The Not-So-Smoother

by

Jennifer Paige Eveson

B.Sc. University of Victoria, 1994

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

Department of Statistics

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

October 1996

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date October 5 / 96

Abstract

In this thesis, a local smoothing method, termed the *not-so-smoother*, designed to estimate discontinuous regression functions is proposed. Local smoothing techniques estimate the regression function at a given point by finding the "best fit" through the observations within a fixed neighbourhood of the point. The "best fit" can be the best constant fit (which gives the moving average smoother), the best linear fit, the best k-degree polynomial fit, *et cetera.* The not-so-smoother finds the best local *broken constant* fit, a piecewise constant function with exactly one simple discontinuity. Unlike any of the traditional local smoothing methods, the not-so-smoother uses discontinuous local fits and, therefore, has the ability to preserve discontinuities in the function.

Consistency of the not-so-smoother under general conditions is proven. Performance of the smoother on simulated data, both continuous and discontinuous, is demonstrated, and an application to a real data set of electric current recordings through an ion channel in a cell membrane is also shown. Variations of the not-so-smoother which can lead to improved performance in certain situations are investigated.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Foremost I want to thank my supervisor, Jean Meloche, for approaching me with this project. Not only has the topic proved interesting and rewarding, but also I couldn't have chosen a better person to work with. Jean was always willing to give his advice, ideas, and encouragement, all of which were much needed.

Secondly, I am grateful to Nancy Heckman for her very useful comments and without whose careful reading my thesis may well have contained erroneous proofs.

Thanks also to Grace Chiu for her invaluable help with Latex, C programming, and various other problems along the way, but most of all for being such a terrific officemate and friend.

Lastly, I would like to acknowledge NSERC for their financial support which has enabled me to enter the real world debt free.

Chapter 1

Introduction

A common problem in statistics is, given a set of noisy data, to estimate the underlying function, also called the signal or regression function. Oftentimes, a parametric form for the signal is assumed. For example, in the case of linear regression, the signal is assumed to belong to the class of linear functions.

A more general problem arises when the regression function is not restricted to any specific form. If limited or no information is known about the underlying function then it is preferable not to place restrictions on the function's form. Nonparametric regression techniques such as kernel smoothers and smoothing splines are typically used to estimate the signal in such situations (Green and Silverman, 1994; Eubank, 1988). These smoothing methods are based on the assumption that the regression function is continuous. Applying the traditional smoothing methods when the function is not continuous tends to smooth away the discontinuities, or "jumps".

In applications where discontinuities are present in the signal, it is important that the jumps be preserved. In fact, identification and preservation of the discontinuities can be the main objective in estimating the signal. For example, in Chapter 4 we consider a data set of electric current recordings through an ion channel in a cell membrane. The current is believed to have two, or perhaps more, conductance levels between which it switches randomly. The data are quite noisy and the goal is to restore the true signal. For data sets such as this, we seek a smoother with the ability to preserve discontinuities. In particular, we are interested in the case where no *a priori* information regarding the parametric form

of the regression function, including the number and location of discontinuities, is known.

In 1986, McDonald and Owen investigated the estimation of a discontinuous regression function and introduced a smoothing algorithm called the split linear smoother which can produce a discontinuous estimate. The main idea of their approach is, for any given point, to obtain linear fits based on points to the left, to the right, and to both sides of the point in question. A smoothed estimate for the point is found by taking a weighted average of the left, right, and central fits, where the weights are chosen based on goodness-of-fit measures for the linear fits. The split linear smoother is quite complicated in concept and in practice, and its statistical properties were only briefly discussed.

Following the work of McDonald and Owen, Hall and Titterington (1992) developed an edge-preserving smoothing algorithm summarized as follows. (Note that an edge is simply a discontinuity in the function.) For each design point, a left, right and central smooth is calculated by taking a weighted average of data to the left, right, and both sides, respectively, of the point. Weights are determined through a procedure which equates leading terms in the Taylor series expansions of the expected smooths. Discontinuities are then identified using various diagnostics to compare the three smooths. Having identified the jumps, a final estimate of the function at each design point is produced. The left (right) smooth is used for points to the left (right) of and sufficiently close to a discontinuity; otherwise, the central smooth is used. As acknowledged by Hall and Titterington, some arbitrariness is involved in the procedure, including the choice of weights, the number of neighbouring data points used in the smooths, and the diagnostics used to identify jumps. Furthermore, properties of their estimator are not known.

Several related problems have also been studied. The case where exactly one discontinuity exists was considered by Müller in 1992. Müller identifies the discontinuity by comparing right and left one-sided kernel smooths, and he also gives results on the rate of convergence of the estimated point of discontinuity to the true point.

In a 1991 paper, Lee presents a method for detecting and measuring the size of change-points, which are discontinuities occurring in the kth order derivative of the regression function. The use of smoothing splines to estimate the function once the change-points have been detected is briefly discussed.

Shiau (1987) proposes a partial spline model to estimate the underlying function of a noisy data set, assuming that the locations of the discontinuities are known. In a case where this information is unavailable, Shiau's approach could be used in conjunction with a procedure for identifying the discontinuities, such as Lee's.

Less recently, Feder (1975) investigated regression functions which have different parametric forms over different regions of the domain. Specifically, he studied the asymptotic distribution of least-squares estimators in this segmented regression problem. Because each segment's parametric form must be specified, this problem is not as general as the one we are interested in.

In this thesis, we propose a smoother which, like those of McDonald and Owen and Hall and Titterington, is capable of preserving discontinuities, but is simpler conceptually as well as in application. Moreover, consistency of the estimator can be shown. No assumptions about the parametric form of the underlying function or the existence of discontinuities are made.

This new smoothing technique is a local smoother. Local smoothers estimate the regression function at a given point by finding the "best fit" through a fixed number of neighbouring observations. The "best fit" depends on the form of the fits considered as well as the criterion used to determine best. In the case of the moving average smoother, only constant fits are considered and best is determined by minimizing the squared error between the observations and the constant. Local linear, quadratic, and k-degree polynomial fits are also commonly used. The smoothing method we propose finds the best "broken constant" fit, a piecewise constant function with exactly one simple

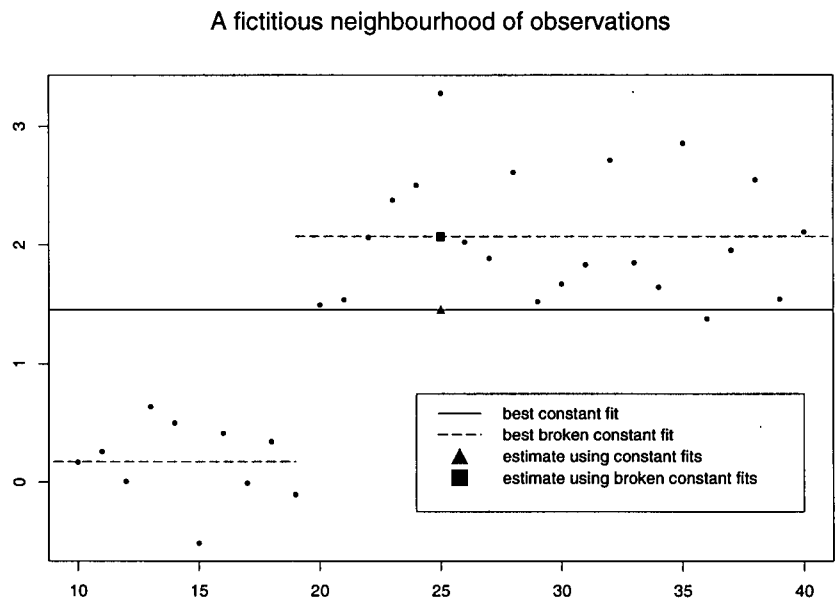


Figure 1.1: The best constant and broken constant fits and their corresponding estimates at the central point

discontinuity, where best is determined by a minimization of squared error criterion. Regardless of the form of local fits used, after determining the best local fit, the estimate of the regression function at the point of interest is taken to be the fitted value at that point.

An illustration should clarify. Suppose we have a data set of 100 observations and we are estimating the regression function at the 25th point. We must first choose the number of neighbouring observations to use in our estimate. For this example, we will use 15 observations to the left and 15 observations to the right of the 25th point, meaning observations 10 through 40. Figure 1.1 shows fictitious data for this situation. The best constant fit, which is just the mean of the observations, and the best broken constant

fit are shown in the figure. To find the best broken constant fit, we consider, for each design point in the neighbourhood, the broken constant fit that breaks at that point. Searching over all these fits, we find that the total squared error between the observed and fitted values is minimized when the breakpoint occurs at the 19th point, as pictured. The estimates at the 25th point corresponding to the best constant fit and best broken constant fit are also marked.

This illustration helps us to see that when there is a discontinuity in the neighbourhood of a point, it is reasonable to expect the best broken constant fit to break at, or near, the point of discontinuity. If so, the function estimate at points close to the discontinuity will only be influenced significantly by observations on the same side of it, thus preserving the edge. Because the result of our smoothing method is an estimate of the function that need not always be smooth, we will refer to it as the *not-so-smoother*.

A detailed description of the not-so-smoother follows in Chapter 2. Consistency of the estimator under general conditions is proven in Chapter 3. In Chapter 4 we demonstrate the performance of the not-so-smoother on simulated as well as real data sets. Chapter 5 contains extensions of the smoother to include the use of local linear, rather than constant, fits and the use of a test to allow the estimator to "break" only if the data provide sufficient evidence of a discontinuity. Finally, concluding remarks and suggestions for further work are found in Chapter 6.

Chapter 2

The Not-So-Smoother

Before introducing the not-so-smoother, we must first define the framework under which we are working. Let f denote the regression function, or signal, being estimated. We assume that f is a bounded, real-valued function on the interval (a, b) with a finite number of unknown discontinuities. Let (t_{nk}, X_{nk}), $k = 0, 1, \ldots, n$, denote the data, where each t_{nk} is a design point in the interval (a, b) and each X_{nk} is the observed regression function at the point t_{nk}.

The model can be expressed as

$$X_{nk} = f(t_{nk}) + \varepsilon_{nk}, \quad k = 0, 1, \ldots, n$$

where $a < t_{n0} < t_{n1} < \cdots < t_{nn} < b$ and the error terms, $\{\varepsilon_{nk}\}$, are independent and identically distributed with mean 0 and finite variance σ^2.

A smoothed estimate of the regression function, f, at each of the design points is calculated according to the following algorithm.

First choose a non-zero *bandwidth*, R_n, such that in estimating f at the point t_{nk} the nearest $2R_n + 1$ observations, namely $\{X_{n,k-R_n}, \ldots, X_{nk}, \ldots, X_{n,k+R_n}\}$, are used. We will refer to the interval $[t_{n,k-R_n}\ ,\ t_{n,k+R_n}]$ as the *neighbourhood* of t_{nk}.

For each design point t_{nk}, $k \in \{0, 1, \ldots, n\}$, define the local *breakpoint*, $\hat{I}_{nk}$, to be the argument that minimizes the function $H_{nk}(J)$ over all $J \in \{-R_n, \ldots, 0, \ldots, R_n - 1\}$, where

$$H_{nk}(J) = \frac{1}{R_n}\left(\sum_{j=-R_n}^{J} \left(X_{n,k+j} - \overline{X}_{-R_n:J}\right)^2 + \sum_{j=J+1}^{R_n} \left(X_{n,k+j} - \overline{X}_{J+1:R_n}\right)^2 \right) \qquad (2.1)$$

and $\overline{X}_{l:m}$ is defined as

$$\overline{X}_{l:m} = \frac{1}{m-l+1}\sum_{j=l}^{m} X_{n,k+j} \, .$$

Then define the estimate of $f(t_{nk})$ to be

$$\hat{f}(t_{nk}) = \hat{f}(t_{nk}, \hat{I}_{nk})$$

where

$$\hat{f}(t_{nk}, \hat{I}_{nk}) = \frac{1}{|S|}\sum_{j \in S} X_{n,k+j} \tag{2.2}$$

and

$$S = \begin{cases} \{-R_n, \ldots, \hat{I}_{nk}\} & \text{if } \hat{I}_{nk} \geq 0 \\ \{\hat{I}_{nk}+1, \ldots, R_n\} & \text{if } \hat{I}_{nk} < 0. \end{cases} \tag{2.3}$$

Note that $|S|$ denotes the number of elements in the set S.

Although the notation may seem daunting, the idea is really quite simple. Suppose we are trying to get a smoothed estimate of f at the point t_{nk}, then consider only the $2R_n+1$ neighbouring observations, which we have denoted $\{X_{n,k-R_n}, \ldots, X_{nk}, \ldots, X_{n,k+R_n}\}$. We want to find the broken constant function that fits the data "best", where best is based on the minimization of squared errors.

Specifically, fix $J \in \{-R_n, \ldots, 0, \ldots, R_n - 1\}$ so that J divides the data in the neighbourhood into two subsets, $S_1 = \{X_{n,k-R_n}, \ldots, X_{n,k+J}\}$ and $S_2 = \{X_{n,k+J+1}, \ldots, X_{n,k+R_n}\}$. The constant line through the observations in S_1 which minimizes the sum of squared errors is the mean of the observations; similarly for S_2. Let us denote the sum of squared errors from the observations in S_1 to their mean and S_2 to their mean as $SSE_{-R_n:J}$ and $SSE_{J+1:R_n}$ respectively. The value of J which minimizes the total squared error, $SSE_{-R_n:J} + SSE_{J+1:R_n}$, or equivalently $H_{nk}(J) = (SSE_{-R_n:J} + SSE_{J+1:R_n})/R_n$, is what we have called the (local) breakpoint. After having identified the best broken constant fit, the estimate of f at t_{nk} is simply taken to be the fitted value at that point.

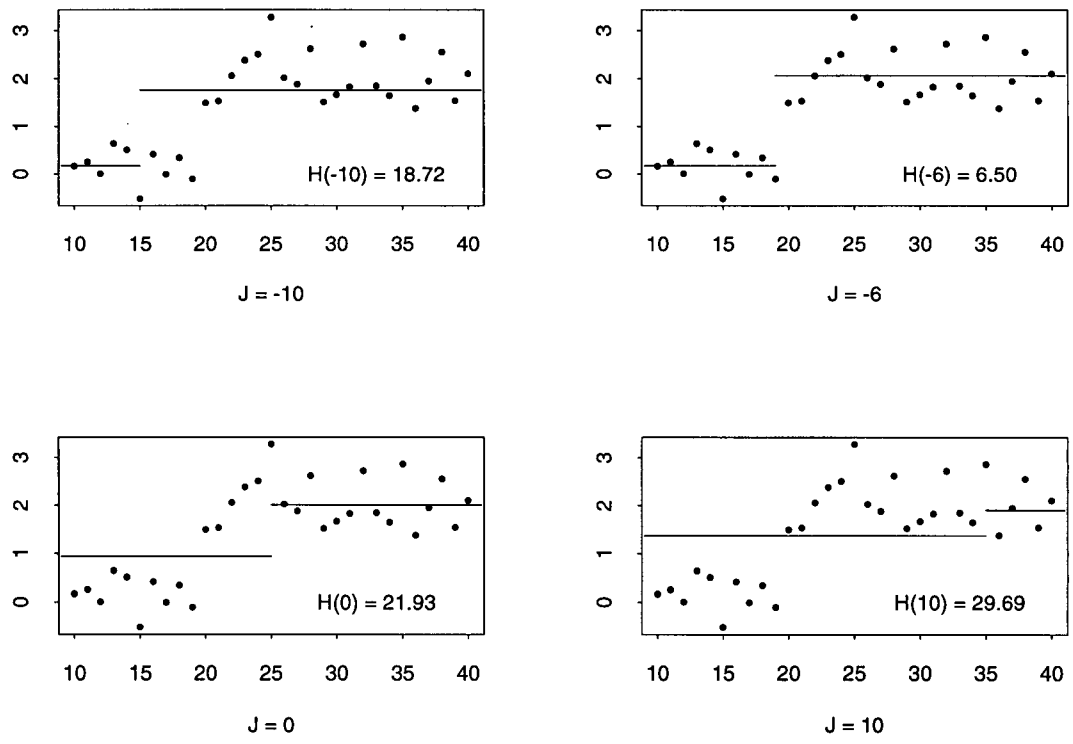


Figure 2.1: Examples of local broken constant fits for the fictitious neighbourhood in Figure 1.1

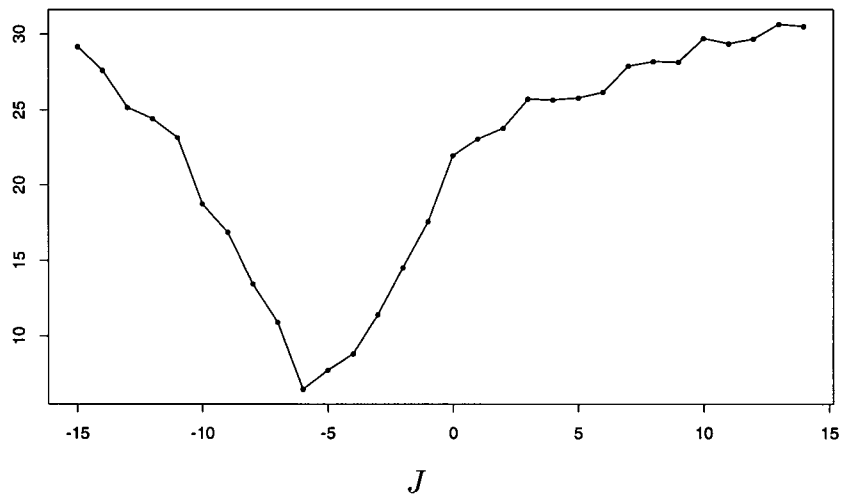


Figure 2.2: The function H corresponding to the fictitious neighbourhood in Figure 1.1

To illustrate this procedure, again consider the fictitious neighbourhood shown in Figure 1.1. For this example, the bandwidth is 15. Thus, there are 30 possible broken constant fits to consider, one corresponding to each value of $J \in \{-15, \ldots, 0, \ldots, 14\}$, where J divides the data in the neighbourhood into two subsets as described above. Broken constant fits for various values of J are shown in Figure 2.1. The function $H_{100,25}$ (recall $n = 100$ and we are estimating the 25th point), for which we will simply write H in this example, is calculated for each value of J. A plot of H is given in Figure 2.2. It is clear from this figure that H is minimized at $J = -6$, meaning the breakpoint is -6 (corresponding to the 19th point). Therefore, the estimate of the regression function at the 25th point is the mean of observation 20 through 40, which is 2.07.

Comments

1. Some modifications must be made for estimating points near the ends of the interval (a, b). If we are estimating f at the point t_{nk} where $k \in \{0, 1, \ldots, R_n - 1\}$, then there are not R_n design points to the left of t_{nk}. Similarly, for estimating f at t_{nk} where $k \in \{n - R_n + 1, \ldots, n - 1, n\}$, there are not R_n design points to the right of t_{nk}. In such cases, we simply replicate the data set to the right and left of itself, connecting the endpoints of the interval (a, b). In doing so, for t_{nk} near a, observation $X_{n,-j}$ becomes observation $X_{n,n-j+1}$, $j = 1, \ldots, R_n$, and for t_{nk} near b, observation $X_{n,n+j}$ becomes observation $X_{n,j-1}$, $j = 1, \ldots, R_n$. More concisely, every X_{nk} can be written as $X_{n,k\%(n+1)}$, where % denotes modulo arithmetic. Furthermore, $t_{n,-j}$ can be defined as $a - (b - t_{n,n+1-j})$, $j = 1, \ldots, R_n$, and $t_{n,n+j}$ as $b + (t_{n,j-1} - a)$, $j = 1, \ldots, R_n$. Although the location of the design points does not affect the estimated value of the function, it is important that all design points be well defined in order that the proofs in Chapter 3 hold near the ends of the interval (a, b). Note that with a usual moving average smoother, replicating the data set could result in biased estimates since the observations near the beginning of the interval need not be consistent with those near the end of the interval. However, using the not-so-smoother there is no such problem since we are just introducing another potential discontinuity which this method is designed to preserve.

2. We chose to use local piecewise constant fits as opposed to local piecewise linear fits, or any more complex piecewise function. This choice stems from the idea that if the width of the neighbourhood is small, then the regression function should not fluctuate too much within a neighbourhood and thus should be adequately approximated by constants. This will not always be true, as a later example will show, leading us to investigate the use of local piecewise linear fits in Chapter 5. The literature on local regression is unanimous to say that local linear and higher order fits perform better than local constant fits for numerous reasons. The example in Chapter 5 suggests that this is also the case for the

estimation of a discontinuous signal.

3. As mentioned previously, the not-so-smoother should detect discontinuities when they exist. If there is a jump within the neighbourhood of the point being estimated, then it is reasonable that the best broken constant fit will break at, or very near, the point of discontinuity. (In Chapter 4 we show this is true in practice and in the next chapter we prove that, under certain conditions, it is true asymptotically.) Therefore, if the point being estimated is to the left (right) of the discontinuity, then the estimate should only be influenced significantly by points to the left (right) of the discontinuity. As a result, the jump will not be smoothed away.

It will be convenient to rewrite the estimator given in equation (2.2) as follows. Since $X_{nk} = f(t_{nk}) + \varepsilon_{nk}$, we can write

$$\hat{f}(t_{nk}, \hat{I}_{nk}) = g_{nk}(t_{nk}, \hat{I}_{nk}) + \gamma_{nk}(\hat{I}_{nk}) \tag{2.4}$$

where

$$g_{nk}(t_{nk}, \hat{I}_{nk}) = \frac{1}{|S|} \sum_{j \in S} f(t_{n,k+j}), \tag{2.5}$$

$$\gamma_{nk}(\hat{I}_{nk}) = \frac{1}{|S|} \sum_{j \in S} \varepsilon_{n,k+j} \tag{2.6}$$

and S is defined in (2.3).

Thus g_{nk} can be thought of as the part of the estimate attributable to the signal whereas γ_{nk} is the part of the estimate attributable to noise in the observations.

Chapter 3

Consistency

This chapter is devoted to proving that the estimator, $\hat{f}$, as defined by (2.4) is consistent under general conditions. The conditions required are:

$(R1)$
$$\max_{i\in\{1,\ldots,n\}}(t_{ni}-t_{n,i-1})=O(1/n)\,,$$
$$t_{n0}-a=O(1/n) \quad \text{and} \quad b-t_{nn}=O(1/n)$$

$(R2)$
$$R_n \overset{n\to\infty}{\longrightarrow} \infty \quad \text{and} \quad \frac{R_n}{n} \overset{n\to\infty}{\longrightarrow} 0$$

$(R3)$ For every discontinuity point t_0, there exists $\delta_1, \delta_2 > 0$ and $M_1, M_2 > 0$ such that

(i)
$$|f(t)-f^-(t_0)| < M_1\,(t_0-t) \quad \forall\ 0<t_0-t<\delta_1$$

and

(ii)
$$|f(t)-f^+(t_0)| < M_2\,(t-t_0) \quad \forall\ 0<t-t_0<\delta_2$$

where $f^-(t_0)=\lim_{t\uparrow t_0} f(t)$ and $f^+(t_0)=\lim_{t\downarrow t_0} f(t)$ are assumed to exist. Furthermore, either $f(t_0)=f^-(t_0)$ or $f(t_0)=f^+(t_0)$.

$(R4)$
$$\mathrm{E}\varepsilon_i^4<\infty \qquad \forall\ i=1,\ldots,n$$

Condition $(R1)$ states that the distance between any two consecutive design points tends to zero at speed $1/n$ as the number of design points tends to infinity. For example,

if the design points are equally spaced over the interval $[0, 1]$, as is commonly the case, then $t_{ni} = i/n$, $i = 0, 1, \ldots, n$. Thus, the distance between any two consecutive points is $1/n$ which is clearly $O(1/n)$ as required.

The additional requirements in $(R1)$ that the left and right endpoints of the interval (a, b) be close to the first and last design points respectively are needed when proving consistency near the ends of the interval. Collectively, the conditions in $(R1)$ state that distance between any two consecutive design points, including those defined by replicating the data set (see Comment 1 on page 10), is $O(1/n)$.

Condition $(R2)$ states that the number of points used in estimating f at any given point must be large in absolute terms, but small relative to the total number of observations. For example, take $R_n = \sqrt{n}$.

An immediate consequence of conditions $(R1)$ and $(R2)$ is that the distance between a given design point and any other point within its neighbourhood tends to zero as n tends to infinity. Formally this can be stated as:

$(R2')$ Let $\mathcal{S}_n = \{-R_n, \ldots, 0, \ldots, R_n\}$. Then

$$\max_{j \in \mathcal{S}_n} |t_{n,k_n+j} - t_{nk_n}| \stackrel{n\to\infty}{\longrightarrow} 0.$$

The proof is simple.

$$\begin{aligned}
\max_{j \in \mathcal{S}_n} |t_{n,k_n+j} - t_{nk_n}| &< t_{n,k_n+R_n} - t_{n,k_n-R_n} \\
&= \sum_{i=-R_n}^{R_n} (t_{n,k_n+i} - t_{n,k_n+(i-1)}) \\
&= \sum_{i=-R_n}^{R_n} O(1/n) \quad \text{by } (R1) \\
&= (2R_n + 1)\, O(1/n) \stackrel{n\to\infty}{\longrightarrow} 0 \quad \text{by } (R2).
\end{aligned}$$

Condition $(R3)$ is a Lipschitz-type condition for the discontinuity points which requires that the underlying function be smooth on either side of a discontinuity. It also

states that f is either right or left continuous to exclude the case where f has a removable discontinuity, that is, $f^-(t_0) = f^+(t_0) \neq f(t_0)$. Although the results presented in this chapter are still true when there is a removable discontinuity, the proofs would need modified to include this case.

Finally, condition $(R4)$ states that the distribution of the error terms must have finite fourth moments. This is true for many distributions, including the commonly assumed normal distribution. The last two requirements, $(R3)$ and $(R4)$, will be needed to prove consistency of $\hat{f}$ at the discontinuity points.

Comment

Note that in the proof of $(R2')$, t_{nk_n} was written instead of t_{nk}; t_{nk} represents a fixed point whereas t_{nk_n} represents a sequence of points. We want to show consistency of $\hat{f}$ at every point t_0 in the interval (a, b). Because $\hat{f}$ is only defined at the design points and t_0 may not correspond to a design point, we must consider a sequence of design points, t_{nk_n}, which converges to t_0. Thus, throughout this chapter, k_n will be written wherever k was formerly used.

The proof that the not-so-smoother is consistent will be presented through a series of lemmas. First it is shown that the part of the estimate attributable to noise, γ_{nk_n} (refer to (2.6)), converges in probability to zero. Next, the term attributable to the signal, g_{nk_n} (refer to (2.5)), is considered and it is shown that the difference between g_{nk_n} and the true function f tends to zero at all points as n tends to infinity. Continuity points and discontinuity points will be considered separately. Finally, since $\hat{f} = g_{nk_n} + \gamma_{nk_n}$, it follows that $\left|\hat{f} - f\right|$ converges in probability to zero, or equivalently that $\hat{f}$ is a consistent estimator of f.

Notation

Before proceeding, some simplifying notation which will be used throughout this chapter is introduced. When considering design point t_{nk_n}, ε_j will be written in place of ε_{n,k_n+j} and X_j in place of X_{n,k_n+j}, for $j = -R_n, \ldots, R_n$. Thus it is important to keep in mind the dependence of these terms on k_n, the position of the current point of interest, and n, the total number of observations.

Lemma 1 *Assume condition (R1) holds. Then*

$$\max_{\hat{I}_{nk_n} \in \{-R_n, \ldots, R_n\}} \left| \gamma_{nk_n}(\hat{I}_{nk_n}) \right| \xrightarrow{\mathcal{P}} 0$$

where $\mathcal{P}$ denotes convergence in probability.

Proof

$$\begin{aligned}
\left| \gamma_{nk_n}(\hat{I}_{nk_n}) \right| &\leq \max \left\{ \frac{|\varepsilon_{-R_n} + \ldots + \varepsilon_0|}{R_n + 1}, \frac{|\varepsilon_{-R_n} + \ldots + \varepsilon_1|}{R_n + 2}, \ldots, \frac{|\varepsilon_{-R_n} + \ldots + \varepsilon_{R_n}|}{2R_n + 1}, \right. \\
&\quad \left. \frac{|\varepsilon_{-R_n+1} + \ldots + \varepsilon_{R_n}|}{2R_n}, \ldots, \frac{|\varepsilon_{-1} + \ldots + \varepsilon_{R_n}|}{R_n + 2}, \frac{|\varepsilon_0 + \ldots + \varepsilon_{R_n}|}{R_n + 1} \right\} \\
&\leq \frac{1}{R_n + 1} \max \{ |\varepsilon_{-R_n} + \ldots + \varepsilon_0|, \ldots, |\varepsilon_{-R_n} + \ldots + \varepsilon_{R_n}|, \\
&\quad |\varepsilon_{-R_n+1} + \ldots + \varepsilon_{R_n}|, \ldots, |\varepsilon_0 + \ldots + \varepsilon_{R_n}| \} \\
&\leq \frac{1}{R_n + 1} \max \{ |\varepsilon_{-R_n}|, |\varepsilon_{-R_n} + \varepsilon_{-R_n+1}|, \ldots, |\varepsilon_{-R_n} + \ldots + \varepsilon_{R_n}|, \\
&\quad |\varepsilon_{-R_n+1} + \ldots + \varepsilon_{R_n}|, \ldots, |\varepsilon_{R_n-1} + \varepsilon_{R_n}|, |\varepsilon_{R_n}| \} \\
&\leq \frac{1}{R_n + 1} \max \{ |\varepsilon_{-R_n}|, |\varepsilon_{-R_n} + \varepsilon_{-R_n+1}|, \ldots, |\varepsilon_{-R_n} + \ldots + \varepsilon_{R_n}| \} \\
&\quad + \frac{1}{R_n + 1} \max \{ |\varepsilon_{R_n}|, |\varepsilon_{R_n-1} + \varepsilon_{R_n}|, \ldots, |\varepsilon_{-R_n} + \ldots + \varepsilon_{R_n}| \} \qquad (3.1)
\end{aligned}$$

By Kolmogorov's Inequality (Chung, 1974), we know for all $\delta > 0$,

$$\begin{aligned}
&\mathrm{P} \left\{ \frac{1}{R_n + 1} \max \{ |\varepsilon_{-R_n}|, |\varepsilon_{-R_n} + \varepsilon_{-R_n+1}|, \ldots, |\varepsilon_{-R_n} + \ldots + \varepsilon_{R_n}| \} > \delta \right\} \\
&= \mathrm{P} \{ \max \{ |\varepsilon_{-R_n}|, |\varepsilon_{-R_n} + \varepsilon_{-R_n+1}|, \ldots, |\varepsilon_{-R_n} + \ldots + \varepsilon_{R_n}| \} > (R_n + 1)\, \delta \}
\end{aligned}$$

$$\leq \frac{\mathrm{Var}(\varepsilon_{-R_n} + \ldots + \varepsilon_{R_n})}{(R_n+1)^2\,\delta^2}$$
$$= \frac{(2R_n+1)\,\sigma^2}{(R_n+1)^2\,\delta^2} \overset{n\to\infty}{\longrightarrow} 0 \quad \text{since by } (R2) \quad R_n \overset{n\to\infty}{\longrightarrow} \infty.$$

By definition, this means that

$$\frac{1}{R_n+1}\max\left\{\left|\varepsilon_{-R_n}\right|, \left|\varepsilon_{-R_n}+\varepsilon_{-R_n+1}\right|, \ldots, \left|\varepsilon_{-R_n}+\ldots+\varepsilon_{R_n}\right|\right\} \xrightarrow{\mathcal{P}} 0. \tag{3.2}$$

Also, by symmetry,

$$\frac{1}{R_n+1}\max\left\{\left|\varepsilon_{R_n}\right|, \left|\varepsilon_{R_n-1}+\varepsilon_{R_n}\right|, \ldots, \left|\varepsilon_{-R_n}+\ldots+\varepsilon_{R_n}\right|\right\} \xrightarrow{\mathcal{P}} 0. \tag{3.3}$$

Now apply the results of (3.2) and (3.3) along with Slutzky's Theorem (Bickel and Doksum, 1977) to statement (3.1) to obtain

$$\begin{aligned}\left|\gamma_{nk_n}(\hat{I}_{nk_n})\right| \quad &\leq \quad \frac{1}{R_n+1}\max\left\{\left|\varepsilon_{-R_n}\right|, \left|\varepsilon_{-R_n}+\varepsilon_{-R_n+1}\right|, \ldots, \left|\varepsilon_{-R_n}+\ldots+\varepsilon_{R_n}\right|\right\} \\ &\quad + \frac{1}{R_n+1}\max\left\{\left|\varepsilon_{R_n}\right|, \left|\varepsilon_{R_n-1}+\varepsilon_{R_n}\right|, \ldots, \left|\varepsilon_{-R_n}+\ldots+\varepsilon_{R_n}\right|\right\} \\ &\xrightarrow{\mathcal{P}} \quad 0+0=0.\end{aligned}$$

□

Consider now convergence of g_{nk_n}. By condition $(R1)$, we know that for any point $t_0 \in (a,b)$ there exists a sequence t_{nk_n} which converges to t_0. Thus, for each t_0, we must show that $\left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) - f(t_0)\right| \overset{n\to\infty}{\longrightarrow} 0$. When f is continuous at t_0, the proof requires only conditions $(R1)$ and $(R2)$, whereas $(R3)$ and $(R4)$ are also needed when f is discontinuous at t_0.

The continuity points are considered first. Suppose we are estimating f at point t_{nk_n}, where $t_{nk_n} \overset{n\to\infty}{\longrightarrow} t_0$. The width of the neighbourhood around t_{nk_n} shrinks to zero asymptotically so that it will not contain a discontinuity when f is continuous at t_0. Thus

$f(t_{n,k_n-R_n}), \ldots, f(t_{nk_n}), \ldots, f(t_{n,k_n+R_n})$ are essentially equal, all converging to $f(t_0)$. Because $g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n})$ is an average of a subset of these terms, it must also converge to $f(t_0)$.

A rigorous proof is given in the following lemma.

Lemma 2 *Assume (R1) and (R2) hold. If* $t_{nk_n} \overset{n\to\infty}{\longrightarrow} t_0$ *and if* f *is continuous at* t_0*, then*

$$\left| g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) - f(t_0) \right| \overset{n\to\infty}{\longrightarrow} 0.$$

Proof

Let S be a subset of $\mathcal{S}_n = \{-R_n, \ldots, 0, \ldots, R_n\}$ and let $|S|$ denote the number of elements in S.

For any $k_n \in \{0, \ldots, n\}$,

$$\begin{aligned} \left| \frac{1}{|S|} \sum_{j\in S} f(t_{n,k_n+j}) - f(t_0) \right| &\leq \frac{1}{|S|} \sum_{j\in S} |f(t_{n,k_n+j}) - f(t_0)| \\ &\leq \max_{j\in S} |f(t_{n,k_n+j}) - f(t_0)| \\ &\leq \max_{j\in \mathcal{S}_n} |f(t_{n,k_n+j}) - f(t_0)| . \end{aligned} \tag{3.4}$$

Regardless of the estimate $\hat{I}_{nk_n}$, $g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n})$ is of the form $\sum_{j\in S} f(t_{n,k_n+j})/|S|$ where $S = \{-R_n, \ldots, \hat{I}_{nk_n}\}$ or $S = \{\hat{I}_{nk_n} + 1, \ldots, R_n\}$. Therefore the result of (3.4) can be applied to get

$$\left| g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) - f(t_0) \right| \leq \max_{j\in \mathcal{S}_n} |f(t_{n,k_n+j}) - f(t_0)| . \tag{3.5}$$

Now, because t_0 is a continuity point of f, we know that for all t such that $|t - t_0| \overset{n\to\infty}{\longrightarrow} 0$, $|f(t) - f(t_0)| \overset{n\to\infty}{\longrightarrow} 0$.

By $(R2')$, we know that $\max_{j\in\mathcal{S}_n} |t_{n,k_n+j} - t_{nk_n}| \overset{n\to\infty}{\longrightarrow} 0$.

Combine this with the fact that $|t_{nk_n} - t_0| \overset{n\to\infty}{\longrightarrow} 0$ to get

$$\begin{aligned} \max_{j\in\mathcal{S}_n} |t_{n,k_n+j} - t_0| &= \max_{j\in\mathcal{S}_n} |t_{n,k_n+j} - t_{nk_n} + t_{nk_n} - t_0| \\ &\leq \max_{j\in\mathcal{S}_n} |t_{n,k_n+j} - t_{nk_n}| + |t_{nk_n} - t_0| \overset{n\to\infty}{\longrightarrow} 0. \end{aligned}$$

Therefore,

$$\max_{j \in \mathcal{S}_n} |f(t_{n,k_n+j}) - f(t_0)| \xrightarrow{n\to\infty} 0.$$

So we can conclude from inequality (3.5) that

$$\left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) - f(t_0)\right| \xrightarrow{n\to\infty} 0. \tag{3.6}$$

□

Note that $f(t_0)$ can be replaced by $f(t_{nk_n})$ in the above lemma for a slightly different statement. The reason this works is as follows:

$$\begin{aligned}\left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) - f(t_{nk_n})\right| &= \left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) - f(t_0) + f(t_0) - f(t_{nk_n})\right| \\ &\leq \left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) - f(t_0)\right| + |f(t_0) - f(t_{nk_n})|.\end{aligned}$$

Because f is continuous at t_0 and $t_{nk_n} \xrightarrow{n\to\infty} t_0$, we know that $|f(t_0) - f(t_{nk_n})| \xrightarrow{n\to\infty} 0$. Using this fact and expression (3.6), we can conclude that

$$\left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) - f(t_{nk_n})\right| \xrightarrow{n\to\infty} 0.$$

To prove convergence at points where f is discontinuous is somewhat more involved. Suppose f is discontinuous at t_0. Ideally, we would like to show that $\left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) - f(t_0)\right| \xrightarrow{\mathcal{P}} 0$. This is unrealistic as we do not usually know if f is right or left continuous at t_0, and even if we did, the most we can expect is that g_{nk_n} will converge to either $f^-(t_0)$ or $f^+(t_0)$. In Lemma 4, this last statement is shown to be true.

In proving Lemma 4, we will need to use a result to be presented in Lemma 3. We saw in the proof of Lemma 2 that the location of the estimated breakpoint is irrelevant asymptotically when f is continuous at the point being estimated. This is not the case for discontinuity points; rather we will need to show that the estimated breakpoint is

sufficiently close to the true location of the jump. To do so requires conditions ($R3$) and ($R4$). This is the content of Lemma 3.

For simplicity, we assume in both Lemmas 3 and 4 that t_0 is a discontinuity point and that it corresponds to some design point; that is, $t_0 = t_{nk_n}$ for some $k_n \in \{0, 1, \ldots, n\}$. Of course, this need not be true; we may only have $t_{nk_n} \overset{n\to\infty}{\longrightarrow} t_0$. Lemma 4 will still be valid; however, the proof becomes much more involved since t_0 can be anywhere inside, or even outside, of the neighbourhood around t_{nk_n}. Numerous cases must then be considered and Lemma 3 needs to be slightly revised.

The essence of the proofs for Lemmas 3 and 4 in the more general setting is similar to when we assume $t_0 = t_{nk_n}$. However, the proofs become lengthy and repetitive, and thus only the more restricted statements are proven here.

Lemma 3 *Assume (R1) - (R4) hold. Consider estimating f at design point t_{nk_n} where $t_{nk_n} = t_0$ for some $k_n \in \{0, 1, \ldots, n\}$ and f is discontinuous at t_0. Then*

$$\left| \frac{\hat{I}_{nk_n}}{R_n} \right| \xrightarrow{\mathcal{P}} 0.$$

Proof

We must show that

$$\mathrm{P}\left\{ \left| \frac{\hat{I}_{nk_n}}{R_n} \right| > \varepsilon \right\} \overset{n\to\infty}{\longrightarrow} 0 \quad \forall\, 0 < \varepsilon < 1.$$

To do so, we will bound the probability by an expression which is easier to work with.

$$\begin{aligned}
\mathrm{P}\left\{ \left| \frac{\hat{I}_{nk_n}}{R_n} \right| > \varepsilon \right\} &= \mathrm{P}\left\{ \left| \hat{I}_{nk_n} \right| > \varepsilon R_n \right\} \\
&= \mathrm{P}\left\{ \bigcup_{|J|>\varepsilon R_n} \{ H_{nk_n}(J) \le H_{nk_n}(K) \quad \forall\, J \neq K \} \right\} \\
&\le \mathrm{P}\left\{ \bigcup_{|J|>\varepsilon R_n} \{ H_{nk_n}(J) \le H_{nk_n}(0) \} \right\}
\end{aligned}$$

$$\leq \sum_{|J|>\varepsilon R_n} \mathrm{P}\left\{H_{nk_n}(J) \leq H_{nk_n}(0)\right\} \tag{3.7}$$

Case 1: $\varepsilon R_n < J < R_n$.

We need to show that

$$\sum_{J>\varepsilon R_n}^{R_n-1} \mathrm{P}\left\{H_{nk_n}(J) < H_{nk_n}(0)\right\} \overset{n\to\infty}{\longrightarrow} 0.$$

The following notation will be convenient: let V_i, $i = 1, \ldots, n$, be any random variables. Then,

$$\overline{V}_{k:l} \equiv \frac{1}{l-k+1} \sum_{j=k}^{l} V_j$$

for any $1 \leq k \leq l \leq n$.

It can be shown that

$$\begin{aligned} \mathrm{P}\left\{H_{nk_n}(J) < H_{nk_n}(0)\right\} &= \mathrm{P}\Big\{R_n(R_n+1)\left(\overline{X}_{-R_n:0} - \overline{X}_{1:J}\right)^2 < \\ &\qquad (J+R_n+1)(R_n-J)\left(\overline{X}_{1:J} - \overline{X}_{J+1:R_n}\right)^2\Big\} \end{aligned} \tag{3.8}$$

through the series of steps that follow.

$$\begin{aligned} H_{nk_n}(J) &= \sum_{j=-R_n}^{J}\left(X_j - \overline{X}_{-R_n:J}\right)^2 + \sum_{j=J+1}^{R_n}\left(X_j - \overline{X}_{J+1:R_n}\right)^2 \\ &= \sum_{j=-R_n}^{0}\left(X_j - \overline{X}_{-R_n:J}\right)^2 + \sum_{j=1}^{J}\left(X_j - \overline{X}_{-R_n:J}\right)^2 + \sum_{j=J+1}^{R_n}\left(X_j - \overline{X}_{J+1:R_n}\right)^2 \\ &= \sum_{j=-R_n}^{0}\left(X_j - \overline{X}_{-R_n:0} + \overline{X}_{-R_n:0} - \overline{X}_{-R_n:J}\right)^2 \\ &\quad + \sum_{j=1}^{J}\left(X_j - \overline{X}_{1:J} + \overline{X}_{1:J} - \overline{X}_{-R_n:J}\right)^2 + \sum_{j=J+1}^{R_n}\left(X_j - \overline{X}_{J+1:R_n}\right)^2 \\ &= \sum_{j=-R_n}^{0}\left(X_j - \overline{X}_{-R_n:0}\right)^2 + (R_n+1)\left(\overline{X}_{-R_n:J} - \overline{X}_{-R_n:0}\right)^2 \\ &\quad + \sum_{j=1}^{J}\left(X_j - \overline{X}_{1:J}\right)^2 + J\left(\overline{X}_{-R_n:J} - \overline{X}_{1:J}\right)^2 + \sum_{j=J+1}^{R_n}\left(X_j - \overline{X}_{J+1:R_n}\right)^2 \end{aligned} \tag{3.9}$$

Now,

$$\begin{aligned}
&\sum_{j=1}^{J}\left(X_j-\overline{X}_{1:J}\right)^2+\sum_{j=J+1}^{R_n}\left(X_j-\overline{X}_{J+1:R_n}\right)^2\\
&=\sum_{j=1}^{J}\left((X_j-\overline{X}_{1:R_n})-(\overline{X}_{1:J}-\overline{X}_{1:R_n})\right)^2+\sum_{j=J+1}^{R_n}\left((X_j-\overline{X}_{1:R_n})-(\overline{X}_{J+1:R_n}-\overline{X}_{1:R_n})\right)^2\\
&=\sum_{j=1}^{R_n}\left(X_j-\overline{X}_{1:R_n}\right)^2-J\left(\overline{X}_{1:J}-\overline{X}_{1:R_n}\right)^2-(R_n-J)\left(\overline{X}_{J+1:R_n}-\overline{X}_{1:R_n}\right)^2
\end{aligned}$$

where the last equality follows by expanding the squares and gathering terms.

Substituting this expression into equation (3.9) and using the fact that

$$H_{nk_n}(0)=\sum_{j=-R_n}^{0}\left(X_j-\overline{X}_{-R_n:0}\right)^2+\sum_{j=1}^{R_n}\left(X_j-\overline{X}_{1:R_n}\right)^2$$

gives

$$\begin{aligned}
H_{nk_n}(J) &= H_{nk_n}(0)+(R_n+1)\left(\overline{X}_{-R_n:J}-\overline{X}_{-R_n:0}\right)^2+J\left(\overline{X}_{-R_n:J}-\overline{X}_{1:J}\right)^2\\
&\quad -J\left(\overline{X}_{1:J}-\overline{X}_{1:R_n}\right)^2-(R_n-J)\left(\overline{X}_{J+1:R_n}-\overline{X}_{1:R_n}\right)^2. \qquad (3.10)
\end{aligned}$$

Therefore

$$\begin{aligned}
&\mathrm{P}\left\{H_{nk_n}(J)<H_{nk_n}(0)\right\}\\
&\quad=\mathrm{P}\left\{(R_n+1)\left(\overline{X}_{-R_n:J}-\overline{X}_{-R_n:0}\right)^2+J\left(\overline{X}_{-R_n:J}-\overline{X}_{1:J}\right)^2\right.\\
&\qquad\left.<J\left(\overline{X}_{1:J}-\overline{X}_{1:R_n}\right)^2+(R_n-J)\left(\overline{X}_{J+1:R_n}-\overline{X}_{1:R_n}\right)^2\right\}. \qquad (3.11)
\end{aligned}$$

More manipulations are needed to get this in the form of equation (3.8). Consider each term in the probability statement on the right-hand side of (3.11) separately.

$$\begin{aligned}
&(R_n+1)\left(\overline{X}_{-R_n:J}-\overline{X}_{-R_n:0}\right)^2\\
&=(R_n+1)\left(\frac{((R_n+1)-(J+R_n+1))\left(X_{-R_n}+\ldots+X_0\right)+(R_n+1)\left(X_1+\ldots+X_J\right)}{(J+R_n+1)(R_n+1)}\right)^2\\
&=\frac{(R_n+1)\,J^2}{(J+R_n+1)^2}\left(\overline{X}_{-R_n:0}-\overline{X}_{1:J}\right)^2
\end{aligned}$$

Similarly,

$$\begin{aligned} J\left(\overline{X}_{-R_n:J} - \overline{X}_{1:J}\right)^2 &= \frac{J\,(R_n+1)^2}{(J+R_n+1)^2}\left(\overline{X}_{-R_n:0} - \overline{X}_{1:J}\right)^2 \\ J\left(\overline{X}_{1:J} - \overline{X}_{1:R_n}\right)^2 &= \frac{J\,(R_n-J)^2}{R_n^2}\left(\overline{X}_{1:J} - \overline{X}_{J+1:R_n}\right)^2 \\ (R_n-J)\left(\overline{X}_{J+1:R_n} - \overline{X}_{1:R_n}\right)^2 &= \frac{(R_n-J)\,J^2}{R_n^2}\left(\overline{X}_{1:J} - \overline{X}_{J+1:R_n}\right)^2. \end{aligned}$$

Substituting these equalities into equation (3.11) gives

$$\begin{aligned} &\mathrm{P}\left\{H_{nk_n}(J) < H_{nk_n}(0)\right\} \\ &= \mathrm{P}\left\{\frac{J\,(R_n+1)}{(J+R_n+1)}\left(\overline{X}_{-R_n:0} - \overline{X}_{1:J}\right)^2 < \frac{J\,(R_n-J)}{R_n}\left(\overline{X}_{1:J} - \overline{X}_{J+1:R_n}\right)^2\right\} \\ &= \mathrm{P}\left\{R_n(R_n+1)\left(\overline{X}_{-R_n:0} - \overline{X}_{1:J}\right)^2 < (J+R_n+1)(R_n-J)\left(\overline{X}_{1:J} - \overline{X}_{J+1:R_n}\right)^2\right\} \end{aligned}$$

which we recognize as equation (3.8).

Thus,

$$\begin{aligned} &\mathrm{P}\left\{H_{nk_n}(J) < H_{nk_n}(0)\right\} \\ &= \mathrm{P}\left\{\left(\overline{X}_{-R_n:0} - \overline{X}_{1:J}\right)^2 < \left(1+\frac{J}{R_n+1}\right)\left(1-\frac{J}{R_n}\right)\left(\overline{X}_{1:J} - \overline{X}_{J+1:R_n}\right)^2\right\} \\ &\le \mathrm{P}\left\{\left(\overline{X}_{-R_n:0} - \overline{X}_{1:J}\right)^2 < \left(1-\left(\frac{J}{R_n}\right)^2\right)\left(\overline{X}_{1:J} - \overline{X}_{J+1:R_n}\right)^2\right\} \\ &= \mathrm{P}\left\{\left|\overline{X}_{-R_n:0} - \overline{X}_{1:J}\right| < m\left|\overline{X}_{1:J} - \overline{X}_{J+1:R_n}\right|\right\} \\ &\qquad\qquad \text{where } m = \sqrt{1-(J/R_n)^2} \\ &\le \mathrm{P}\left\{\left|\mathrm{E}\overline{X}_{-R_n:0} - \mathrm{E}\overline{X}_{1:J}\right| - \left|\overline{X}_{-R_n:0} - \mathrm{E}\overline{X}_{-R_n:0}\right| - \left|\overline{X}_{1:J} - \mathrm{E}\overline{X}_{1:J}\right|\right. \\ &\qquad \left. < m\left(\left|\overline{X}_{J+1:R_n} - \mathrm{E}\overline{X}_{J+1:R_n}\right| + \left|\overline{X}_{1:J} - \mathrm{E}\overline{X}_{1:J}\right| + \left|\mathrm{E}\overline{X}_{J+1:R_n} - \mathrm{E}\overline{X}_{1:J}\right|\right)\right\} \\ &\qquad\qquad \text{using the fact that } |a|-|b|-|c| \le |a+b+c| \le |a|+|b|+|c| \\ &= \mathrm{P}\left\{(1+m)\left|\overline{X}_{1:J} - \mathrm{E}\overline{X}_{1:J}\right| + m\left|\overline{X}_{J+1:R_n} - \mathrm{E}\overline{X}_{J+1:R_n}\right| + \left|\overline{X}_{-R_n:0} - \mathrm{E}\overline{X}_{-R_n:0}\right|\right. \\ &\qquad \left. > \left|\mathrm{E}\overline{X}_{-R_n:0} - \mathrm{E}\overline{X}_{1:J}\right| - m\left|\mathrm{E}\overline{X}_{J+1:R_n} - \mathrm{E}\overline{X}_{1:J}\right|\right\} \\ &\le \mathrm{P}\left\{(1+m)\left|\overline{X}_{1:J} - \mathrm{E}\overline{X}_{1:J}\right| > 1/3\left(\left|\mathrm{E}\overline{X}_{-R_n:0} - \mathrm{E}\overline{X}_{1:J}\right| - m\left|\mathrm{E}\overline{X}_{J+1:R_n} - \mathrm{E}\overline{X}_{1:J}\right|\right)\right\} \end{aligned}$$

$$+\mathrm{P}\left\{m\left|\overline{X}_{J+1:R_n}-\mathrm{E}\overline{X}_{J+1:R_n}\right|>1/3\left(\left|\mathrm{E}\overline{X}_{-R_n:0}-\mathrm{E}\overline{X}_{1:J}\right|-m\left|\mathrm{E}\overline{X}_{J+1:R_n}-\mathrm{E}\overline{X}_{1:J}\right|\right)\right\}$$
$$+\mathrm{P}\left\{\left|\overline{X}_{-R_n:0}-\mathrm{E}\overline{X}_{-R_n:0}\right|>1/3\left(\left|\mathrm{E}\overline{X}_{-R_n:0}-\mathrm{E}\overline{X}_{1:J}\right|-m\left|\mathrm{E}\overline{X}_{J+1:R_n}-\mathrm{E}\overline{X}_{1:J}\right|\right)\right\}.$$

To simplify the notation, define

$$\begin{aligned}\mu_1 &\equiv \mathrm{E}(\overline{X}_{-R_n:0})=\frac{1}{R_n+1}\sum_{j=-R_n}^{0}f(t_{n,k_n+j})\\ \mu_2 &\equiv \mathrm{E}(\overline{X}_{1:J})=\frac{1}{J}\sum_{j=1}^{J}f(t_{n,k_n+j})\\ \mu_3 &\equiv \mathrm{E}(\overline{X}_{J+1:R_n})=\frac{1}{R_n-J}\sum_{j=J+1}^{R_n}f(t_{n,k_n+j}).\end{aligned}$$

Thus, we have

$$\begin{aligned}&\mathrm{P}\left\{H_{nk_n}(J)<H_{nk_n}(0)\right\}\\ &\quad\leq\ \mathrm{P}\left\{(1+m)\left|\overline{X}_{1:J}-\mu_2\right|>1/3\left(|\mu_1-\mu_2|-m\,|\mu_3-\mu_2|\right)\right\}\\ &\qquad+\mathrm{P}\left\{m\left|\overline{X}_{J+1:R_n}-\mu_3\right|>1/3\left(|\mu_1-\mu_2|-m\,|\mu_3-\mu_2|\right)\right\}\\ &\qquad+\mathrm{P}\left\{\left|\overline{X}_{-R_n:0}-\mu_1\right|>1/3\left(|\mu_1-\mu_2|-m\,|\mu_3-\mu_2|\right)\right\}.\end{aligned}\tag{3.12}$$

Recall that we need to show that

$$\sum_{J>\varepsilon R_n}^{R_n-1}\mathrm{P}\left\{H_{nk_n}(J)<H_{nk_n}(0)\right\}\ \stackrel{n\to\infty}{\longrightarrow}\ 0.$$

To do so, we consider each probability statement on the right-hand side of (3.12) separately. We apply to each a generalization of Markov's inequality (Bickel and Doksum, 1977) and also use the the facts that $m=\sqrt{1-(J/R_n)^2}<\sqrt{1-\varepsilon^2}$, since $J>\varepsilon R_n$, and

$$\mathrm{E}\left(\overline{X}_{k:l}-\mathrm{E}\overline{X}_{k:l}\right)^4=\mathrm{E}(\overline{\varepsilon}_{k:l})^4=\frac{\mathrm{E}\varepsilon_1^4-2\sigma^4}{(l-k+1)^3}+\frac{2\sigma^4}{(l-k+1)^2}.$$

In deriving this expression, we used the fact that the ε_i's are independent and identically distributed with mean 0 and variance σ^2.

$$\begin{aligned}
&\mathrm{P}\left\{(1+m)\left|\overline{X}_{1:J}-\mu_2\right| > 1/3\left(|\mu_1-\mu_2| - m\,|\mu_3-\mu_2|\right)\right\} \\
&\quad\leq \frac{(3(1+m))^4\,\mathrm{E}\left(\overline{X}_{1:J}-\mu_2\right)^4}{\left(|\mu_1-\mu_2| - m\,|\mu_3-\mu_2|\right)^4} \\
&\quad\leq \frac{(3(1+\sqrt{1-\varepsilon^2}))^4\,\mathrm{E}\left(\overline{X}_{1:J}-\mu_2\right)^4}{\left(|\mu_1-\mu_2| - \sqrt{1-\varepsilon^2}\,|\mu_3-\mu_2|\right)^4} \\
&\quad= \frac{(3(1+\sqrt{1-\varepsilon^2}))^4}{\left(|\mu_1-\mu_2| - \sqrt{1-\varepsilon^2}\,|\mu_3-\mu_2|\right)^4}\left[\frac{\mathrm{E}\varepsilon_1^4-2\sigma^4}{J^3}+\frac{2\sigma^4}{J^2}\right] \qquad (3.13)
\end{aligned}$$

$$\begin{aligned}
&\mathrm{P}\left\{m\left|\overline{X}_{J+1:R_n}-\mu_3\right| > 1/3\left(|\mu_1-\mu_2| - m\,|\mu_3-\mu_2|\right)\right\} \\
&\quad\leq \frac{(3m)^4\,\mathrm{E}\left(\overline{X}_{J+1:R_n}-\mu_3\right)^4}{\left(|\mu_1-\mu_2| - m\,|\mu_3-\mu_2|\right)^4} \\
&\quad= \frac{(3m)^4}{\left(|\mu_1-\mu_2| - \sqrt{1-\varepsilon^2}\,|\mu_3-\mu_2|\right)^4}\left[\frac{\mathrm{E}\varepsilon_1^4-2\sigma^4}{(R_n-J)^3}+\frac{2\sigma^4}{(R_n-J)^2}\right] \qquad (3.14)
\end{aligned}$$

$$\begin{aligned}
&\mathrm{P}\left\{\left|\overline{X}_{-R_n:0}-\mu_1\right| > 1/3\left(|\mu_1-\mu_2| - m\,|\mu_3-\mu_2|\right)\right\} \\
&\quad\leq \frac{3^4\,\mathrm{E}\left(\overline{X}_{-R_n:0}-\mu_1\right)^4}{\left(|\mu_1-\mu_2| - m\,|\mu_3-\mu_2|\right)^4} \\
&\quad\leq \frac{3^4\,\mathrm{E}\left(\overline{X}_{-R_n:0}-\mu_1\right)^4}{\left(|\mu_1-\mu_2| - \sqrt{1-\varepsilon^2}\,|\mu_3-\mu_2|\right)^4} \\
&\quad= \frac{3^4}{\left(|\mu_1-\mu_2| - \sqrt{1-\varepsilon^2}\,|\mu_3-\mu_2|\right)^4}\left[\frac{\mathrm{E}\varepsilon_1^4-2\sigma^4}{(R_n+1)^3}+\frac{2\sigma^4}{(R_n+1)^2}\right] \qquad (3.15)
\end{aligned}$$

Assume without loss of generality that f is left continuous.[1] Then, by $(R3)$ (i), there exists $\delta_1 > 0$ and $M_1 > 0$ such that

$$|f(t_{n,k_n+j}) - f^-(t_0)| < M_1(t_0 - t_{n,k_n+j}) \qquad (3.16)$$

[1]Note that the function need not be left continuous. If f is right continuous then the proof will follow in the same manner except back at equation (3.8), we would have split into the terms $\overline{X}_{-R_n:-1}$, $\overline{X}_{0:J}$, and $\overline{X}_{J+1:R_n}$.

whenever $0 \leq t_0 - t_{n,k_n+j} < \delta_1$. Recalling that $t_0 = t_{nk_n}$ and using $(R2')$, we know $0 \leq \max_{j\in\{-R_n,\ldots,0\}} t_0 - t_{n,k_n+j} \overset{n\to\infty}{\longrightarrow} 0$. By (3.16),

$$\max_{j\in\{-R_n,\ldots,0\}} |f(t_{n,k_n+j}) - f^-(t_0)| \overset{n\to\infty}{\longrightarrow} 0. \tag{3.17}$$

Analagously, using (R3) (ii) and $(R2')$,

$$\max_{j\in\{1,\ldots,R_n\}} |f(t_{n,k_n+j}) - f^+(t_0)| \overset{n\to\infty}{\longrightarrow} 0. \tag{3.18}$$

Therefore,

$$\begin{aligned}
|\mu_1 - f^-(t_0)| &= \left|\frac{1}{R_n+1}\sum_{j=-R_n}^{0} f(t_{n,k_n+j}) - f^-(t_0)\right| \overset{n\to\infty}{\longrightarrow} 0 \\
|\mu_2 - f^+(t_0)| &= \left|\frac{1}{J}\sum_{j=1}^{J} f(t_{n,k_n+j}) - f^+(t_0)\right| \overset{n\to\infty}{\longrightarrow} 0 \quad \forall J \\
|\mu_3 - f^+(t_0)| &= \left|\frac{1}{R_n-J}\sum_{j=J+1}^{R_n} f(t_{n,k_n+j}) - f^+(t_0)\right| \overset{n\to\infty}{\longrightarrow} 0 \quad \forall J.
\end{aligned}$$

Thus, for all $\varepsilon > 0$, there exists N_0 such that if $n \geq N_0$ then $|\mu_1 - f^-(t_0)| < \varepsilon$. Take $\varepsilon = \delta/4$ where δ is defined as

$$\delta \equiv |f^-(t_0) - f^+(t_0)|,$$

which is a positive constant.

Then there exists N_1 such that if $n \geq N_1$ then $|\mu_1 - f^-(t_0)| < \delta/4$. Similarly, there exists N_2 such that if $n \geq N_2$ then $|\mu_2 - f^+(t_0)| < \delta/4$ for all J, and there exists N_3 such that if $n \geq N_3$ then $|\mu_3 - f^+(t_0)| < \delta/4$ for all J.

Let $N = \max\{N_1, N_2, N_3\}$ and we have, for all $n \geq N$,

$$|\mu_1 - f^-(t_0)| < \delta/4 \tag{3.19}$$

$$|\mu_2 - f^+(t_0)| < \delta/4 \quad \forall J \tag{3.20}$$

$$|\mu_3 - f^+(t_0)| < \delta/4 \quad \forall J. \tag{3.21}$$

So, for all J and for all $n \geq N$,

$$\begin{aligned}
|\mu_1 - \mu_2| &= |(f^-(t_0) + \mu_1 - f^-(t_0)) - (f^+(t_0) + \mu_2 - f^+(t_0))| \\
&= |(f^-(t_0) - f^+(t_0)) + (\mu_1 - f^-(t_0)) + (f^+(t_0) - \mu_2)| \\
&\geq |f^-(t_0) - f^+(t_0)| - |\mu_1 - f^-(t_0)| - |\mu_2 - f^+(t_0)| \\
&\geq \delta - \delta/4 - \delta/4 \qquad \text{using (3.19) and (3.20)} \\
&= \delta/2 \qquad (3.22)
\end{aligned}$$

and

$$\begin{aligned}
|\mu_3 - \mu_2| &= |(\mu_3 - f^+(t_0)) - (\mu_2 - f^+(t_0))| \\
&\leq |\mu_3 - f^+(t_0)| + |\mu_2 - f^+(t_0)| \\
&\leq \delta/4 + \delta/4 \qquad \text{using (3.20) and (3.21)} \\
&= \delta/2. \qquad (3.23)
\end{aligned}$$

Therefore, using (3.22) and (3.23) along with (3.13) gives, for all $n \geq N$,

$$\begin{aligned}
&\sum_{J>\varepsilon R_n}^{R_n-1} \mathrm{P}\left\{(1+m)\left|\overline{X}_{1:J} - \mu_2\right| > 1/3\left(|\mu_1 - \mu_2| - m\,|\mu_3 - \mu_2|\right)\right\} \\
&\quad\leq \frac{\left(3(1+\sqrt{1-\varepsilon^2})\right)^4}{\left(\delta/2(1-\sqrt{1-\varepsilon^2})\right)^4} \sum_{J>\varepsilon R_n}^{R_n-1} \left[\frac{\mathrm{E}\varepsilon_1^4 - 2\sigma^4}{J^3} + \frac{2\sigma^4}{J^2}\right] \\
&\quad\leq \frac{\left(3(1+\sqrt{1-\varepsilon^2})\right)^4}{\left(\delta/2(1-\sqrt{1-\varepsilon^2})\right)^4} R_n(1-\varepsilon)\left[\frac{\mathrm{E}\varepsilon_1^4 - 2\sigma^4}{(\varepsilon R_n)^3} + \frac{2\sigma^4}{(\varepsilon R_n)^2}\right] \\
&\quad= \frac{\left(3(1+\sqrt{1-\varepsilon^2})\right)^4}{\left(\delta/2(1-\sqrt{1-\varepsilon^2})\right)^4} (1-\varepsilon)\left[\frac{\mathrm{E}\varepsilon_1^4 - 2\sigma^4}{\varepsilon^3 R_n^2} + \frac{2\sigma^4}{\varepsilon^2 R_n}\right] \overset{n\to\infty}{\longrightarrow} 0 \qquad (3.24)
\end{aligned}$$

since $R_n \overset{n\to\infty}{\longrightarrow} \infty$ by $(R2)$ and $\mathrm{E}\varepsilon_1^4 < \infty$ by $(R4)$.

Using (3.22) and (3.23) along with (3.14) gives, for all $n \geq N$,

$$\sum_{J>\varepsilon R_n}^{R_n-1} \mathrm{P}\left\{m\left|\overline{X}_{J+1:R_n} - \mu_3\right| > 1/3\left(|\mu_1 - \mu_2| - m\,|\mu_3 - \mu_2|\right)\right\}$$

$$
\begin{aligned}
&\leq \frac{3^4}{\left(\delta/2(1-\sqrt{1-\varepsilon^2})\right)^4} \sum_{J>\varepsilon R_n}^{R_n-1} \left(1-\left(\frac{J}{R_n}\right)^2\right)^2 \left[\frac{\mathrm{E}\varepsilon_1^4-2\sigma^4}{(R_n-J)^3}+\frac{2\sigma^4}{(R_n-J)^2}\right] \\
&= \frac{3^4}{\left(\delta/2(1-\sqrt{1-\varepsilon^2})\right)^4} \sum_{J>\varepsilon R_n}^{R_n-1} \left(\left(1-\frac{J}{R_n}\right)\left(1+\frac{J}{R_n}\right)\right)^2 \left[\frac{\mathrm{E}\varepsilon_1^4-2\sigma^4}{R_n^3\left(1-J/R_n\right)^3}+\frac{2\sigma^4}{R_n^2\left(1-J/R_n\right)^2}\right] \\
&= \frac{3^4}{\left(\delta/2(1-\sqrt{1-\varepsilon^2})\right)^4} \sum_{J>\varepsilon R_n}^{R_n-1} \left(1+\frac{J}{R_n}\right)^2 \left[\frac{\mathrm{E}\varepsilon_1^4-2\sigma^4}{R_n^3\left(1-J/R_n\right)}+\frac{2\sigma^4}{R_n^2}\right] \\
&\leq \frac{3^4}{\left(\delta/2(1-\sqrt{1-\varepsilon^2})\right)^4} \sum_{J>\varepsilon R_n}^{R_n-1} 2^2 \left[\frac{\mathrm{E}\varepsilon_1^4-2\sigma^4}{R_n^3\left(1/R_n\right)}+\frac{2\sigma^4}{R_n^2}\right] \\
&= \frac{3^4}{\left(\delta/2(1-\sqrt{1-\varepsilon^2})\right)^4} 2^2(1-\varepsilon) \left[\frac{\mathrm{E}\varepsilon_1^4-2\sigma^4}{R_n}+\frac{2\sigma^4}{R_n}\right] \overset{n\to\infty}{\longrightarrow} 0 \qquad (3.25)
\end{aligned}
$$

since $R_n \overset{n\to\infty}{\longrightarrow} \infty$ by $(R2)$ and $\mathrm{E}\varepsilon_1^4 < \infty$ by $(R4)$.

Finally, using (3.22) and (3.23) along with (3.15) gives, for all $n \geq N$,

$$
\begin{aligned}
&\sum_{J>\varepsilon R_n}^{R_n-1} \mathrm{P}\left\{\left|\overline{X}_{-R_n:0}-\mu_1\right| > 1/3\left(\left|\mu_1-\mu_2\right|-m\left|\mu_3-\mu_2\right|\right)\right\} \\
&\quad\leq \frac{3^4}{\left(\delta/2(1-\sqrt{1-\varepsilon^2})\right)^4} \sum_{J>\varepsilon R_n}^{R_n-1} \left[\frac{\mathrm{E}\varepsilon_1^4-2\sigma^4}{(R_n+1)^3}+\frac{2\sigma^4}{(R_n+1)^2}\right] \\
&\quad= \frac{3^4}{\left(\delta/2(1-\sqrt{1-\varepsilon^2})\right)^4} R_n(1-\varepsilon) \left[\frac{\mathrm{E}\varepsilon_1^4-2\sigma^4}{(R_n+1)^3}+\frac{2\sigma^4}{(R_n+1)^2}\right] \\
&\quad= \frac{3^4(1-\varepsilon)}{\left(\delta/2(1-\sqrt{1-\varepsilon^2})\right)^4} \left[\frac{\mathrm{E}\varepsilon_1^4-2\sigma^4}{R_n^2\left(1+1/R_n\right)^3}+\frac{2\sigma^4}{R_n\left(1+1/R_n\right)^2}\right] \overset{n\to\infty}{\longrightarrow} 0 \qquad (3.26)
\end{aligned}
$$

since $R_n \overset{n\to\infty}{\longrightarrow} \infty$ by $(R2)$ and $\mathrm{E}\varepsilon_1^4 < \infty$ by $(R4)$.

Thus, by (3.12) and (3.24) – (3.26), we can conclude that

$$
\sum_{J>\varepsilon R_n}^{R_n-1} \mathrm{P}\left\{H_{nk_n}(J) < H_{nk_n}(0)\right\} \overset{n\to\infty}{\longrightarrow} 0.
$$

Case 2: $-R_n \leq J < -\varepsilon R_n$.

We need to show that

$$
\sum_{J\geq -R_n}^{-\varepsilon R_n} \mathrm{P}\left\{H_{nk_n}(J) < H_{nk_n}(0)\right\} \overset{n\to\infty}{\longrightarrow} 0.
$$

The proof is analagous to case 1 and thus will be omitted.

Combining cases 1 and 2, we can conclude that

$$\sum_{|J|>\varepsilon R_n} \mathrm{P}\left\{H_{nk_n}(J) \leq H_{nk_n}(0)\right\} \overset{n\to\infty}{\longrightarrow} 0,$$

and therefore, using (3.7), that

$$\mathrm{P}\left\{\left|\frac{\hat{I}_{nk_n}}{R_n}\right| > \varepsilon\right\} \overset{n\to\infty}{\longrightarrow} 0.$$

□

Lemma 4 *Assume (R1) – (R4) hold. Let $t_0 = t_{nk_n}$ for some $k_n \in \{0, 1, \ldots, n\}$ where f is discontinuous at t_0. Then*

$$\min\left\{\left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) - f^-(t_0)\right| \ , \ \left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) - f^+(t_0)\right|\right\} \xrightarrow{\mathcal{P}} 0.$$

Proof

Case 1: Suppose $\hat{I}_{nk_n} \geq 0$.

That is, the estimated breakpoint is to the right of t_0 $(= t_{nk_n})$. Then

$$\begin{aligned} g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) &= \frac{1}{\left|\hat{I}_{nk_n}\right| + R_n + 1} \sum_{j=-R_n}^{\hat{I}_{nk_n}} f(t_{n,k+j}) \\ &= \frac{1}{\left|\hat{I}_{nk_n}\right| + R_n + 1} \sum_{j=-R_n}^{-1} f(t_{n,k_n+j}) + \frac{1}{\left|\hat{I}_{nk_n}\right| + R_n + 1} \sum_{j=0}^{\hat{I}_{nk_n}} f(t_{n,k_n+j}) \end{aligned}$$

Therefore,

$$\begin{aligned} &\left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) - f^-(t_0)\right| \\ &\leq \left|\frac{1}{\left|\hat{I}_{nk_n}\right| + R_n + 1} \sum_{j=-R_n}^{-1} \left(f(t_{n,k_n+j}) - f^-(t_0)\right)\right| + \left|\frac{1}{\left|\hat{I}_{nk_n}\right| + R_n + 1} \sum_{j=0}^{\hat{I}_{nk_n}} \left(f(t_{n,k_n+j}) - f^-(t_0)\right)\right| \\ &\leq \frac{1}{\left|\hat{I}_{nk_n}\right| + R_n + 1} \sum_{j=-R_n}^{-1} \left|f(t_{n,k_n+j}) - f^-(t_0)\right| + \frac{1}{\left|\hat{I}_{nk_n}\right| + R_n + 1} \sum_{j=0}^{\hat{I}_{nk_n}} \left|f(t_{n,k_n+j}) - f^-(t_0)\right| \end{aligned}$$

$$\leq \frac{1}{R_n}\sum_{j=-R_n}^{-1}|f(t_{n,k_n+j})-f^-(t_0)|+\frac{1}{R_n}\sum_{j=0}^{\hat{I}_{nk_n}}|f(t_{n,k_n+j})-f^-(t_0)|$$

$$\leq \max_{j\in\{-R_n,\ldots,-1\}}|f(t_{n,k_n+j})-f^-(t_0)|+\frac{\left|\hat{I}_{nk_n}\right|+1}{R_n}\max_{j\in\{0,\ldots,\hat{I}_{nk_n}\}}|f(t_{n,k_n+j})-f^-(t_0)|$$

$$\leq \max_{j\in\{-R_n,\ldots,-1\}}|f(t_{n,k_n+j})-f^-(t_0)|+\frac{\left|\hat{I}_{nk_n}\right|+1}{R_n}\max_{j\in\{0,\ldots,R_n\}}|f(t_{n,k_n+j})-f^-(t_0)|. \tag{3.27}$$

Case 2: Suppose $\hat{I}_{nk_n}<0$.

That is, the estimated breakpoint is to the left of t_0 $(=t_{nk_n})$. Then

$$\begin{aligned} g_{nk_n}(t_{nk_n},\hat{I}_{nk_n}) &= \frac{1}{R_n+\left|\hat{I}_{nk_n}\right|}\sum_{j=\hat{I}_{nk_n}+1}^{R_n}f(t_{n,k_n+j}) \\ &= \frac{1}{R_n+\left|\hat{I}_{nk_n}\right|}\sum_{j=\hat{I}_{nk_n}+1}^{0}f(t_{n,k_n+j})+\frac{1}{R_n+\left|\hat{I}_{nk_n}\right|}\sum_{j=1}^{R_n}f(t_{n,k_n+j}) \end{aligned}$$

Therefore,

$$\begin{aligned} &\left|g_{nk_n}(t_{nk_n},\hat{I}_{nk_n})-f^+(t_0)\right| \\ &\leq \left|\frac{1}{R_n+\left|\hat{I}_{nk_n}\right|}\sum_{j=\hat{I}_{nk_n}+1}^{0}(f(t_{n,k_n+j})-f^+(t_0))\right|+\left|\frac{1}{R_n+\left|\hat{I}_{nk_n}\right|}\sum_{j=1}^{R_n}(f(t_{n,k_n+j})-f^+(t_0))\right| \\ &\leq \frac{1}{R_n+\left|\hat{I}_{nk_n}\right|}\sum_{j=\hat{I}_{nk_n}+1}^{0}|f(t_{n,k_n+j})-f^+(t_0)|+\frac{1}{R_n+\left|\hat{I}_{nk_n}\right|}\sum_{j=1}^{R_n}|f(t_{n,k_n+j})-f^+(t_0)| \\ &\leq \frac{1}{R_n}\sum_{j=\hat{I}_{nk_n}+1}^{0}|f(t_{n,k_n+j})-f^+(t_0)|+\frac{1}{R_n}\sum_{j=1}^{R_n}|f(t_{n,k_n+j})-f^+(t_0)| \\ &\leq \frac{\left|\hat{I}_{nk_n}\right|}{R_n}\max_{j\in\{\hat{I}_{nk_n}+1,\ldots,0\}}|f(t_{n,k_n+j})-f^+(t_0)|+\max_{j\in\{1,\ldots,R_n\}}|f(t_{n,k_n+j})-f^+(t_0)| \\ &\leq \frac{\left|\hat{I}_{nk_n}\right|}{R_n}\max_{j\in\{-R_n,\ldots,0\}}|f(t_{n,k_n+j})-f^+(t_0)|+\max_{j\in\{1,\ldots,R_n\}}|f(t_{n,k_n+j})-f^+(t_0)|. \end{aligned} \tag{3.28}$$

Combining the two cases, we have for any $\hat{I}_{nk_n}$,

$$\min\left\{\left|g_{nk_n}(t_{nk_n},\hat{I}_{nk_n})-f^-(t_0)\right|\ ,\ \left|g_{nk_n}(t_{nk_n},\hat{I}_{nk_n})-f^+(t_0)\right|\right\}$$

$$\leq \max_{j\in\{-R_n,\ldots,-1\}} |f(t_{n,k_n+j}) - f^-(t_0)| + \frac{\left|\hat{I}_{nk_n}\right| + 1}{R_n} \max_{j\in\{0,\ldots,R_n\}} |f(t_{n,k_n+j}) - f^-(t_0)|$$
$$+ \frac{\left|\hat{I}_{nk_n}\right|}{R_n} \max_{j\in\{-R_n,\ldots,0\}} |f(t_{n,k_n+j}) - f^+(t_0)| + \max_{j\in\{1,\ldots,R_n\}} |f(t_{n,k_n+j}) - f^+(t_0)| . \quad (3.29)$$

By Lemma 3, we have $\left|\hat{I}_{nk_n}\right|/R_n \xrightarrow{\mathcal{P}} 0$, so clearly $(\left|\hat{I}_{nk_n}\right| + 1)/R_n \xrightarrow{\mathcal{P}} 0$ as well. Also, $\max_{j\in\{0,\ldots,R_n\}} |f(t_{n,k_n+j}) - f^-(t_0)| < \infty$ and $\max_{j\in\{-R_n,\ldots,0\}} |f(t_{n,k_n+j}) - f^+(t_0)| < \infty$ since f is bounded. Therefore, in (3.29),

$$\frac{\left|\hat{I}_{nk_n}\right| + 1}{R_n} \max_{j\in\{0,\ldots,R_n\}} |f(t_{n,k_n+j}) - f^-(t_0)| \xrightarrow{\mathcal{P}} 0$$

and

$$\frac{\left|\hat{I}_{nk_n}\right|}{R_n} \max_{j\in\{-R_n,\ldots,0\}} |f(t_{n,k_n+j}) - f^+(t_0)| \xrightarrow{\mathcal{P}} 0.$$

Now consider the remaining terms on the right-hand side of (3.29). In the proof of Lemma 3, statement (3.17), we showed that

$$\max_{j\in\{-R_n,\ldots,0\}} |f(t_{n,k_n+j}) - f^-(t_0)| \xrightarrow{n\to\infty} 0$$

and therefore that

$$\max_{j\in\{-R_n,\ldots,-1\}} |f(t_{n,k_n+j}) - f^-(t_0)| \xrightarrow{n\to\infty} 0.$$

Also in the proof of Lemma 3, statement (3.18), we showed that

$$\max_{j\in\{1,\ldots,R_n\}} |f(t_{n,k_n+j}) - f^+(t_0)| \xrightarrow{n\to\infty} 0.$$

Thus, using inequality (3.29), we can conclude that

$$\min\left\{\left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) - f^-(t_0)\right| , \left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) - f^+(t_0)\right|\right\} \xrightarrow{n\to\infty} 0$$

and Lemma 4 is proven.

□

Consistency of the not-so-smoother, $\hat{f}$, at continuity points now follows directly from Lemmas 1 and 2, and at discontinuity points from Lemmas 1, 3 and 4. This is stated formally in the two theorems which follow.

Theorem 1 *Assume (R1) and (R2) hold. If* $t_{nk_n} \overset{n\to\infty}{\longrightarrow} t_0$ *where* f *is continuous at* t_0*, then*

$$\left|\hat{f}(t_{nk_n}) - f(t_0)\right| \xrightarrow{\mathcal{P}} 0.$$

Proof

$$\begin{aligned}\left|\hat{f}(t_{nk_n}) - f(t_0)\right| &= \left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) + \gamma_{nk_n}(\hat{I}_{nk_n}) - f(t_{nk_n})\right| \\ &\leq \left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) - f(t_{nk_n})\right| + \left|\gamma_{nk_n}(\hat{I}_{nk_n})\right|\end{aligned}$$

By Lemma 2 we have that $\left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) - f(t_{nk_n})\right| \overset{n\to\infty}{\longrightarrow} 0$, and by Lemma 1 we have that $\left|\gamma_{nk_n}(\hat{I}_{nk_n})\right| \xrightarrow{\mathcal{P}} 0$. Thus we can conclude that $\left|\hat{f}(t_{nk_n}) - f(t_0)\right| \xrightarrow{\mathcal{P}} 0$.

□

By the remarks following Lemma 2, we can replace $f(t_0)$ by $f(t_{nk_n})$ in the above theorem to obtain the slightly different statement that $\left|\hat{f}(t_{nk_n}) - f(t_{nk_n})\right| \xrightarrow{\mathcal{P}} 0$.

Theorem 2 *Assume (R1) - (R4) hold. If* $t_{nk_n} = t_0$ *for some* $k_n \in \{0, 1, \ldots, n\}$ *where* f *is discontinuous at* t_0*, then*

$$\min\left\{\left|\hat{f}(t_{nk_n}) - f^-(t_0)\right|, \left|\hat{f}(t_{nk_n}) - f^+(t_0)\right|\right\} \xrightarrow{\mathcal{P}} 0.$$

Proof

$$\begin{aligned}&\min\left\{\left|\hat{f}(t_{nk_n}) - f^-(t_0)\right|, \left|\hat{f}(t_{nk_n}) - f^+(t_0)\right|\right\} \\ &= \min\left\{\left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) + \gamma_{nk_n}(\hat{I}_{nk_n}) - f^-(t_0)\right|, \left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) + \gamma_{nk_n}(\hat{I}_{nk_n}) - f^+(t_0)\right|\right\} \\ &\leq \min\left\{\left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) - f^-(t_0)\right|, \left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) - f^+(t_0)\right|\right\} + \left|\gamma_{nk_n}(\hat{I}_{nk_n})\right|\end{aligned}$$

By Lemma 4 we have that $\min\left\{\left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) - f^-(t_0)\right|, \left|g_{nk_n}(t_{nk_n}, \hat{I}_{nk_n}) - f^+(t_0)\right|\right\} \xrightarrow{n\to\infty} 0$, and by Lemma 1 we have that $\left|\gamma_{nk_n}(\hat{I}_{nk_n})\right| \xrightarrow{\mathcal{P}} 0$. Thus we can conclude that

$$\min\left\{\left|\hat{f}(t_{nk_n}) - f^-(t_0)\right|, \left|\hat{f}(t_{nk_n}) - f^+(t_0)\right|\right\} \xrightarrow{\mathcal{P}} 0.$$

□

It should be noted that Theorem 2 is true if we replace $t_{nk_n} = t_0$ with the more general statement $t_{nk_n} \xrightarrow{n\to\infty} t_0$. This is because, as stated preceding Lemma 3, Lemma 4 is still valid in this more general setting. However, the details of the proof will not be given in this thesis.

Comment

Under the condition that the bandwidth goes to infinity as n goes to infinity but goes to zero relative to n, the usual moving average smoother is consistent except at the discontinuity points. Briefly, this is because the neighbourhoods around all continuity points get smaller and smaller until they eventually do not contain a discontinuity. At a point of discontinuity, the moving average estimate converges to the midpoint of the jump, that is, $(f^-(t_0) + f^+(t_0))/2$ where t_0 denotes the discontinuity point. The not-so-smoother, on the other hand, is infinitely close to either $f^-(t_0)$ or $f^+(t_0)$. The failure to converge at only one point (or a few points) may not seem significant asymptotically, but in practice where n is finite, the implications are important. In the vicinity of a discontinuity and for finite n, the moving average smoother is upset by the presence of a discontinuity while the not-so-smoother is not. This is illustrated in the next chapter.

Chapter 4

Performance of The Not-So-Smoother

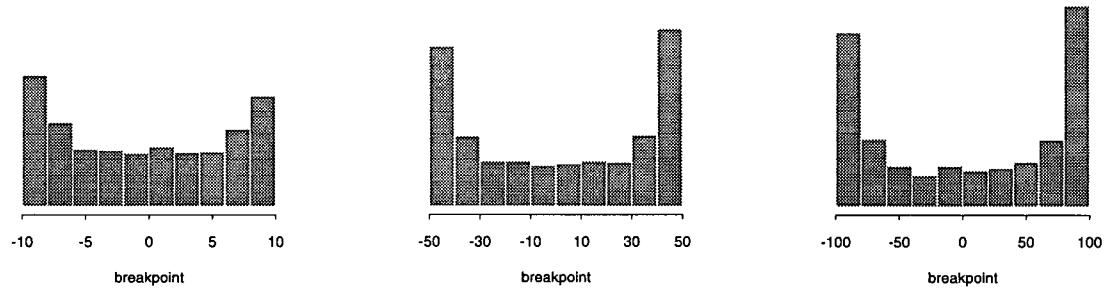
In this chapter, the performance of the not-so-smoother in application is demonstrated. Before considering the global performance, the local behaviour within a fixed neighbourhood is investigated.

4.1 Local Performance

Recall that in estimating the function, f, at a given design point, say t_{ni}, we find the best broken constant fit in a neighbourhood around the point. (The neighbourhood includes $2R_n + 1$ points — the R_n points to both the right and left of t_{ni} as well as t_{ni} itself.) The point at which the best broken constant "splits" is called the breakpoint, and the function estimate is taken to be the average of all observations in the neighbourhood to the left or right of the breakpoint, whichever contains the central point, t_{ni}. In order to demonstrate the characteristics of the breakpoint, consider the simple case where the local model is truly two constants.

$$X_{n,i+j} = \begin{cases} c1 + \varepsilon_{n,i+j} & \text{for } j = -R_n, \ldots, I_{ni} \\ c2 + \varepsilon_{n,i+j} & \text{for } j = I_{ni} + 1, \ldots, R_n \end{cases} \tag{4.1}$$

When there is a discontinuity within the neighbourhood, $c1 \neq c2$ and I_{ni} indexes the point at which the discontinuity occurs. When the function is continuous within the neighbourhood, $c1 = c2$ and I_{ni} becomes irrelevant. Of course, in practice this local model will be violated. However, within a small enough neighbourhood, the function f should not fluctuate too much and the local model should be an adequate approximation.



$R_n = 10$ $R_n = 50$ $R_n = 100$

Figure 4.1: Histograms of breakpoints when no discontinuity exists

Suppose we are in the case where no discontinuity occurs. Then we would prefer that the average used as the estimate of $f(t_{ni})$ includes as many observations as possible. Regardless of the location of the estimated breakpoint, the estimate will have no bias since all observations have the same mean, but the more observations used in the average, the smaller the variance of the estimate and thus the greater the efficiency. Consequently, we would like the breakpoint to be at either extreme of the neighbourhood.

We do not know the theoretical distribution of the breakpoint but the empirical distribution can be studied. We considered the case where the error terms are independent, standard normal random variables (mean 0 and standard deviation $\sigma = 1$) and we take $c1$ $(= c2)$ arbitrarily to be zero. We randomly generated $2R_n + 1$ standard normal observations and determined the breakpoint by finding the argument that minimizes equation (2.2). After repeating this 1000 times, we looked at the histogram of the breakpoints. This was done for varying neighbourhood sizes, namely $R_n = 10$, 50, and 100. Histograms of the results are given in Figure 4.1.

We can see that the breakpoint does have a tendency to be located near the edges of

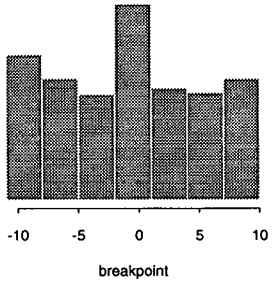
the neighbourhood and this tendency becomes stronger as the number of points in the neighbourhood increases, suggesting that $\min\left\{\left|\hat{I}_{ni}/R_n - 1\right|, \left|\hat{I}_{ni}/R_n + 1\right|\right\} \xrightarrow{\mathcal{P}} 0$.

At this point we should note that in order for the locally constant model to be approximately true, the neighbourhood must be small, yet to get an efficient estimate, a large number of points in the neighbourhood is needed. This can be achieved if the number of observations in the data set is very large and the observations are close together, since this would allow us to define a neighbourhood size that is small relative to the total amount of data but still contains numerous observations. Note that the asymptotic equivalents of these conditions, stated in requirements $(R1)$ and $(R2)$ at the beginning of Chapter 3, were needed to prove consistency.

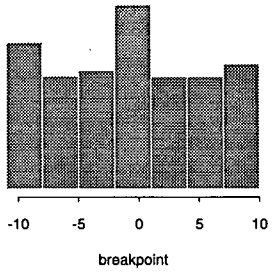
Now, when we are in the case where the neighbourhood contains a discontinuity, we expect the breakpoint to be at, or at least very near, the location of the discontinuity. This is the fundamental idea behind the not-so-smoother. If the breakpoint is correctly located near the jump, then the estimate of the central point will be an average of points all or most of which lie on the same side of the jump.

Intuitively, the larger the magnitude of the jump and the smaller the variability in the data, the more likely the breakpoint will be located at the jump. Denote the jump size by δ, which in the local model is just $|c1 - c2|$, and the standard deviation of the data by σ as before. Then if we consider the ratio δ/σ, we expect to be more successful identifying the jumps as the ratio increases. Of course, the number of points in the neighbourhood also affects the results. Presumably, with more observations the jump is easier to identify.

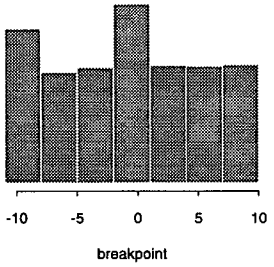
We can investigate empirically the behaviour of the breakpoint in the presence of a discontinuity. The discontinuity can occur anywhere within the neighbourhood, but we chose the center for our simulations, meaning $I_{ni} = 0$ in the local model. We considered the case where the error terms are independent, normal random variables with mean 0



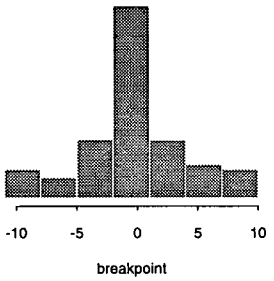
$\delta/\sigma = 0.5; \sigma = 1$



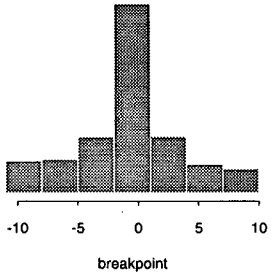
$\delta/\sigma = 0.5; \sigma = 10$



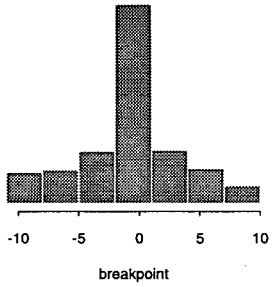
$\delta/\sigma = 0.5; \sigma = 50$



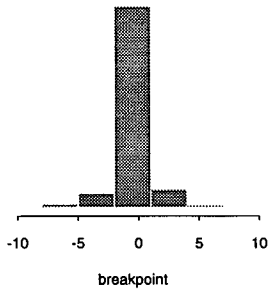
$\delta/\sigma = 1; \sigma = 1$



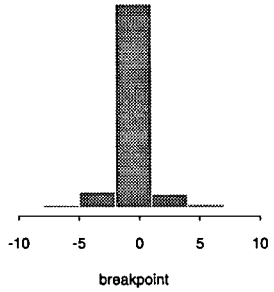
$\delta/\sigma = 1; \sigma = 10$



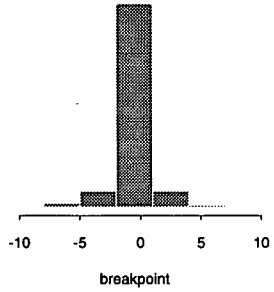
$\delta/\sigma = 1; \sigma = 50$



$\delta/\sigma = 2; \sigma = 1$



$\delta/\sigma = 2; \sigma = 10$



$\delta/\sigma = 2; \sigma = 50$

Figure 4.2: Histograms of breakpoints for various δ and σ values keeping the ratio δ/σ constant

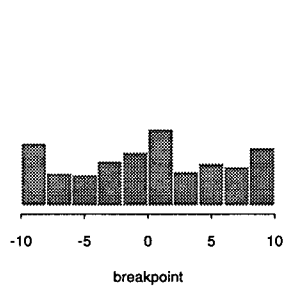
and standard deviation $\sigma = 1$. Data were generated corresponding to a given δ/σ value and a given neighbourhood size, and the breakpoint was determined. Specifically we looked at the cases where R_n equals 10, 50, and 100, and for each value of R_n we looked at δ/σ equal to 0.5, 1, and 2. Intuition suggests that the distribution of the breakpoint depends only on the ratio δ/σ and not the individual values of δ and σ. Our experiences seem to confirm this. Refer to Figure 4.2 where histograms of the breakpoint for varying δ and σ values keeping the ratio constant are presented. Note that the bandwidth is taken to be 10 in all results shown. The histograms look almost identical when the ratio δ/σ is the same. Thus, the actual values of δ and σ used in our simulations appear to be irrelevant; however, for completeness, note that we consistently took σ to be 1.

In each case, 1000 replications were done and the resulting breakpoints reported in histograms. These results are shown in Figure 4.3. As we expect, the breakpoint is correctly located at the center of the neighbourhood more frequently as R_n increases and as δ/σ increases.

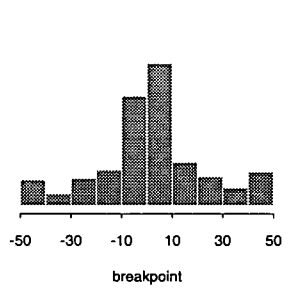
Recall that Lemma 3 in the previous chapter states that when the function is discontinuous at the point being estimated, meaning the discontinuity occurs in the center of the neighbourhood, the ratio of the breakpoint to the bandwidth size, R_n, converges to the location of the discontinuity. The results of our simulations are consistent with this lemma. The rate of convergence, however, is highly dependent on the value of δ/σ. We see that when δ/σ equals 0.5 the rate is considerably slower than when it equals 1 or 2.

4.2 Global Performance

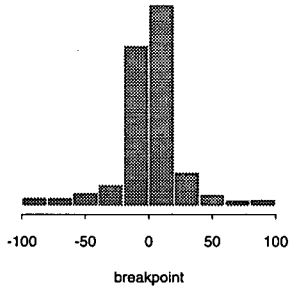
Now that the local behaviour has been considered, we will look at the not-so-smoother's global performance on some data sets. As with other smoothing methods, a bandwidth must be chosen before applying the method, and, as with other smoothing methods, the



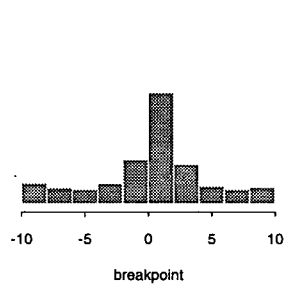
$\delta/\sigma = 0.5$; $R_n = 10$



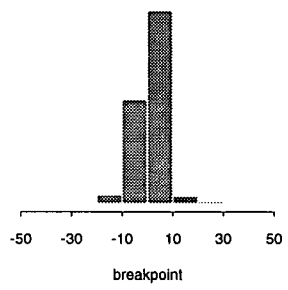
$\delta/\sigma = 0.5$; $R_n = 50$



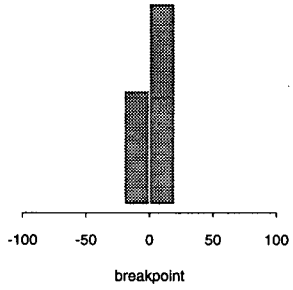
$\delta/\sigma = 0.5$; $R_n = 100$



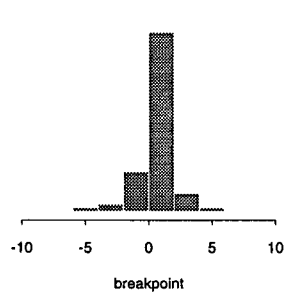
$\delta/\sigma = 1$; $R_n = 10$



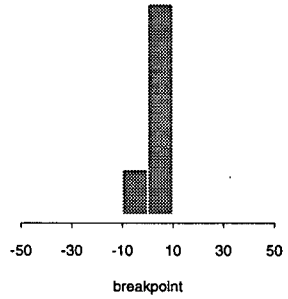
$\delta/\sigma = 1$; $R_n = 50$



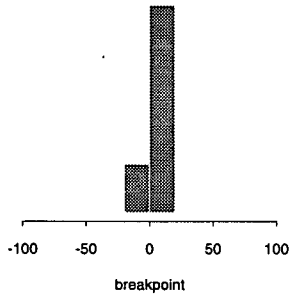
$\delta/\sigma = 1$; $R_n = 100$



$\delta/\sigma = 2$; $R_n = 10$



$\delta/\sigma = 2$; $R_n = 50$



$\delta/\sigma = 2$; $R_n = 100$

Figure 4.3: Histograms of breakpoints when a discontinuity exists in the center of the neighbourhood

best way to make this choice is not clear. In general, a large bandwidth leads to a small variance but a large bias in the estimate, whereas a small bandwidth leads to a small bias but a large variance. This is often referred to as the "bias-variance trade-off". We will not discuss the problem of bandwidth selection in depth in this thesis. Suffice it to say that cross-validation and plug-in methods, which are used in the case of continuous functions, can presumably be derived and applied here.

Simulated Data

To begin with, we will evaluate the not-so-smoother's performance on simulated data in which case the true function is known. The bandwidth which minimizes the squared error between the estimated function and the true function at the design points can be determined; we will refer to this as the optimal bandwidth. The optimal performance of the not-so-smoother will be compared with the optimal performance of a simple (unweighted) moving average smoother.

A continuous function is considered first. We expect the optimal performance of the not-so-smoother to be worse than the optimal performance of a simple moving average smoother when the true function is continuous — how much worse is of interest. The function chosen was one cycle of the sine curve. One hundred observations were generated at equal increments over the interval $[0,\ 2\pi)$ and the noise generated was normal with standard deviation 0.3. The model can be written as

$$X_i = \sin\left(\frac{2\pi i}{100}\right) + \varepsilon_i, \qquad i = 0, 1, \ldots, 99$$

where $\varepsilon_i \sim Normal(0, \sigma = 0.3)$. Note that the function was chosen such that no discontinuities are introduced by replicating the data set when estimating near the ends.

Five hundred data sets were generated according to this model. For each simulated data set, the not-so-smooth estimate of the function was calculated using every bandwidth from 1 to 50 (half the total number of observations). For each bandwidth, the mean

Table 4.1: Summary results of smoothing methods on simulated data (500 replications in each case)

	Moving Average Smooth		Not-So-Smooth	
function	mean optimal bandwidth (sd)	mean optimal mse (sd)	mean optimal bandwidth (sd)	mean optimal mse (sd)
sine N(0,.3) noise	8.458 (1.561)	0.006 (0.003)	4.390 (0.942)	0.027 (0.005)
split cube-root N(0,.2) noise	1.650 (0.604)	0.020 (0.002)	7.930 (2.379)	0.008 (0.002)

squared error between the estimate and the function at the design points was calculated. The optimal bandwidth is the one leading to the smallest mean squared error, which we will call the optimal mse. Specifically, for each bandwidth $R \in \{1, \ldots, 50\}$,

$$MSE(R) = \sum_{i=0}^{n} \left(\hat{f}_R(t_i) - f(t_i)\right)^2$$

was calculated, where $n = 99$, $t_i = (2\pi i)/100$, and $\hat{f}_R$ denotes the not-so-smooth corresponding to bandwidth R. We call the minimum value of MSE(R) the optimal mse and the value of R minimizing it the optimal bandwidth. In the exact same way, the optimal bandwidth and optimal mse using a simple moving average smooth were determined for each data set, where the only difference is $\hat{f}_R$ now refers to the moving average estimator.

The results of the 500 simulations confirm that the moving average smooth consistently achieves a smaller minimum mse than the not-so-smooth. Also, the bandwidth corresponding to the minimum mse is consistently larger for the moving average smooth than the not-so-smooth. A summary of the average optimal mse and bandwidth for the two methods is presented in the first row of Table 4.1.

The optimal not-so-smooth and the optimal moving average smooth for one simulation are shown in Figure 4.4. The not-so-smoother appears to be suitably named. Although

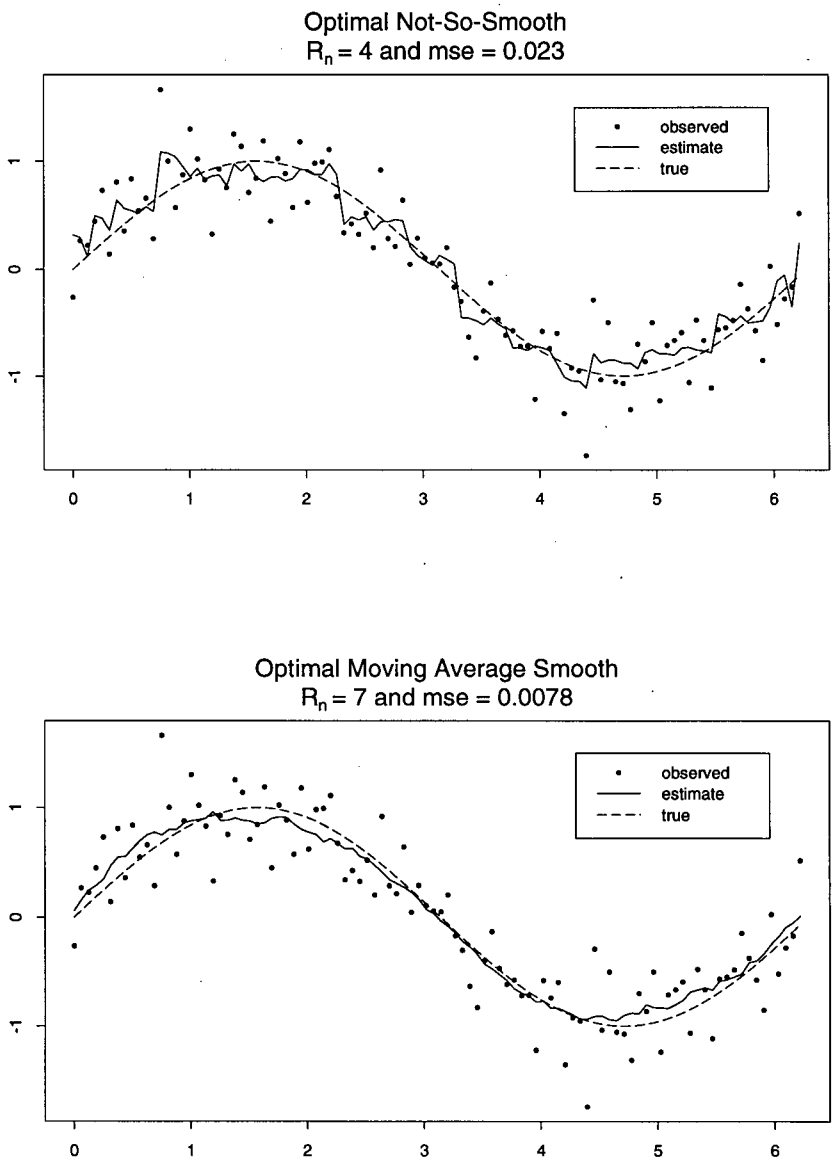


Figure 4.4: Optimal smooths of one cycle of the sine function with Normal(0, 0.3) errors

both smooths follow the curve of the data well, the not-so-smooth is much jumpier due to the fact that it "breaks" within every neighbourhood even when a discontinuity is not present. Methods to improve upon this will be presented in Chapter 5.

Next we consider a function with one discontinuity. We expect the not-so-smoother to perform better than a moving average smoother in such a case. The function chosen was the cube root function in the range -1 to 1, with the section from -1 to 0 shifted upwards by 2. One hundred observations were generated at equal increments over the interval $[-1, 1)$ and the noise generated was normal with standard deviation 0.2. The model can be written as

$$X_i = \begin{cases} \left(\frac{i-50}{50}\right)^{1/3} + 2 + \varepsilon_i & i = 0, \ldots, 49 \\ \left(\frac{i-50}{50}\right)^{1/3} + \varepsilon_i & i = 50, \ldots, 99 \end{cases}$$

where $\varepsilon_i \sim Normal(0, \sigma = 0.2)$. Again note that the function was chosen such that no discontinuities are introduced when estimating at design points near the ends of the data set.

Five hundred data sets were generated according to this model. As before, the optimal mse and the corresponding optimal bandwidth for both the not-so-smoother and the moving average smoother were found for each data set. In each of the 500 simulations the not-so-smoother achieved a smaller optimal mse and also had a substantially larger optimal bandwidth than the moving average smoother. A summary of the mean results is given in the second row of Table 4.1. Note that, on average, the optimal bandwidth for the moving average smoother was just 1.65. If the bandwidth is large, then all points near the discontinuity will have highly biased estimates. Because the increase in bias which results from using a large bandwidth is greater than the decrease in variability, the mse is minimized at a very small bandwidth. This leads to a very rough smooth overall, as we illustrate next.

The optimal not-so-smooth and the optimal moving average smooth for one simulation

are shown in Figure 4.5. The optimal moving average smooth uses a bandwidth of 1 and therefore does not smooth away the discontinuity too much, but it is indeed very rough. The not-so-smoother preserves the discontinuity even better and gives a much smoother estimate overall, as reflected by the smaller mse.

The ideal situation in which to apply the not-so-smoother is when the underlying function is two distinct constants. To give a complete evaluation of our method's performance, we should investigate such a situation. Data were generated according to the following model, which sets one of the constants equal to 0 for convenience:

$$X_i = \begin{cases} \varepsilon_i & i = 0, \ldots, 24 \\ \delta + \varepsilon_i & i = 25, \ldots, 49 \end{cases}$$

where $\varepsilon_i \sim Normal(0, \sigma = 1)$.

The magnitude of the jump, δ, relative to the standard deviation in the noise, σ, which in this model is 1, was varied. Similar to when we investigated local performance of the not-so-smoother, we considered δ/σ equal to 0.5, 1 and 2, and in addition we considered δ/σ equal to 0 since this represents the continuous case. One thousand simulations were carried out for each δ/σ value. For every data set, the optimal mse and optimal bandwidth using the not-so-smoother and the moving average smoother were found. Table 4.2 summarizes the results of the simulations.

As we expect, the moving average smoother performs better than the not-so-smoother in the continuous case (when $\delta/\sigma = 0$). It has a smaller average optimal mse, and more specifically, its optimal mse is smaller in 932 of the 1000 cases. When $\delta/\sigma = 0.5$, the moving average smoother again performed better with respect to having a smaller average optimal mse. Also, in 685 of the 1000 replications, the optimal moving average mse was smaller than the optimal not-so-smooth mse. This is not too surprising because we saw in our investigation of the not-so-smoother's local behaviour that when $\delta/\sigma = 0.5$ and

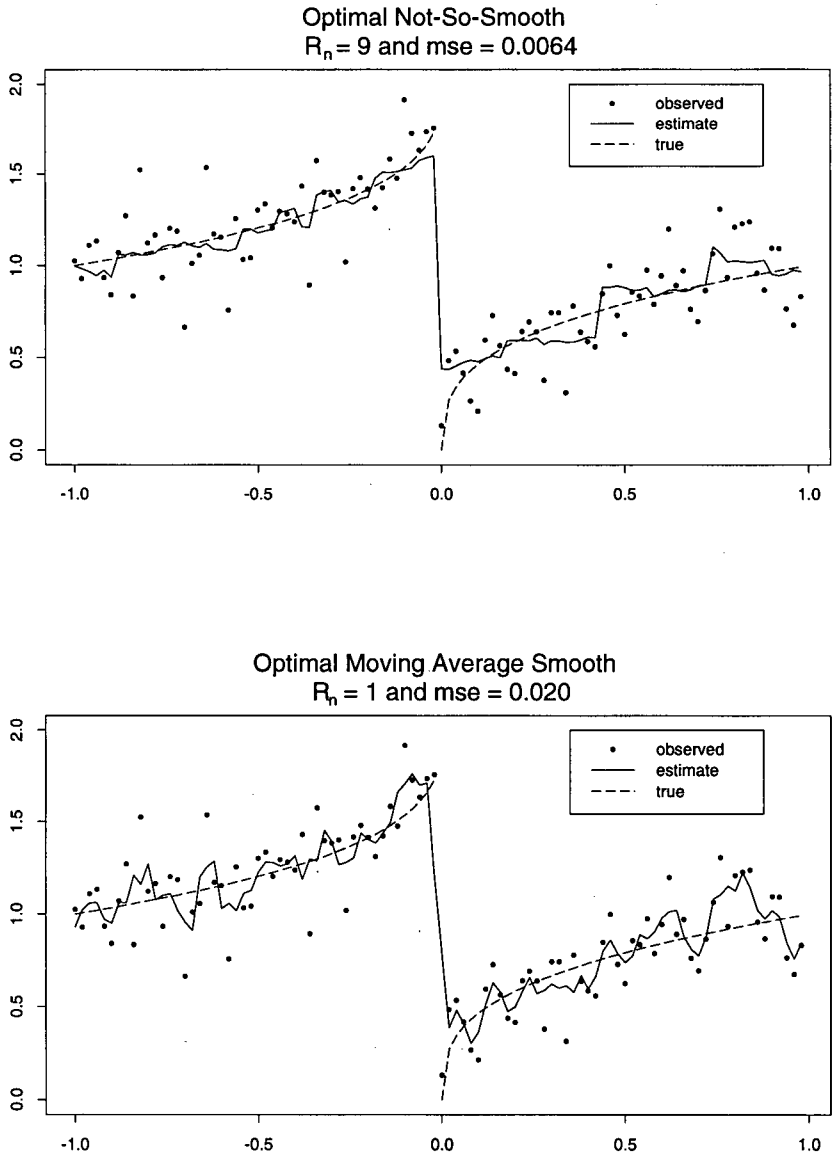


Figure 4.5: Optimal smooths of a split cube-root function with Normal(0, 0.2) errors

Table 4.2: Summary results of smoothing methods for data simulated according to the two-constant model (1000 replications in each case)

	Moving Average Smooth		Not-So-Smooth	
δ/σ	mean optimal bandwidth (sd)	mean optimal mse (sd)	mean optimal bandwidth (sd)	mean optimal mse (sd)
0	25.000 (0.000)	0.019 (0.025)	23.381 (2.432)	0.041 (0.032)
0.5	13.734 (4.542)	0.209 (0.126)	20.846 (4.677)	0.272 (0.159)
1.0	8.170 (3.004)	0.106 (0.047)	17.118 (5.304)	0.109 (0.061)
2.0	4.288 (1.858)	0.213 (0.054)	14.291 (3.702)	0.120 (0.094)

the bandwidth is small, the discontinuity is not located very accurately. Furthermore, the moving average smoother will not be too biased in the vicinity of the discontinuity when the jump size is relatively small. When $\delta/\sigma = 1$, the two methods perform equally well in terms of mse. We see from the table that the average optimal mse is almost identical for both methods (taking the standard deviations into account, the difference is insignificant). In addition, the optimal not-so-smooth mse was smaller in approximately half of the replications, 524 out of 1000 to be exact. When $\delta/\sigma = 2$, the not-so-smoother performs much better, having a smaller optimal mse than the moving average smooth in 861 of the 1000 replications and also having a smaller optimal mse on average (refer to Table 4.2). It seems clear from these results that as the ratio δ/σ increases, the not-so-smoother becomes increasingly preferable to the moving average smoother.

Lastly, we compare the optimal bandwidth sizes. For both smoothing methods, the average optimal bandwidth decreased as the ratio δ/σ increased. Moreover, when $\delta/\sigma = 0$, the moving average smooth tended to have a slightly larger optimal bandwidth than the not-so-smooth. For each non-zero value of δ/σ, the not-so-smooth tended to have

a larger optimal bandwidth. This is consistent with the results in Table 4.1 in which the moving average smooth had a larger mean optimal bandwidth in the continuous case (sine curve data), and the not-so-smooth had a larger mean optimal bandwidth in the discontinuous case (split cube-root function).

Real Data

Using a real data set for which the true signal is unknown will give a more practical demonstration of the not-so-smoother's performance. We considered data for which each observation is the measurement of current flowing through an ion channel in a cell membrane (Fredkin and Rice, 1990). Theoretically, the current has two, and perhaps more, conductance levels between which it seems to switch randomly. Due to the method used to measure the current, the noise can be great compared to the current size. The data set has a total of 16000 current recordings from which we chose a subset of 500 observations to investigate here. A plot of this subset is shown in Figure 4.6.

As with the simulated data, we compared the performance of the not-so-smoother and the moving average smoother on the current recordings data. To avoid the problem of bandwidth selection, we applied both smoothing methods using a range of bandwidths, namely $R_n = 5$, 15 and 30. Quantitative methods to determine the best choice of bandwidth could be used, but we simply evaluated the smooths visually. Plots of all the smooths are given in Figure 4.7.

Consistent with the results of our simulations on discontinuous data, a much smaller bandwidth appears best for the moving average smoother than for the not-so-smoother since there are seemingly several discontinuities. With a bandwidth of 5, the moving average smooth identifies the jumps quite clearly but produces jagged output as the cost. As the bandwidth increases, the output becomes progressively smoother but the jumps

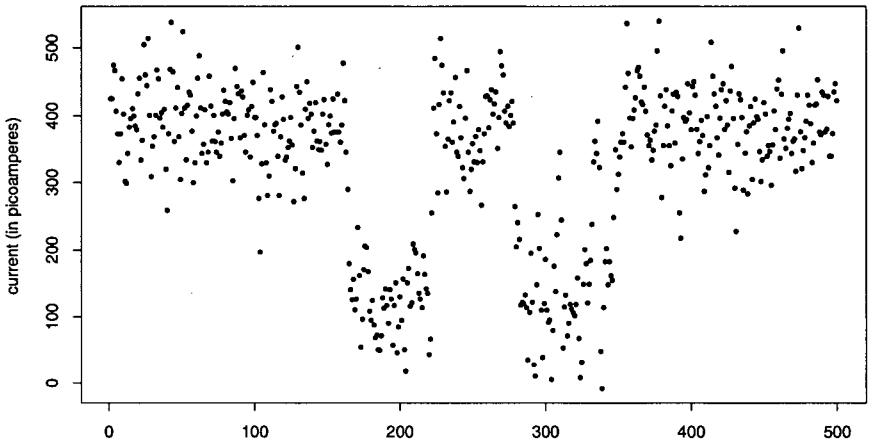


Figure 4.6: Measurements of current flowing through an ion channel in a cell membrane

become more blurred. On the other hand, the not-so-smoother also produces progressively smoother output as the bandwidth increases yet the jumps are still preserved and, in fact, become sharper.

Although the true signal is unknown, the results strongly suggest that the current switched between a high and low conductance level at four times indexed approximately by 165, 222, 280 and 348. In addition, the not-so-smooth using a bandwidth of 5 provides some evidence that the current switched levels rapidly, perhaps to an intermediate level, a couple of times between indices 280 and 348.

This data set provides a good opportunity to see how the not-so-smoother behaves when the bandwidth is so large that a neighbourhood contains two or more discontinuities. In such a case, the local two-constant model is grossly violated and we can expect poor results. If the minimum number of observations between any two discontinuities is

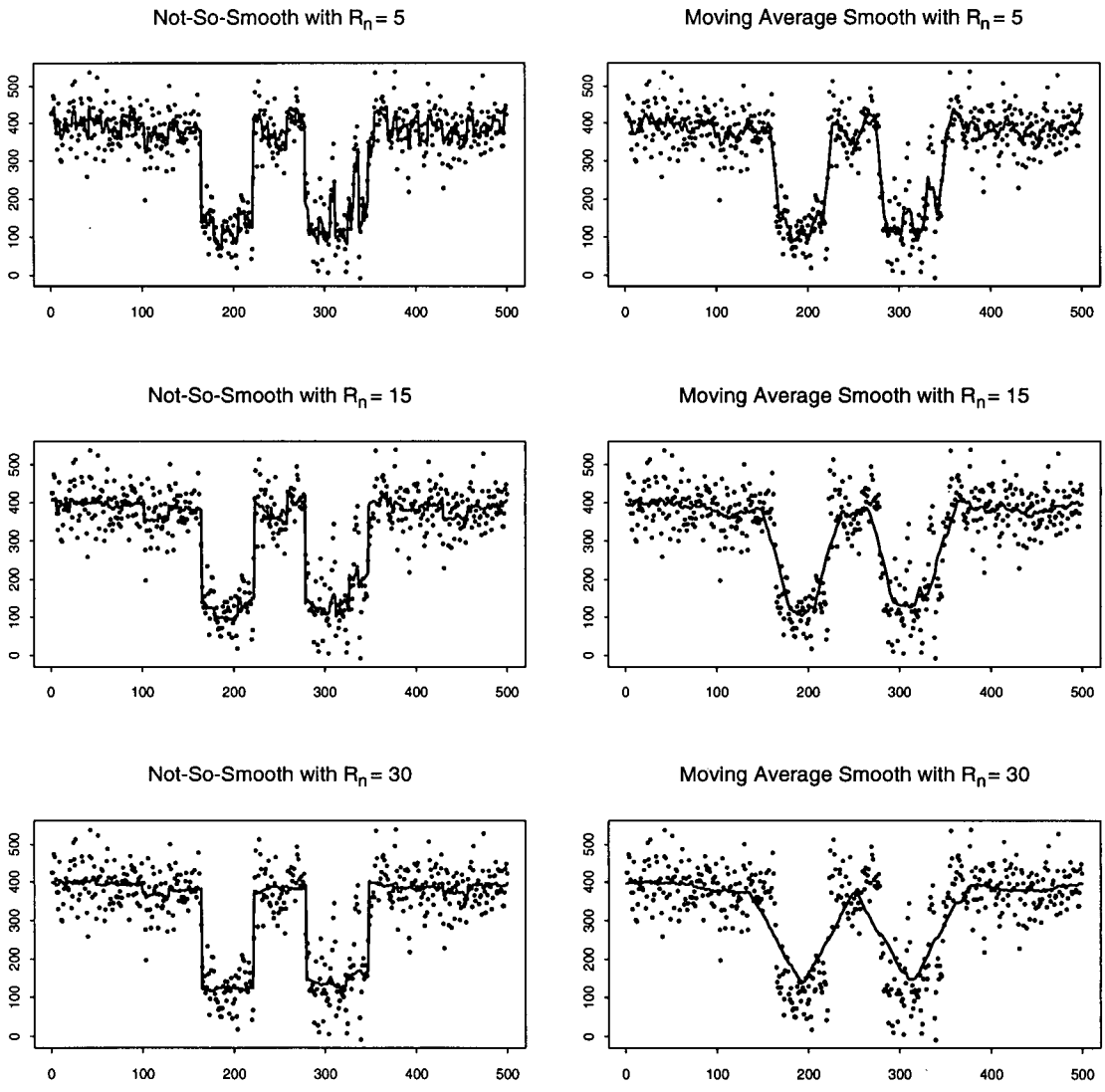


Figure 4.7: Smooths of the current data using the not-so-smoother and the moving average smoother for various bandwidths

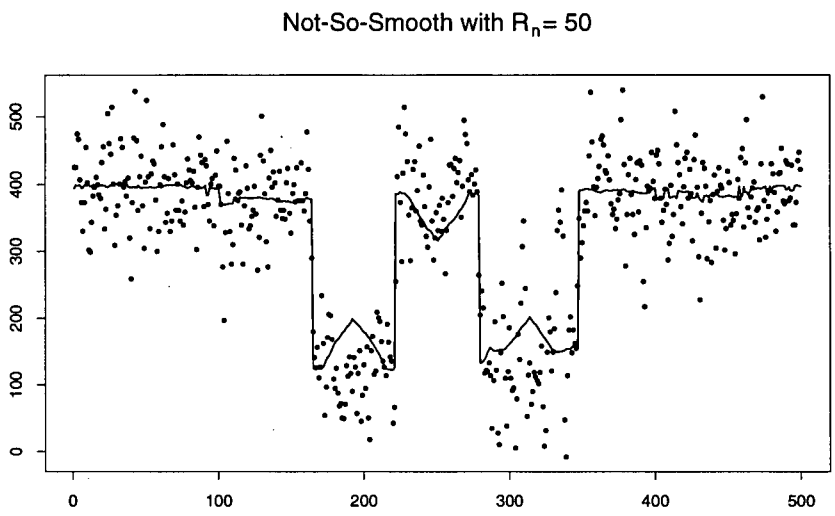


Figure 4.8: Not-so-smooth of the current data using a bandwidth of 50

d, then taking the bandwidth to be greater than $d/2$ will violate the local model. For the ion current data, the minimum number of observations between discontinuities is about 60 (assuming that the discontinuity points are in fact indexed approximately by 165, 222, 280 and 348). Thus, using any bandwidth exceeding 30 will result in neighbourhoods containing more than one discontinuity. We calculated the not-so-smooth for $R_n = 50$ and the result is shown in Figure 4.8. Undesirable peaks in the estimate occur between the discontinuities because in neighbourhoods containing two discontinuities, no matter where the best broken constant fit breaks, the estimate will still be an average of observations on both sides of one of the discontinuities.

Chapter 5

Extensions

5.1 The Somewhat-Smoother

The examples in the previous chapter illustrated that the not-so-smoother performs well at preserving discontinuities, but often produces rough output. This is due to the nature of the smoothing method in that the best local fit always breaks, even when no discontinuity exists, so only a subset of the neighbourhood data is averaged for each estimate.

One way to improve the roughness of the not-so-smoother is, for each neighbourhood, to compare the fit of the best broken constant with the fit of the best constant, which is just the mean of all observations in the neighbourhood. Unless the broken constant fit is substantially better, the mean should be used as the estimate. In other words, the not-so-smooth estimate should be used if the broken constant fit is much better since this is evidence of a break, but otherwise the moving average estimate should be used.

A hypothesis test can be used to determine if the broken constant fit is substantially better. Consider a fixed neighbourhood about a point t_{ni}. Assume that the local model given by equation (4.1), which says that the true function within the neighbourhood is two constants, holds. For convenience, we repeat the model here:

$$X_{n,i+j} = \begin{cases} c1 + \varepsilon_{n,i+j} & \text{for } j = -R_n, \ldots, I_{ni} \\ c2 + \varepsilon_{n,i+j} & \text{for } j = I_{ni} + 1, \ldots, R_n. \end{cases}$$

Recall that the ε's are assumed to be independent and identically distributed with mean

0 and variance σ^2. We want to test the null hypothesis that the two constants are equal versus the alternative that they are not equal. If the null is rejected, then the not-so-smooth estimate is used.

The natural estimators of $c1$ and $c2$ are the sample means of $X_{n,i-R_n}, \ldots, X_{n,i+I_{ni}}$ and $X_{n,i+I_{ni}+1}, \ldots, X_{n,i+R_n}$ respectively, which in keeping with the notation of Chapter 3 are denoted by $\overline{X}_{-R_n:I_{ni}}$ and $\overline{X}_{I_{ni}+1:R_n}$. Of course, I_{ni} is unknown and must be replaced by its estimate, $\hat{I}_{ni}$. The variance, σ^2, is also unknown in most situations and therefore must be estimated.

A two-sample t-test for the difference in means can be used.

$H_0 : c1 = c2 \quad \text{vs.} \quad H_a : c1 \neq c2$

$$t_s = \frac{\overline{X}_{-R_n:\hat{I}_{ni}} - \overline{X}_{\hat{I}_{ni}+1:R_n}}{\sqrt{s_p^2 \left(\frac{1}{\hat{I}_{ni} + R_n + 1} + \frac{1}{R_n - \hat{I}_{ni}} \right)}}$$

where

$$s_p^2 = \frac{1}{2R_n - 1} \left(\sum_{j=-R_n}^{\hat{I}_{ni}} \left(X_j - \overline{X}_{-R_n:\hat{I}_{ni}} \right)^2 + \sum_{j=\hat{I}_{ni}+1}^{R_n} \left(X_j - \overline{X}_{\hat{I}_{ni}+1:R_n} \right)^2 \right)$$

If we assume that the error terms are normally distributed (or alternatively, if the number of points in each average is large), then t_s has (approximately) a t-distribution with $2R_n - 1$ degrees of freedom. Therefore, we reject H_0 at level α and conclude that there is significant evidence of a break in the neighbourhood if $|t_s| > t_{2R_n-1}(1-\alpha)$, the $(1-\alpha)$th quantile of the t_{2R_n-1} distribution.

When $\alpha = 0$, we never reject the null hypothesis and the moving average smooth is the result. On the contrary, when $\alpha = 1$, the null is always rejected and the not-so-smooth is the result. Thus, by varying the significance level we can control the amount of smoothing done. Because the amount of smoothing ranges between that of the moving average smoother and the not-so-smoother, we will appropriately name this modified smoothing method the *somewhat-smoother.*

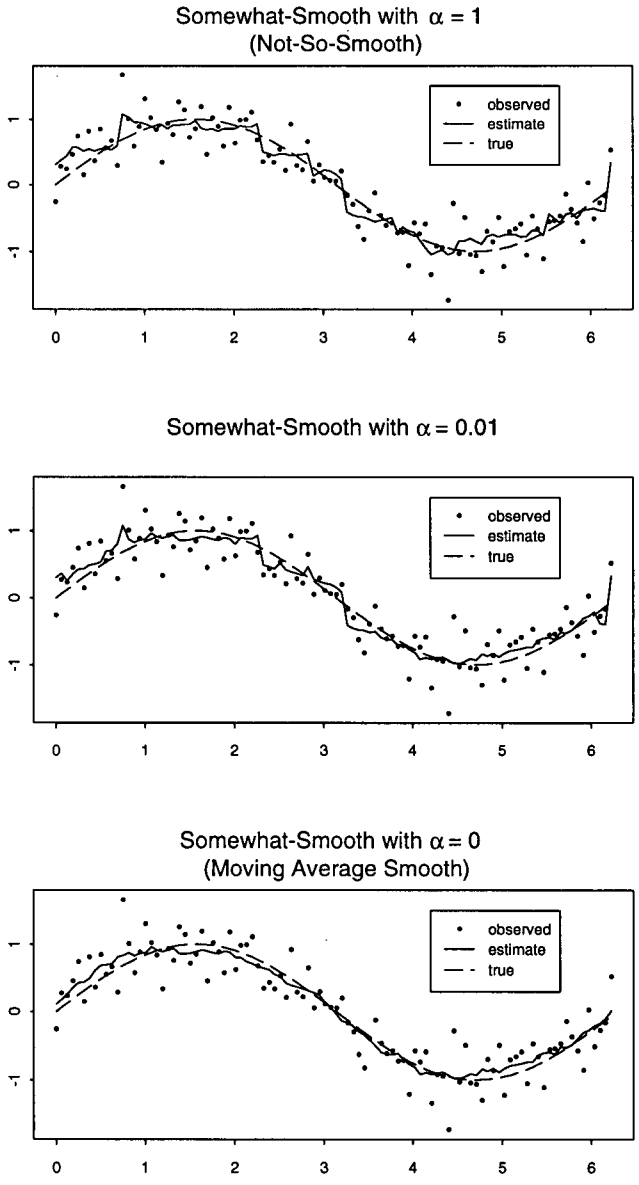


Figure 5.1: Somewhat-smooths of the sine data using $R_n = 5$ and various levels of α

In the previous chapter, we considered data generated from a sine curve with normally distributed noise. The not-so-smoother produced very rough output. The purpose of the somewhat-smoother is to improve upon this. To demonstrate, we use the same data shown in Figure 4.4 and apply the somewhat-smoother using α equal to 0, 0.01, and 1 and using a bandwidth of 5 in all cases. Graphs of the smooths are given in Figure 5.1 on the previous page. The somewhat-smoother with $\alpha = 0.01$ provides a less jagged estimate than the not-so-smoother ($\alpha = 1$). Of course, when the underlying function is continuous, as is the case with the sine data, the moving average smoother ($\alpha = 0$) still performs best.

However, when the possibility of a discontinuity exists, the somewhat-smoother allows for the preservation of the jump while producing smoother results in the continuous stretches. To illustrate, we will again consider a data set used previously, namely the split cube-root data as shown in Figure 4.5. This figure shows that the optimal not-so-smooth preserved the discontinuity perfectly but produced fairly rough output. Thus, we applied the somewhat-smoother to the data using the same bandwidth as the optimal not-so-smooth ($R_n = 9$) for the purpose of comparison. The level of significance used was $\alpha = 0.0001$. Although this level may seem very small, it was chosen because it produced quite smooth output yet the test still detected the jump. The somewhat-smooth is given in Figure 5.2, as is a replicate of the optimal not-so-smooth, and the results look much improved.

The drawback to using the somewhat-smoother is that there are now two parameters, R_n and α, to be selected. In our examples, α was chosen visually but, as with the selection of an optimal bandwidth, we would like to develop quantitative methods for choosing the optimal α value.

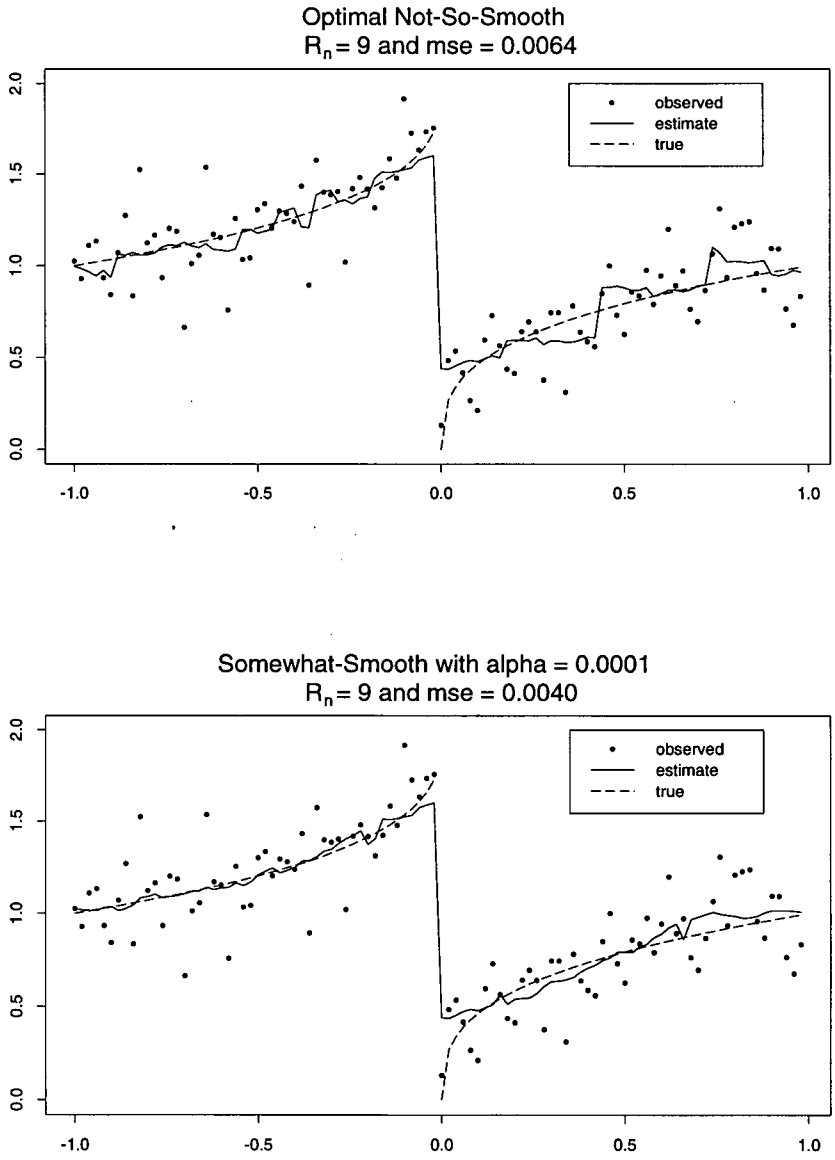


Figure 5.2: Not-so-smooth and somewhat-smooth with $\alpha = 0.0001$ of the split cube-root data using $R_n = 9$

5.2 The Not-So-Smoother using Local Linear Fits

There are situations when using local constant fits within each neighbourhood will not provide a good fit. As an example, we consider the "sawtooth" function used in the paper by McDonald and Owen as well as Hall and Titterington. The function consists of two line segments rising from 0 to 1 — one between 0 and 0.5 and the other between 0.5 and 1. There are 256 equally spaced data points with normal noise added. The standard deviation of the noise is taken to be half that of the function; that is $\sigma = 1/2 \left\{ \int_0^{0.5} (x - 0.5)^2 \, 2x \, dx + \int_{0.5}^1 (x - 0.5)^2 (2x - 1) \, dx \right\}^{1/2} = 1/2 \sqrt{1/12}$. Because the true function is nowhere approximately constant, applying the not-so-smoother gives a very jagged estimate. Figure 5.3 shows the optimal not-so-smooth according to the minimum mse criterion. Although it is less jagged than the optimal moving average smooth would be, we would still prefer a smoother estimate. Using the somewhat-smoother would only marginally improve the results because it still assumes that the local two-constant model holds.

This example leads us naturally to consider using local linear fits rather than local constant fits. Within each neighbourhood, the two lines which minimize the mean squared error are found. This involves minimizing the function $H^*_{nk_n}(J, \alpha_1, \alpha_2, \beta_1, \beta_2)$ with respect to its five parameters where

$$H^*_{nk_n}(J, \alpha_1, \alpha_2, \beta_1, \beta_2) = \frac{1}{R_n} \left(\sum_{j=-R_n}^{J} (X_{n,k_n+j} - \alpha_1 - \beta_1 \, t_{n,k_n+j})^2 + \sum_{j=J+1}^{R_n} (X_{n,k_n+j} - \alpha_2 - \beta_2 \, t_{n,k_n+j})^2 \right).$$

The value of J which minimizes the function is the estimated breakpoint, the values of α_1 and β_1 which minimize the function are estimates of the intercept and slope of the first line respectively, and the values of α_2 and β_2 which minimize the function are estimates of the intercept and slope of the second line respectively.

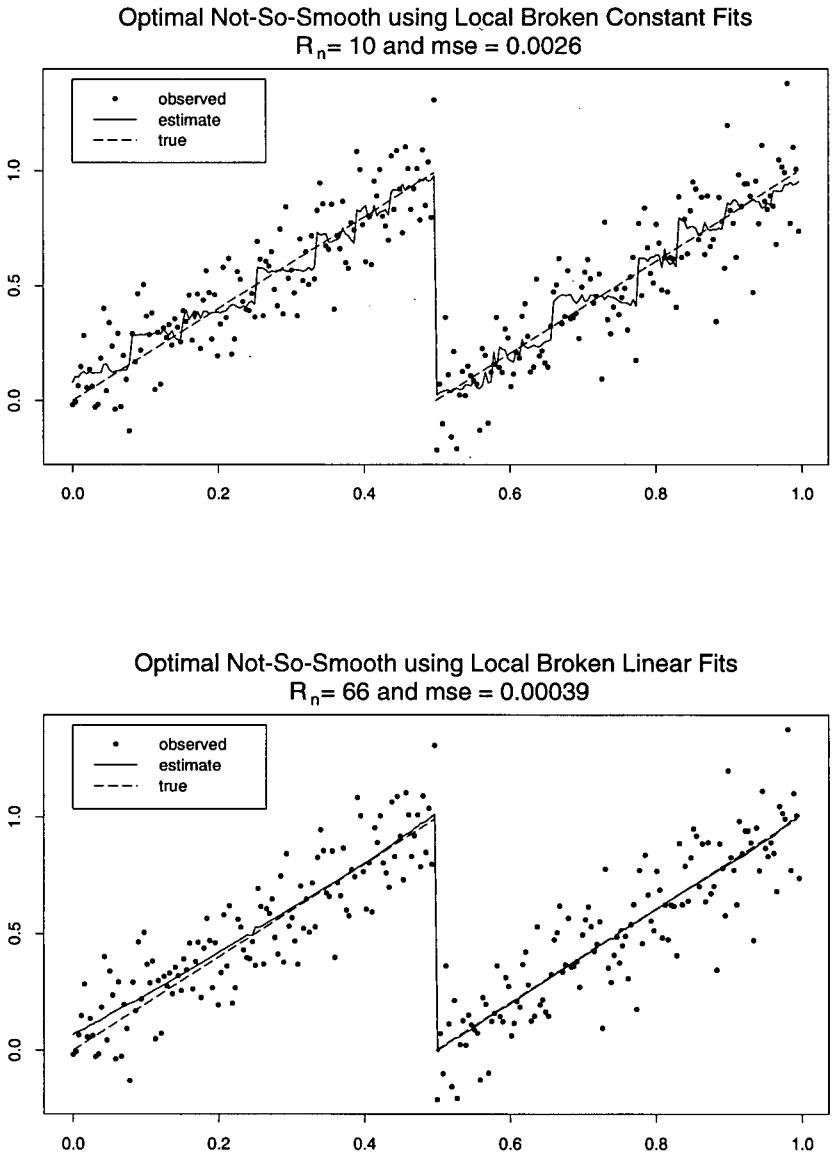


Figure 5.3: Optimal not-so-smooths of the sawtooth function with noise using local constant fits and local linear fits

The not-so-smoother using local linear fits was applied to the sawtooth data and the optimal smooth was found (see Figure 5.3). The benefits of using linear rather than constant fits is evident in this case. The estimated function is almost indistinguishable from the true function over most of the range.

Because the sawtooth function consists of two line segments, it is not surprising that using linear fits leads to a good estimate. However, this method involves estimating a greater number of parameters. If the optimal neighbourhood size using linear fits is sufficiently larger than that using constant fits, then no loss incurs from the additional parameter estimation. Whether this will in fact be the case depends on the data. When the data are such that the locally constant model is approximately true, the original not-so-smoother or somewhat-smoother may be preferable, but otherwise, using the not-so-smoother with local linear fits is likely to give better results.

Chapter 6

Conclusion and Discussion

In Chapter 2, we proposed a smoothing method which, when applied in situations where the regression function being estimated is discontinuous, is designed to preserve the discontinuities. This smoother, termed the not-so-smoother, was shown to be consistent under very general conditions.

The performance of the not-so-smoother, both locally (within a neighbourhood) and globally (on an entire data set), was thoroughly investigated. Using simulated data for which the underlying function was known, we could evaluate the accuracy with which the discontinuities were located and the function was estimated. Overall, the not-so-smoother was successful at estimating functions with discontinuities, with greater success as the number of points within the neighbourhood increased and as the ratio of the size of the discontinuity to the variability in the data increased.

The performance of a smoothing method is highly dependent on the choice of the neighbourhood size, or bandwidth. By using simulated data sets, we were able to determine the optimal bandwidth, meaning the one which minimized the mean squared error. Thus we could compare the optimal performance of the not-so-smoother with the optimal performance of a moving average smoother. When the regression function was continuous, the moving average smoother performed better, as the not-so-smoother produced a much rougher estimate. When the regression function was discontinuous, the relative performance of the two methods depended on the ratio of the discontinuity

size to the standard deviation in the data. Our simulations suggest that when the ratio is less than 1, the moving average smoother performs better. As long as the ratio is greater than 1, the not-so-smoother is superior, becoming increasingly better as the ratio increases. In this case, not only does the optimal not-so-smooth preserve the edges more sharply than the optimal moving average smooth, but it also produces smoother output since the optimal moving average smooth requires such a small bandwidth that it is very jumpy. Note that we are evaluating the smooths based solely on minimum mean squared error criterion; if the ability to preserve an edge was the criterion used, then the not-so-smoother would be preferable whenever a discontinuity is present.

The not-so-smoother is designed such that it assumes, essentially, that a discontinuity exists within each neighbourhood. This property can result in jagged output over continuous segments of the data, and a desire to reduce this jaggedness led us to consider a modified method called the somewhat-smoother. The somewhat-smoother, presented in Chapter 5, combines the not-so-smoother with the moving average smoother. A test is performed within each neighbourhood to determine if the data provide sufficient evidence that a discontinuity exists. If so, the not-so-smooth estimate of the point is used; otherwise, the moving average estimate is used. Examples illustrated that the somewhat-smoother can achieve the goal of producing smooth output while still preserving the discontinuities.

We considered another modification to the not-so-smoother in which the best piecewise linear, rather than constant, function with exactly one simple discontinuity is identified within each neighbourhood. When the regression function is not approximately flat, even within small segments, as with the sawtooth function, using constant fits gives a rough estimate. Using linear fits can improve the results greatly, but requires more parameter estimation.

In summary, if it is known that the underlying function is continuous, then there is

no advantage to applying the not-so-smoother. The not-so-smoother will not perform as well as, say, a moving average smoother, and computationally it takes much longer. However, if it is suspected that the function being estimated is discontinuous, or if no information about the function is available, then using the not-so-smoother or one of its modifications can be highly beneficial.

As always, there is further work to be done. Consistency of the not-so-smoother was proven, but more work is required to determine the asymptotic behaviour. We expect that the not-so-smoother will converge much faster than the traditional smoothing techniques when there is a discontinuity, and at the same speed when there is not.

Bandwidth selection was mentioned briefly in Chapter 4. How to choose the best bandwidth first requires a criterion on which to base evaluation. Typically mean squared error is used, but perhaps this does not tell the whole story in the case of discontinuous regression functions. When the size of the discontinuity is small, a moving average smooth may have a smaller mean squared error than a not-so-smooth even though the edge is defined more sharply with the latter. Depending on the situation, edge preservation may be more important. Next, an evaluation criterion involves knowing the true function, and since the function is unknown, methods of dealing with this problem must be developed. Traditionally, cross-validation and plug-in techniques are used in the selection of a bandwidth, and ways to apply these techniques in the discontinuous case can be derived.

Lastly, the not-so-smoother was introduced in a one-dimensional setting where one might argue that the discontinuities can be identified visually, at least for some data sets. When a higher dimensional function is being estimated, discontinuities would be very difficult to identify without some quantitative technique such as a smoother. The not-so-smoother can be extended quite naturally to include higher dimensional cases. Consider a two-dimensional function. At each point in the sample, define its neighbourhood to

be all points within a fixed area around the point. For simplicity, the area could be taken to be a square. Then find the line dividing the neighbourhood into the two subsets which minimize the distance between the mean and the data of the first subset plus the distance between the mean and the data of the second subset. To consider all possible lines could be lengthy and not even desirable (since very curvy optimal lines would likely result), so a subset such as all vertical, horizontal and diagonal lines could be considered. After determining the "best" line, the estimate of the point under consideration would be the mean of the subset to which it belongs. Modifications analagous to the somewhat-smoother and the use of linear fits in the one-dimensional case can also be extended to higher dimensions.

Bibliography

[1] Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics.* Prentice Hall, New Jersey.

[2] Chung, K. L. (1974). *A Course in Probability Theory.* 2nd edition. (Probability and Mathematical Statistics: A Series of Monographs and Textbooks.) Academic Press, New York.

[3] Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* (Statistics, Textbooks and Monographs, 90.) M. Dekker, New York.

[4] Feder, P. I. (1975). On Asymptotic Distribution Theory in Segmented Regression Problems – Identified Case. *The Annals of Statistics* **3** 49-83.

[5] Fredkin, D. R. and Rice, J. A. (1990). Bayesian Restoration of Single Channel Patch Clamp Recordings. University of California, San Diego.

[6] Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* 1st edition. (Monographs on Statistics and Applied Probability, 58.) Chapman and Hall, New York.

[7] Hall, P. and Titterington, D. M. (1992). Edge-Preserving and Peak-Preserving Smoothing. *Technometrics* **34** 429-440.

[8] Lee, D. (1991). Detection, Classification, and Measurement of Discontinuities. *SIAM Journal of Scientific and Statistical Computing* **12** 311-341.

[9] McDonald, J. A. and Owen, A. B. (1986). Smoothing with Split Linear Fits. *Technometrics* **28** 195-208.

[10] Müller, H. G. (1992). Change-Points in Nonparametric Regression Analysis. *The Annals of Statistics* **20** 737-761.

[11] Shiau, J. H. (1987). A Note of MSE Coverage Intervals in a Partial Spline Model. *Communications in Statistics – Theory and Methods* **16** 1851-1866.