter of 1990 and reported within six years of diagnosis will be used. These data will be treated as being free of reporting delays and under-reporting, i.e, these data will be assumed to represent the true numbers of new AIDS cases. The Canadian data are displayed in Table 7.1 with quarters of a year labelled Q1-Q4.

Table 7.1: Quarterly numbers of AIDS cases reported within six years of diagnosis in Canada, 79Q4-90Q1.

	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Q1	NA	0	2	6	17	34	65	125	190	267	351	372
Q2	NA	4	2	4	16	36	82	147	223	254	317	NA
Q3	NA	0	2	7	13	39	99	163	261	295	350	NA
Q4	1	0	2	6	17	50	117	186	261	304	328	NA

The UK data as shown in Table 7.2 are taken from Table 6 in [20] in which the monthly numbers of reported new AIDS cases and Healy and Tillett's estimates of number of unreported new AIDS cases are displayed. Healy and Tillett [20] imputed the estimates of numbers of unreported new AIDS cases based on the delay distribution estimated from the data from 1984 to 1986. Thus the corrected data are the sum of the reported number and the estimate of the number of unreported new AIDS cases.

Table 7.2: Monthly numbers of AIDS cases reported to the end of September 1987 in UK. For time periods May 1986 to September 1987, the first number is the reported and the second number in parenthesis is the estimate of number of unreported new AIDS cases.

	1982	1983	1984	1985	1986	1987
Jan	1	1	6	16	25	33(7.5)
Feb	0	0	5	15	30	44(9.7)
Mar	2	2	8	18	26	35(12.2)
Apr	1	0	4	16	35	35(15.8)
May	0	1	5	14	38(0.1)	25(19.8)
Jun	0	1	6	12	27(0.4)	48(25.9)
Jul	1	5	10	17	24(0.9)	27(32.8)
Aug	1	4	14	24	29(1.4)	27(44.2)
Sep	1	3	7	23	39(2.3)	14(66.1)
Oct	1	0	14	25	34(3.3)	NA
Nov	1	5	8	22	38(4.3)	NA
Dec	2	7	16	21	46(5.8)	NA

The Canadian data and the corrected UK data are plotted in Figure 7.1. To test our forecasting method, for the Canadian data, we consider 88Q1 as the present time and for the UK data we consider December 1986 as the present time. In each plot, time has been rescaled so that the data with t in $[0, 1]$ are "known" and the data with $t > 1$ are "unknown" and will be predicted. For the Canadian data, data with t in $[0, 1]$ are quarterly numbers of new AIDS cases from the fourth quarter of 1979 (79Q4) to the first quarter of 1988 (88Q1) and are used to "forecast" the quarterly numbers of new AIDS cases from 88Q2 to 90Q1. For the UK data, data with t in $[0, 1]$ are monthly numbers of new AIDS cases from January 1982 to December 1986 and are used to "forecast" the monthly numbers of new AIDS cases from January 1987 to September 1987.

Healy and Tillett [20] fit a log-linear model in t to the most recent two years data assuming the Y_t's to be Poisson counts,

$$\log(E(Y_t)) = \alpha_0 + \alpha_1 t, \tag{7.1}$$

and then extrapolate the fit to get forecasts.

The local linear forecasting estimator with FCV, and Poisson regression using the most recent data will be applied to each data set to produce forecasts for specified time periods. Healy and Tillett [20] chose a time span of two years in the Poisson regression for the UK AIDS data. The choice of a time span of data on which Poisson regression is computed is equivalent to the choice of a bandwidth in local regression. To show that forecasts are highly dependent on the time span, that is, the bandwidth, Poisson regression will be computed using data from the most recent half year, one year and two years and then the fit will be extrapolated to yield forecasts. This choice of the spans of "most recent" data is arbitrary. The forecasts will then be compared to the corresponding true numbers of new AIDS cases for the Canadian data or to the corresponding corrected numbers of new AIDS cases for the UK data.

Tables 7.3-7.4 display the "forecasts" by the local linear method and by Poisson re-

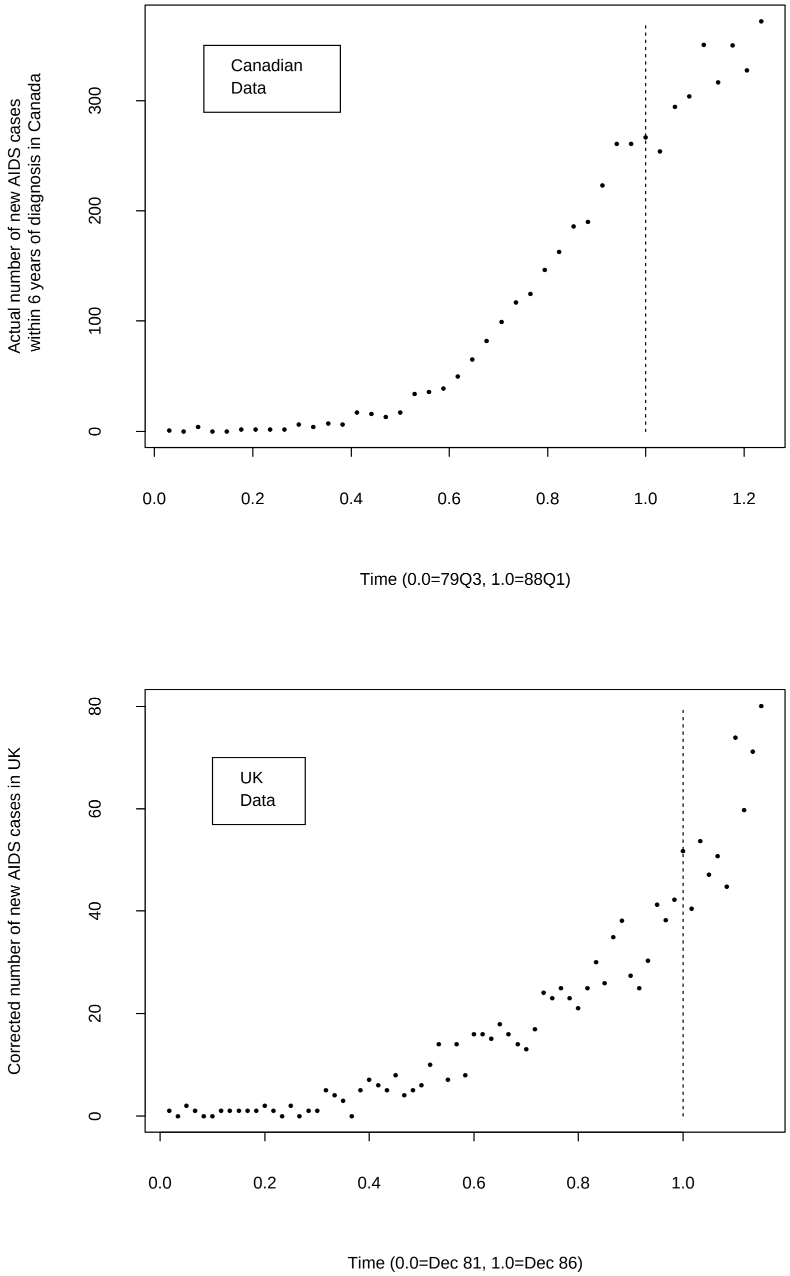

Figure 7.1: Canadian AIDS data and UK AIDS data: forecasting beyond $t = 1$.

gression using the most recent data with the three time spans. The optimal bandwidths estimated by FCV are rescaled in years so that it is easy to see the "time span" that the local linear forecasting estimator uses. The average of the squared forecasting errors ($ASFE$), displayed in the last line of these tables, serves as a measure of the precision of forecasts by each method. For the local linear method, the FCV score and estimated bandwidth (in years) are also displayed.

For the local linear forecast at $t = 1 + \Delta_n$ for each Δ_n, $AMSE(\delta)$ with $\delta = \Delta_n / h_{opt}$ can be used as a measure of accuracy of the forecast. However, $AMSE$ is unknown. Recall that $FCV - \sigma^2$ is approximately $AMSE$, provided that errors in Ys are homoscedastic. Since Figure 7.1 shows that the assumption of homoscedastic errors is reasonable for the Canadian data with t in $[0, 1]$ and for the UK data with t in $[0.5, 1.0]$, the Rice estimator can be applied to estimate σ^2 in each case. The Rice estimator as in (4.13), when applied to the Canadian data in $[0, 1]$ and the UK data in $[0.5, 1.0]$ yields 177 and 14, respectively. Although the distribution of FCV is unknown, FCV scores are displayed as an indication of forecasting errors.

Table 7.3 shows that the local linear forecasting estimator with FCV gives the closest forecasts for the Canadian data, with the smallest $ASFE$, 716. Poisson regression from data of different time spans clearly gives very different forecasts, showing that the choice of a time span of most recent data is a practical problem. This sensitivity to the choice of the bandwidth is usual for smoothing methods. Poor performance of Poisson regression to the Canadian data also suggests the possibility that Y_is may not be Poisson and thus a specific assumption on the distribution of Y_is may not be justifiable for different AIDS data sets.

Table 7.4 shows that Poisson regression using data of the most recent year gives the closest forecasts, with the smallest $ASFE$, 62. However, Poisson regression using data of the most recent half year gives the worst result, with the largest $ASFE$, 2389. In

Table 7.3: Forecasts ($\hat{m}$s) for 8 quarters, 88Q2 to 90Q1, on Canadian AIDS data (79Q4-88Q1) by local linear forecasting estimator (LL) with FCV, and by Poisson regression (PR) using most recent data respectively. The optimal bandwidth estimated by FCV is rescaled in years and denoted by h_{opt}^{yr}. The FCV score and h_{opt}^{yr} are displayed together in parenthesis. The σ^2 estimated by the Rice estimator is 177.

Time	Method				Actual #
	$\hat{m}$ by LL	$\hat{m}$ by PR on data from most recent			Y
	$\hat{m}$ (FCV,h_{opt}^{yr})	0.5 year	1.0 year	2.0 years	
88Q2	293 (64, 3.2)	273	288	311	254
88Q3	315 (483, 2.4)	279	303	341	295
88Q4	309 (234, 0.9)	286	319	373	304
89Q1	349 (312, 1.0)	292	337	408	351
89Q2	370 (403, 1.2)	299	355	446	317
89Q3	346 (685, 0.9)	306	374	488	350
89Q4	359 (1816, 0.9)	313	394	534	328
90Q1	371 (2474, 0.9)	320	415	584	372
$ASFE =$ $\sum(\hat{m}-Y)^2/8$	716	1191	1222	17033	/

Table 7.4: Forecasts ($\hat{m}$s) for 9 months, January 1987 to September 1987, on UK AIDS data (January 1982-December 1986) by local linear forecasting estimator (LL) with FCV, and by Poisson regression (PR) using most recent data respectively. The optimal bandwidth estimated by FCV is rescaled in years and denoted by h^{yr}_{opt}. The FCV score and h^{yr}_{opt} are displayed together in parenthesis. The σ^2 estimated by the Rice estimator is 14.

Time	Method				Delay-
	$\hat{m}$ by LL	$\hat{m}$ by PR on data from most recent			corrected #
	$\hat{m}$ (FCV,h^{yr}_{opt})	0.5 year	1.0 year	2.0 years	Y^c
87Jan	53 (54, 0.6)	58	47	47	41
87Feb	45 (60, 1.4)	66	49	49	54
87Mar	48 (57, 1.3)	75	52	51	47
87Apr	49 (53, 1.3)	85	55	54	51
87May	52 (44, 1.2)	96	58	57	45
87Jun	55 (27, 1.1)	109	61	60	74
87Jul	57 (20, 1.0)	124	64	63	60
87Aug	59 (46, 1.0)	141	67	66	71
87Sep	61 (68, 0.9)	160	71	69	80
$ASFE =$ $\sum(\hat{m} - Y^c)^2/9$	129	2389	62	65	/

comparison, the local linear forecasting estimator with FCV gives decent forecasts with $ASFE$, 129. Figure 7.2 shows the forecasts by the local linear method and the forecasts by the Poisson regression with the smallest $ASFE$. For the Canadian AIDS data, the Poisson regression with the smallest $ASFE$ uses the data from the most recent half year; for the UK AIDS data the Poisson regression with the smallest $ASFE$ uses the data from the most recent year. The plots in Figure 7.2 confirm our observation from Tables 7.3-7.4 that the local linear method gives very decent forecasts for both data sets.

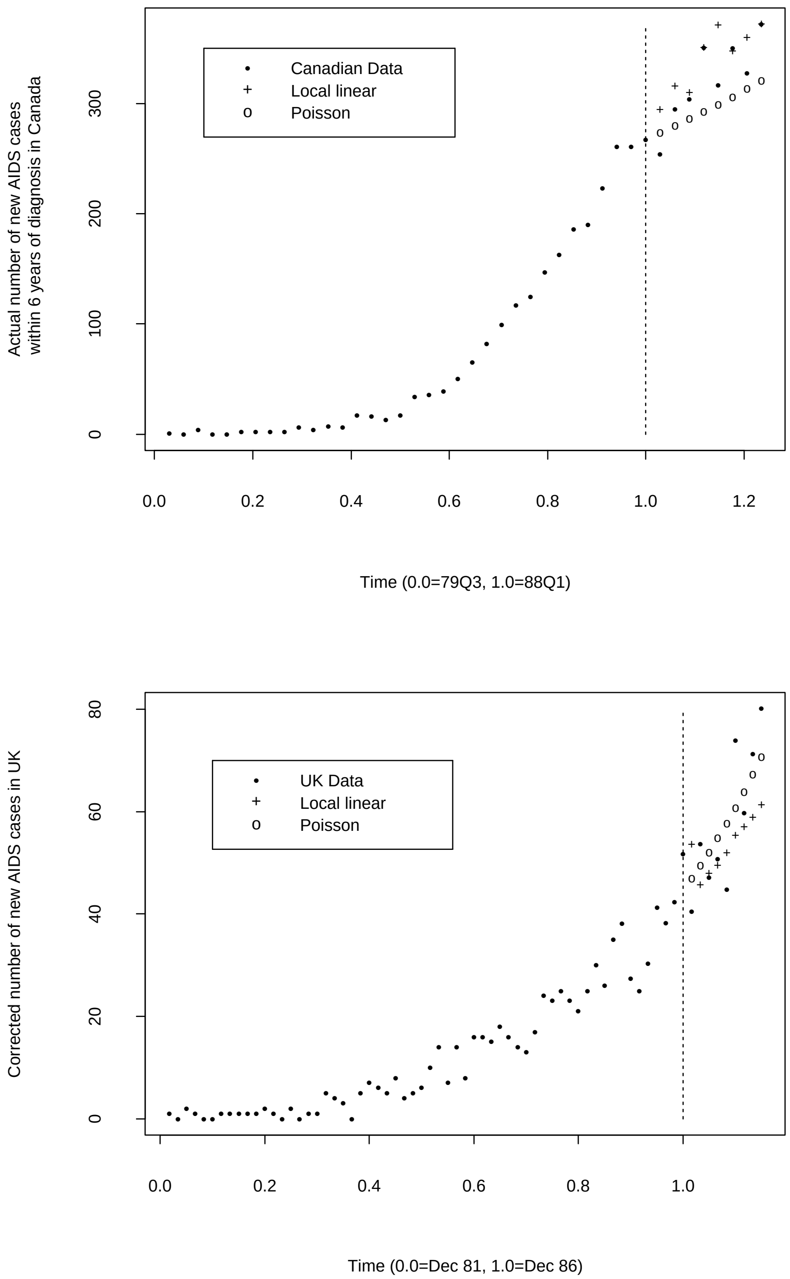

Figure 7.2: Forecasts by the local linear method and by Poisson regression using the most recent data. Among three sets of forecasts by Poisson regression, the set with the smallest $ASFE$ is plotted.

Chapter 8

Concluding remarks

In this thesis, the local linear forecasting estimator is proposed as an alternative to either backcalculation or parametric regression in the context of forecasting for AIDS data. The asymptotic theory of this estimator is developed and applied to the automatic implementation of this method, i.e., the data driven estimation of an optimal bandwidth for forecasting. For the estimation of an optimal bandwidth, two plug-in procedures and two cross-validation procedures are investigated theoretically and by simulations. Simulations clearly show the advantage of FCV over the two plug-in procedures, $CAMSE$ and $HAMSE$, and the other cross-validation procedure, BCV. The simulation study of Chapter 6 shows that the local linear method provides better forecasts than the backcalculation approach and that backcalculation can not accomplish the two tasks, that is, to provide a reasonable estimate of the HIV infection curve and to provide forecasts of AIDS incidence. Analyses of the Canadian AIDS data and the UK AIDS data show that the local linear method forecasts well. In addition, these analyses expose the practical problem of choosing an appropriate size of a time span of most recent data for use in parametric regression.

The theoretical results for the local linear forecasting estimator are derived based on the assumption of independent data. This assumption is not entirely realistic and the statistical properties of the local linear forecasting estimator using correlated data and the estimation of an optimal bandwidth will be interesting topics for future research. In the literature of AIDS research, many authors (e.g., [1], [20] and [23]) assumed the numbers of AIDS cases to be Poisson. For example, Bacchetti [1] assumed that the numbers of people infected with HIV follow a nonhomogeneous Poisson process. Under appropriate assumptions, the process of numbers of AIDS cases can be shown to be Poisson. It would be interesting to simulate the HIV infection process and the incubation distribution to get simulated AIDS counts and then to analyse the simulated AIDS counts by the backcalculation approach and the local linear forecasting estimator. Another interesting extension of the local linear forecasting estimator is to include covariates in the regression, for example, sexual practices, age and gender, and to estimate the expected numbers of new AIDS cases among a group with specific covariate values. Moreover, construction of point-wise confidence intervals of the forecasts by the local linear method is an interesting problem.

Bibliography

[1] Bacchetti, P., Segal, M.R., and Jewell, N.P. (1993), "Backcalculation of HIV Infection Rates," *Statistical Science* **8**, 82-119.

[2] Bacchetti, P., Segal, M.R., Hessol, N.A., and Jewell, N.P. (1993), "Different AIDS Incubation Periods and Their Impacts on Reconstructing Human Immunodeficiency Virus Epidemics and Projecting AIDS Incidence," *Proc. Natl. Acad. Sci. USA* **90**, 2194-2196.

[3] Brookmeyer, R. (1991), "Reconstruction and Future Trends of the AIDS Epidemic in the United States," *Science* **253**, 37-42.

[4] Brookmeyer, R., and Quinn, T.C. (1995), "Estimation of Current Human Immunodeficiency Virus Incidence Rates from a Cross-sectional Survey Using Early Diagnostic Tests," *American Journal of Epidemiology* **141**, 166-172.

[5] Brookmeyer, R., Quinn, T.C., Shepherd, M., et al (1995), "The AIDS Epidemic in India: A New Method for Estimating Current Human Immunodeficiency Virus (HIV) Incidence Rates," *American Journal of Epidemiology* **142**, 709-713.

[6] Cohen, P.T., Sande, M.A., and Bolberding, P.A. (Editors), *The AIDS Knowledge Base* (2nd Ed, 1994), Little, Brown and Company.

[7] Fan, J., and Gijbels, I. (1994), "Data-driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation," *JRSSB*, to appear.

[8] Fan, Jianqing (1992), "Design-adaptive Nonparametric Regression," *JASA* **87**, 998-1004.

[9] Gail, M.H., and Brookmeyer, R. (1988), "Methods for Projecting Course of Acquired Immunodeficiency Syndrome Epidemic (Review)," *Journal of the National Cancer Institute* **80**, 900-911.

[10] Gail, M.H., and Rosenberg, P.S. (1991), "Perspectives on Using Backcalculation to Estimate HIV Prevalence and Project AIDS Incidence" manuscript.

[11] Gasser, T., Sroka, L., and Jennen-Steinmetz, C. (1986), "Residual Variance and Residual Pattern in Nonlinear Regression, " *Biometrika* **73**, 625-633.

[12] Gasser, T., Kneip, A. and Köhler, W. (1991), "A Flexible and Fast Method for Automatic Smoothing," *JASA* **86**, 643-652.

[13] Hall, P., Sheather, S.J., Jones, M.C. and Marron, J.S. (1991), "On Optimal Data-based Bandwidth Selection in Kernal Density Estimation," *Biometrika* **78**, 263-271.

[14] Härdle, W., and Marron, J.S. (1985), "Optimal Bandwidth Selection in Nonparametric Regression Function Estimation," *The Annals of Statistics* **13**, 1465-1481.

[15] Härdle, W. (1989), *Applied Nonparametric Regression*, SIAM Cambridge University Press.

[16] Hart, J. D. and Wehrly, T. E. (1986), "Kernel Regression Estimation Using Repeated Measurements Data," *JASA* **81**, 1080-1088.

[17] Hart, J. D. (1994), "Automated Kernel Smoothing of Dependent Data by Using Time Series Cross-validation," *Journal of the Royal Statistical Society*, Ser. B. **56**, 529-542.

[18] Hart, J. D. and Yi, S. (1996), "One-sided Cross-validation," manuscript.

[19] Hastie, T. and Loader, C. (1993), "Local regression: Automatic Carpentry," *Statist. Sci.* **8**, 120-143.

[20] Healy, M. J. R., and Tillett, H. E. (1988), "Short-term Extrapolation of the AIDS Epidemic," *JRSSA* **151**, 50-61.

[21] Lehmann, E. L. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, Holden Day.

[22] Linz, Peter (1985), *Analytical and Numerical Methods for Volterra Equations*, SIAM Philadelphia.

[23] Marion, S. A., and Schechter M. T. (Dec., 1990), "Estimation of the Number of Persons in Canada Infected with Human Immunodeficiency Virus," *A report commissioned by the Minister of National Health and Welfare.*

[24] Marion, S. A., and Schechter M. T. (Mar., 1992), "Human Immunodeficiency Virus Infection in Canada: An Updated Analysis Using Backcalculation," *A report commissioned by the Minister of National Health and Welfare.*

[25] Rice, J. (1984), "Bandwidth Choice for Nonparametric Regression," *The Annals of Statistics* **12**, 1215-1230.

[26] Ruppert, D., Sheather, S.J., and Wand, M.P. (1993), "An Effective Bandwidth Selector for Local Least Squares Regression," manuscript.

[27] Stone, C. J. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics* **10**, 1040-1053.

[28] Wahba, G. (1977), "Practical Approximate Solutions to Linear Operator Equations When the Data Are Noisy," *SIAM J. Numer. Anal.* **Vol. 14**, 651-667.

[29] Wahba, G. (1979), "Smoothing and Ill posed Problems," *Solution Methods for Integral Equations with Applications*, Michael Golberg (Editor), Plenum Press, 183-194.

[30] Wahba, G. (1982), "Constrained Regularization for Ill Posed Linear Operator Equations, with Application in Meteorology and Medicine," *Statistical Decision Theory and Related Topics III* **Vol. 2**, Shanti S. Gupta and J. O. Berger (Editors), 383-418.

[31] Wahba, G. (1990), *Spline Models for Observational Data*, SIAM.