11-12-2018 3:00 PM

Bias Assessment and Reduction in Kernel Smoothing

Wenkai Ma, *The University of Western Ontario*

Supervisor: Braun, W. John, *The University of Western Ontario*
Co-Supervisor: Davison, Matt, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Biostatistics

Abstract

When performing local polynomial regression (LPR) with kernel smoothing, the choice of the smoothing parameter, or bandwidth, is critical. The performance of the method is often evaluated using the Mean Square Error (MSE). Bias and variance are two components of MSE. Kernel methods are known to exhibit varying degrees of bias. Boundary effects and data sparsity issues are two potential problems to watch for.

There is a need for a tool to visually assess the potential bias when applying kernel smooths to a given scatterplot of data. In this dissertation, we propose pointwise confidence intervals for bias and demonstrate a software tool to implement the confidence bands in practice. The effectiveness of the proposed bias assessment tool is demonstrated using simulated data and is illustrated by its application to classical data sets.

To reduce the bias of LPR while keeping other good properties of it, as well as to mitigate the sparsity and boundary issues, this thesis extends the technique of double-smoothing from the local linear regression context to higher order local polynomial regression, with particular focus on local quadratic and local cubic regression. Double-smoothing is a technique that involves two levels of smoothing and is known to reduce bias in nonparametric regression while maintaining control over variance. What was not known is that the method is often sub-optimal when the bandwidths at both levels of smoothing are equal. In this thesis, we propose a simple method for obtaining the second-level smoothing bandwidth while using a cross-validation method for the first level.

The proposed tools in this dissertation are employed to address real problems related to wildfire management: one is for modelling the time to initial attack using fire case records for the years from 1930 to 2012 in a Northeastern Ontario area. Another one is for the experimental results of burning debris as a randomized component in a firebrand spotting simulator.

Keywords: Local Polynomial, bias assessment, bias reduction, double-smoothing

This thesis is dedicated to my family

for their love, support and encouragement.

Acknowledgements

I would like to thank my supervisor Dr. John Braun for his relentless effort and continuous guidance. Without his encouragement and support, this work would have not any possibility. My appreciation also goes to faculty members in the Department of Statistical and Actuarial Sciences at Western University, for their academic support including but not limited to Dr. Matt Davison, Dr. Duncan Murdoch, Dr. Serge Provost, Dr. Wenqing He, Dr. Reg Kulperger, Dr. Hao Yu. and Dr. Ian McLeod. Being the Teaching Assistant of Dr. Bethany White, Ms. Mary Millard and Steve Kopp was an honor to me. The help from staff members including Ms. Jennifer Dungavell, Ms. Jane Bai, Ms. Lisa Hunt and Ms. Erin Woolnouogh is appreciated. Great appreciation is extended to the thesis examiners (Alphabetically ordered): Dr. Yun-Hee Choi, Dr. Serge Provost, Dr. Zilin Wang and Dr. Douglas Woolford. I would also like to thank my fellow colleagues in Western for their friendship and support.

I acknowledge the use of data products from the Ontario Ministry of Natural Resources and Forestry, compiled by Dr. David Martell of the University of Toronto and his helpful discussions. I also appreciate the conversations about fire-spotting that I had with Dr. Greg Kopp of the Boundary Layer Wind Tunnel.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Local Polynomial Regression (LPR) is a set of data visualization procedures designed to fit flexible curves to noisy data where the objective is to smooth out the noise but retain what is believed to be the underlying signal. Thus, applying LPR is often a first step in exploratory data analysis, in that it has potential to provide important information about the process by which the data were generated, such as monotonicity, amount of curvature, flat or sharp peaks and so on.

The main challenge in LPR is the degree of smoothing to apply. The Mean Square Error (MSE) is usually employed to assess LPR performance. Bias and variance are the two components of MSE, and LPR has elements of both. Minimizing MSE corresponds to balancing bias and variance. When a scatterplot is over-fitted, the bias term is large and the variance is small. When a scatterplot is under-fitted, the bias term is small and the variance is large. This is called the bias-variance tradeoff.

Currently, there is no software tool which can be used to indicate a potential bias problem in a given LPR application in any of the literature, as far as we know. In this thesis, we present a new diagnostic tool to accompany a local polynomial regression plot to signal how much bias

there may be in the fitted curve. With fitting onto real study data, the effectiveness of the tool will be demonstrated.

Since severe and variable levels of bias can lead to distorted views of the curve underlying the data, a number of bias-reduction strategies have been proposed in recent decades, including the double-smoothing approach of He and Huang (2009). This method was to reduce bias in local linear regression, a special case of LPR. Extending double-smoothing to higher degrees in LPR to achieve bias reduction is another objective of this thesis.

This thesis is also motivated by several real-life statistical challenges in the wildland fire management environment. The first one is about a trend hypothesis for the initial attack problem in an area of Northeastern Ontario (more details in Section 2.2). Another one is related to a mechanism to simulate firebrand spotting distances for a given wildland fire (more details in Section 2.3).

1.2 Outline of Thesis

In Chapter 2, we provide a review of the relevant literature for the thesis. In Chapter 3, we propose a bias assessment tool for local polynomial regression, and in Chapter 4, we show how double-smoothing can be extended to higher degree local polynomial regression. Chapters 5 and 6 are concerned with application of our techniques to the two fire management applications referred to above.

Chapter 2

Background and Literature Review

2.1 Local Polynomial Regression

2.1.1 Definitions

Local Polynomial Regression (LPR) is a method of nonparametrically smoothing bivariate scatter plots. Nadaraya (1964) and Watson (1964) were early pioneers of this method, focusing on what is now recognized to be local constant regression. In LPR, a weighted least squares (WLS) polynomial is fit at each point of interest, x using data within a neighborhood of x. The books by Wand and Jones (1994) and Fan and Gijbels (1996) provide excellent introductions to the methodology.

Applications of the method abound. For example, the paper by Li et al. (2003) studied the usage of LPR on estimating average derivatives.

Suppose X and Y are two random variables, from which independent data are collected, denoted as $(x_i, y_i), i = 1, 2, ..., n$. We let $\mathbf{x}$ and $\mathbf{y}$ denote column vectors whose elements are x_i and y_i, respectively.

The model we consider is

$$y = g(x) + \varepsilon \tag{2.1}$$

where $g(x)$ is assumed to be smooth enough for a local polynomial estimator to be consistent, as described below and in Wand and Jones (1994) for example. The ε term is assumed to be a random noise element with mean 0 and variance σ^2.

The local pth degree polynomial estimator is (Fan and Gijbels, 1996)

$$\widehat{g_1}(x) = e_1^T (X_x^T K_x X_x)^{-1} X_x^T K_x \mathbf{y} \tag{2.2}$$

where e_1^T is the first standard $(p + 1)$ dimensional unit vector, i.e. $e_1^T = [1\ 0\ 0\ \cdots\ 0]$, K_x is a diagonal matrix with ith diagonal element as $(K_h(x_i - x))$. K_h is a symmetric density function with scale parameter h; h is referred to as the bandwidth or smoothing parameter. The kernel could be chosen to have compact support (as in the cases of uniform, triangular, or Epanechnikov) or non-compact support (as in the case of gaussian).

The bandwidth could be fixed or adaptive (Fan and Gijbels, 1995b; Loader, 1999). When the bandwidth is large, the bias is large and the variance is small. On the other hand, if the bandwidth is small, the bias is smaller and the variance is larger. The preference usually depends on the context. Bias reduction, while keeping the asymptotic variance to a similar order of magnitude, has been studied widely (Choi and Hall, 1998, 1999; He and Huang, 2009).

The local polynomial approach readily extends to estimation of derivatives of the regression function, when they exist (Fan and Gijbels, 1996). In this case,

$$\widehat{\frac{dg_1^k}{dx^k}}(x) = k! e_k^T (X_x^T K_x X_x)^{-1} X_x^T K_x \mathbf{y} \tag{2.3}$$

where $e_{(k+1)}^T$ is the $k + 1$th standard unit vector, with $k \in \{1, 2, \ldots, p\}$.

Other smoothing techniques exist as well. Spline smoothing is a popular method. Compared with LPR, spline smoothing requires a choice of knots which is computationally intensive. LPR has the advantage of simplicity in implementation since only 1 bandwidth is needed for one-dimensional problems. Eilers and Marx (1996) introduced the method of penalized splines which goes some way to reduce the effects of poor knot choice. Generalized additive

models (GAM), e.g. Wood (2006), employ the methodology of splines as well as kernel methods, and allows for ready extension of one-dimensional techniques to higher dimensional data. Functional data analysis, e.g. Ramsay et al. (2009), is a set of techniques for handling more complex data structures, including random effects, longitudinal and multilevel designs, again largely using splines.

2.1.2 Double-Smoothing as a Bias Reduction Strategy

The paper of He and Huang (2009) proposed double-smoothing for local linear regression with the purpose of achieving bias reduction. In contrast to using only the intercept estimates in local linear regression as in (2.2) with $p = 1$, double-smoothing makes use of both the intercept and slope estimates, where the slope estimate is obtained from (2.3) with $k = 1$ and $p = 1$. This has been done by employing a weighted integral as a second step on the fitted values at each gridpoint t of the fitted lines from a first level of smoothing.

The essential idea is that we can estimate $g(x)$, with varying levels of accuracy, using the linear least-squares estimate $\widehat{g}_{x_0}(x) = \widehat{\beta}_{0,x_0} + \widehat{\beta}_{1,x_0}(x - x_0)$, for any real value of x_0. When $x_0 = x$, we recover the conventional estimate, but for other values of x_0, particularly those near x, we can realistically hope to obtain useful alternate estimators for $g(x)$. Thus, a weighted average of $\widehat{g}_{x_0}(x)$, indexed by x_0, and where higher weights are associated with values of x_0 near x, has potential for improved accuracy over the conventional estimate. The integral referred to at the end of the preceding paragraph essentially computes this weighted average, and He and Huang (2009) rigorously demonstrated the improvement in asymptotic accuracy for this technique.

He and Huang (2009) also employed simulation to study the finite-sample performance of double-smoothing local linear regression. It was compared with the earlier proposal of Choi and Hall (1998) as well as conventional local linear and local cubic regression using different target functions with various levels of noise to signal ratio. The conclusion was that the bias reduction effect of double-smoothing is considerable, with little effect on variance. The method proposed by Choi and Hall (1998) is a special case of He and Huang (2009). An analogous

idea is adapted to the wavelet estimator in the study of Chen (2011).

2.1.3 Data Sharpening as a Bias Reduction Strategy

Data sharpening procedures perturb data in order to improve estimation performance of an estimator over the performance of that same estimator on the raw data. The focus has often been on bias reduction. Choi and Hall (1999) introduced data sharpening for kernel density estimation (KDE). Choi et al. (2000) considered data sharpening on low order kernel regression. The study of Braun and Hall (2001) extended the technique so that the good properties of data sharpening are retained and certain constraints can be imposed, such as unimodality or monotonicity. The R package CHsharp (Woolford and Braun, 2015) implements the idea of Choi and Hall (1999).

2.2 Initial Attack Problem

Wildland fires, due to a variety of causes including lightning, occur in many locations around the world, including Canada. When a fire is discovered and reported, there is usually a delay until a fire-fighting crew arrives on the scene to begin the task of extinguishing the fire. The response time is the time of report to the initial fire fighting (initial attack) time.

The length of the response time can be an indicator of the effectiveness of fire management strategies in a region or province. It is reasonable to surmise that shorter response times might be related to fewer large, long-lasting fires. Thus, examining the historical record of response times is of interest in fire management. One specific objective is to determine whether there was a significant change in response time from a wildland fire management perspective around the year 1950 when approaches to fire suppression changed in the Province of Ontario.

The data we consider are from fire case logs in a study area in northeastern Ontario from the year of 1930 to the year of 2012. The data source was the Ontario Ministry of Natural Resources and Forestry (OMNRF), compiled by Dr. David Martell of the University of Toronto.

This dataset was produced by merging data from two individual databases. The first database includes data on fires reported in a Northeastern (NE) area of Ontario (see Figures 2.1 and 2.2 for study area) from the years 1930 to 1959, and the second includes data from OMNRF's digital fire archive, which covers the whole province of Ontario from 1960 to 2012. The OMNRF digital fire archive data were therefore clipped so that only fires occurring within the northeastern study area were included in the current dataset. Fire management zone data (FMZ) were obtained using OMNRF shapefiles and an FMZ was attributed to every fire included in this data using Geographic Information Systems (GIS).

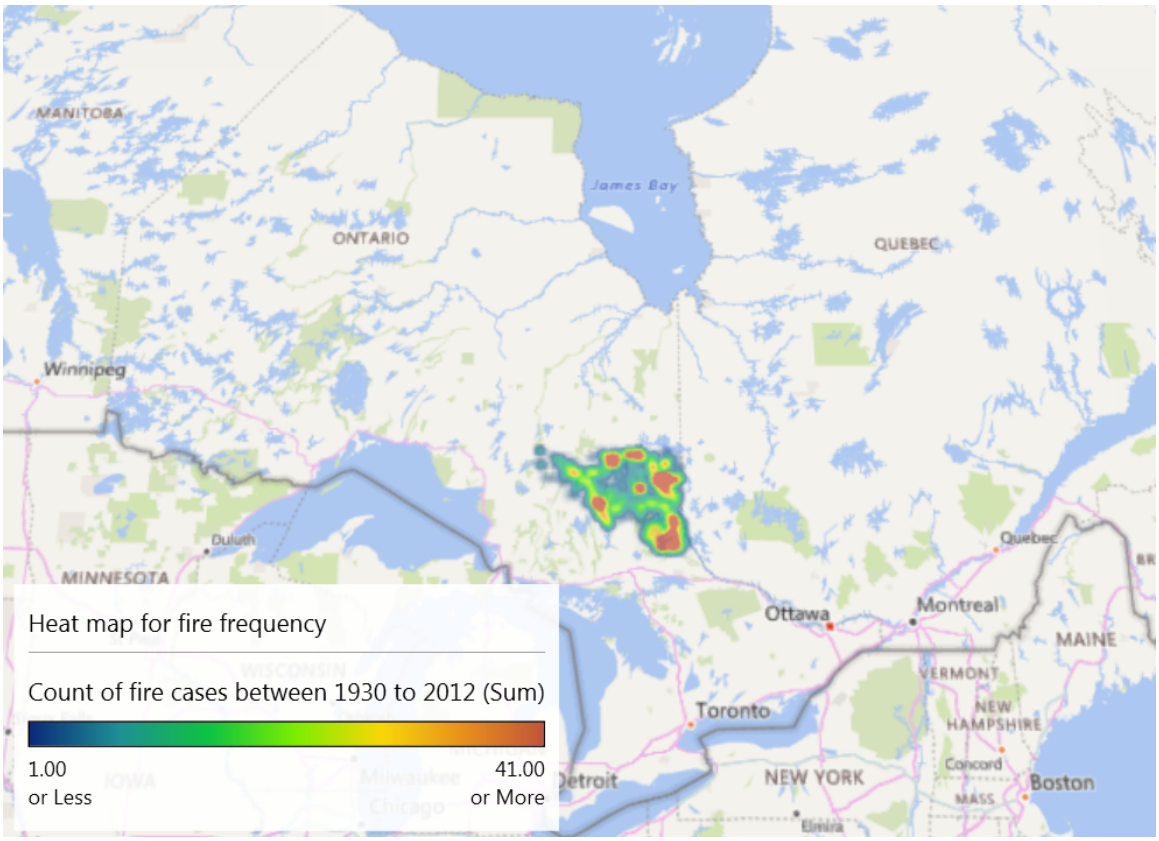

Figure 2.1: Heat map of fire cases under study in northeastern Ontario. Red areas are with high fire frequency (up to 41 fire cases at the same longitude and latitude), and green areas are less frequent.

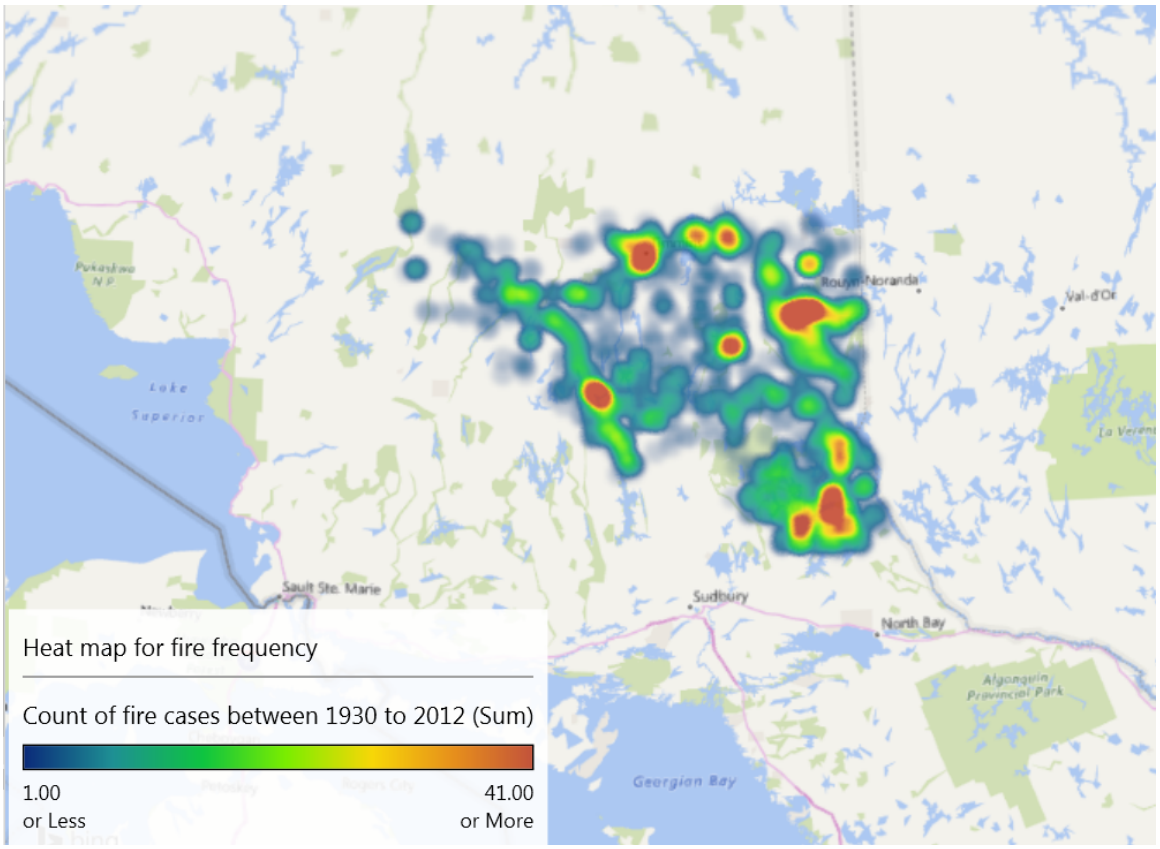

Figure 2.2: Heat map of fire cases under study (zooming in). Red areas are with high fire frequency (up to 41 fire cases at the same longitude and latitude), and green areas are less frequent.

There were 9,187 fire cases reported for that area during the study period.

The time stamps for the sequence of events log was documented. And the events include the following activities.

1. Discovery : the forest fire was first detected by the first reporting agency, either on the basis of a member of the public or by wildland fire management staff.

2. Report : the fire was reported to the fire headquarters or organized initial response group.

3. Fire fighter get away: the first initial attack group left their base location.

4. Fighting began : the initial attack began.

5. Fire being held: the fire was classified as being held (not likely to further spread).

6. Fire under control: the fire was declared under control for the final time.

7. Fire fight ended: by the late 1930s, Ontario Department of Lands and Forests (ODLF, predecessor to OMNRF) changed the name of this field to "Fire Out" but still referred to the same event type.

8. Fire out: the fire was declared to be out.

In the actual fire records, the event flow shows some departure from the standard event process outlined above, primarily because of errors in event-logging for various reasons, some of which we can identify. In Chapter 5, we will explain more on the details of the procedure on how different data issue scenarios were handled.

Using an integrated R packages suite for analyzing process workflow, bupaR (stands for Business Process Analysis Using R, (Janssenswillen, 2018) to examine the fire events, the event pathways or traces (meaning the process from the first event log of a fire case, to the last event log of the same fire case) of fires can be identified using different process mining methods. For the purpose of introduction, some brief summaries of the fires are explained as follows.

The Figure 2.3 shows all possible traces or pathways of those fire cases following the time order documented in the record. Since there are many types of departure from the expected event flow, clear patterns cannot be seen in Figure 2.3. Hence, the most frequent (80% by volume of the pathways or traces) fire events workflow patterns in Figure 2.4 are utilized to generate insights.

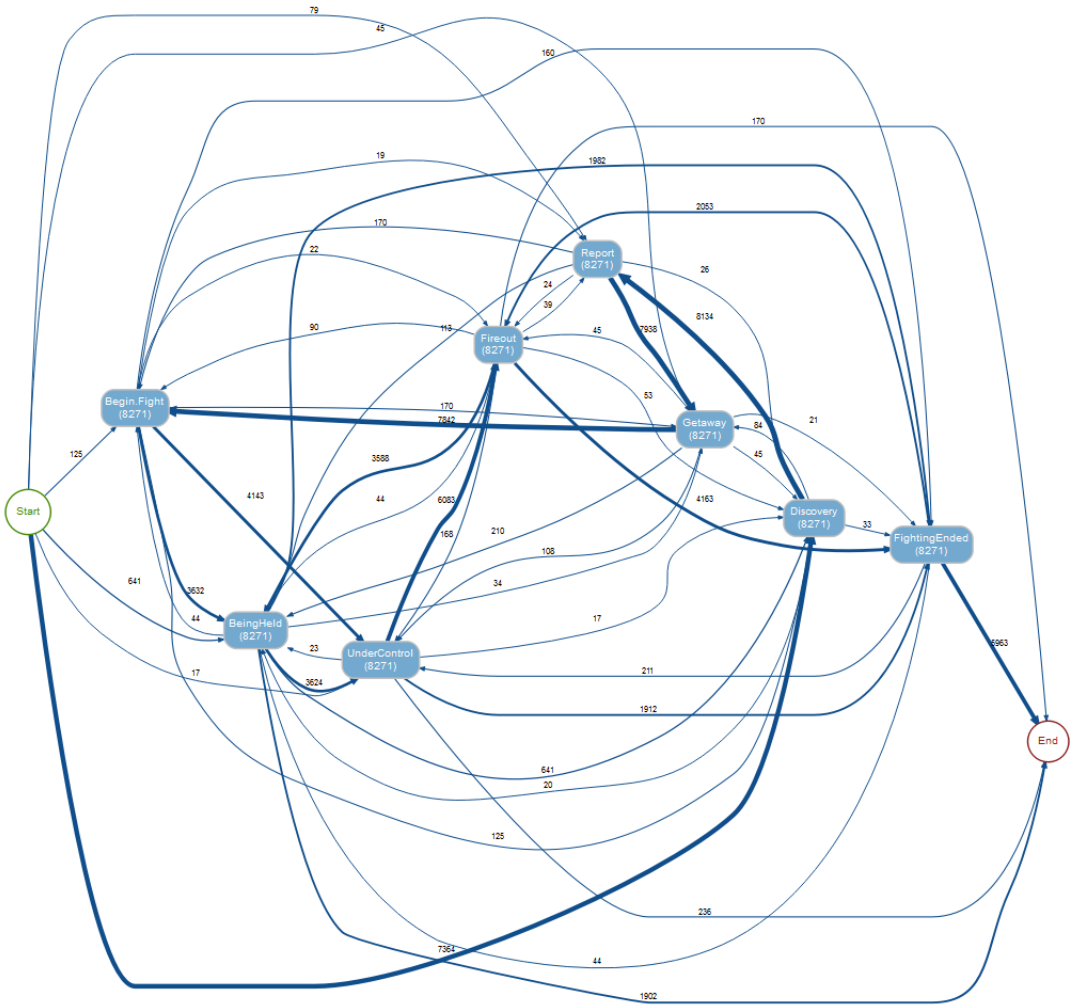

Figure 2.3: All possible pathways of the fire cases. A blue box is an event in the management process, a line is a path in time and arrows suggest the chronological order. The numbers on the line are the counts of the fire cases going through that path. Darker line represent larger numbers of fire cases.

Another perspective on these traces is displayed in Fig.2.5, showing the percentage of associated traces occur in the whole collection of fire management records.

The tools for visualizing event traces from Fig 2.3 to 2.5 can help a researcher investigate data quality and related data issues.

In wildland fire management, the initial attack strategy is the first effort of a firefighting crew to assess a fire and complete the tasks needed to manage a fire. The Ontario Ministry of Natural Resources and Forestry (OMNRF, 2018) lays out the initial attack steps. A “hit-them-

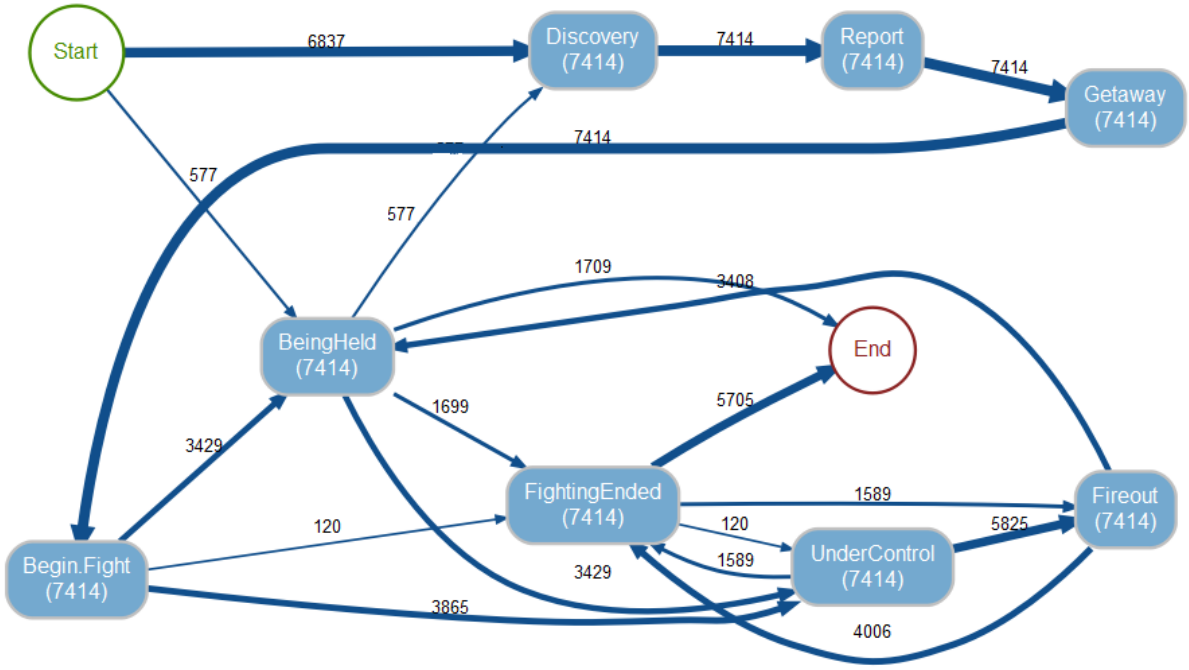

Figure 2.4: Most frequent (covering 80% of fire cases) pathways of the fire cases for clearer pattern discovery.

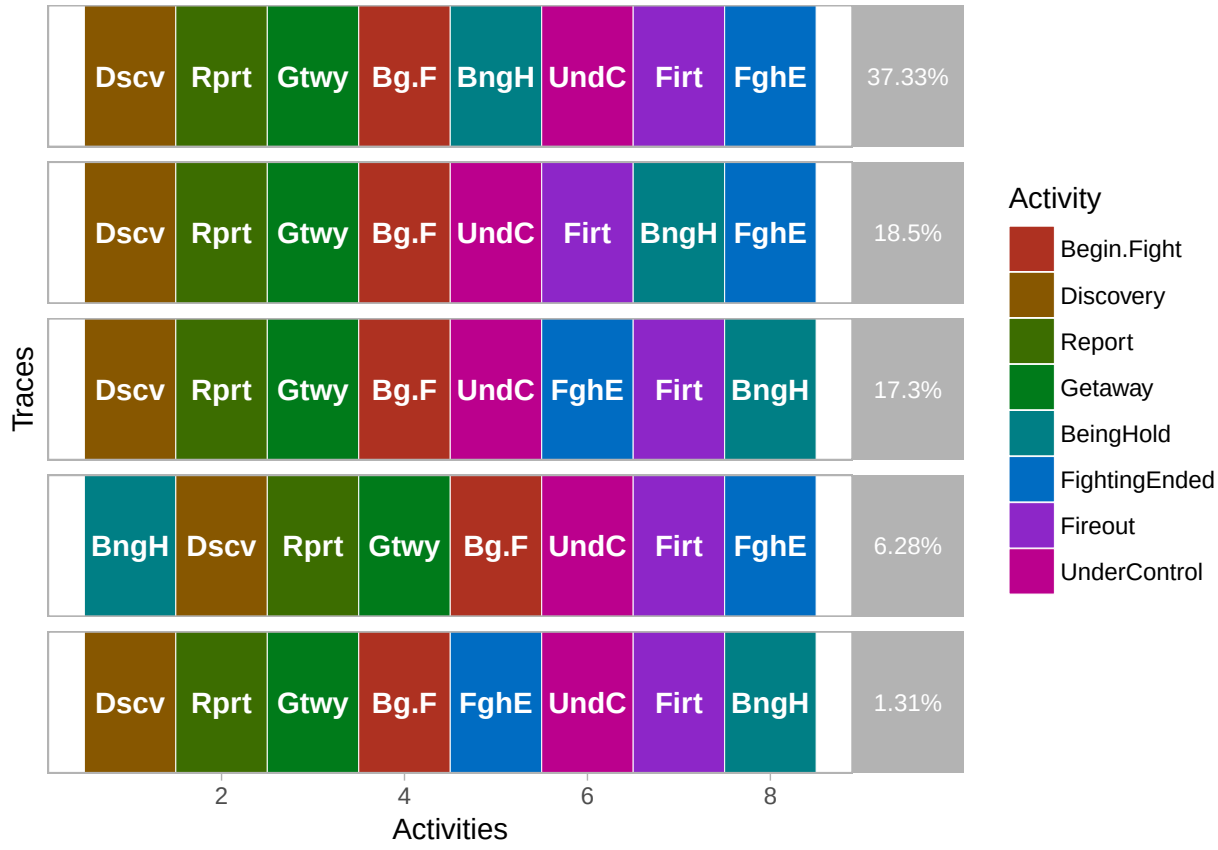

Figure 2.5: Pathways of the events associated with the fire management process. Each row is a trace or pathway of fire cases, and the number in the last column is the percentage of occurrences of the associated pathway in all of the fire data records under study.

hard, hit-them-fast" approach (Canadian Council of Forest Ministers, 2005) is critical for fire control objectives.

The initial attack time is defined as the time difference between fire report time and the time at which a crew begins to suppress the fire. According to the event log visualization, the order is following the correct logical order for almost all of the fire cases.

There was a belief that around 1950, due to the increased usage of airplanes (Stocks, 2013), the initial attack time should be reduced. In Chapter 5, we investigate whether this is a believable statement. The research hypothesis we will investigate relates to a trend change in the median initial attack time around the year 1950.

2.3 Spotting in the Context of Fire Growth Models

Spotting is a dangerous and unpredictable feature of wildfire where firebrands (hot wood embers) are sent aloft by the convection column and carried by the wind, often starting new fires (Albini, 1979). Physical spotfire models are required by fire management agencies, to ensure fire crew safety and to manage fire suppression resource deployment.

The prediction of spot fires involves two separate issues: Spotting distance distribution and probability of ignition by the firebrands.

Regarding spotting distance, some extreme spotting records from Canadian Forest Service Research officer Martin E. Alexander in a talk on the MITACS/GEOIDE conference in 2009 at Hinton, Alberta, Canada are listed as followed:

- Spot fire distance of about 5 km was reported during 2003 fire season in British Columbia (Alexander, 2009).

- On the fires near Gippsland, Victoria, Australia, in March 1965,

  > "One well authenticated spot fire started 29 kilometers ahead of the main fire. This is an Australian record."

(McArthur, 1968).

- Spotting for 34 kilometers happened on February 7th 2009 (McGourty, 2009), the famous catastrophe (known as Black Saturday) happened in Australia. In this series of fires, 505 hectares forest was burned because of ignition of spotting in Maroondah and Upper Yarra.

In fires at the Wildland Urban Interface (WUI), spotting is a very dangerous source of ignition of fires in structures. Spotting also reduces the effectiveness of fire prevention, and makes the control of prescribed burns more complicated.

An accurate stochastic spotting mechanism in a fire spread model can aid in prediction of forest fire propagation. The model proposed by Boychuk et al. (2009) incorporated spotting. Several assumptions from this study are as follows:

- When a firebrand becomes airborne, its time aloft is determined. They assumed an exponentially-distributed time aloft for simplicity.

- The likelihood a firebrand is still burning or smoldering upon landing is determined. They again assume that the time to burnout is exponentially distributed. Only if it has not burned out do we need to proceed to the next step.

- The landing location of a burning or smoldering firebrand is a function of its time aloft and the predominant wind direction, subject to a random jitter that is bivariate-normally distributed.

To underline the importance of the spotting mechanism within the context of fire spread simulation, we summarize the findings of Boychuk et al. (2009). Their model for fire growth which attempts to incorporate spotting explicitly is: $r_{(i,j)}(F,B) = \sum \lambda_S((i,j),x_{(k,l)},w,g)$ as the spotting rate, added to the original transition equation. $r_{(i,j)}(F,B)$ is the rate from mode 'Fuel (F)' to mode 'Burn (B)' for cell (i,j), λ is the transition rate, $x_{(k,l)}$ is the mode of cell (k,l), which is the neighbor of interested cell (i,j), w contains weather information and g includes

the terrain information. Based on rough physical grounds, they assumed the time a firebrand aloft was exponentially distributed with a mean of 10s, and the mean amount of time a firebrand alight was 100m, under a wind speed of 43 km/h. They obtained a mean aloft distance about 119.4 meters and standard deviation of 1 grid cell width (about 12.5 meters in both lateral directions). With the support of the spotting, the simulation for the Dogrib fire jumped Red Deer River as a barrier. For comparison purposes, we note that the Prometheus Wildland Fire Growth Model (Tymstra et al., 2010) is incapable of replicating this same behaviour.

2.3.1 Albini's Model for Maximal Spotting Distance

One possible model for spotting distance that is used by fire management agencies in North America is based on the work of Albini (1979), a deterministic model which provides an upper bound on the spotting distance. The problem addressed by Albini (1979) is under intermediate fire severity, and assumes that an individual tree or a group of trees is burning, to predict the maximum spotting distance, with other conditions being given. For example, tree diameter at breast height (D.B.H) is a factor. Maximum spotting distance is defined as the greatest possible distance that could be traveled by a lofted firebrand which remains burning. The approach in Albini (1979) was to examine and develop a mathematical description for each phase of the process of spotting. Albini states

> "A small group of trees torches out. The flame, and the buoyant plume above it, exist for a brief period of time. This fluid flow field is capable of lofting potential firebrands into air. The flow structures are described by separate, steady state models, joined at the tip of the flame. These models assume still ambient air for simplicity, and to ensure that a maximum height is predicted for the firebrand particles". (Albini, 1979, page 7)

The Maximum Spotting Distance Submodels and their outputs are:

- **Flame Structure Model**: flame length and gas flow velocity

- **Buoyant Plume Model**: fluid flow field in the buoyant plume
- **Firebrand Burning Rate Model**: time until firebrand burns out, given initial firebrand density, size and wind velocity
- **Firebrand Lofting Model**: height of the firebrand when it leaves the buoyant plume and is carried by the wind
- **Wind Field Model**: distance traveled by a firebrand until landing
- **Burning Firebrand Model**: maximum distance that could be traveled by a *burning* firebrand

Terrain would affect the flight distance. On flat terrain, the maximum horizontal distance X is given by (Albini, 1979)

$$X = \frac{2U_H}{v_0(0)} \frac{\sqrt{.1313Hz(0)}}{\log(7.62)} \left(\sqrt{\frac{z(0)}{.1313H}} \left(\log\left(\frac{z(0)}{.1313H} \right) - 2 \right) + 2 \right) \tag{2.4}$$

where H is the height of the trees, assuming it is homogeneous in a given area of forest. The numbers in this result were based on empirical analysis: 0.1313 is from assuming the roughness height (z_o) is at the height of $0.1313H$. U_H is the wind speed at height H. $z(t)$ is the height of the firebrand, $v_0(t)$ is the terminal velocity of the firebrand at time t. It is a function of time because when the brand is burning and flying, the size and mass of brand will decrease, which updates the terminal velocity of the firebrand:

- Albini shows that

$$v_0(0) = \sqrt{z(0)K\pi g/C_D} \tag{2.5}$$

  - K is the burning rate of the firebrand
  - C_D is the drag coefficient (assumed to be 1.2 to obtain maximum potential landing distance for a cylindrical firebrand)
  - $z(0)$ is the same as above, the height at initial time;

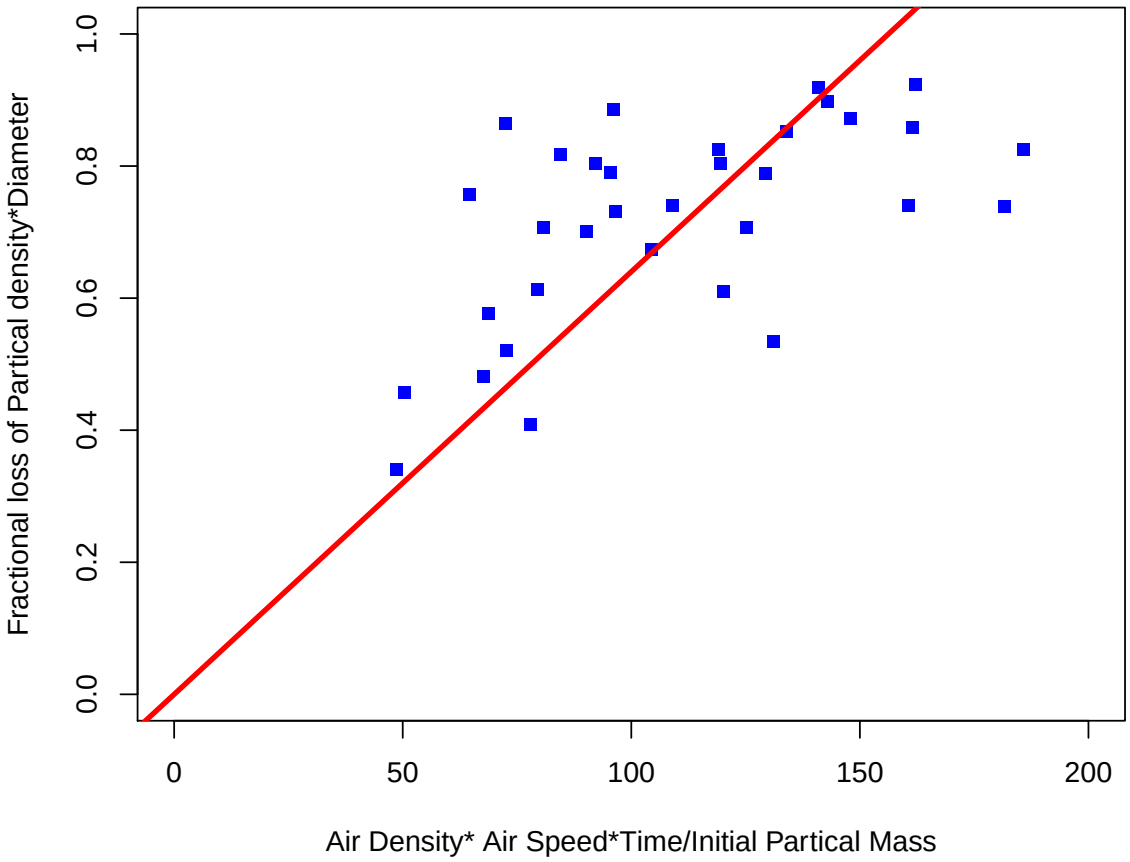

Figure 2.6: Fit of regression line on labratory burning data. Appendix C Albini, 1979.

- g is acceleration due to gravity

In the above mentioned parameters, the burning rate K is of interest to us. Albini (1979) employed a linear regression model for the decay rate. This approach is reasonable when the index and response are in the appropriate domain and range. Please refer to Fig 2.6, which shows a certain degree of goodness of fit even just by visualizing it. The issue is that the when burning time is long enough, the particle material will surely burn out. Under this scenario the linear decay model will not apply. The alternatives will be discussed in Chapter 6

The following example is from Albini's work: The fuel type is Douglas Fir, and other conditions are:

- Tree height is 30 meters
- Wind speed at tree top height is 36 km/h
- Flame height and duration about 70 meters and 4 seconds
- Initial firebrand height and velocity determined by submodels

Maximum spot distance over flat terrain about 0.98 miles was obtained by Albini's investigations.

2.3.2 Building on the Work of Albini

Some factors are omitted in Albini (1979). They are

- The likelihood of tree burning
- Availability of optimum firebrand material
- The probability of spot fire ignition
- The number of spot fires

Albini et al. (2012) proposed a mathematical model for predicting the maximum potential spotting distance from a crown fire. That paper refers only to level terrain, and provides an estimate of the maximum potential spotting distance from active crown fire as a function of the firebrand particle diameter at alighting based on three inputs, namely, canopy top height, free flame height (i.e. flame distance above the canopy top height), and wind speed at the height of the canopy. The model described in Albini et al. (2012) built upon a sub-modular approach similar to the approach to torching trees (Albini, 1979), a pile of wood (Albini, 1981b) and spreading surface fires (Albini, 1981a), but the core components describing the heat source

and buoyant plume dynamics in the 2012 study departs considerably from the researchers' earlier efforts. In earlier research, the flame duration was assumed to be short, and the particles climbed vertically first in the column above the original piles of trees before being released from the column. Under the new approach, the flame was no longer vertical. With ambient wind, the center-line of the plume inclined in the direction of wind most extremely at the top. The simulation result obtained in their research for maximum spotting distance was higher than earlier efforts without surprise considering their different physical basics: With 25 km/h wind speed at 10 meters height, a firebrand from a crown fire can maximally fly around 1.5 km (Albini et al., 2012) , while for a single torching tree, the comparable value was 0.5 km.

Sardoy et al. (2008) proposed a model tackling transport of burning brands by plumes above line fires in a crosswind. In particular, the characteristics of firebrands at landing and their ground distribution were investigated. Calculations were performed with disk-shaped firebrands for fire intensities and wind speeds representative of moderate- to high-intensity surface wildfire scenarios. Model results reveal a bimodal ground-level distribution of the released firebrands when both pyrolysis and char oxidation are present in the firebrand. Some of the brands, mostly in a flaming state, land at a short distance from the fire and others, in charring state, land at a long distance. The product $\rho_{w0}\tau_{f0}$ determines which firebrands will land in the short-or long-distance regions, where ρ_{w0} is the density of the particle, and τ_{f0} is the firebrand thickness. The "short-distance" firebrand distribution can be approximated by a lognormal function of the landing distance. Fixed release position and random release position for fire brands were both considered, and the results in two different scenarios remained consistent (Sardoy et al., 2008).

Some research on the landing of airborne debris, are relevant to spotting, in the sense of updating the drift distances of burning or smoldering embers after they are released from the buoyant plume. Tachikawa (1983) was a pioneer in the field of wind-borne debris who studied experimentally and numerically the trajectories of generic debris types. Tachikawa's contribution to this questions was recognized by Holmes et al. (2006). The parameter $k = \frac{\rho_a U^2 l}{2mg}$

defined by Tachikawa, representing the ratio of aerodynamic to gravity forces, is the main non-dimensional parameter determining the trajectories of debris items of all types.

In Wills et al. (2002)'s study, shapes of debris are classified into three categories: compact, rod and sheet. For 'compact' objects, only the drag force is working on the object, but not the lift effect. Auto-rotation influences spherical objects, but this will not affect the trajectory and landing distance of spherical objects. 'Rod' and 'sheet' like objects would be more complicated, since not only is the drag force considered, but also the lift force and rotation. Fire missiles would follow a similar classification: tree stem exposure would generate some compact objects, with spheres being the simplest compact objects. Firebrands are thought to be rod like objects. And sheet objects would be fragments of barks that are generated from strong wind and strong fire intensity.

Others researchers support the classification in the above study, while focusing on different shapes of debris. Visscher and Kopp (2007) commented that purely translational and auto-rotational modes lead to distinctly different lengths of flight and to different flight speeds, hence different impact energy. The researchers were dealing with panel objects and designed some experiments to get data for longitudinal and lateral distributions of flying panel objects. Turbulence was incorporated into their experiments. The spectral density of von Karman form for the horizontal component of the wind profile (Von Karman, 1948), and the Busch and Panofsky form for the vertical component are usually assumed (Busch and Panofsky, 1968).

Holmes (2004) focused on spherical objects. Three different kinds of scenarios were considered:

- Neglecting vertical air resistance
- Not neglecting vertical air resistance
- With effect of turbulence

The conclusion of that study was that the effect of vertical air resistance was significant and should be included to predict both horizontal and vertical displacements. With the settings in

the study of Holmes (2004), the effect of turbulence was not significant with respect to mean of debris landing distribution. However, from further experiments, Karimpour and Kaye (2012) confirmed that turbulence would play an important role when flying time is long enough. With a flame plume, the small object that fire generates would flow high with a probability distribution unless they are ejected from the plume or burn out before reaching the flame top. In this case, turbulence would significantly affect the horizontal displacement.

Lin et al. (2007) conducted some experiments on rod-like debris. With small angle of attack (0° to 30°), the horizontal displacement of rod-like debris would be similar to 2D plates. Empirical equations were fitted for mean horizontal displacement based on given traveling time.

Chapter 3

Bias Assessment in Local Regression

3.1 The Need for Visualizing Potential Bias in a Smoother

The problem of smoothing a scatterplot has been considered by numerous authors, going back to the pioneering work of Nadaraya (1964) and Watson (1964). Those papers introduced kernel methods for regression. The books by Wand and Jones (1994) and Fan and Gijbels (1996) provide excellent introductions to the methodology. Research continues in the area. For example, the paper by Li et al. (2003) studied the usage of local polynomial regression (LPR) on estimating average derivatives. Other research focused on the performance of LPR, and how to achieve bias reduction (e.g. Choi and Hall (1998) and He and Huang (2009)) and efficient computation (e.g. Samarov (2015)). Numerous packages exist for computing LPR in modern software. In R, the `locfit` (Loader, 2013) and `KernSmooth` packages (Wand, 2015), and the `loess` function in package `stats` are capable of fitting LPR (R Development Core Team 2018). In SAS version 9 (SAS Institute Inc, 2013), PROCedure LOESS is used to fit LPR.

The main challenge in LPR is how much smoothing to do. Mean Square Error (MSE) is usually employed to assess LPR performance. Bias and variance are the two components of MSE. When a scatterplot is over-fitted, the bias term is large and the variance is small. When a scatterplot is under-fitted, the bias term is small and the variance is large. This is called the bias-

variance tradeoff. Minimizing MSE corresponds to balancing bias and variance. Currently, however, there is no software tool which can be used to indicate a potential bias problem in a given LPR application in any of the literature, as far as we can tell.

In the present study, we establish a new diagnostic tool to accompany the local polynomial regression plot to signal how much bias there may be in the fitted curve. The organization of the chapter is as follows. After a brief review of local polynomial regression, we explain the methodology for pointwise confidence intervals of LPR fitting bias in Section 3.2. In Section 3.3 we discuss how the bandwidth selection procedure can be applied. We use simulations to demonstrate the use of our new tool; some examples illustrate the application of this proposed method. 3.4 contains real data analysis and diagnosis using our tool. The last section will wrap up our study with a conclusion and discussion.

3.2 Pointwise Confidence Intervals for Bias

To motivate the assessment plot technique, and to make the present work self-contained, it is useful to outline the main results of the local polynomial regression methodology. Suppose X and Y are two random variables, from which independent data are collected, denoted as $(x_i, y_i), i = 1, 2, ..., n$. x and y are vectors.

The model we consider is

$$\mathbf{y} = g(\mathbf{x}) + \varepsilon \tag{3.1}$$

where $g(\mathbf{x})$ is assumed to be smooth enough for a local polynomial estimator to be consistent.

The pth local polynomial estimator is (Fan and Gijbels, 1996)

$$\widehat{g_1}(x) = \mathbf{e_1}^T (X_x^T K_x X_x)^{-1} X_x^T K_x \mathbf{y} \tag{3.2}$$

where $\mathbf{e_1}^T = [1\ 0\ 0\ \cdots\ 0]$ is a $(p_1 + 1)$-vector.

K_x is a diagonal matrix with ith diagonal element as $(K_{h_1}(x_i - x))$, which is a symmetric density function, and h_1 is the bandwidth. The Gaussian kernel is a reasonable choice, though not necessarily the very best; the choice of kernel function does not have as big effect as the bandwidth h_1 and the degree of the polynomial k_1.

$$X_x = \begin{bmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^{k_1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x & \cdots & (x_n - x)^{k_1} \end{bmatrix}$$

where k_1 is the degree of the polynomials.

Confidence Interval for Bias and the Assessment Plot

The estimator of $g(x)$ at point x_0 is

$$\widehat{g}_1(x_0) = \mathbf{e_1}^T (X_{x_0}^{1T} K_x^1 X_{x_0}^1)^{-1} X_{x_0}^{1T} K_x^1 \mathbf{y} \tag{3.3}$$

$$= \mathbf{e_1}^T C_{x_0}^1 \mathbf{y} \tag{3.4}$$

This estimation is based on bandwidth h_1 and k_1 degree of polynomials. The bias of this estimator at x_0 is

$$B(x_0) = E(\mathbf{e_1}^T (X_{x_0}^{1T} K_x^1 X_{x_0}^1)^{-1} X_{x_0}^{1T} K_x^1 y) - g(x_0) \tag{3.5}$$

$$= \mathbf{e_1}^T (X_x^{1T} K_x^1 X_x^1)^{-1} X_x^{1T} K_x^1 g(x) - g(x_0) \tag{3.6}$$

Since

$$E(\mathbf{y}) = \begin{bmatrix} g(x_1) \\ \dots \\ g(x_n) \end{bmatrix} = \mathbf{g}(\mathbf{x}) \tag{3.7}$$

This study proposes to employ a less biased estimator $\widehat{g_2}(x)$, either with a smaller bandwidth or a higher degree or both to estimate the bias. The less biased estimator is denoted as

$$\widehat{g_2}(x) = \mathbf{e_1}^T (X_x^{2T} K_x^2 X_x^2)^{-1} X_x^{2T} K_x^2 \mathbf{y}$$

$$= \mathbf{e_1}^T C_x^2 \mathbf{y}, \tag{3.8}$$

and the corresponding bandwidth and degree of polynomial are h_2 and k_2. Then an estimator of bias $B(x)$ at x_0 is

$$\widehat{B}(x_0) = \mathbf{e_1}^T (X_{x_0}^{1T} K_x^1 X_{x_0}^1)^{-1} X_{x_0}^{1T} K_x^1 \widehat{g_2}(x) - \widehat{g_2}(x_0) \tag{3.9}$$

$$= \mathbf{e_1}^T C_{x_0}^1 \widehat{g_2}(x) - \widehat{g_2}(x_0) \tag{3.10}$$

To make the expressions concise, we use the notation

$$C_{x_0,1}^1 = e_1^T C_{x_0}^1$$

and

$$C_{x_0,1}^2 = \mathbf{e_1}^T C_{x_0}^2,$$

then

$$\widehat{g_1}(x_0) = C_{x_0,1}^1 \mathbf{y}, \tag{3.13}$$

$$\widehat{g_2}(x_0) = C_{x_0,1}^2 \mathbf{y}, \tag{3.14}$$

$$\widehat{\mathbf{g_2}}(\mathbf{x}) = \begin{bmatrix} \widehat{g_2}(x_1) \\ \dots \\ \widehat{g_2}(x_n) \end{bmatrix} = \begin{bmatrix} C_{x_1,1}^2 \\ \dots \\ C_{x_n,1}^2 \end{bmatrix} \mathbf{y} = C^2 \mathbf{y}. \tag{3.15}$$

So with all of the notation defined above, the estimator of bias in $\widehat{g_1}(x_0)$ is

$$\widehat{B}(x_0) = C^1_{x_0,1} C^2 \mathbf{y} - C^2_{x_0,1} \mathbf{y} \tag{3.16}$$

$$= (C^1_{x_0,1} C^2 - C^2_{x_0,1}) \mathbf{y}. \tag{3.17}$$

We assume that $\text{Var}(\mathbf{y}) = \sigma^2$. Thus,

$$\text{Var}(\widehat{B}(x_0)) = \sigma^2 (C^1_{x_0,1} C^2 - C^2_{x_0,1})(C^1_{x_0,1} C^2 - C^2_{x_0,1})^T. \tag{3.18}$$

With the variance of the bias estimator, this study is able to construct a pointwise confidence interval based on the Central Limit Theorem. With a large sample, the approximate α-level confidence interval would be

$$\left[\widehat{B}(x_0) - z_{\alpha/2} * \sqrt{\widehat{\text{Var}}(\widehat{B}(x_0))}, \widehat{B}(x_0) + z_{\alpha/2} * \sqrt{\widehat{\text{Var}}(\widehat{B}(x_0))} \right] \tag{3.19}$$

If the variance of the bias is written in vector form, it would be

$$\text{Var}(\widehat{B}(x)) = \sigma^2 (C^1 C^2 - C^2)(C^1 C^2 - C^2)^T. \tag{3.20}$$

In order to evaluate the confidence interval and the variance of the bias, and estimate of σ^2 is needed. A number of candidates are available in the literature, but a simple, approximately unbiased estimator, which seems to work well for moderate to large samples is one based on perturbing the (ordered) covariate observations slightly. Specifically, we set

$$x^*_{2j-1} = x^*_{2j} = (x_{2j} + x_{2j-1})/2$$

for $j = 1, 2, \ldots, n/2$, ignoring x_n if n is odd. The perturbed data set $\{x^*, y\}$ now has replicates

so that an estimate of the pure error sum of squares (SSPE) can be obtained.

$$\text{SSPE} = \frac{1}{2}\sum_{j=1}^{n/2}(y_{2j} - y_{2j-1})^2.$$

Dividing SSPE by $n/2$ gives an unbiased estimate of σ^2 when the amount of perturbation required is 0. Otherwise, the estimator is biased, though not substantially, in practice.

In the next section, we will showcase the assessment plot along with the product of the variance of the bias estimator. The study of Fan and Gijbels (1995a) inferred a method for choosing the order of the polynomials using the asymptotic characteristics of the bias and variance of the local polynomials. That study used a fixed bandwidth. On the other hand, the study of Fan and Gijbels (1995b) suggested a relatively refined data-driven bandwidth selection procedure, based on approximation of bias and variance of estimator. This procedure is capable of choosing constant bandwidth or variable bandwidth in a pilot estimation procedure. The approximation of bias for pth order of local polynomial fit used $(p + a)$th order of local polynomial fit which used in the pilot estimation. Here a is a positive integer. Fan and Gijbels (1995b) used $a = 2$ in their presentation to demonstrate this "two stage" method outperforms the usual "one stage" method. In our current study, we demonstrate the assessment tool using a fixed bandwidth and make comparisons under various scenarios.

3.3 Simulation Study

In this section, the finite-sample behavior for the proposed local polynomial bias estimator with pointwise confidence interval for the estimate of bias is investigated. This study used

$$h = O(n^{-\frac{1}{2p+3}})$$

to determine the appropriate bandwidth for the pilot estimator mentioned in section 3.2. This is to minimize asymptotic MSE.

The simulation study design is summarized in Table 3.1. Target functions from He and Huang (2009) are employed since they represent different noise to signal ratio magnitude. Noise to signal is defined as $\sigma/(\text{Var}(g(x))+\sigma^2)^{1/2}$. The higher noise to ratio statistic, the harder the estimation will be in terms of choosing the optimal degree and bandwith.

Example	$g(x)$	σ	Noise/Signal
1	$x + 2e^{(-16x^2)}$	0.4	1/3
2	$sin(2x) + 2e^{(-16x^2)}$	0.3	1/3
3	$0.3e^{(-4(x+1)^2)} + 0.7e^{(-16*(x-1)^2)}$	0.1	1/2

Table 3.1: Target functions, σ and noise to signal ratios considered in the simulation study.

The simulation study follows these steps:

1. Randomly generate x and y according to target functions on the predefined support $[-2, 2]$.

2. Fit LPR to the simulated ordered pairs. Generate the confidence interval for the potential bias.

3. Calculate the real bias at each grid point.

4. The bias confidence interval is also generated according to the methodology described in section 3.2.

5. Check whether the bias confidence interval contains the true bias at each x points.

6. Repeat steps 1 to 4 for a large number of times. This study used 1000 times. Obtain the true confidence interval coverage, by computing the proportion of confidential intervals which contains the the true bias out of the simulation number. The true confidence level should be close to the nominal coverage when the method works well.

There are several factors affecting our simulation behavior, including: bandwidth h_1, degree of polynomials k_1, kernel for $\widehat{g_1}(x)$, and correspondence for $\widehat{g_2}(x)$, nominal coverage level and sample size. This study focuses on nominal coverage level, bandwidth for $\widehat{g_2}(x)$ and sample size as the setup as shown in Table 3.2.

An example of the Bias assessment plot is shown in Fig 3.1.

According to the eight scenarios in Table 3.2, simulation results for the first target function are shown in Fig 3.2 to Fig 3.9, for the second target function, results are in Fig 3.10 to Fig 3.17 and for the second target function, results are in Fig 3.18 to Fig 3.25. In particular, we ploted the pointwise coverages for each of the values of x in the domain of interest. We observed that almost all of these simulation worked well, in that the actual coverage randomly fluctuates about the nominal coverage, though in a few cases, it might be somewhat conservative in that actual coverage slightly exceeds the nominal level.

Scenario	Nominal Level	h_2	Sample Size	k_2
s1	0.95	h	100	$k+1$
s2	0.95	h	200	$k+1$
s3	0.95	$h/2$	100	k
s4	0.95	$h/2$	200	k
s5	0.99	h	100	$k+1$
s6	0.99	h	200	$k+1$
s7	0.99	$h/2$	100	k
s8	0.99	$h/2$	200	k

Table 3.2: Simulation design table containing nominal confidence level, relative bandwidth, the number of simulated observations per sample and relative degree. Each row represents a scenario.

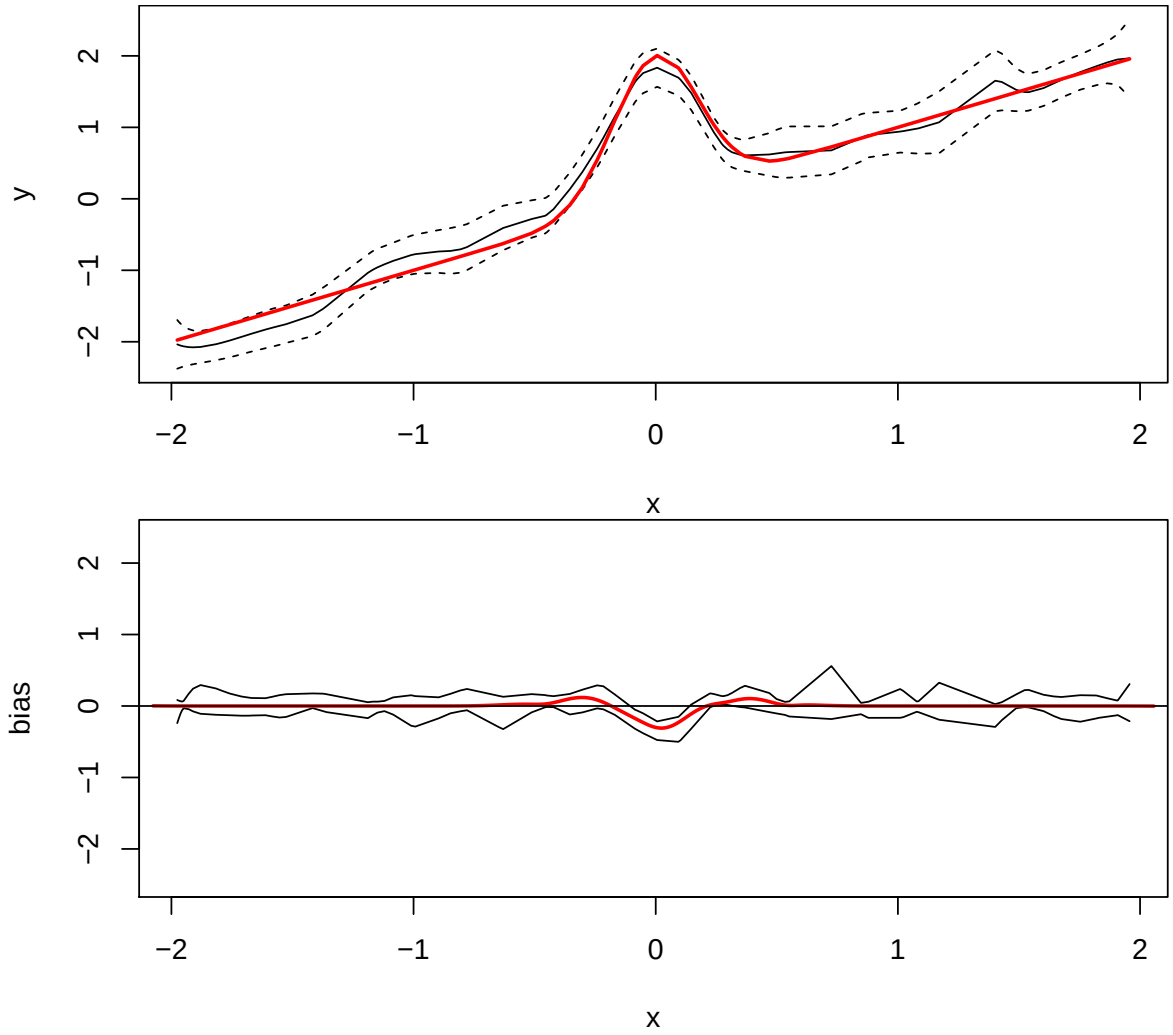

Figure 3.1: Pointwise Bias assessment tool for Example 1 according to the setup in Scenario 1. Top Panel: Target function (red curve), Local Polynomial regression (solid black) and variance (dashed black). Bottom Panel: True bias (red curve), confidence interval for bias in one simulation (black curve)

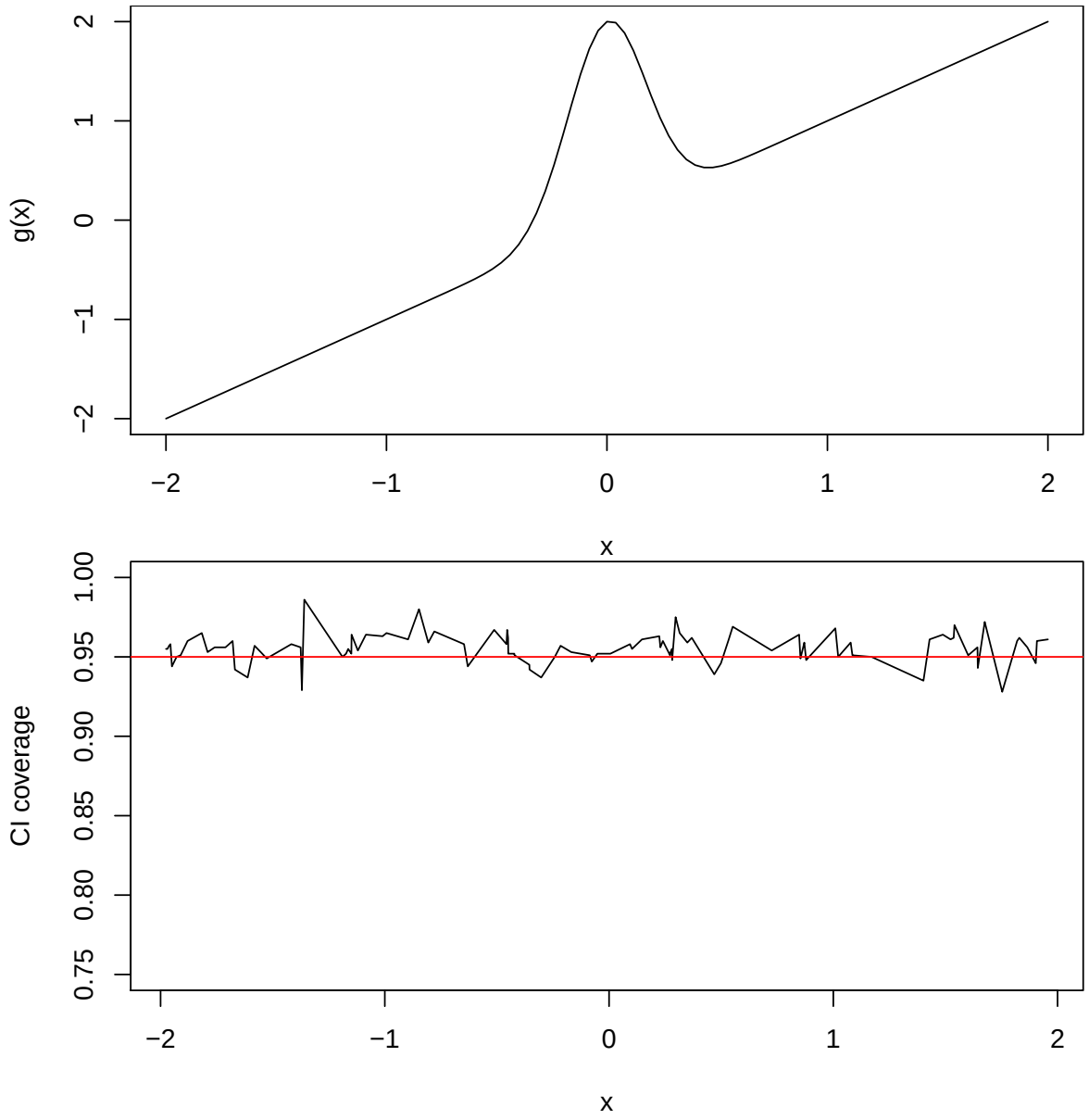

Figure 3.2: Pointwise Bias assessment tool performance for local linear for Example 1 according to the setup in Scenario 1. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

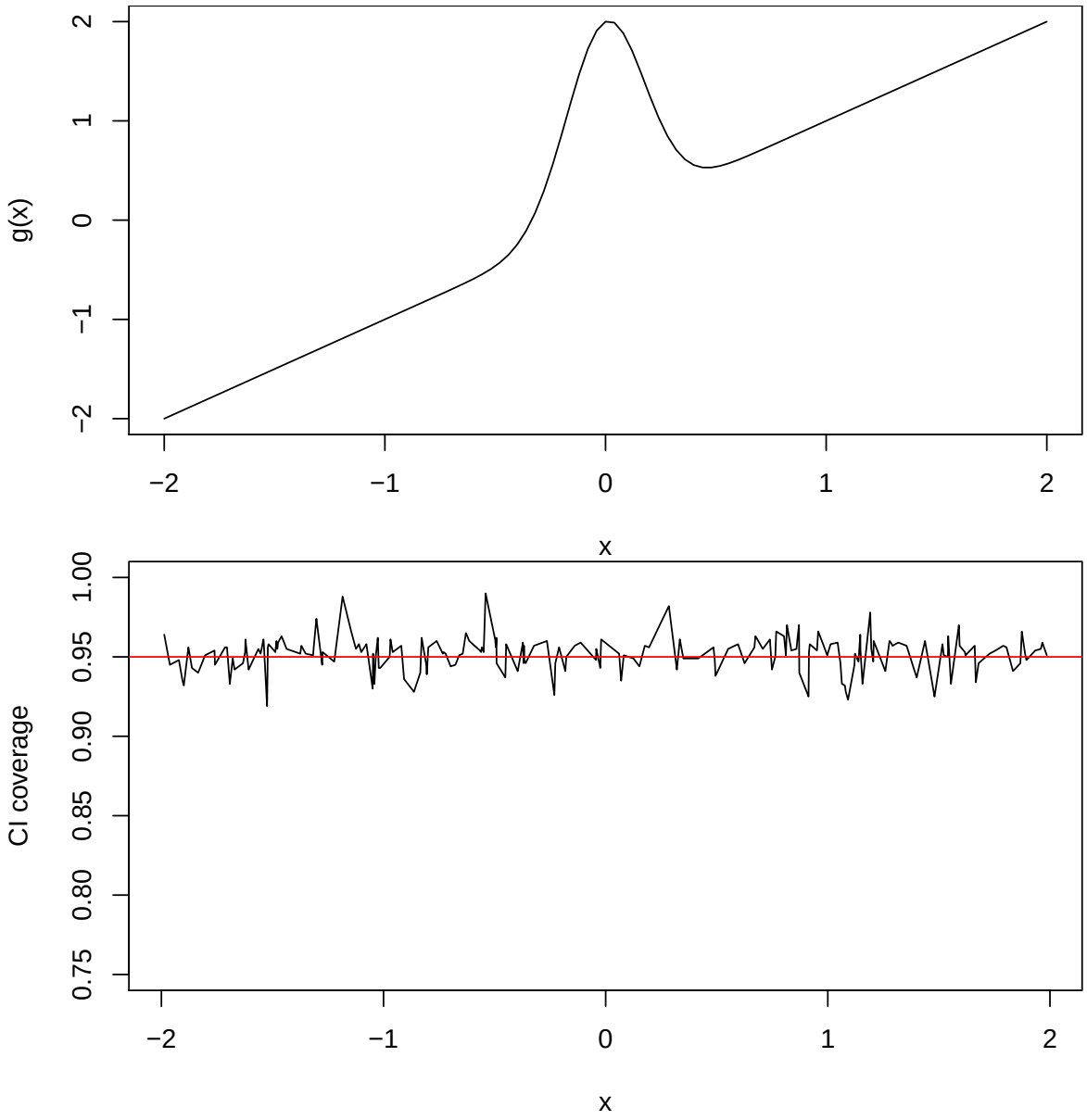

Figure 3.3: Pointwise Bias assessment tool performance for local linear for Example 1 according to the setup in Scenario 2. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

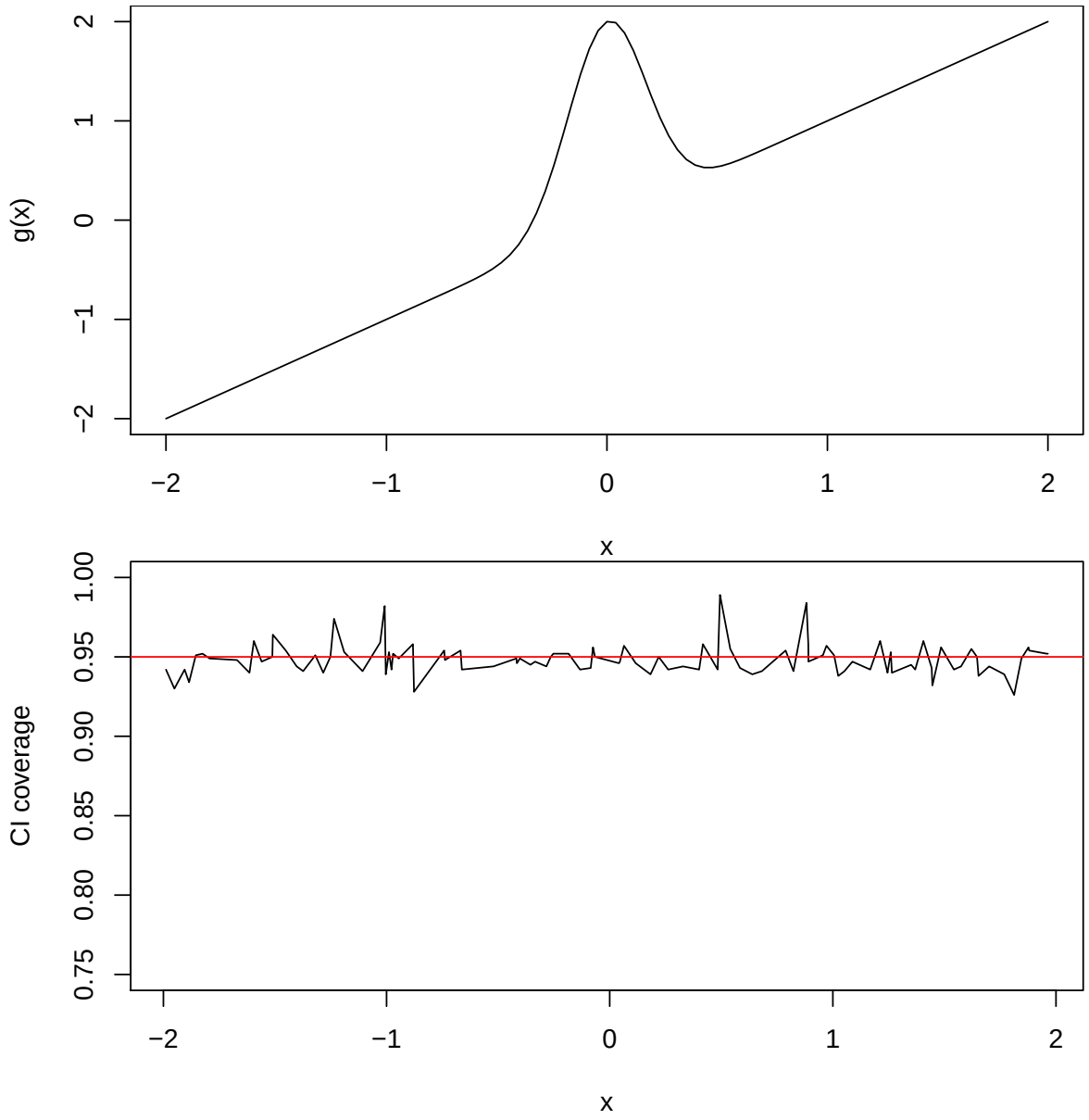

Figure 3.4: Pointwise Bias assessment tool performance for local linear for Example 1 according to the setup in Scenario 3. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

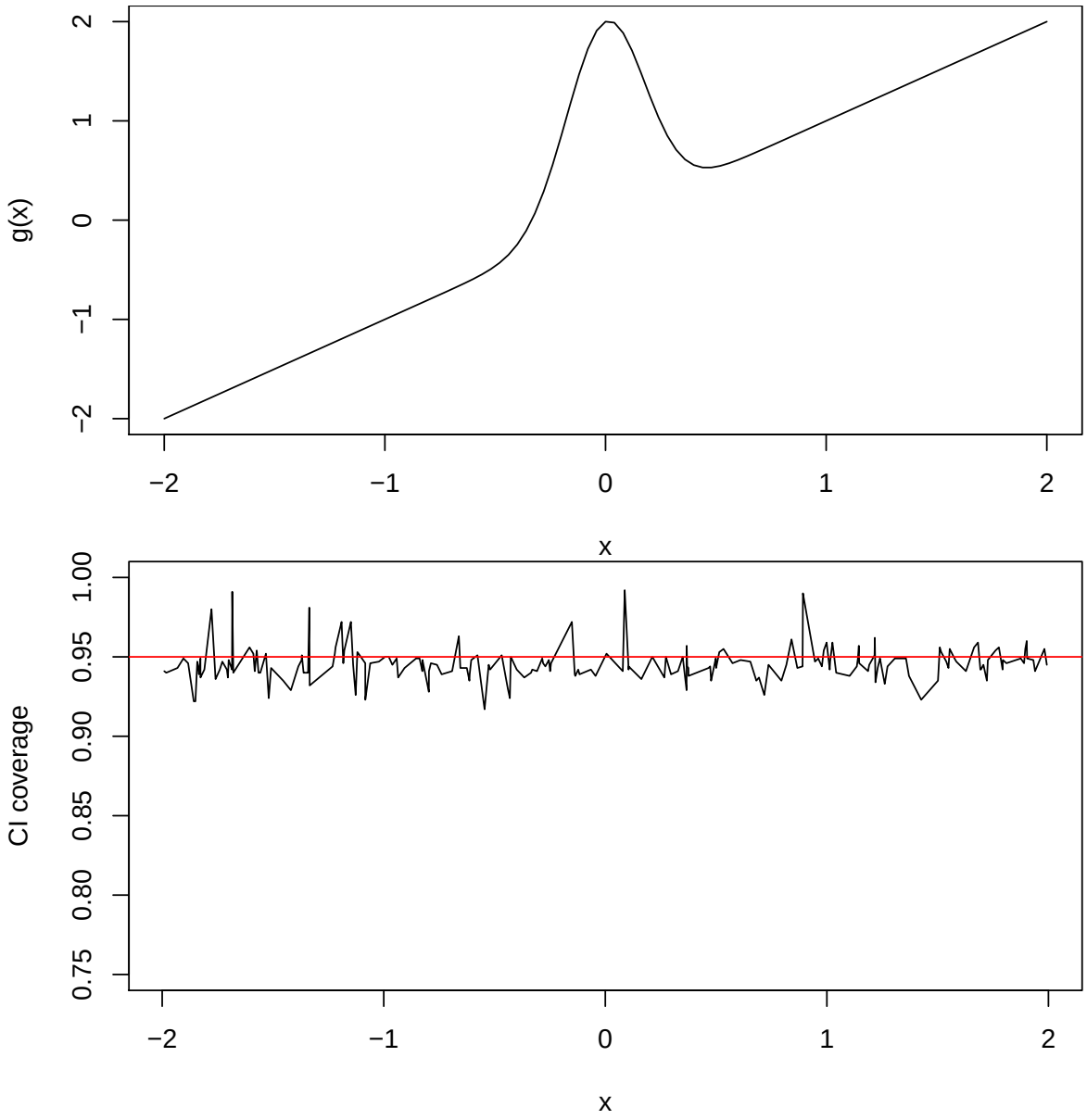

Figure 3.5: Pointwise Bias assessment tool performance for local linear for Example 1 according to the setup in Scenario 4. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

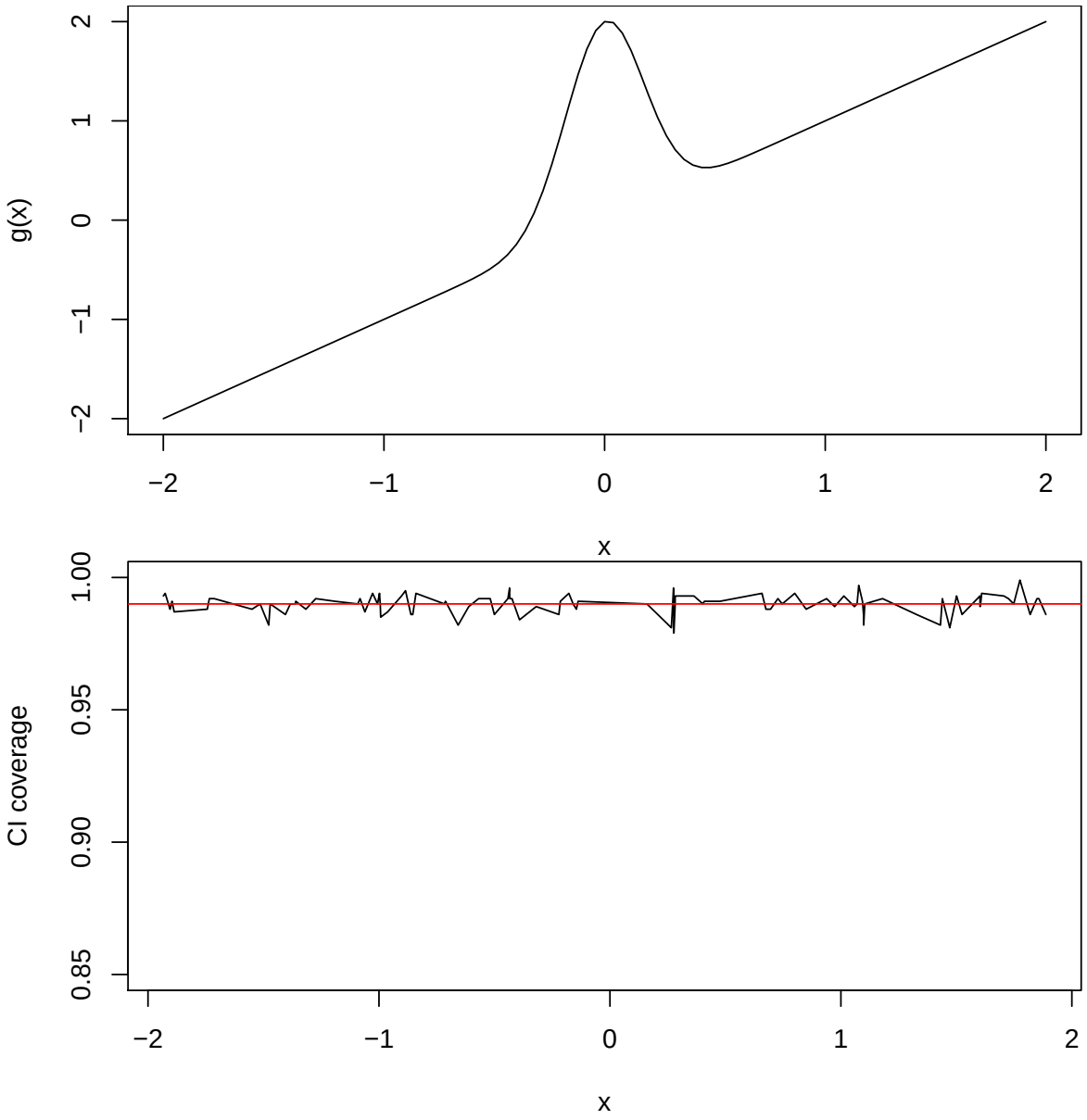

Figure 3.6: Pointwise Bias assessment tool performance for local linear for Example 1 according to the setup in Scenario 5. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

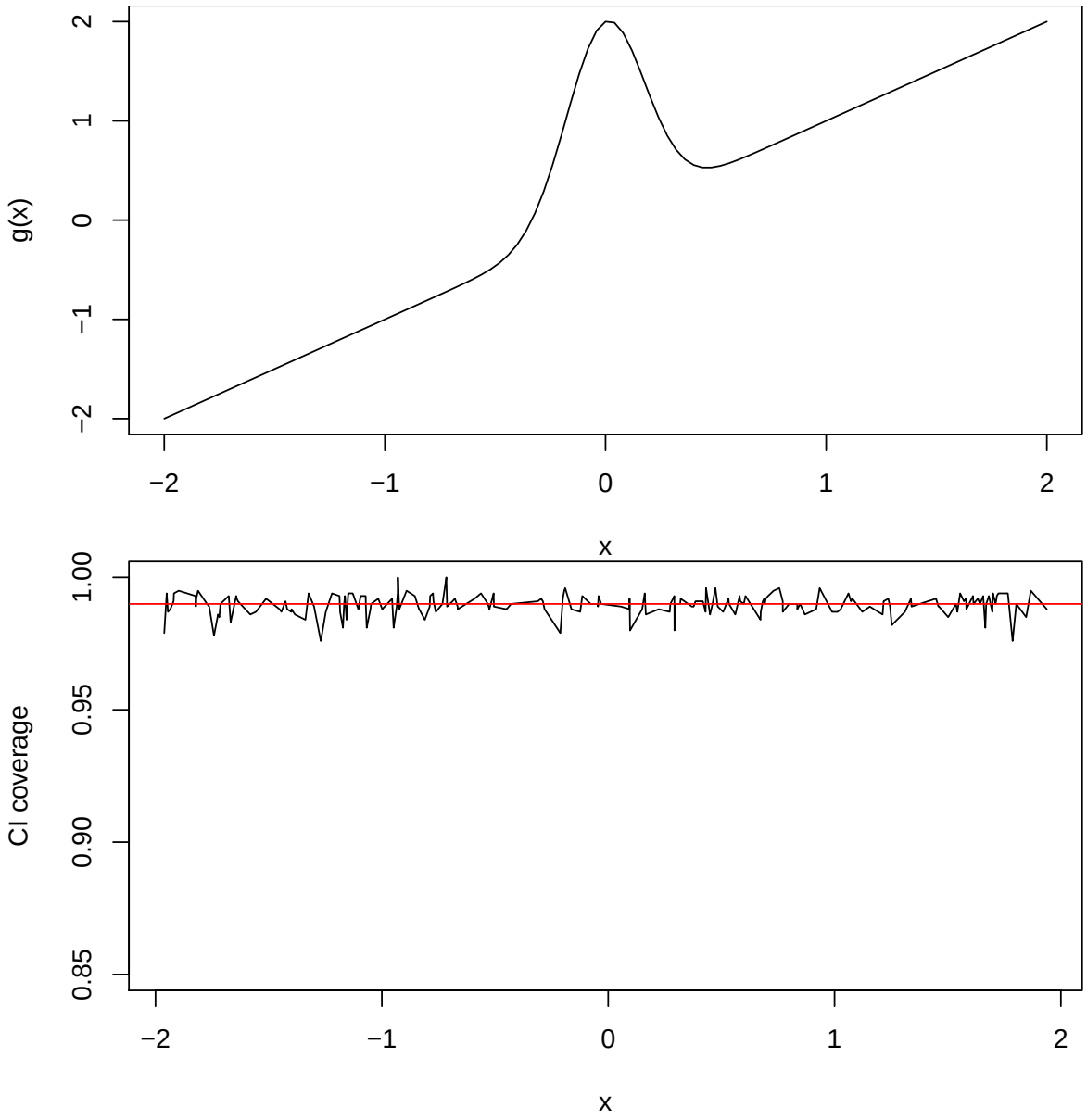

Figure 3.7: Pointwise Bias assessment tool performance for local linear for Example 1 according to the setup in Scenario 6. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

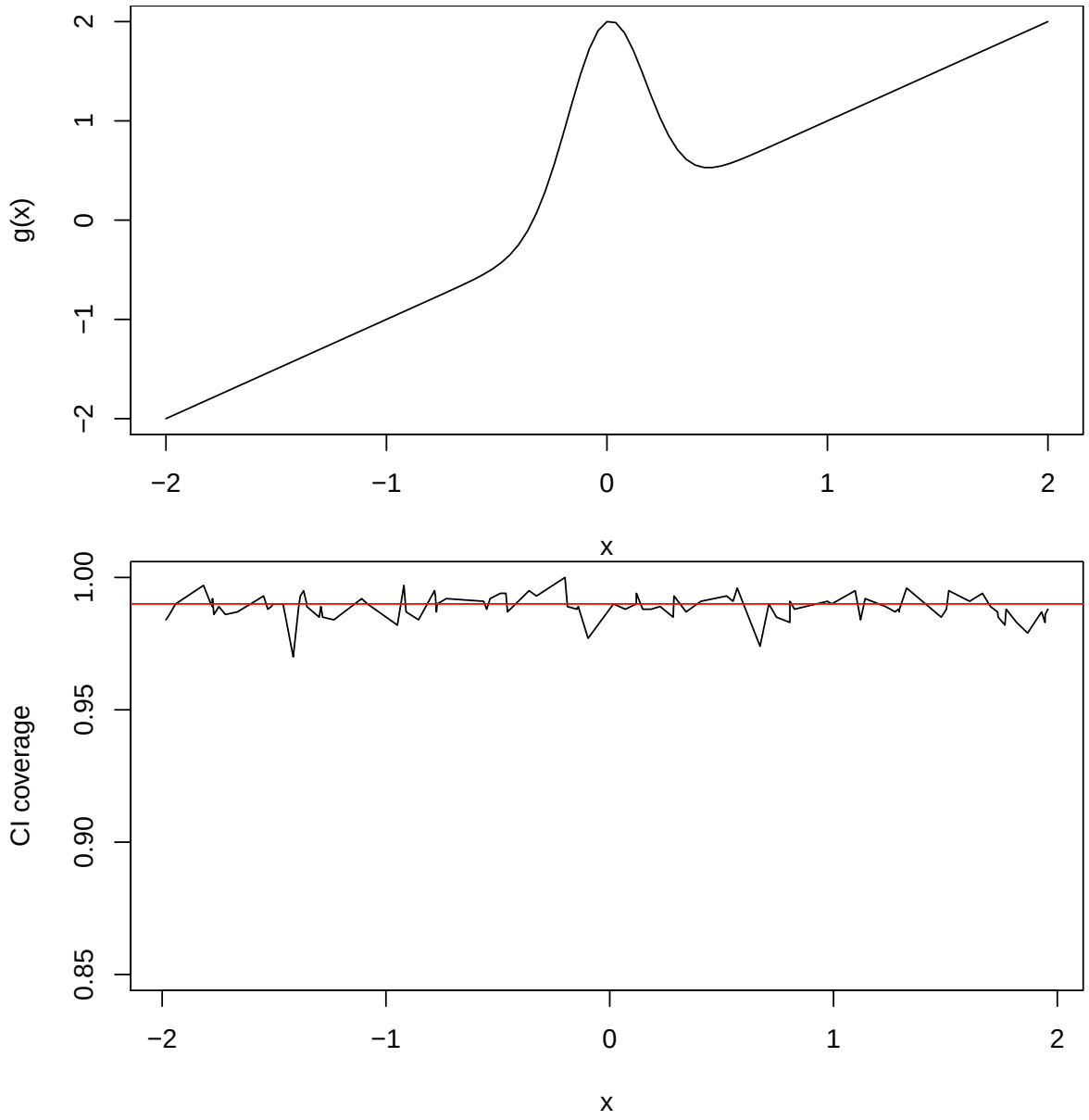

Figure 3.8: Pointwise Bias assessment tool performance for local linear for Example 1 according to the setup in Scenario 7. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

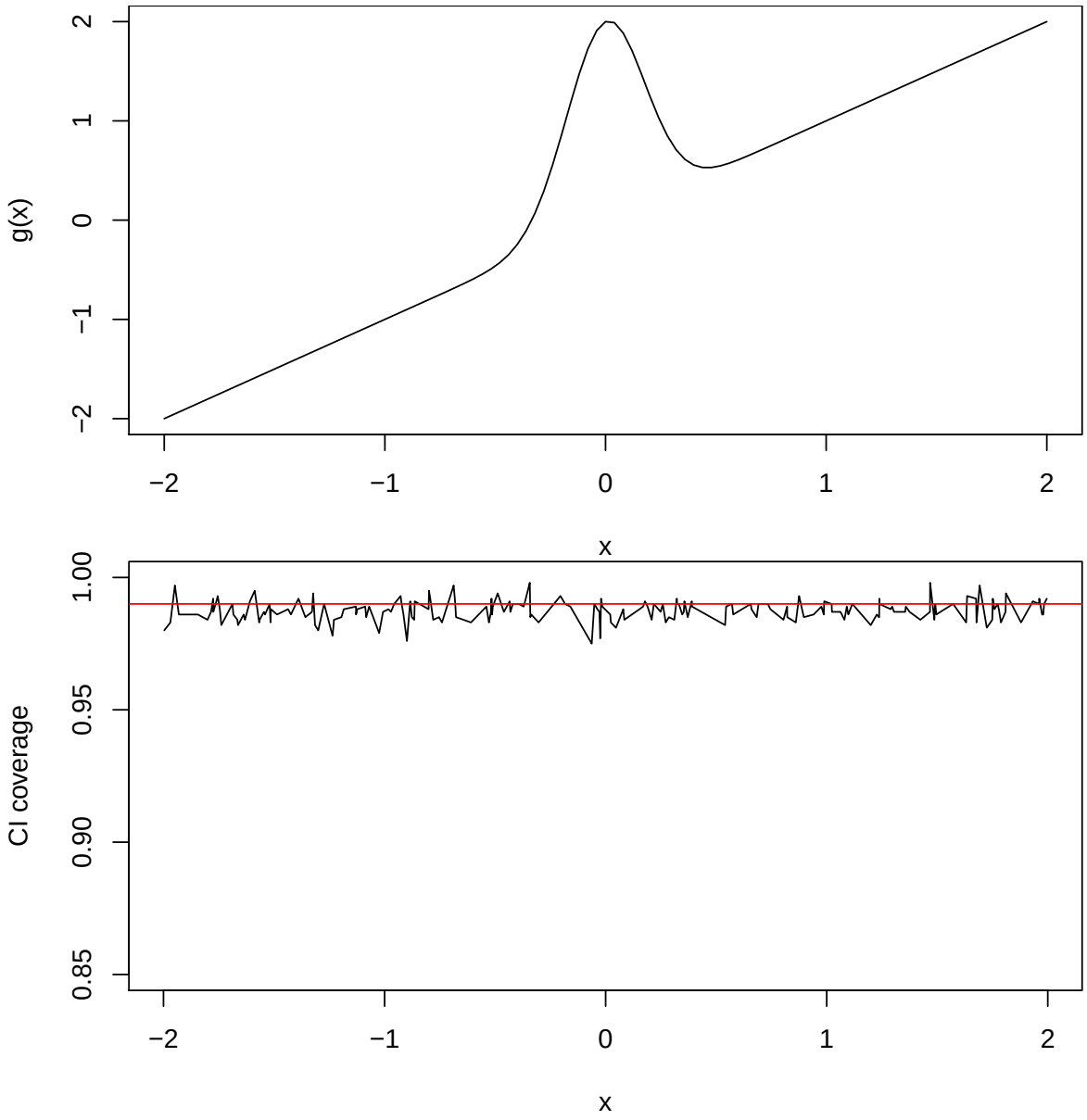

Figure 3.9: Pointwise Bias assessment tool performance for local linear for Example 1 according to the setup in Scenario 8. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

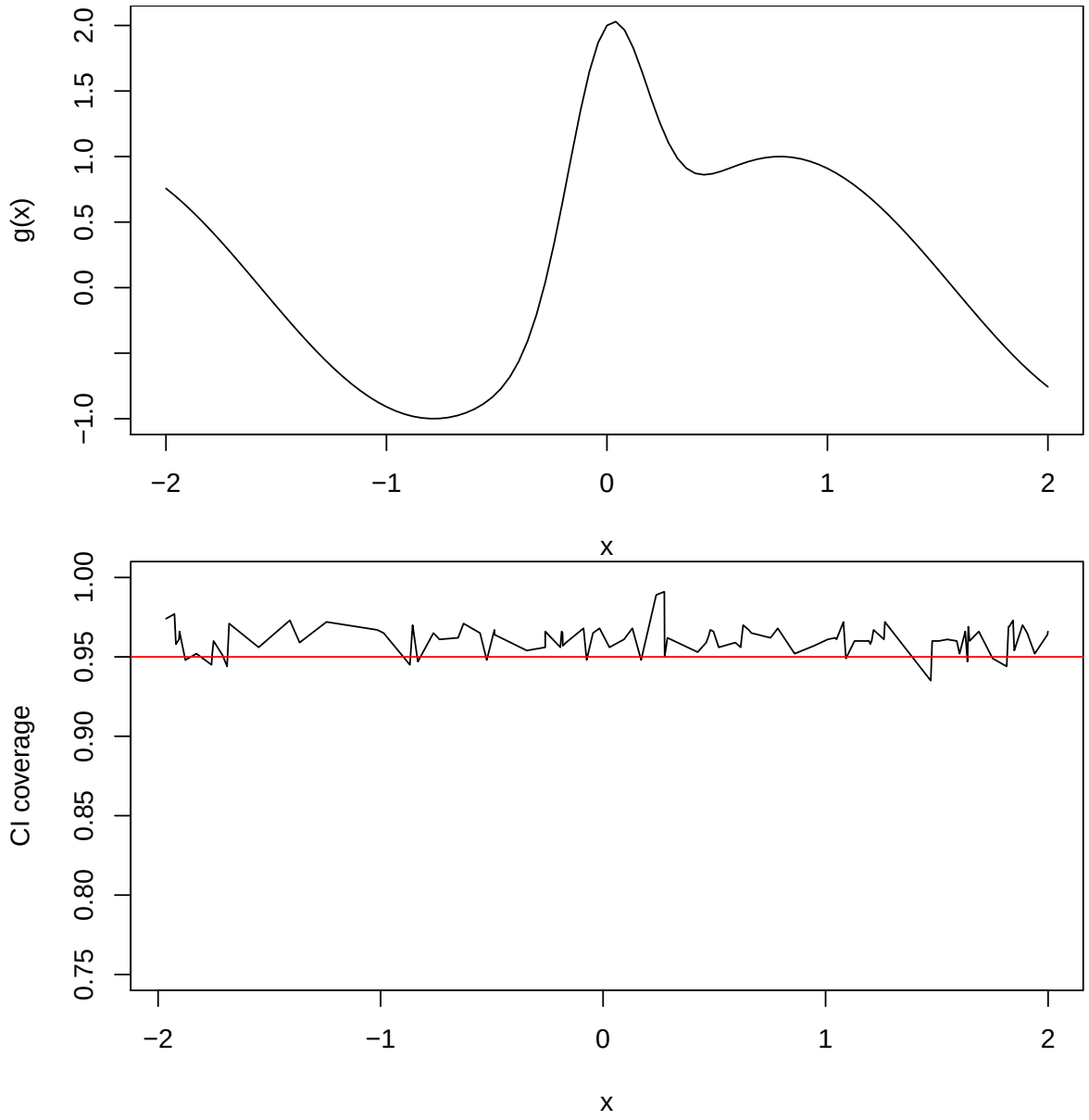

Figure 3.10: Pointwise Bias assessment tool performance for local linear for Example 2 according to the setup in Scenario 1. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

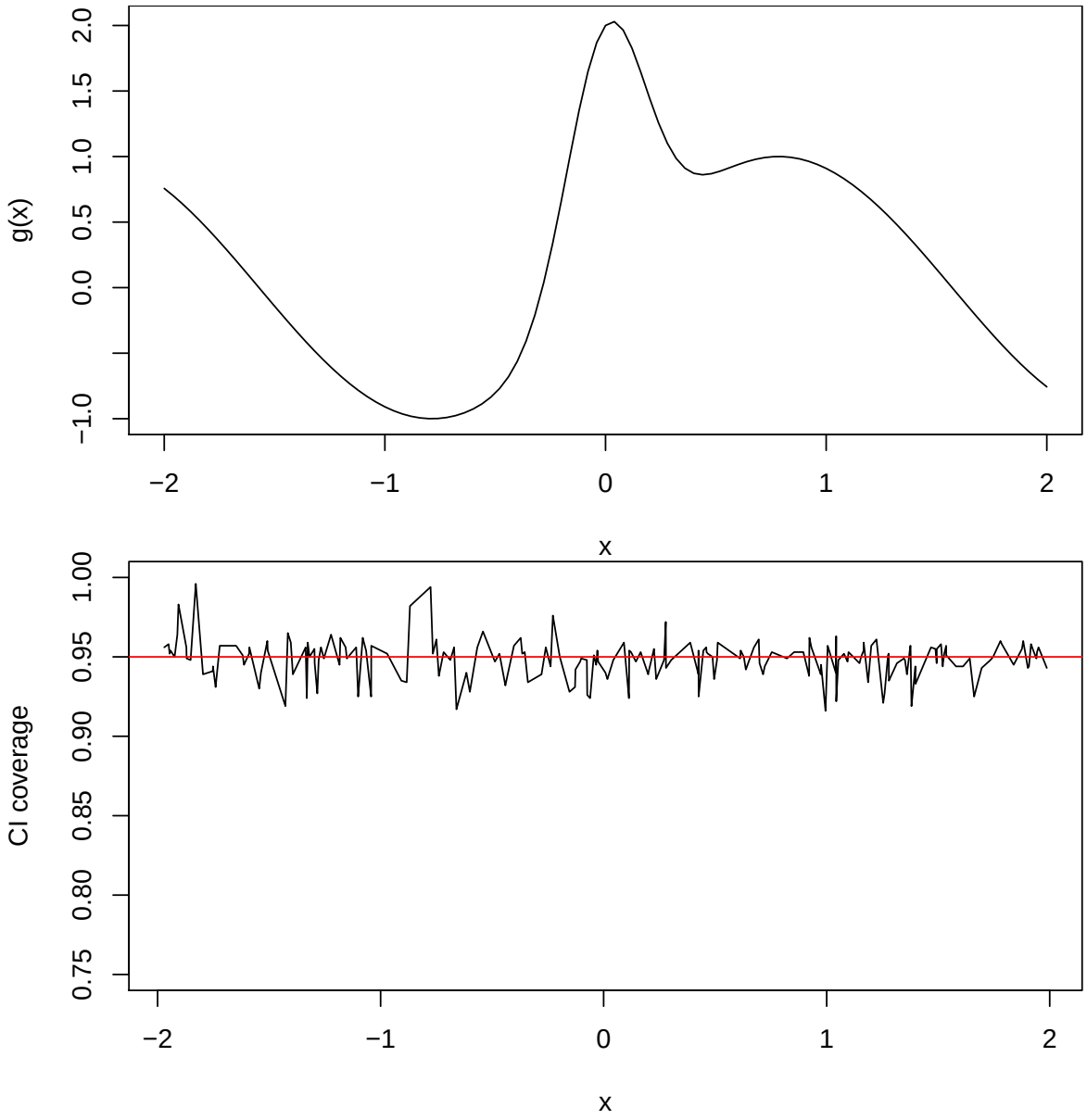

Figure 3.11: Pointwise Bias assessment tool performance for local linear for Example 2 according to the setup in Scenario 2. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

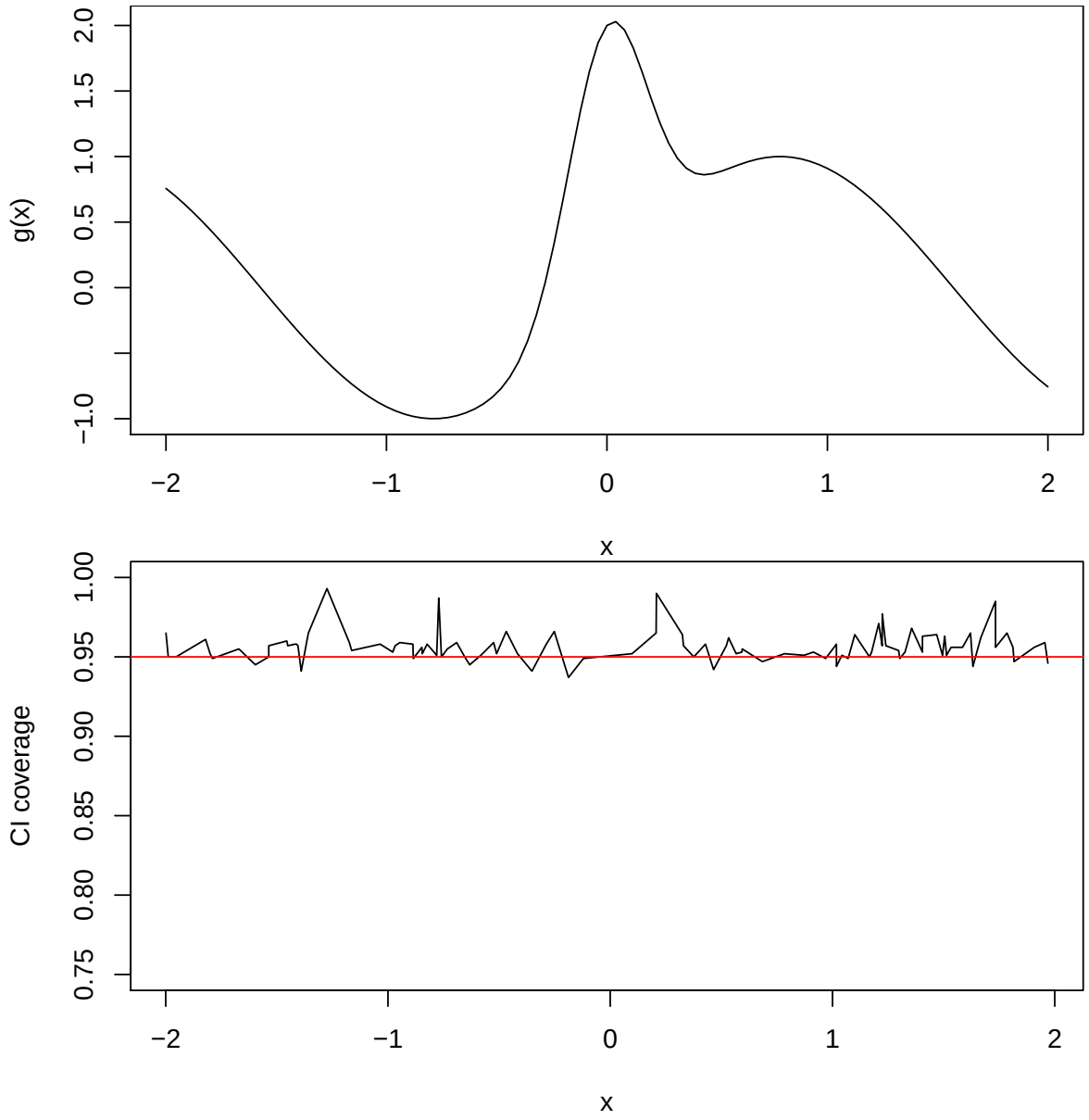

Figure 3.12: Pointwise Bias assessment tool performance for local linear for Example 2 according to the setup in Scenario 3. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

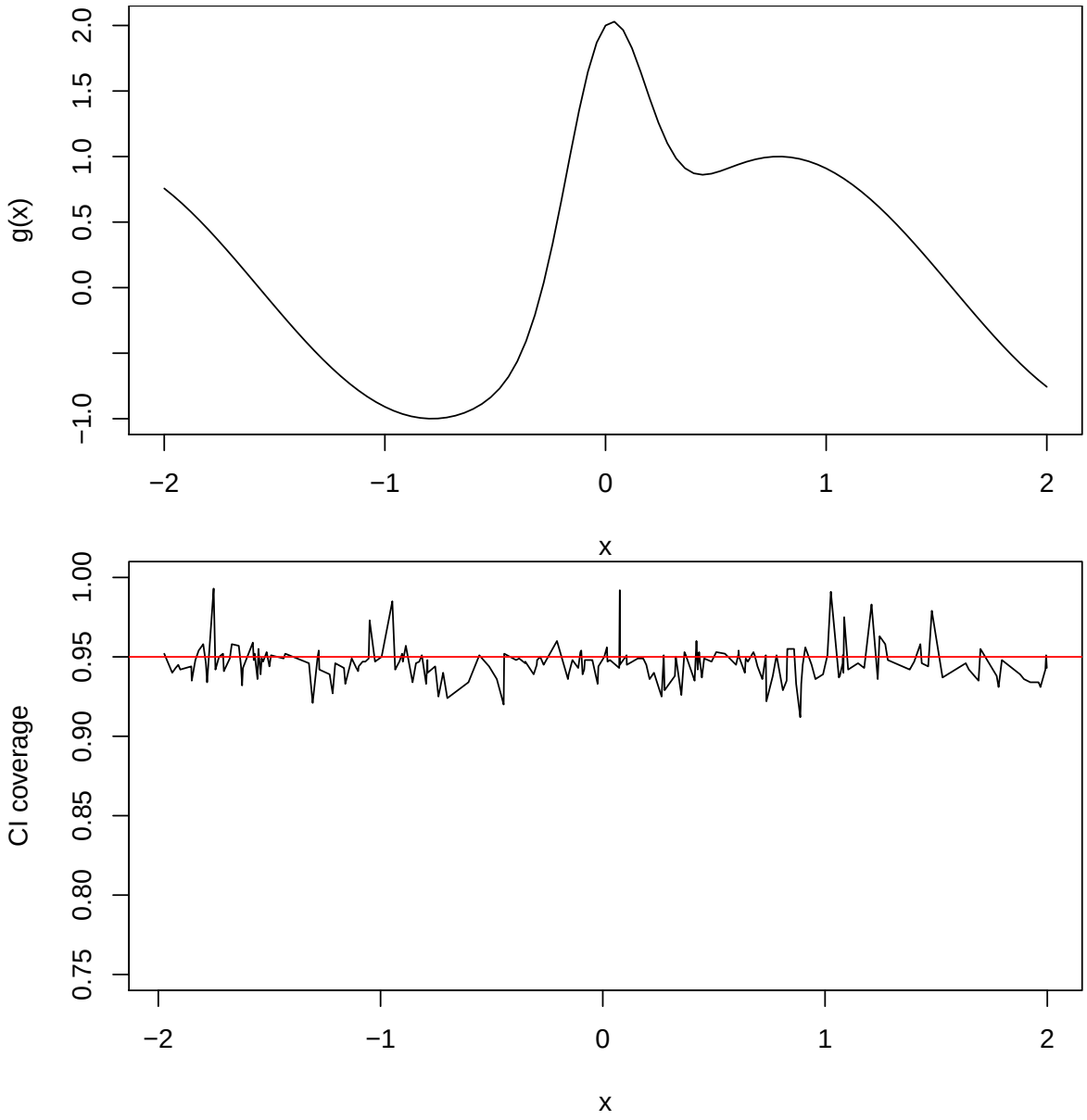

Figure 3.13: Pointwise Bias assessment tool performance for local linear for Example 2 according to the setup in Scenario 4. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

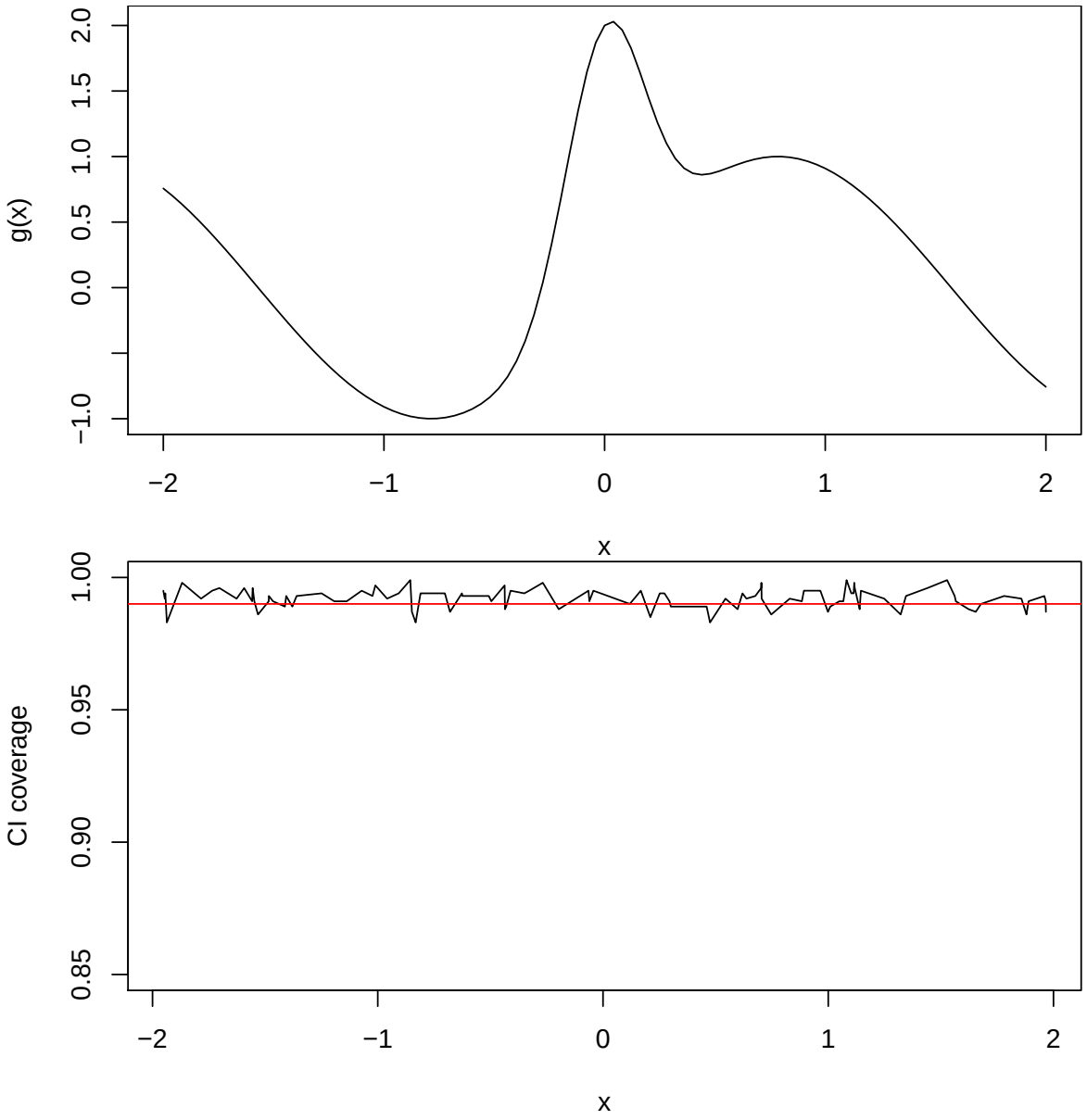

Figure 3.14: Pointwise Bias assessment tool performance for local linear for Example 2 according to the setup in Scenario 5. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

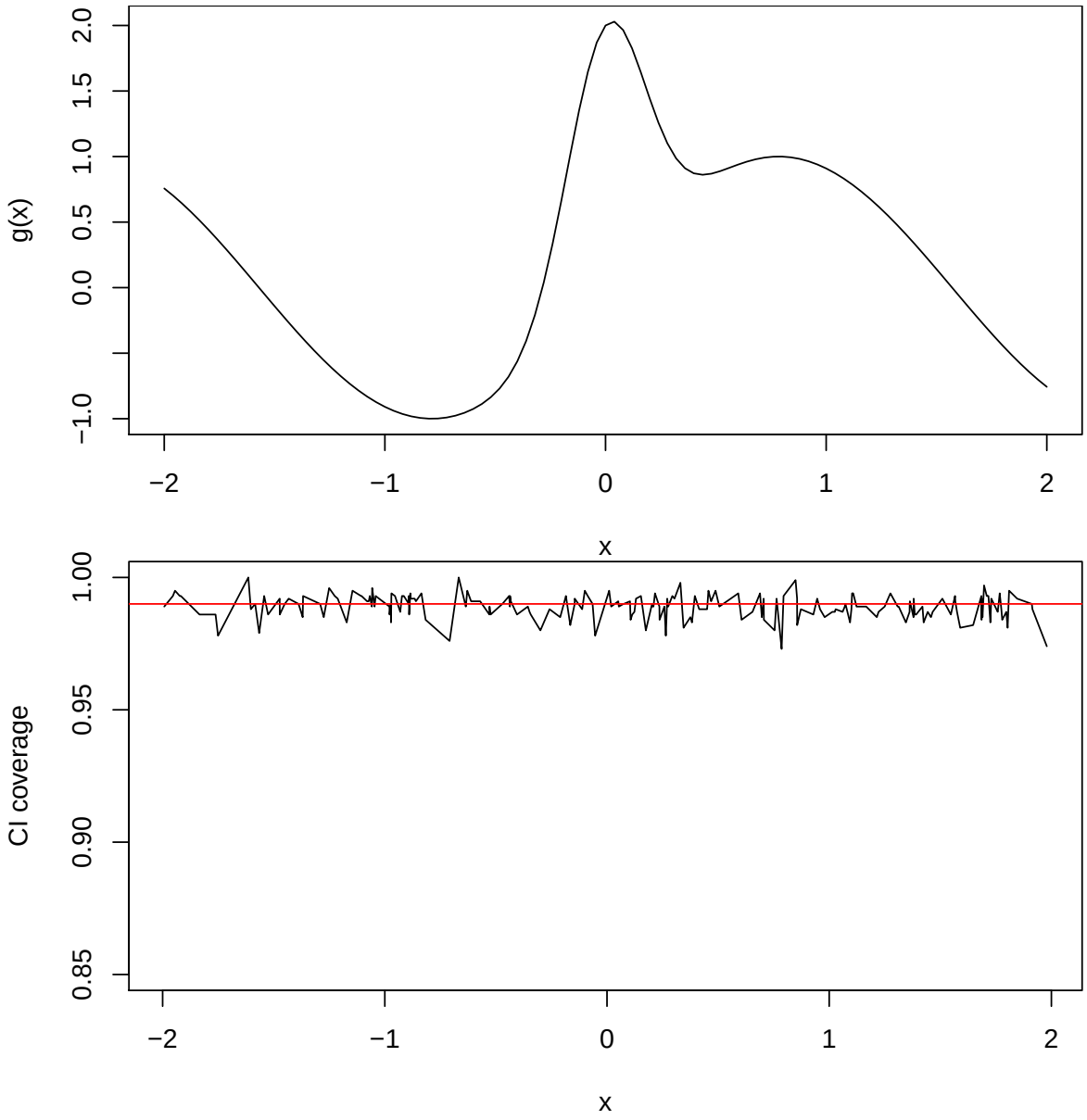

Figure 3.15: Pointwise Bias assessment tool performance for local linear for Example 2 according to the setup in Scenario 6. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

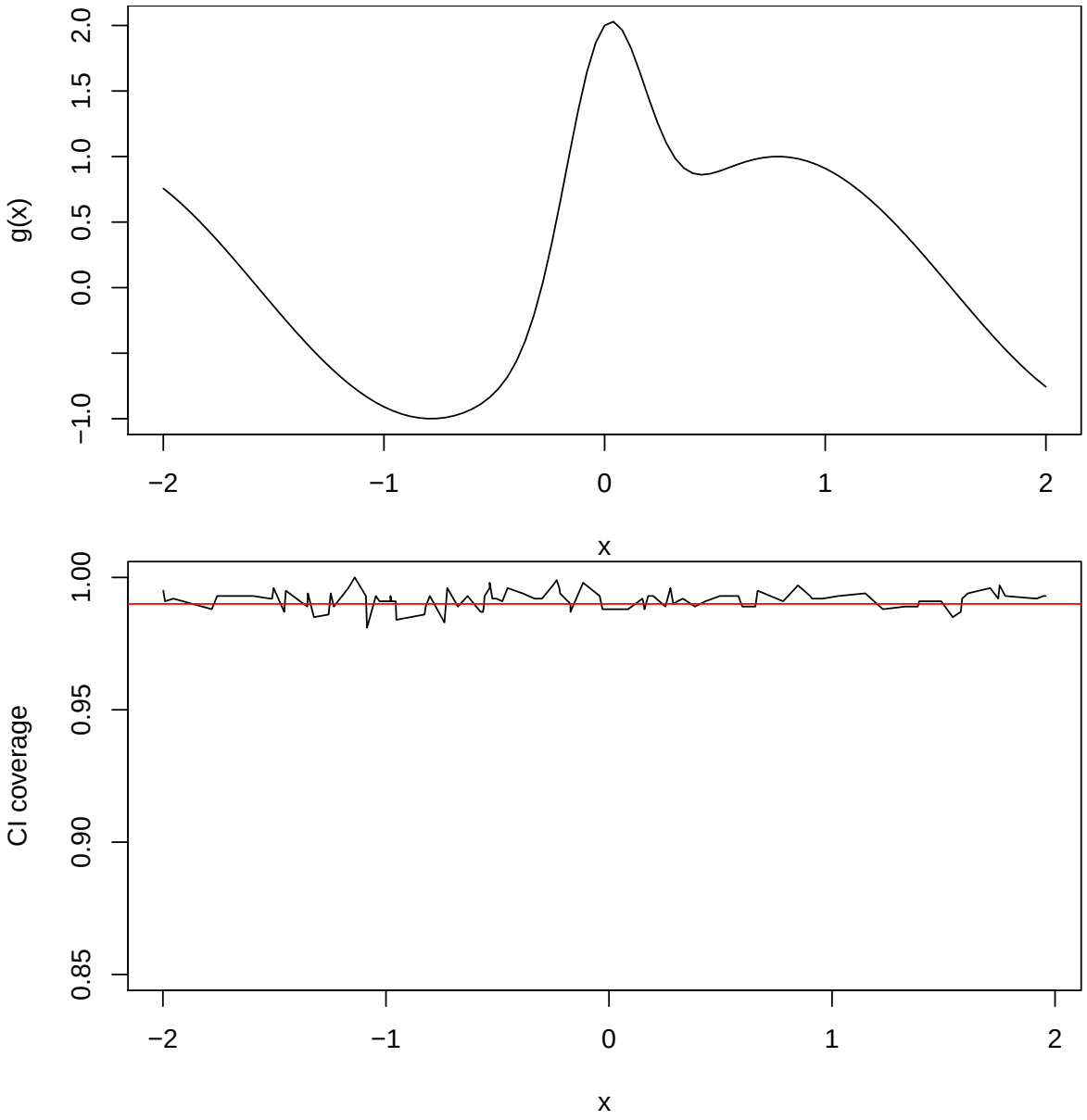

Figure 3.16: Pointwise Bias assessment tool performance for local linear for Example 2 according to the setup in Scenario 7. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

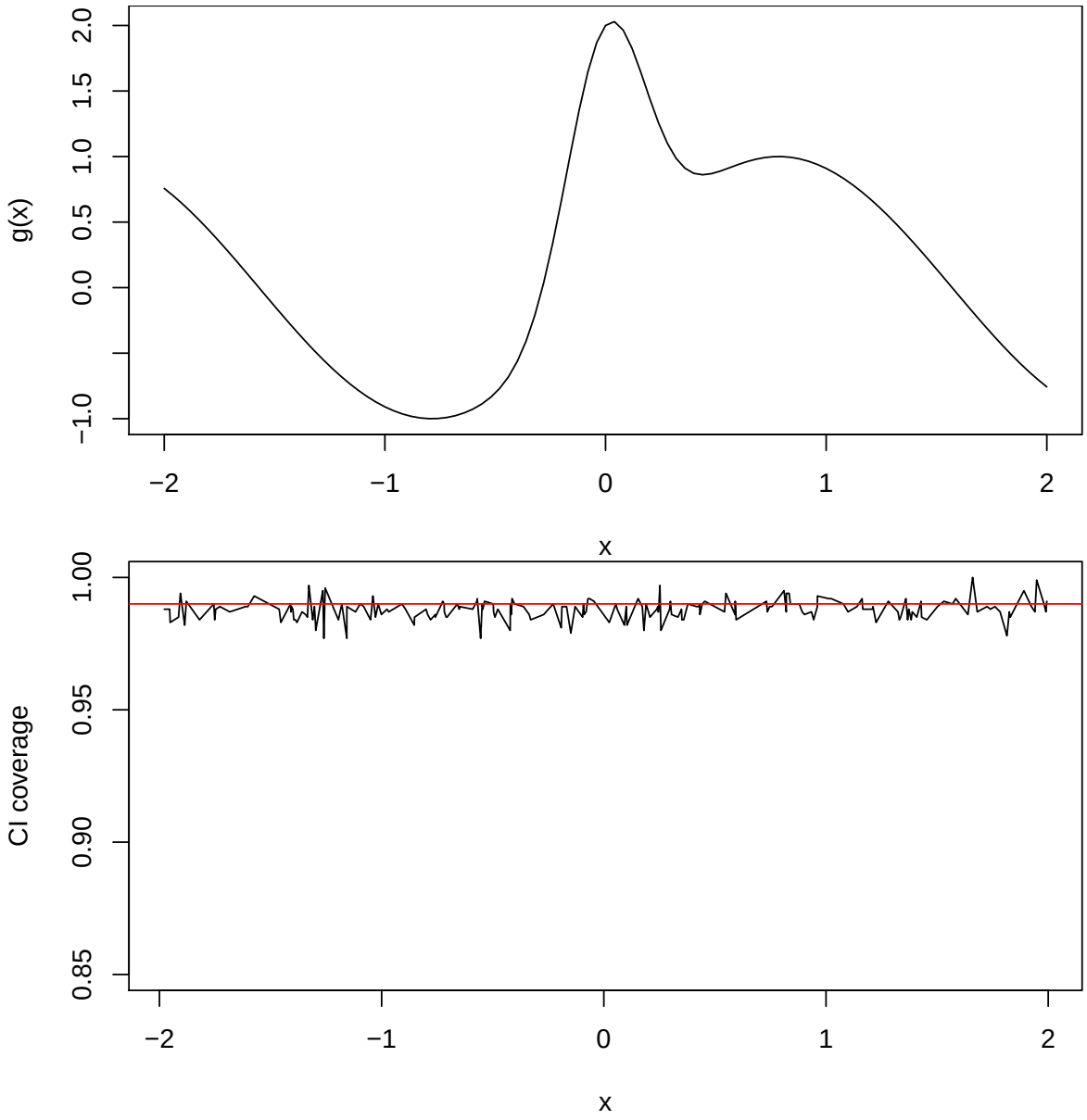

Figure 3.17: Pointwise Bias assessment tool performance for local linear for Example 2 according to the setup in Scenario 8. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

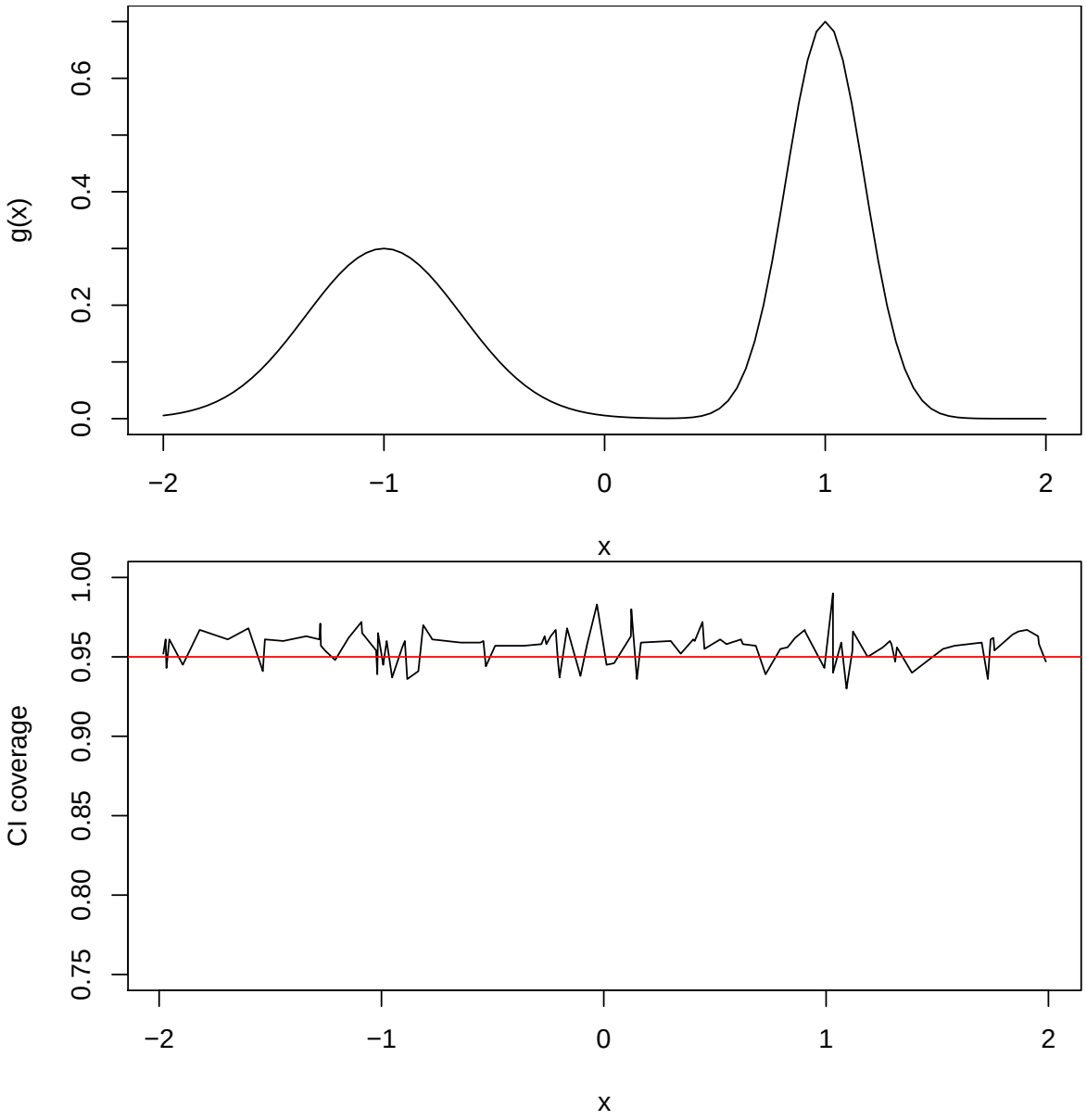

Figure 3.18: Pointwise Bias assessment tool performance for local linear for Example 3 according to the setup in Scenario 1. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

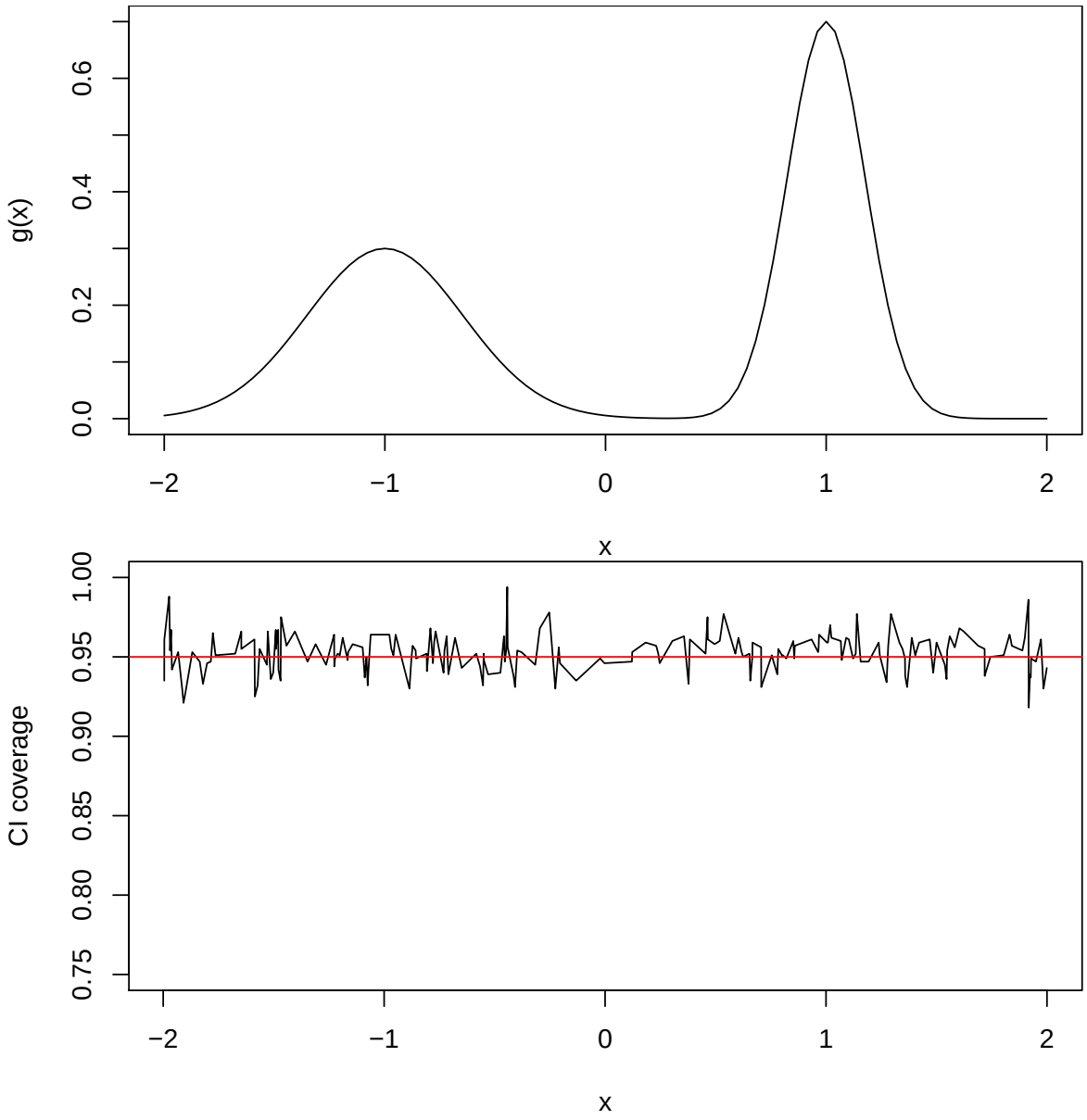

Figure 3.19: Pointwise Bias assessment tool performance for local linear for Example 3 according to the setup in Scenario 2. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

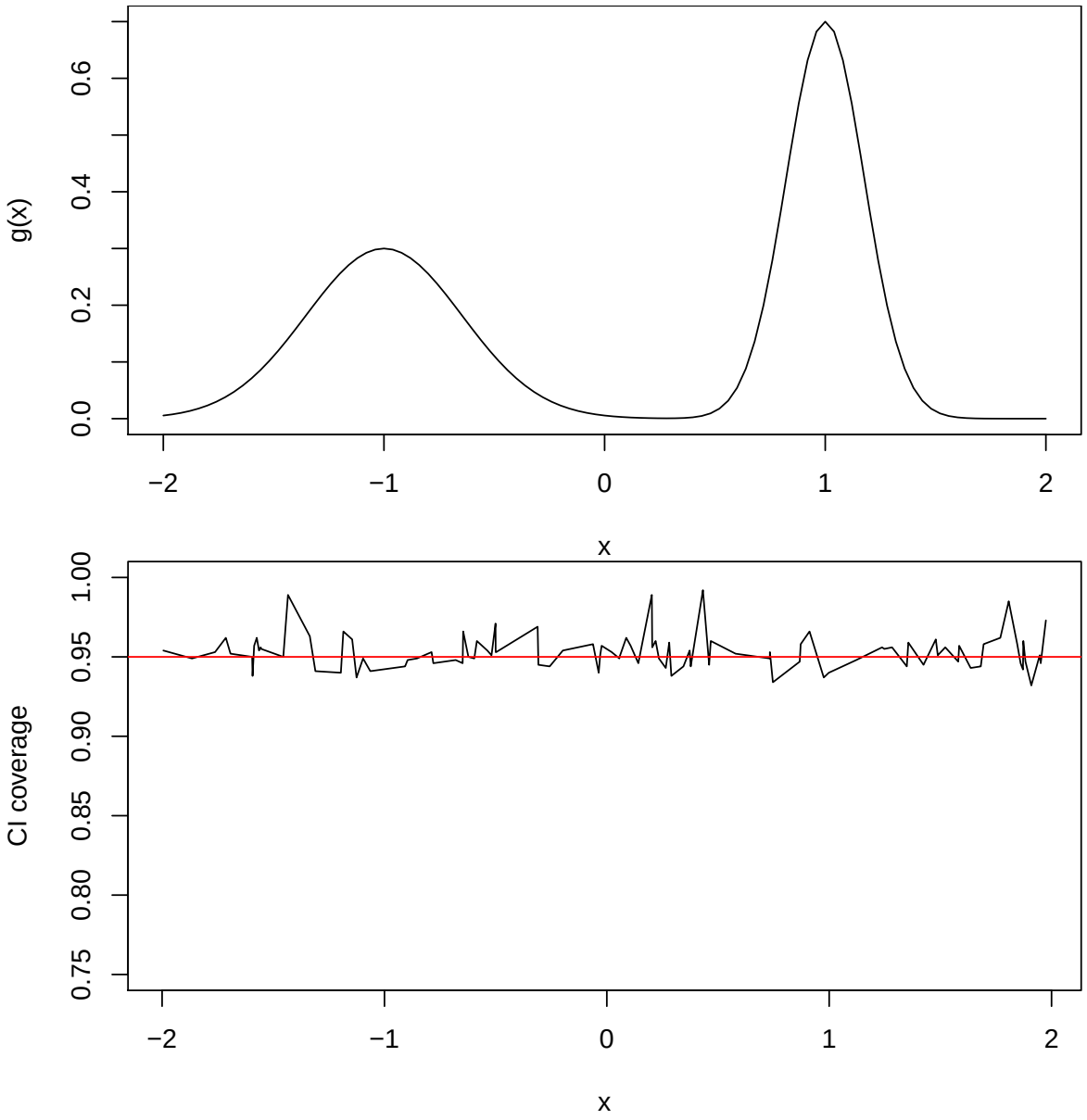

Figure 3.20: Pointwise Bias assessment tool performance for local linear for Example 3 according to the setup in Scenario 3. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

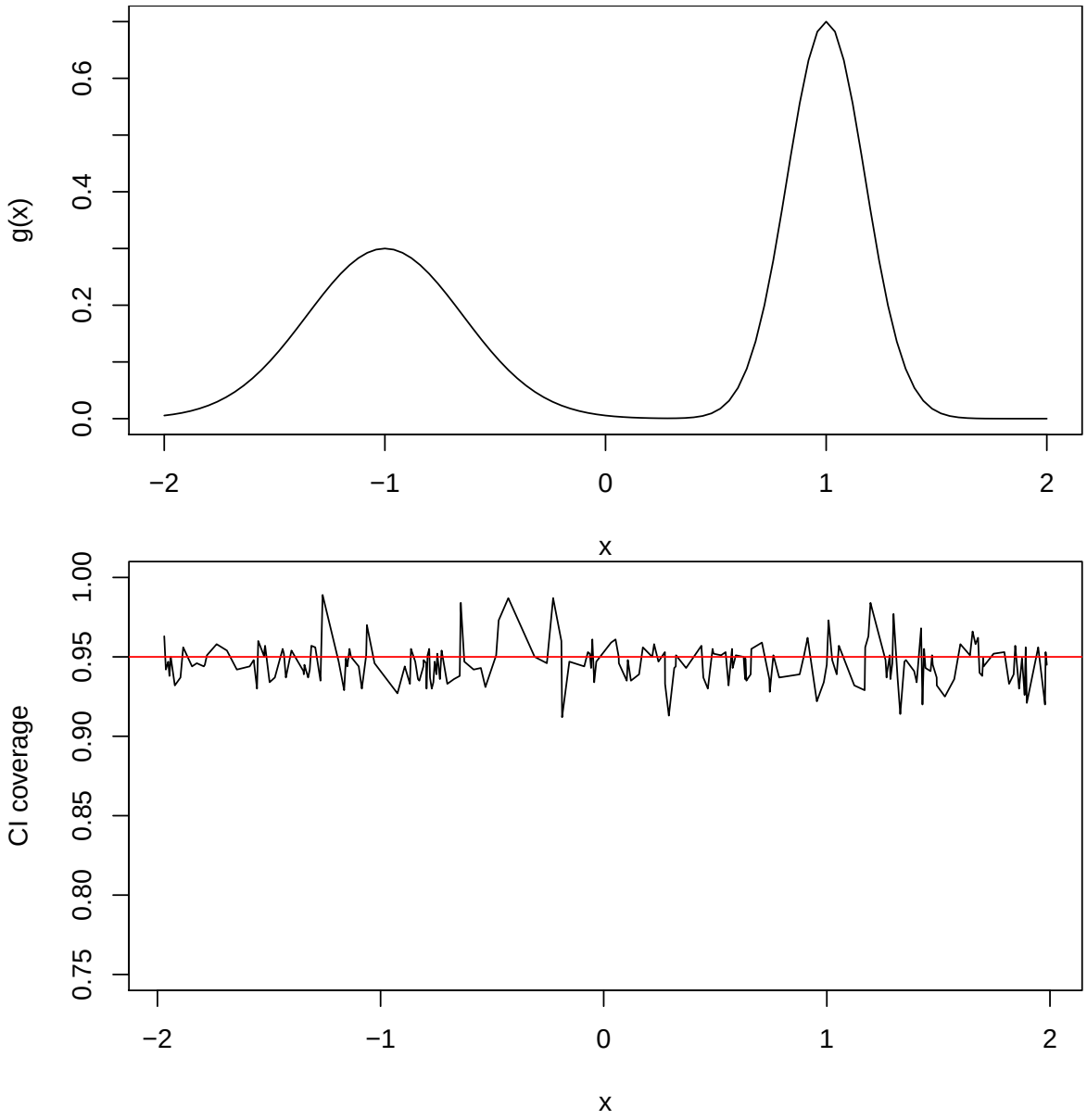

Figure 3.21: Pointwise Bias assessment tool performance for local linear for Example 3 according to the setup in Scenario 4. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

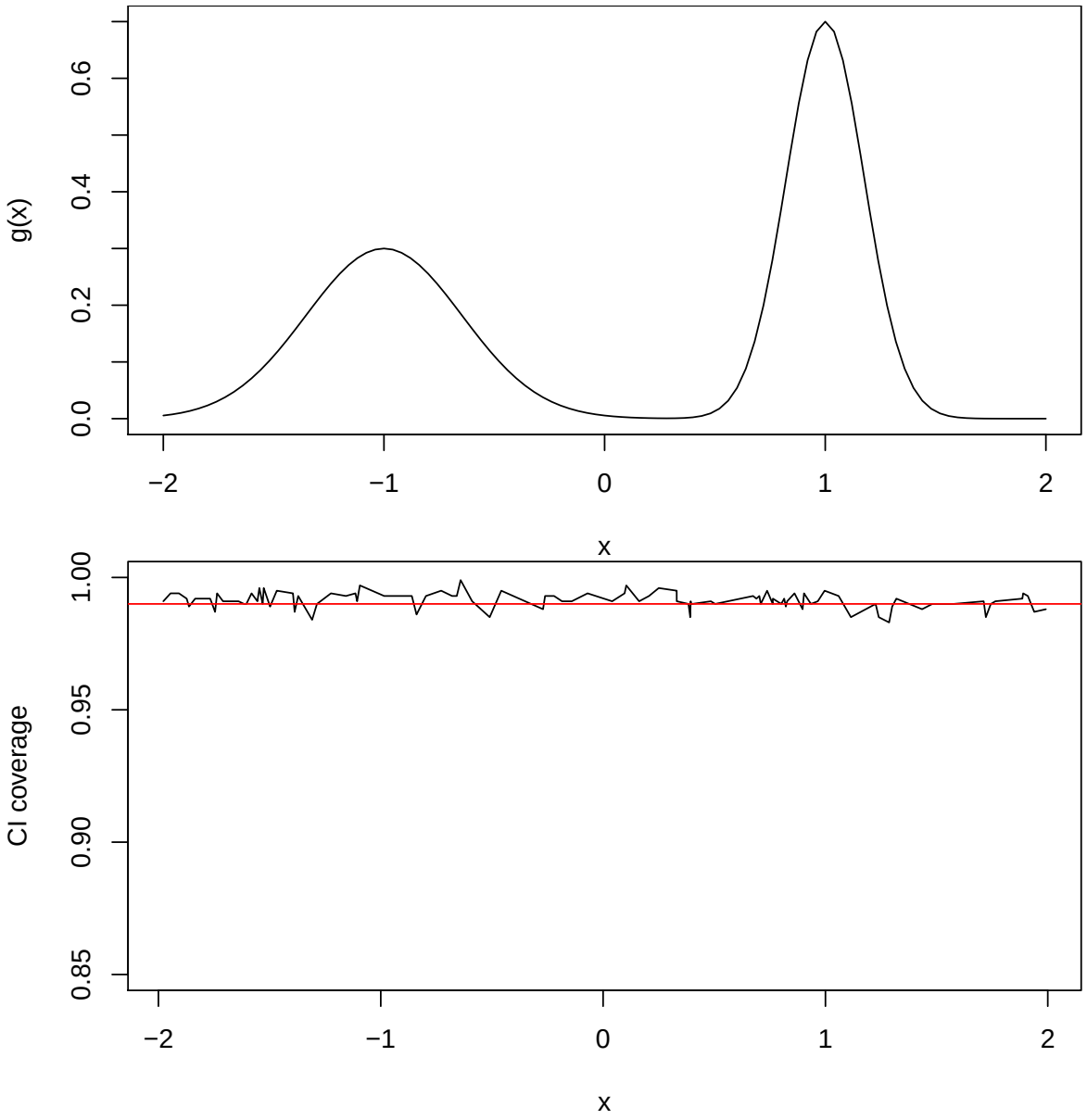

Figure 3.22: Pointwise Bias assessment tool performance for local linear for Example 3 according to the setup in Scenario 5. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

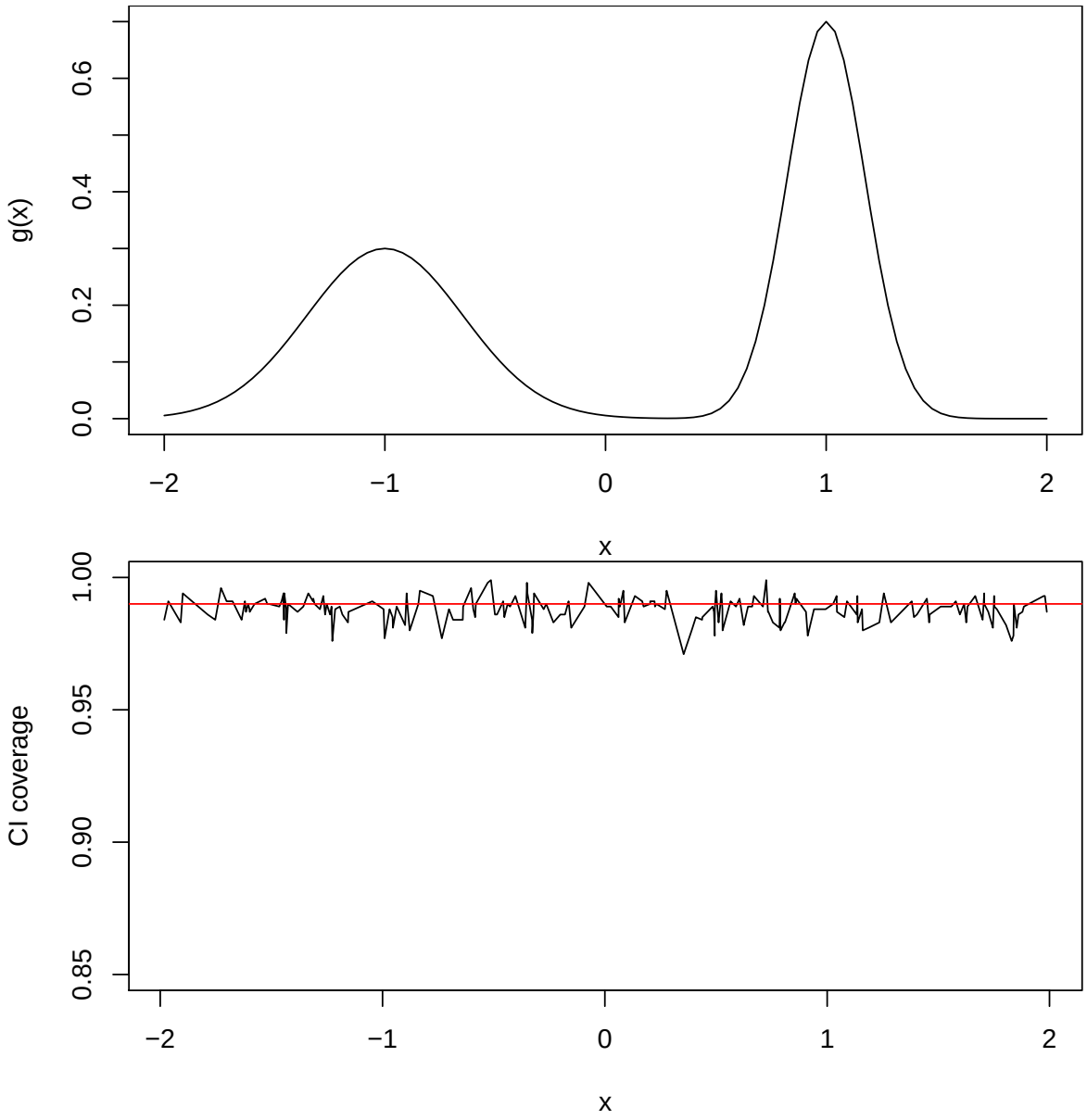

Figure 3.23: Pointwise Bias assessment tool performance for local linear for Example 3 according to the setup in Scenario 6. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

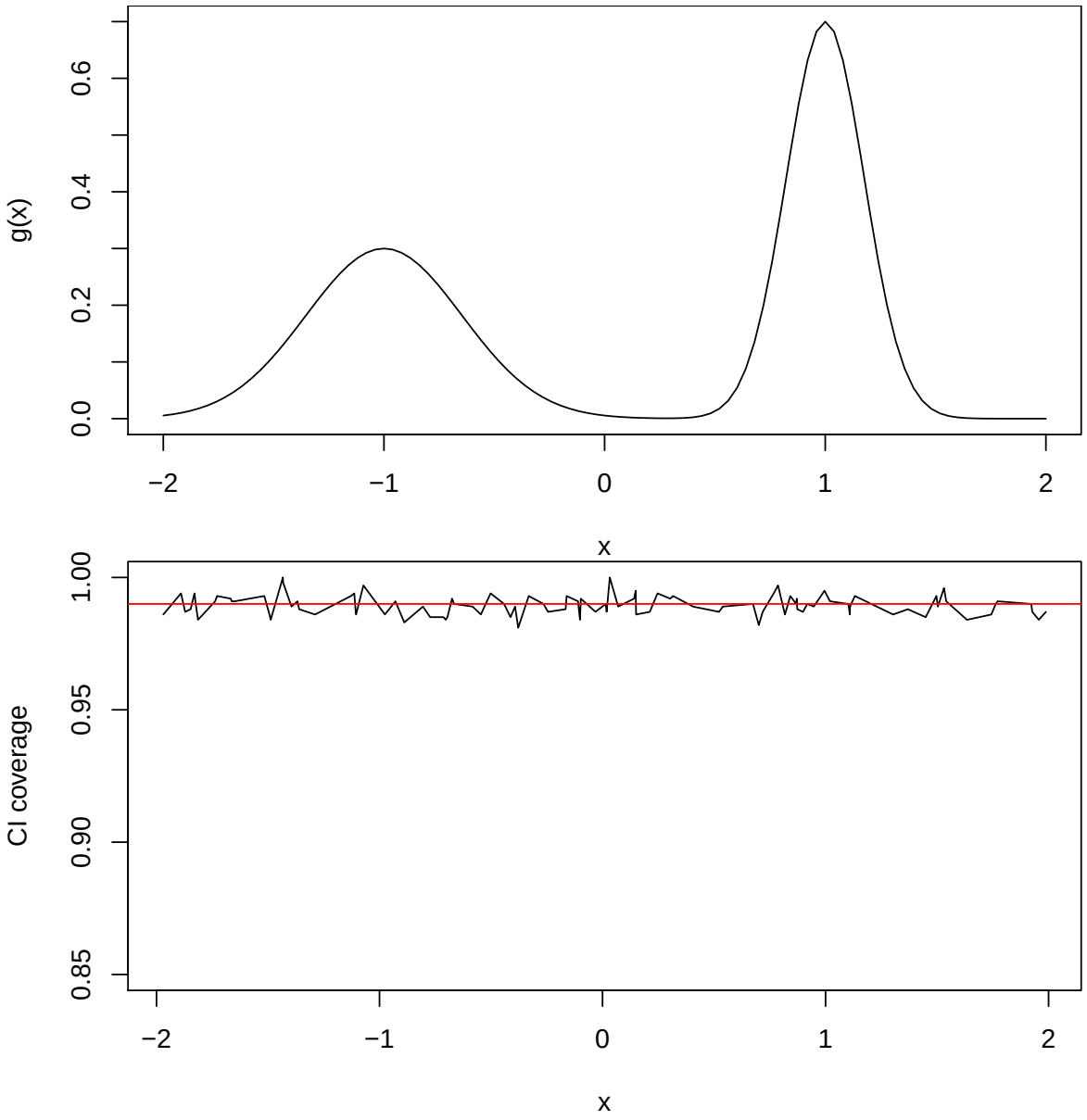

Figure 3.24: Pointwise Bias assessment tool performance for local linear for Example 3 according to the setup in Scenario 7. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

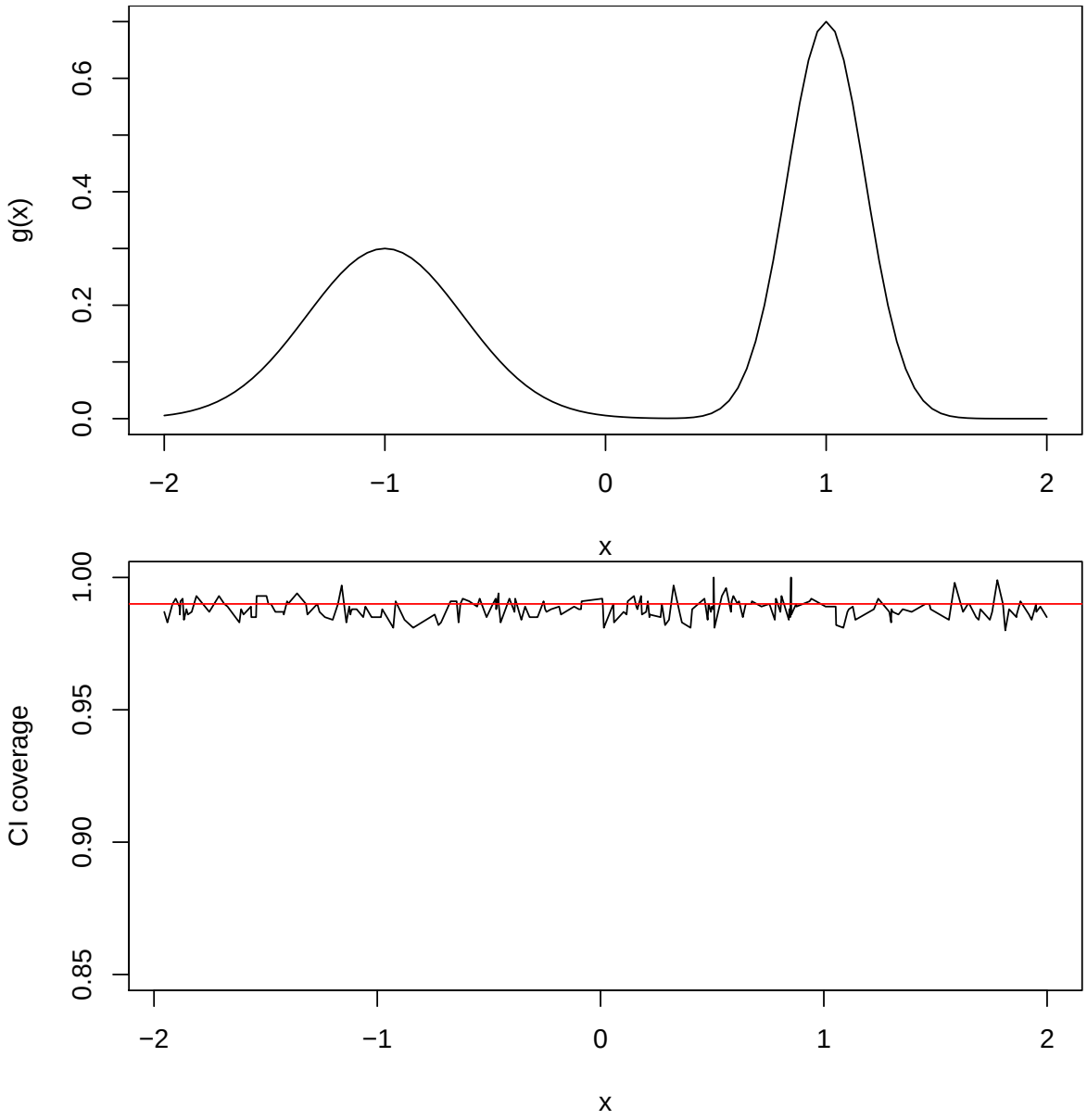

Figure 3.25: Pointwise Bias assessment tool performance for local linear for Example 3 according to the setup in Scenario 8. For reference, the target function is pictured in the top panel. Actual confidence interval coverage is displayed in the bottom panel, with the nominal level plotted as a horizontal line in red.

3.4 Illustrative Examples

In this section, we apply the proposed method to some real data. In each example, a local polynomial regression is fit and the diagnostic tool is used to visualize the possible extent of the bias in the fit.

3.4.1 Application to Old Faithful Data

The first example is the Old Faithful data set discussed in Härdle (1991): this data set contains waiting times between eruptions (in minutes) and the duration of the eruption for the Old Faithful Geyser in Yellowstone National Park, Wyoming, USA. There are two variables in this data set: "eruptions" for eruption time in minutes, and "waiting" for waiting time to the next eruption. Figure 3.26 to Figure 3.28 display the plots of the local polynomial fitting along with 95% nominal pointwise confidence intervals for both of the fitting (top) and the estimates of bias (bottom). In each plot, different powers of the polynomial k_1, k_2 and bandwidths h_1, h_2 are used. After seeing the plots, we compared the change of efficiency (variance) for the estimate of the bias.

When keeping the bandwidth fixed, changing the degree of the polynomials to a higher degree will produce smaller confidence bands of bias estimate and also smaller bias. This can be visualized from the bottom panels of Figure 3.27 and Figure 3.28. On the top of these two plots, we can see the variance of the estimates are getting larger as would be expected. This demonstrates the bias-variance trade-off and our bias confidence bands plot is a helpful support for assessing the bias while utilizing local polynomial regression.

Holding the degree of the local polynomials constant and increasing the bandwidths gives the results shown in Figures 3.29 and 3.30. A clear pattern is easy to see: a larger bandwidth is associated with more bias and wider confidence bands for bias. We also see that the confidence bands for the regression function are tighter when the bandwidth is larger, an effect of reduction in variance.

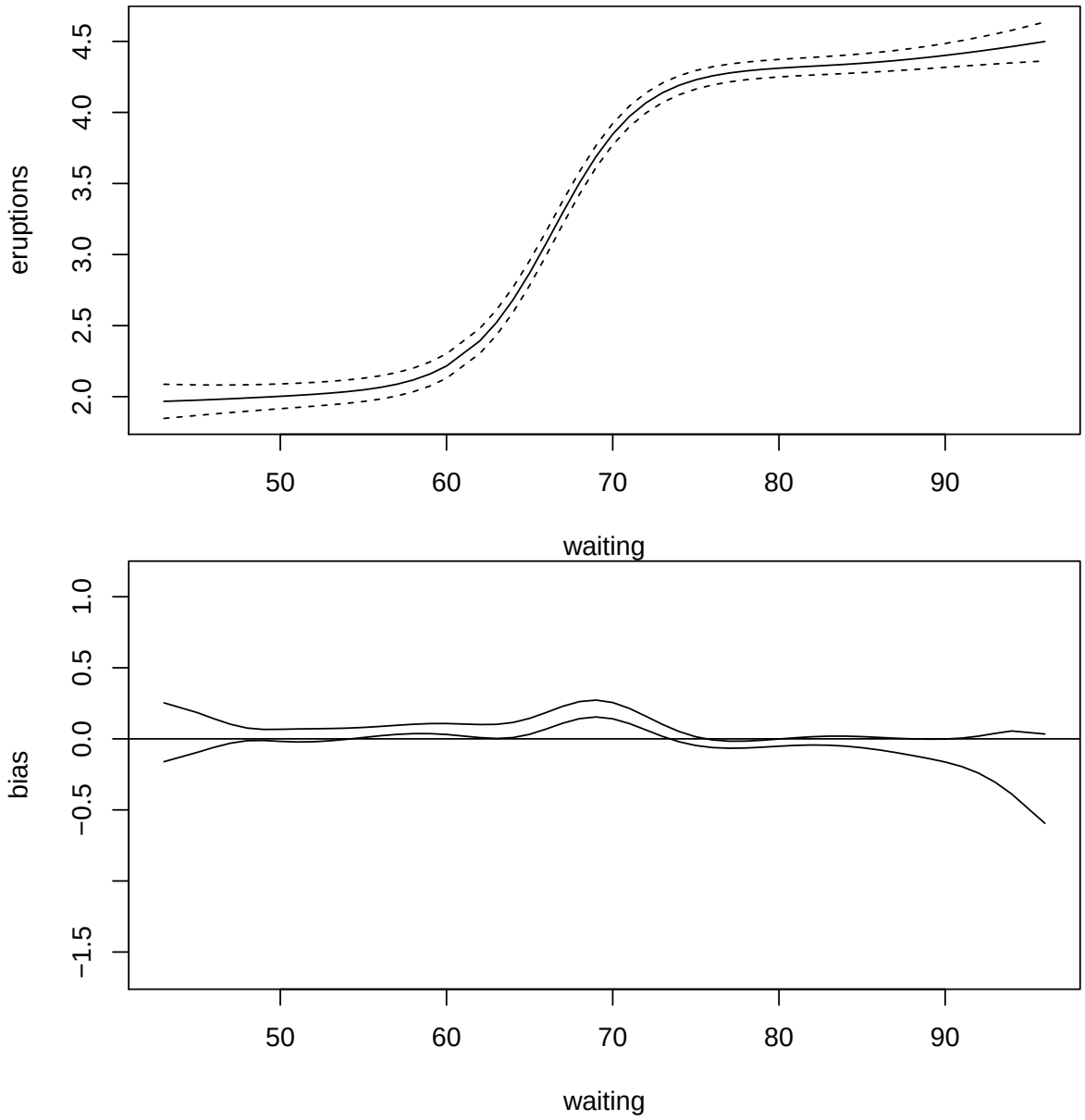

Figure 3.26: Old Faithful data, Top Panel: local polynomial fitting with $k_1 = 0$ and $h_1 = 5$. Bottom Panel: Pointwise confidence interval plot for bias with $k_2 = 1$ and $h_2 = 4$

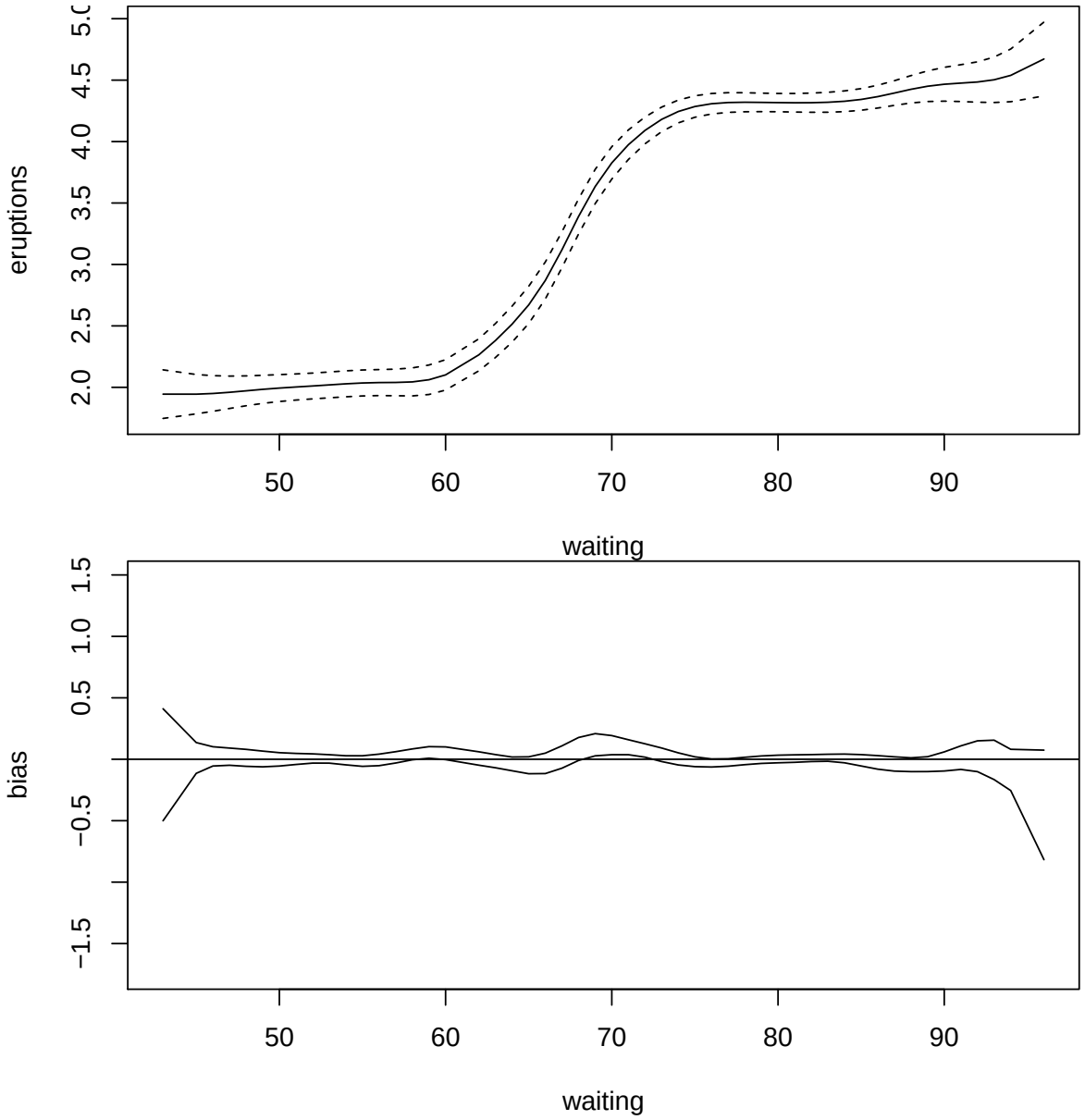

Figure 3.27: Old Faithful data, Top Panel: local polynomial fitting with $k_1 = 0$ and $h_1 = 2.5$. Bottom Panel: Pointwise confidence interval plot for bias with $k_2 = 1$ and $h_2 = 2$

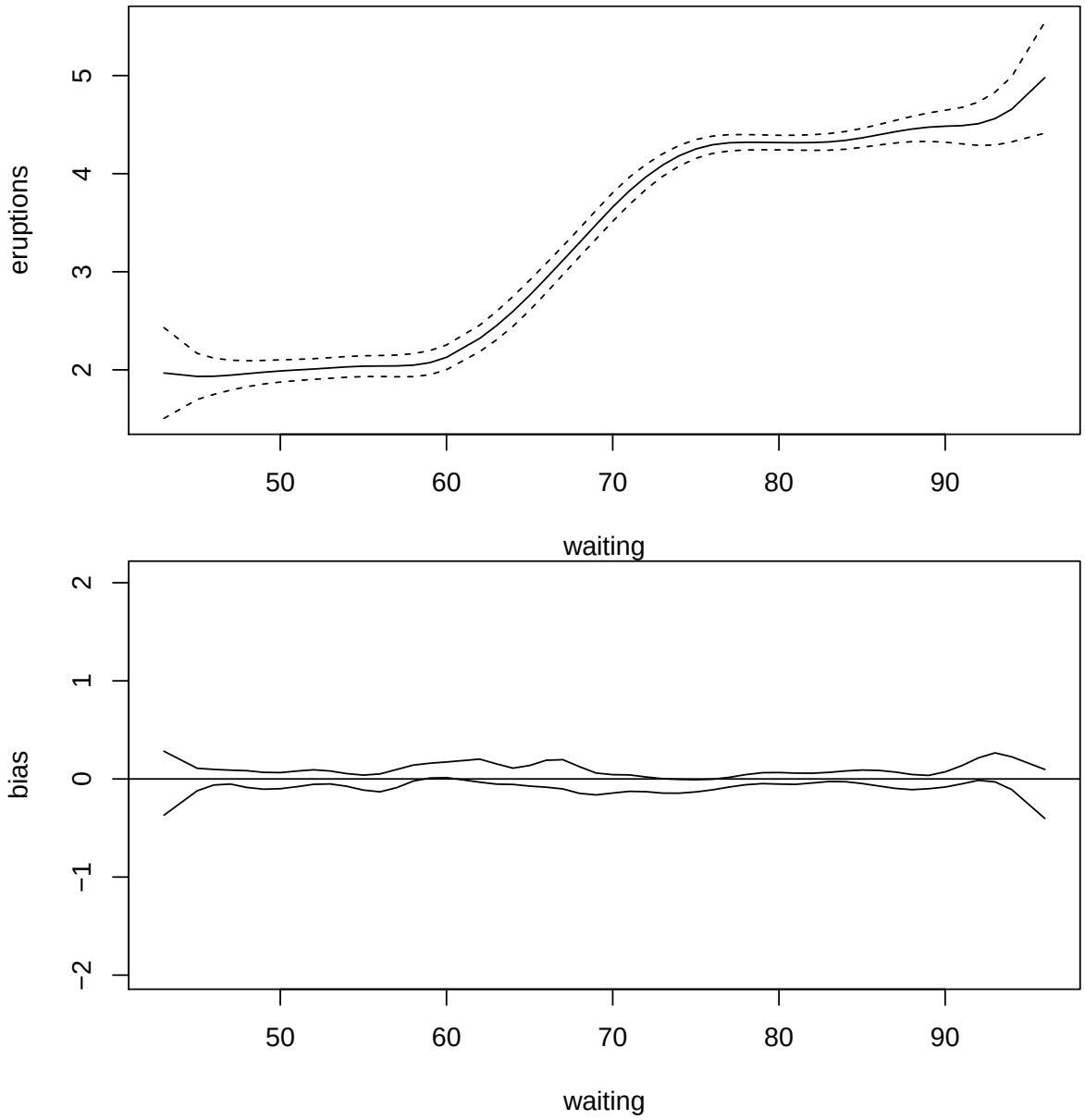

Figure 3.28: Old Faithful data, Top Panel: local polynomial fitting with $k_1 = 1$ and $h_1 = 2.5$. Bottom Panel: Pointwise confidence interval plot for bias with $k_2 = 2$ and $h_2 = 2$

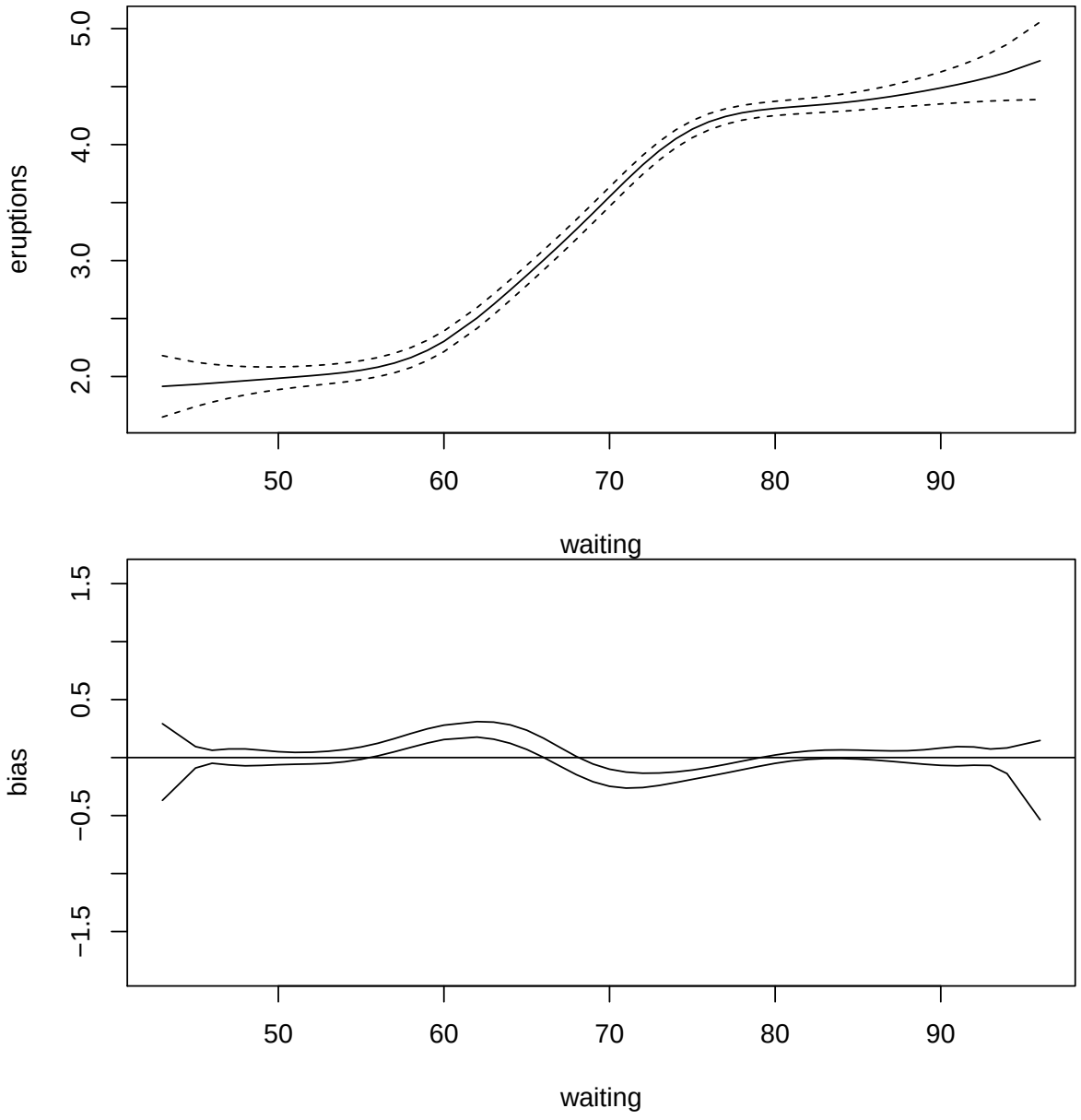

Figure 3.29: Old Faithful data, Top Panel: local polynomial fitting with $k_1 = 1$ and $h_1 = 5$. Bottom Panel: Pointwise confidence interval plot for bias with $k_2 = 2$ and $h_2 = 4$

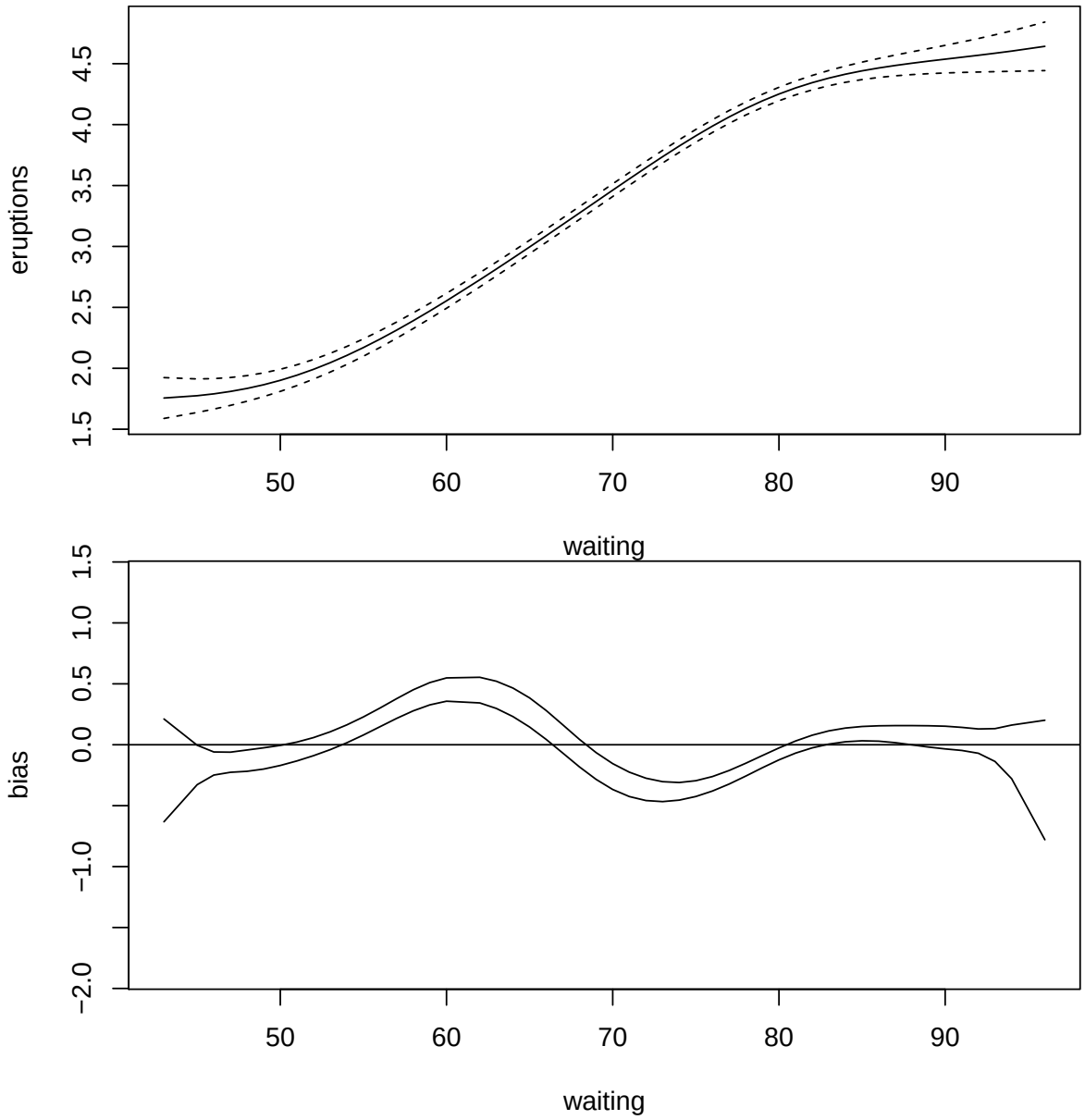

Figure 3.30: Old Faithful data, Top Panel: local polynomial fitting with $k_1 = 1$ and $h_1 = 10$. Bottom Panel: Pointwise confidence interval plot for bias with $k_2 = 2$ and $h_2 = 8$

3.4.2 Application to Beluga Whale Data

Our next example concerns a data set for the study of the development of beluga whales that focuses on the nursing behavior of mother whales. There are two variables: "time" is for the the index of time period. "nursing" is for the square root of the number of seconds spent successfully nursing during the period. Different bandwidths are tested with the same degrees of the polynomials. The plots are shown from Figure 3.31 to Figure 3.33. Similar patterns to those seen in the Old Faithful Geyser example are evident here.

In Fig 3.31, we see what happens when we use a bandwidth close to the "plug in" choice (Loader, 1999). The bias confidence interval fails to contain the 0-reference line at several locations, suggesting notable bias exists at each location. There appears to be substantial bias at the highest peak, and the confidence interval also suggests significant bias near both the left and the right boundary. The effect of using a bandwidth $h_1 = 5$ is seen in Fig 3.32. The confidence bands for the bias still fail to contain 0 at the highest peak location, indicating that bias is still a problem there, but it is reduced, due to the use of the smaller bandwidth being used.

Further reducing the bandwidth is associated with relatively smaller bias (Fig 3.33). However, the bias assessment results in Fig 3.31 to Fig 3.33 suggested an adaptive bandwidth is probably necessary to handle this problem.

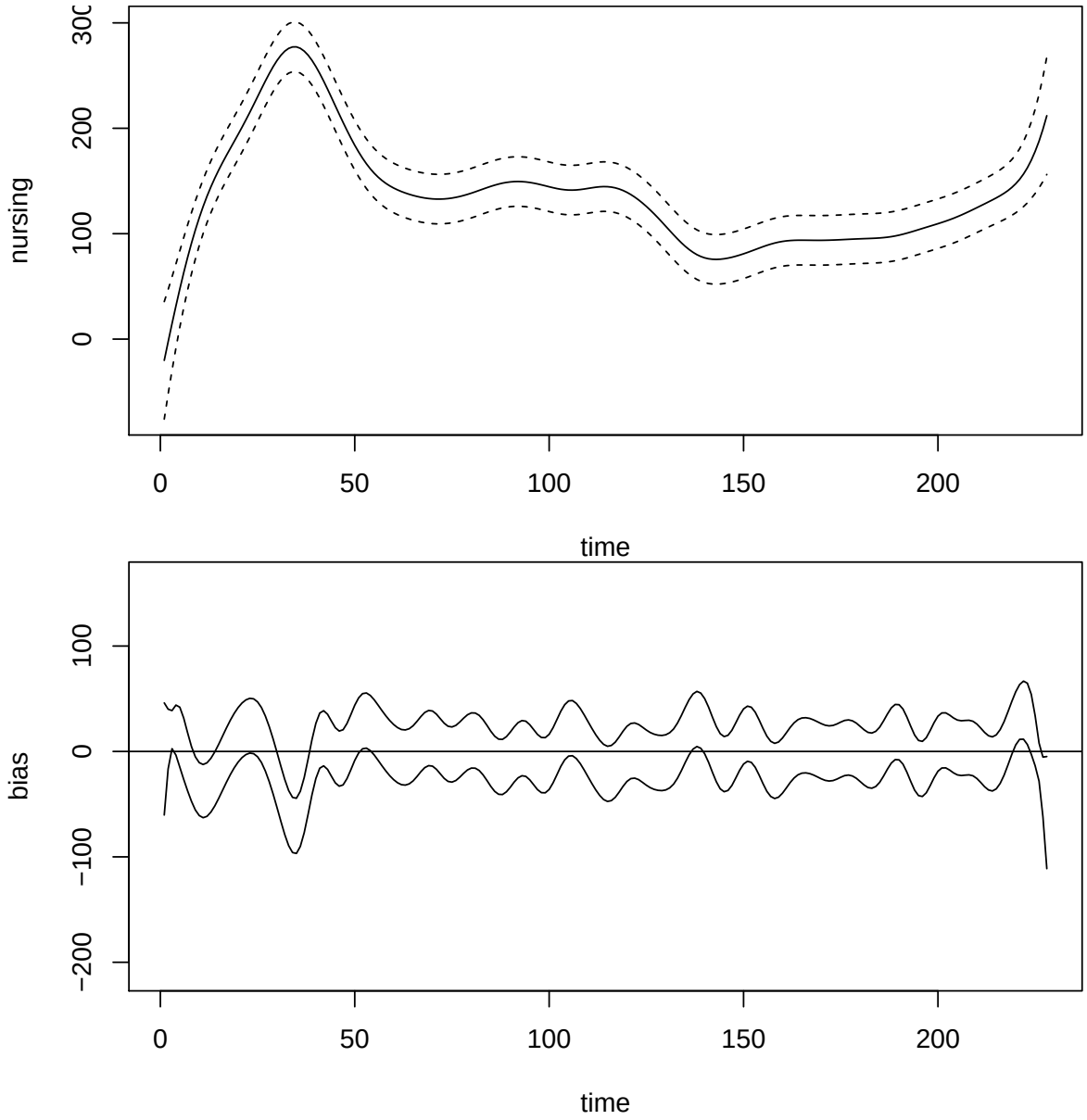

Figure 3.31: Beluga data, Top Panel: local polynomial fitting with $k_1 = 1$ and $h_1 = 7$. Bottom Panel: Pointwise confidence interval plot for bias using $k_2 = 2$ and $h_2 = 4$

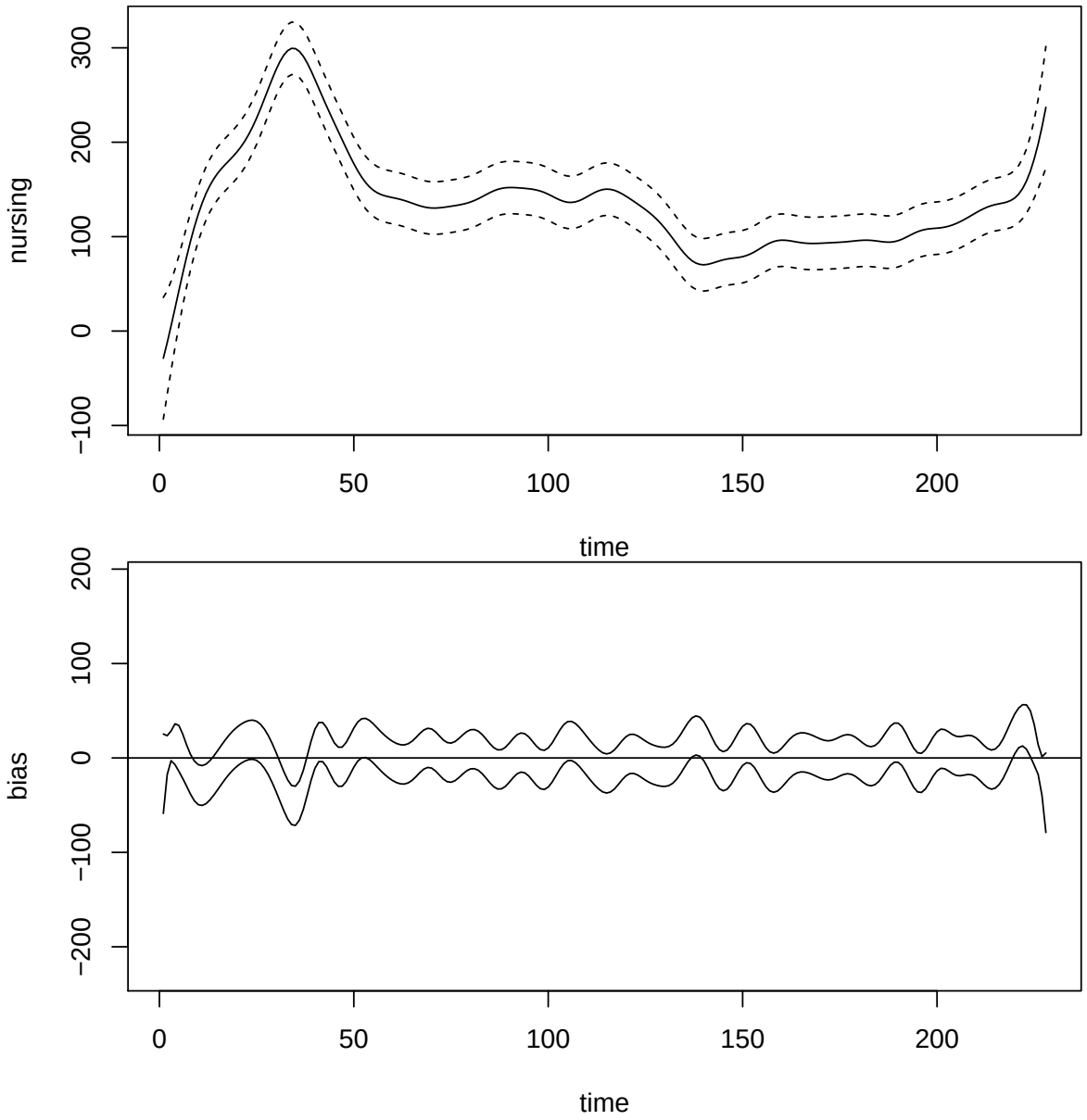

Figure 3.32: Beluga data, Top Panel: local polynomial fitting with $k_1 = 1$ and $h_1 = 5$. Bottom Panel: Pointwise confidence interval plot for bias using $k_2 = 2$ and $h_2 = 4$

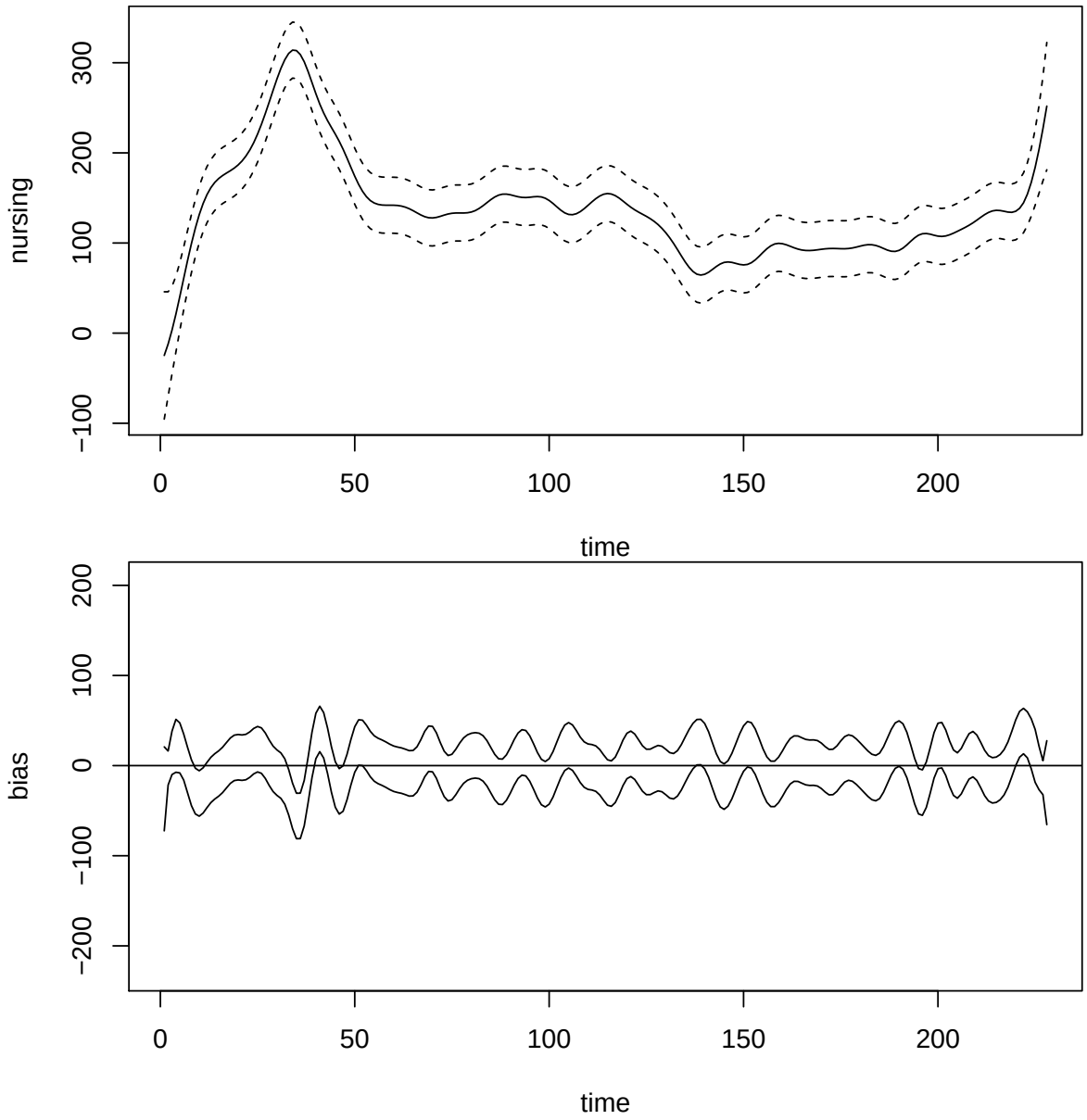

Figure 3.33: Beluga data, Top Panel: local polynomial fitting with $k_1 = 1$ and $h_1 = 4$. Bottom Panel: Pointwise confidence interval plot for bias using $k_2 = 2$ and $h_2 = 3$

3.4.3 Application to Ethanol Data

The last real data example is the ethanol data set from Simonoff (2012). The study of He and Huang (2009) used this data set to compare their proposed method with the local linear estimator, local cubic estimator and the data sharpened estimator in the study of Choi and Hall (1998). In the current study, with the bandwidth adopted from the double-smoothing local linear estimation procedure in the study of He and Huang (2009), several local polynomial fittings are displayed in Fig 3.34 and Fig 3.35. Without varying $k1$, $k2$ but varying the bandwidths, it is clear that a smaller bandwidth generates smaller bias and narrower confidence bands. This is the same effect as in the previous illustrative examples.

While reducing the bandwidth as in Fig 3.35, bias is much smaller compared to Fig 3.34.

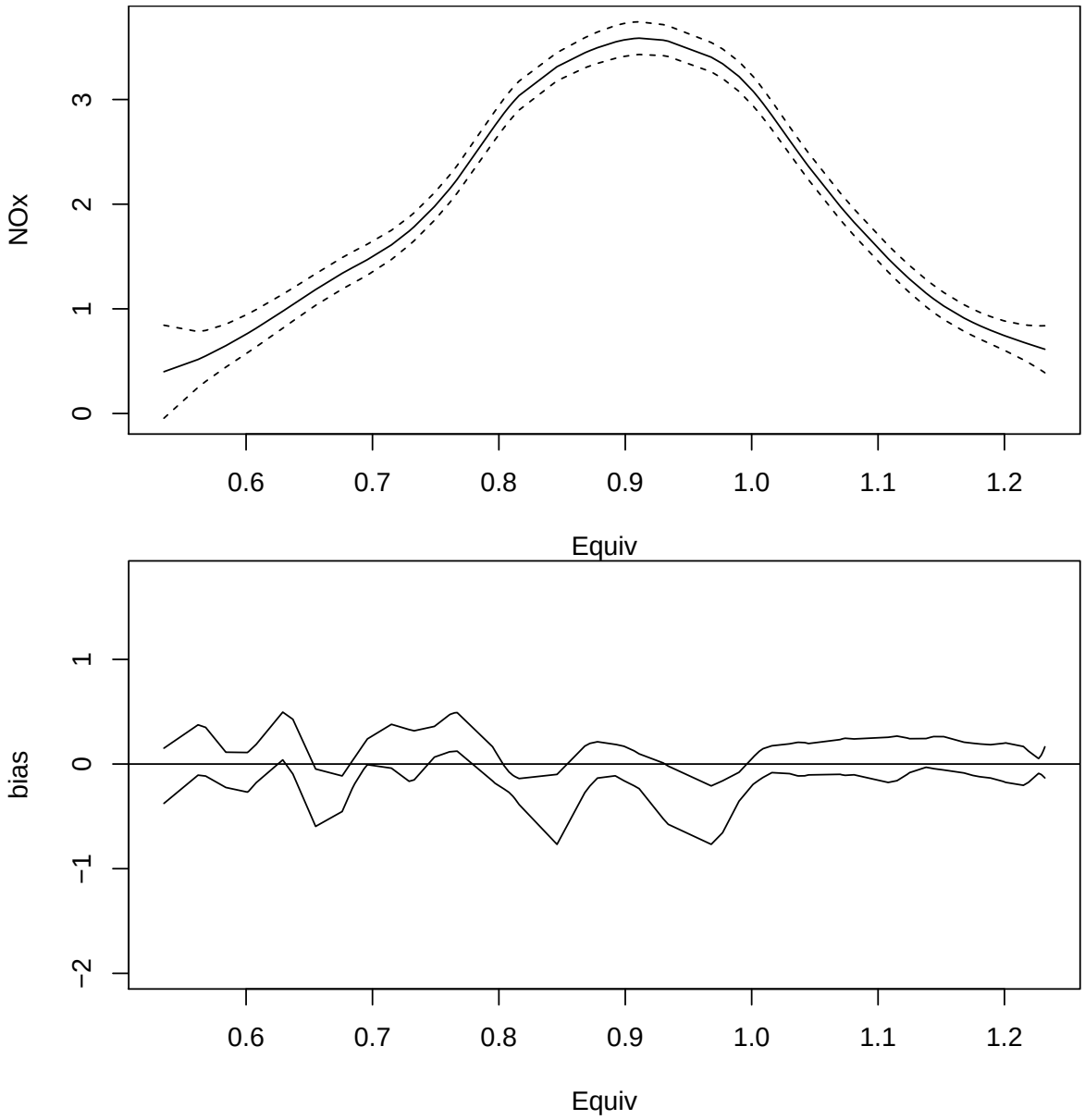

Figure 3.34: Ethanol data, Top Panel: local polynomial fitting with $h_1 = 0.4$ which He and Huang 2009 used for local cubic regression and $k_1 = 1$.
Bottom Panel: Pointwise confidence interval plot for bias using $k_2 = 2$ and $h_2 = 0.02$

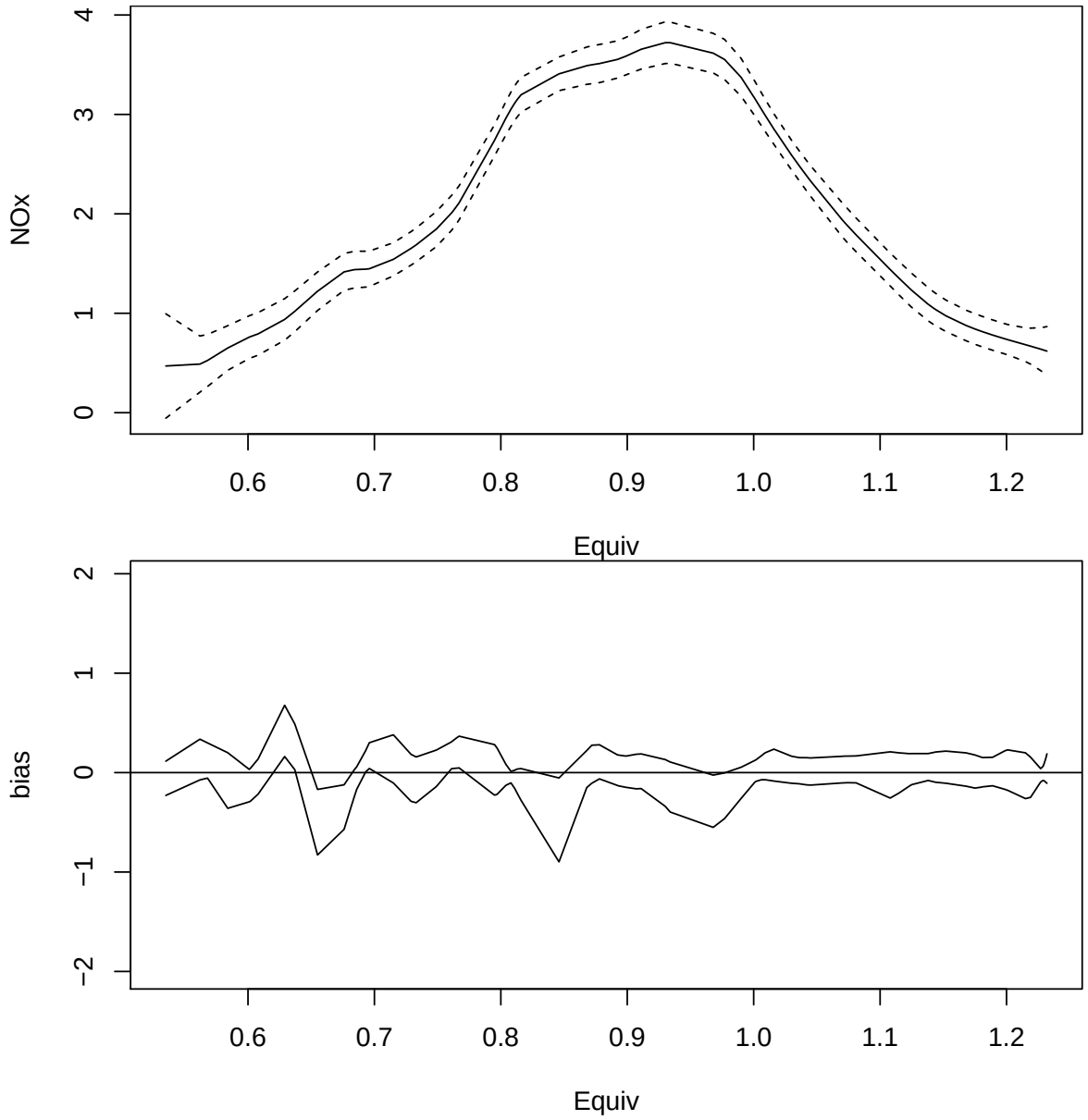

Figure 3.35: Ethanol data, Top Panel: local polynomial fitting with bandwidth as 0.0253 as He and Huang used for local linear regression and $k1 = 1$; Bottom Panel: Pointwise confidence interval plot for bias using $k_2 = 2$ and $h_2 = 0.0153$

3.5 Concluding Remarks

This study proposed a new support tool when fitting local polynomial regressions to scatterplot data. We have proposed an estimator of bias based on the difference between a less biased pilot estimator and the expected value of the local polynomial estimator applied to the pilot estimator. By calculating the exact variance of the bias estimator, we constructed pointwise confidence bands for the bias as a function of the covariate. Simulations at a variety of parameter settings for three different target functions provide evidence for the accuracy of the tool, in that the actual confidence interval coverage appears to match the nominal coverage very well.

In three real data examples, the usefulness of our assessment plots is clear. It can help us to visualize the possible magnitude of the bias in particular local polynomial estimates. The tool also demonstrates the well known bias-variance trade off. It also gives clear indications of where the local polynomial regression could fail in that the bias would be large: that is, the regions having high curvature.

Chapter 4

Double-smoothing

4.1 Introduction

Double-smoothing is a bias reduction technique for local linear regression that was introduced by He and Huang (2009), generalizing a method proposed by Choi and Hall (1998). The crux of the method is to take a weighted average of multiple local linear least-squares estimates of the regression function at each grid point t. By doing so, bias reduction can be achieved, because the first level estimates can be obtained using a somewhat smaller bandwidth than would normally be used leading to a smaller bias but more variance; the averaging process reduces the variance. The effect is an estimate whose variance has not increased substantially but whose bias has. A side effect of the method is that mild data sparsity issues are somewhat mitigated.

A natural question to ask is whether the technique of double-smoothing can be applied to higher dimensional local polynomial regression. We are motivated by the fact that local quadratic regression and local cubic regression do not work well with sparse data. Double-smoothing techniques might mitigate this. Thus, this chapter will expand the study of He and Huang (2009) to higher degree local polynomials.

4.2 Higher Degree Double-Smoothing

The possibility of extending double-smoothing to higher degree local polynomial regression is now considered for the case of local quadratic regression methods where we wish to estimate a smooth function $m(x)$ at a point $x_0 \in X$, (where X is interior to the support of the estimator), using n independent observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. The assumed model is

$$y = m(x) + \varepsilon \tag{4.1}$$

where ε is noise with constant standard deviation σ. We assume that $K_h(x) = K(x/h)/h$ where $K(x)$ is a sufficiently smooth symmetric positive density function, such as the Gaussian kernel. Assumptions made here are standard (see, for example, Fan and Gijbels (1996)).

By Taylor's theorem, we know that minimizing

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1(x_i - x) - \beta_2(x_i - x)^2)^2 K_h(x_i - x) \tag{4.2}$$

with respect to β_0, β_1 and β_2 allows us to estimate $m(x_0)$ as

$$\widehat{m(x_0)} = \widehat{\beta_0} + \widehat{\beta_1}(x_0 - x) + \frac{1}{2}\widehat{\beta_2}(x_0 - x)^2 \tag{4.3}$$

with an asymptotic bias (as $n \to \infty$ and $h \to 0$, $nh \to \infty$) of order $O(h^2)$, when $x_0 = x$. When $x_0 \neq x$, an additional bias of order $O((x_0 - x)^2)$ is induced. This additional bias can be reduced or eliminated by averaging a pair of estimates of $m(x_0)$ based on minimizations at x^* and x^+ where $(x_0 - x^*) = (x^+ - x_0)$.

The asymptotic variance for estimates of $\widehat{m(x_0)}$ based on any target point x is of the order $O((nh)^{-1})$. Thus, as long as x_0 is not far from x, we can hope to achieve an improvement in estimation accuracy for $m(x_0)$ by taking a weighted average of estimates of the form $\widehat{m(x_0)}$. As discovered by He and Huang (2009), we can accomplish this averaging through the use of integrating estimates of the form $\widehat{m(x_0)}$ over a range of x values, using a symmetric kernel

function L_{h_2} to control the weighting.

Thus, the double-smoothing local quadratic estimator can be expressed as:

$$\widehat{m(x_0)} = \int_a^b \Big\{\widehat{\beta_0(x)} + \widehat{\beta_1(x)} * (x_0 - x) + \frac{\widehat{\beta_2(x)} * (x_0 - x)^2}{2!}\Big\} L_{h_2}(x - x_0)dx. \tag{4.4}$$

We will assume L is the same as K for simplicity. We will distinguish the bandwidth h_2 for the second level of smoothing from the bandwidth h used for the first level of smoothing. a and b define the support of the regression function estimate.

A similar idea can be used to extend double-smoothing to the local cubic case. Thus, the double-smoothing local cubic estimator can be expressed as:

$$\int_a^b \Big\{\widehat{\beta_0(x)} + \widehat{\beta_1(x)} * (x_0 - x) + \frac{\widehat{\beta_2(x)} * (x_0 - x)^2}{2!} + \frac{\widehat{\beta_3(x)} * (x_0 - x)^3}{3!}\Big\} L_{h_2}(x - x_0))dx. \tag{4.5}$$

Even higher degree local polynomial estimation can be considered but usually not in practice.

4.3 Bandwidth Selection

In He and Huang (2009), bandwidth selection is considered, but really only from a theoretical point of view. The bandwidths used in the simulation studies described there were chosen to minimize the asymptotic MISE, but there was no guidance provided for selecting a bandwidth for a given set of data. To remedy this situation, a form of cross-validation selection is likely to give the best results, but it must be modified, to overcome the computationally intensive nature of the double-smoothing algorithm. If leave-one-out cross-validation were to be applied to double-smoothing local polynomial regression directly, a large number of regressions would need to be computed. Furthermore, two parameters must be selected instead of only one.

In their simulation work and their examples, He and Huang employ the same bandwidth

for both steps of the double-smoothing algorithm. This reduces the dimensionality of the bandwidth selection but, as we will later see, this also reduces the quality of the estimates obtained.

In the past, there has been debate in the literature regarding how to choose the bandwidth in local polynomial regression (summarized in the paper Loader (1999)). Plug-in approaches have been popular, since they tend to be fairly quick to compute, and they have the appearance of being asymptotically accurate, balancing the trade-off between bias and variance. Loader argues that cross-validation type approaches have their place and that they can sometimes be superior to plug-in methods. Cross-validation can also be used when the asymptotic expressions for the bias and variance are not available.

In the simulation studies and illustrative examples that follow in this chapter, we use a form of cross-validation to select the smoothing parameter in the local linear, quadratic and cubic regression estimators. We do this in order to have a basis for comparison with the double-smoothing algorithm where direct-plug-in methods are more difficult to derive. In order for the method to be quick in simulations, we repeatedly randomly remove 90% of the sample and use the resulting fitted model to predict the remaining 10% of the data. The bandwidth h that minimizes the sum of the squared predicted residuals is used.

In the double-smoothing algorithm, we consider differing bandwidths for the two levels of smoothing as well as the use of the same bandwidth at both levels. The bandwidth at the first stage of smoothing is taken to be kh where k is a multiplier, usually taken to be less than 1.0. When a different bandwidth is required for the second level of smoothing, it is taken to be the average of the successive differences in the predictor values. This admittedly ad hoc method has the virtue of being quick and simple, and there is some justification: the double-smoothing method is essentially based on a weighted average of kernel regression estimates in the neighbourhood of the conventional estimate, and this rule ensures that, if a Gaussian kernel is used, only a relatively small neighbourhood will be used. This prevents the method from causing serious distortions.

Choosing the multiplier k to be less than 1 has the effect of decreasing the bias in the first

level of smoothing while increasing the variance. The weighted averaging of the kernel estimates through the second level of smoothing should have the effect of decreasing the variance, while modestly reintroducing bias (induced by the use of regressions at neighbouring points).

4.4 Simulation Study

Using the target functions from He and Huang (2009), we can evaluate performance of double-smoothing local polynomial regression by simulation. The target functions are

$$m_1(x) = x + 2\mathrm{e}^{-16x^2}, \quad -2 < x < 2 \tag{4.6}$$

and

$$m_2(x) = \sin(2x) + 2\mathrm{e}^{-16x^2}, \quad -2 < x < 2. \tag{4.7}$$

In our simulation study, we compare the double-smoothed local polynomial estimators with their local polynomial counterparts, using both of these targets, one at a time. Gaussian kernels were used in all cases and at both levels of double-smoothing. In each case, we have simulated 1000 sets of data, computed the estimates and plotted their biases, MSEs, and absolute deviation errors as functions of x.

4.4.1 Target Function 1

In our simulations of target function 1, we generated 1000 samples of size 50 from the model

$$y = m_1(x) + \varepsilon$$

where ε is independent normal noise with standard deviation 0.4.

For each sample, we computed the cross-validation bandwidths h corresponding to local linear, local quadratic and local cubic regressions, and we computed the corresponding regressions. Then we computed the double-smoothed counterparts for each polynomial degree, using a multiplier of $k = .7, .8, .9$, successively, to obtain the first level smoothing parameter values. We then used the ad hoc technique to choose the second level smoothing parameter, resulting in three sets of double-smoothed estimates for each polynomial degree. We next computed an additional three sets of double-smoothed estimates (corresponding to the three values of k) by using the same bandwidth h at both levels of smoothing.

Bias Results

Figure 4.1 shows the pointwise biases which were calculated as the difference between $m_1(x)$ and the regression estimates for predictor values between -1.375 and 1.375. This interval avoids the boundaries – see the comments at the end of this section.

The bandwidth used for the first level of smoothing was $h = 0.1709386$ for the local linear regression estimates. The bandwidths used at the first level of double-smoothing were, respectively, $.7h$, $.8h$ and $.9h$. The corresponding bandwidths used at the second level of double-smoothing were all 0.0759514. The plots of the pointwise bias appear in the top panel of the figure, where the dashed curves correspond to the local linear bias, and the solid curves correspond to the double smoothed local linear bias.

The lower panel of Figure 4.1 shows the pointwise local linear and double smoothed local linear bias under almost the same conditions as above, except that the bandwidths used at the second level of double-smoothing are $.7h$, $.8h$ and $.9h$, respectively.

What is evident from the figure is that the improvement of double-smoothing bias over conventional local linear decreases as the multiplier k increases. What is clear from the lower panel of the figure is that using the same smoothing bandwidth at the second level of double-

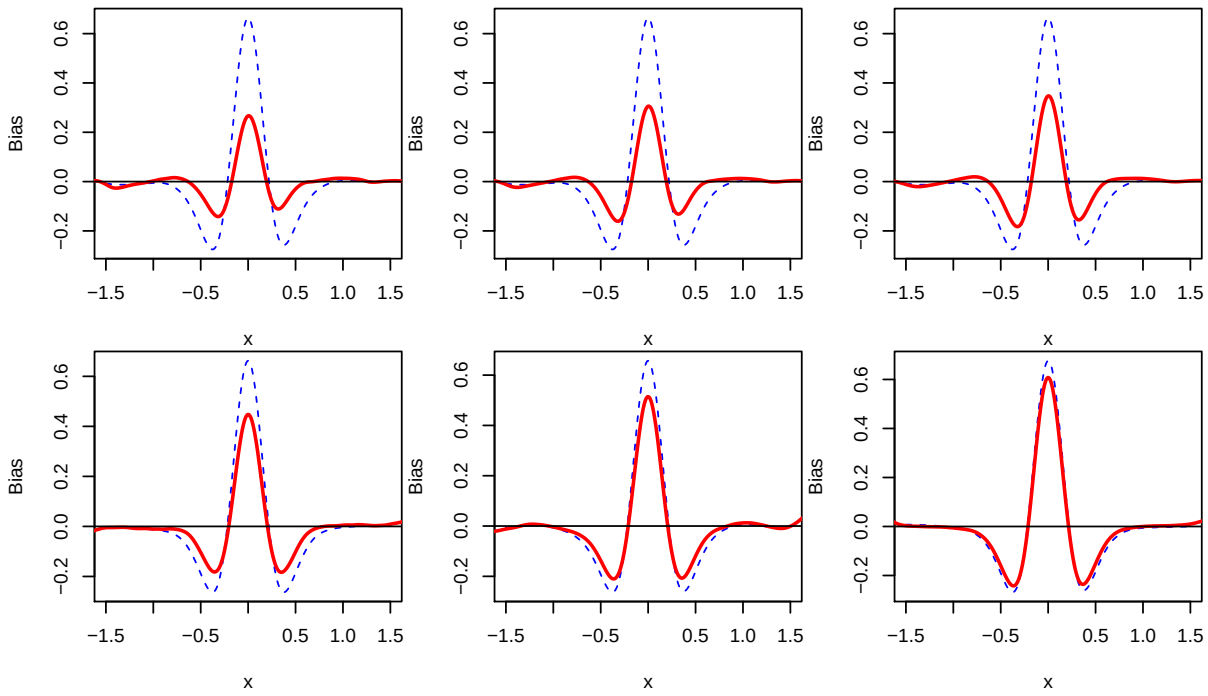

Figure 4.1: Pointwise Bias for local linear (blue dashed curve) and double-smoothed local linear (red solid curve) regression applied to simulated data from Example 1 model. Left Column: Multiplier is 0.7; Middle Column: Multiplier is 0.8; Right Column: Multiplier is 0.9. Top Row: Bandwidths at both levels of double-smoothing are different; Bottom Row: Bandwidths are the same.

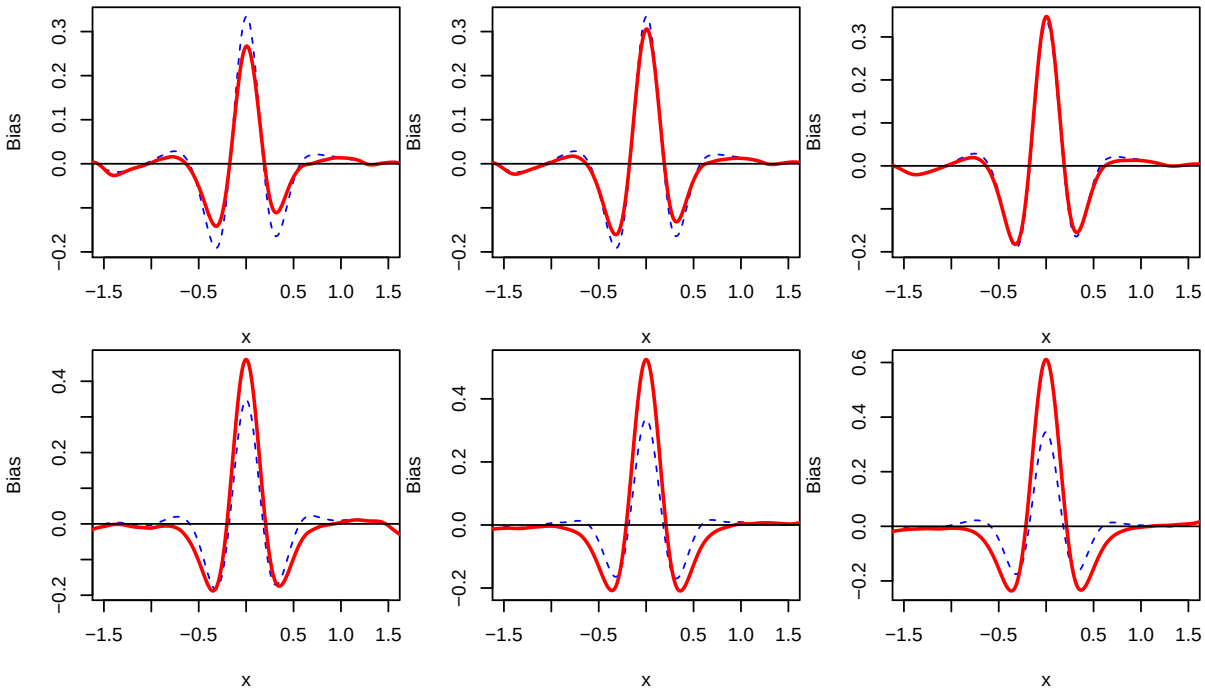

Figure 4.2: Pointwise Bias for local quadratic (blue dashed curve) and double-smoothed local quadratic (red solid curve) regression applied to simulated data from Example 1 model. Left Column: Multiplier is 0.7; Middle Column: Multiplier is 0.8; Right Column: Multiplier is 0.9. Top Row: Bandwidths at both levels of double-smoothing are different; Bottom Row: Bandwidths are the same.

smoothing can lead to deterioration in performance. The bias tends to be larger for double-smoothing than for conventional local linear smoothing, in this example, when the first and second level bandwidths are the same. Again, the bias tends to increase as the multiplier k increases.

In a similar manner to Figure 4.1, Figures 4.2 and 4.3 contain pointwise bias comparison plots for local quadratic and local cubic and their double-smoothed counterparts.

The bandwidth used for the first level of smoothing was $h = 0.1709386$ for the local quadratic regression estimates. The bandwidths used at the first level of double-smoothing were, respectively, $.7h$, $.8h$ and $.9h$. The corresponding bandwidths used at the second level of double-smoothing were all 0.0759514. The plots of the pointwise bias appear in the top panel of the figure, where the dashed curves correspond to the local quadratic bias, and the solid curves correspond to the double smoothed local quadratic bias.

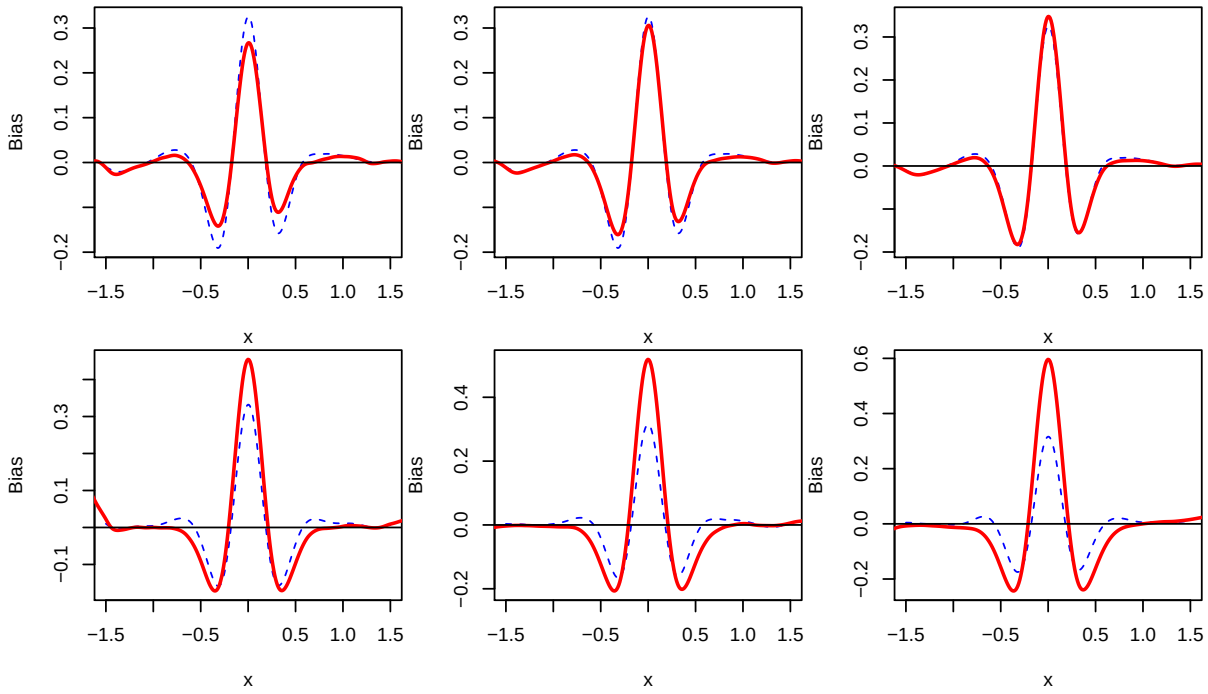

Figure 4.3: Pointwise Bias for local cubic (blue dashed curve) and double-smoothed local cubic (red solid curve) regression applied to simulated data from Example 1 model. Left Column: Multiplier is 0.7; Middle Column: Multiplier is 0.8; Right Column: Multiplier is 0.9. Top Row: Bandwidths at both levels of double-smoothing are different; Bottom Row: Bandwidths are the same.

The lower panel of Figure 4.2 shows the pointwise local quadratic and double smoothed local quadratic bias under almost the same conditions as above, except that the bandwidths used at the second level of double-smoothing are $.7h$, $.8h$ and $.9h$, respectively.

The bandwidth used for the first level of smoothing was $h = 0.1709386$ for the local cubic regression estimates. The bandwidths used at the first level of double-smoothing were, respectively, $.7h$, $.8h$ and $.9h$. The corresponding bandwidths used at the second level of double-smoothing were all 0.0759514. The plots of the pointwise bias appear in the top panel of the figure, where the dashed curves correspond to the local cubic bias, and the solid curves correspond to the double smoothed local cubic bias.

The lower panel of Figure 4.3 shows the pointwise local cubic and double smoothed local cubic bias under almost the same conditions as above, except that the bandwidths used at the second level of double-smoothing are $.7h$, $.8h$ and $.9h$, respectively.

In both Figures 4.2 and 4.3, the patterns observed in the top panels are similar to the pattern observed in the top panel of Figure 4.1: the bias tends to increase as the first level bandwidth multiplier increases, but the improvement in bias between conventional and double smoothed estimates is less. In the bottom panels of these figures, we see that the bias is actually larger for the double smoothed estimates, and quite substantially so for the largest bandwidth multiplier. This is another case where using the same smoothing parameter for the first and second levels of smoothing is not to be recommended.

MSE Results

Figures 4.4 through 4.6 provide the pointwise MSE comparisons between the conventional and double smoothed estimates in a completely analogous way to the bias plots of the last section.

The pattern observed in the bias essentially repeats itself in the case of MSE. We see that double-smoothing with a first level bandwidth of h together with a bandwidth of $.7h$ provides the best MSE performance among all possibilities considered in the simulation study. This is true of all three polynomial degrees. The relative improvement over conventional smoothing decreases with degree, though the cubic double smoothed estimate looks to be the best overall. Again, using the same bandwidths for both levels of double-smoothing gives worse MSE performance than the corresponding conventional smoothers.

Absolute Deviation Error (ADE) Results

Figures 4.7 through 4.9 provide the pointwise absolute deviation error comparisons between the conventional and double smoothed estimates in a completely analogous way to the bias and MSE plots of the last two sections.

These plots show the same kinds of results as we saw in the MSE plots, but they have the advantage of showing the amount of improvement in accuracy in the scale of the response variable units. Thus, we see, for example, that if we use double smoothed local linear regression, we will still have about 0.3 units of error on average, at the origin (the peak), while with dou-

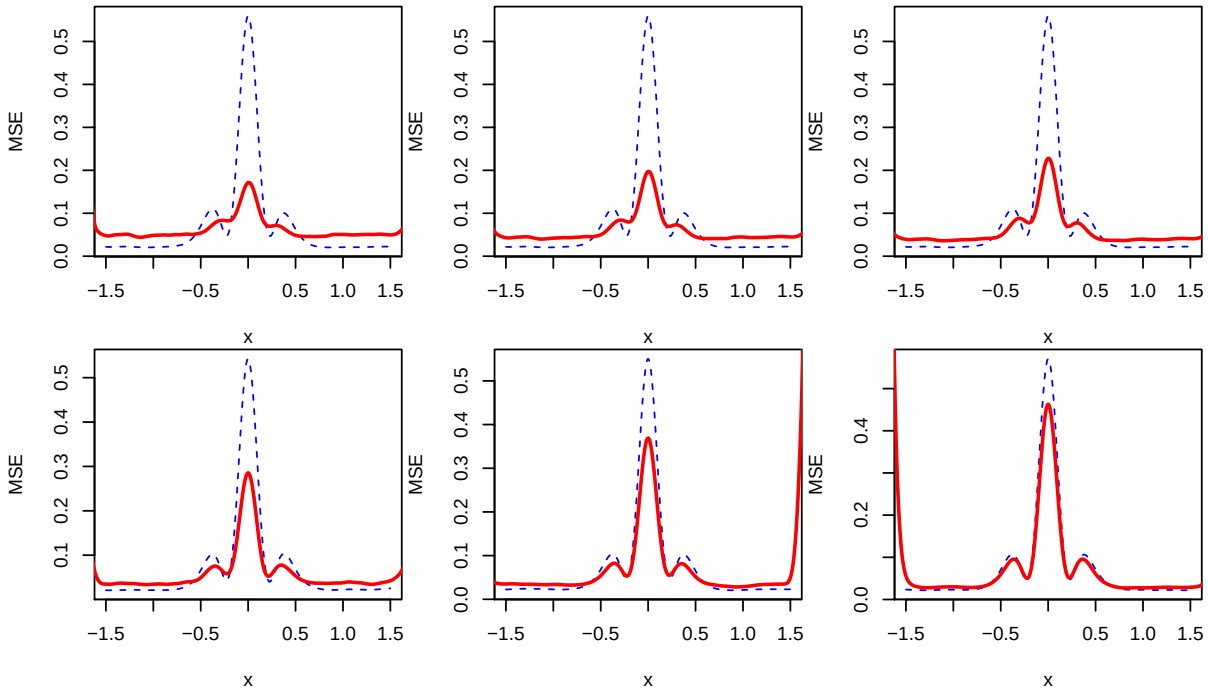

Figure 4.4: Pointwise MSE for local linear (blue dashed curve) and double-smoothed local linear (red solid curve) regression applied to simulated data from Example 1 model. Left Column: Multiplier is 0.7; Middle Column: Multiplier is 0.8; Right Column: Multiplier is 0.9. Top Row: Bandwidths at both levels of double-smoothing are different; Bottom Row: Bandwidths are the same.

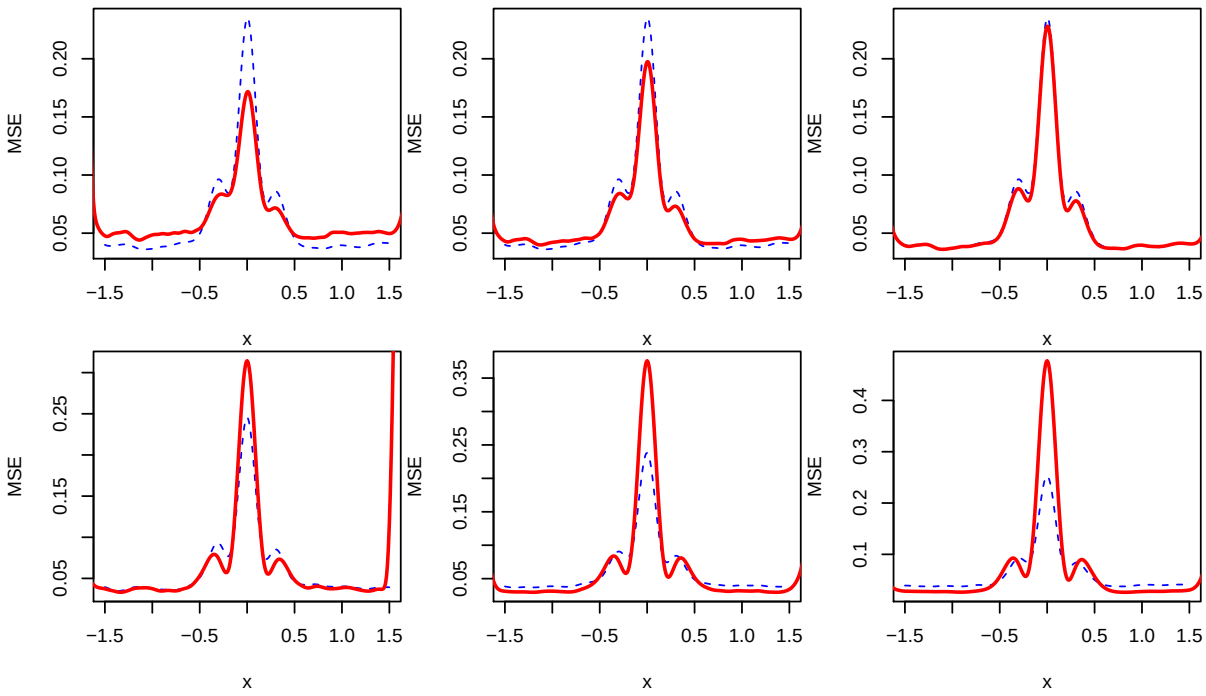

Figure 4.5: Pointwise MSE for local quadratic (blue dashed curve) and double-smoothed local quadriatic (red solid curve) regression applied to simulated data from Example 1 model. Left Column: Multiplier is 0.7; Middle Column: Multiplier is 0.8; Right Column: Multiplier is 0.9. Top Row: Bandwidths at both levels of double-smoothing are different; Bottom Row: Bandwidths are the same.

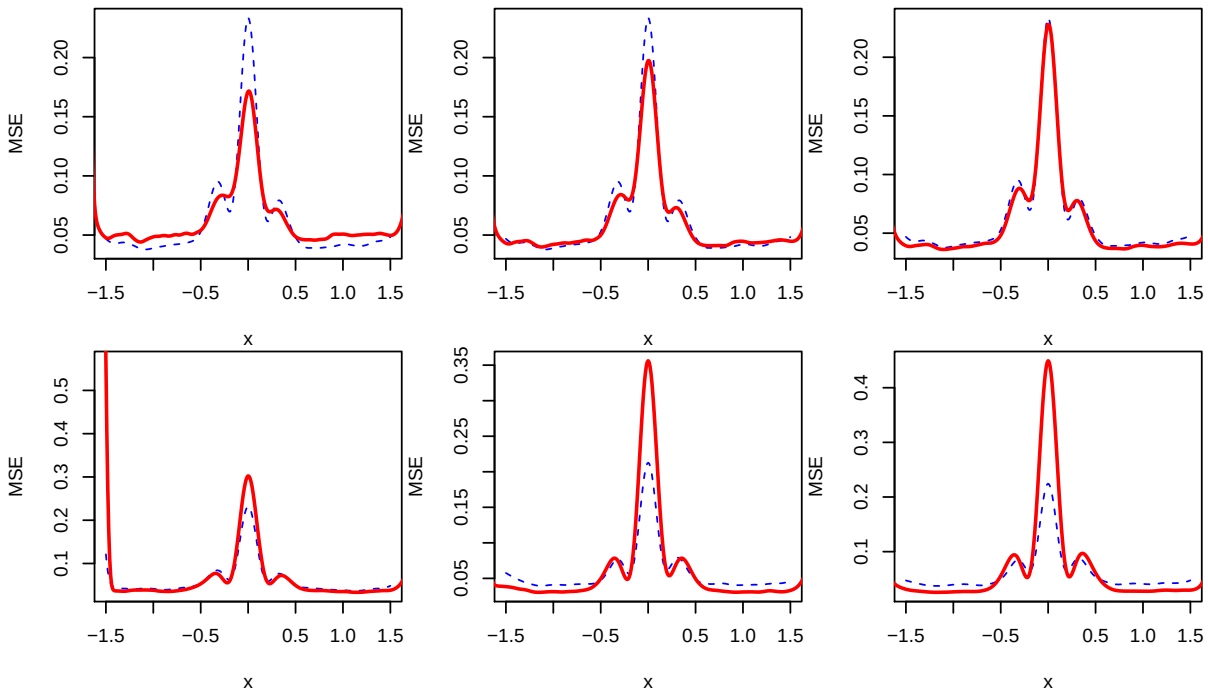

Figure 4.6: Pointwise MSE for local cubic (blue dashed curve) and double-smoothed local cubic (red solid curve) regression applied to simulated data from Example 1 model. Left Column: Multiplier is 0.7; Middle Column: Multiplier is 0.8; Right Column: Multiplier is 0.9. Top Row: Bandwidths at both levels of double-smoothing are different; Bottom Row: Bandwidths are the same.

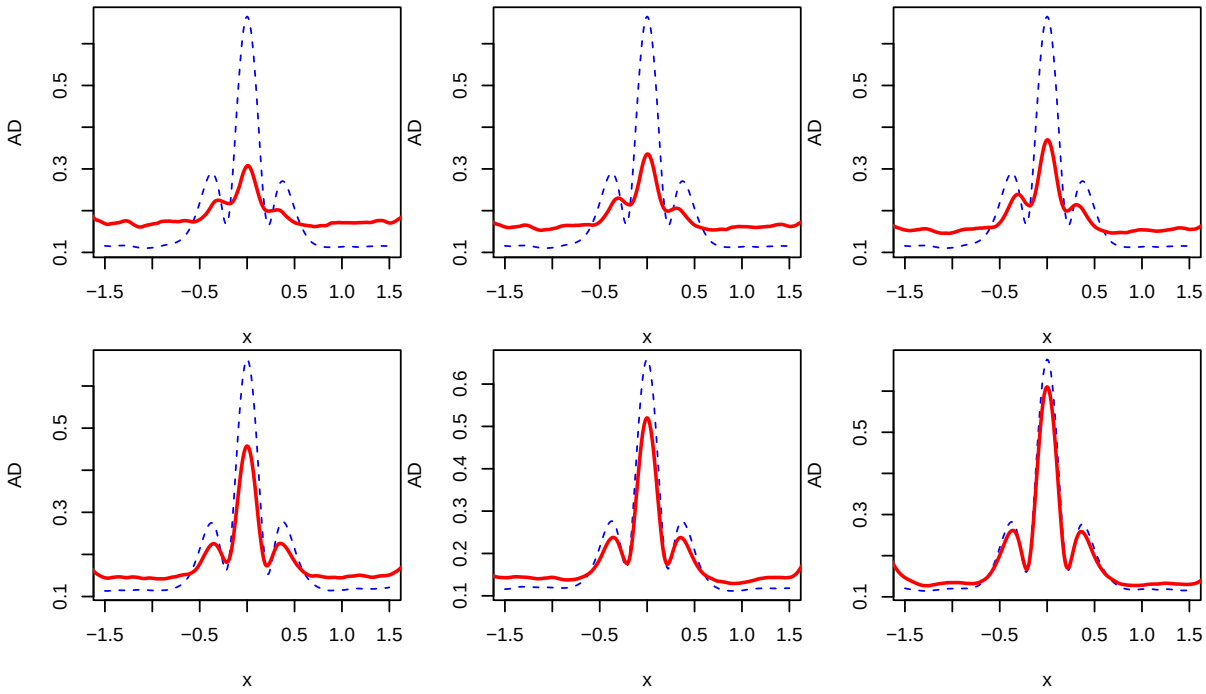

Figure 4.7: Pointwise ADE for for local linear (blue dashed curve) and double-smoothed local linear (red solid curve) regression applied to simulated data from Example 1 model. Left Column: Multiplier is 0.7; Middle Column: Multiplier is 0.8; Right Column: Multiplier is 0.9. Top Row: Bandwidths at both levels of double-smoothing are different; Bottom Row: Bandwidths are the same.

ble smoothed local cubic regression, the absolute error is about the same at the origin, but it appears to be slightly less near the boundaries.

4.4.2 Target Function 2

In our simulations of target function 2, we generated 1000 samples of size 50 from the model

$$y = m_2(x) + \varepsilon$$

where ε is independent normal noise with standard deviation 0.3.

For each sample, we computed the cross-validation bandwidths h corresponding to local linear, local quadratic and local cubic regressions, and we computed the corresponding regressions. Then we computed the double-smoothed counterparts for each polynomial degree, using

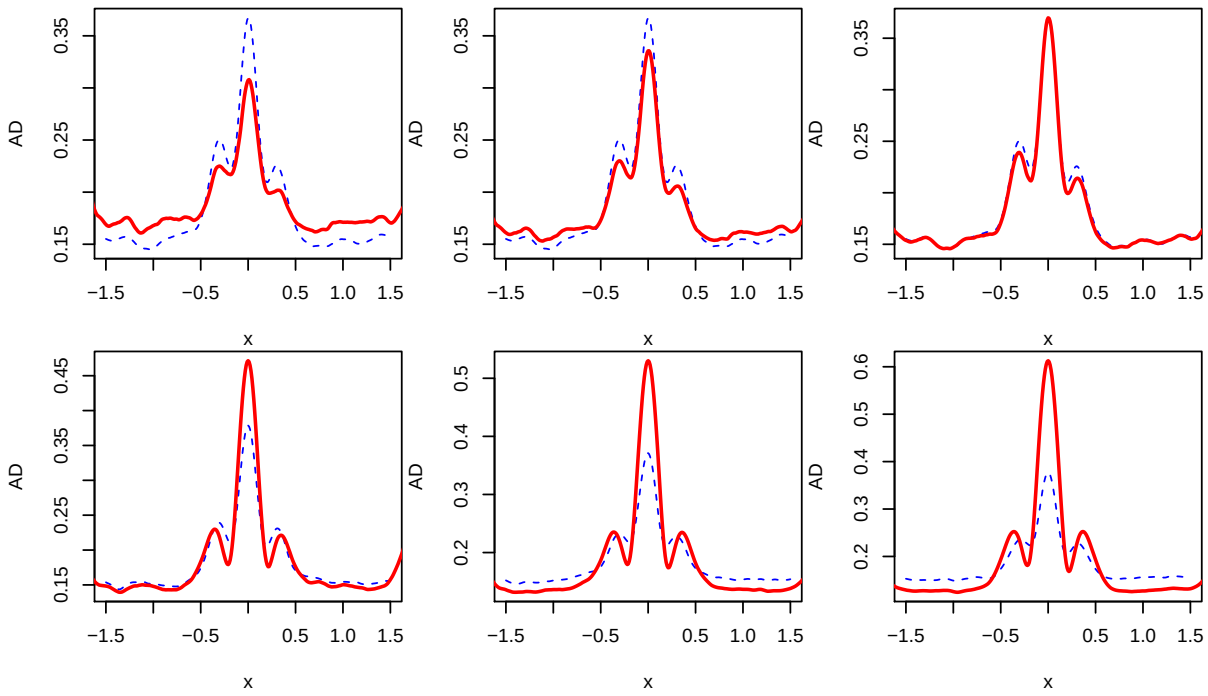

Figure 4.8: Pointwise ADE for for local quadratic (blue dashed curve) and double-smoothed local quadratic (red solid curve) regression applied to simulated data from Example 1 model. Left Column: Multiplier is 0.7; Middle Column: Multiplier is 0.8; Right Column: Multiplier is 0.9. Top Row: Bandwidths at both levels of double-smoothing are different; Bottom Row: Bandwidths are the same.

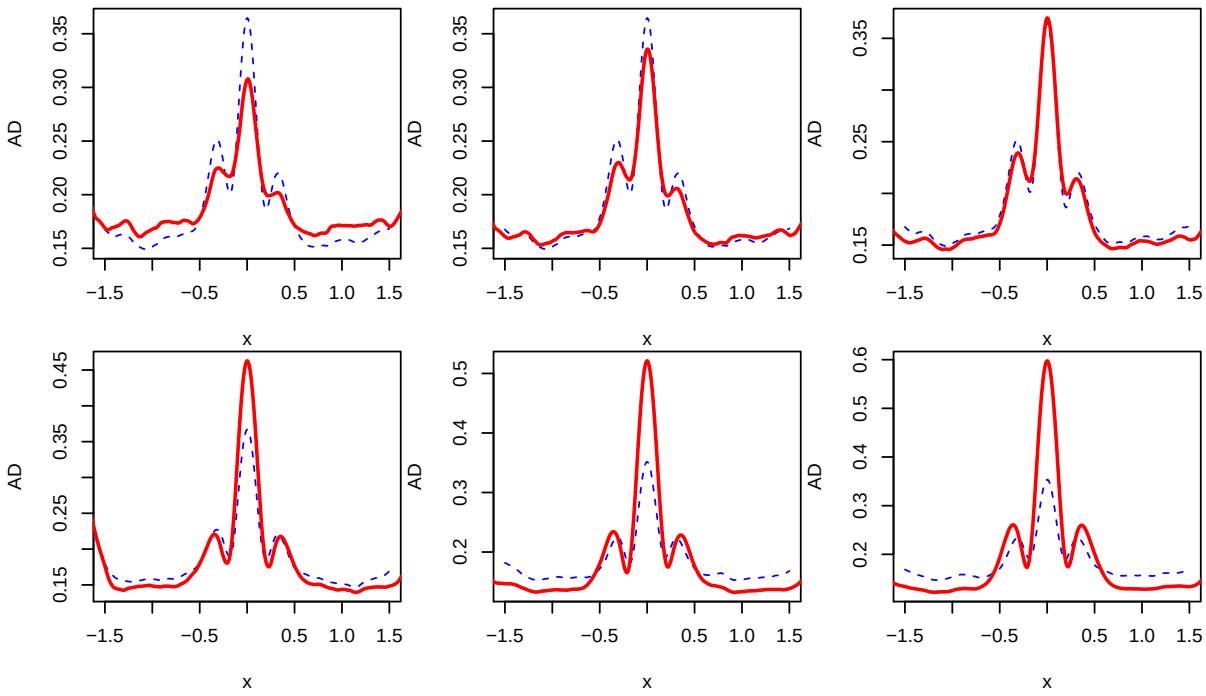

Figure 4.9: Pointwise ADE for for local cubic (blue dashed curve) and double-smoothed local cubic (red solid curve) regression applied to simulated data from Example 1 model. Left Column: Multiplier is 0.7; Middle Column: Multiplier is 0.8; Right Column: Multiplier is 0.9. Top Row: Bandwidths at both levels of double-smoothing are different; Bottom Row: Bandwidths are the same.

a multiplier of $k = .7, .8, .9$, successively, to obtain the first level smoothing parameter values. We then used the ad hoc technique to choose the second level smoothing parameter, resulting in three sets of double-smoothed estimates for each polynomial degree. We next computed an additional three sets of double-smoothed estimates (corresponding to the three values of k) by using the same bandwidth h at both levels of smoothing. Information about the selected bandwidths is provided in the next subsection.

Figures 4.10 through 4.18 show the pointwise bias, MSE and ADE plots local polynomial regression and double-smoothing for this target function. The messages to take home from these plots are similar to those that can be inferred from the plots for the first target function. double-smoothing can provide an improvement over conventional smoothing, where the improvement tends to be largest when the bandwidths are different and where the first level bandwidth is taken to be somewhat smaller than the bandwidth for the conventional estimator. As the degree increases, the amount of relative improvement over the conventional estimator tends to decrease.

Bias Results

The bandwidth used for the first level of smoothing was $h = 0.1709386$ for the local linear regression estimates. The bandwidths used at the first level of double-smoothing were, respectively, $.7h$, $.8h$ and $.9h$. The corresponding bandwidths used at the second level of double-smoothing were all 0.0759514.

The bandwidth used for the first level of smoothing was $h = 0.1709386$ for the local quadratic regression estimates. The bandwidths used at the first level of double-smoothing were, respectively, $.7h$, $.8h$ and $.9h$. The corresponding bandwidths used at the second level of double-smoothing were all 0.0759514.

The bandwidth used for the first level of smoothing was $h = 0.1709386$ for the local cubic regression estimates. The bandwidths used at the first level of double-smoothing were, respectively, $.7h$, $.8h$ and $.9h$. The corresponding bandwidths used at the second level of double-

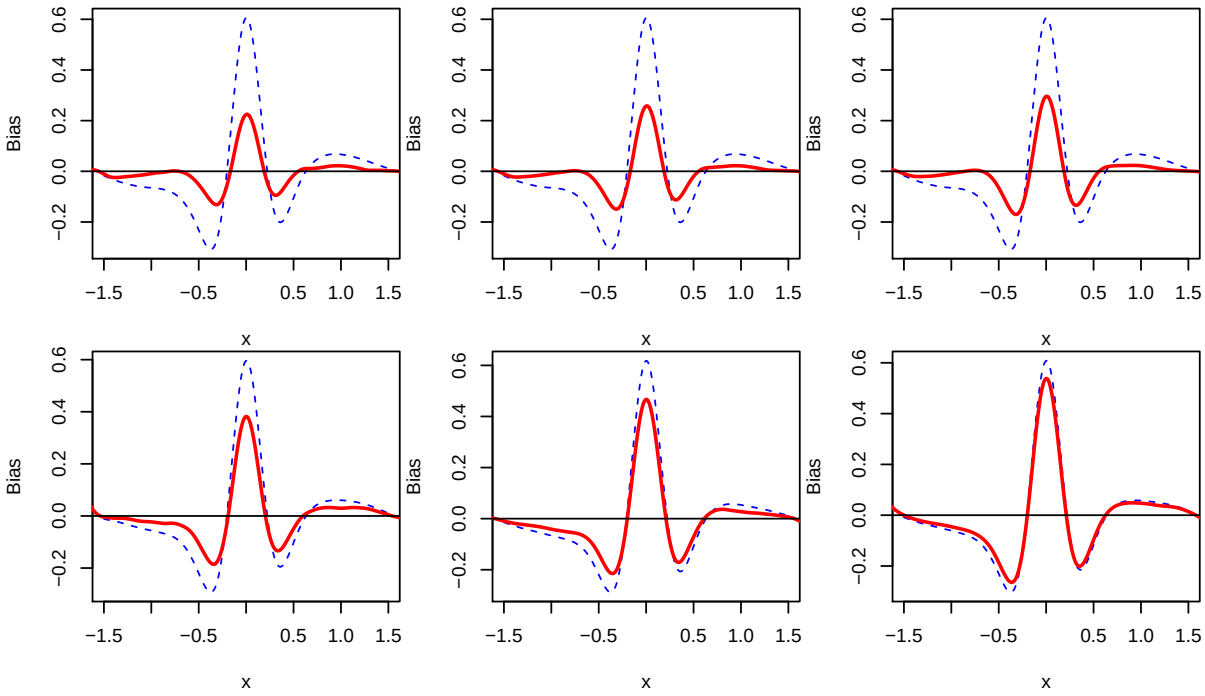

Figure 4.10: Pointwise Bias for local linear (blue dashed curve) and double-smoothed local linear (red solid curve) regression applied to simulated data from Example 2 model. Left Column: Multiplier is 0.7; Middle Column: Multiplier is 0.8; Right Column: Multiplier is 0.9. Top Row: Bandwidths at both levels of double-smoothing are different; Bottom Row: Bandwidths are the same.

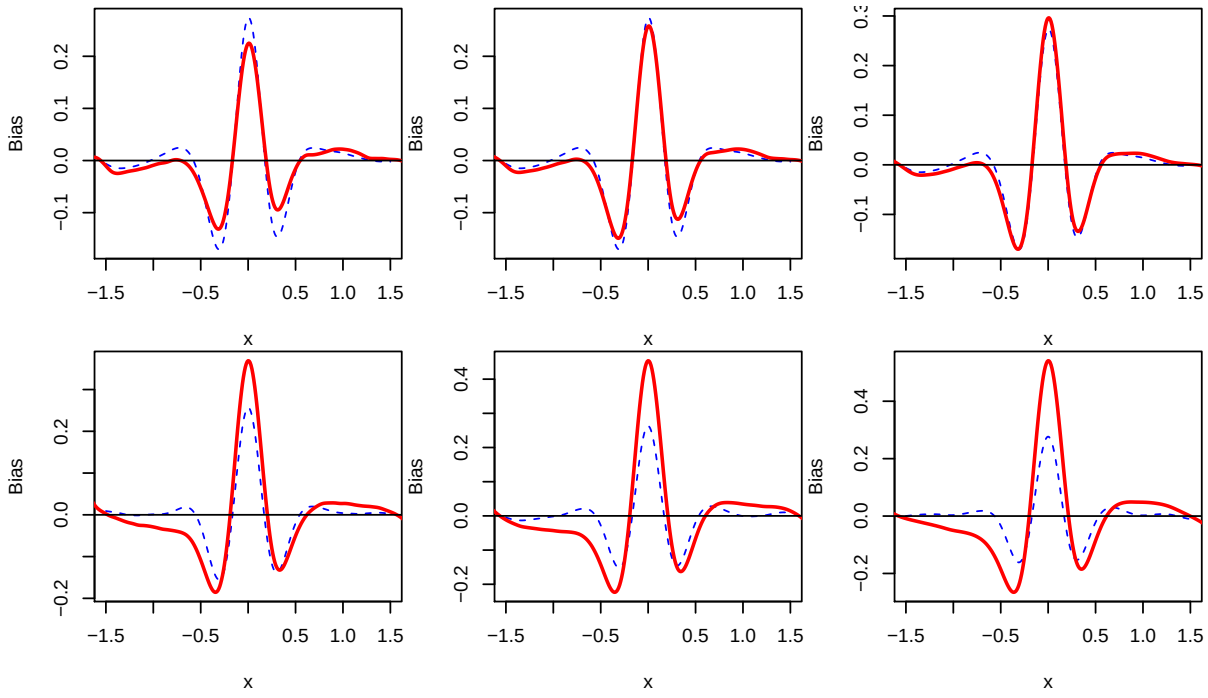

Figure 4.11: Pointwise Bias for local quadratic (blue dashed curve) and double-smoothed local quadratic (red solid curve) regression applied to simulated data from Example 2 model. Left Column: Multiplier is 0.7; Middle Column: Multiplier is 0.8; Right Column: Multiplier is 0.9. Top Row: Bandwidths at both levels of double-smoothing are different; Bottom Row: Bandwidths are the same.

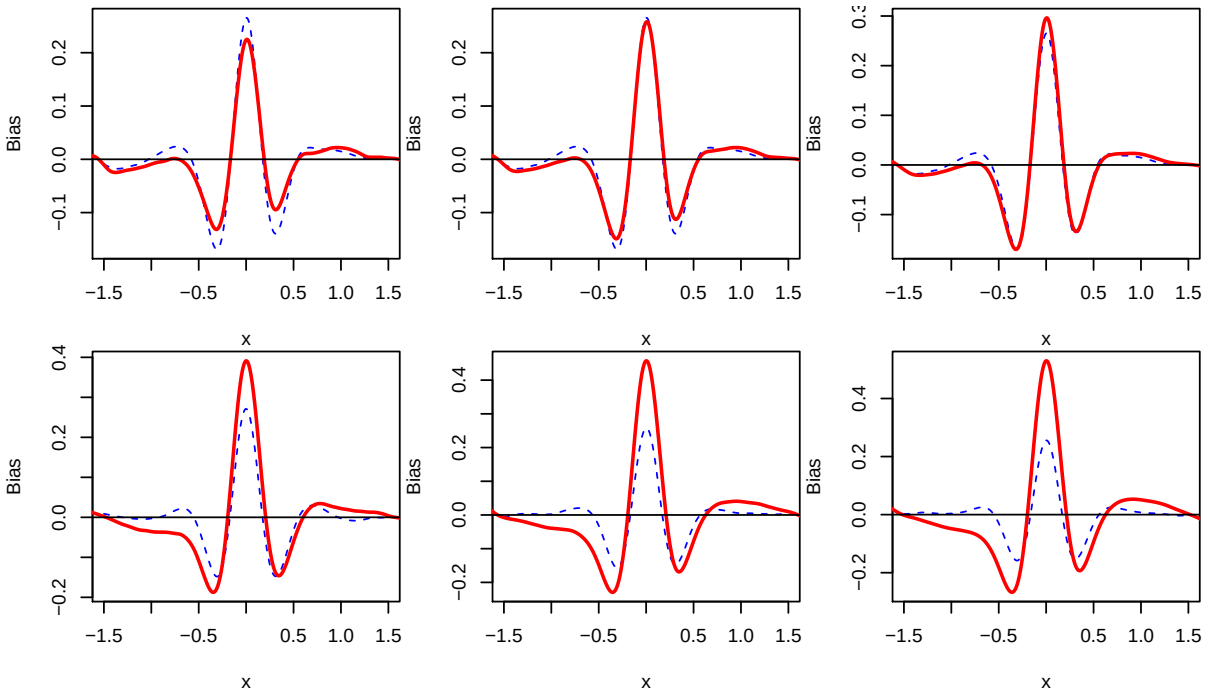

Figure 4.12: Pointwise Bias for local cubic (blue dashed curve) and double-smoothed local cubic (red solid curve) regression applied to simulated data from Example 2 model. Left Column: Multiplier is 0.7; Middle Column: Multiplier is 0.8; Right Column: Multiplier is 0.9. Top Row: Bandwidths at both levels of double-smoothing are different; Bottom Row: Bandwidths are the same.

smoothing were all 0.0759514.

MSE Results

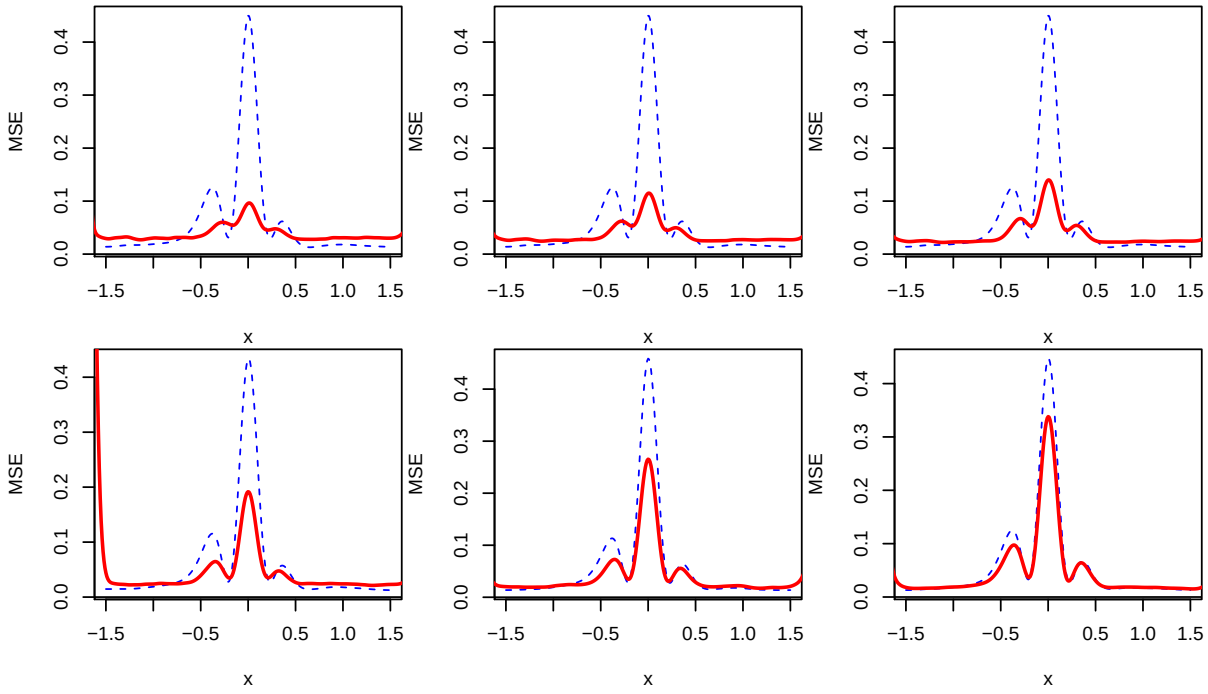

Figure 4.13: Pointwise MSE for local linear (blue dashed curve) and double-smoothed local linear (red solid curve) regression applied to simulated data from Example 2 model. Left Column: Multiplier is 0.7; Middle Column: Multiplier is 0.8; Right Column: Multiplier is 0.9. Top Row: Bandwidths at both levels of double-smoothing are different; Bottom Row: Bandwidths are the same.

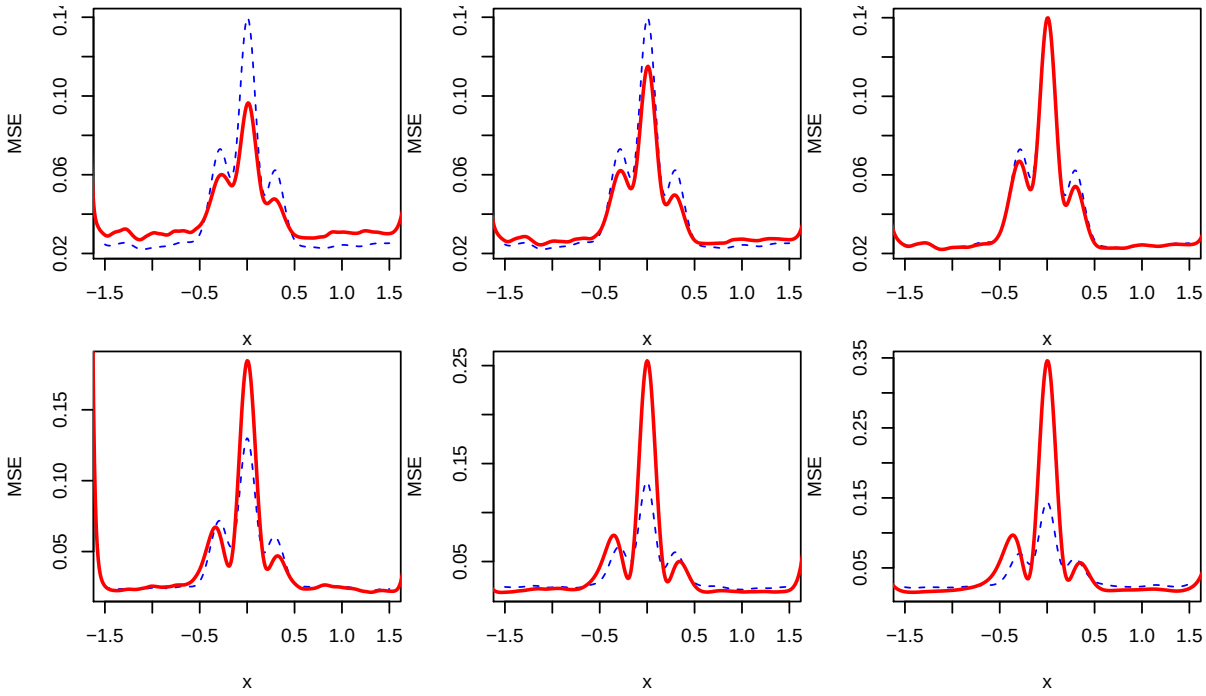

Figure 4.14: Pointwise MSE for local quadratic (blue dashed curve) and double-smoothed local quadratic (red solid curve) regression applied to simulated data from Example 2 model. Left Column: Multiplier is 0.7; Middle Column: Multiplier is 0.8; Right Column: Multiplier is 0.9. Top Row: Bandwidths at both levels of double-smoothing are different; Bottom Row: Bandwidths are the same.

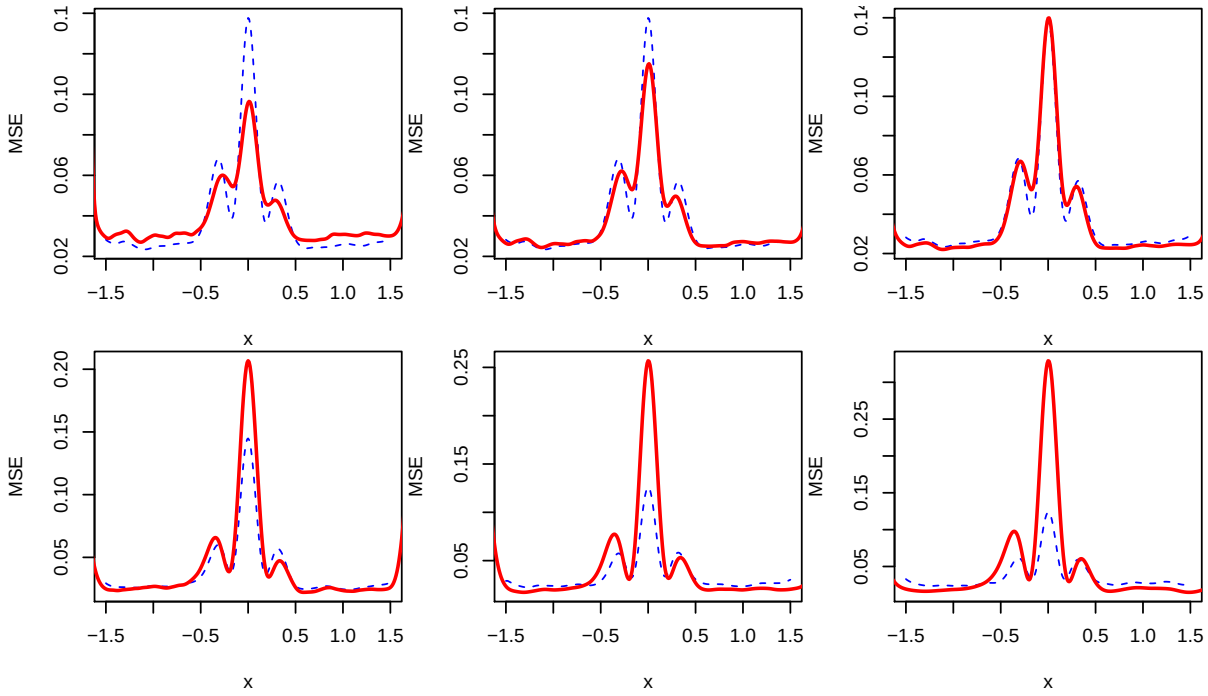

Figure 4.15: Pointwise MSE for local cubic (blue dashed curve) and double-smoothed local cubic (red solid curve) regression applied to simulated data from Example 2 model. Left Column: Multiplier is 0.7; Middle Column: Multiplier is 0.8; Right Column: Multiplier is 0.9. Top Row: Bandwidths at both levels of double-smoothing are different; Bottom Row: Bandwidths are the same.

Absolute Deviation Error (ADE) Results

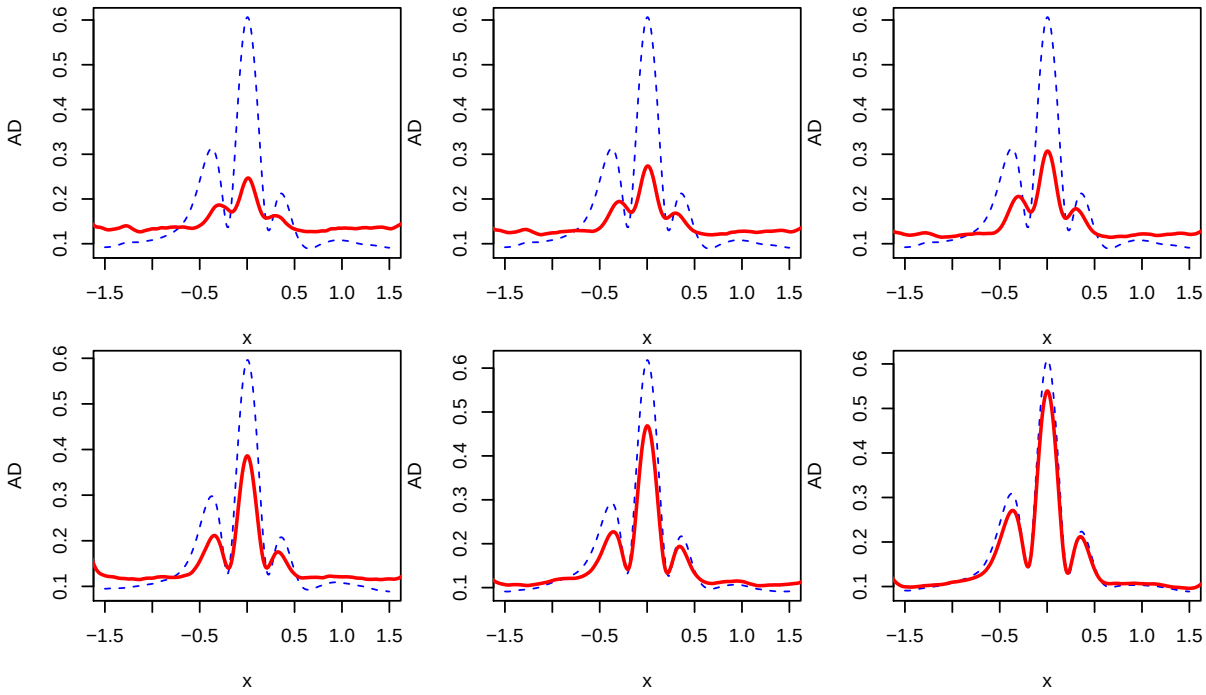

Figure 4.16: Pointwise ADE for for local linear (blue dashed curve) and double-smoothed local linear (red solid curve) regression applied to simulated data from Example 2 model. Left Column: Multiplier is 0.7; Middle Column: Multiplier is 0.8; Right Column: Multiplier is 0.9. Top Row: Bandwidths at both levels of double-smoothing are different; Bottom Row: Bandwidths are the same.

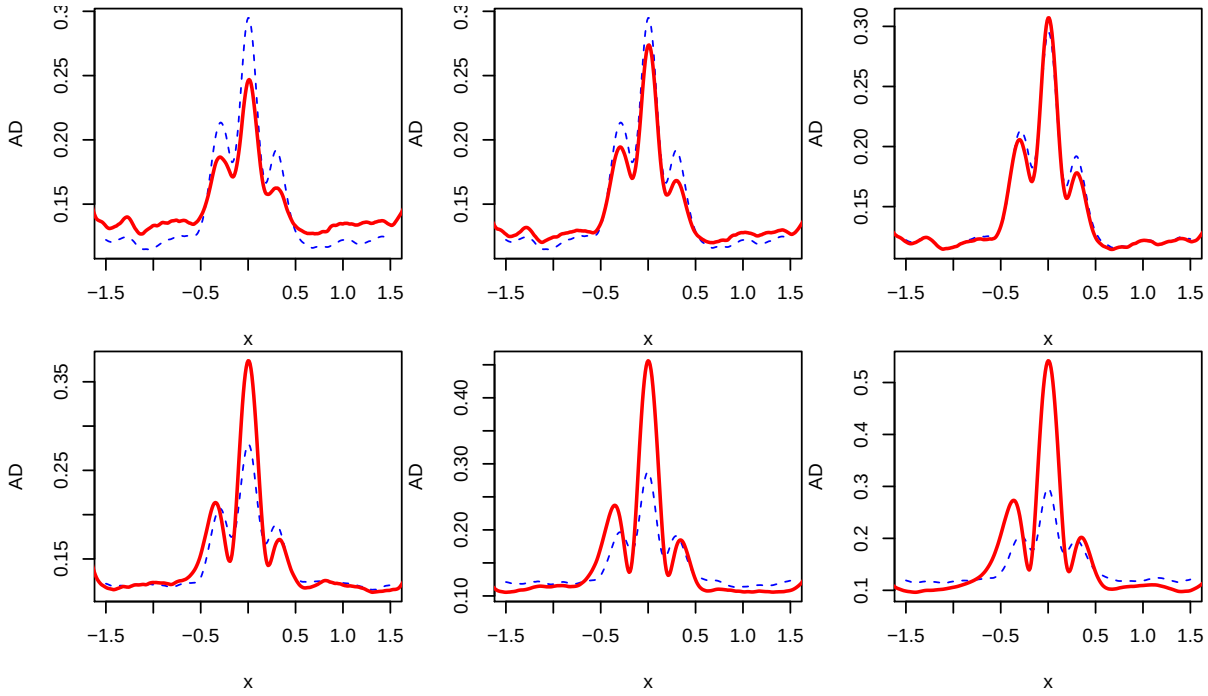

Figure 4.17: Pointwise ADE for for local quadratic (blue dashed curve) and double-smoothed local quadratic (red solid curve) regression applied to simulated data from Example 2 model. Left Column: Multiplier is 0.7; Middle Column: Multiplier is 0.8; Right Column: Multiplier is 0.9. Top Row: Bandwidths at both levels of double-smoothing are different; Bottom Row: Bandwidths are the same.

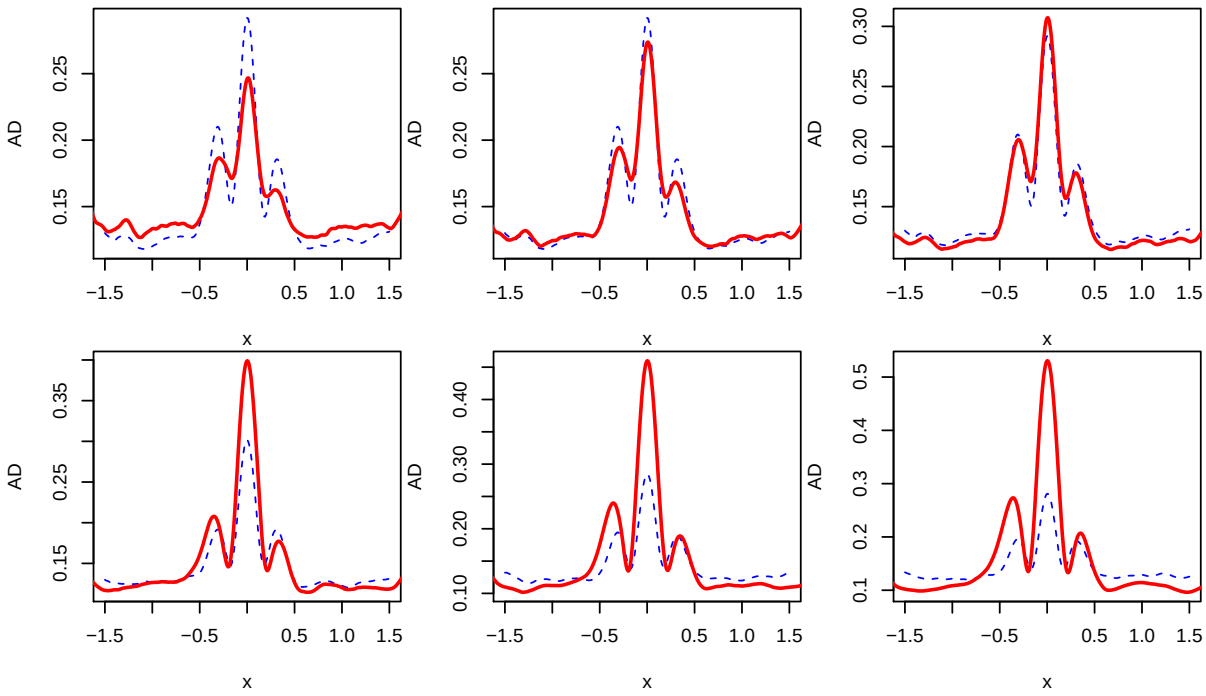

Figure 4.18: Pointwise ADE for for local cubic (blue dashed curve) and double-smoothed local cubic (red solid curve) regression applied to simulated data from Example 2 model. Left Column: Multiplier is 0.7; Middle Column: Multiplier is 0.8; Right Column: Multiplier is 0.9. Top Row: Bandwidths at both levels of double-smoothing are different; Bottom Row: Bandwidths are the same.

4.4.3 Additional Observations

The bias, MSE and AD comparisons were made in the interior of the range, avoiding boundary effects - a subject that is beyond the scope of our investigations here. However, we should make some comments about what was observed in the simulations. In the local linear simulations, we tended to see reasonable boundary behaviour in both the local linear and double smoothed estimates. In the case of local quadratic, there were often instances where we observed poor boundary behaviour in the conventional estimate but much improved behaviour in the double smoothed estimate. Similarly, in the case of local cubic regression, there were cases where the boundary behaviour was quite poor, while the double smoothed version overcame any such difficulties.

This points out an additional benefit of the double smoothed approach, at degree 1 or higher: data sparsity conditions caused by proximity to the boundary for example can be mitigated by the double-smoothing technique. This could render higher order local polynomial regression (or more accurately, their double smoothed versions) as more practically useful.

4.5 Illustrative Examples

In the next three subsections, we consider three standard examples: the Old Faithful data set, the beluga whale nursing data set and the ethanol data set.

4.5.1 Old Faithful Data Set

For the Old Faithful data set, we applied the local polynomial regression and double smoothed local polynomial regression for degrees 1, 2 and 3. For local linear regression, we used a bandwidth of $h = 3.0701754$. For local quadratic regression, we used a bandwidth of $h = 5.122807$. For local cubic regression, we used a bandwidth of $h = 5.8070175$. We used a bandwidth multiplier of $k = .7$ to obtain the first level of double-smoothing in each case, and we used a second level double-smoothing bandwidth of 0.195572.

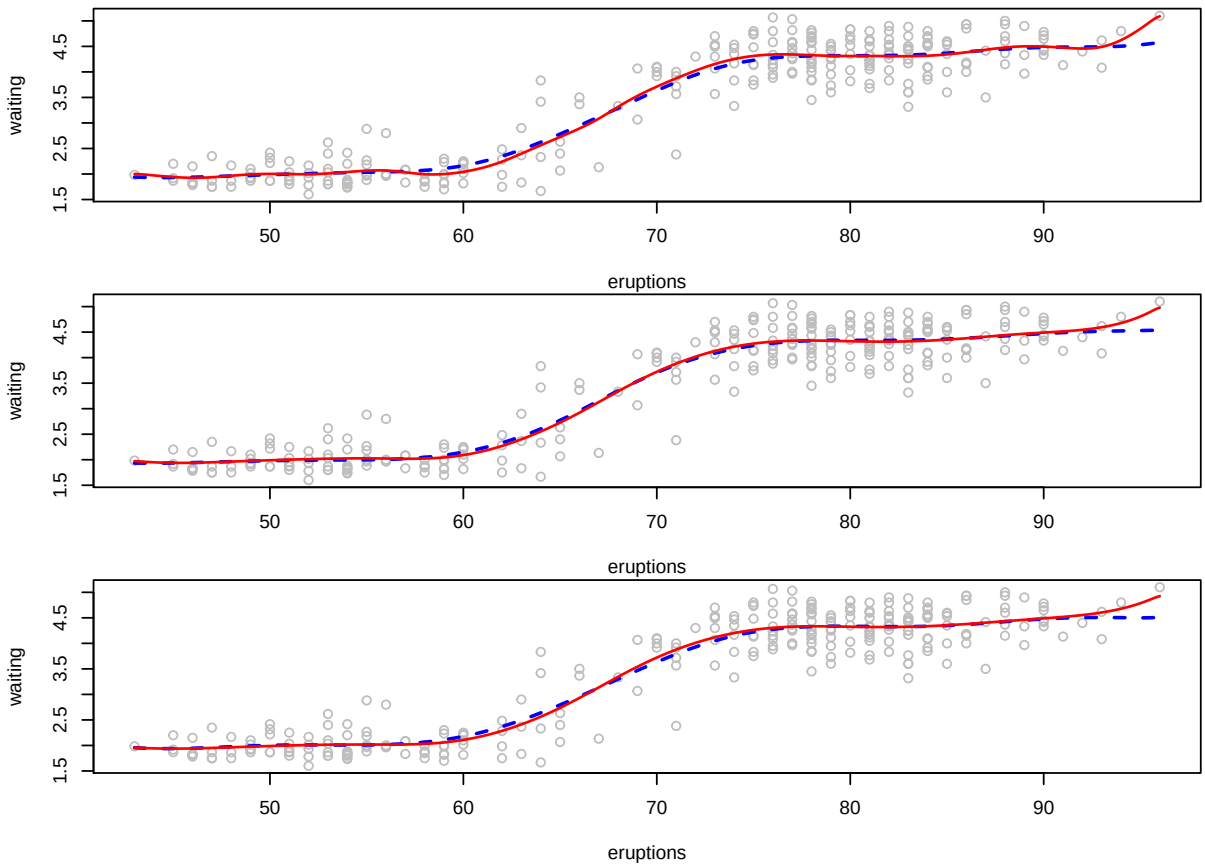

Figure 4.19: Old Faithful waiting times and corresponding eruption durations with overlaid local polynomial and double-smoothed local polynomial regression curves, using a bandwidth multiplier of 0.7 in the double-smoothed regression. The double-smoothed curve is solid red and the local polynomial curve is dashed blue. Upper Panel: local linear; Middle Panel: local quadratic; Lower Panel: local cubic.

Figure 4.19 shows the scatterplot overlaid with the local polynomial regression estimators (dashed curves) and double smoothed regression estimators (solid curves) for the three degrees. As expected, based on the simulation results, we see the biggest difference when comparing local linear with double smoothed local linear. Based on the simulation results, we might trust the double smoothed regression function estimate in the range from 60 through 80 somewhat more than the local polynomial estimate. Boundary behaviour is different in the local linear and local quadratic cases; again, we might trust the double smoothed estimates somewhat more; all three double smoothed estimates say essentially the same thing at the rightmost boundary of the data set.

4.5.2 Beluga Whale Nursing Data Set

For the beluga whale nursing data set, we applied the local polynomial regression and double smoothed local polynomial regression for degrees 1, 2 and 3. For local linear regression, we used a bandwidth of $h = 3.245614$. For local quadratic regression, we used a bandwidth of $h = 6.3245614$. For local cubic regression, we used a bandwidth of $h = 5.2982456$. We used a bandwidth multiplier of $k = .7$ to obtain the first level of double-smoothing in each case, and we used a second level double-smoothing bandwidth of 1.

Figure 4.20 shows the scatterplot overlaid with the local polynomial regression estimators (dashed curves) and double smoothed regression estimators (solid curves) for the three degrees. Again, we see the biggest difference when comparing local linear with double smoothed local linear. In fact, there is a lot of oscillation in both the double-smoothed and conventional local linear estimates, likely due to a bandwidth that has been chosen to be too small. The double smoothed local quadratic and cubic cases are smoother, and are possibly more believable than the local linear counterparts.

The peak at 35 is more pronounced in all of the double smoothed estimates with the sharpest peak observed in the double smoothed local linear and double smoothed local cubic. Behaviour

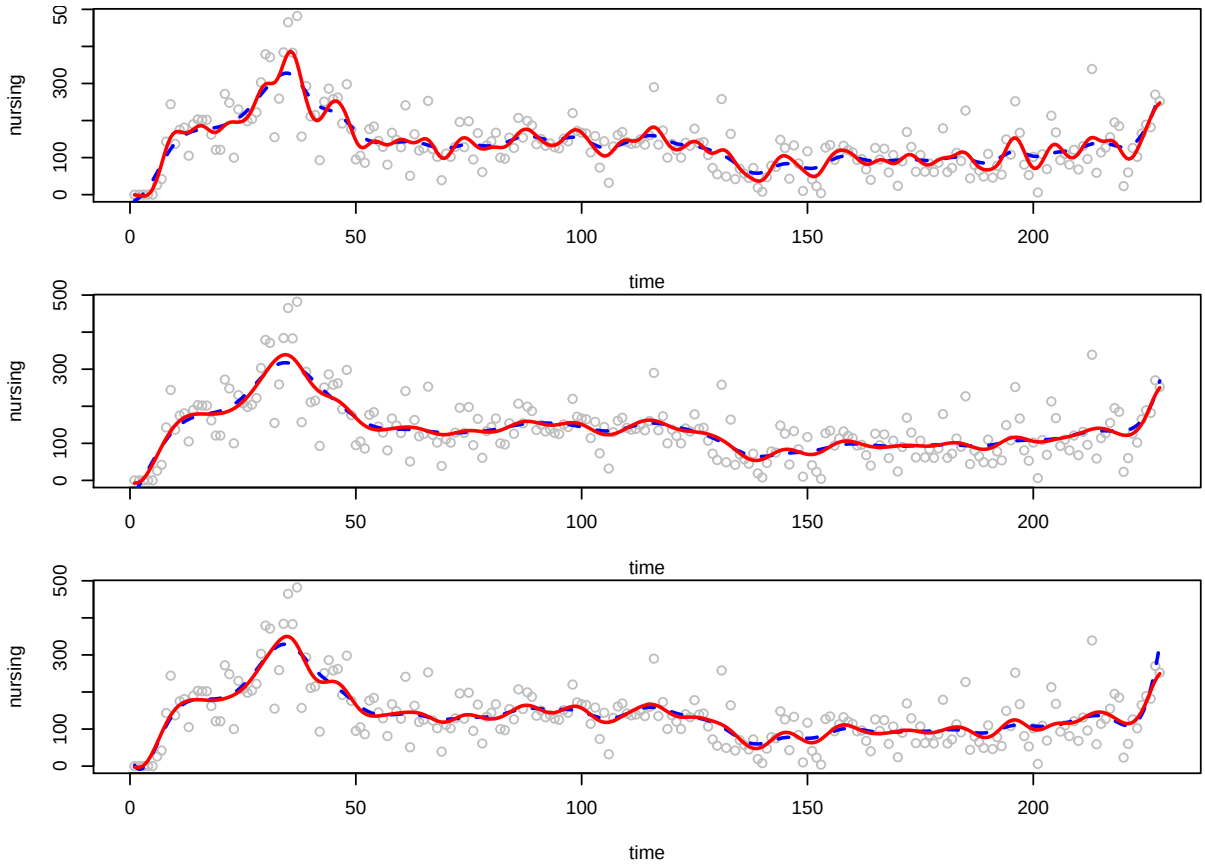

Figure 4.20: Beluga whale nursing duration versus elapsed time with overlaid local polynomial and double-smoothed local polynomial regression curves, using a bandwidth multiplier of 0.7 in the double-smoothed regression. The double-smoothed curve is solid red and the local polynomial curve is dashed blue. Upper Panel: local linear; Middle Panel: local quadratic; Lower Panel: local cubic.

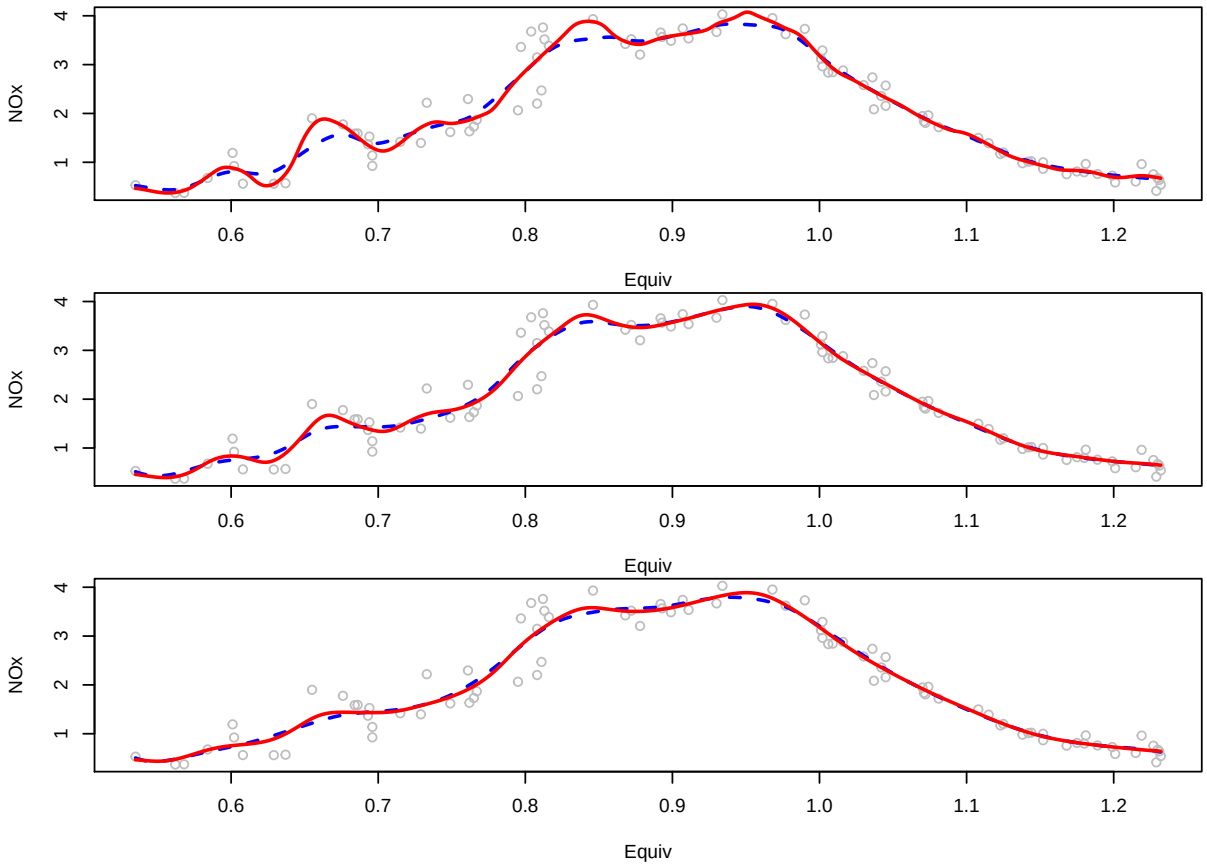

Figure 4.21: Ethanol data with overlaid local polynomial and double-smoothed local polynomial regression curves, using a bandwidth multiplier of 0.7 in the double-smoothed regression. The double-smoothed curve is solid red and the local polynomial curve is dashed blue. Upper Panel: local linear; Middle Panel: local quadratic; Lower Panel: local cubic.

at the rightmost boundary is somewhat different for all three cases; again, we might trust the double smoothed estimates somewhat more; all three double smoothed estimates say essentially the same thing at the rightmost boundary of the data set.

4.5.3 Ethanol Data Set

For the ethanol data set, we applied the local polynomial regression and double smoothed local polynomial regression for degrees 1, 2 and 3. For local linear regression, we used a bandwidth of $h = 0.0172982$. For local quadratic regression, we used a bandwidth of $h = 0.0289298$. For local cubic regression, we used a bandwidth of $h = 0.0405614$. We used a bandwidth multiplier

of $k = .7$ to obtain the first level of double-smoothing in each case, and we used second level double-smoothing bandwidths of $0.0172982, 0.0289298, 0.0405614$ for each respective degree.

Figure 4.21 shows the scatterplot overlaid with the local polynomial regression estimators (dashed curves) and double smoothed regression estimators (solid curves) for the three degrees. Again, we see the biggest difference when comparing local linear with double smoothed local linear. Again, the bandwidths might have been selected to be too small here. For the local quadratic and cubic cases, the curves are smoother, with less of an apparent attempt to model the noise. The double smoothed local cubic, in particular, seems to have few spurious wiggles while possibly capturing the main features of the data.

4.6 Summary

We have found that double-smoothing can be extended to higher order local polynomial regression, and the technique can offer an improvement in terms of bias and MSE. The improvement in accuracy is somewhat less dramatic than that obtained when applying double-smoothing to local linear regression. In all cases, bandwidth choice is critical. It is possible for double-smoothing to cause of a loss of accuracy. Using the same bandwidth for both levels of double-smoothing can also lead to a loss of accuracy. We have found that a simple method for choosing the second level bandwidth based on the average of the successive differences in the predictor values works reasonably well. We used a cross validation method for the first level bandwidth. This could probably be improved upon.

Chapter 5

Application to the Initial Attack Problem

In this chapter, we apply our techniques to data on initial attack response times introduced in Section 2.2 to demonstrate the usefulness of the techniques.

We briefly outline the reason why this data set is interesting in fire management in Section 5.1. Then we illustrate our bias assessment tool as it applies to the median initial attack response time data set. Next we apply the double-smoothing bias reduced kernel smoother to the data set. Finally, in Section 5.4, the process of data cleaning with regarding to all possible event time stamps is described as well as issues that arose during this process.

5.1 Objectives and Data Visualization

From a wildland fire management perspective, one objective behind studying this data set is to determine whether there was a significant change in response time at 1950. The response time is the time between the report time to the initial fire fighting (initial attack) time.

Based on Figure 2.4 in Chapter 1, we know that although not all fires followed the standard event log sequences, almost all of the reporting time to initial attack time of the fire cases indeed followed that sequence.

There are several ways to visualize the time differences between these two events (reporting and beginning of initial attack). An elementary view of the event duration time can be obtained

using the business process management model (Janssenswillen, 2018). In Figure 5.1, the grey dots represent recorded reporting time and the green dots represent the time that initial attack began. The fire cases were partitioned into two categories: before 1950 and after. On the vertical axis in Figure 5.1, the fire cases are sorted by initial attack time which is the duration from reporting time to beginning of initial attack. We only considered initial attack response times up to 20 hours for each year group, removing several cases that took unrealistically large amounts of time.

The original level and log transformation of the initial attack time were both checked and they do not show normality. Thus, we chose to use the median of initial attack time grouped by each year to continue the study due to the robustness of the median to skewness and to outliers.

Figure.5.2 shows the plots for the median (in hours) of each year together with a very rudimentary smoothing (linear fit) overlaid on it.

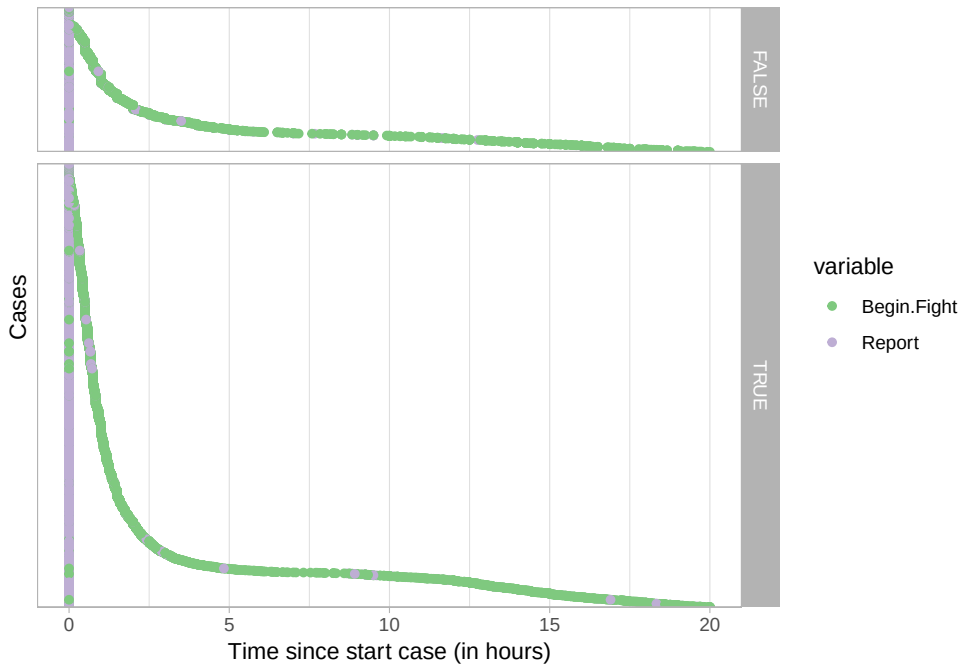

Figure 5.1: Dotted plot for the initial attack time of fire cases, sorted by the length of the initial attack time (neglecting those times beyond 20 hours). Grey dots are for the reporting time, and green dots are for the beginning of initial attack. Top Panel: fire cases before and including the year 1950; Bottom Panel: fire cases after 1950.

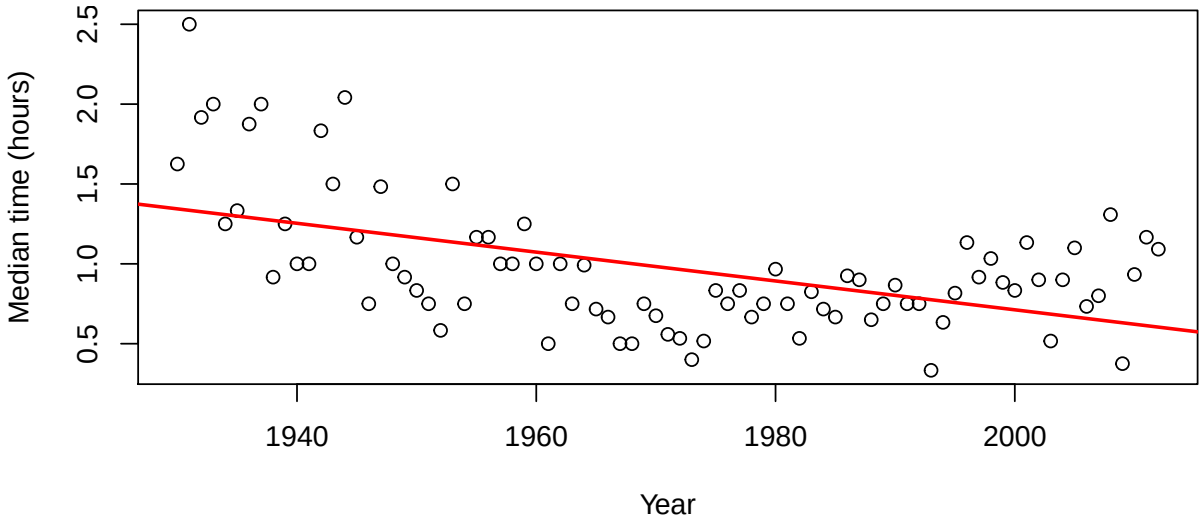

Figure 5.2: Median of initial attack time for each year with linear smoothing overlaid

5.2 Kernel Smoothing and Bias Assessment

In Chapter 3, we proposed a new method of visualizing the bias for local polynomial regression. We now utilize it in the initial attack problem. The result of applying local linear regression to the median initial attack response times, using a bandwidth of 5 and then of 10 is displayed in Figure 5.3 and Figure 5.4, respectively. Corresponding bias assessment plots are shown in the lower panels of each figure.

The top panels of Figures 5.3 and 5.4 show the local linear regression estimate, together with pointwise 95% confidence bounds for the true regression function, assuming no bias. The bottom panels of the figures show the 95% pointwise confidence bounds for the bias. It is clear from this plot that there is significant bias in a number of regions, since the horizontal axis falls outside the confidence bounds at times. The assessment plots for the bias in both scenarios are in relative agreement with each other, with years around 1952, 1970 and 1992 having relatively larger bias than other years. The magnitude of the bias is clearly larger when the larger bandwidth is used.

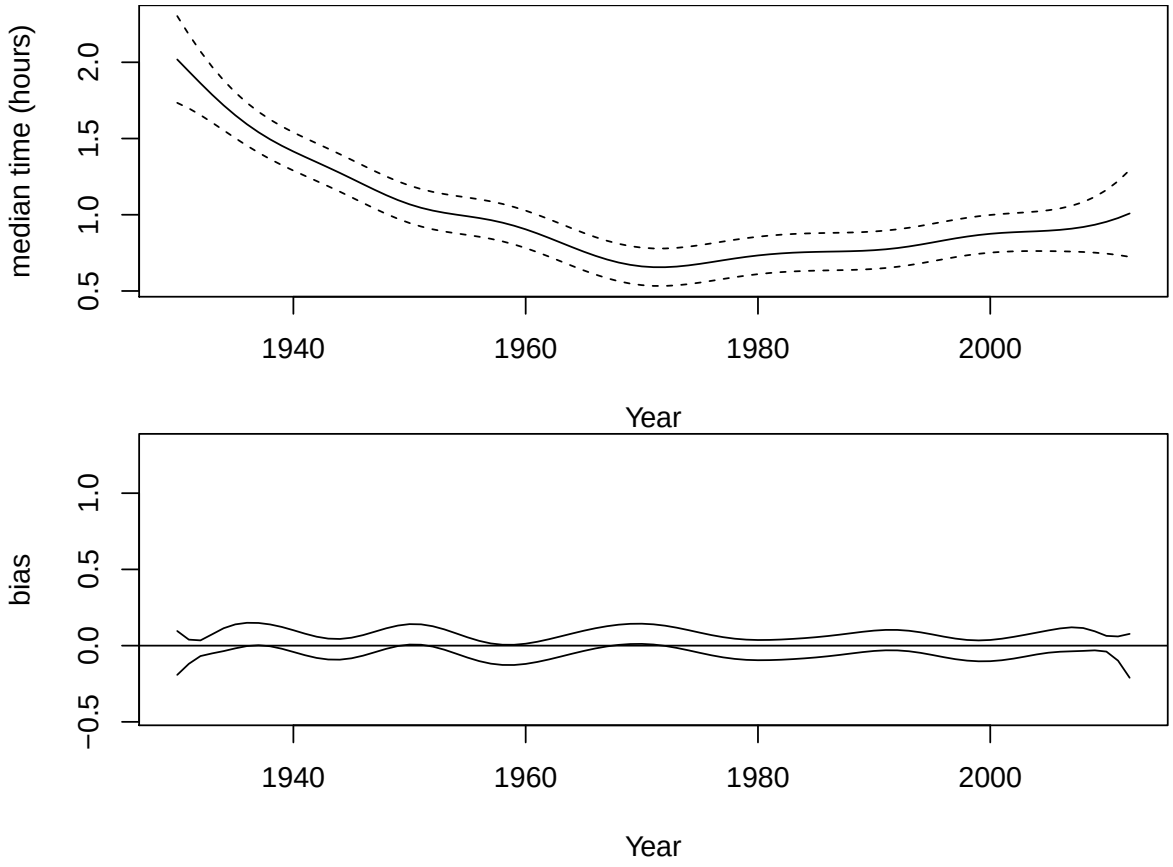

Figure 5.3: Top Panel: Local linear fitting using bandwidth 5 and degree to be 1. Bottom Panel: Bias assessment plot with same bandwidth, degree to be 2.

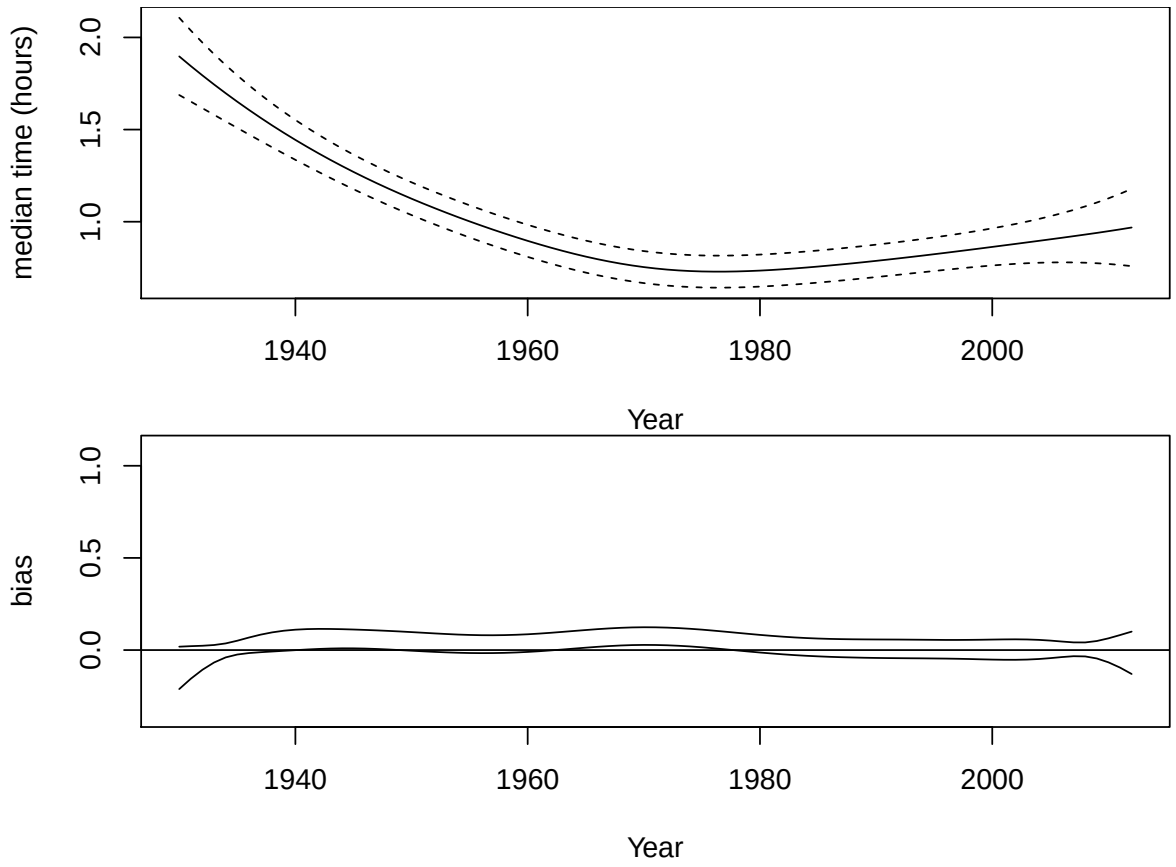

Figure 5.4: Top Panel: Local linear fitting using bandwidth 10 and degree to be 1. Bottom Panel: Bias assessment plot with same bandwidth, degree to be 2.

5.3 Application of Double-Smoothing

In Chapter 4, we reviewed double-smoothing local linear methods based on He and Huang (2009), and extended the double-smoothing procedure to local quadratic and local cubic regression. We now apply these techniques to the fire record data records for our analysis of initial attack time.

We first applied double-smoothing using the cross-validation bandwidth for the local polynomial regressions of degrees 1, 2 and 3, with a multiplier $k = 0.7$ for the first level of smoothing, and using our quick and simple bandwidth for the second level of smoothing. Since the data are equally spaced in this application, the second level bandwidth was simply the length

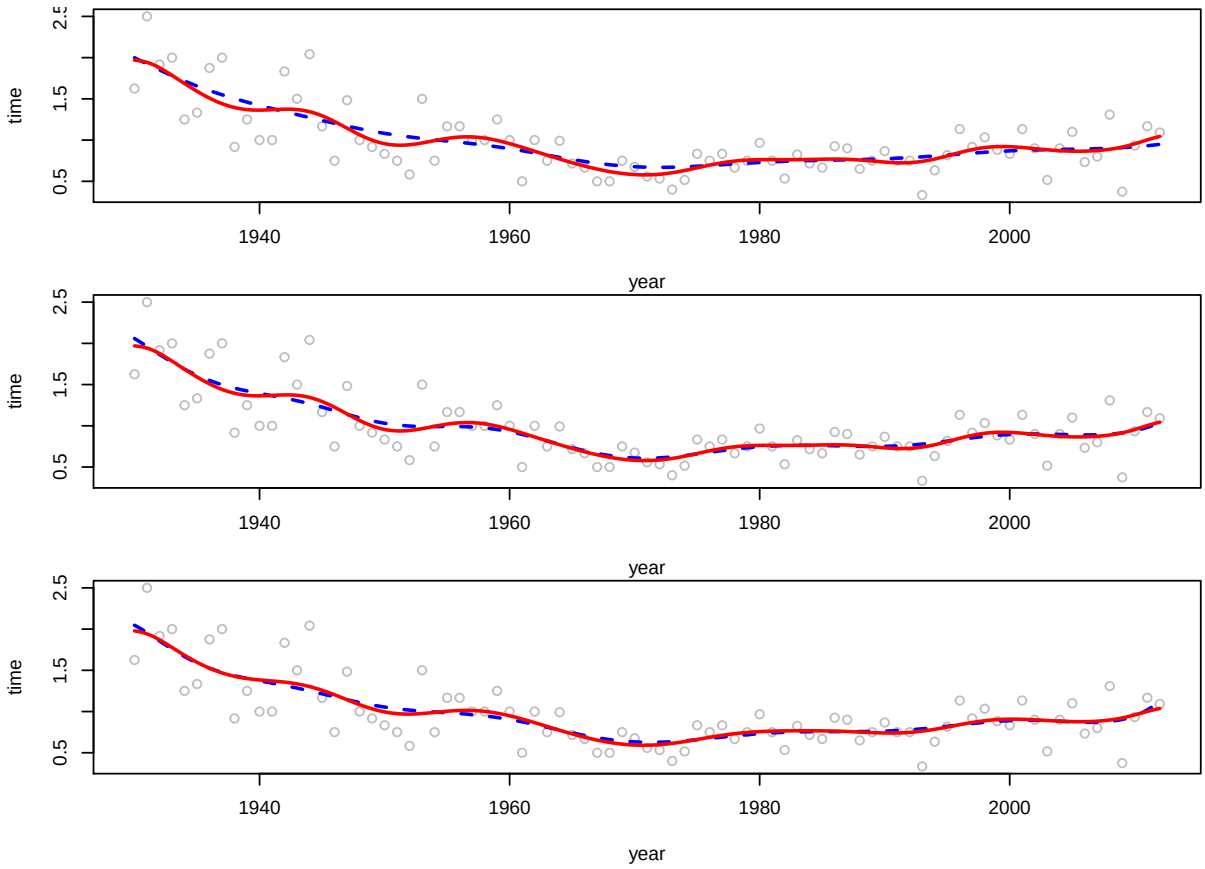

Figure 5.5: Initial attack data with overlaid local polynomial and double-smoothed local polynomial regression curves, using a bandwidth multiplier of 0.7 in the double-smoothed regression. The double-smoothed curve is solid red and the local linear curve is dashed blue. Top Panel: local linear; Middle Panel: local quadratic; Bottom Panel: local cubic.

of the interval between observations, namely, $h_2 = 1$.

Bandwidths for the the local polynomial regressions were $5.64, 5.64, 6.667$ for degrees 1, 2 and 3, respectively. The first level bandwidths in the double-smoothing algorithm were $3.948, 3.948, 4.667$, respectively.

What we observe in Figure 5.5 is that the double-smoothed local linear regression provides a much different curve from local linear regression, as compared with the differences between the higher order double-smoothed curves and their local polynomial counterparts. The biggest differences in the local linear case correspond to where we saw the largest bias issues earlier.

We also experimented with some other possibilities. In particular, we reduced the pilot bandwidths and the corresponding first level bandwidths in double-smoothing in order to ob-

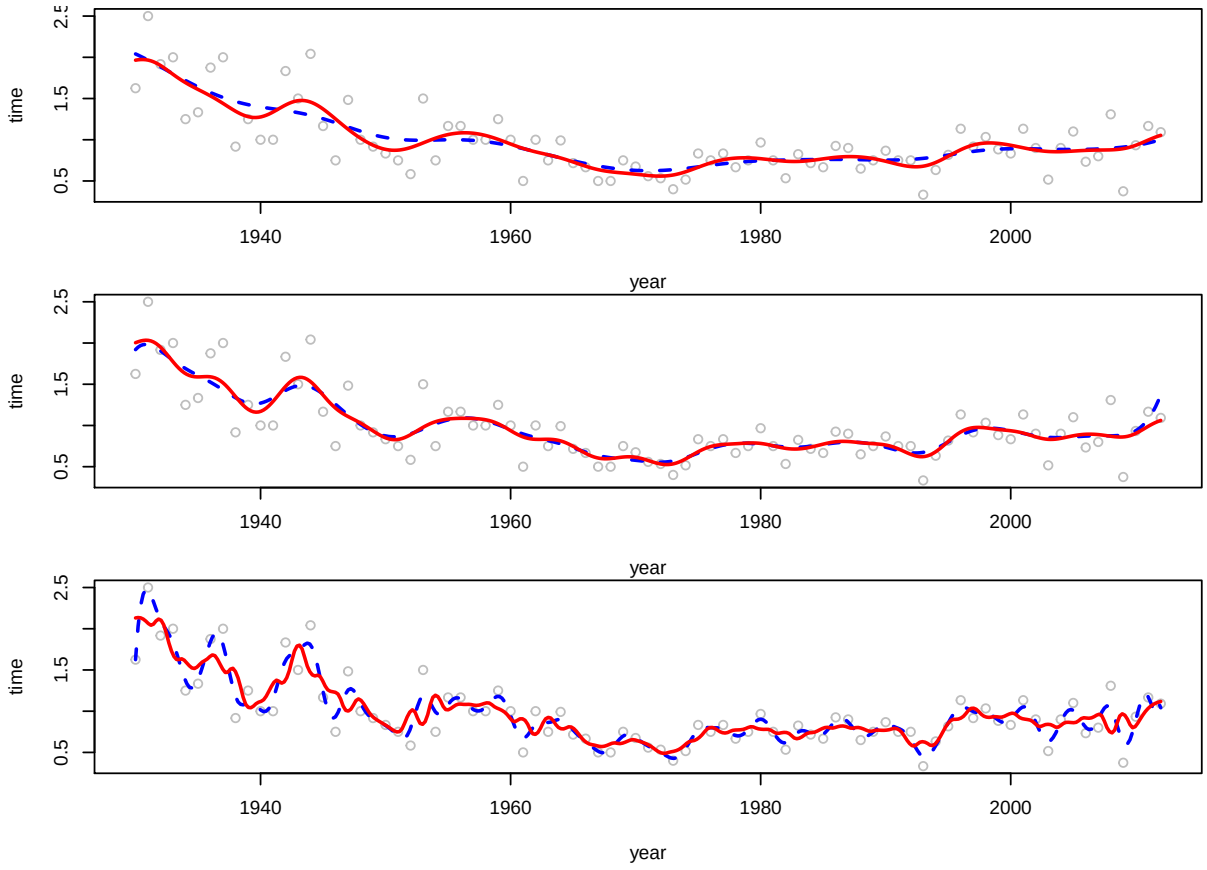

Figure 5.6: Initial attack data with overlaid local polynomial and double-smoothed local polynomial regression curves, using smaller first level bandwidths in the double-smoothed regression for respective degrees 1, 2, and 3. The double-smoothed curve is solid red and the local linear curve is dashed blue. Top Panel: local linear; Middle Panel: local quadratic; Bottom Panel: local cubic.

tain the results plotted in Figure 5.6. Bandwidths for the the local polynomial regressions were $3.784, 5.421, 2.711$ for degrees 1, 2 and 3, respectively. The first level bandwidths in the double-smoothing algorithm were $2.649, 3.795, 1.897$, respectively.

Again, the local quadratic double-smoothed result is not very different from its conventional counterpart, but now we see a big difference between double-smoothed local cubic and local cubic regression. There is a lot of spurious wiggliness in the local cubic regression which is only partially corrected in the double smoothed version; the bandwidth is likely much too small here.

As in the local linear case, we see the biggest changes in the region where the bias assessment tool highlighted bias issues. When the bandwidth has been chosen appropriately, the double-smoothed regressions appear to have corrected the bias to some extent.

5.4 Discussion and Data Issues

We have applied our bias assessment tool to the median initial attack response time data set. When applying local regression to the data, using a reasonable fixed bandwidth, we see fairly serious bias at a few time points, including the time (the year 1950) where there is interest in whether a change point has occurred. By applying double-smoothing, we can reduce the bias, in an automatic way with local linear double-smoothing, but also by reducing the first level bandwidth in higher degree double-smoothing.

5.4.1 Description of Data Issues

In preparing the data set, there were a number of interesting issues that should be highlighted. We found that there were some calculated time values that were either missing or negative (which does not make physical sense). Potential reasons are as follows:

1. **Missing initial attack record.** Since initial attack time has been used to get the time metric of initial attack response time, when initial attack time was missing, a default date 1900-01-01 was employed as a data resource description. This generated negative initial attack response time. We have treated these records as missing instead. This problem occurred in 295 fire cases out of 9,187 fire cases.

2. **There are fire cases out of logical sequence.** These are relatively easy to correct, since we can see that the logical events (see the introductory chapter should follow or closely follow a time flow. If just one procedure was recorded improperly, it can be corrected with confidence. At least we can set up an interval for the correct time. There are about 20 fire cases in this category.

3. **There was an overnight recording problem.** As metadata resources indicated, some 12:00:01 AM (formatted as hour:minute:second AM/PM) may not be valid, since these values may just indicate some time of the day. There about 30 fire cases in this category.

4. **AM/PM problem.** This means the recorder misclassified AM/PM. If just one of the "events" is out of logical sequence by a half day, it is easily corrected. There are about 8 fire cases that can be so identified.

5. **Negative values.** Some cases with 2 or 3 of the event procedures were recorded with a discrepancy from other procedures. It would be hard to determine the causes. 57 cases are in this category. For analysis, we have deleted these negative values.

5.4.2 Data Cleaning Procedure

Based on the reasoning given in the preceding section, the following data cleaning process was undertaken. After consulting with David Martell of the University of Toronto, to make statistical inferences, we took the next several steps to get the desired data:

1. All data points with missing values in the following variables were excluded: Initial

attack time, travel time, indicator of getaway date in the same day of report date and indicator of fire-being-held time is before the noon time of next day from report time, final size is smaller than or equal to 4 acres. This step reduced the number of records from 9,187 to 8,717.

2. Travel time should be non-negative, so any negative data points with negative travel time were excluded. The number of records was so reduced from 8,717 to 8,629. Excluding negative initial attack time further reduced the number of records to 8,322. Excluding negative getaway time resulted in 7,907 records.

3. Upper cut-offs for each time variable were set, so that unreasonably high records could be excluded: Travel time and getaway time should be less than or equal to 48 hours, and initial attack time should be less than 200 hours. These last omissions gave us the number of records as 7,809.

When the research question only concerns the trend in initial attack response time, there is no need to clean all data records like travel time and getaway time. The initial attack time analysis used a data set with 8,353 records. However, in terms of sensitivity analysis, since the median is robust to outliers, there is limited impact on the results above with and without eliminating the missing or negative initial attack time values.

In future work, it would be of interest to consider quantiles, other than the median. For such purposes, the data must be cleaned more thoroughly, depending on the level of quantile to be studied. Extreme quantiles must be analyzed with caution, while moderate values such as the quartiles could be analyzed in the manner we have described. Double-smoothing quantile regression itself is another topic for future research.

Chapter 6

Application to Albini's Spotting Model

The goal of this chapter is to provide a way of handling the uncertainty in the maximal spotting distance model proposed by Albini (1979). Ultimately, the desired outcome is a simulation model which outputs the distribution of the maximal horizontal displacement of a burning particle (i.e. a firebrand) of a particular shape (we consider cylinders here to be consistent with Albini (1979), but other shapes could be considered as well), given a number of input parameters, such as tree height, wind speed, and characteristics of the fuel and air.

As part of this, we use the methods developed in this thesis to help in the study of the data set on firebrand burning rate provided in Albini (1979) to obtain statistical models for the conditional distribution of the burning rate, given an index (based on a ratio of air density to particle density, time burned and wind speed) proposed by Albini as the covariate. Unlike Albini, we will transform the response with a logit before doing the model fitting to take into account the fact that the response is constrained to be within the interval (0, 1). In the Chapter 2, a plot of Albini's data has been included in Figure 2.6, where the inappropriateness of doing linear regression through the origin on data with the given type of response is evident. The models we consider here are

1. a parametric random coefficient regression model

2. a nonparametric local linear regression model

3. a data sharpened monotonically increasing local linear model

The results from the parametric model are used to demonstrate how a firebrand spotting mechanism could be simulated. The nonparametric models could be used as alternatives to the parametric model and provide an opportunity to illustrate the bias assessment tool in a practical setting.

The rest of the chapter will proceed as follows. First, we describe the randomized maximal firebrand spotting distance model, noting the variables to which we will be attaching probability distributions, as well as a description of the burning rate data used in Albini (1979); this section will conclude with an outline of the simulation procedure in order to give a precise and detailed notion as to how the application can proceed. In the subsequent section, we describe the random coefficient approach to the burning rate data set. We next consider local linear regression and its bias assessment, and apply double-smoothing in order to reduce the bias. Finally, we consider a data sharpened monotonically increasing local linear model.

6.1 Spotting Simulation Model

The goal of this section is to provide context for the use of the kinds of models that we can develop, in an important practical setting. The problem of firebrand spotting, the spontaneous production of new fires at a location which is fairly remote from a currently burning wildland fire, has serious safety and insurance ramifications. As we have seen, an important early model for maximal firebrand spotting distance was developed by Albini (1979); there have been several updates over the years, but we focus on this version, because it is the most basic and represents a possible starting point for future developments.

In a conference talk ((Alexander, 2009) in Hinton, Alberta, Canada), Senior Fire Behavior Research Officer Dr. Martin E. Alexander mentioned that Albini's early work never underestimated the landing distances when compared with observations from the field. Stochastic spotfire models are desired in order to identify the most likely locations to expect spotfires.

The current research is built on the work of Albini (1979) by using much of the physics he developed. However, the main focus of our study is on the following:

- studying the distributional approximation to the maximal spotting distance
- treating Albini's fixed parameters as appropriately random variables where possible

6.1.1 A Firebrand Spotting Distance Simulator

As Boychuk et al. (2009) mentioned, not all the factors affecting spotting distance are readily observable, but the main factors include:

- wind speed profile (assumed to be logarithmic in some scenarios)
- initial firebrand density, size and shape
- initial firebrand height and velocity

Moisture is an important factor, and we assume that the wood is dry enough to burn with intermediate severity. This is about 5 - 15% fuel stick moisture, 5 - 20% forest litter moisture, and 15 - 60% relative humidity ((Albini, 1979). Heavier moisture will result in low ignition hazards, while lighter moisture corresponds to extreme situations which is also of interest here.

Following Albini, we use the definitions:

- $X(t)$: horizontal displacement at t
- $z(t)$: vertical displacement at t
- H: treetop height
- U_H: horizontal windspeed at treetop height H
- W: vertical windspeed
- z_o: friction length, estimated to be 0.13H for forest covered terrain under stable conditions

- $v_o(t)$: terminal velocity
- ρ_s: density of particle at time t
- ρ_a: density of air
- D: the diameter of the firebrand

As in Albini (1979)'s study, the governing equations are:

$$\frac{dX}{dt} = U_H ln(z/z_o)/ln(H/z_0) \tag{6.1}$$

$$\frac{dz}{dt} = -v_o(t) \tag{6.2}$$

$$\frac{d(\rho_s D)}{dt} = -K\rho_a v_0 \tag{6.3}$$

$$v_o(0) = \sqrt{\pi g(\rho_s D)/(2C_D\rho_a)} \tag{6.4}$$

Equation (6.1) describes the change in horizontal displacement of a burning firebrand at time t, assuming a logarithmic wind profile; Equation (6.2) describes the change in vertical displacement, assuming $W = 0$; Equation (6.3) describes the decay of the burning firebrand; Equation (6.4) describes the terminal fall velocity in a still fluid. Note that (2.5) in Chapter 2 was derived from (6.4) using the maximal horizontal displacement assumption. We do not make that assumption here.

Based on these assumptions, Albini (1979) obtained a simple form for the terminal fall velocity as it changes with decaying mass. He first noted that the solution to the mass decay equation (6.3) is

$$\sqrt{\rho_s D} = -\frac{1}{2}K\rho_a \sqrt{\pi g/(2C_D\rho_a)}t + (\rho_s D)_o/2 \tag{6.5}$$

This allowed him to write

$$v_o(t) = v_o(0)(1 - t/\tau) \tag{6.6}$$

where $\tau = 4C_D v_0(0)/(K\pi g)$ as a shorthand notation.

From the solution of the mass decay equation (6.5), we can determine the time when the firebrand burns out, assuming this happens when the mass disappears. That is, find t such that $(\rho_s D)_t = 0$:

$$t_{\text{burnout}} = \frac{(\rho_s D)_0 \sqrt{2C_D}}{K\sqrt{\pi g \rho_a}} \tag{6.7}$$

Integrating the vertical displacement equation (6.2), using (6.6) gives

$$z(t) = z(0) - v_o(0)(t - t^2/(2\tau)) \tag{6.8}$$

To find the landing time, we set $z(t) = 0$ and get

$$t_{\text{landing}} = \tau \pm \sqrt{\tau^2 - 2\frac{z(0)}{v_o(0)}} \tag{6.9}$$

The solution with the minus sign corresponds to the time that a firebrand lands. The one with the plus sign is a theoretical solution which is not relevant.

According to this model, the firebrand will land and still burn when

$$t_{\text{landing}} \leq t_{\text{burnout}}.$$

Otherwise, the firebrand will burn out during flight and not be able to start a new fire upon landing. For the firebrands that land and still burn, numerical integration of the horizontal displacement equation over the interval $[0, t_{\text{landing}}]$ gives the horizontal displacement of the firebrand in a maximal approximation sense.

An important component of this process is the factor K which Albini derived from the data set already described in the literature review chapter. He used a regression through the origin model to estimate K, but it is clear that there is a great deal of uncertainty surrounding K, so treating it as a random variable seems more appropriate. Furthermore, the response variable

for deriving K is necessarily constrained to be between [0,1], so a simple regression model is not appropriate unless the dependent variable is in the valid support. In general, a logistic regression is more useful in such a situation. The logit function $logit(y) = \frac{y}{1-y}$ can be used to transform the response variable so that it could take any real number.

In a nonparametric modelling context, for simulation, to find the distribution for t_{burnout}, we can resample from residuals of the form

$$\epsilon = y_t - \frac{\rho_s D_o}{\rho_s U}.$$

For each observation, find the minimum value of t for which

$$\epsilon + k(t\frac{\rho_s D_o}{\rho_s U}),$$

this will be t_{burnout} for that observation.

If t is too large, then we estimate the derivative of $k()$ using local linear (1st derivative) regression and use the value of $k'(t)$ for large t as k then use Albini's equation

$$(\rho_s D)_t = -Kt\rho_a U + (\rho_s D)_o \Rightarrow \frac{(\rho_s D)_o - (\rho_s D)_t}{k\rho_a U} = t_{\text{burnout}}$$

The parameters that we randomize are: C_D, D and K. For a specific simulation, comparison of t_{landing} and t_{burnout} will determine whether the firebrand is still alight when it lands on a fuel bed. The whole simulation procedure goes as follows:

1. Input: the initial tree height, initial densities and wind speed.

2. Randomly generate C_D, D and K. Randomization of C_D is to be realized by θ which is an angle between wind direction and the firebrand. Details about these parameters will be discussed in the next section.

3. Calculate the landing time and burn out time for each run.

4. Identify the simulated firebrands for which $t_{\text{burnout}} \geq t_{\text{landing}}$. These are the ones that are still burning after landing.

5. Integrate the horizontal displacement function from 0 to t_{landing} for the firebrands still alight from the previous step.

A maximal initial tree height can be chosen, based on observation of the burning forest stand, under the assumption that firebrands emitted from taller trees have potential to spot at longer distances than those coming from shorter trees. Maximal wind speed could be obtained from measurements of wind gusts near the fire. The initial density of the air is assumed to not be highly variable. The initial particle density will depend on the type of tree in the forest stand to some extent, but the particles are all made of essentially the same type of cellulose-based material.

6.1.2 Modelling the C_D, D and K as Random Variables

Albini (1979) used the maximum drag coefficient for a cylinder $C_D = 1.2$. In our work, the drag coefficient is taken as a function of angle of attack for cylindrical objects. The relation $C_D = 0.0112 * \theta + 0.162$ (Sardoy et al., 2008) is employed to run the simulation in the current study. θ is the angle of attack (in degrees) between the wind and the object particles which are assumed to be either a cylinder or a cone. With 90° of angle of attack, the drag coefficient is 1.17, which is very close to the number that the study of Albini (1979) adopted to maximize the horizontal displacement.

The diameter, D, is assumed to be normally distributed with mean 1.1 and standard deviation 0.1.

All of Albini's assumptions are made in the current study as well. In Albini (1979), Submodel C describes the firebrand burning rate model. Based on data Albini collected from the

Northern Forest Fire Laboratory, the model was fitted as:

$$Y = bX$$

with $Y = 1 - (\rho_s D)/(\rho_s D)_o$ and $X = \rho_a U t/(\rho_s D)_o$ where ρ_s is the density of firebrand (Kg/m^3), ρ_a is the air density (Kg/m^3), D is the diameter of brand (m). Subindex o means initial conditions. U is wind speed (m/s). t is time (s). X and Y are both dimensionless quantities. b is correlated with burning rate. This simple regression was fitted in Albini (1979) with output: b is estimated as 0.0064, with a Root Mean Square Error (RMSE) = 0.24.

The current study assumes that the decay rate should change as the density and diameter of the firebrand change. So an alternative random slope regression is incorporated and fitted here. Based on the notation Albini introduced, the Random Slope Regression model in this study is:

$$Y = Kx + \epsilon, \quad K \sim N(k, \sigma_K^2), \quad \epsilon \sim N(0, \sigma_\epsilon^2)$$

Then $Y \sim N(kx, \sigma_K^2 x^2 + \sigma_\epsilon)$. The same data set from the Northern Forest Fire Laboratory is employed.

The distribution of K is estimated to be $N(0.005620, 0.001619)$. This result will be used in the following simulation.

6.1.3 An Example

With initial conditions given as in the example from Albini (1979) but using the fitted random burning rate model as discussed in the previous section, we obtained output from one simulation with 10000 runs: the random landing locations of the simulated firebrands that remained burning until after they landed. The results are displayed in Figure 6.1.

This simulation shows that the time to burn out and the time to land are approximately lognomally distributed. The burn out duration distribution is more apparently right skewed. With landing time shorter than burnout time, most of the firebrands are alight when they land

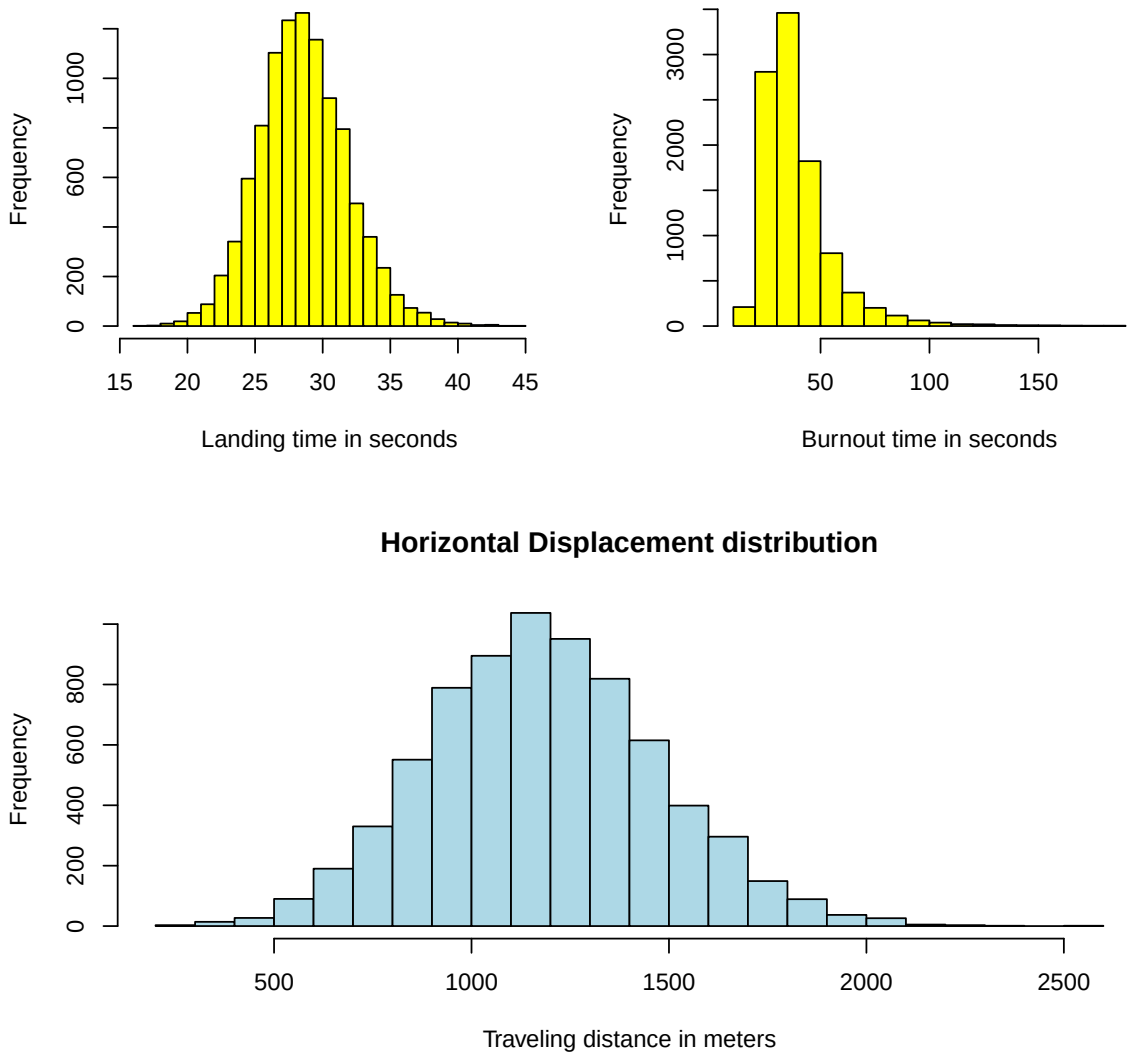

Figure 6.1: Simulation of 10000 runs for approximation to maximal horizontal displacement based on analysis of landing time and burn out time in seconds.

on the ground. The mean levels of both times are somewhat consistent with the conjectures in Boychuk et al. (2009). The histogram of the horizontal displacements is right skewed and the median is around 1183 meters. A log normal fit seems reasonable. Under the same initial conditions, Albini (1979) predicted the maximal distance to be about 1500 meters (0.92 miles).

6.1.4 Limitations of the Simulation Model

This simulator will systematically overestimate the actual spotting distance, because it is still using some of the assumptions from Albini (1979) which lead to a maximal type result. However, this approach to simulation is still very useful, since the deterministic approach to the maximal value is, itself, limited, and providing information about this extremal statistic is of use in practice. As an example, Albini (1979) assumed that the particle travels at wind speed immediately when it is released from the vertical plume zone. This will lead to a larger horizontal displacement than necessary. This is an important consideration for any future development of such a simulator. Randomness of the wind speed is not considered in the crude model considered here. The decay rate model may be based on shape of the object, and we studied cylindrical objects only. The distribution of the drag coefficient is in need of validation. Experimental data should be employed to model the distribution of the drag coefficient.

6.2 Local Regression and Bias Assessment

As noted in the introduction to this chapter, we could model the burning rate nonparametrically, using the residuals from the fitted model in a residual-based bootstrap simulation in place of the random coefficient component of the firebrand simulator. Here we consider kernel regression.

In Chapter 3, we proposed a bias assessment tool to check for potential bias in local polynomial regression. The decay data is a suitable candidate to utilize the tool. The result with local linear regression is shown in Figure 6.2. In particular, local linear regression was applied,

with a gaussian kernel and bandwidth of 40, to the burning rate data that was studied in Section 6.1.2 as a function of burning index. The result is shown in Figure 6.2. The bias assessment is shown in the lower panel of the Figure. It indicates that there is some significant bias and that there is scope for reduction in bias, either through double-smoothing or through reduction in the bandwidth.

In the next section, we consider double-smoothing. In the subsequent section, we consider a reduction of the bandwidth which results in features which need to be corrected through data sharpening.

6.3 Double-Smoothing

Double-smoothing techniques proposed in Chapter 4 and demonstrated in Chapter 5 can also be applied to the experimental data from Albini.

We first applied double-smoothing using the cross-validation bandwidth for the local polynomial regressions of degrees 1, 2 and 3, with a multiplier $k = 0.7$ for the first level of smoothing, and using our quick and simple bandwidth for the second level of smoothing.

Bandwidths for the the local polynomial regressions were $50.295, 117.47, 36.86$ for degrees 1, 2 and 3, respectively. The first level bandwidths in the double-smoothing algorithm were $35.206, 82.229, 25.802$, respectively. The bandwidth for the second level of smoothing was 4.289.

We see that the double-smoothed local linear appears to correct the bias problem in the vicinity of a burning index of 90. The local quadratic and local cubic smooths introduce boundary issues that are re-corrected using double-smoothing, resulting in curves for all three degrees that resemble each other strongly.

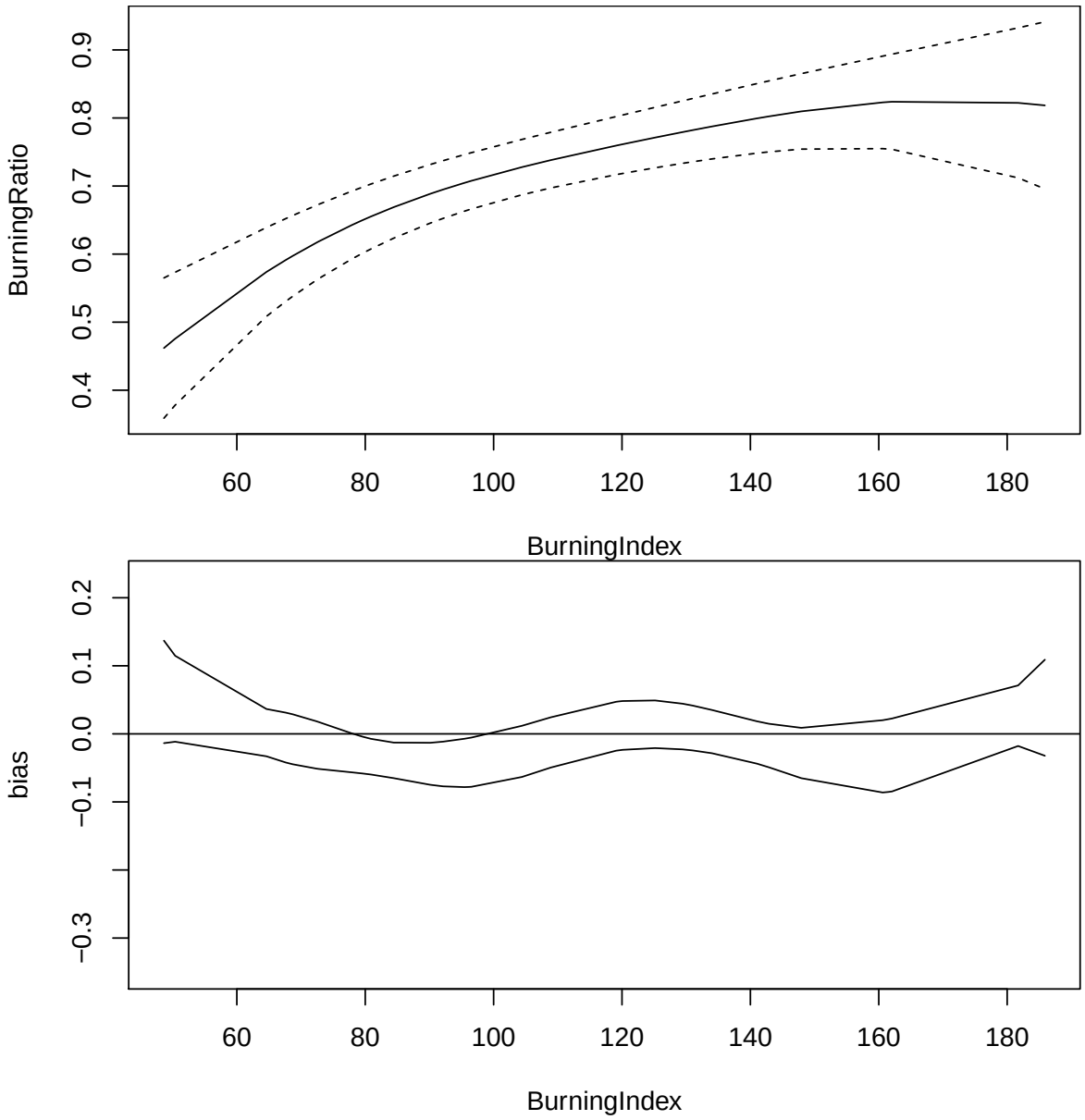

Figure 6.2: Bias Assessment of local polynomial fitting. Data is burning rate data from Albini 1979. Top: local linear regression with $k_1 = 1$ and $h_1 = 30$. Bottom: Pointwise confidence interval assessment plot for bias with $k_2 = 2$ and $h_2 = 20$.

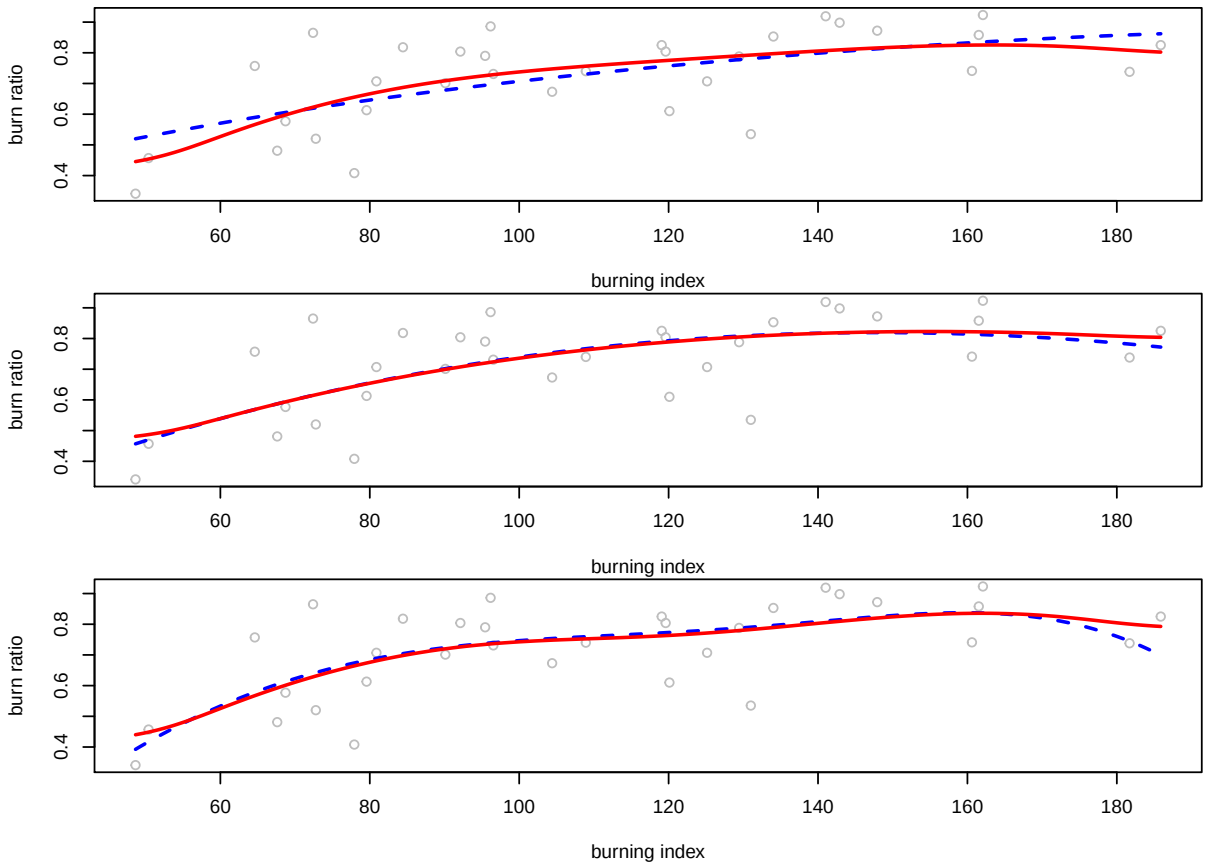

Figure 6.3: Albini's data with overlaid local polynomial and double-smoothed local polynomial regression curves, using a bandwidth multiplier of 0.7 in the double-smoothed regression. The double-smoothed curve is solid red and the local linear curve is dashed blue. Top Panel: local linear; Middle Panel: local quadratic; Bottom Panel: local cubic.

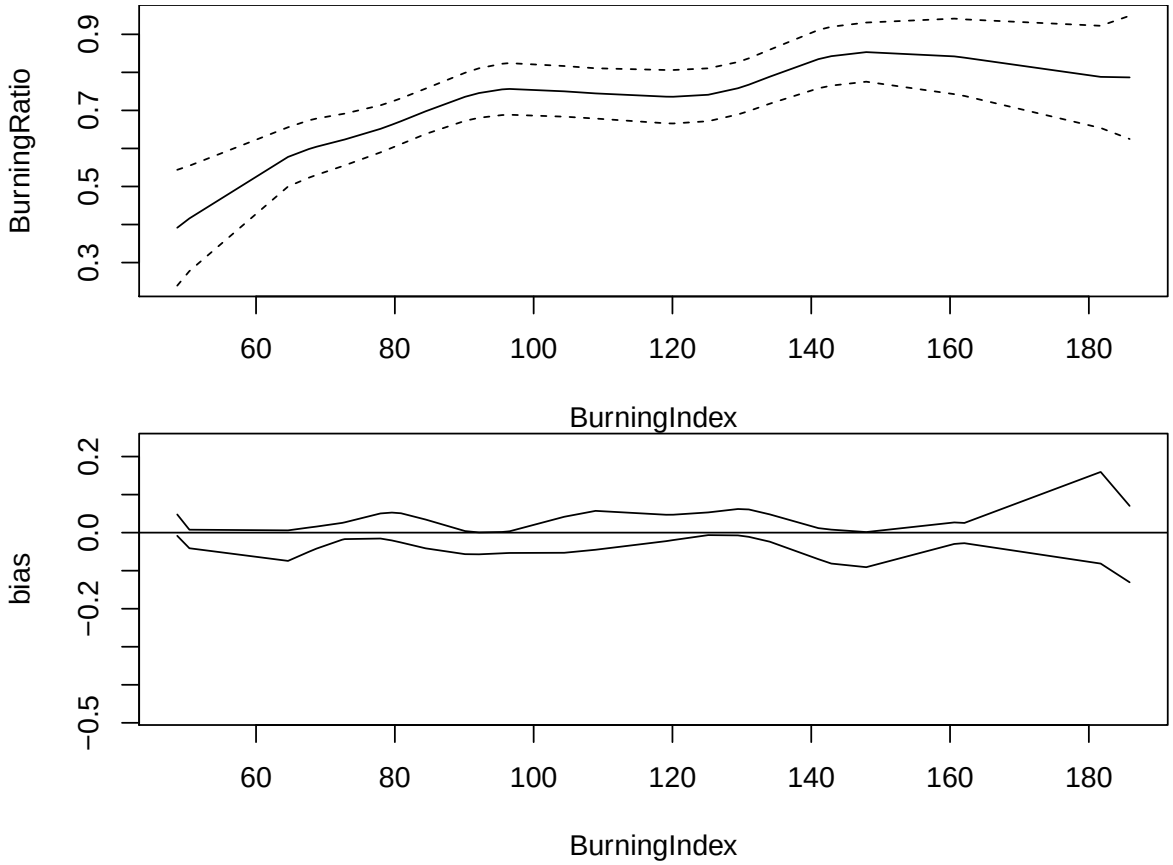

Figure 6.4: Bias Assessment of local polynomial fitting. Data is burning rate data from the study of Albini 1979. Top: local polynomial fitting with $k_1 = 1$ and $h_1 = 10$. Bottom: Pointwise confidence interval assessment plot for bias with $k_2 = 2$ and the $h_2 = 10$.

6.4 Monotone Local Polynomial Fitting

Another way to handle the bias problem is to reduce the bandwidth. This increases the variance and has other consequences as we see in Figure 6.4, where we have employed a smaller bandwidth ($h = 10$), and we have applied our bias assessment tool.

In terms of bias, the assessment tool is indicating that we have gone a long way to reducing bias. However, the cost is a curve which is no longer realistic. We should expect a monotonically increasing function over most or all of the support. Here we see spurious dips which are a result of the increased variance in the estimation. Theoretically, it should increase over the whole of its support, with a possible plateau or mild deterioration towards the right boundary.

This is motivation to incorporate the data sharpening method. As reviewed in Chapter 2,

this is a method that slightly shifts the response points so that the performance can be improved relative to raw data when doing statistical modeling or estimating (Choi et al., 2000). In this case, we aim to improve performance in estimation by inducing the requirement that the regression function does not decrease anywhere on its support. In terms of data sharpening, we aim to minimize the distance between the observed responses y_i and sharpened responses y_i^*,

$$\sum_{i=1}^{n}(y_i - y_i^*)^2$$

subject to the constraint that $\widehat{m}'(x)$ is nonnegative on a fine mesh overlapping the support. Here, we consider the following version of the local constant estimator

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} K_h(x_i - x)y_i^*}{\sum_{i=1}^{n} K_h(x_i - x)}$$

with the Gaussian kernel.

Quadratic programming can be applied to do the minimization to determine the sharpened responses. The resulting monotone increasing local constant regression estimate for the burning rate data is shown in Figure 6.5. The curve no longer has regions of decrease, but it is not very smooth.

Combining the data sharpening and double smoothing is a way to overcome this difficulty, and is worth systematic exploration in future work. In the present work, we simply apply the double smoothed local linear, quadratic and cubic regressions to the sharpened data, coming from the constrained local constant regression problem. In all cases, the local polynomial estimate with $h = 10$ has been computed using the raw data, and the corresponding double smoothed regression estimates are obtained for the sharpened data, using a first level bandwidth of $h_1 = 7$ and the second level bandwidth $h_2 = 4.2886199$, used earlier. and The results are pictured in Figure 6.6, where we see monotonic increasing curves for the double smoothed estimates which are much smoother than the wiggly non-monotonic curves coming from the conventional estimates, having been applied in cases where the bandwidth has possibly been

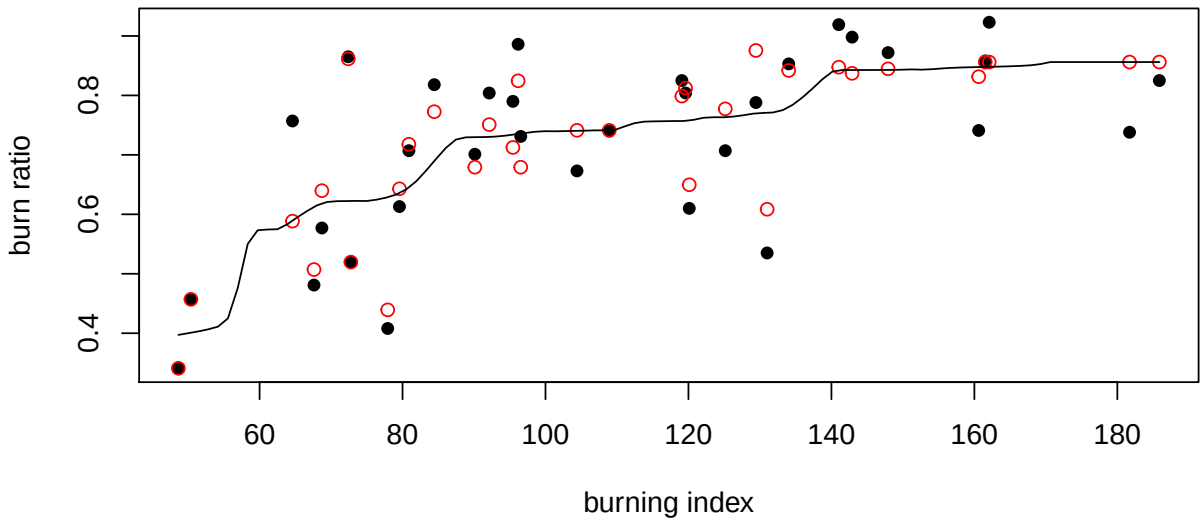

Figure 6.5: Monotone local estimation. Data is burning rate data from Albini 1979. The black dots are observed records; red circles are sharpened records; the blue curve is the monotone local constant estimate.

taken too small.

6.5 Discussion

In this chapter, we have considered the firebrand spotting simulation problem in some level of detail in order to illustrate the issues and to give an indication of the complexity of the problem as well as to set the context for the use of the various techniques that we have been considering in this thesis. The original random coefficient version of the simulator has its appeal but is based on parametric assumptions which may not be true.

In addition to double smoothing, we have also considered constrained data sharpening in order to preserve monotonicity in the regression estimate for the burning rate data. To overcome the lack of smoothness in the result, we have proposed a combination of double smoothing and data sharpening to render a smoother monotonic fit. The results of our ad hoc approach look promising and provide motivation for future work where the sharpening algorithm should be incorporated directly into the double smoothing procedure.

The residuals from either of the nonparametric smoothing models could be used in a boot-

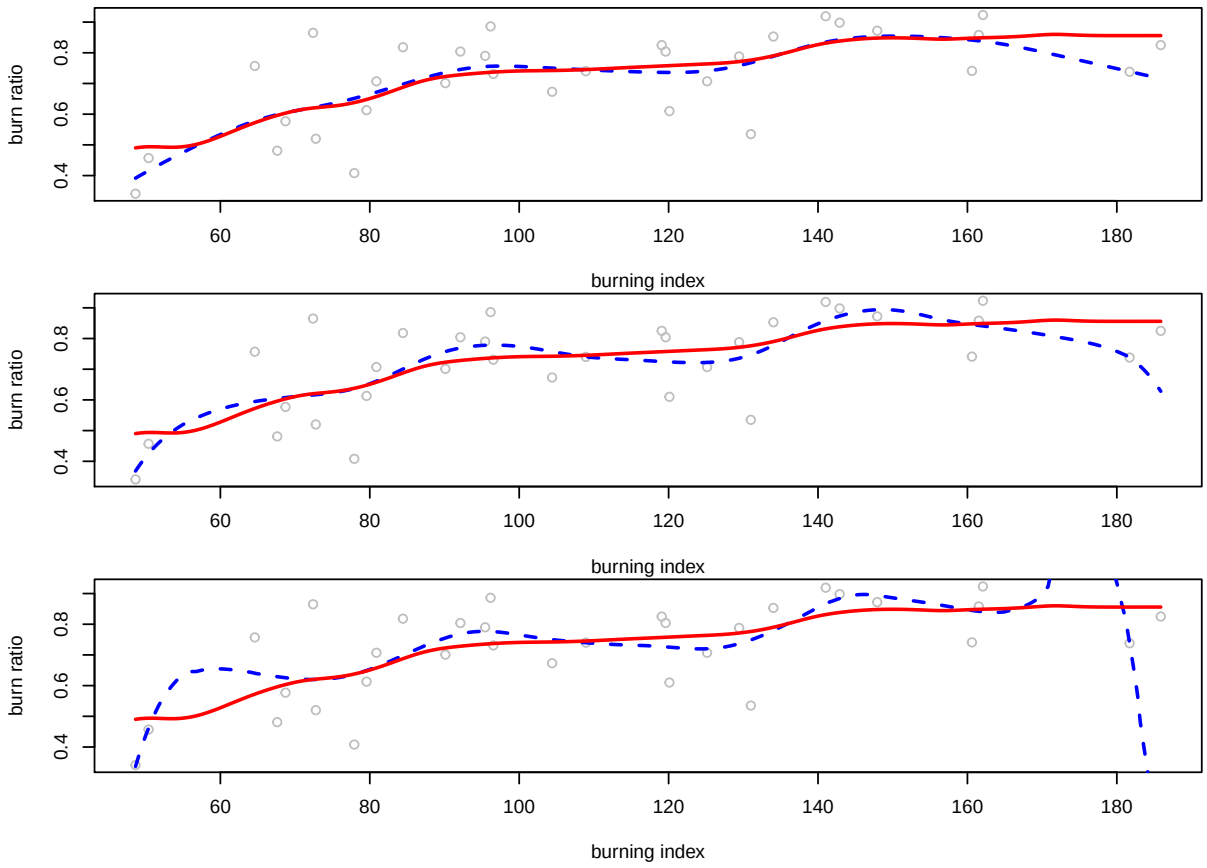

Figure 6.6: Albini's data with overlaid local polynomial and sharpened double-smoothed local polynomial regression curves, using a bandwidth multiplier of 0.7 in the double-smoothed regression. The double-smoothed curve is solid red and the local linear curve is dashed blue. Top Panel: local linear; Middle Panel: local quadratic; Bottom Panel: local cubic.

strap procedure as alternatives to the random coefficient procedure to produce simulators that are less dependent on modelling assumptions. In a future study, output from the three simulators should be compared and brought to fire science experts for their consideration.

Chapter 7

Conclusions and Remarks

In this thesis, we developed a bias assessment tool for Local Polynomial Regression and extended double smoothing local linear regression to higher order local polynomials regression in a practical way. This assessment tool and double-smoothing local quadratic and cubic regressions were applied using simulation scenarios and classical standard data sets to demonstrate their practical performance. In addition, they were employed to analyze real data sets to investigate the trend in an initial attack problem in the fire control environment of Northeastern Ontario over the years of 1930 to 2012, and also for nonparametric estimation of the burning rate for spotting fire brands based on laboratory experiment results. Other bias reduction techniques, such as the data sharpening technique was incorporated to enforce the constraint of monotonicity.

With the constructed bias assessment tool, we concluded that caution should be given to the area or support segment where high curvature was evident. The pointwise confidence bands appear to be quite accurate in moderate to large sample sizes, and highlight difficulties in estimation of regression functions that have varying levels of curvature. In the real data examples, the usefulness of our assessment plots was illustrated. It helped us to visualize the efficiency of the bias estimate. In the initial attack problem, with the bias assessment tool, we saw that the bias around the change point of interest (year 1950) was relatively high. This led

us to further investigate reasonable methods to reduce the bias.

The extension of double-smoothing to higher order local polynomial regression offers an improvement in terms of bias and MSE comparing to ordinary local polynomial regressions. The improvement in accuracy was less dramatic than that obtained when applying double-smoothing to local linear regression. We found that using the same bandwidths for both levels of double-smoothing was not always to be recommended. We have included a simple method for choosing the second level bandwidth based on the average of the successive differences in the predictor values that works reasonably well. A cross validation method for the first level bandwidth was utilized. This could probably be improved upon.

The various techniques considered in this thesis were incorporated into firebrand spotting problem to assess and reduce the bias while doing smoothing of experimental data. A framework for stochastic simulation of firebrand spotting was proposed, based on a random coefficient regression approach, applied to burning rate data produced in a laboratory. In future work, simulators of the maximal spotting mechanism can be constructed using methods other than random coefficient regression, such as the double smoothing procedure proposed in this thesis.

Further investigation of the burning rate data using local polynomial regression with smaller bandwidths motivated consideration of data sharpening, further motivating an approach which combines data sharpening with double smoothing. We have shown, for the first time, that double smoothing of data which have been sharpened to provide monotonicity in local constant regression also leads to regression estimates which are monotonic, or approximately so. A systematic investigation of this combined approach should be a fruitful line of future research.

Bibliography

Albini, F. A. (1979). Spot fire distance from burning trees-a predictive model. *USDA Forest Service General Technical Report*, 75:1–80.

Albini, F. A. (1981a). A model for the wind-blown flame from a line fire. *Combustion and Flame*, 43:155–174.

Albini, F. A. (1981b). *Spot fire distance from isolated sources: extensions of a predictive model*, volume 309. US Dept. of Agriculture, Forest Service, Intermountain Forest and Range Experiment Station.

Albini, F. A., Alexander, M. E., and Cruz, M. G. (2012). A mathematical model for predicting the maximum potential spotting distance from a crown fire. *International Journal of Wildland Fire*, 21:609–627.

Alexander, M. E. (2009). Some pragmatic thoughts on the prediction of spotting in wildland fires. In *MITACS/GEOIDE Conference on Forest Fire Modelling June 22, 2009, Hinton, AB*.

Boychuk, D., Braun, W. J., Kulperger, R. J., Krougly, Z. L., and Stanford, D. A. (2009). A stochastic forest fire growth model. *Environmental and Ecological Statistics*, 16(2):133–151.

Braun, W. J. and Hall, P. (2001). Data sharpening for nonparametric inference subject to constraints. *Journal of Computational and Graphical Statistics*, 10(4):786–806.

Busch, N. E. and Panofsky, H. A. (1968). Recent spectra of atmospheric turbulence. *Quarterly Journal of The Royal Meterorological Society*, 94(132-148).

Canadian Council of Forest Ministers (2005). Canadian wildland fire strategy a vision for an innovative and integrated approach to managing the risks. Technical report, Canadian Council of Forest Ministers.

Chen, H. (2011). Local polynomial wavelet estimation of local average treatment effect. Technical report, Working Paper, Vanderbilt University.

Choi, E. and Hall, P. (1998). On bias reduction in local linear smoothing. *Biometrika*, 85(2):333–345.

Choi, E. and Hall, P. (1999). *Data sharpening as a prelude to density estimation*, volume 86. Oxford University Press.

Choi, E., Hall, P., and Rousson, V. (2000). Data sharpening methods for bias reduction in nonparametric regression. *Annals of statistics*, pages 1339–1355.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102.

Fan, J. and Gijbels, I. (1995a). Adaptive order polynomial fitting: bandwidth robustification and bias reduction. *Journal of Computational and Graphical Statistics*, 4(3):213–227.

Fan, J. and Gijbels, I. (1995b). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 371–394.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. London Chapman and Hall.

Härdle, W. (1991). *Smoothing techniques: with implementation in S*. Springer-Verlag New York.

He, H. and Huang, L.-S. (2009). Double-smoothing for bias reduction in local linear regression. *Journal of Statistical Planning and Inference*, 139(3):1056–1072.

Holmes, J. (2004). Trajectories of spheres in strong winds with application to wind-borne debries. *Journal of wind engineering and industrial aerodynamics*, 92:9–22.

Holmes, J., Bakerb, C., and Tamura, Y. (2006). Tachikawa number: A proposal. *Journal of Wind Engineering And Industrial aerodynamics*, 94:41–47.

Janssenswillen, G. (2018). *bupaR: Business Process Analysis in R*. R package version 0.4.1.

Karimpour, A. and Kaye, N. (2012). On the stochastic nature of compact debris fight. *Journal of wind engineering and industrial aerodynamics*, 100:77–90.

Li, Q., Lu, X., and Ullah, A. (2003). Multivariate local polynomial regression for estimating average derivatives. *Journal of Nonparametric Statistics*, 15(4-5):607–624.

Lin, N., Letchford, C., and Gunn, T. (2007). Investigation of flight mechanics of 1d (rod-like) debris. *ASCE J.Structural Eng.*, 2:274–282.

Loader, C. (2013). *locfit: Local Regression, Likelihood and Density Estimation.* R package version 1.5-9.1.

Loader, C. R. (1999). Bandwidth selection: Classical or plug-in? *Annals of Statistics*, 27:415–438.

McArthur, A. (1968). The fire control problem and fire research in australia. Technical report, Proc. 6th World Forestry Congress.

McGourty, J. (2009). *Black Saturday: Stories of love, loss and courage from the Victorian bushfires*. HarperCollins.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

OMNRF (2018). Forest fire management. `https://www.ontario.ca/page/forest-fire-management#section-9`. Online; accessed 19-Aug-2018.

Ramsay, J., Hooker, G., and Graves, S. (2009). *Functional data analysis with R and MATLAB*. Springer Science & Business Media.

Samarov, D. V. (2015). The fast rodeo for local polynomial regression. *Journal of Computational and Graphical Statistics*, 24(4):1034–1052.

Sardoy, N., Consalvi, J., Kaiss, A., Fernandez-Pello, A., and Porterie, B. (2008). Numerical study of ground-level distribution of firebrands generated by line fires. *Combustion and Flame*, 154:478–488.

SAS Institute Inc (2013). *SAS 9.4 Language Reference: Concepts*. SAS Institute Inc. Cary, NC, USA.

Simonoff, J. S. (2012). *Smoothing methods in statistics*. Springer Science & Business Media.

Stocks, B. (2013). Evaluating past, current and future forest fire load trends in canada. Technical report, B.J. Stocks Wildfire Investigations Ltd.

Tachikawa, M. (1983). Trajectories of flat plates in uniform flow with applications to wind-generated missiles. *J. Wind End. Ind. Aerodyn.*, 14:443–453.

Tymstra, C., Bryce, R., Wotton, B., Taylor, S., Armitage, O., et al. (2010). Development and structure of prometheus: the canadian wildland fire growth simulation model. *Natural Resources Canada, Canadian Forest Service, Northern Forestry Centre, Information Report NOR-X-417.(Edmonton, AB)*.

Visscher, B. and Kopp, G. (2007). Trajectories of roof sheathing panels under high winds. *Journal of Wind engineering and industrial aerodynamics*, 95:796–713.

Von Karman, T. (1948). Progress in the statistical theory of turbulence. *Proceedings of the National Academy of Sciences of the United States of America*, 34(11):530.

Wand, M. (2015). *KernSmooth: Functions for Kernel Smoothing Supporting Wand and Jones (1995)*. R package version 2.23-15.

Wand, M. P. and Jones, M. C. (1994). *Kernel smoothing*. Chapman and Hall/CRC.

Watson, G. S. (1964). Smooth regression analysis. *Sankhy: The Indian Journal of Statistics, Series A (1961-2002)*, 26(4):359–372.

Wills, J., Lee, B., and Wyatt, T. (2002). A model of wind-borne derbies damage. *J.Wind Eng. Ind. Aerodynamic*, 90:555–565.

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

Woolford, D. and Braun, J. (2015). *CHsharp: Choi and Hall Style Data Sharpening*. R package version 0.4.

Curriculum Vitae

Name:	Wenkai Ma
Post-Secondary Education and Degrees:	Bachelor of Science in Statistics, 2004 -2008 Nankai University Tianjin, China
	M.Sc. Biostatistics, 2009 - 2010 University of Western Ontario, London, Ontario, Canada
	Ph.D Biostatistics, 2010 - 2018 University of Western Ontario, London, Ontario, Canada
Related Work Experience:	Teaching Assistant, 2009 - 2014 The University of Western Ontario, London, Ontario, Canada
	Research Assistant, 2009 - 2014 The University of Western Ontario, London, Ontario, Canada
	Research Associate, 2014-2015 Lawson Health Science Center, London, Ontario, Canada
	Statistician, 2015 - 2017 Workplace Safety and Insurance Board, Toronto, Ontario, Canada
	Data Scientist, 2017 - present Workplace Safety and Insurance Board, Toronto, Ontario, Canada