Bandwidth Selection Concerns for Jump Point Discontinuity Preservation in the Regression Setting Using M-smoothers and the Extension to Hypothesis Testing

David Allan Burt

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Clint W. Coakley, Chair

Jeffrey B. Birch

George R. Terrell

Eric P. Smith

Robert V. Foutz

March 23, 2000

Blacksburg, Virginia

Keywords: Bandwidth, M-Smoother, Nonparametric Regression, Critical Bandwidth Testing, Bootstrap

Bandwidth Selection Concerns for Jump Point Discontinuity Preservation in the Regression Setting Using M-smoothers and the Extension to Hypothesis Testing

by

David Allan Burt

Clint W. Coakley, Chairman

(ABSTRACT)

Most traditional parametric and nonparametric regression methods operate under the assumption that the true function is continuous over the design space. For methods such as ordinary least squares polynomial regression and local polynomial regression the functional estimates are constrained to be continuous. Fitting a function that is not continuous with a continuous estimate will have practical scientific implications as well as important model misspecification effects. Scientifically, breaks in the continuity of the underlying mean function may correspond to specific physical phenomena that will be hidden from the researcher by a continuous regression estimate. Statistically, misspecifying a mean function as continuous when it is not will result in an increased bias in the estimate.

One recently developed nonparametric regression technique that does not constrain the fit to be continuous is the jump preserving M-smooth procedure of Chu, Glad, Godtliebsen & Marron (1998),'Edge-preserving smoothers for image processing', *Journal of the American Statistical Association* **93**(442), 526-541. Chu et al.'s (1998) M-smoother is defined in such a way that the noise about the mean function is smoothed out while jumps in the mean function are preserved. Before the jump preserving M-smoother can be used in practice the choice of the bandwidth parameters must be addressed. The jump preserving M-smoother requires two bandwidth parameters h and g. These two parameters determine the amount of noise that is smoothed out as well as the size of the jumps which are preserved. If these parameters are chosen haphazardly the resulting fit could exhibit worse bias properties than traditional regression methods

which assume a continuous mean function. Currently there are no automatic bandwidth selection procedures available for the jump preserving M-smoother of Chu et al. (1998).

One of the main objectives of this dissertation is to develop an automatic data driven bandwidth selection procedure for Chu et al.'s (1998) M-smoother. We actually present two bandwidth selection procedures. The first is a crude rule of thumb method and the second is a more sophistocated direct plug in method. Our bandwidth selection procedures are modeled after the methods of Ruppert, Sheather & Wand (1995) with two significant modifications which make the methods robust to possible jump points.

Another objective of this dissertation is to provide a nonparametric hypothesis test, based on Chu et al.'s (1998) M-smoother, to test for a break in the continuity of an underlying regression mean function. Our proposed hypothesis test is nonparametric in the sense that the mean function away from the jump point(s) is not required to follow a specific parametric model. In addition the test does not require the user to specify the number, position, or size of the jump points in the alternative hypothesis as do many current methods. Thus the null and alternative hypotheses for our test are: H0: The mean function is continuous (i.e. no jump points) vs. HA: The mean function is not continuous (i.e. there is at least one jump point).

Our testing procedure takes the form of a critical bandwidth hypothesis test. The test statistic is essentially the largest bandwidth that allows Chu et al.'s (1998) M-smoother to satisfy the null hypothesis. The significance of the test is then calculated via a bootstrap method. This test is currently in the experimental stage of its development. In this dissertation we outline the steps required to calculate the test as well as assess the power based on a small simulation study. Future work such as a faster calculation algorithm is required before the testing procedure will be practical for the general user.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Traditional regression techniques operate under the assumption that the "true" function is continuous. However, there are numerous examples of real world problems that do not satisfy the continuity assumption. One of the most famous examples of a data set containing a break in the continuity of the underlying mean function or a "jump point" is the Nile river flow data (Cobb 1978). Another application where the jump point problem is addressed is the Bombay sea-level pressure data (Qiu & Yandell 1998). In addition to the scientific fields of hydrology and meteorology, jump point applications can be found in quality control, economics, medicine, signal and image processing, and many physical sciences.

One jump point application that the author has recently encountered is the relative light transmittance data provided by Mike Battaglia from the Department of Forestry at Virginia Tech. In this data set sun light from various stations in plots with different cut treatments is compared to the sunlight in a nearby open plot. Relative transmittance data were recorded at equal time intervals throughout the daylight hours for numerous days (see Figure 1.1 for a plot of the relative light transmittance data for one station during one

day). Cloud interference and overstory pattern (the shade produced by each tree) are the two most common phenomena that cause jump points in the relative transmittance data. Jump points that remain consistent across days may be attributed to overstory pattern while jump points that do not remain consistent across days are probably just cloud interference. It is the objective of the researcher to relate overstory pattern to growth rates.

Figure 1.1: Relative Light Transmittance Data

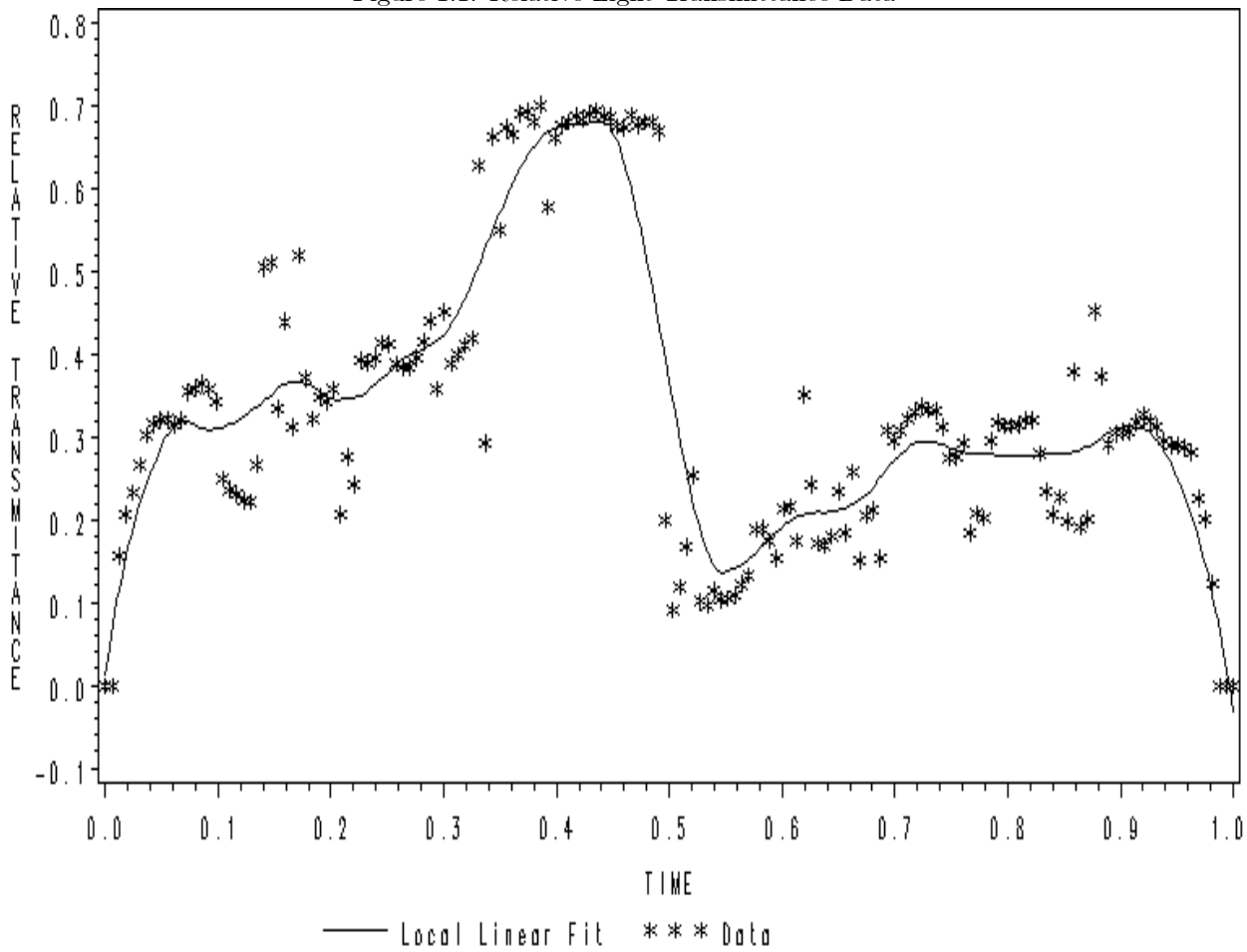


Figure 1.1 displays the relative light transmittance data (data plotted as stars) provided by Mike Battaglia from the Forestry Department of Virginia Tech. Figure 1.1 also includes a nonparametric local linear regression fit (solid line) to the transmittance data.

It is widely known that when jump point discontinuities exist in a regression setting the traditional parametric and nonparametric methods of curve estimation encounter serious consistency problems related to over smoothing (Müller 1992). This can be seen in the local linear nonparametric regression estimates in

Figure 1.2: Broken Sine Simulated Data

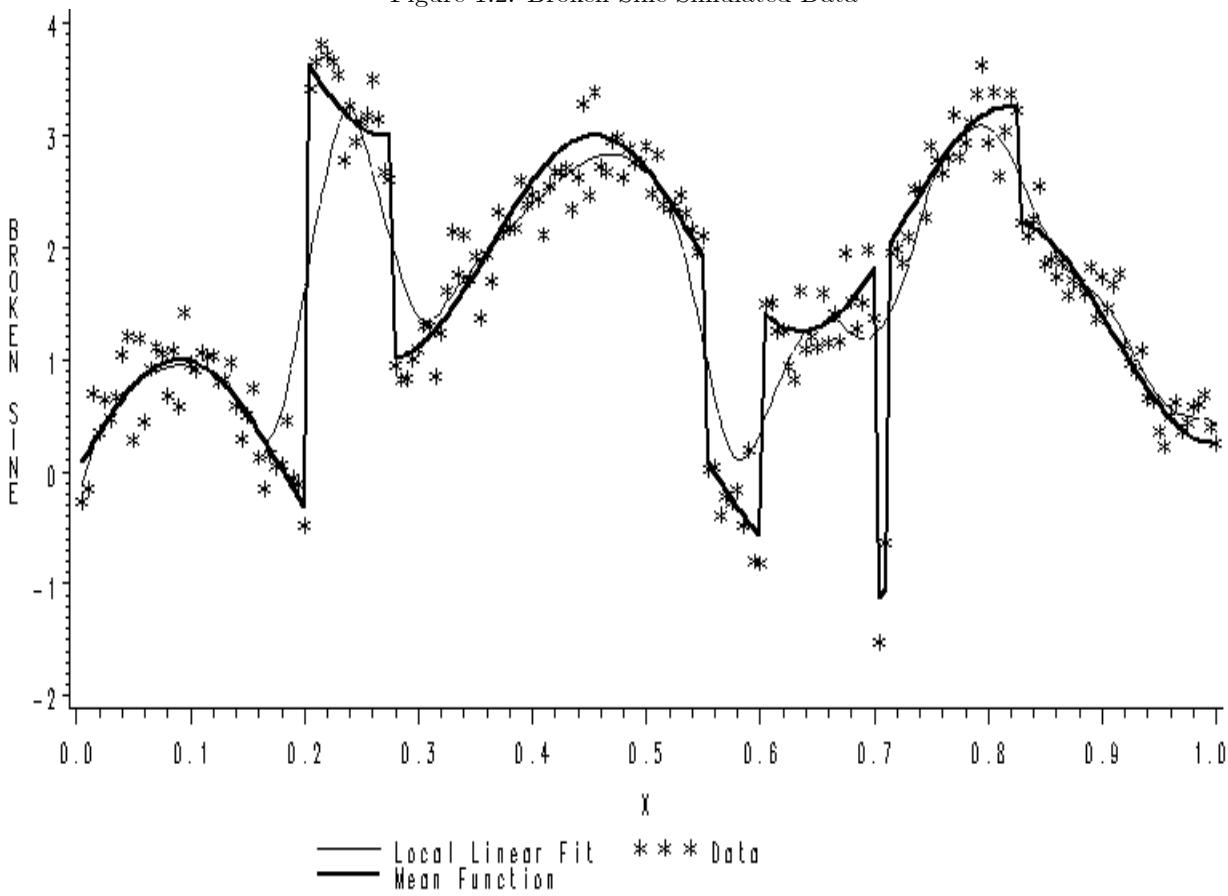


Figure 1.2 shows a simulated data set about the mean function $\sin(5.5\pi x)$ with jump points added at positions $x = 0.2, 0.275, 0.55, 0.6, 0.7, 0.71$ and 0.825 of sizes $4, -2, -1.75, 2, -3, 3$ and -1.75 respectively. The error distribution used to simulate the data about this mean function was Gaussian with mean zero and variance $\sigma^2 = (0.25)^2$. A local linear nonparametric regression was fit to the simulated data using the Direct Plug In (DPI) method of Ruppert et al. (1995).

Figures 1.1 and 1.2. Figure 1.1 demonstrates the local linear regression applied to the relative transmittance data, and Figure 1.2 demonstrates the local linear regression applied to a simulated data set with known jump points (this simulated data set will be described in more detail in Chapter 3). Smoothing over jump points causes inflated bias in the regression estimate as well as a loss of important scientific information. Therefore a jump preserving regression estimate should be employed when the continuity assumption is not met. This leads to the following research question. How does one determine if the continuity assumption is reasonable?

1.2 Literature Review

Much work has been done in regression on detecting and estimating the shifts in the model parameters. For time series models there are well documented methods of testing for interventions in time. There are also nonparametric methods based on asymmetric kernel densities for locating and estimating jumps that are known to exist. Shaban (1980) provides a good annotated bibliography of earlier methods that tackle the jump-point problem. The procedures for estimating and testing change points in regression data discussed in Shaban's (1980) bibliography include parametric, nonparametric, classical, and Bayesian methods.

In the parametric regression setting, Brown, Durbin & Evans (1975) have described and compared a group of methods for estimating and testing for a shift in the coefficients. The first two of these methods rely on calculating recursive residuals. Recursive residuals are similar to regular regression residuals except that they are modified so that their covariances are zero. Under normal theory these residuals are then independent, and through a simple transformation follow a Gaussian process (Brown et al. 1975). The other methods discussed in Brown et al. (1975) are the moving regressions technique and the Log-likelihood Ratio technique. The moving regressions method is similar to nonparametric nearest neighbor regression. As its name implies the Log-likelihood method uses the assumed distribution of errors to calculate a likelihood ratio test. Modifying these methods to the model free, distribution free setting seems intractable. Smooth regression residuals are usually dependent upon the bandwidth chosen, thus are not as informative as recursive residuals in parametric regression. The likelihood ratio test requires that stringent assumptions be applied to the distribution of the errors, which we wish to avoid. Although these methods are not directly adaptable to our nonparametric setting they do serve as useful tools in parametric regression and Brown et al. (1975) even provide the software to implement their proposed methods. For further work including power approximations on these methods see James, James & Siegmund (1987) and Kim & Siegmund (1989).

When the data follow the more specific model of a time series, jumps and sharp cuspids are referred to as interventions. Intervention analysis not only estimates and tests for changes in the general trend, but

also provides a tool for interpreting what type of shift is occuring. Interventions may be a gradual drift away from the original trend, or perhaps a sudden shift that eventually returns to the original trend. The basic method of intervention analysis is to add terms to the original time series model and test for significant parameters. For a more detailed discussion of intervention analysis see Box, Jenkins & Reinsel (1994), Box & Tiao (1965), Box & Tiao (1975), and Glass, Wilson & Gottman (1975). Intervention analysis is very model specific and requires that the data follow a time series structure. Therefore time series intervention analysis cannot easily be modified to a model and error distribution free setting.

More recently Müller (1992) and Loader (1996) have developed two different nonparametric methods of estimating jump points based on local polynomial regression with asymmetric kernel density functions. Müller (1992) and Loader (1996) also provide confidence intervals for both the size and position of the jump point based on the asymptotic properties of their one sided kernel estimates. However these one sided kernel estimates were developed to only consider a single jump point and our research objective is to test for an unknown number of discontinuities.

Another kernel-type estimator of jump point discontinuities is given by Wu & Chu (1993). Here Wu & Chu (1993) adapt the one sided kernel estimates so as to capture multiple jumps. They also provide a hypothesis test for the number of jump point discontinuities. However it is shown that these estimates and the corresponding test are very sensitive to the choice of the bandwidth, and no automatic bandwidth procedure is given.

Recently a jump preserving regression technique based on M-Estimation and extremely useful in the area of image processing was developed by Chu et al. (1998). Chu et al.'s (1998) regression procedure preserves jump point discontinuities in an automatic way that does not require the user to specify the number of jump points or their positions. This technique has been shown to preserve a remarkably large number of jump points. It is conjectured that the jump preserving M-smoother of Chu et al. (1998) has nicer robustness and asymptotic properties than the methods based on one sided kernel estimates. This conjecture originates from the fact that one sided kernel estimates have an added asymptotic bias that the M-smoother

is not suspected to have when there is no jump point discontinuity. Chu et al.'s (1998) jump preserving M-smoother shows promising advantages in a wide variety of applications. However, it is quite sensitive to the choice of the two bandwidth parameters g and h. The bandwidth g is analogous to the tuning constant in most M-estimates and determines the size of jump that can be preserved by Chu et al.'s (1998) regression technique. As in usual kernel and local polynomial regression techniques the bandwidth h determines the amount of local averaging in the jump preserving M-smoother. Currently there is no automatic bandwidth selection procedure available for the jump preserving M-smoother.

1.3 Research Objective

The two primary research objectives of this dissertation are:

i. To provide an automatic bandwidth selection procedure for the jump preserving M-smoother.

ii. To develop a hypothesis test for the presence of a jump point discontinuity in the mean function based on the M-smooth procedure of Chu et al. (1998).

The jump preserving regression technique requires a good automatic bandwidth selection procedure. Thus in the following dissertation a crude Robust Rule of Thumb (RROT) automatic bandwidth selection procedure is developed to estimate g and h. In addition a more sophisticated Direct Plug In (DPI) automatic bandwidth selection procedure is developed which uses the RROT estimates. These algorithms are modeled after the methods of Ruppert et al. (1995) with two significant modifications which make the algorithms robust to possible jump points. The first modification is a variable width blocking scheme and the second is an M-regression estimate within each block. The details behind these modifications will be presented in Chapter 3.

Our jump point hypothesis test follows a critical bandwidth method previously used to test for the number of modes in a kernel density estimate (Silverman 1981) and to test for monotonicity in a regression

mean function (Bowman, Jones & Gijbels 1998). The critical bandwidth procedures of Silverman (1981) and Bowman et al. (1998) operate under the principle that nonparametric estimates generally exhibit more features with smaller bandwidths than do estimates with larger bandwidths. Features in the kernel density estimation setting are multiple modes while features in kernel and local linear regression are wiggles.

The features of the jump preserving M-smoother we wish to exploit are the jumps. Thus to apply the critical bandwidth testing procedure to the jump point discontinuity setting we need to find the smallest value for the bandwidth g such that the jump preserving M-smooth estimate is continuous. Under the null hypothesis (no jumps in the mean function) this critical g will be quite small, and under the alternative hypothesis (at least one jump in the mean function) the critical g will be much larger. A bootstrap method is then used to calculate a p-value based on the critical g found above.

For introductory purposes we have over simplified the steps required to implement our hypothesis test. A step by step outline of the testing procedure is presented in Chapter 4. Many of the tedious details required by the testing procedure lead to methods that are of practical importance outside the hypothesis testing setting. For example Chu et al.'s (1998) original paper did not include a method of determining if the M-smooth fit contained a jump point. This is an important step in our testing procedure and a novel method for determining this is developed in Chapter 4. This method leads to estimates of the number, size(s), and position(s) of all jumps points in the mean function. Furthermore Chu et al.'s (1998) jump preserving M-smooth regression method was originally limited to fit the mean function at x_i's that were in the data set. Since our method supposes that jumps would be easiest to detect between the data points, we present a method to extend the regression technique to fit all x within the design space (particularly the x's located in the center of two adjacent data points).

Before we describe the details behind our bandwidth selection procedure and hypothesis testing procedure, we must first present some background behind M-smoothing. This background, including the basic M-estimation technique, traditional nonparametric regression, and the combination of the two, is presented in Chapter 2. Since bandwidth selection is such a crucial part of nonparametric regression, we provide a brief

survey of existing data driven methods in the nonparametric section of Chapter 2. The last two sections of Chapter 2 describe M-smooth procedures that combine the ideas of M-estimation and nonparametric regression together. The most traditional of these does not preserve jump points while the more recent version does preserve jumps. The data driven bandwidth selection procedures RROT and DPI mentioned above are developed in Chapter 3. The jump point discontinuity hypothesis test is presented in Chapter 4. Finally future research ideas as well as our conclusions from the current research project are presented in Chapter 5.

Chapter 2

Traditional M-Smoothers

In this chapter we give the background behind traditional M-smoothers. In Section 2.1 we build up M-estimation in the location model, then in Section 2.2 we give a review of smoothing methods. In Section 2.3 we provide a survey of the methods that put the two ideas of M-estimation and smoothing together.

2.1 M-Estimation

The term "M-Estimation" refers to a "generalized maximum likelihood estimation" and was coined by Huber (1964) when he first proposed the method. In maximum likelihood estimation we wish to maximize the likelihood, or minimize the negative log likelihood; thus for estimating a simple location parameter we have

$$\hat{\theta}_{MLE} = \min_{\theta} \sum_{i=1}^{n} [-\ln(L_\theta(X_i))].$$

Here $L_\theta(X_i)$ is the likelihood of observing the random variable X_i given the parameter θ. We can think of the $-\ln(L_\theta(X_i))$ as a loss function to be minimized. M-estimators generalize this idea by allowing the loss

function to be modified; thus we have the following "objective function"

$$\sum_{i=1}^{n}[\rho(X_i, \theta)].$$

As different loss functions $\rho(X_i, \theta)$ are chosen the estimate θ_M will exhibit different properties. The property most commonly desired in an M-estimate is robustness to outliers.

The relationship between outliers and regression jump-point discontinuities makes M-estimation relevant to our project. This similarity can be best seen by comparing the jump-point model to the mean shift outlier model. In the single jump-point discontinuity case the model would be

$$Y_j = m(x_j) + \varepsilon_j,$$

for $j = 1, 2, ..., n$ with ε_j iid random errors from a symmetric unimodal distribution F centered about zero, and where the mean function $(m(x))$ is of the form

$$m(x) = \mu(x) + \Delta \cdot I(x > t).$$

Here μ is a continuous function, t is the position of the jump-point, and Δ is the size of the jump-point. Thus the model for the jump-point problem is

$$Y_j = \mu(x_j) + \Delta \cdot I(x_j > t) + \varepsilon_j. \tag{2.1}$$

If we relax the identical distribution assumption above and let ε_j^* $j = 1, 2, ..., n$ be random variables from the distribution $F^* = F_1 \cdot I(x_j \leq t) + F_2 \cdot I(x_j > t)$, where F_1 is a symmetric unimodal distribution centered at zero and $F_2(x)$ is a unimodal distribution centered at Δ, then we can generalize 2.1 to

$$Y_j = \mu(x_j) + \varepsilon_j^*. \tag{2.2}$$

Notice that for the mean shift outlier model, where the outlier occurs at the i^{th} data point, we usually assume the errors ϵ_j come from the distribution $G = F_1 \cdot I(j \neq i) + F_2 \cdot I(j = i)$ where F_1 and F_2 are as described above. The mean shift outlier model can then be written as

$$Y_j = \mu(x_j) + \epsilon_j.$$

Indeed if 2.2 is generalized to the multiple jump-point discontinuity model, then it can be shown that the mean shift outlier model is a special case.

Notice that in (2.2) and the mean shift outlier model the errors follow a mixture distribution. If the regression estimation problem is reduced to the location problem by assuming μ does not depend on the x_j, then we are simply estimating the center of a bimodal mixture of two unimodal distributions. That is, we are seeking the estimate of the location of the distribution $F = \delta F_1 + (1-\delta)F_2$, where $\delta \in [0,1]$. The similarity between regression jump points, outliers, and bimodal mixture distributions is the inspiration behind using M-estimation to model and test for jump-point discontinuities in the regression setting.

2.1.1 Ψ Functions

There are many reasonable choices for the loss function ρ in the M-estimation procedure. If ρ is differentiable, say $\Psi(x,\theta) = \frac{\partial \rho(x,\theta)}{\partial \theta}$, then minimizing the "objective function"

$$\sum_{i=1}^{n} [\rho(X_i, \theta)]$$

is equivalent to finding the appropriate root to the "defining equation"

$$\sum_{i=1}^{n} \Psi(X_i, \theta) = 0. \tag{2.3}$$

Notice that for iid Gaussian errors and the negative log likelihood loss function (2.3) is the ordinary least squares normal equation. Due to the convenience of solving the defining equation, throughout this paper we will consider only those loss functions that are differentiable. Furthermore we shall refer to specific M-estimators via their Ψ function.

If the Ψ function is continuous, monotonic, and odd then the optimization procedure is simplified further. Monotonically increasing continuous odd Ψ functions correspond to symmetric convex loss functions. Symmetry and convexity of the loss function guarantee that the only local minimum of the objective function is also the global minimum. Out of all plausible monotonically increasing continuous odd Ψ functions we wish

to focus on two types: unbounded and limiting. A final class of Ψ functions we consider are the redescending Ψ functions. As their name implies, redescending Ψ functions relax the monotonicity criteria and introduce the possibility of multiple local minima in the objective function.

Unbounded Ψ Functions

For the class of unbounded Ψ functions $\Psi(x, \theta)$ tends to $\pm\infty$ as x and θ grow far apart. We state this formally in the following definition.

Definition 2.1 *An unbounded Ψ function is any monotonically increasing odd function of $x - \theta$ such that*

$$\lim_{(x-\theta)\uparrow\infty} \Psi(x, \theta) = \infty,$$

and

$$\lim_{(x-\theta)\downarrow-\infty} \Psi(x, \theta) = -\infty.$$

The most common unbounded Ψ function is the identity Ψ function which corresponds to the negative log likelihood loss function (Quadratic loss function) when the random variable is from a Gaussian distribution. The identity Ψ function has the following form and a plot of this Ψ function can be seen in Figure 2.2.

$$\begin{aligned} \Psi(x, \theta) &= \frac{\partial}{\partial\theta}[-\ln(\exp -\frac{1}{2}(x-\theta)^2)] \\ &= x - \theta. \end{aligned} \tag{2.4}$$

Unbounded Ψ functions work well when sampling iid random variables from a Gaussian distribution, because they place large penalties to points away from θ. However when the distribution is not Gaussian such as in the location version of the mean shift outlier model or heavy tailed distributions, the unbounded Ψ function penalizes extreme values to the extent that those extreme values strongly impact the optimal estimate of θ. Therefore unbounded Ψ functions have poor robustness properties in the presence of outliers. To remedy this a Ψ function must limit the penalty assigned to extreme data values. The following two classes of

Ψ functions were specifically designed to robustly estimate a location parameter by restricting the penalty assigned to extreme data values.

Figure 2.1: Quadratic ρ Function

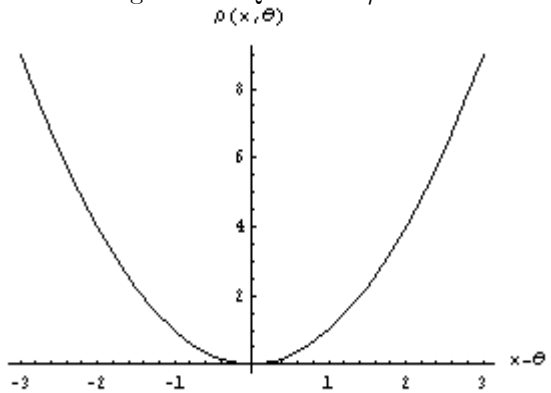


Figure 2.2: Identity Ψ Function

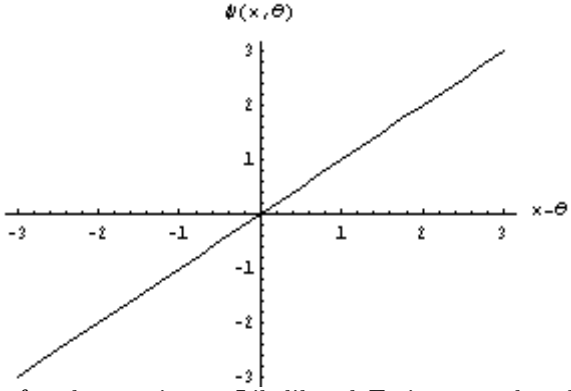


Figure 2.1 and Figure 2.2 plot the ρ and Ψ functions for the maximum Likelihood Estimator when likelihood is based upon the Gaussian distribution.

Limiting Ψ Functions

Limiting Ψ functions approach a constant bound as x and θ grow far apart. Thus we have the following formal definition.

Definition 2.2 *A limiting Ψ function is any monotonically increasing odd function of $x - \theta$ such that*

$$\lim_{(x-\theta)\uparrow\infty} \Psi(x, \theta) = c,$$

and

$$\lim_{(x-\theta)\downarrow-\infty} \Psi(x, \theta) = -c$$

for some positive constant c.

The following is an example of a very common limiting Ψ function.

Example 2.1 *One very important example of a limiting* Ψ *function is the Huber* Ψ *(Huber 1964) defined as*

$$\Psi_H(x,\theta) = \begin{cases} -c_H & : \quad x-\theta < -c_H \\ \frac{x-\theta}{\hat{s}} & : \quad |\frac{x-\theta}{\hat{s}}| \leq c_H \\ c_H & : \quad x-\theta > c_H \end{cases}, \tag{2.5}$$

where c_H *is an appropriately chosen constant and* $\hat{s}$ *is a scale factor. See Figure 2.3 and 2.4 for a plot of the Huber* ρ *and* Ψ *functions.*

Although this function simply takes the identity Ψ function for x close to θ and a constant elsewhere, it has many nice properties. One of the nicest properties of the Huber M-estimate is its robustness to outliers. Here robustness is defined in terms of the supremum of the asymptotic variance with respect to a bimodal mixture distribution such as described in 2.2.

Intuitively the Huber M-estimate and all limiting Ψ functions limit the impact that extreme data points have on the optimal θ estimate. It is important to realize that all extreme data points still have an impact on the optimal θ estimate, but much less than that of the unbounded Ψ functions.

For the Huber M-estimate we now have a method of quantifying what data points are "extreme". Any data point further than c_H units away from the θ estimate is considered extreme. The next logical question is how to choose the c_H. Often c_H is chosen to be 1.345 in order for the Huber M-estimate to maintain a 95% asymptotic relative efficiency vs. the sample mean when the X_i are iid Gaussian.

Since the limiting Ψ functions can control the influence an extreme outlier has on the θ estimate, one might ask if there exist Ψ functions that reduce or even eliminate the influence of outliers on the θ estimate. These Ψ functions do exist and will be the last class of Ψ functions we consider.

Redescending Ψ Functions

The most relevant class of Ψ functions for our research is the class of redescending Ψ functions. The definition of a redescending Ψ function is somewhat convoluted, but basically they follow the same properties of the

Figure 2.3: Huber ρ Function

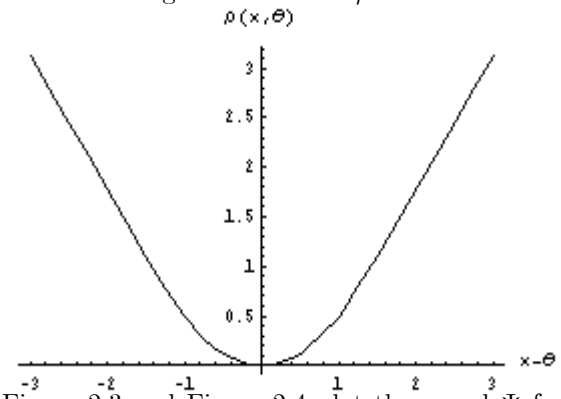


Figure 2.4: Huber Ψ Function

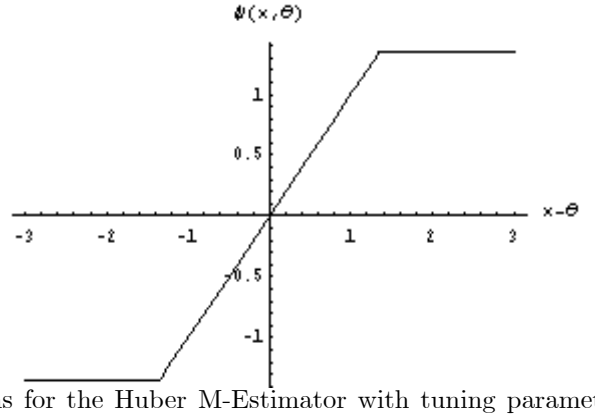


Figure 2.3 and Figure 2.4 plot the ρ and Ψ functions for the Huber M-Estimator with tuning parameter $c_H = 1.345$.

limiting Ψ functions for data points where x_i is close to θ and then "descends" to zero as x_i grows far away from θ. We state the formal definition below.

Definition 2.3 *A redescending Ψ function is any continuous odd function of $(x - \theta)$ that satisfies the following four criteria:*

i. *If $|x - \theta| \leq c$, $|x' - \theta| \leq c$, and $x < x'$ then $\Psi(x, \theta) \leq \Psi(x', \theta)$,*

ii. *If $x - \theta < -c$, $x' - \theta < -c$, and $x < x'$ then $\Psi(x, \theta) > \Psi(x', \theta)$,*

iii. *If $x - \theta > c$, $x' - \theta > c$, and $x < x'$ then $\Psi(x, \theta) > \Psi(x', \theta)$,*

iv. $\lim_{|x-\theta| \uparrow \infty} \Psi(x, \theta) = 0$,

where c is a positive constant.

Example 2.2 *The first example of a redescending Ψ function we wish to demonstrate mimics the Huber Ψ*

function. This Ψ function was first introduced by Hampel (1968) and has the following form.

$$\Psi(x,\theta) = \begin{cases} \frac{x-\theta}{\hat{s}} & : \quad 0 \leq |\frac{x-\theta}{\hat{s}}| < a, \\ a & : \quad a \leq |\frac{x-\theta}{\hat{s}}| < b, \\ a\frac{c-|\frac{x-\theta}{\hat{s}}|}{c-b} & : \quad b \leq |\frac{x-\theta}{\hat{s}}| < c, \\ 0 & : \quad |\frac{x-\theta}{\hat{s}}| \geq c, \end{cases} \tag{2.6}$$

where a, b, and c are positive constants and $\hat{s}$ is a scale factor. See Figure 2.5 and Figure 2.6 for a plot of the Hampel ρ and Ψ functions.

Figure 2.5: Hampel ρ Function

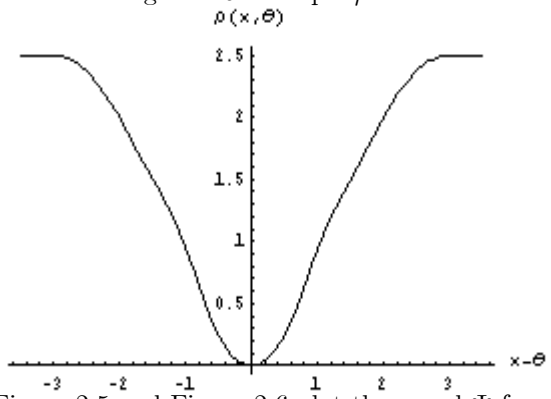


Figure 2.6: Hampel Ψ Function

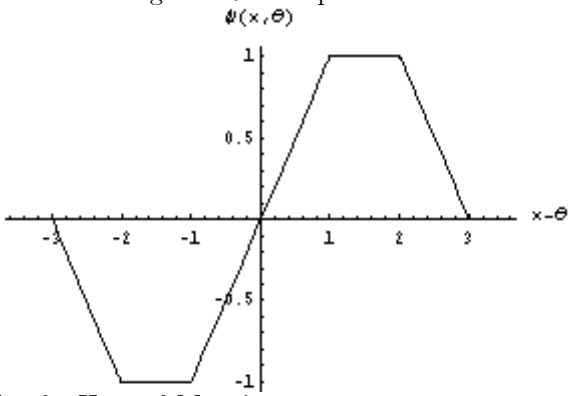


Figure 2.5 and Figure 2.6 plot the ρ and Ψ functions for the Hampel M-estimator.

Notice that for all $|\frac{x-\theta}{\hat{s}}| < b$ Hampel's Ψ function is equal to Huber's Ψ function if a is chosen to be c_H. Outside of that range the function redescends to zero.

It is undesirable to require the user to choose three tuning parameters in the Ψ function as does the Hampel Ψ function. Ψ functions that have continuous derivatives are easier to solve than those such as the Hampel Ψ function.

One Ψ function that has only one tuning parameter to choose and has a continuous derivative everywhere except at two points is the sine Ψ function by Andrews, Bickel, Hampel, Huber, Rogers & Tukey (1972). This Ψ function is based upon the sine function and is shown in the following example.

Example 2.3 *The sine Ψ function is defined by*

$$\Psi(x,\theta) = \begin{cases} c\sin(\frac{x-\theta}{c}) & : \quad |x-\theta| \le c\pi \\ 0 & : \quad |x-\theta| > c\pi \end{cases}, \tag{2.7}$$

where c is a positive constant. For a plot of the sine ρ and Ψ functions see Figure 2.7 and Figure 2.8.

Figure 2.7: Sine ρ Function

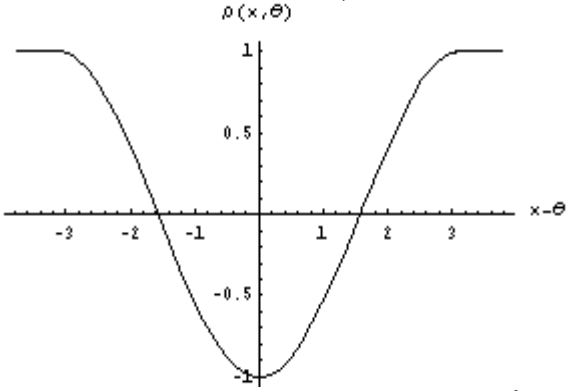


Figure 2.8: Sine Ψ Function

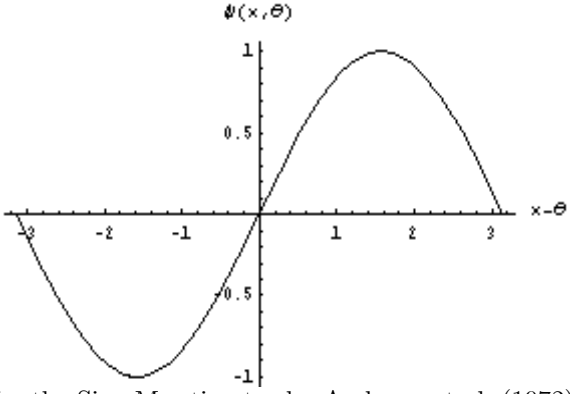


Figure 2.7 and Figure 2.8 plot the ρ and Ψ functions for the Sine M-estimator by Andrews et al. (1972).

Another popular redescending Ψ function by Beaton & Tukey (1974) is the "bisquare" Ψ function which has the following form.

Example 2.4

$$\Psi(x,\theta)_b = \begin{cases} (x-\theta)\big(1-\big(\frac{x-\theta}{c_b}\big)^2\big)^2 & : \quad |x-\theta| \le c_b \\ 0 & : \quad |x-\theta| > c_b \end{cases}, \tag{2.8}$$

where c_b is a positive constant. For a plot of the Bisquare ρ and Ψ functions see Figure 2.9 and 2.10.

The last example of a redescending Ψ function we wish to discuss has a continuous derivative everywhere. This Ψ function arrises when an negative Gaussian distribution is used for the loss function ρ and was mentioned in a comment by Simpson, He & Liu (1998).

Example 2.5 *So when we use the loss function*

$$\rho(x,\theta) = -\frac{1}{\sqrt{2\pi}\sigma}\exp\big[-\frac{1}{2}\big(\frac{x-\theta}{\sigma}\big)^2\big], \tag{2.9}$$

Figure 2.9: Bisquare ρ Function

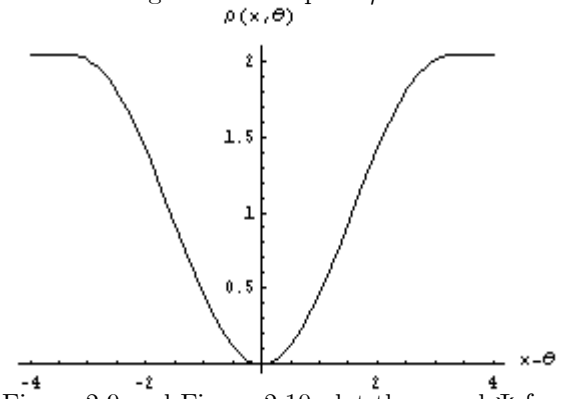


Figure 2.10: Bisquare Ψ Function

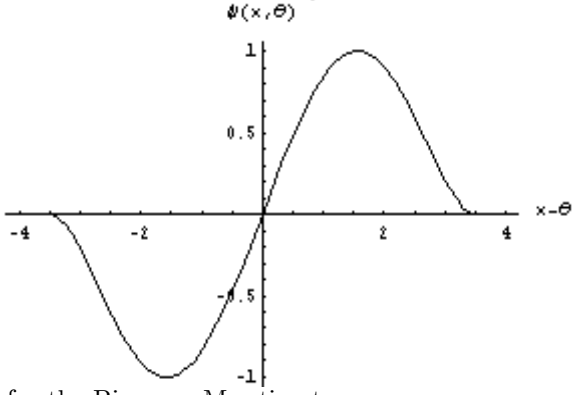


Figure 2.9 and Figure 2.10 plot the ρ and Ψ functions for the Bisquare M-estimator.

we generate the "Gaussian" Ψ function

$$\Psi(x,\theta) = -\frac{x-\theta}{\sqrt{2\pi\sigma^3}} \exp\left[-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2\right]. \tag{2.10}$$

See Figure 2.11 and Figure 2.12 for a plot of the Gaussian ρ and Ψ functions.

Figure 2.11: Gaussian ρ Function

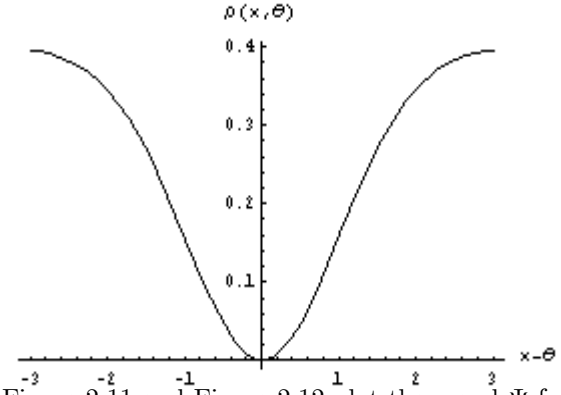


Figure 2.12: Gaussian Ψ Function

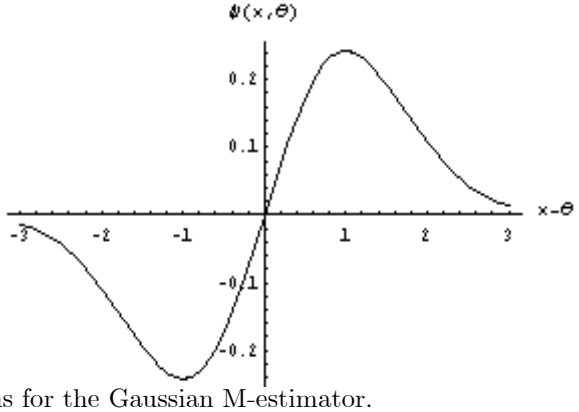


Figure 2.11 and Figure 2.12 plot the ρ and Ψ functions for the Gaussian M-estimator.

The Gaussian Ψ is much more convienient than the previously mentioned Ψ functions because we can write it in closed form without the use of an indictor function. This saves a few steps in programming an iterative method to minimize the objective function.

When the Gaussian Ψ is used to estimate a location parameter in M-estimation there is an analogy to univariate kernel density estimation with Gaussian kernel. Thus the tuning parameter σ could be chosen based on existing data driven methods of selecting a bandwidth. Some of these methods are explained in

detail in Silverman (1986). Although this analogy between kernel density estimation and M-estimation has been known for some time it was first referenced in Rue, Chu, Godtliebsen & Marron (1998). The Gaussian Ψ is not the only Ψ function based on a kernel function. There is also a commonly used kernel function that gives rise to the Bisquare Ψ. It can be shown that when inverted kernel functions are used for the ρ function the corresponding Ψ functions are redescending Ψ functions.

There has been much controversy concerning why one should use a redescending Ψ function because it has been shown that the most robust estimator is the Huber M-estimator. Intuitively, it does seems reasonable that the extreme values should maintain some influence on the optimal parameter estimate. Additionally redescending Ψ functions introduce the problem of multiple roots in the defining equation when the underlying distribution is a bimodal mixture distribution such as described in 2.2. Two of the roots will correspond to the two modes and one will correspond to the center of the data. Usually in M-estimation we seek the root corresponding to the center of the data, but many iterative procedures have difficulty finding the correct root.

These disadvantages in usual M-estimation become advantages in estimating jump-point discontinuities. Estimating the jump point is analogous to estimating the two modes of the bimodal mixture distribution mentioned above. Thus the multiple root problem of redescending Ψ functions is the tool we propose using to test for jump point discontinuities.

2.1.2 Iterative Algorithms

For the identity Ψ function there is a closed form solution to 2.3 (the defining equation) that can easily be computed. For most other Ψ functions, including all other examples previously mentioned, there is no closed form solution to the defining equation. Therefore an iterative method must be employed to find most M-estimates. Two iteration methods commonly used include the Newton-Raphson method and Iterated Reweighted Least Squares (IRLS).

Newton-Raphson

Newton-Raphson or Newton's method was first used to solve a cubic polynomial in a paper by Newton in 1687 in the Principia. Since then many modifications have been made to enhance the ability of Newton's method to solve for the root of a nonlinear equation. It can also be shown that Newton's method has a quadratic rate of convergence (Birch 1997), the fastest possible rate of convergence. Newton's method is based on the near linear local properties of nonlinear functions and their first order Taylor series approximation. A detailed discussion of Newton's method applied to M-estimators in the location parameter setting can be found in Birch (1997) and is outlined below.

1. Choose a starting value for the location parameter θ, say $\hat{\theta}_0$.

2. Use the $\hat{\theta}_0$ to estimate scaled residuals r_i^*, where

$$r_i^* = \frac{x_i - \hat{\theta}_0}{\hat{s}}.$$

3. Calculate the updated estimate of θ by

$$\hat{\theta}_{k+1} = \hat{\theta_k} + \hat{s}\frac{\sum_{i=1}^{n} \Psi(r_i^*)}{\sum_{i=1}^{n} \Psi'(r_i^*)},$$

   where Ψ' is the derivative

$$\Psi'(r) = \frac{\partial}{\partial r}\Psi(r).$$

4. Use the updated $\hat{\theta}_{k+1}$ to calculate scaled residuals and repeat the process until convergence.

Note that the $\hat{s}$ above is an appropriately chosen estimate of the scale parameter.

IRLS

Iterated Reweighted Least Squares (IRLS), sometimes called Iterated Weighted Least Squares (IWLS), is based upon the solution to the "normal equations" in Ordinary Least Squares (OLS) estimation. The idea

is that if we can obtain a "normal equation" type of solution for θ in terms of a function of θ, then we can use this function to update an initial estimate until convergence. A more detailed description of the (IRLS) algorithm as it applies to M-estimation and M-smoothing can be seen in Birch (1997) and Simpson et al. (1998) and is outlined below.

1. Choose a starting value for the location parameter θ, say $\hat{\theta}_0$.

2. Use the $\hat{\theta}_0$ to estimate scaled residuals r_i^*, where

$$r_i^* = \frac{x_i - \hat{\theta}_0}{\hat{s}}.$$

3. Calculate the updated estimate of θ by

$$\hat{\theta}_{k+1} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i},$$

   where

$$w_i = \frac{\Psi(r_i^*)}{r_i^*}.$$

4. Use the updated $\hat{\theta}_{k+1}$ to calculate scaled residuals and repeat the process until convergence.

Implementing IRLS in the M-smoothing regression setting is straightforward because M-smoothing is associated with solving a type of "normal equation" where the weight function incorporates the Ψ weight and the local weight together. For the standard choice of starting values for M-estimation see Birch (1997). The modified M-smooth estimate uses untraditional starting values as will be explained in Chapter 2.4.

2.2 Smoothing Review

The term smoothing refers to an assortment of methods used to visually summarize a scatter plot of paired (x, y) data. Some of these methods include Kernel Estimators, Spline Smoothing, Wavelet Thresholding, Locally Weighted Scatter Plot Smoothing (LOWESS), and Local Polynomial Fitting. Although M-smoothing

could be included in this list we postpone our discussion of M-smoothing for the next section. We will briefly describe and reference each of the above methods with an added emphasis on the local polynomial fitting method since it is most relevant to our research. For a more detailed survey of the above methods see Chapter 2 of Fan & Gijbels (1996).

Two of the most popular kernel estimators include the Nadaraya-Watson estimator (Nadaraya 1964 and Watson 1964) and the Gasser-Müller estimator (Gasser & Müller 1979 and Gasser & Müller 1984). These two kernel estimators are very similar and both use a kernel weighting function to obtain a weighted local average. The LOWESS method proposed by Cleveland (1979) provides a robust scatter plot smooth by down weighting the residuals of a local polynomial regression. The LOWESS method is similar in design and performance to the traditional M-smoothing method of the next section. Wavelet Thresholding is a smoothing method related to Fourier series type approximations. For an introduction to wavelet-based methods see Chui (1992), Daubechies (1992), and Strang (1989). For a more extensive overview of wavelet-based methods see Donoho (1995). Spline smoothing is based on fitting piecewise polynomials to the scatter plot data. For a more detailed discussion on spline smoothing see Eubank (1988), Wahba (1990), and Green & Silverman (1994).

2.2.1 Local Polynomial Regression

Local polynomial regression is similar to kernel regression in that they both use a kernel weight function to locally estimate the mean function. However, local polynomial regression fits a least squares polynomial locally rather than a weighted average. Thus the usual regression setup is as follows: The Y_i's $i = 1, 2, \cdots, n$ are modeled as

$$Y_i = m(X_i) + \varepsilon_i$$

where ε_i are iid random errors from a unimodal symmetric density centered about 0 and the $m(x)$ is a continuous mean function with continuous derivative. We then write

$$\mathbf{X}_i = \begin{pmatrix} 1 & (X_1 - X_i) & \cdots & (X_1 - X_i)^p \\ 1 & (X_2 - X_i) & \cdots & (X_2 - X_i)^p \\ & \vdots & & \\ 1 & (X_n - X_i) & \cdots & (X_n - X_i)^p \end{pmatrix}, \mathbf{y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \text{and } \mathbf{W}_i = \begin{pmatrix} w_{i,1} & & & \\ & w_{i,2} & & \\ & & \ddots & \\ & & & w_{i,n} \end{pmatrix}.$$

Here $w_{i,j} = \frac{1}{h}K(\frac{X_j - X_i}{h})$ where K is a kernel function (usually a symmetric unimodal density function) and h is a positive constant referred to as a bandwidth. Then the local polynomial regression estimate of the conditional mean function $\hat{m}(X_i)$ is the first element of the $\hat{\beta}_i$ vector given by

$$\hat{\beta}_i = (\mathbf{X}_i'\mathbf{W}_i\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{W}_i\mathbf{y}.$$

Sometimes the w_{ij} above is calculated as

$$w_{ij} = \frac{\frac{1}{h}K(\frac{X_j - X_i}{h})}{\frac{1}{h}\sum_{j=1}^{n} K(\frac{X_j - X_i}{h})}$$

so that the sum of the w_{ij} will be constrained to be 1. However, for the purpose of solving for β above, any constant multiple within the calculation of the w_{ij}'s will cancel. Therefore to simplify all computer programming statements we have dropped the $\sum K(\frac{X_j - X_i}{h})$ above in all local polynomial regression calculations. In addition we have dropped the constant $\frac{1}{\sqrt{2\pi}}$ when using the Gaussian kernel K.

It can be shown that kernel regression is a special case of local polynomial regression where the degree of the polynomial is zero. Some of the advantages of using local polynomial regression over the other previously mentioned methods can be found in Chapter 3 of Fan & Gijbels (1996). Asymptotic variance and asymptotic bias formulas for the local polynomial regression can also be found in Chapter 3 of Fan & Gijbels (1996).

One of the biggest setbacks of local polynomial regression and all other smoothing procedures is that the estimated conditional mean function is quite sensitive to the choice of the smoothing parameter(s). This problem is compounded by the fact that there is no one universally accepted simple method of choosing the bandwidth. The current methods available for choosing the bandwidth in the local polynomial regression

setting range from crude and simple to sophisticated and complex. To help motivate the novel (based on modifications of existing methods) bandwidth selection procedure for the M-smoother developed in Chapter 3 we provide a brief survey of some of the most commonly used automatic data driven bandwidth selectors for local polynomial regression below. We focus our discussion of bandwidth selection to four classes: Rule of Thumb, Residual Squares Criterion, Direct Plug In, and Solve the Equation. Overshadowing each of the above classes is the question of whether the bandwidth chosen should be the same for all grid points (constant bandwidth), or be allowed to change from one grid point to the next (variable bandwidth).

Rule of Thumb Bandwidth Selection

The Rule Of Thumb (ROT) class of bandwidth selectors can also be thought of as the first generation Direct Plug In (DPI) methods, because the idea is to plug in estimates of unknown quantities into the optimal asymptotic bandwidth formula. The optimal bandwidth formula can be found by minimizing the Integrated Asymptotic Mean Squared Error (IAMSE). This derivation can be found in Fan & Gijbels (1996). The ambiguous part of this process is what method to use to estimate the unknown quantities, which include the variance, certain derivative functions, and the density function.

Fan & Gijbels (1996) suggest fitting a fourth degree polynomial by the method of ordinary least squares and using the derivatives of this polynomial and the standardized residual sum of squares as estimates to their unknown analogs. Ruppert et al. (1995) suggest using a similar technique except that their polynomial is fit in blocks. The number of blocks is chosen based on Mallows' C_p statistic (Mallows 1973).

It is important to note that the rule of thumb bandwidth selectors assume a constant variance, due to a single variance estimate, and thus are not available (without modification) in the variable bandwidth type.

Residual Squares Criterion Bandwidth Selection

The Residual Squares Criterion (RSC) class of bandwidth selectors encompasses a large group of distinct methods. This class in general attempts to select a bandwidth that minimizes a multiple of the residual sum of squares defined by

$$p(h) = n^{-1} \sum_{i=1}^{n} [Y_i - \hat{m}_h(x_i)]^2 w(x_i).$$

A multiple is tacked on to preserve certain asymptotic properties of the bandwidth selector and is a function of $n^{-1}h^{-1}$, which we shall refer to as $\Xi(n^{-1}h^{-1})$. Numerous quantities have been suggested for this multiple, which lead to numerous methods.

Craven & Wahba (1979) suggest using

$$\Xi(n^{-1}h^{-1}) = (1 - n^{-1}h^{-1}K(0))^{-2},$$

which leads to the generalized cross-validation method. Another cross-validation method by Clark (1975) uses the multiple

$$\Xi(n^{-1}h^{-1}) = 1 + 2n^{-1}h^{-1}K(0) + O_p(n^{-2}h^{-2}).$$

Akaike (1970) and Akaike (1974) propose two distinct multiples

$$\Xi(n^{-1}h^{-1}) = exp(2n^{-1}h^{-1}K(0)),$$

and

$$\Xi(n^{-1}h^{-1}) = \frac{1 + n^{-1}h^{-1}K(0)}{1 - n^{-1}h^{-1}K(0)}.$$

Rice (1984) suggest the multiple

$$\Xi(n^{-1}h^{-1}) = (1 - 2n^{-1}h^{-1}K(0))^{-1}.$$

Härdle, Hall & Marron (1988) show that all these RSC methods possess the same asymptotic properties, but they demonstrate through simulation the small sample differences of each method. Although it appears

that Rice's (1984) RSC method is preferred, there may exist special situations where other RSC methods are superior.

Generally these RSC methods produce a single bandwidth which is used as a constant bandwidth for all grid points, but Fan & Gijbels (1996) provide an algorithm based on k subintervals that adapts the residual squares criterion to yield a variable bandwidth.

Direct Plug In Bandwidth Selection

The Direct Plug In (DPI) method of bandwidth selection accomplishes basically the same idea as the ROT selectors, namely that they substitute estimates for unknown quantities into the optimal bandwidth formula based on the AIMSE. The primary difference in the next generation DPI methods and ROT methods is that the DPI methods use smooth functional estimates to estimate the unknowns. The problem here is that the smooth functional estimates also require some type of pilot bandwidth. This pilot bandwidth can be chosen by a ROT method such as Ruppert et al. (1995) suggest or by a RSC method and described in Fan & Gijbels (1996).

We include two important quotes here to tie the previous three classes (ROT, RSC, and DPI) together. Härdle et al. (1988) note that the RSC bandwidths have the same rate of convergence as the optimal bandwidth which the DPI is based on even if the unknown quantities were known. They go on to say, "Hence the additional noise involved in estimating these unknown parts in practice, especially the second derivative part in the case where m is not very smooth, seems to cast considerable doubt on the applicability of the plug-in estimator" (Härdle et al. 1988). It is assumed that the DPI bandwidth selectors referenced here are first generation DPI or ROT selectors. In their section of refined bandwidth selection where (Fan & Gijbels 1996, page) develop their next generation DPI they say, "In order to calculate this MSE we need however a pilot bandwidth and this will be obtained via the RSC. The resulting more sophisticated estimation procedures clearly outperform the ones relying only on RSC." (Fan & Gijbels 1996). Thus the controversy between the DPI and RSC methods is not yet resolved.

Solve the Equation Bandwidth Selection

The Solve The Equation (STE) class of methods makes use of the fact that when a smooth functional is used to estimate the unknowns, the optimal bandwidth formula used in the DPI can essentially be written as an equation involving the bandwidth. Thus there is a bandwidth on both sides of the equation. The idea is thus to solve the equation for the bandwidth. Obviously there is no closed form solution and an iterative solving technique must be employed.

The STE methods, although more sophisticated, are computationally cumbersome and difficult to program. It is speculated based on the results of Ruppert et al. (1995) that the additional accuracy of the STE method over the DPI method will be minimal compared to the added computational and programming difficulty. Ruppert et al. (1995) suggest that variable bandwidth selector adaptations to their constant bandwidth STE method are straightforward, but they do not provide any details.

It has been the experience (howbeit limited) of the author that the constant DPI bandwidth selection procedure of Ruppert et al. (1995) is as good as any of the above mentioned methods. Because of this and various logistical limitations of RSC methods in the jump preserving M-smoother case, we have chosen to use a modified version of Ruppert et al.'s (1995) method for our current research purposes. These modifications will be discussed in Chapter 3.

2.3 Traditional M-Smoothers

M-estimation and nonparametric smoothing were first combined by Härdle & Gasser (1984) for the purpose of robustly fitting an unknown regression mean function in the presence of possible outliers in the errors. The possible outliers were assumed to originate from a heavy tailed error distribution and not the mean shift outlier model (2.2). The local regression estimate at x by Härdle & Gasser (1984) is defined to be the zero

of $H_n(x,.)$, where

$$H_n(x,.) = \sum_{i=1}^{n} \alpha_i(x)\Psi(Y_i - .).$$

Here the $\alpha_i(x)$ are weights that incorporate a kernel function and a bandwidth. Thus the kernel function does the local averaging and the Ψ function down weights outliers. Härdle & Gasser (1984) restrict the Ψ function to the class of limiting Ψ functions for the proof of consistency, but mention that if the proper root is selected their method is consistent when redescending Ψ functions are used.

Figure 2.13: Härdle's M-smooth Fit to The Broken Sine Data

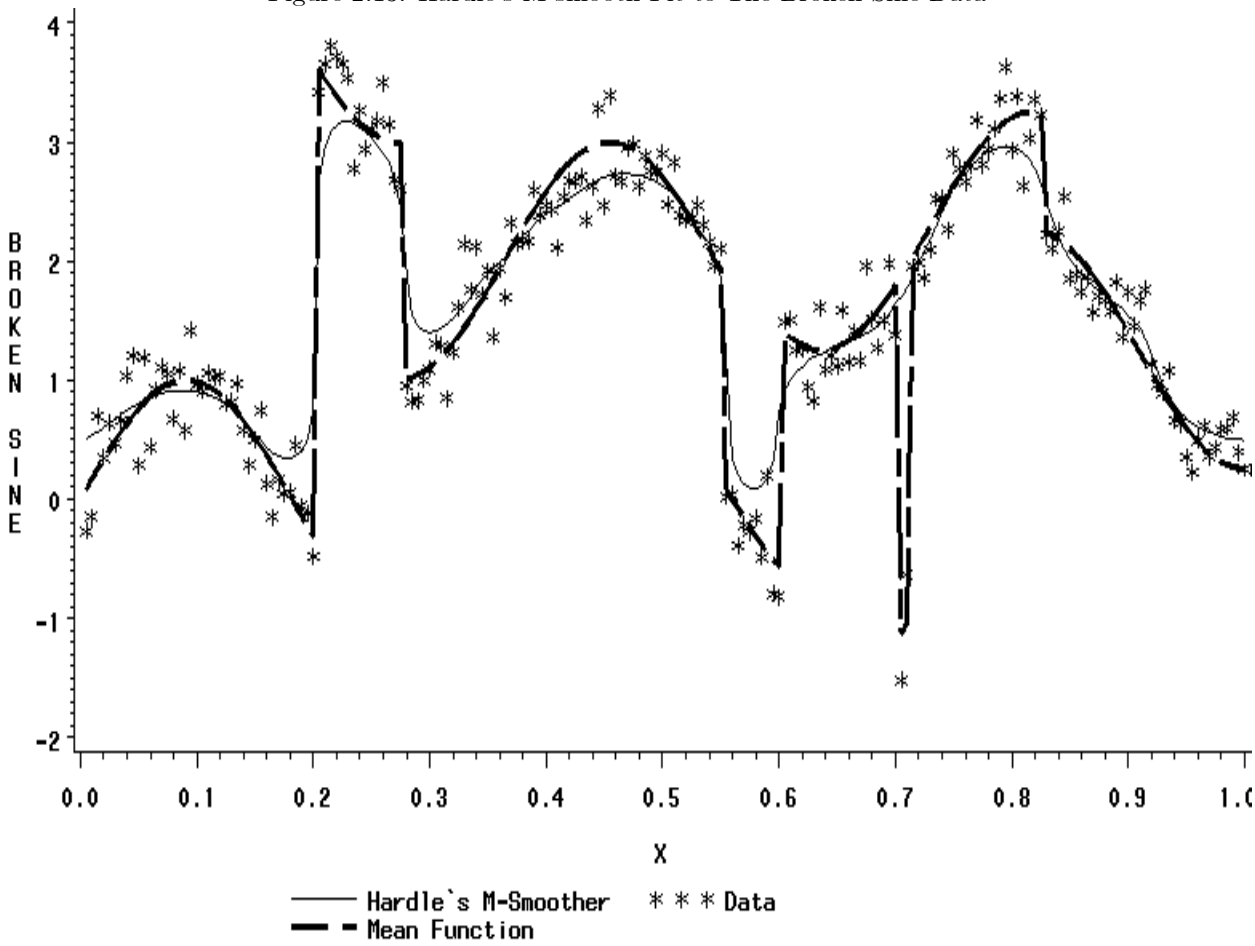


Figure 2.13 shows a simulated data set about the mean function $\sin(5.5\pi x)$ with jump points added at positions $x = 0.2, 0.275, 0.55, 0.6, 0.7, 0.71$ and 0.825 of sizes $4, -2, -1.75, 2, -3, 3$ and -1.75 respectively. The error distribution used to simulate the data about this mean function was Gaussian with mean zero and variance $\sigma^2 = (0.25)^2$. Härdle & Gasser's (1984) M-smooth regression was fit to the simulated data using the Huber Ψ function and the Gaussian kernel function. The bandwidth associated with the Gaussian kernel function was fixed at $h = 0.03$ and the tuning constant associated with the Huber Ψ was chosen to be $C_H = 1.345$.

Since the purpose of Härdle & Gasser's (1984) M-smoother is to be robust against outliers it accomplishes the opposite effect we desire in the presence of a jump-point discontinuity. To demonstrate this we have fit Härdle & Gasser's (1984) M-smooth to the broken sine simulated data first described in Section 1.1. This fit is shown in Figure 2.13. Two successive jumps of equal magnitude in opposite directions within a small segment of the design space will often mimic one outlier or a small group of outliers. We shall refer to this type of jump scenario as an impulse jump. Such is the jump scenario seen in Figure 2.13 between $x = 0.7$ and $x = 0.71$. For such jump scenarios Härdle & Gasser's (1984) M-smoother will ignore the data between the two successive jumps as if they were outliers. Thus the jump information is essentially washed out of the regression fit. For other jump points that are somewhat isolated from their neighboring jump points (such as demonstrated in Figure 2.13 at position $x = 0.825$) Härdle & Gasser's (1984) M-smoother will experience the same over smoothing problem as does local linear regression.

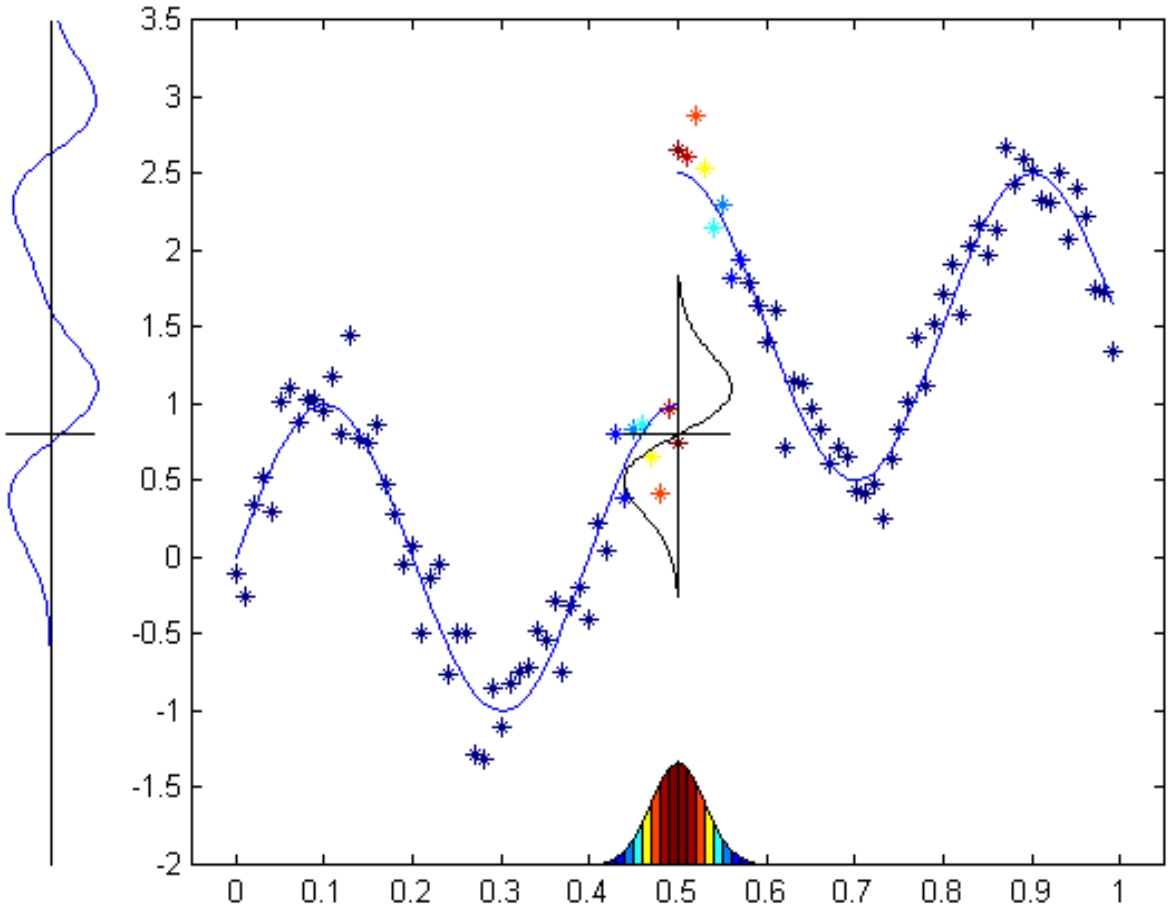
Leung, Marriott & Wu (1993) provide a more extensive list of M-smoothers that accomplish the same purpose as Härdle & Gasser's (1984) M-smoother along with a comparison of different kernel bandwidth selection methods.

M-smoothers that model jump-point discontinuities are a relatively recent advance in the literature and will be the focus of the next section.

2.4 The Jump Preserving M-smoother

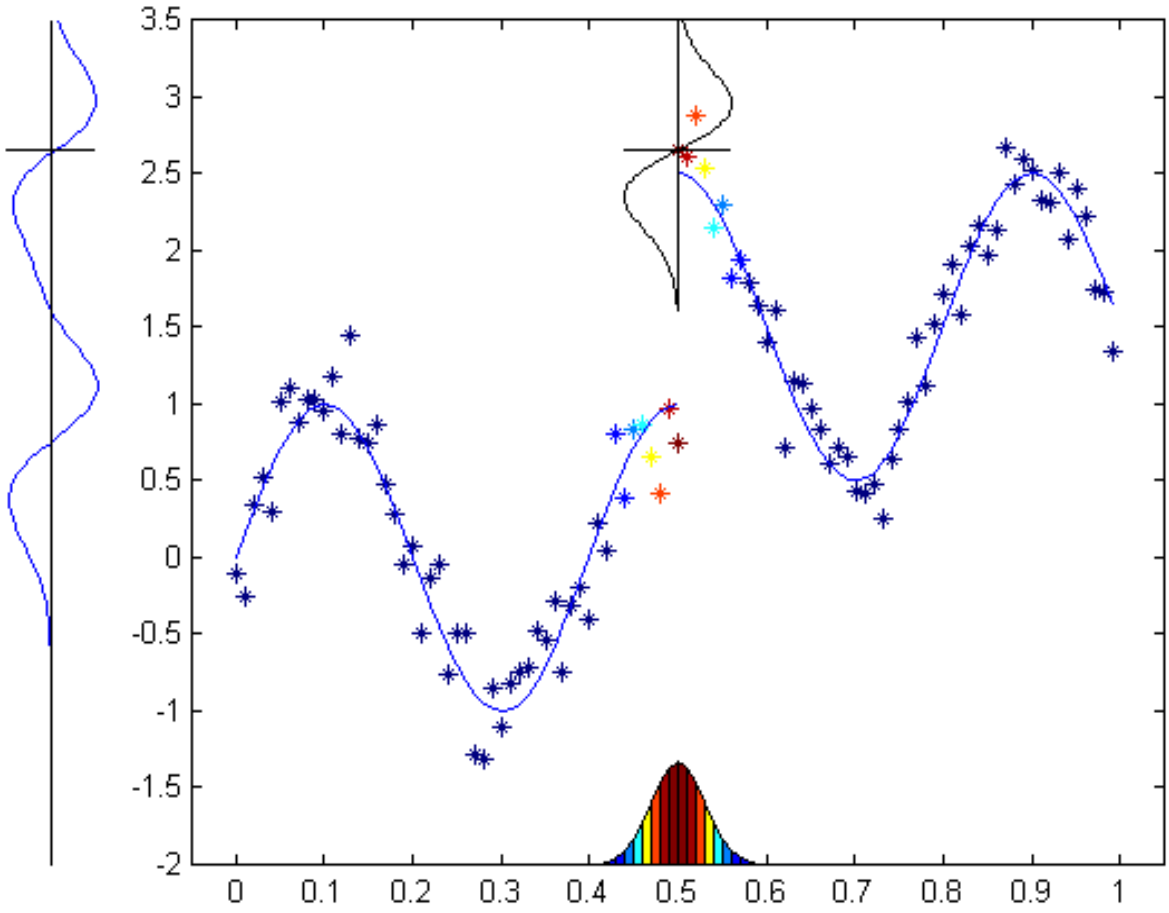
Chu et al. (1998) developed a local constant version of a M-smoother based on a modification of Härdle & Gasser's (1984) M-smoother that smooths regression data while preserving edges or jumps. In an unpublished article Rue et al. (1998) provide the local linear version of this M-smoother. The M-smoother of Chu et al. (1998) capitalizes on the multiple root dilemma encountered when using a redescending Ψ function. Using a redescending Ψ function Chu et al.'s (1998) M-smoother searches for the root of the defining equation closest to the raw response data point Y_i. Figures 2.14 and 2.15 display the relationship between the redescending Ψ function, the kernel weight function, and the multiple roots used to preserve jump point discontinuities. Both Figure 2.14 and 2.15 contain two graphical displays. The graph on the right of each figure displays the simulated data set, the mean function, the Gaussian kernel function, and the Ψ weight function. The Gaussian kernel function is plotted along the x (horizontal) axis. The simulated mean function is $sin(5\pi x) + (1.5)I(x > 0.5)$. The error distribution used to simulate data around this mean function was Gaussian with mean zero and variance $\sigma^2 = (0.25)^2$. The data (plotted as stars) are color coded according to the Gaussian kernel function to reflect their corresponding weight in the local averaging. The Ψ weight function is plotted vertically in the center of the graph. The graphical display on the left of Figure 2.14 and Figure 2.15 show the function whose roots define Chu et al.'s (1998) M-smooth fit. As we pick the root closest to the raw response data point for $(x = 0.5, y = 0.729)$ the fit will be near the mean function before the jump (see Figure 2.14). Likewise, as we pick the root closest to the raw response data point for $(x = 0.5005, y = 0.267)$ the fit will be near the mean function after the jump (see Figure 2.15). In this manner Chu et al.'s (1998) M-smoother preserves jumps. This method has been shown to have the extraordinary ability to estimate mean functions with many of the impulse type jumps described in Section 2.3. For a demonstration of the M-smoother's ability to fit a mean function with many impulse jumps see Chu et al. (1998).

Figure 2.14: Demonstration of Chu's Jump Preserving M-smoother



The graphical display on the right of Figure 2.14 shows a simulated data set about the mean function $\sin(5\pi x)$ with one jump point added at position $x = 0.5$ of size 1.5. The error distribution used to simulate the data about this mean function was Gaussian with mean zero and variance $\sigma^2 = (0.25)^2$. Along the x axis centered at $x = 0.5$ is plotted the Gaussian kernel function used by the M-smoother. Additionally the data are color coded according to the kernel function to reflect their corresponding weight in the local averaging (at $x = 0.5$). The Ψ weight function is plotted vertically in the center of the graph. The graphical display on the left of Figure 2.14 shows the function whose roots determine the M-smooth fit.

Figure 2.15: Demonstration of Chu's Jump Preserving M-smoother



The graphical display on the right of Figure 2.15 shows a simulated data set about the mean function $\sin(5\pi x)$ with one jump point added at position $x = 0.5005$ of size 1.5. The error distribution used to simulate the data about this mean function was Gaussian with mean zero and variance $\sigma^2 = (0.25)^2$. Along the x axis centered at $x = 0.5$ is plotted the Gaussian kernel function used by the M-smoother. Additionally the data are color coded according to the kernel function to reflect their corresponding weight in the local averaging (at $x = 0.5005$). The Ψ weight function is plotted vertically in the center of the graph. The graphical display on the left of Figure 2.15 shows the function whose roots determine the M-smooth fit.

We now give a few of the specific mathematics behind the M-smoother. Recall from Section 2 that the regression setting we wish to model is given by

$$Y_j = m(x_j) + \epsilon_j$$

for $j = 1, 2, ..n$. For simplicity we also assume that we have an equally spaced fixed design with ϵ_j iid random variables with zero mean and finite variance. Note that the equally spaced assumption is not required for most of the theoretical calculations, but there are programming concerns that this assumption greatly simplifies. Two of the most important concerns that must be addressed before the equal spaced fixed design can be relaxed are: (1) The possibility of obtaining two responses at the same regressor value. This would pose a problem with fitting the root closest to the raw response data point since there would then be two to chose from. (2) The design space should be sufficiently dense. It is not known how dense the design space must be in order to implement the jump preserving M-smoother, but certain matrices required to calculate the estimate will be singular when the data are sparse. Assuming an equally spaced fixed design eradicates the previous two concerns. Thus throughout this dissertation an equally spaced fixed design will be assumed. The mean function $m(x_j)$ above is of the form

$$m(x) = \mu(x) + d_k \cdot I(x > t_k)$$

for $k = 1, 2, ...T$, where T is the total number of jump point discontinuities, t_k is the position of the k^{th} jump point, and d_k is the corresponding size of the k^{th} jump point. The local constant M-Smooth estimate of $m(x_i)$ is then found by taking the local minimizer (with respect to a) of

$$S(a, x_i) = (-1) \sum_{j=1}^{n} \frac{1}{h} K(\frac{x_i - x_j}{h}) \frac{1}{g} L(\frac{Y_j - a}{g}) \tag{2.11}$$

that is closest to Y_i. Similarly the local linear M-Smooth estimate of $m(x_i)$ is then found by taking the local minimizer (with respect to a and b) of

$$S(a, b, x_i) = (-1) \sum_{j=1}^{n} \frac{1}{h} K(\frac{x_i - x_j}{h}) \frac{1}{g} L(\frac{Y_j - a - b(x_i - x_j)}{g}) \tag{2.12}$$

that has the closest a to Y_i, see Rue et al. (1998) for further details. Here L and K are kernel functions and h and g are two bandwidths. Using the kernel function L (along with the negative in front of the $\sum$)

is analogous to using a redescending Ψ function as was shown in Section 2.1.1. We also suggest using the Gaussian distribution for both kernel functions L and K. This is equivalent to using the Gaussian Ψ function defined in Section 2.1.1. Choosing L to be Gaussian may be somewhat less common in M-smoothing than other choices such as the loss function that gives rise to the Bisquare Ψ. However, L will need to be chosen as the Gaussian distribution in order to satisfy certain theoretical concerns of our hypothesis test to be presented in Chapter 4.

$S(a, x_i)$ and $S(a, b, x_i)$ above are too complicated to find a closed form for their local minimum, so numerical algorithms must be implemented. Chu et al. (1998) provide sophisticated modifications to Newton's iterative method to find the local minimum closest to Y_i. Rue et al. (1998) suggest using a method that considers all possible roots and then chooses the correct one based on an iterative algorithm. These methods, however, are complicated to program and do not generalize to higher order polynomials. For the local constant M-Smoother Simpson et al. (1998) suggest using Iterated Reweighted Least Squares with Y_i as the starting value. Although there is no guarantee that the IRLS algorithm will converge to the correct local minimum, Simpson et al. (1998) found that using reasonable bandwidths and the data as starting values resulted in correct convergence. Unlike previously implemented solving algorithms for the jump preserving M-smoother the IRLS algorithm can easily be extended to higher order polynomials. The IRLS iterative algorithm for Chu et al.'s (1998) M-smoother is explained in detail in the next chapter.

Another important issue required to implement the jump preserving M-smoother is the choice of the two bandwidths g and h. Due to the impact this choice has on both the regression fit and our proposed hypothesis test, a majority of our current research has been devoted to developing an automatic data driven selection procedure. Thus the following chapter is devoted to the detailed description of our bandwidth selection procedure for the jump preserving M-smoother.

Chapter 3

Automatic Bandwidth Selection for the Jump Preserving M-Smoother

Generally nonparametric regression techniques mistake background noise for structure in the data when the smoothing parameter is chosen too small. Likewise when the smoothing parameter is chosen too large the underlying structure in the data is smoothed out. Chu et al.'s (1998) M-smoother differs from most nonparametric regression techniques in that it requires two smoothing parameters or bandwidths instead of only one. In addition, traditional nonparametric regression methods find structure in the data as bumps, whereas the jump preserving M-smoother finds structure in bumps and/or jumps. Thus developing an automatic bandwidth selection procedure for the two bandwidths of Chu et al.'s (1998) M-smoother is a more difficult problem than developing bandwidth selection procedures for traditional nonparametric regression methods.

Currently there are no data driven automatic bandwidth selection procedures available for choosing the h and g required for the jump preserving M-smoother. All examples in Chu et al. (1998) employ fixed bandwidths chosen by the authors, but in a later paper they suggest using a simple cross-validation method

(Rue et al. 1998). In a comment Simpson et al. (1998) suggest using an L_1 norm method over an L_2 norm in the cross-validation procedure. However, no complete cross validation algorithm has been developed for Chu et al.'s (1998) M-smoother.

Although Rue et al. (1998) and Simpson et al. (1998) have suggested applying cross validation methods to the jump preserving M-smoother of Chu et al. (1998), there are many practical difficulties that have not been solved. These difficulties are listed below.

1. Recall that the local linear jump preserving M-smooth fit at a given x_i is defined as the minimizer of (2.12) closest to the regressor data point Y_i. Most cross validation methods fit the $m(x_i)$ without using the information from the point (x_i, Y_i). The jump preserving M-smoother requires the data point (x_i, Y_i) to determine which root to choose when estimating $m(x_i)$. Thus it is unclear how to calculate the one point removed version of the jump preserving M-smoother.

2. The usual cross validation method does not produce polynomial coefficient estimates which are needed for starting values in the IRLS algorithm. Therefore any cross validation method would require an additional estimation procedure if the IRLS algorithm is to be employed.

3. Finally the cross validation method requires a candidate list of bandwidths to compare. Currently there is no defined method of choosing a candidate list.

Instead of addressing each of the above difficulties we chose to adapt the methods of Ruppert et al. (1995) to the jump preserving M-smoother setting. In Section 2.2.1 we described a number of different automatic bandwidth selection procedures for the local polynomial regression estimator (assuming continuity). The ROT is the simplest of these methods. We propose a similar method, which is soon to appear in the article Burt & Coakley (2000), that provides crude bandwidth estimates for the g and h in the jump preserving M-smoother. Our method shall be referred to as the Robust Rule of Thumb (RROT) method. The RROT requires optimal bandwidth formulas for both h and g as well as a robust method of estimating the unknowns in these formulas. This procedure is outlined in Section 3.3 and evaluated by a simulation

study in Section 3.4.

The way that the RROT method estimates the unknowns in the optimal bandwidth formulas is quite crude. It is likely that using the jump preserving M-smoother to estimate these quantities would result in a much better bandwidth selection procedure. However, to do this would require pilot bandwidths for g and h. In Section 3.3 we outline how the RROT bandwidths can be used as pilot bandwidths to produce a more sophistocated DPI bandwidth selection procedure for Chu et al.'s (1998) M-smoother. Our DPI method is modeled after the method of Ruppert et al. (1995) described in Section 2.2.1.

3.1 Robust Rule of Thumb Bandwidth Selection

As we alluded to earlier, ROT bandwidth selectors work by taking asymptotically optimal bandwidth formulas and substituting crude estimates for the unknown parameters in the formulas. The optimal bandwidth formula is usually based upon minimizing a global loss function made up of bias and variance pieces.

For our proposed RROT method a global loss function is defined and minimized with respect to the bandwidth h in Section 3.1.1. An optimal bandwidth formula for g is also discussed in Section 3.1.1. Crude estimates to the unknowns in the bandwidth formulas are obtained via a modified version of Härdle & Marron's (1995) Fast and Simple Scatterplot Smoothing method. Härdle & Marron's (1995) method must be altered because of the sporadic behavior of the scatter plot smooth when jump points are present. Section 3.1.2 provides the details behind our modifications that make the estimates of the unknown quantities robust to jump points.

3.1.1 Asymptotic Optimal Bandwidth Formula

Our strategy for developing optimal bandwidth formulas for h and g centers around the role each one plays in the M-smoother. Since h is similar to a nonparametric regression bandwidth the formula for $\mathsf{h}_{\mathsf{opt}}$ is found

by minimizing a global loss function made up of bias and variance components. Since g acts as a combination of scale parameter and tuning constant in usual M-estimation we set $\mathsf{g}_{\mathsf{opt}} = \sigma\alpha$ where σ is a measure of scale and α is determined by an Asymptotic Relative Efficency (ARE) calculation. To calculate an ARE we return to the location parameter version of the M-estimate. We therefore use the following theorem to determine the formula for $\mathsf{g}_{\mathsf{opt}}$.

Theorem 3.1 *For the* Ψ *function*

$$\Psi(r) = \frac{-r}{g} \exp \frac{-r^2}{g^2}$$

which corresponds to choosing L *as the Gaussian kernel density, the following formula for* $\mathsf{g}_{\mathsf{opt}}$ *yields a location M-estimate that has an Asymptotic Relative Efficiency ARE(M-estimate,Least Squares) of 95% when the true distribution is Normal*$(0, \sigma^2)$.

$$\mathsf{g}_{\mathsf{opt}} = 2.11\sigma. \tag{3.1}$$

The proof of Theorem 3.1 is given in Appendix A.

A common global loss function for ROT bandwidth selection procedures is the Integrated Asymptotic Mean Square Error (IAMSE), see Ruppert et al. (1995). The IAMSE is the integral over the design space of the asymptotic conditional bias squared plus the asymptotic conditional variance. Unfortunately, the asymptotic bias given x_i and the asymptotic variance given x_i of the M-Smoother depend upon what region x_i is in (Chu et al. 1998, Rue et al. 1998). There are three types of regions of the design space that define the asymptotic bias and asymptotic variance of the M-Smoother: smooth regions, boundary regions, and neighborhoods of jump points. Technically, in order to integrate over all regions of the design space, the number and position of all jump points must be known. Asymptotically (assuming a finite number of jump points, and arbitrarily small neighborhoods), the union of all neighborhoods of jump points constitutes a set of measure zero. Similarly the union of all boundary regions asymptotically approaches a set of measure zero because to ensure consistency h should go to zero as $n \to \infty$. Therefore, it is reasonable to take the IAMSE to be the integral assuming the entire design space is in a smooth region.

Therefore we propose minimizing the IAMSE (assuming the entire design space is in a smooth region) to obtain an optimal bandwidth formula for h. The following theorem gives us the IAMSE needed for the $\mathsf{h_{opt}}$ formula.

Theorem 3.2 *The Integrated Asymptotic Mean Squared Error for the local linear version of Chu's M-Smoother assuming an equally spaced design space is*

$$\begin{aligned} \mathsf{IAMSE} &= (1/4)h^4k_2^2\int m^{(2)}(x)^2dx \\ &+ (1/4)\left(\frac{h^3}{ng^5}\right)\beta^*[2k_2^2\tau_0+\tau_2]\int m^{(2)}(x)^2dx \\ &+ (nhg^3)^{-1}\beta\tau_0 + \mathsf{Op}\left(\frac{h^2}{n^2g^{10}}\right), \end{aligned} \tag{3.2}$$

where $k_j = \int u^jK(u)du$, *and* $\tau_j = \int u^jK(u)^2du$; $\beta = f(0)f^{(2)}(0)^{-2}\int L^{(1)}(u)^2du$, *and* $\beta^* = f(0)f^{(2)}(0)^{-2}\int L^{(2)}(u)^2du$ *where* $f(\cdot)$ *is the density function of the random errors. The* $m^{(2)}(x)$ *above refers to* $\frac{\partial^2}{\partial x^2}m(x)$.

For the proof of Theorem 3.2 see Appendix B.

Notice that (3.2) cannot be minimized with respect to the bandwidth g, because all occurences of g appear in the denominator. Thus choosing g based on the IAMSE would result in choosing g as large as possible. As the bandwidth g is increased the M-Smooth regression estimate grows closer to the usual local polynomial regression estimate. The reason g only appears in the denominator of (3.2) is because we only considered the smooth region of the design space. We conclude that the usual local polynomial regression would be asymptotically preferable to the M-smoother in the absence of jump point discontinuities (or outliers). The inability to minimize the IAMSE for g is another reason we chose to calculate $\mathsf{g_{opt}}$ based on the ARE calculation above.

Unlike g, the bandwidth h can be chosen to minimize the IAMSE. Thus the optimal bandwidth formula for h is found by simply differentiating (3.2), setting the result equal to zero and solving for h. Thus we have

$$\mathsf{h_{opt}} = \left(\frac{\beta\tau_0}{k_2^2\int m^{(2)}(x)^2dx}\right)^{\frac{1}{5}} n^{\frac{-1}{5}}g^{\frac{-3}{5}}. \tag{3.3}$$

The unknown parameters in (3.3) are $\int m^{(2)}(x)^2 dx$, β, and g. We suggest using the normal distribution with mean zero and variance σ^2 for f, rather than trying to estimate the density of the error terms used in calculating β. Therefore, using 2.11σ for g, and the normal density for f in β, we still need to estimate $\int m^{(2)}(x)^2 dx$ and σ. The estimation procedures for these will be described in the following subsection.

3.1.2 Estimation of Unknown Quantities

Ruppert et al. (1995) propose using the fast and simple block polynomial smoothing of Härdle & Marron (1995) to obtain estimates of unknown quantities in the asymptotic optimal bandwidth formula. This block polynomial smoothing technique splits the design space into N equally spaced blocks and fits an ordinary least squares polynomial for each block. Ruppert et al. (1995) suggest choosing N (the number of blocks) from a candidate list using Mallows' C_p criteria (Mallows 1973). The candidate list of possible n values ranges from $N = 1$ to $N = Nmax$, where $Nmax = \max(\min([n/20], 5), 1)$. Although this candidate list may seem somewhat arbitrary it does limit the maximum number of blocks depending on the sample size. The idea behind the choice of $Nmax$ is that each block should have at least 20 data points to fit a reasonable polynomial regression. The 5 in the above $Nmax$ calculation was chosen by Ruppert et al. (1995) because their experience suggested that the largest number of blocks that a well behaved regression function should need is about 5. Ruppert et al. (1995) also suggest increasing the 5 to some larger number if the mean function is suspected to be extremely bumpy (i.e. not well behaved).

We suggest extending the equal width block smoothing technique of Härdle & Marron (1995) in a way that is robust to jump points. The following are two proposed extensions that will make the block smoothing technique able to adapt to jump point discontinuities.

Robust Polynomial Estimates

The first modification to the fast and simple blocked polynomial method is to use robust M-estimates to calculate the different blocked regressions. Calculating an M Regression estimate using an Iterated Reweighted Least Squares algorithm with Least Trimmed (sum of) Squares (LTS) (Rousseeuw 1984), starting values can be substantially slower than the Ordinary Least Squares method, but the added robustness properties are well worth the computational cost when severe jump point discontinuities are present in the data. Using M-polynomial estimates will greatly reduce the influence a jump point has on the regression estimate within a block.

We use $\frac{\partial^\nu}{\partial x^\nu}\hat{m}_{BM}(x)$ to estimate the $\frac{\partial^\nu}{\partial x^\nu}m(x)$ for all positive integer values of ν, where $\hat{m}_{BM}(x)$ refers to the M-regression estimates used within each block. Therefore, in order to obtain a nonzero estimate of $m^{(2)}(x_i)$ it is necessary to fit at least a second order polynomial within each block. If we desire an estimate of $m^{(2)}(x_i)^2$ that is not constant across all x_i's we need to fit at least a third order polynomial. Each added degree in the polynomial becomes more and more difficult to estimate using M-regressions with LTS starting values, and the reduction of the sample size due to the blocking only adds to this difficulty. So although we desire higher order polynomials to enhance our estimate of $m^{(2)}(x_i)^2$ our ability to obtain even crude estimates is limited. Thus we suggest using cubic M-regressors within each block, and all examples and simulations throughout the remainder of this section use third order blocked M-polynomials to estimate the unknowns in the bandwidth formulas.

Variable Blocking Scheme

The second modification to the fast and simple blocking method is the variable blocking scheme. The first step in the new blocking scheme is to split the design space into $Nmax$ equally spaced blocks, where $Nmax$ is chosen by $Nmax = \max(\min([n/20], 7), 1)$ and $[\cdot]$ refers to the greatest integer function. Notice that for our calculation of $Nmax$ we have chosen to replace the 5 in Ruppert et al.'s (1995) equation with 7. This is

due to the added features in the mean function that may exist due to the jump points. We call these equally spaced blocks basic building blocks (BBB). We then group adjacent BBB together by taking the union of individual BBB. Next we examine all possible combinations of groupings that partition the design space. Finally we choose the combination of groupings that minimizes a robust version of Mallows C_p (see (3.4) for the formula for the robust version of Mallows C_p). This process will be demonstrated in the following example.

Example 3.1 *Given the design space* $x_i = \frac{i}{99}$ $i = 1, 2, 3, ..., 99$*, we calculate* $Nmax = \max(\min([99/20], 7), 1) = 4$. *The basic building blocks are*

$$\begin{aligned} A &= [\frac{1}{99}, \frac{24}{99}], \\ B &= [\frac{25}{99}, \frac{49}{99}], \\ C &= [\frac{50}{99}, \frac{74}{99}], \\ D &= [\frac{75}{99}, \frac{99}{99}]. \end{aligned}$$

The groupings include all the above basic building blocks as well as

$$\begin{aligned} E &= A \cup B, \\ F &= B \cup C, \\ G &= C \cup D, \\ H &= A \cup B \cup C, \\ I &= B \cup C \cup D, \\ J &= A \cup B \cup C \cup D. \end{aligned}$$

The list of all possible combinations of adjacent groupings that partition the design space is

1 A and B and C and D

2 E and C and D

3	A and F and D
4	A and B and G
5	H and D
6	A and I
7	J.

The grouping with the lowest robust C_p (calculated in (3.4)) would be the desired blocking arrangement.

The formula for the robust version of Mallow's C_p (first derived by Sutherland (1992)) is given by

$$\mathsf{Cp} = \mathsf{trace}(H'H) + \frac{(S^2_{\mathsf{pooled}} - S^2_{\mathsf{full}})\mathsf{trace}[(I-H)'(I-H)]}{S^2_{\mathsf{full}}}, \tag{3.4}$$

where H is the hat matrix for the M − estimate;

$$S^2_{Pooled} = \frac{\frac{\hat{s}^2 \sum_{k=1}^{N} \sum_{i=1}^{n_k} \Psi^2(r_i^*)}{n_k - p}}{\left(\frac{\sum_{k=1}^{N} \sum_{i=1}^{n_k} \Psi'(r_i^*)}{\sum_{k=1}^{N} n_k}\right)^2}, \tag{3.5}$$

where N is the total number of blocks,

n_k is the number of data points in the k_{th} block,

$\hat{s}^2$ is the Median Absolute Deviation of the residuals pooled across all blocks;

and S^2_{full} is just S^2_{pooled} when the blocking is partitioned with all $\mathsf{N_{max}}$ BBB′s.

For more details behind the calculation of the robust pooled variance estimate see Birch (1997). Note that the $\sqrt{S^2_{pooled}}$ will be used for the estimate $\hat{\sigma}$ used in both the formula for $\mathsf{g_{opt}}$ and in the density f (needed to calculate $\mathsf{h_{opt}}$).

It can be shown that the number of groupings is $\frac{Nmax(Nmax+1)}{2}$, and the number of possible combinations of groupings that constitute a partition of the design space is 2^{Nmax-1}. Thus this Variable Width Blocking Scheme requires $\frac{Nmax(Nmax+1)}{2}$ different regressions. The original Equal Spaced Blocking Scheme also requires $\frac{Nmax(Nmax+1)}{2}$ different regressions, but only considers $\mathsf{N_{max}}$ combinations. Thus since the majority of the estimation time is used in calculating different regressions, the computational efficiency of

the Variable Width Blocking Scheme is comparable to the computational efficiency of the Equal Spaced Blocking Scheme.

Practically, the Variable Width Blocking Scheme has a much larger candidate set of blocked polynomial regressions than the Equal Width Blocking Scheme for about the same cost in computer time. Intuitively, the advantage the Variable Width Blocking Scheme has over the Equal Width Blocking Scheme is its ability to isolate jump point discontinuities. Often with a practical design space the optimal number of blocks is one, because the more blocks we have the harder it is to estimate the regressions within each block. The boundary effects encountered by blocked polynomial fitting also make large numbers of blocks undesirable. However, if the data contain a jump point at about one fifth the way from a boundary point, then the Equal Spaced Blocking Scheme might choose the optimal number of blocks to be 5 in order to break the data near the jump. With the Variable Blocking Scheme we can break the data near the jump, but combine all the other data into one block.

The M-regression estimates within each block and the variable width blocking scheme work in combination in the following manner. First the variable width blocking scheme isolates jump points. Next, within blocks containing jumps, the M-regression estimate fits the portion of the block which contains the largest continuous segment of the mean function. The remainder of the block will then be down weighted by the M-regressions. This combination will produce estimates of the unknown quantities that are robust to possible jump points without having to find the jumps. To demonstrate this robustness to jump points the broken sine simulated data first described in Section 1.1 are fit with the variable width blocked M-polynomial scheme in Figure 3.1.

Figure 3.1 contains two graphical displays. The top graph contains a scatter plot of the data, the underlying mean function, and the variable width blocked M-estimate fit described above. The bottom graph displays the weight each data point has in the M-regression estimates within each block. These weights are also used to calculate the robust version of the C_P statistic and the robust estimate of the error variance S^2_{pooled}.

At first glance the variable block M-regression fit shown in Figure 3.1 appears quite poor. For example the fit is extremely far away from the true mean function in the regions $x \in [0.2, 0.275] \cup [0.55, 0.6] \cup [0.7, 0.71]$. However, each of these regions is given little weight in the calculation of the error variance estimate S^2_{pooled}. Furthermore by ignoring these points we do not destroy our estimate of the second derivative around the jumps $x = 0.2, 0.275, 0.55, 0.6, 0.7,$ and 0.71. The estimate of the second derivative calculated by the variable block M-regression is shown in Figure 3.2. Note that using a blocked cubic regression estimate leads to a blocked linear estimate of the second derivative. The second derivative estimate provided by our variable blocked M-regression technique appears to provide more reasonable estimates than would a local polynomial estimate or a traditional M-regression estimate.

Figure 3.1: Variable Width M-polynomial Fit to the Broken Sine Simulated Data

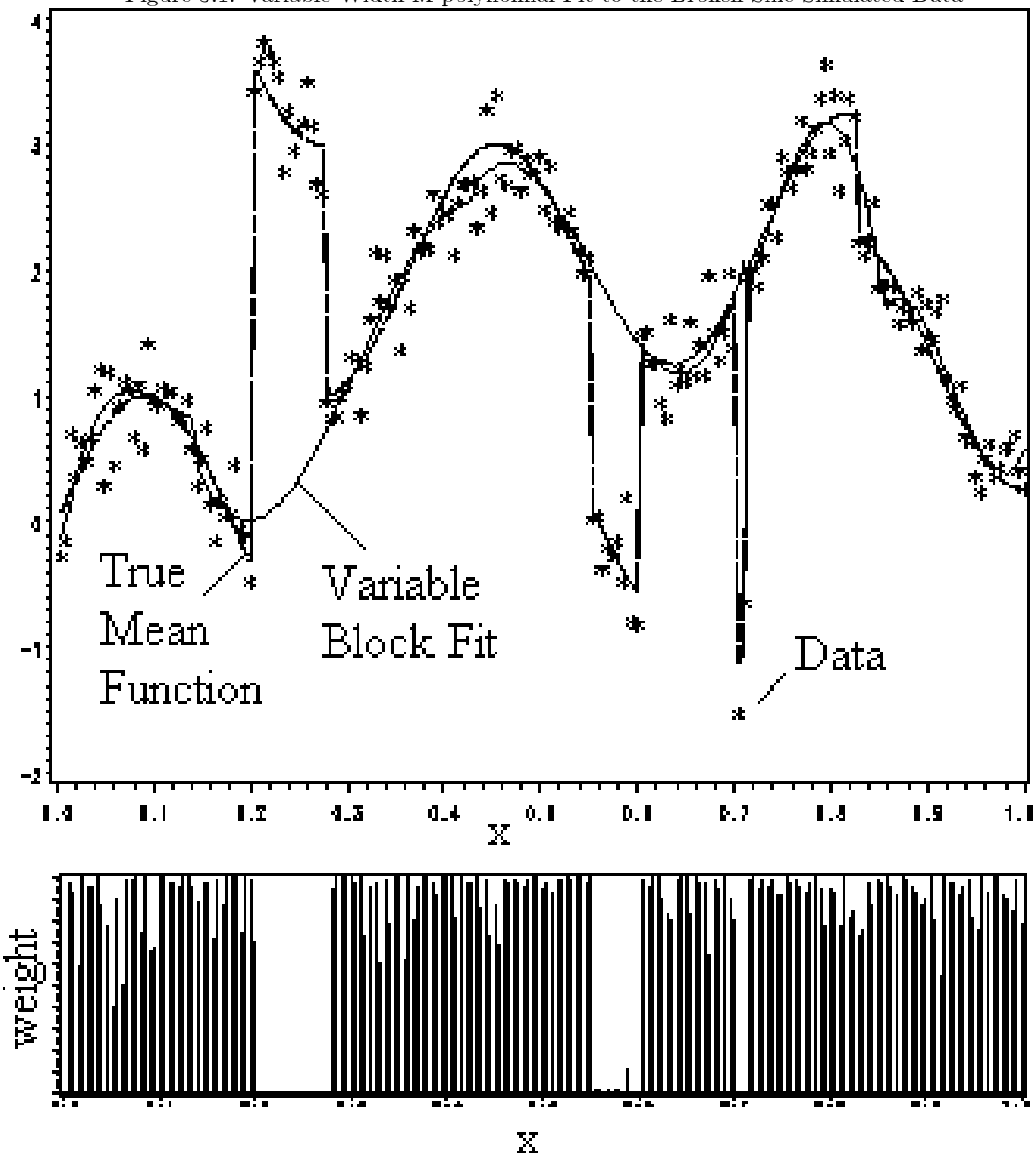


Figure 3.1 shows a simulated data set about the mean function $\sin(5.5\pi x)$ with jump points added at positions $x = 0.2, 0.275, 0.55, 0.6, 0.7, 0.71$ and 0.825 of sizes $4, -2, -1.75, 2, -3, 3$ and -1.75 respectively. The error distribution used to simulate the data about this mean function was Gaussian with mean zero and variance $\sigma^2 = (0.25)^2$. .

Figure 3.2: Variable Width M-polynomial Second Derivative Fit to the Broken Sine Simulated Data

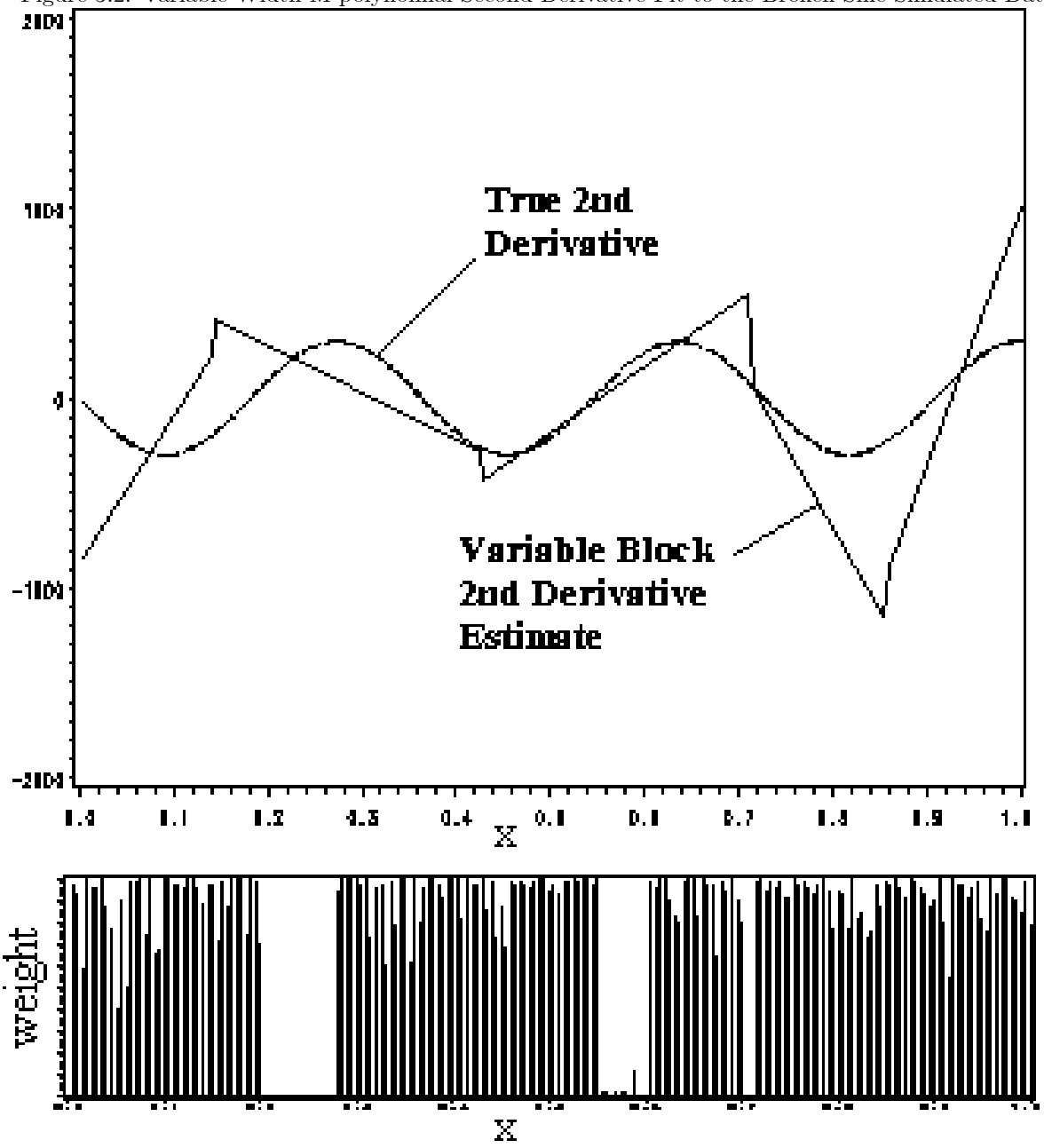


Figure 3.2 shows a simulated data set about the mean function $\sin(5.5\pi x)$ with jump points added at positions $x = 0.2, 0.275, 0.55, 0.6, 0.7, 0.71$ and 0.825 of sizes $4, -2, -1.75, 2, -3, 3$ and -1.75 respectively. The error distribution used to simulate the data about this mean function was Gaussian with mean zero and variance $\sigma^2 = (0.25)^2$. .

3.2 IRLS Procedure

We now leave the topic of the optimal bandwidth formulas and the estimates of the unknowns to give a short discourse on how to implement IRLS for the local linear version of Chu et al.'s (1998) M-smoother. Recall that the local linear M-smooth estimate at x_i is found by first finding the minima of $S(a, b, x_i)$ with respect to (a, b). If we write $S(a, b, x_i)$ in matrix form using the Gaussian kernel for both K and L we have

$$S(\underline{\beta}, x_i) = (-1)\sum_{j=1}^{n} \frac{1}{h}\exp(-0.5(\frac{x_i - x_j}{h})^2)\frac{1}{g}\exp(-0.5(\frac{Y_j - \underline{x}'_{ij}\underline{\beta}}{g})^2) \tag{3.6}$$

where $\underline{\beta} = [a \quad b]'$, and $\underline{x}'_{ij} = [1 \quad (x_i - x_j)]$. Differentiating 3.6 with respect to $\underline{\beta}$ and setting equal to $\underline{0} = [0 \quad 0]'$ yields

$$\underline{0} = \sum_{j=1}^{n} \exp(-0.5(\frac{x_i - x_j}{h})^2)(Y_j - \underline{x}'_{ij}\underline{\beta})\underline{x}_{ij}\exp(-0.5(\frac{Y_j - \underline{x}'_{ij}\underline{\beta}}{g})^2). \tag{3.7}$$

If we now write $w_{ij} = \exp(-0.5(\frac{x_i - x_j}{h})^2)\exp(-0.5(\frac{Y_j - \underline{x}'_{ij}\underline{\beta}}{g})^2)$ then 3.7 becomes

$$\underline{0} = \sum_{j=1}^{n} \underline{x}_{ij} w_{ij} Y_j - \sum_{j=1}^{n} \underline{x}'_{ij}\underline{\beta}\underline{x}_{ij}. \tag{3.8}$$

In matrix form this is

$$X'_i W_i \underline{Y} = X'_i W_i X_i \underline{\beta}, \tag{3.9}$$

where $X_i = [\underline{x}_{i1} \quad \underline{x}_{i2} \quad \cdots \quad \underline{x}_{in}]'$, $\underline{Y}$ is the vector of responses, and W_i is the diagonal matrix made up of the w_{ij}'s. Given an equally spaced design the $X'_i W_i X_i$ above will be a nonsingular matrix. Thus we can write

$$\underline{\beta} = (X'_i W_i X_i)^{-1} X'_i W_i \underline{Y}. \tag{3.10}$$

Although $\underline{\beta}$ appears by itself on the left hand side of 3.10 we have not solved the equation for $\underline{\beta}$ because W_i is a function of $\underline{\beta}$. However, 3.10 does provide the formula needed to update each estimate in the IRLS iterative solving technique. Thus to update the estimate of $\underline{\beta}$ in the IRLS algorithm we use the following formula.

$$\underline{\beta}_{\nu+1} = (X'_i W_i^{\nu} X_i)^{-1} X'_i W_i^{\nu} \underline{Y},$$

where W_i^{ν} is the diagonal weight matrix obtained when substituting $\underline{\beta}_{\nu}$ for $\underline{\beta}$.

The only other element needed to implement the IRLS algorithm is the starting value for the $\underline{\beta}_0 = [a_0 \quad b_0]$. Since we now have a variable blocked cubic M-regression estimate at x_i we suggest using $\frac{\partial}{\partial x}\hat{m}_{BM}(x_i)$ as a starting value for the slope parameter b_0. For the starting value for the intercept parameter a_0 we suggest choosing Y_i when fitting the M-smoother at x_i. Choosing the Y_i as the starting value for the intercept parameter helps the IRLS algorithm converge to the correct root. However, using the above mentioned starting values does not guarantee correct convergence. Generally we have found that using reasonable bandwidth parameters for g and h (such as calculated by our RROT method) and using the above mentioned starting values results in correct convergence provided jumps in the mean function are large enough to be distinguished from the noise. When the jumps are too small to be distinguished from the noise the root closest to the response data point may not correspond to the root that preserves the jump. When unreasonably small bandwidth parameters are used the IRLS method will break down and convergence to the correct root is unlikely.

3.3 Direct Plug In Bandwidth Selection Procedure

A higher order polynomial version of Chu et al.'s (1998) M-smoother would be useful to estimate derivatives of the mean function in the presence of jump point discontinuities. In Section 3.2 we described how to use IRLS to obtain a local linear version of Chu et al.'s (1998) M-smoother. The extention to higher order polynomials is relatively straight forward. It is conjectured that a derivative estimate based on the jump preserving M-smoother would have less bias and variance than the crude derivative estimates obtained by the variable width blocked M-regression described in Section 3.1.2. The reason the blocking procedure is a cruder method than the jump preserving M-smooth procedure is because it tends to downweight more data points. Advanced derivative estimates based upon a jump preserving cubic M-smoother are the key to our second generation Direct Plug In (DPI) bandwidth selection procedure.

To fit a third degree version of Chu et al.'s (1998) M-smoother we will need choices for the bandwidth parameters g and h as well as starting values for each of the four polynomial coefficients required to implement IRLS. Recall that the local linear version only required starting values for the two polynomial coefficients. We obtain these in a similar manner as was described in the RROT method above. First we need a variable blocked polynomial M-regression fit to the data. This fit is then used to estimate the unknowns $\int m^{(2)}(x)$ and σ. To obtain the starting values for each x_i we simply set the first, second, and third derivatives of a generic cubic polynomial equal to the corresponding derivatives of the blocked polynomial fit $\hat{m}_{BM}(x_i)$ at x_i and solve for the coefficients. The outline of this procedure is presented below as well as the definition of the cubic M-smoother fit.

Definition 3.1 *Jump Preserving Cubic M-smooth Fit*

We define the jump preserving cubic M-smoother fit at x_i by first finding all local minima with respect to $\underline{\beta}$ of

$$S(\underline{\beta}, x_i) = (-1)\sum_{j=1}^{n} \frac{1}{h}\exp(-0.5(\frac{x_i - x_j}{h})^2)\frac{1}{g}\exp(-0.5(\frac{Y_j - \underline{x}_j'\underline{\beta}}{g})^2),$$

where $\underline{\beta}$ is a $4x1$ column vector of coefficients and $\underline{x}_j' = [1 \quad x_j \quad x_j^2 \quad x_j^3]$. The M-smooth fit is then the $\underline{x}_i'\underline{\beta}$ that is the closest to the response data point Y_i. Note that the centering notation has been dropped to simplify the derivative calculations below.

Using the definition above, the DPI bandwidth selection procedure is listed below.

1. Calculate a variable width blocked quartic M-regression $\hat{m}_{BM}(x)$ exactly as described in Section 3.1.2 except using a quartic rather than a cubic regression within each block.

2. Estimate $\int m^{(2)}(x)^2 dx$ by

$$\sum_{i=1}^{n} \hat{m}_{BM}^{(2)}(x_i),$$

where

$$\hat{m}^{(2)}(x_i) = \frac{\partial^2}{\partial x^2}\hat{m}_{BM}(x)\big|_{x_i}.$$

3. Calculate S^2_{Pooled} as described in Equation (3.4).

4. Substitute $\sqrt{S^2_{Pooled}}$ and $\sum_{i=1}^{n} \hat{m}^{(2)}_{BM}(x_i)$ in for the unknowns in the bandwidth formulas h_{opt} and g_{opt} to obtain pilot bandwidth parameter estimates for g and h.

5. For each x_i calculate the starting value

$$\underline{\beta}_0 = \begin{bmatrix} Y_i - x_i(\hat{m}^{(1)}_{BM}(x_i) - \hat{m}^{(2)}_{BM}(x_i)) - x_i^2(\frac{1}{2}\hat{m}^{(2)}_{BM}(x_i) - \hat{m}^{(3)}_{BM}(x_i)) + \frac{5}{6}x_i^3\hat{m}^{(3)}_{BM}(x_i) \\ \hat{m}^{(1)}_{BM}(x_i) - \hat{m}^{(2)}_{BM}(x_i) - x_i\hat{m}^{(3)}_{BM}(x_i) - \frac{1}{2}x_i^2\hat{m}^{(3)}_{BM}(x_i) \\ \frac{1}{2}\left(\hat{m}^{(2)}_{BM}(x_i) - x_i\hat{m}^{(3)}_{BM}(x_i)\right) \\ \frac{1}{6}\left(\hat{m}^{(3)}_{BM}(x_i)\right) \end{bmatrix},$$

where

$$\hat{m}^{(\nu)}_{BM}(x_i) = \frac{\partial^\nu}{\partial x^\nu}\hat{m}_{BM}(x)\Big|_{x_i},$$

and $\hat{m}_{BM}(x)$ is the variable width blocked quartic M-regression estimate calculated in Step 1.

6. For each x_i update the $\underline{\beta}$ by applying

$$\underline{\beta}_{\nu+1} = (X_i'W_i^\nu X_i)^{-1}X_i'W_i^\nu \underline{Y},$$

where W_i^ν is the diagonal weight matrix obtained when substituting $\underline{\beta}_\nu$ for $\underline{\beta}$ in

$$w_{ij} = \exp(-0.5(\frac{x_i - x_j}{h})^2)\exp(-0.5(\frac{Y_j - \underline{x}_j'\underline{\beta}}{g})^2).$$

7. Repeat Step 6 for each x_i until convergence.

8. For each x_i take the $\underline{\beta}_c$ obtained after the IRLS algorithm has converged and store the following quantities

$$\hat{m}_{ROT}(x_i) = \underline{x}_i'\underline{\beta}_c \tag{3.11}$$

$$\hat{m}^{(1)}_{ROT}(x_i) = \underline{x}_i'\left(\underline{\beta}_c'\begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \end{bmatrix}\right)' \tag{3.12}$$

$$\hat{m}^{(2)}_{ROT}(x_i) \quad = \quad \underline{x}_i' \left(\underline{\beta}_c' \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 \end{bmatrix} \right)' \tag{3.13}$$

9. Estimate the error standard deviation to be $\hat{\sigma}_{ROT} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_{ROT}(x_i))^2}$.

10. Estimate the unknown $\int m^{(2)}(x)^2 dx$ by

$$\hat{\theta}_{22} = \sum_{i=1}^{n} \hat{m}^{(2)}_{ROT}(x_i)^2$$

.

11. Substitute both $\hat{\sigma}$ and $\hat{\theta}_{22}$ into the optimal bandwidth formulas derived in Section 3.1.2 to obtain the DPI bandwidth estimates h_{DPI} and g_{DPI}.

The performance of our RROT and DPI automatic bandwidth selection procedures are evaluated and compared via simulation studies in the following section.

3.4 Simulation Results

The fact that there are no current automatic bandwidth selection procedures available for the jump preserving M-smoother makes evaluating our methods somewhat difficult. The ideal comparison here would be between our methods and a cross validation method as suggested by Rue et al. (1998) and Simpson et al. (1998), but as we mentioned before there is no complete cross-validation method available for Chu et al.'s (1998) M-smoother.

Another natural simulation comparison would be between our methods and the methods of Ruppert et al. (1995). However, since Ruppert et al.'s (1995) bandwidth selection procedures were developed for local linear nonparametric regression (assuming no jump points) they have no estimation technique for our

additional smoothing parameter g. Furthermore, Ruppert et al.'s (1995) formula for h is optimal only in the local linear regression setting. The DPI method and the STE method of Ruppert et al. (1995) use local linear nonparametric regression fits to estimate unknowns in the optimal bandwidth formulas, which would include model misspecification (the continuity assumption) to the selection of g and h if used in the jump point problem.

3.4.1 RROT Simulation Results

Since the direct comparisons above are not possible, we chose to compare a slightly altered version of the ROT method of Ruppert et al. (1995) with our RROT method. To achieve a version of the ROT that can be applied to the jump point problem we took our optimal bandwidth formulas for g and h (from Section 3.1.1) and used an equal width blocking scheme with Least Squares regressions within each block (ELS) to estimate the unknowns in our bandwidth formulas. It seems that this is currently the most logical competitor to our RROT method. For reference purposes we have included three other competitors. Each of the reference competitors use our bandwidth formulas and vary only in the way the unknowns are estimated. For the EM method we used an equal width blocking scheme with M-regressions within each block, for the VLS method we used the Variable width blocking scheme with Least Squares regressions within each block, and for the OPT method we used the optimal bandwidths found by substituting the true values (which are known since it is a simulation study) for the unknowns in the optimal bandwidth formulas. The results of these comparisons should shed light on how much advantage there is in each of our proposed modifications described in Section 3.1.2.

The 3000 simulated sample data sets were generated from Normal$(0, 0.25^2)$ errors added to the broken sine function (see Figure 1.2) first displayed in the original paper on the local linear jump preserving M-smoothing by Rue et al. (1998). This sine function is broken by jump points at the positions $x = 0.2, 0.275, 0.55, 0.6, 0.7, 0.71$ and 0.825 of sizes $4, -2, -1.75, 2, -3, 3$ and -1.75 respectively. Each data set contained 200 paired data points (x_i, Y_i), where the x_i's were equally spaced from 0 to 1.

To evaluate the bandwidths selected by our RROT method and the above competitors we took the M-smooth fit using the bandwidths and calculated an Integrated Mean Square Error (IMSE) by

$$\mathsf{IMSE} = \sum_{i=1}^{n} (m(x_i) - \hat{m}(x_i))^2, \tag{3.14}$$

where $m(x_i)$ is the true mean function at the point x_i and $\hat{m}(x_i)$ is the M-smooth fit at x_i using one of the above bandwidth methods. The IMSE's for each simulation were averaged by method and the averages are given in Column 2 of Table 3.1, with Column 1 listing the method row labels. Similarly the standard deviation of the IMSE for all simulations by method are listed in Column 3 of Table 3.1. Columns 4 through 8 of Table 3.1 display the ranking of the methods per simulation. For example Column 4 gives the number of times out of all simulations that each method had the lowest IMSE. Thus the OPT method had a lower IMSE than all other methods 1974 times out of 3000 as indicated by Row 1 Column 4.

Table 3.1: Summary of Integrated Mean Square Errors for the Simulation Study

Method	Average IMSE	Standard Deviation	# Best	# 2nd	# 3rd	# 4th	# Last
OPT	5.41	1.51	1974	716	234	48	28
RROT	6.44	1.94	783	1309	372	276	260
EM	7.58	1.91	198	553	1139	120	990
ELS	8.12	1.00	43	371	886	1628	72
VLS	8.21	1.02	2	51	369	928	1650

In the table above the row labeled OPT refers to the optimal bandwidths ($g = 0.5275$ and $h = 0.0255$) calculated by the optimal bandwidth formulas (from Section 3.1.1) using the true values of σ and $\int m^{(2)}(x)^2 dx$ (which are only known since this is a simulation study). RROT is our proposed bandwidth selection procedure using the variable width blocking scheme and M-regression estimates within each block to estimate the unknowns in the bandwidth formulas. The EM method uses an Equal width blocking scheme and M-regression estimates within each block to estimate the unknowns in the bandwidth formulas. The ELS uses an equal width blocking scheme and Least Squares regression estimates within each block to estimate the unknowns in the bandwidth formulas. Similarly the VLS method uses the Variable width blocking scheme with Least Squares regression estimates within each block to estimate the unknowns in the bandwidth formulas. The column labeled # Best contains the number of times each corresponding method had the lowest IMSE. Similarly the Columns labeled #2nd, #3rd, #4th, and # Last correspond to the number of simulations in which the IMSE for each method had the appropriate rank.

From the results in Table 3.1 we see that our proposed RROT bandwidth selection procedure out performed all competitors except the OPT competitor which has the unfair advantage of substituting true values in for the unknowns in the bandwidth formulas. When the OPT method is not included in the analysis our RROT bandwidth selection procedure has the lowest IMSE a majority of the time (2015 times out of

3000). Furthermore our RROT method beat the ELS competitor over 80% of the time (2437 out of 3000). Table 3.1 also demonstrates the advantage the RROT method has over the ELS competitor in terms of average IMSE (6.44 vs 8.12).

An ANOVA to test for differences in the average IMSE for each of the methods above indicates a significant difference (p-value < 0.0001). Friedman's nonparametric test also resulted in a significant difference (p-val < 0.0001) between at least one of the IMSE's for the methods. Note the Friedman test is based on the sum of the ranks per simulation and the approximate significance (p-value) is calculated by $p = P(\chi^2_3 \geq 6008)$. The results of a Tukey multiple comparison indicate that the RROT method was significantly better than all other methods except the OPT method. Furthermore, the nonparametric multiple comparison using the sum of the ranks (Hollander & Wolfe 1973) indicated that the RROT method was significantly better than all other methods except the OPT method. For the results of the two multiple comparison procedures see Table 3.2.

Table 3.2: Multiple Comparison Results of the IMSE for Each Bandwidth Selection Procedure

Method	Average IMSE	Tukey Grouping	Sum of Ranks	LSD for Ranks	Grouping for Ranks
OPT	5.41	A	4440	335	A
RROT	6.44	B	6921	335	B
EM	7.58	C	10151	335	C
ELS	8.12	D	10315	335	C
VLS	8.20	D	13173	335	D

Table 3.2 lists the results of Tukey's multiple comparison as well as the nonparametric multiple comparison method of Hollander & Wolfe (1973) based on the sum of the ranks. The method labels in the first column are explained in the note for Table 3.1. The third column contains letters that correspond to the grouping based upon Tukey's multiple comparison method. Rows that contain the same letter indicate that the IMSE for those methods were not significantly different at the 0.05 level based upon Tukey's multiple comparison. Note that the fifth column contains the least significant differences for the sums of the ranks. Thus the EM and ELS methods would not be found to be significantly different based on Hollander & Wolfe's (1973) method. The grouping letters for the method based on the sums of ranks are given in Column 6.

For a summary of the average of all bandwidths $\hat{g}$ and $\hat{h}$ chosen by each method above see Table 3.3.

The results in Table 3.3 show that the selected bandwidths $\hat{g}$ and $\hat{h}$ of our RROT method are closer to the optimal than any other competing method.

To demonstrate the advantage of our RROT method in terms of ability to detect and preserve jump

Table 3.3: Summary of Selected Bandwidths for the Simulation Study

Method	Average $\hat{g}$	STD $\hat{g}$	Average $\hat{h}$	STD $\hat{h}$
OPT	0.5275	0	0.0255	0
RROT	0.6022	0.05831	0.01854	0.003011
EM	0.6406	0.05605	0.01805	0.004261
ELS	0.9890	0.03766	0.01037	0.0002221
VLS	0.9879	0.03697	0.01055	0.0002359

In Table 3.3 the method labels in the first column are explained in the note for Table 3.1. Column 2 and 3 above contain the average and standard deviation of the bandwidth parameter g selected by the corresponding methods. Similarly, column 4 and 5 above contain the average and standard deviation of the bandwidth parameter h selected by the corresponding methods.

points we plotted the results of the 3000 simulations in Figures 3.4 through 3.7. For reference purposes we have plotted the broken sine mean function in Figure 3.3. Figure 3.4 shows the 3000 fits with bandwidths chosen using our proposed RROT bandwidth selection procedure. The curves from top to bottom represent the maximum, 95th percentile, median, 5th percentile, and minimum of all 3000 simulated fits. Similarly Figure 3.5 shows the 3000 fits with bandwidths chosen using the ELS competitor with the curves representing the same quantiles as Figure 3.4.

Figure 3.3: Broken Sine Simulated Data

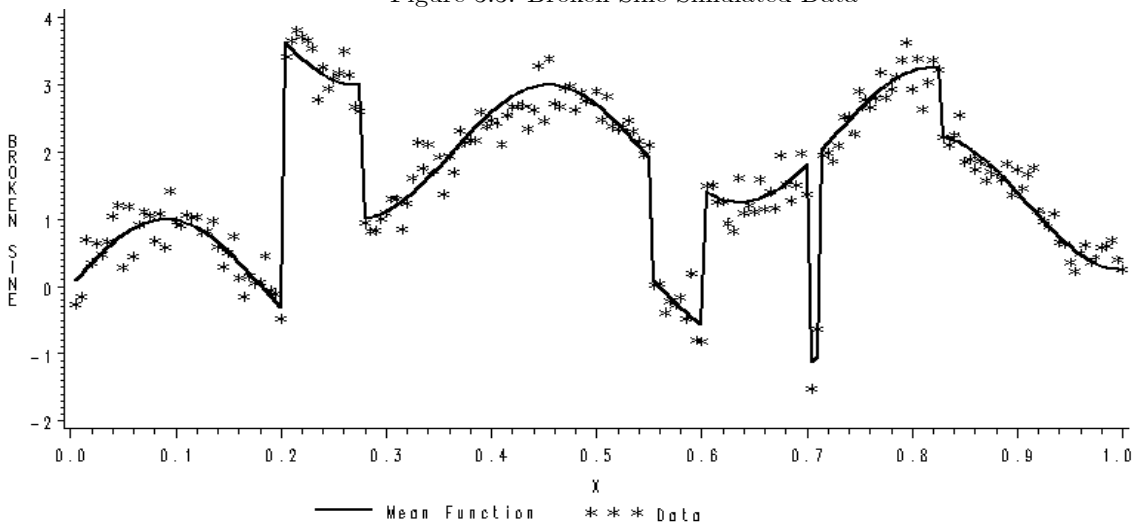


Figure 3.3 shows a simulated data set about the mean function $\sin(5.5\pi x)$ with jump points added at positions $x = 0.2, 0.275, 0.55, 0.6, 0.7, 0.71$ and 0.825 of sizes $4, -2, -1.75, 2, -3, 3$ and -1.75 respectively. The error distribution used to simulate the data about this mean function was Gaussian with mean zero and variance $\sigma^2 = (0.25)^2$.

Figure 3.4: RROT Bandwidth Selection Procedure Simulation Results

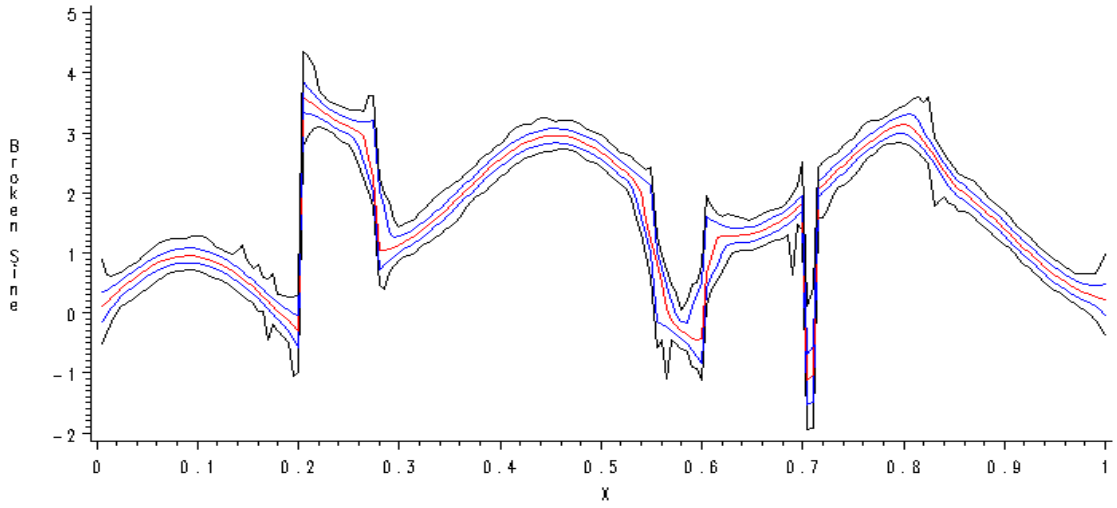


Figure 3.4 shows the results of the RROT bandwidth selection and the jump preserving M-smoother applied to 3000 simulated data sets. The simulated data was generated around the mean function described in Figure 1.2. The error distribution used to simulate the data about this mean function was Gaussian with mean zero and variance $\sigma^2 = (0.25)^2$. The five lines represent the maximum, 95th percentile, median, 5th percentile, and minimum of fits per x value.

The jump preserving performances of each of the two methods are difficult to compare based on Figures 3.4 and 3.5. Thus we took each of the 3000 simulated fits using the RROT method and the ELS method and subtracted the true mean function. This deviation from the true mean function is plotted in Figure 3.6 for the RROT method and in Figure 3.7 for the ELS method. In both Figures 3.6 and 3.7 the curves (from top to bottom) represent the quantiles: Maximum, 95th percentile, median, 5th percentile, and minimum.

To see how many of the simulated fits preserved the jump points one can simply observe which curves break the center line (Bias = 0) of Figures 3.6 and 3.7. If the median line remains flat around the center line at the point of a known jump, then most jump points were preserved. If the maximum and minimum curves both break the center line, then that jump was not preserved by any of the simulations.

Close inspection of Figures 3.6 and 3.7 shows that our RROT method preserved more jump points in a larger number of simulations than did the ELS competitor. To see this, note that in Figure 3.6 the maximum and minimum lines never break the center line and the 95th and 5th percentile lines only break the center

Figure 3.5: ROT Bandwidth Selection Procedure Simulation Results

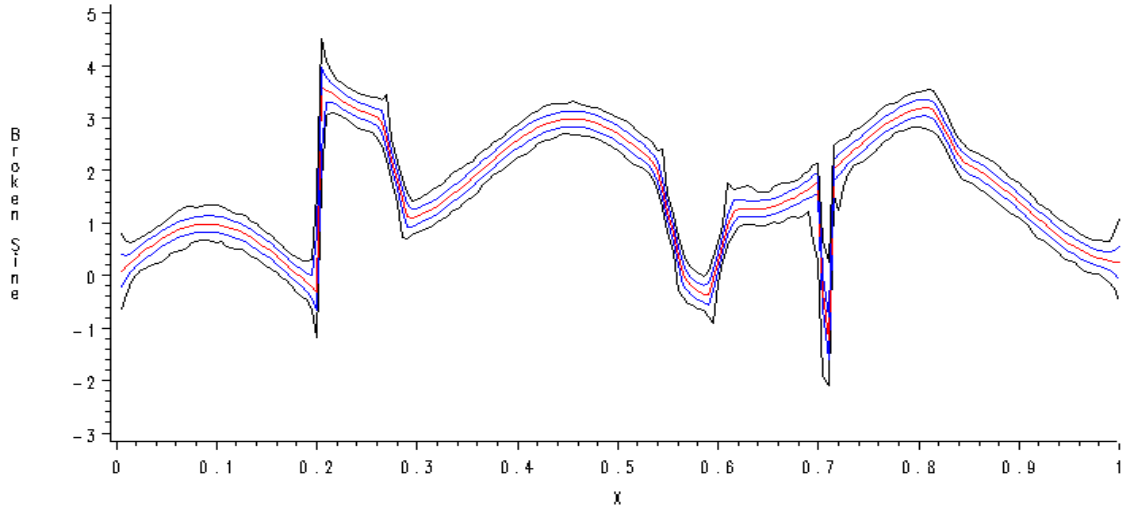


Figure 3.5 shows the results of Ruppert et al.'s (1995) ROT bandwidth selection procedure used with the jump preserving M-smoother applied to 3000 simulated data sets. The simulated data was generated around the mean function described in Figure 1.2. The error distribution used to simulate the data about this mean function was Gaussian with mean zero and variance $\sigma^2 = (0.25)^2$. The five lines represent the maximum, 95th percentile, median, 5th percentile, and minimum of fits per x value.

Figure 3.6: RROT Bandwidth Selection Procedure Bias

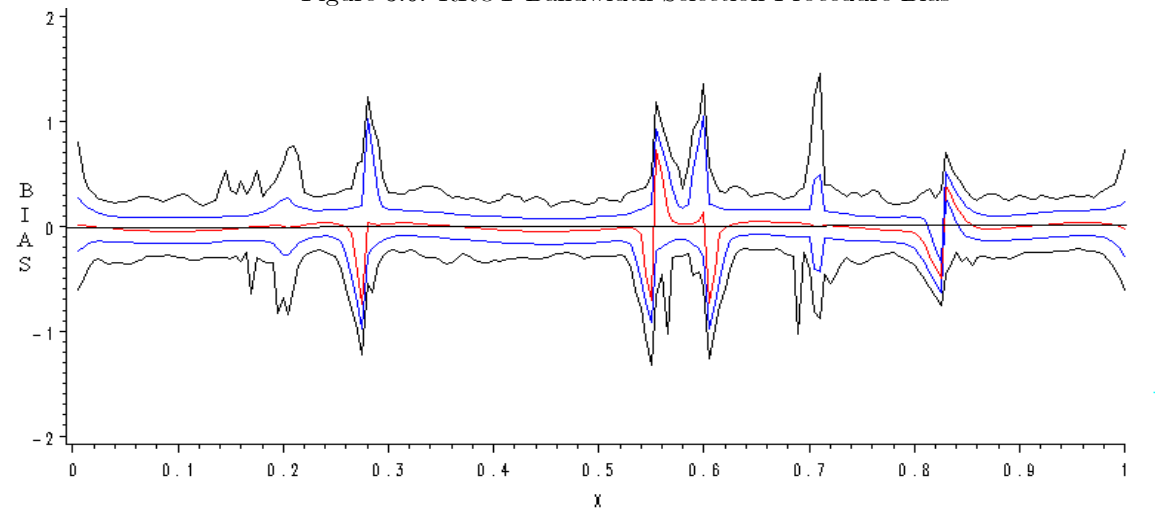


Figure 3.6 shows the bias results of the RROT bandwidth selection and the jump preserving M-smoother applied to 3000 simulated data sets. The simulated data was generated around the mean function described in Figure 1.2. The error distribution used to simulate the data about this mean function was Gaussian with mean zero and variance $\sigma^2 = (0.25)^2$. The five lines represent the maximum, 95th percentile, median, 5th percentile, and minimum of the fit deviations from the true mean function per x value.

line at the last jump point ($x_i = 0.825$), whereas the maximum and minimum (as well as all other percentile lines) of Figure 3.7 break the center line at jump points ($x = 0.275, 0.55, 0.6$ and 0.825). We assume that this increase in jump point detection and preservation is the reason our RROT method also had a lower average IMSE in Table 3.1.

Figure 3.7: ROT Bandwidth Selection Procedure Bias

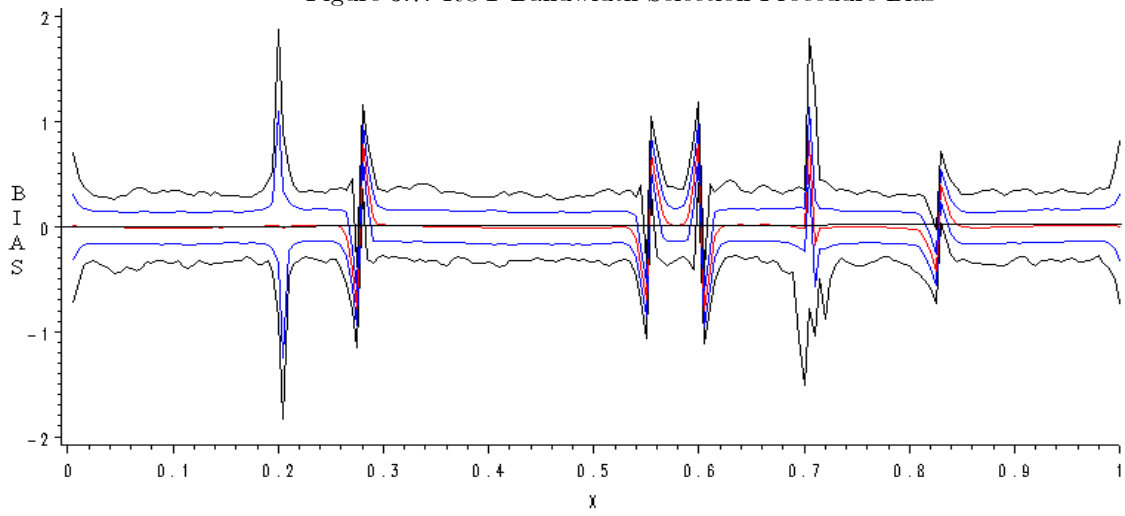


Figure 3.7 shows the bias results of Ruppert et al.'s (1995) ROT bandwidth selection and the jump preserving M-smoother applied to 3000 simulated data sets. The simulated data was generated around the mean function described in Figure 1.2. The error distribution used to simulate the data about this mean function was Gaussian with mean zero and variance $\sigma^2 = (0.25)^2$. The five lines represent the maximum, 95th percentile, median, 5th percentile, and minimum of the fit deviations from the true mean function per x value.

3.4.2 DPI Simulation Results

The DPI bandwidth selection procedure developed in Section 3.3 is more sophistocated than the RROT method developed in Section 3.1. The added sophistication translates to more complicated programming and slightly longer computer time. It is important to determine if the advantages of the added sophistication will be worth the added programming difficulty and computer time. To answer this question we applied the our DPI bandwidth selection procedure to the same 3000 simulated data sets used to evaluate the RROT method. The results are shown in Table 3.4.

Thus far we have only considered competing bandwidth selection procedures for the jump preserving M-smoother. In the following simulation results we wish to add a comparison between Chu et al.'s (1998) M-smooth procedure (using our bandwidth selection method) with another method of preserving jump points. The competitor we chose was the method of wavelet shrinkage.

The method of wavelet analysis is based on a wavelet transformation of the data. A wavelet transformation is similar to a Fourier series transformation in that it decomposes a function into localized oscillating components. When the data are decomposed the noise will affect each of the coefficients of the transformation. Wavelet shrinkage then shrinks the coefficients toward zero by a thresholding method. The inverse transformation using the smaller coefficients will then smooth the data. For a more detailed discussion of wavelet shrinkage see Donoho & Johnstone (1994), Donoho & Johnstone (1995), and Donoho, Kerkyacharian & Picard (1995).

The method of wavelet shrinkage has been shown to adapt locally to jumps and sharp cusps in the mean function (Donoho & Johnstone 1994). Therefore, wavelet thresholding is another method of jump preservation. Since most wavelet analysis is performed on signals, the design is usually assumed to be equally spaced in time. Furthermore the number of jumps in the signal (or mean function) does not need to be specified to obtain the smooth wavelet shrinkage fit. These similarities make the wavelet shrinkage method a much closer competitor to the jump preserving M-smoother than the methods of Müller (1992)

and Loader (1996) based on asymmetric kernels.

The amount of smoothing that is accomplished by the wavelet shrinkage method is determined by the size of the coefficients that are reduced toward zero. This size is referred to as the threshold. Unfortunately there is currently no universally accepted method of choosing this threshold. Two of the most common methods of choosing this threshold are the universal method (although it is not universally accepted) and the Stein Unbiased Risk Estimator (SURE) method. The universal method sets the threshold to $\sqrt{2log(n)}$, where n is the sample size. This method will often over smooth the data. The SURE method was developed by Donoho & Johnstone (1995) and is based on optimizing an estimate of the MSE. For the purpose of our comparison we have included both the universal method and the SURE method.

Another issue involved in wavelet shrinkage is the choice of the wavelet function used in the transformation. The most common choice of a wavelet function in nonparametric smoothing is the S8 wavelet function. The S8 wavelet refers to a symmlet function constructed by Daubechies (1992). The S8 wavelet provides a good compromise between feature detection and denoising. Therefore we used the S8 wavelet in all the following wavelet shrinkage simulations.

The results of the comparison simulation between the M-smoother (with RROT bandwidth selection), the M-smoother (with DPI bandwidth selection), wavelet shrinkage (with universal threshold rule), and wavelet shrinkage (using the SURE threshold rule) are shown in Table 3.4. For reference purposes we included the M-smooth method using the OPT bandwidth to show the best that the M-smoother could achieve. To compare each of these an Integrated Mean Square Error (IMSE) was calculated by

$$\mathsf{IMSE} = \sum_{i=1}^{n} (m(x_i) - \hat{m}(x_i))^2, \tag{3.15}$$

where $m(x_i)$ is the true mean function at the point x_i and $\hat{m}(x_i)$ is the fit at x_i using one of the above methods. The IMSE's for each simulation were averaged by method and the averages are given in Column 2 of Table 3.4, with Column 1 listing the method row labels. Similarly the standard deviation of the IMSE for all simulations by method are listed in Column 3 of Table 3.4. Columns 4 through 8 of Table 3.4 display

the ranking of the methods per simulation. For example Column 4 gives the number of times out of all simulations that each method had the lowest IMSE.

Table 3.4: Summary of Integrated Mean Square Errors for the DPI Simulation Study

Method	Average IMSE	Standard Deviation	# Best	# 2nd	# 3rd	# 4th	# Last
OPT	5.41	1.51	1382	989	569	60	0
DPI	5.64	1.32	1128	1046	794	32	0
RROT	6.44	1.94	475	877	1349	299	0
SURE	9.31	2.20	15	88	288	2609	0
UNIV	19.06	3.06	0	0	0	0	3000

In the table above the row labeled OPT refers to the optimal bandwidths ($g = 0.5275$ and $h = 0.0255$) calculated by the optimal bandwidth formulas (from Section 3.1.1) using the true values of σ and $\int m^{(2)}(x)^2 dx$ (which are only known since this is a simulation study). DPI and RROT are our proposed bandwidth selection procedures described in Sections 3.1 and 3.3. The row labeled SURE refers to a wavelet fit to the simulated data where the wavelet function was the S8 function and the thresholding method was the SURE method by Donoho & Johnstone (1994). The row labeled UNIV represents a wavelet fit to the simulated data where the wavelet function was the S8 function and the thresholding method was the Universal method. The column labeled # Best contains the number of times each corresponding method had the lowest IMSE. Similarly the Columns labeled #2nd, #3rd, #4th, and # Last correspond to the number of simulations in which the IMSE for each method had the appropriate rank.

Table 3.4 demonstrates how for the broken sine function used in our simulation study the jump preserving M-smoother out performed the wavelet based methods regardless of the bandwidth selection procedure (DPI or RROT). Furthermore, the DPI bandwidth selection procedure produced an average IMSE much closer to the OPT average IMSE than did the RROT method. Once again we preformed an ANOVA and a Friedman test to determine differences among the methods. Both the ANOVA and the Friedman test rejected the null hypothesis (p-value < 0.0001 for both tests) that all methods were equal in their performance based on the average IMSE. Both a Tukey multiple comparison and a nonparametric multiple comparison were calculated to determine differences between individual methods. Both the Tukey method and the method of Hollander & Wolfe (1973) showed the DPI method to be significantly better than all other methods except the OPT method. The results of the multiple comparison procedures are listed in Table 3.5.

To compare the ability of the jump preserving M-smoother with the wavelet shrinkage method we have plotted the deviation from the true mean function for all simulation in Figure 3.8 for the M-smoother (with DPI bandwidths) and Figure 3.9 for the wavelet shrinkage method (with SURE threshold rule).

Table 3.5: Multiple Comparison Results of the IMSE for Each Bandwidth Selection Procedure

Method	Average IMSE	Tukey Grouping	Sum of Ranks	LSD for Ranks	Grouping for Ranks
OPT	5.41	A	5307	335	A
DPI	5.64	B	5730	335	B
RROT	6.44	C	7472	335	C
SURE	9.31	D	11491	335	D
UNIV	19.06	E	15000	335	E

Table 3.5 lists the results of Tukey's multiple comparison as well as the nonparametric multiple comparison method of Hollander & Wolfe (1973) based on the sum of the ranks. The method labels in the first column are explained in the note for Table 3.4. The third column contains letters that correspond to the grouping based upon Tukey's multiple comparison method. Rows that contain the same letter indicate that the IMSE for those methods were not significantly different at the 0.05 level based upon Tukey's multiple comparison. Note that the fifth column contains the least significant differences for the sums of the ranks. The grouping letters for the method based on the sums of ranks are given in Column 6.

Figure 3.8: M-Smooth With DPI Bandwidth Bias

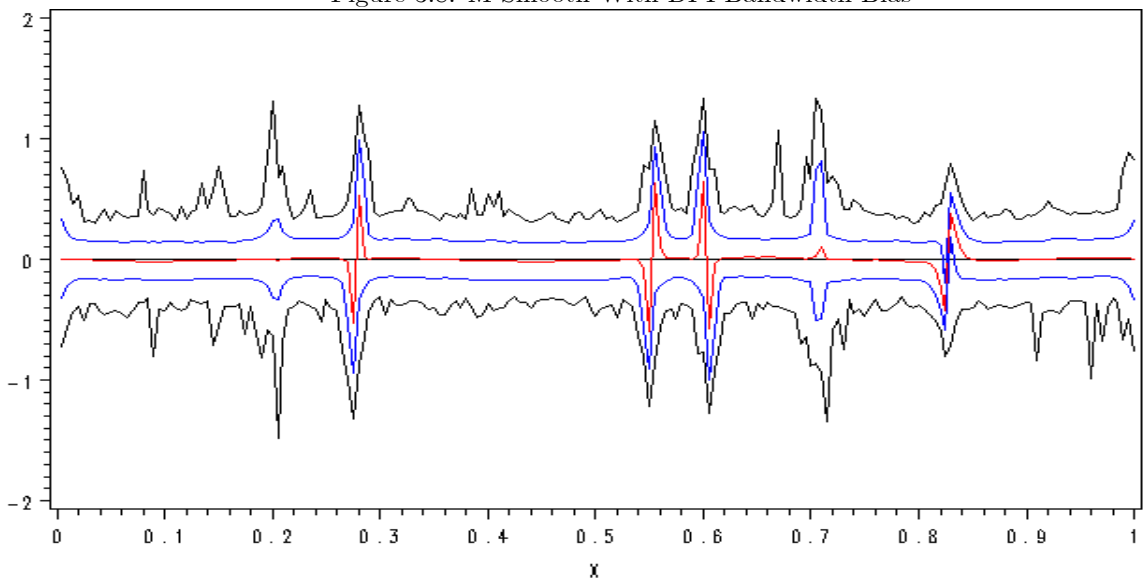


Figure 3.8 shows the bias results of the jump preserving M-smoother with our DPI bandwidth selection procedure applied to 3000 simulated data sets. The simulated data was generated around the mean function described in Figure 1.2. The error distribution used to simulate the data about this mean function was Gaussian with mean zero and variance $\sigma^2 = (0.25)^2$. The five lines represent the maximum, 95th percentile, median, 5th percentile, and minimum of the fit deviations from the true mean function per x value.

The M-Smooth method was clearly better able to preserve the jumps at points $x = 0.2, 0.275, 0.55, 0.6, 0.7,$ and 0.71 in our simulated data set than was the wavelet shrinkage method. Furthermore, the M-smooth method had a much lower absolute bias in the smooth regions $[0.2, 0.275]$, $[0.55, 0.6]$, and $[0.7, 0.71]$. This leads us to speculate that the jump preserving M-smooth method will provide much better fits than the wavelet shrinkage method when the true mean function contains impulse jumps. We also believe, due to the

Figure 3.9: Wavelet Shrinkage Procedure Bias

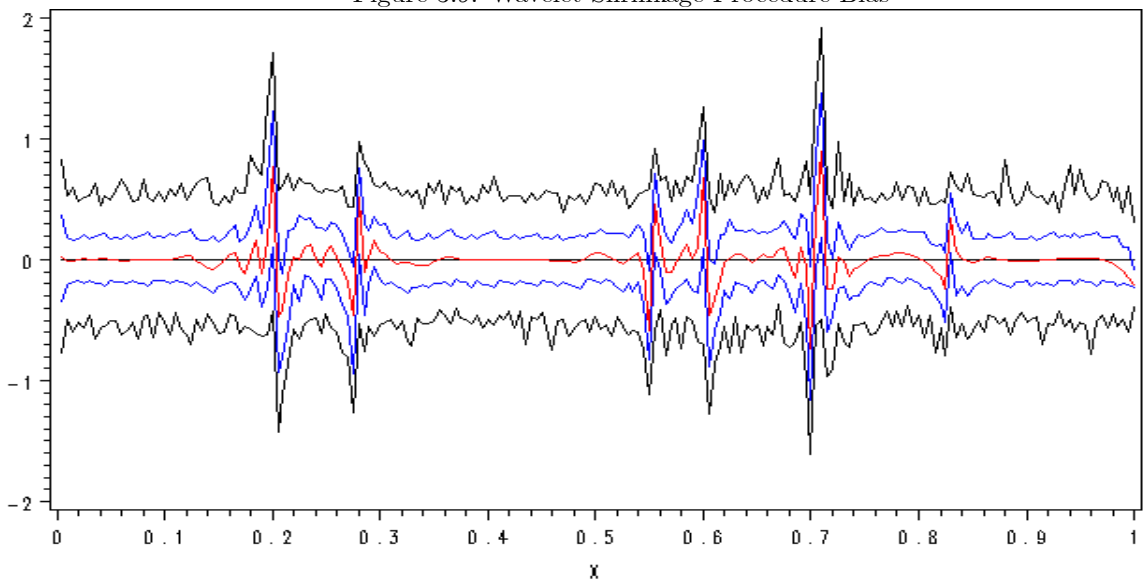


Figure 3.9 shows the bias results of the wavelet shrinkage method with SURE threshold rule applied to 3000 simulated data sets. The simulated data was generated around the mean function described in Figure 1.2. The error distribution used to simulate the data about this mean function was Gaussian with mean zero and variance $\sigma^2 = (0.25)^2$. The five lines represent the maximum, 95th percentile, median, 5th percentile, and minimum of the fit deviations from the true mean function per x value.

nature of asymmetric kernels, that the jump preserving M-smoother will out perform the methods of Müller (1992) and Loader (1996) when the mean function contains impulse jumps.

Based upon the IMSE results for our simulation, we conclude that our DPI method provides enough advantage over our RROT method to make it worth the added programming effort and computer time. Furthermore, with the bandwidths chosen from our DPI method, the M-smoother of Chu et al. (1998) provides a competitive method of preserving jump points. Therefore it is thought that a hypothesis testing procedure for jump points based on the M-smooth method (using our DPI bandwidths) would be an innovative solution to the jump point problem. The development of such a hypothesis test is given in Chapter 4.

Chapter 4

The Jump Point Critical Bandwidth Test

To introduce our proposed jump point critical bandwidth testing procedure we first provide the historical background behind the critical bandwidth method. This historical background contains two hypothesis tests. The first critical bandwidth hypothesis test was developed by Silverman (1981) to test for multiple modes in a kernel density estimate. This application of the critical bandwidth method is of particular relevance to the jump point problem because of the similarities between modal regression and M-estimation alluded to in Section 2.1. The second critical bandwidth hypothesis test was developed by Bowman et al. (1998) to test the monotonicity of a regression function. Bowman et al.'s (1998) hypothesis test is also relevant to our research because it adapts the method to the regression setting. Silverman's (1981) critical bandwidth test is presented in Section 4.1. Bowman et al.'s (1998) test is outlined in Section 4.2. Our hypothesis test is described in Section 4.3. Section 4.4 evaluates our testing procedure based on a simulation study.

4.1 Multimodality Test

Consider a sequence of independent random variables $X_1, X_2, \cdots, X_n$ identically distributed with distribution F which has density f. We then wish to test the null hypothesis that f has k modes, against the alternative that f has more than k modes.

The first element of the critical bandwidth testing procedure is a nonparametric smoothing estimate that is a function of a bandwidth. For the test of multimodality this estimate is the following kernel density estimate

$$\hat{f}(t;h) = n^{-1}h^{-1}\sum_{i=1}^{n} K\{h^{-1}(t - X_i)\},$$

where K is a kernel function and h is a bandwidth parameter.

The second element of the critical bandwidth testing procedure is a critical bandwidth or in the terminology of Silverman (1981) a "critical window width". The critical bandwidth is the critical value of the bandwidth parameter such that the nonparametric smoothing estimate satisfies the null hypothesis. In the test for multimodality the critical bandwidth h_{crit} is

$$h_{crit} = \inf\{h; \hat{f}(.,h) \quad \text{has at most} \quad k \quad \text{modes}\}.$$

An example of the calculation of this critical bandwidth is shown in Figure 4.1. Figure 4.1 shows a number of nonparametric density estimates of a set of data generated from a truly bimodal density. The red curve represents the unimodal density estimate with the least amount of smoothing. The bandwidth chosen here was $h = 0.85$. Thus we would calculate $h_{crit} = 0.85$.

The critical bandwidth is only of use in the testing procedure if it separates all nonparametric smoothing estimates that satisfy the null hypothesis from the set of all nonparametric smoothing estimates that do not satisfy the null hypothesis. This will be true in the multimodality setting if and only if the number of modes in the kernel density estimate $\hat{f}(t;h)$ is a monotonically decreasing function of the bandwidth h. It can be shown (Silverman 1981) that this is true when the kernel function K is chosen as the Gaussian density.

Figure 4.1: Demonstration of Critical Bandwidth for Kernel Density Estimation

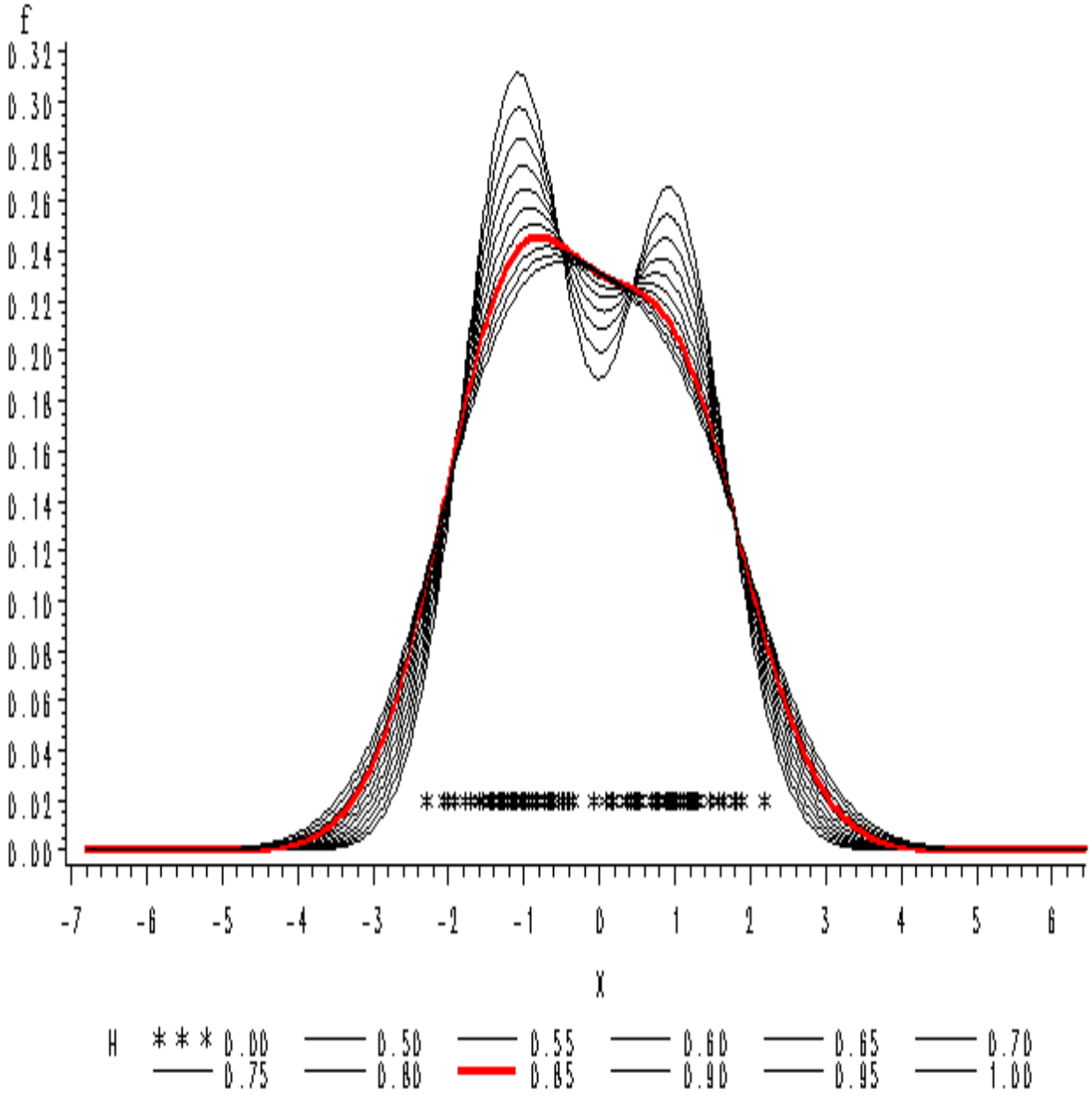


Figure 4.1 shows a simulated data set (data plotted as stars) from a bimodal mixture distribution. Each line corresponds to a normal kernel density estimate of the same data set with different bandwidths h. The critical bandwidth $h_{crit} = 0.85$ for $k = 0$ modes (the smallest bandwidth such that the estimate is unimodal) is highlighted in red.

If there is strong evidence for the alternative hypothesis it will require a large amount of smoothing to force the nonparametric estimate to satisfy the null hypothesis. Thus, we would expect the critical bandwidth to be quite large under the alternative hypothesis. Conversely, under the null hypothesis little smoothing would be needed to make the nonparametric estimate satisfy the null hypothesis. In this way the critical bandwidth is similar to a test statistic in the traditional hypothesis testing setting.

If the distribution of the critical bandwidth were known under the null hypothesis then the testing procedure would be complete. However, for the test of multimodality this distribution is not currently known. Silverman's (1981) method and all existing critical bandwidth techniques rely on a bootstrap technique to determine the significance of the critical bandwidth test.

For Silverman's (1981) method a bootstrap sample data set is generated by

$$y_i = (1 - h_{crit}^2/\hat{\sigma}^2)^{\frac{-1}{2}}(X_{I(i)} + h_{crit}\epsilon_i),$$

where $X_{I(i)}$ are sampled uniformly, with replacement, from the data $X_1, X_2, \cdots, X_n$, $\hat{\sigma}^2$ is the sample variance of the data, and ϵ_i is an independent sequence of standard normal random variables. It can be shown that this bootstrap sample will be independently and identically distributed with density $\hat{f}(.; h_{crit})$ (Efron 1979).

The significance (or p-value) of h_{crit} is calculated by

$$\text{prob}\{\hat{f}(.; h_{crit}) \quad \text{has more than k modes} | \{X_1, X_2, \cdots, X_n\} \quad \text{is drawn from} \quad f\}.$$

This quantity can be estimated by the fraction of bootstrap samples for which $\hat{f}(.; h_{crit})$ has more than k modes. For more details behind this method as well as a real world example see Silverman (1981).

One of the primary virtues of Silverman's (1981) method and critical bandwidth techniques in general is their use of nonparametric smoothing estimates where the amount of smoothing is chosen in a natural and unambiguous manner. Another virtue of the method is that the critical bandwidth needs only to be calculated once to implement the test. One drawback of the existing critical bandwidth methods is that the level of the tests tends to be overly conservative. That is, rejecting the null hypothesis when the p-value calculated above is 0.05 or smaller will usually result in a hypothesis test of level much lower than 0.05.

There are many circumstances where the number of features in a nonparametric smoothing estimate tends to decrease as the level of smoothing is increased. For example, the number of wiggles in a nonparametric regression estimate tends to decrease as the smoothing is increased. It is conceivable that the critical bandwidth method could be adapted to this example as well as many other examples. Bowman et al.'s (1998) adaptation of the critical bandwidth test to the number of wiggles in a regression function is presented in Section 4.2.

4.2 Test of Monotonicity of Regression

Bowman et al. (1998) developed a version of the critical bandwidth testing procedure to test the number of wiggles in a regression function. Rather than test for k wiggles versus more than k wiggles Bowman et al. (1998) consider the more specific case of 0 wiggles (i.e. monotonicity) versus more than 0 wiggles (i.e. nonmonotonicity). This test contains each of the elements of a critical bandwidth test listed in Section 4.1. The nonparametric smoothing estimate is a local constant or local linear regression which is a function of the bandwidth h. A critical h is calculated such that for all $h > h_{crit}$ the smooth regression fit is monotonic. Finally, the significance of the h_{crit} is determined based on a bootstrap resampling technique. Each of these elements are explained further in the following outline of the procedure.

1. Consider the Gasser-Müller kernel regression estimator

$$\hat{m}(x;h) = \sum_{i=1}^{n} \int_{T_{i-1}}^{T_i} \phi_h(x-y)dyY_i,$$

where ϕ is the Gaussian density function and $\phi_h(.) = h^{-1}\phi(h^{-1}.)$ (Härdle 1990). Here, $T_j = \frac{1}{2}(X_j + X_{j+1}), j = 1, \cdots, n-1$, where, without loss of generality, the X's are ordered, and T_0 and T_n are the ends of the range of interest.

2. Find the critical bandwidth h_{crit} where

$$h_{crit} = \inf\{h; \hat{m}(x;h) \quad \text{is monotone}\}.$$

3. For $i = 1, \cdots, n$, calculate $\hat{\varepsilon}_i = Y_i - \hat{m}(X_i; h_0)$ where h_0 is the bandwidth selected by the method of Ruppert et al. (1995)

4. Generate a bootstrap sample of errors $\hat{\varepsilon}_1^*, \cdots, \hat{\varepsilon}_n^*$ by resampling, with replacement from $\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n$. Apply these bootstrap errors to the critical fit to obtain a bootstrap sample response data set $Y_i^* = \hat{m}(X_i; c_{crit}) + \hat{\varepsilon}_i^*, i = 1, \cdots, n$.

5. Apply $\hat{m}$ using h_{crit} to $\{(X_i, Y_i^*), i = 1, \cdots, n\}$ and observe whether or not the result is monotone.

6. Repeat the bootstrap sample a large number of times and calculate the p-value to be the proportion of estimates in Step 6 which are NOT monotonic.

For a demonstration of the steps of this test of monotonicity see Figure 4.2. In Figure 4.2 we have a simulated data set (data plotted as stars) about the mean function $m(x) = 1 + x - 0.45\exp\{-0.5(x - 0.5)^2/0.1^2\}$. The noise around this mean function was generated by a Normal distribution with mean zero and variance $\sigma^2 = (0.15)^2$. The blue line represents the local linear nonparametric fit using the bandwidth chosen by Ruppert et al.'s (1995) selection procedure. The bootstrap errors would then be found by resampling, with replacement, from the differences between this blue line and the stars. These bootstrap errors would then be applied to the critical fit (shown as a red line in Figure 4.2) to obtain the bootstrap sample $\{(x_i, Y_i^*)\}$. Each bootstrap sample would then be fit with a local linear nonparametric regression using the critical bandwidth $h_{crit} = 0.97$ and monotonicity assessed. The proportion of bootstrap fits that were not monotonic would then be the p-value for the hypothesis test. Although this specific data set has not been tested using the method, Bowman et al. (1998) provide a simulation study using a similar data set that rejected monotonicity.

The last issue to be resolved before we move on is the question of monotonicity of the wiggles with respect to the bandwidth parameter h. That is, is the number of wiggles in a nonparametric regression function a monotonically decreasing function of the bandwidth h? The answer is yes provided the nonparametric regression is of the kernel type (local polynomial of degree 0) with Gaussian kernel. Otherwise there is no guarantee that the wiggles are a monotonic function of the bandwidth. However, Bowman et al. (1998)

claim (based on simulations) that this lack of monotonicity does not often cause a problem when using a local linear version of the nonparametric regression with Gaussian kernel.

It is important to note that in the regression setting a bootstrap resample can only be calculated given a good method of estimating the errors $\hat{\varepsilon}_i, i = 1, \cdots, n$. In nonparametric regression this will depend on a good method of choosing the bandwidth h. Bowman et al.'s (1998) bootstrap errors use the automatic bandwidth selection procedure of Ruppert et al. (1995). The calculation of the test statistic h_{crit} is automatic and does not require a data driven bandwidth selection procedure just as in the multimodality testing procedure. However, the calculation of the p-value which requires the bootstrap samples does require a data driven bandwidth selection procedure. This will also be true in the regression jump point discontinuity case. Thus the development of our RROT and DPI methods in Chapter 3 are crucial to adapting the critical bandwidth testing procedure to the jump point problem. The steps for our critical bandwidth bootstrap hypothesis test are modeled after the 6 steps above and are outlined in the following section.

Figure 4.2: Demonstration of the Critical Bandwidth for the Test of Monotonicity

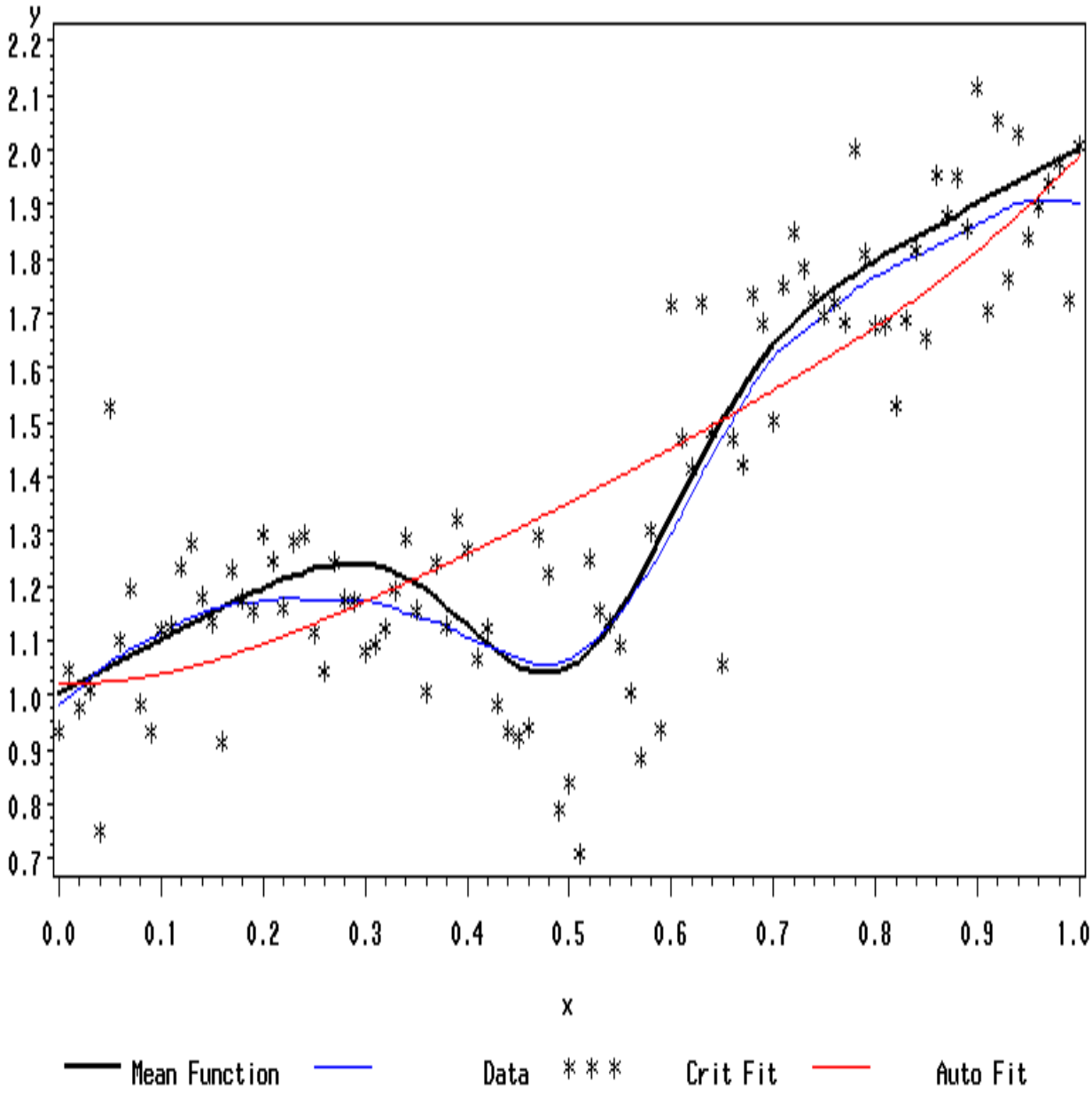


Figure 4.2 shows a simulated regression data set (data plotted as stars) from a mean function of $m(x) = 1 + x - 0.45\exp\{-0.5(x-0.5)^2/0.1^2\}$. The noise around this mean function was generated by a Normal distribution with mean zero and variance $\sigma^2 = (0.15)^2$. The black line represents the true mean function, the blue line represents the nonparametric fit using Ruppert et al.'s (1995) automatic bandwidth selection procedure, and the red line represents the critical fit (using the smallest bandwidth $h_{crit} = 0.97$ that produces a monotonic fit).

4.3 Jump Point Critical Bandwidth Test

The hypothesis test of monotonicity in regression leads to a hypothesis test for a jump point discontinuity in the regression setting. One slight glitch in the direct application of the critical bandwidth test to the jump point problem is the interaction between the bandwidth parameters h and g in terms of the M-smoother's ability to preserve jumps. If h is allowed to vary there is no value for the minimum bandwidth g such that the M-smooth fit does not contain a jump. Furthermore, the number of jumps in the M-smooth fit is not necessarily a monotonically decreasing function of the bandwidth h. However, we claim that if h is fixed a critical value for g can be found, provided the kernel L is chosen as the Gaussian density. This claim is based upon empirical evidence and Theorem 4.1. We defer the argument for the existence of a critical bandwidth g_{CRIT} to Section 4.3.2 so that we may first develop methodology for detecting jumps in the M-smooth fit.

Our hypothesis test for the presence of at least one jump point discontinuity is outlined in the following steps.

1. Consider the local constant version of the jump preserving M-smoother by Chu et al. (1998) defined in Section 2.4.

2. Use the DPI bandwidth selection procedure developed in Section 3.3 to obtain the bandwidth parameters g_{DPI} and h_{DPI}.

3. Fix the bandwidth parameter h at h_{DPI} and find the value

$$g_{CRIT} = \inf\{g| \quad \text{The M-smooth regression does not contain a jump}\}.$$

4. Calculate $\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n$ by

$$\hat{\varepsilon}_i = Y_i - \hat{m}_{DPI}(x_i),$$

where $\hat{m}_{DPI}(x)$ is the jump preserving M-smooth using the bandwidths g_{DPI} and h_{DPI}.

5. Generate a bootstrap sample of errors $\hat{\varepsilon}_1^*, \cdots, \hat{\varepsilon}_n^*$ by resampling, with replacement from $\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n$.

Apply these bootstrap errors to the critical fit to obtain a bootstrap sample response data set $Y_i^* = \hat{m}(X_i; h_{DPI}, g_{CRIT}) + \hat{\varepsilon}_i^*, i = 1, \cdots, n$.

6. Apply the M-smoother using h_{DPI} and g_{CRIT} to $\{(X_i, Y_i^*), i = 1, \cdots, n\}$ and observe whether or not the result contains a jump point discontinuity.

7. Repeat the bootstrap sample a large number of times and calculate the p-value to be the proportion of estimates in Step 6 which contain at least one jump point discontinuity.

Observing whether the jump preserving M-smooth fit contains a jump is not as simple as observing a dip in a local linear nonparametric regression fit. The jump preserving M-smoother was designed to preserve jumps but not to estimate the number, position, or size of the jump(s). Therefore the jump preserving M-smoother does not contain an internal automatic method for flagging whether the fit contains a jump. When the fit is plotted graphically it is difficult with the human eye to detect the jumps. Since the M-smooth fit is only calculated at the data points $x_1, x_2, \cdots, x_n$ the M-smooth fit is discrete. When observing a discrete fit it is easy to mistake the step from one data point to another as a jump point. If the discrete fit is connected (by just connecting the dots) the result will appear continuous. Flagging an M-smooth fit that contains a jump is an important part of the critical bandwidth testing procedure and is the topic of Section 4.3.1.

4.3.1 Detecting Jumps in the M-smoother

In order to determine if the M-smooth fit contains a jump point we suggest extending the method to calculate the fit between the data points x_i and x_{i+1} $i = 1, 2, \cdots, n-1$. The positions that will be most informative for flagging jump points will be the midpoints $(x_i + x_{i+1})/2$ $i = 1, 2, \cdots, n-1$. To fit the M-smoother at any x value we redefine the local linear jump preserving M-smooth fit at x as follows: Consider all the local minima $\hat{a}$ of

$$S(a, x) = (-1) \sum_{j=1}^{n} \frac{1}{h} K(\frac{x - x_j}{h}) \frac{1}{g} L(\frac{Y_j - a}{g}). \tag{4.1}$$

Find the regressor data point x_i closest to x. Calculate the M-smooth fit to be the local minimum $\hat{a}$ closest to the response data point Y_i that corresponds to the x_i closest to x. In the case where there are two regressor data points x_i and x_{i+1} of equal distance to x (i.e. $x = (x_i + x_{i+1})/2$) we calculate two M-smooth fit values at x. One fit $\hat{m}_l(x)$ is found by the local minimum closest to Y_i and the other $\hat{m}_u(x)$ is found by the local minimum closest to Y_{i+1}. We flag $x = (x_i + x_{i+1})/2$ as the position of a jump point discontinuity in the M-smooth fit when the two M-smooth fits at x correspond to different local minima of (4.1).

This procedure can be used to flag the M-smooth fits that contain at least one jump point. Additionally, this procedure greatly enhances the graphical display when connecting the dots of the M-smooth fit. If the dots are connected in the appropriate order jumps will appear as vertical lines. This procedure of fitting the midpoints also provides estimates of the number, position, and size of jump point(s) based on the jump preserving M-smoother. The number of jump points in the mean function can be estimated by the number of midpoints that have two distinct M-smooth fit values. The midpoints with two distinct fit values provide estimates of the position of the jump points. The size of the jump point at x can be estimated by the difference $\hat{m}_u(x) - \hat{m}_l(x)$. These estimates of the number, position, and size of the jump(s) are quite sensitive to the choice of the bandwidth parameters g and h. We suspect that using the bandwidths chosen by our proposed DPI method in the midpoint M-smooth fit procedure above will provide reasonable estimates of the number, position, and size of the jump(s). However, no studies have been made to determine the asymptotic or small sample properties of these estimates.

Now that we can flag the M-smooth fits that contain at least one jump point we can develop an algorithm for finding the g_{CRIT}. This algorithm along with the argument for the existence of a critical bandwidth g_{CRIT} is presented in Section 4.3.2.

4.3.2 Algorithm For Calculating g_{CRIT}

We first present a theorem that helps support the claim that a critical bandwidth exists.

Theorem 4.1 *Given a fixed set of data* $\{(x_i, Y_i)\}, i = 1, \cdots, n$, *the midpoint* $x = (x_i + x_{i+1})/2$ *for some* i, *the constant* h, *and the two constants* $0 < g_l < g_u$ *we have*

$$\begin{aligned} & \xi\Big(\sum_{j=1}^{n} \frac{-1}{2\pi h g_u^3} \exp\{-0.5(\frac{x - x_j}{h})^2\} \exp\{-0.5(\frac{Y_j - a}{g_u})^2\}(Y_j - a)\Big) \\ \leq \quad & \xi\Big(\sum_{j=1}^{n} \frac{-1}{2\pi h g_l^3} \exp\{-0.5(\frac{x - x_j}{h})^2\} \exp\{-0.5(\frac{Y_j - a}{g_l})^2\}(Y_j - a)\Big), \end{aligned} \tag{4.2}$$

where $\xi(\cdot)$ *represents the number of sign changes in* $(\cdot)$ *as the argument (in our case the* a*) traverses the domain of definition from left to right.*

The proof of Theorem 4.1 is based on the variation-diminishing property of totally positive kernel functions and is presented in Appendix C. The fact that the kernel function L was chosen as the totally positive Gaussian kernel is a crucial element of the proof of Theorem 4.1. Other kernels can be used provided they satisfy the total positive criteria. For a discussion of the total positive criteria for kernel functions as well as examples see Karlin (1968 pp 11-22).

Notice that Theorem 4.1 shows that the number of local minima of $S(a, x)$ is a decreasing function of the bandwidth g, provided h is fixed and the kernel function L is chosen as the Gaussian density. Thus, once a g is found such that the defining equation $\frac{\partial}{\partial a} S(a, x) = 0$ has a single root, all such defining equations with larger g will also contain a single root. A single root in the defining equation $\frac{\partial}{\partial a} S(a, x) = 0$ indicates that the M-smooth fit at x is continuous (does not contain a jump at x). However, the presence of multiple roots in the defining equation $\frac{\partial}{\partial a} S(a, x) = 0$ does not necessarily indicate a jump point in the M-smooth fit at x. It is difficult to calculate a formula for the difference between the closest root to Y_i and the closest root to Y_{i+1} in terms of g. Thus a complete proof that the number of jumps in Chu et al.'s (1998) M-smooth fit is a decreasing function of the bandwidth g is left for further research. Empirical evidence indicates that even if there exist examples where this is not the case they are rare.

To find g_{CRIT} we implement a sequence of doubling and averaging different values of g and checking the M-smooth fit to see if there is a jump. This sequence is outlined below.

1. Let i = 1, $h = h_{DPI}$, and $g = g_{DPI}$. Set $g_{MIN} = 0$.

2. Calculate the two M-smooth fits at the midpoint $x = (x_i + x_{i+1})/2$ using g and h for the bandwidth parameters.

3. If the two M-smooth fits are unequal then double g. Repeat this doubling of g until the M-smooth fits at $x = (x_i + x_{i+1})/2$ are equal. Set g_{MAX} equal to the first g found such that the M-smooth fits at the midpoint $x = (x_i + x_{i+1})/2$ are equal.

4. Set $g = (g_{MAX} + g_{MIN})/2$ and compute the two M-smooth fits at the midpoint $x = (x_i + x_{i+1})/2$. If the M-smooth fits are equal then set $g_{MAX} = g$ otherwise set $g_{MIN} = g$.

5. Repeat Step 4 until convergence in g is obtained.

6. Increment i by 1 (provided $i < n - 1$) and return to Step 2.

7. Once $i = n - 1$ we set $g_{CRIT} = g$.

Until this point we have been using the IRLS algorithm to find the local minimum of $S(a, x)$ needed for the M-smoother. However, in the above algorithm the IRLS has convergence problems. As we mentioned in Section 3.2 the IRLS algorithm tends to converge to the correct root provided reasonable values for the bandwidth parameters g and h are used. The above algorithm requires the M-smooth fit be obtained for extremely small values of the bandwidth parameter g. These small values of g can cause the IRLS to have difficulty converging to the root closest to the starting value. This lack of convergence in the IRLS causes the above algorithm to calculate g_{CRIT} too large when the true mean function does not contain a jump point. The inflation of g_{CRIT} is due to the algorithm forcing the g large enough for the IRLS to have correct convergence (rather than forcing the g just large enough that no jump is detected).

Since the IRLS method has trouble with convergence, another method of finding the local minima of $S(a, x)$ is needed to calculate g_{CRIT}. Currently there is no fast and simple method of finding the correct root when g is chosen extremely small. For our current research we have used a grid search to locate the

local minima of $S(a, x)$. Although the grid search will find the correct root, it is much slower than the IRLS algorithm. The development of a reliable algorithm that is faster than the grid search is an interesting and open research topic that would greatly enhance the current jump point critical bandwidth testing procedure.

The number of computer intensive grid searches required by the jump point critical bandwidth test is quite large. The majority of these grid searches occur in the nested loops required to calculate g_{CRIT}. Additionally to evaluate each of the bootstrap samples one grid search must be calculated per each of the $(n-1)$ midpoints. This limits the number of simulation evaluations that can be completed in a reasonable time frame. Therfore, even though it represents a large number of computer hours, the simulation study presented in Section 4.4 was limited in scope and size.

4.4 Hypothesis Test Simulation Results

Each mean function we used in our simulation study was constructed by adding a continuous function with a step function. We used two different continuous functions along with two different step functions to generate four different mean functions. The two continuous functions we chose were (1) $sin(\pi x)$ and (2) $sin(3\pi x)$ for all $x \in [0, 1]$. The two step functions we chose were (1) $\Delta\sigma I_{x\in[0.5,1]}$ and (2) $\Delta\sigma I_{x\in[0.45,0.55]}$. The Δ in each step function controls the size of the jump or jumps in the mean function relative to the noise σ in the simulated data. We chose three values for the Δ in our simulation study: $\Delta = 0, 4$, and 8. The errors applied to each mean function were independently generated from a Normal$(0, \sigma^2)$ distribution, where $\sigma = 0.15$. The sample size and bootstrap sample size were each set to 100. For each of the 16 different combinations above we replicated 100 simulations.

Thus our simulation study takes the form of a factorial experiment with 3 factors. The first factor curvature in the mean function, has two levels (the low level corresponds to $sin(\pi x)$ and high level corresponds to $sin(3\pi x)$). For the second factor, number of jump points in the mean function, we again have two levels (i.e. one jump vs two jumps). The last factor, jump size, has 3 levels corresponding to the 3 values of Δ.

As we calculate the proportion of the simulations for which the jump point critical bandwidth test rejects continuity (p-value < 0.05) we can estimate the power curve for each mean function. The proportions required for this calculation are presented in Table 4.1. The rows of Table 4.1 correspond to the values of Δ and each column represents a different mean function. Within each cell is the proportion of the simulations that rejected continuity.

Table 4.1: Critical Bandwidth Test Results P-val < 0.5

Δ	$\sin(\pi x)+\Delta I_{x\in[0.5,1]}$	$\sin(\pi x)+\Delta I_{x\in[0.45,0.55]}$	$\sin(3\pi x)+\Delta I_{x\in[0.5,1]}$	$\sin(3\pi x)+\Delta I_{x\in[0.45,0.55]}$
0	1/100	1/94	0/86	0/100
4	9/100	11/100	2/100	0/100
8	48/100	70/97	46/100	68/88

Within each cell of Table 4.1 is the proportion of simulated data sets for which the jump point critical bandwidth hypothesis test rejected continuity (p-value less than 0.05).

The estimated power curves for the jump point critical bandwidth hypothesis test are plotted in Figure 4.3. The estimated power curves of Figure 4.3 suggest that the ability of our test to detect small jump points may hinge more on the level of curvature in the mean function than on the number of jumps. Likewise the ability of our test to detect two large jumps will be easier than one large jump regardless of the level of the curvature in the mean function (provided the size of the jumps are large with respect to both the level of curvature in the mean function and the error variance).

From Figure 4.3 we see that the estimated level of our test is well below the desired $\alpha = 0.05$. Thus our jump point critical bandwidth testing procedure exhibits the same conservative features that Bowman et al.'s (1998) test of monotonicity demonstrated. Additionally, it seems from this simulation study that the power of our test is also quite low. This may be an artifact of the conservative nature of the test or a specific aspect of the simulation example we chose. To determine more information about the power of our proposed hypothesis test further investigation is required.

Figure 4.3: Power Curves for the Jump Point Critical Bandwidth Hypothesis Test

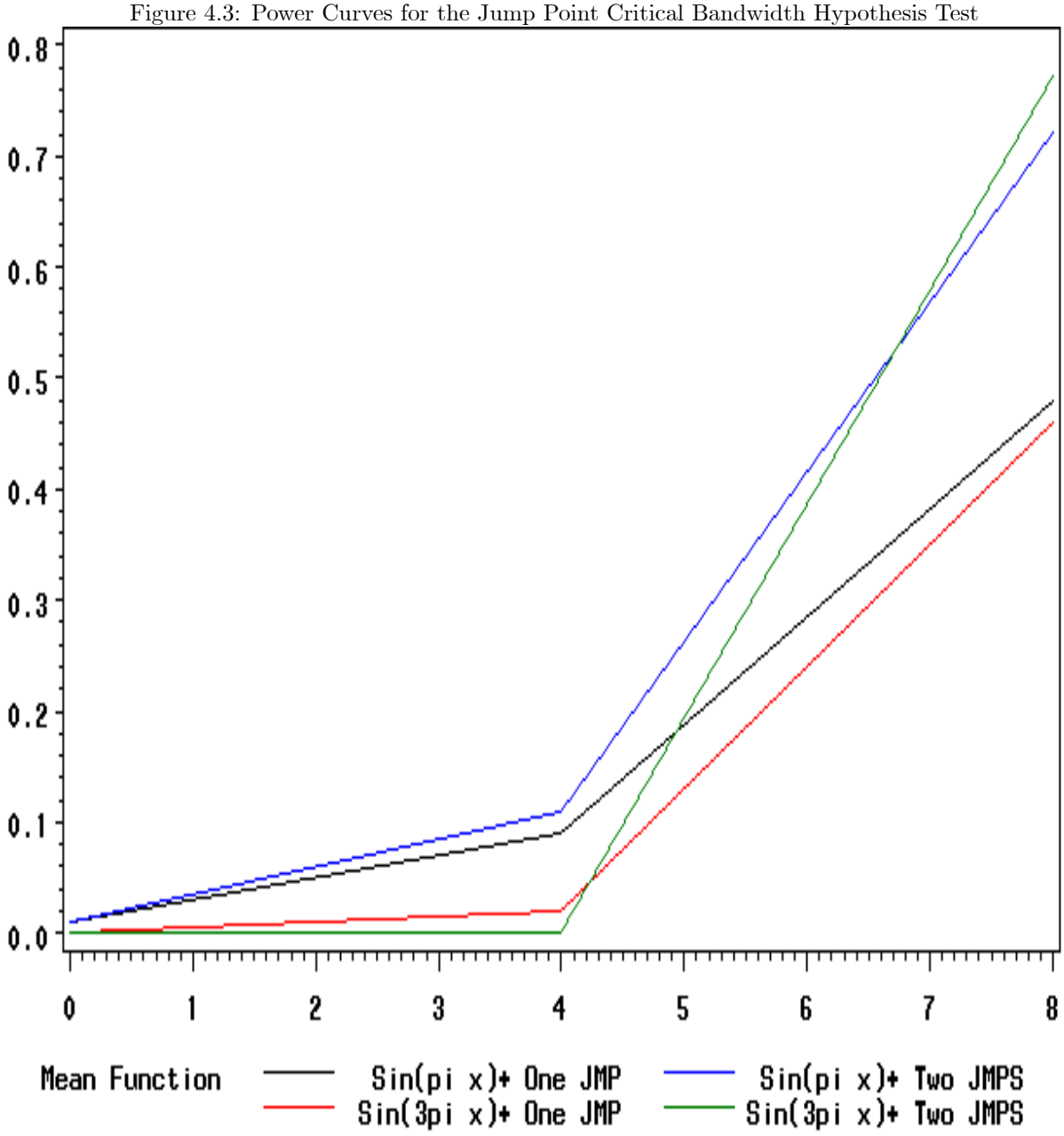


Figure 4.3 displays the estimated power curves for the four different mean functions with respect to Δ.

Chapter 5

Conclusions and Future Research Areas

5.1 Bandwidth Selection Summary

The jump preserving M-smoother of Chu et al. (1998) shows a lot of promise for image processing and many other scientific fields where jump point discontinuities are common. However, without an automatic bandwidth selection procedure this regression method cannot be applied without arbitrary user input. Until now, there have been no automatic bandwidth selection procedures available for the jump preserving M-smoother, thus evaluating its performance compared to other methods has not been possible.

The RROT method presented in Chapter 3 represents the first fully automatic bandwidth selection procedure available for Chu et al.'s (1998) M-smoother. The development of this RROT method includes a number of original procedures and calculations. The first of these calculations is the optimal bandwidth formula for h, which is based on the asymptotic bias and variance of the local linear version of the jump preserving M-smoother. We also provided a formula for g based on an asymptotic relative efficiency calculation.

In Section 3.1.2 we presented a crude method for estimating the second derivative of a mean function that contains jump points. This estimation procedure includes a variable width blocking scheme in conjunction with M-regression estimates within each block.

The procedure used to obtain estimates of the second derivative of the mean function gives rise to other derivative estimates as well. This leads to starting values for certain parameter coefficients required by the IRLS algorithm. Additionally, higher order derivative estimates provided the tools needed to develop the more sophistocated DPI bandwidth selection procedure of Section 3.3.

The DPI bandwidth selection procedure uses the jump preserving M-smooth technique in place of the crude blocking procedure of the RROT. Thus, bandwidth values chosen by our DPI selection procedure are thought to be much better than those chosen by the RROT method. The simulation study in Section 3.4.2 comparing the RROT and DPI method provided confirming evidence to the superiority of the DPI method.

After developing the DPI automatic bandwidth selection procedure we presented the first simulation study comparing the jump preserving M-smoother to the very popular wavelet shrinkage method. Our simulation study demonstrates one example where the jump preserving M-smoother of Chu et al. (1998) (allong with our DPI selection procedure) out performed the wavelet shrinkage method in terms of IMSE.

Our bandwidth selection procedure also provides the groundwork for many possible extensions of the jump preserving M-smoother. One such extention is in the area of jump point hypothesis testing. In Section 5.2 we summarize our proposed jump point hypothesis test.

5.2 Critical Bandwidth Test Summary

In Section 4.3 we made the first steps toward the implementation of a jump point hypothesis testing procedure based on the M-smoother of Chu et al. (1998). The test statistic for our method is essentially the smallest possible bandwidth that allows the jump preserving M-smoother of Chu et al. (1998) to satisfy the null

hypothesis (i.e. has no jumps). A bootstrap method is then used to calculate the probability that this smallest possible bandwidth would be observed given a mean function with no jump-point discontinuities. This "critical" bandwidth technique is similar to a test for multimodality in density estimation (Silverman 1981) and a test for monotonicity of regression (Bowman et al. 1998).

Many of the specific details required to implement the proposed critical bandwidth jump point test addressed in this dissertation are of practical value outside of the testing setting. For example, determining if the jump preserving M-smoother satisfies the null hypothesis (i.e. has no jumps) leads to estimates of all the jump point positions and sizes. We also suggest a method for extending the jump preserving M-smoother of Chu et al. (1998) to fit any x whereas the original M-smooth fit was limited to the data points x_i.

The proposed hypothesis test in its current form has a number of short comings. The jump point critical bandwidth hypothesis test is extremely computer intensive due to the large number of grid searches required. The level of the test cannot be adjusted and seems to be fixed at a point much lower than usually desired ($\alpha = 0.05$). As with many extremely conservative methods our hypothesis test seems to suffer lower power than one might desire. Despite these downfalls, through enhanced modifications, the jump point critical bandwidth testing procedure may have the potential to be a faster and more powerful method.

5.3 Areas for Future Research

Much work is required to determine when an M-smooth fit will contain at least one jump. A mathematical formula in terms of g and h for the number of jump points in the M-smoother would be quite helpful in the advancement of our testing procedure. For example, it can be shown that for small fixed h $N^* = \sum_{i=1}^{n-1} I_{|Y_{i+1}-Y_i|>cg}$, where N^* represents the number of jump points in the M-smooth fit and c is a function of h. If a similar function could be constructed for larger h we could then develop a much faster algorithm for calculating g_{CRIT}. Furthermore, such a formula might provide information into the null distribution of g_{CRIT}, thus eliminating the need for a bootstrap sample. This would certainly increase the speed of the

testing procedure and might positively affect the level and power of the test. However the search for such a formula will be very difficult and may not be mathematically tractable. If such an exact formula cannot be calculated, then perhaps an approximate version based on a Taylor series expansion may be of use.

Another future research idea would be the development of a fast root searching technique that would not break down for extremely small g. Perhaps the IRLS or Newton methods could be modified so that they converge to the correct root even when g is small.

Once the speed of our test is addressed, larger more extensive simulation studies need to be conducted to assess the power of our test compared to other existing methods.

Another future research area briefly discussed in Section 4.3.1 is the estimation of the number, size(s), and position(s) of the jump points in a mean function. Although we outlined a method for calculating these estimates no studies have been conducted to evaluate their small sample or asymptotic properties.

One subtlety that is involved in the bandwidth formula for h_{OPT} is the way that the error density $f(\cdot)$ is replaced by a normal density with mean zero and variance σ. All simulation examples studied thus far have had normal error variances. Thus, the robustness properties of our bandwidth selection procedures RROT and DPI to other error variances are currently unknown. Of particular interest would be the performance of our methods and the M-smoother when the distribution of the error variance is a heavy tailed distribution. It may be quite difficult for the M-smoother to distinguish noise from a heavy tailed distribution from jump points in the underlying mean function.

Another area of interesting future research is the extension of the M-smoother and our testing procedure to design spaces that are not equally spaced. Some accommodation would need to be made to ensure that all determinants in the M-smooth calculation are non-singular. We would also need to redefine the M-smoother to be able to handle multiple response data points for the same regressor.

Jump point discontinuities in the derivative of a mean function often correspond to sharp cusps in the mean function. It seems feasible that the M-smoother and our hypothesis test could be extended to jumps

in the derivative of a mean function.

The above future research ideas include just a fraction of the questions that relate to the jump preserving M-smoother and our hypothesis test. Many of the future research questions behind Chu et al.'s (1998) M-smoother could not begin to be answered if it were not for the development of our automatic bandwidth selection procedure. In that sense, our DPI bandwidth selection procedure acts as a key opening many other practical applications and modifications of the M-smooth method of Chu et al. (1998).

Bibliography

Akaike, H. (1970), 'Statistical predictor information', *Annals of the Institute of Statistical Mathematics* **22**, 203–217.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **19**, 716–723.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. & Tukey, J. W. (1972), *Robust Estimates of Location: Survey and Advances*, Princeton University Press, Princeton, New Jersey.

Beaton, A. E. & Tukey, J. W. (1974), 'The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data', *Technometrics* **16**, 147–185.

Birch, J. B. (1997), Exploratory and robust data analysis. Unpublished Lecture Notes.

Bowman, A. W., Jones, M. C. & Gijbels, I. (1998), 'Testing monotonicity of regression', *Journal of Computational and Graphical Statistics* **7**(4), 489–500.

Box, G. E. P., Jenkins, G. M. & Reinsel, G. C. (1994), *Time Series Analysis*, 3 edn, Prentice Hall, Englewood Cliffs, New Jersey.

Box, G. E. P. & Tiao, G. C. (1965), 'A change in level of a non-stationary time series', *Biometrika* **52**, 181–192.

Box, G. E. P. & Tiao, G. C. (1975), 'Intervention analysis with applications to economic and enviromental problems', *Journal of the American Statistical Association* **70**, 70–79.

Brown, R. L., Durbin, J. & Evans, J. M. (1975), 'Techniques for testing the constancy of regression relationships over time', *Journal of the Royal Statistical Society, Series B* **37**(2), 149–163.

Burt, D. A. & Coakley, C. W. (2000), 'Automatic bandwidth selection for modified m-smoother', *The Journal of Statistical Computation and Simulation* . To Appear.

Chu, C. K., Glad, I. K., Godtliebsen, F. & Marron, J. S. (1998), 'Edge-preserving smoothers for image processing', *Journal of the American Statistical Association* **93**(442), 526–541.

Chui, C. (1992), *An Introduction to Wavelets*, Academic Press, Boston.

Clark, R. (1975), 'A calibration curve for radio carbon dates', *Antiquity* **49**, 251–266.

Cleveland, W. S. (1979), 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association* **74**, 829–836.

Cobb, G. W. (1978), 'The problem of the Nile: Conditional solution to a changepoint problem', *Biometrika* **65**(2), 243–251.

Craven, P. & Wahba, G. (1979), 'Smoothing noisy data with spline functions', *Numerische Mathematik* **31**, 377–403.

Daubechies, I. (1992), *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Donoho, D. L. (1995), 'Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition', *Appl. Comput. Harm. Anal.* **2**, 101–126.

Donoho, D. L. & Johnstone, I. M. (1994), 'Ideal spatial adaptation via wavelet shrinkage', *Biometrika* **81**, 425–455.

Donoho, D. L. & Johnstone, I. M. (1995), 'Adapting to unknown smoothness via wavelet shrinkage', *Journal of the American Statistical Association* **90**(432), 1200–1224.

Donoho, D. L., Kerkyacharian, G. & Picard, D. (1995), 'Wavelet shrinkage: Asymptopia?', *Journal of the Royal Statistical Society, Series B* **57**, 301–369.

Efron, B. (1979), 'Bootstrap methods: another look at the jack-knife', *Annals of Statistics* **7**, 1–26.

Eubank, R. L. (1988), *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York.

Fan, J. & Gijbels, I. (1996), *Local polynomial modelling and its applications*, Chapman and Hall, London.

Gasser, T. & Müller, H. G. (1979), Kernel estimation of regression functions, *in* 'Smoothing Techniques for Curve Estimation', Vol. 757, Springer-Verlag, New York, pp. 23–68. Lecture Notes in Mathematics.

Gasser, T. & Müller, H. G. (1984), 'Estimating regression functions and their derivatives by the kernel method', *Scand. J. of Statist.* **11**, 171–185.

Glass, G. V., Wilson, V. L. & Gottman, J. M. (1975), *Design and Analysis of Time Series Experiments*, Colorado Associated University Press, Boulder, Colorado.

Green, P. J. & Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*, Chapman and Hall, London.

Hampel, F. R. (1968), Contributions to the theory of robust estimation, PhD thesis, University of California, Berkeley, California.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.

Härdle, W. & Gasser, T. (1984), 'Robust nonparametric function fitting', *Journal of the Royal Statistical Society, Ser. B* **46**, 42–51.

Härdle, W., Hall, P. & Marron, J. S. (1988), 'How far are automatically chosen regression smoothing parameters from thier optimum?', *Journal of the American Statistical Association* **83**, 86–95.

Härdle, W. & Marron, J. S. (1995), 'Fast and simple scatterplot smoothing', *Computational Statistics and Data Analysis* **20**, 1–17.

Hollander, M. & Wolfe, D. (1973), *Nonparametric Statistical Methods*, Wiley, New York, New York.

Huber, P. J. (1964), 'Robust estimation of a location parameter', *Annals of Mathematical Statistics* **35**, 73–101.

James, B., James, K. L. & Siegmund, D. (1987), 'Tests for a change-point', *Biometrika* **74**(1), 71–83.

Karlin, S. (1968), *Total Positivity*, Vol. 1, Stanford University Press, Stanford, California.

Kim, H.-J. & Siegmund, D. (1989), 'The likelihood ratio test for a change-point in simple linear regression', *Biometrika* **76**(3), 409–423.

Leung, D. H. Y., Marriott, F. H. C. & Wu, E. K. H. (1993), 'Bandwidth selection in robust smoothing', *Journal of Nonparametric Statistics* **2**, 333–339.

Loader, C. R. (1996), 'Change point estimation using nonparametric regression', *The Annals of Statistics* **24**, 1667–1678.

Mallows, C. L. (1973), 'Some comments on Cp', *Technometrics* **15**, 661–675.

Müller, H. G. (1992), 'Change-points in nonparametric regression analysis', *The Annals of Statistics* **20**, 737–761.

Nadaraya, E. A. (1964), 'On estimating regression', *Theory of Probablility and Its Applications* **9**, 141–142.

Qiu, P. & Yandell, B. (1998), 'Local polynomial jump-detection algorithm in nonparametric regression', *Technometrics* **40**(2), 141–152.

Rice, J. (1984), 'Bandwidth choice for nonparametric regression', *The Annals of Statistics* **12**, 1215–1230.

Rousseeuw, P. J. (1984), 'Least median of squares regression', *Journal of the American Statistical Association* **79**, 871–880.

Rue, H., Chu, C. K., Godtliebsen, F. & Marron, J. S. (1998), M-smoother with local linear fit. Unpublished manuscript.

Ruppert, D., Sheather, S. J. & Wand, M. P. (1995), 'An effective bandwidth selector for local least squares regression', *Journal of the American Statistical Association* **90**(432), 1257–1270.

Schoenberg, I. J. (1950), 'On pólya frequency functions. II: Variation-diminishing integral operators of the convolution type', *Acta Scientarum Mathematicarum Szeged* **12B**, 97–106.

Shaban, S. A. (1980), 'Change point problem and two-phase regression: an annotated bibliography', *International Statistical Review* **48**, 83–93.

Silverman, B. W. (1981), 'Using kernel density estimates to investigate multimodality', *Journal of the Royal Statistical Society B* **43**(1), 97–99.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Simpson, D. G., He, X. & Liu, Y.-T. (1998), 'Comment on edge-preserving smoothers by Chu, Glad, Godtliebsen, and Marron', *Journal of the American Statistical Association* **93**(442), 544–548.

Strang, G. (1989), 'Wavelets and dilation equations: a brief introduction', *SIAM review* **31**, 614–627.

Sutherland, S. S. (1992), Sequential Design Augmentation With Model Misspecification, PhD thesis, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Wahba, G. (1990), *Spline models for observational data*, SIAM, Philadelphia.

Watson, G. S. (1964), 'Smooth regression analysis', *Sankhya Ser. A* **26**, 359–372.

Wu, J. S. & Chu, C. K. (1993), 'Kernel-type estimators of jump points and values of a regression function', *The Annals of Statistics* **21**, 1545–1566.

Appendix A

Proof of Theorem 3.1

Since the bandwidth parameter g in the M-smooth defining equation cannot be chosen using asymptotic bias and variance formulas, we investigate the asymptotic relative efficency of the Gaussian Psi M-estimate versus the Least Squares estimate in the location estimation setting. In the location parameter estimation setting the asymptotic relative efficiency of the M-estimate with Gaussian Ψ versus the least squares estimate assuming the data are normally distributed is given by

$$ARE[M_G, LS] = \sigma^2 / \left(\frac{\int \Psi^2 dF}{(\int \Psi' dF)^2} \right).$$

If the loss function $-L_g$ is chosen as the negative of the Gaussian distribution with scale parameter g as suggested above, then we have

$$-L_g(r) = \frac{-1}{\sqrt{2\pi} g} \exp\left(\frac{-r^2}{2g^2}\right).$$

This choice of a loss function gives rise to the following Ψ function

$$\Psi(r) = \frac{-\partial}{\partial r} L_g(r) = \frac{r}{\sqrt{2\pi} g^3} \exp\left(\frac{-r^2}{2g^2}\right).$$

Similarly

$$\Psi'(r) = \frac{\partial}{\partial r} \Psi(r) = \frac{1}{\sqrt{2\pi} g^3} \exp\left(\frac{-r^2}{2g^2}\right) - \frac{r^2}{\sqrt{2\pi} g^5} \exp\left(\frac{-r^2}{2g^2}\right).$$

Thus

$$\int \Psi^2 dF = \int \big(\frac{r}{\sqrt{2\pi}g^3}\exp\big(\frac{-r^2}{g^2}\big)\big)^2 dF.$$

Since the underlying distribution is assumed to be Normal(0,σ^2), we have

$$\begin{aligned}
\int \Psi^2 dF &= \int \big(\frac{r}{\sqrt{2\pi}g^3}\exp\big(\frac{-r^2}{2g^2}\big)\big)^2 \big(\frac{1}{\sqrt{2\pi}\sigma}\exp\big(\frac{-r^2}{2\sigma^2}\big)\big)dr \\
&= \frac{1}{2\pi g^6\sigma}\int \frac{r^2}{\sqrt{2\pi}}\exp\big(\frac{-r^2}{g^2}+\frac{-r^2}{2\sigma^2}\big)dr \\
&= \frac{1}{2\pi g^6\sigma}\sqrt{\big(\frac{\sigma^2 g^2}{2\sigma^2+g^2}\big)}\int \frac{r^2}{\sqrt{2\pi\big(\frac{\sigma^2 g^2}{2\sigma^2+g^2}\big)}}\exp\left(-r^2/\left(2\big(\frac{\sigma^2 g^2}{2\sigma^2+g^2}\big)\right)\right)dr \\
&= \frac{\sigma^2}{2\pi g^3(2\sigma^2+g^2)^{3/2}}. \qquad \text{(A.1)}
\end{aligned}$$

Similarly we have

$$\begin{aligned}
\big(\int \Psi' dF\big)^2 &= \left(\int \frac{1}{\sqrt{2\pi}g^3}\exp\big(\frac{-r^2}{2g^2}\big)dF - \int \frac{r^2}{\sqrt{2\pi}g^5}\exp\big(\frac{-r^2}{2g^2}\big)dF\right)^2 \\
&= \left(\frac{1}{\sqrt{2\pi(g^2+\sigma^2)}g^2}\int \frac{\exp\big(\frac{-r^2}{2\left(\frac{g^2\sigma^2}{g^2+\sigma^2}\right)}\big)}{\sqrt{2\pi\big(\frac{g^2\sigma^2}{g^2+\sigma^2}\big)}}dr - \frac{1}{\sqrt{2\pi(g^2+\sigma^2)}g^4}\int \frac{r^2\exp\big(\frac{-r^2}{2\left(\frac{g^2\sigma^2}{g^2+\sigma^2}\right)}\big)}{\sqrt{2\pi\big(\frac{g^2\sigma^2}{g^2+\sigma^2}\big)}}dr\right)^2 \\
&= \left(\frac{1}{\sqrt{2\pi(g^2+\sigma^2)}g^2} - \frac{1}{\sqrt{2\pi(g^2+\sigma^2)}g^2}\frac{\sigma^2}{g^2+\sigma^2}\right)^2 \\
&= \frac{1}{2\pi(g^2+\sigma^2)g^4}\left(1-\frac{\sigma^2}{g^2+\sigma^2}\right)^2 \\
&= \frac{1}{2\pi(g^2+\sigma^2)^3}. \qquad \text{(A.2)}
\end{aligned}$$

Substituting we have

$$\begin{aligned}
ARE[M_G, LS] &= \sigma^2/\left(\frac{\frac{\sigma^2}{2\pi g^3(2\sigma^2+g^2)^{3/2}}}{\frac{1}{2\pi(g^2+\sigma^2)^3}}\right) \\
&= \frac{g^3(2\sigma^2+g^2)^{3/2}}{(g^2+\sigma^2)^3}. \qquad \text{(A.3)}
\end{aligned}$$

If we set $g = \alpha\sigma$ where α is a positive constant we have

$$\begin{aligned}
ARE[M_G, LS] &= \frac{\alpha^3\sigma^3(2\sigma^2+\alpha^2\sigma^2)^{3/2}}{(\alpha^2\sigma^2+\sigma^2)^3} \\
&= \frac{\alpha^3\sigma^3(2+\alpha^2)^{3/2}\sigma^3}{\sigma^6(\alpha^2+1)^3} \\
&= \frac{\alpha^3(2+\alpha^2)^{3/2}}{(\alpha^2+1)^3}. \qquad \text{(A.4)}
\end{aligned}$$

Solving $ARE[M_G, LS] = 0.95$ for the constant α yields $\alpha = 2.11$. Therefore, in order for the M_G location estimate to maintain a 95% asymptotic relative efficiency with the least squares estimate when the distribution is Normal(0,σ) the tuning/scale parameter g should be chosen as 2.11σ. When σ is not known we suggest choosing $g = 2.11\hat{\sigma}$.

Appendix B

Proof of Theorem 3.2

To find the optimal bandwidth parameters g and h, we expand the Taylor series expansion of the asymptotic bias and asymptotic variance given in Rue et al. (1998). As mentioned before we calculate the IAMSE assuming the entire design space is in the smooth region R_1. Recall the following notation from Rue et al. (1998).

For each $x_i \in R_1$, set

$$\begin{aligned}
U_0 &= \hat{a} - m(x_i), \\
U_1 &= \hat{b} + m^{(1)}(x_i), \\
D_j &= Y_j - m(x_i) + m^{(1)}(x_i)(x_i - x_j), \\
\hat{D}_j &= (-1)U_0 - U_1(x_i - x_j), \\
S_k &= n^{-1} \sum_{j=1}^{n} K_h(x_i - x_j) L^{(2)}(g^{-1} D_j)(x_i - x_j)^k, \\
T_k &= n^{-1} \sum_{j=1}^{n} K_h(x_i - x_j) L^{(1)}(g^{-1} D_j)(x_i - x_j)^k.
\end{aligned}$$

We will also make use of the expected value and variance formulas given by Rue et al. (1998), which are as follows:

$$E[S_k] = h^k g^3 f^{(2)}(0) k_k [1 + \mathcal{O}(n^{-1}h^{-1} + g^2)], \tag{B.1}$$

$$Var[S_k] = n^{-1}h^{2k-1} g f(0) \int_{-1}^{1} L^{(2)}(u)^2 du \tau_{2k} [1 + \mathcal{O}(n^{-1}h^{-1} + g^2)], \tag{B.2}$$

$$E[T_k] = h^{k+2} g^2 f^{(2)}(0)(1/2) m^{(2)}(x_i) k_{k+2} [1 + \mathcal{O}(n^{-1}h^{-1} + h^{-2}g^3)], \tag{B.3}$$

$$Var[T_k] = n^{-1}h^{2k-1} g f(0) \int_{-1}^{1} L^{(1)}(u)^2 du \tau_{2k} [1 + \mathcal{O}(n^{-1}h^{-1} + g^2)]. \tag{B.4}$$

Lemma B.1 *The asymptotic bias of the M-smoother developed by Rue et al. (1998) for $x_i \in R_1$ is given by*

$$AB = \frac{1}{2}h^2 m^{(2)}(x_i) k_2 + \frac{1}{2}\left(\frac{h}{ng^5}\right) m^{(2)}(x_i) \beta^* [k_2 \tau_0 + \tau_2] + \mathsf{op}\left(\frac{h}{ng^5}\right). \tag{B.5}$$

where $\beta^* = f(0) f^{(2)}(0)^{-2} \int L^{(2)}(u)^2 du$.

Proof:

First note that since the M-smoother is estimated by $\hat{a}$ the bias is equal to $E(U_0)$. Recall Rue et al. (1998) showed that U_0 can be approximated by

$$U_0 \approx (gT_0S_2 - gT_1S_1)/(S_0S_2 - S_1^2).$$

Thus the second order Taylor expansion for $E(U_0)$ is given by:

$$\begin{aligned}
E(U_0) \approx\ & U_0\Big|_{E(S_0),E(S_1),E(S_2),E(T_0),E(T_1)} \\
&+ \frac{1}{2}\left(\frac{\partial^2 U_0}{\partial T_0^2}\Big|_{E(S_0),E(S_1),E(S_2),E(T_0),E(T_1)}\right) Var(T_0) \\
&+ \frac{1}{2}\left(\frac{\partial^2 U_0}{\partial T_1^2}\Big|_{E(S_0),E(S_1),E(S_2),E(T_0),E(T_1)}\right) Var(T_1) \\
&+ \frac{1}{2}\left(\frac{\partial^2 U_0}{\partial S_0^2}\Big|_{E(S_0),E(S_1),E(S_2),E(T_0),E(T_1)}\right) Var(S_0) \\
&+ \frac{1}{2}\left(\frac{\partial^2 U_0}{\partial S_1^2}\Big|_{E(S_0),E(S_1),E(S_2),E(T_0),E(T_1)}\right) Var(S_1) \\
&+ \frac{1}{2}\left(\frac{\partial^2 U_0}{\partial S_2^2}\Big|_{E(S_0),E(S_1),E(S_2),E(T_0),E(T_1)}\right) Var(S_2).
\end{aligned} \tag{B.6}$$

Throughout this proof it will be important to note that $E(S_1) = 0$ *due to the fact that the all the odd moments of the kernel* K *are zero. Thus the first term of (B.6) is*

$$\begin{aligned} U_0\Big|_{E(S_0),E(S_1),E(S_2),E(T_0),E(T_1)} &\approx \left(\frac{gE(T_0)E(S_2) - gE(T_1)E(S_1)}{E(S_0)E(S_2) - [E(S_1)]^2}\right) \\ &= \frac{gE(T_0)}{E(S_0)}. \end{aligned}$$

Using the approximations in (B.1) and (B.3) we have

$$\begin{aligned} U_0\Big|_{E(S_0),E(S_1),E(S_2),E(T_0),E(T_1)} &\approx g\frac{h^2g^2f^{(2)}(0)(1/2)m^{(2)}(x_i)k_2}{g^3f^{(2)}(0)k_0} \\ &= (1/2)h^2m^{(2)}(x_i)k_2. \end{aligned}$$

Notice that $\frac{\partial^2 U_0}{\partial T_0^2} \approx 0$, *and* $\frac{\partial^2 U_0}{\partial T_1^2} \approx 0$. *Thus the second and third term of (B.6) drop out.*

To find the fourth term we first calculate $\frac{\partial U_0}{\partial S_0}$ *to be*

$$\frac{\partial U_0}{\partial S_0} = (-S_2)\frac{gT_0S_2 - gT_1S_1}{(S_0S_2 - S_1^2)^2}.$$

Thus

$$\frac{\partial^2 U_0}{\partial S_0^2} = (2S_2^2)\frac{gT_0S_2 - gT_1S_1}{(S_0S_2 - S_1^2)^3}.$$

Replacing the S_k*'s and* T_k*'s with their expected values, keeping in mind that* $E[S_1] = 0$ *due to the zero odd moments of the kernel* K*, we have*

$$2[E(S_2)]^2\frac{gE(T_0)E(S_2)}{(E(S_0)E(S_2))^3} = \frac{2gE(T_0)}{(E(S_0))^3}.$$

Applying the formulas given in (B.1) and (B.3) and multiplying by $(1/2)$ *and the* $Var[S_0]$ *given by (B.2), we see that the fourth term of (B.6) is asymptotically*

$$\begin{aligned} (1/2)\frac{2gh^2g^2f^{(2)}(0)(1/2)m^{(2)}(x_i)k_2}{[g^3f^{(2)}(0)k_0]^3}n^{-1}h^{-1}gf(0)\int_{-1}^{1}L^{(2)}(u)^2du\tau_0 \\ = (1/2)\Big(\frac{h}{ng^5}\Big)m^{(2)}(x_i)k_2\beta^*\tau_0, \end{aligned}$$

where $\beta^* = f(0)f^{(2)}(0)^{-2}\int L^{(2)}(u)^2du$.

For the fifth term of (B.6) we focus on $\frac{\partial^2 U_0}{\partial S_1^2}$. *Notice that*

$$\begin{aligned}\frac{\partial U_0}{\partial S_1} &\approx \frac{(S_0S_2 - S_1^2)(-gT_1) - (gT_0S_2 - gT_1S_1)(-2S_1)}{(S_0S_2 - S_1^2)^2} \\ &= \frac{gT_1S_1^2 - gT_1S_0S_2 + 2gT_0S_1S_2 - 2gT_1S_1^2}{(S_0S_2 - S_1^2)^2}.\end{aligned}$$

Thus

$$\begin{aligned}\frac{\partial^2 U_0}{\partial S_1^2} &\approx \frac{-(gT_1S_1^2 - gT_1S_0S_2 + 2gT_0S_1S_2 - 2gT_1S_1^2)}{(S_0S_2 - S_1^2)^4}2(S_0S_2 - S_1^2)(-2S_1) \\ &+ \frac{(S_0S_2 - S_1^2)^2(2gT_1S_1 + 2gT_0S_2 - 4gT_1S_1)}{(S_0S_2 - S_1^2)^4}.\end{aligned}$$

When we substitute the expected values in (keeping in mind $E(S_1) = 0$*) we have*

$$\begin{aligned}\left.\frac{\partial^2 U_0}{\partial S_1^2}\right|_{E(S_0),E(S_1),E(S_2),E(T_0),E(T_1)} &\approx \frac{[E(S_0)E(S_2)]^2[2gE(T_0)E(S_2)]}{[E(S_0)E(S_2)]^4} \\ &= \frac{2gE(T_0)}{[E(S_0)]^2E(S_2)}.\end{aligned}$$

Thus from (B.1) and (B.3) we have

$$\begin{aligned}\left.\frac{\partial^2 U_0}{\partial S_1^2}\right|_{E(S_0),E(S_1),E(S_2),E(T_0),E(T_1)} &\approx \frac{2gh^2g^2f^{(2)}(0)(1/2)m^{(2)}(x_i)k_2}{g^6f^{(2)}(0)^2k_0^2h^2g^3f^{(2)}(0)k_2} \\ &= g^{-6}f^{(2)}(0)^{-2}m^{(2)}(x_i).\end{aligned}$$

Multiplying by $(1/2)$ *and* $Var(S_1)$ *yields the following for the fifth term of (B.6)*

$$\begin{aligned}&(1/2)g^{-6}f^{(2)}(0)^{-2}m^{(2)}(x_i)n^{-1}hgf(0)\int_{-1}^{1}L^{(2)}(u)^2du\tau_2 \\ = \;& (1/2)\Big(\frac{h}{ng^5}\Big)m^{(2)}(x_i)\beta^*\tau_2.\end{aligned}$$

For the last term of (B.6) we have

$$\begin{aligned}\frac{\partial^2 U_0}{\partial S_2^2} &\approx \frac{\partial}{\partial S_2}\left(\frac{(S_0S_2 - S_1^2)(gT_0) - (gT_0S_2 - gT_1S_1)S_0}{(S_0S_2 - S_1^2)^2}\right) \\ &= \frac{\partial}{\partial S_2}\left(\frac{gT_0S_0S_2 - gT_0S_1^2 - gT_0S_0S_2 + gT_1S_0S_1}{(S_0S_2 - S_1^2)^2}\right)\end{aligned}$$

$$
\begin{aligned}
&= \frac{\partial}{\partial S_2}\left(\frac{gT_1S_0S_1 - gT_0S_1^2}{(S_0S_2 - S_1^2)^2}\right) \\
&= -2\left(\frac{gT_1S_0S_1 - gT_0S_1^2}{(S_0S_2 - S_1^2)^3}(S_0)\right).
\end{aligned}
$$

Notice that when $E(S_1) = 0$ is substituted in for S_1, this term drops out.

By collecting terms, we have shown (B.5) to be the asymptotic bias of the M-smoother estimator.

The next step in calculating $\mathsf{h}_{\mathsf{opt}}$ and $\mathsf{g}_{\mathsf{opt}}$ is to expand the asymptotic variance, which we do in the next lemma.

Lemma B.2 *The asymptotic variance of the M-smoother developed by Rue et al. (1998) for $x_i \in R_1$ is given by*

$$
AV = (nhg^3)^{-1}\beta\tau_0 + (1/4)\left(\frac{h^3}{ng^5}\right)m^{(2)}(x_i)^2\beta^*(k_2^2\tau_0) + \mathsf{op}\left(\frac{h^3}{ng^5}\right) \tag{B.7}
$$

where $\beta = f(0)f^{(2)}(0)^{-2}\int L^{(1)}(u)^2du$.

Proof:

We first observe that $Var(\hat{a}) = Var(\hat{a} - m(x_i)) = Var(U_0)$. From the delta method we have

$$
\begin{aligned}
Var(U_0) \approx & \left(\frac{\partial U_0}{\partial T_0}\Big|_{E(S_0),E(S_1),E(S_2),E(T_0),E(T_1)}\right)^2 Var(T_0) \\
&+ \left(\frac{\partial U_0}{\partial T_1}\Big|_{E(S_0),E(S_1),E(S_2),E(T_0),E(T_1)}\right)^2 Var(T_1) \\
&+ \left(\frac{\partial U_0}{\partial S_0}\Big|_{E(S_0),E(S_1),E(S_2),E(T_0),E(T_1)}\right)^2 Var(S_0) \\
&+ \left(\frac{\partial U_0}{\partial S_1}\Big|_{E(S_0),E(S_1),E(S_2),E(T_0),E(T_1)}\right)^2 Var(S_1) \\
&+ \left(\frac{\partial U_0}{\partial S_2}\Big|_{E(S_0),E(S_1),E(S_2),E(T_0),E(T_1)}\right)^2 Var(S_2)
\end{aligned} \tag{B.8}
$$

To find the first term of (B.8) we calculate $\frac{\partial U_0}{\partial T_0}$ to be

$$
\frac{\partial U_0}{\partial T_0} \approx \frac{gS_2}{S_0S_2 - S_1^2}
$$

Thus

$$\begin{aligned}\left.\frac{\partial U_0}{\partial T_0}\right|_{E(S_0),E(S_1),E(S_2),E(T_0),E(T_1)} &\approx \frac{gE(S_2)}{E(S_0)E(S_2)-[E(S_1)]^2}\\ &= \frac{g}{E(S_0)}\\ &= \frac{g}{g^3 f^{(2)}(0)k_0}\\ &= g^{-2}f^{(2)}(0)^{-1}.\end{aligned}$$

Squaring the above and multiplying by $Var(T_0)$ yields

$$\begin{aligned}& g^{-4}f^{(2)}(0)^{-2}n^{-1}h^{-1}gf(0)\int_{-1}^{1}L^{(1)}(u)^2du\tau_0\\ = \;& (nhg^3)^{-1}\beta\tau_0,\end{aligned}$$

where $\beta = f(0)f^{(2)}(0)^{-2}\int L^{(1)}(u)^2du$.

To find the second term of (B.8) we calculate $\frac{\partial U_0}{\partial T_1}$ to be

$$\frac{\partial U_0}{\partial T_1} \approx \frac{gS_1}{S_0S_2-S_1^2}$$

Thus

$$\begin{aligned}\left.\frac{\partial U_0}{\partial T_1}\right|_{E(S_0),E(S_1),E(S_2),E(T_0),E(T_1)} &\approx \frac{gE(S_1)}{E(S_0)E(S_2)-[E(S_1)]^2}\\ &= 0.\end{aligned}$$

Therefore the second term of (B.8) drops out.

To find the third term we find

$$\frac{\partial U_0}{\partial S_0} \approx -S_2\frac{gT_0S_2-gT_1S_1}{(S_0S_2-S_1^2)^2}$$

When we evaluate the above at the expected values found in (B.1) and (B.3) we have

$$\begin{aligned}& -E(S_2)\frac{gE(T_0)E(S_2)}{[E(S_0)E(S_2)]^2}\\ = \;& \frac{-gE(T_0)}{[E(S_0)]^2}\\ = \;& \frac{-gh^2g^2f^{(2)}(0)(1/2)m^{(2)}(x_i)k_2}{g^6f^{(2)}(0)^2k_0^2}\\ = \;& (-1/2)h^2g^{-3}f^{(2)}(0)^{-1}m^{(2)}(x_i)k_2.\end{aligned}$$

Squaring and multiplying by $Var(S_0)$ yields

$$\begin{aligned} & (1/4)h^4 g^{-6} f^{(2)}(0)^{-2} m^{(2)}(x_i)^2 k_2^2 n^{-1} h^{-1} g f(0) \int_{-1}^{1} L^{(2)}(u)^2 du \tau_0 \\ = \quad & (1/4)\left(\frac{h^3}{ng^5}\right) m^{(2)}(x_i)^2 k_2^2 \beta^* \tau_0, \end{aligned}$$

where $\beta^* = f(0) f^{(2)}(0)^{-2} \int L^{(2)}(u)^2 du$.

For the fourth term we find

$$\frac{\partial U_0}{\partial S_1} \approx \frac{(S_0 S_2 - S_1^2)(-gT_1) + 2(gT_0 S_2 - gT_1 S_1) S_1}{(S_0 S_2 - S_1^2)^2}.$$

Evaluating the above with the expected values from (B.1) and (B.3) yields the following

$$\begin{aligned} & \frac{-gE(S_0)E(S_2)E(T_1)}{[E(S_0)E(S_2)]^2} \\ = \quad & \frac{-gh^3 g^2 f^{(2)}(0)^{-2}(1/2) m^{(2)}(x_i) k_3}{g^3 f^{(2)}(0) k_0 * h^2 g^3 f^{(2)}(0) k_2}. \end{aligned}$$

Since the odd moments of the kernel K (specifically k_3) are zero, the above term drops out.

The final term of (B.8) is found by

$$\frac{\partial U_0}{\partial S_2} \approx \frac{(S_0 S_2 - S_1^2)(gT_0) - (gT_0 S_2 - gT_1 S_1) S_0}{(S_0 S_2 - S_1^2)^2}.$$

Substituting the expected values (keeping in mind $E(S_1) = 0$) yields

$$\begin{aligned} & \frac{[E(S_0)E(S_2)](gE(T_0)) - [gE(T_0)E(S_2)]E[S_0]}{[E(S_0)E(S_2)]^2} \\ = \quad & \frac{gE(S_0)E(S_2)E(T_0) - gE(S_0)E(S_2)E(T_0)}{[E(S_0)E(S_2)]^2} \\ = \quad & 0. \end{aligned}$$

Thus the last term drops out as well.

Combining the terms gives us (B.7) and completes the proof.

By squaring the asymptotic bias and adding to the asymptotic variance we calculate the asymptotic mean squared error in the next lemma.

Lemma B.3 *The asymptotic mean squared error (AMSE) for the M-smoother can be shown to be*

$$\begin{aligned} AMSE &= \left((1/2)h^2 m^{(2)}(x_i)k_2 + (1/2)\left(\frac{h}{ng^5}\right)m^{(2)}(x_i)\beta^*[k_2\tau_0+\tau_2]\right)^2 + \mathsf{op}\left(\frac{h^2}{n^2g^{10}}\right) \\ &+ (nhg^3)^{-1}\beta\tau_0 + (1/4)\left(\frac{h^3}{ng^5}\right)m^{(2)}(x_i)^2\beta^*(k_2^2\tau_0) + \mathsf{op}\left(\frac{h^3}{ng^5}\right) \\ &= (1/4)h^4 m^{(2)}(x_i)^2k_2^2 + (1/4)\left(\frac{h^3}{ng^5}\right)m^{(2)}(x_i)^2\beta^*[k_2^2\tau_0+\tau_2] + (1/4)\left(\frac{h}{ng^5}\right)^2 m^{(2)}(x_i)^2(\beta^*)^2[k_2\tau_0+\tau_2]^2 \\ &+ (nhg^3)^{-1}\beta\tau_0 + (1/4)\left(\frac{h^3}{ng^5}\right)m^{(2)}(x_i)^2\beta^*(k_2^2\tau_0) + \mathsf{op}\left(\frac{h^2}{n^2g^{10}}\right). \end{aligned}$$

If we ignore the $\left(\frac{h}{ng^5}\right)^2$ *term above which is quite small compared to the other terms, we have*

$$\begin{aligned} AMSE &= (1/4)h^4 m^{(2)}(x_i)^2k_2^2 + (1/4)\left(\frac{h^3}{ng^5}\right)m^{(2)}(x_i)^2\beta^*[2k_2^2\tau_0+\tau_2] \\ &+ (nhg^3)^{-1}\beta\tau_0 + \mathsf{Op}\left(\frac{h^2}{n^2g^{10}}\right). \end{aligned} \tag{B.9}$$

Since we wish to construct global constant optimal bandwidths for g and h we need a global loss function. The above AMSE is a local loss function for x_i's away from the boundary and jump points. The global loss function we suggest is the Integrated Asymptotic Mean Squared Error (IAMSE). Here we simply integrate the AMSE over the range of the x_i's. Since the jump points and the boundary points make up a set of measure zero, they will minimally effect the calculation of the IAMSE. Thus we will integrate the AMSE assuming all $x_i \in R_1$. The proposed IAMSE is found by

$$\begin{aligned} \mathsf{IAMSE} &= \int\left((1/4)h^4 m^{(2)}(x)^2k_2^2 + (1/4)\left(\frac{h^3}{ng^5}\right)m^{(2)}(x)^2\beta^*[2k_2^2\tau_0+\tau_2] + (nhg^3)^{-1}\beta\tau_0\right)w(x)dx \\ &+ \mathsf{Op}\left(\frac{h^2}{n^2g^{10}}\right) \\ &= \left[(1/4)h^4k_2^2(1/4)\left(\frac{h^3}{ng^5}\right)\beta^*[2k_2^2\tau_0\tau_2]\right]\int m^{(2)}(x)^2w(x)dx \\ &+ (nhg^3)^{-1}\beta\tau_0\int w(x)dx + \mathsf{Op}\left(\frac{h^2}{n^2g^{10}}\right). \end{aligned}$$

The $w(x)$ in the above equation represents the underlying design density of the x's. If we assume

the x_i's are equally spaced from 0 to 1, then $w(x) = 1$ for all $x \in [0,1]$ and $w(x) = 0$ elsewhere. Thus $\int w(x)dx = 1$ for equal spaced x's from 0 to 1.

Thus for equal spaced x_i's we have

$$\begin{aligned} \mathsf{IAMSE} &= \left[(1/4)h^4k_2^2 + (1/4)\left(\frac{h^3}{ng^5}\right)\beta^*[2k_2^2\tau_0 + \tau_2]\right]\int m^{(2)}(x)^2dx \\ &+ (nhg^3)^{-1}\beta\tau_0 + \mathsf{Op}\left(\frac{h^2}{n^2g^{10}}\right). \end{aligned} \tag{B.10}$$

The proof of Theorem 3.2 is now complete.

Appendix C

Proof Of Theorem 4.1

We begin the proof of Theorem 4.1 by stating two definitions and the main theorem presented in a paper by Schoenberg (1950).

Definition C.1 *A function* $\Lambda(x), -\infty < x < \infty$, *is called a Pólya frequency function if it satisfies the following three characteristic conditions (Schoenberg 1950):*

i. $\Lambda(x)$ *is measurable.*

ii. If $x_1 < x_2 < \cdots < x_n$ *and* $t_1 < t_2 < \cdots < t_n$ *then*

$$\det \begin{pmatrix} \Lambda(x_1 - t_1) & \Lambda(x_2 - t_1) & \cdots & \Lambda(x_n - t_1) \\ \Lambda(x_1 - t_2) & \Lambda(x_2 - t_2) & \cdots & \Lambda(x_n - t_2) \\ \vdots & \vdots & \vdots & \vdots \\ \Lambda(x_1 - t_n) & \Lambda(x_2 - t_n) & \cdots & \Lambda(x_n - t_n) \end{pmatrix} \geq 0.$$

iii.

$$0 < \int_{-\infty}^{\infty} \Lambda(x) dx < +\infty.$$

Definition C.2 *Consider the integral transformation*

$$g(x) = \int_{-\infty}^{\infty} f(x-t)dL(t), \tag{C.1}$$

where $f(x)$ is an arbitrary continuous and bounded function. We say that (C.1) is variation-diminishing if it always implies the inequality $\xi(g) \leq \xi(f)$, where $\xi(\cdot)$ represents the number of sign changes in $(\cdot)$ as the argument traverses the domain of definition from left to right (Schoenberg 1950).

Theorem C.1 *The transformation C.1 is variation-diminishing if and only if $L(t)$ is either, up to the sign, a cumulative Pólya frequency function*

$$L(t) = \varepsilon \int_{-\infty}^{t} \Lambda(u)du,$$

where $\varepsilon = \pm 1$ and $\Lambda(x)$ is a Pólya frequency function (Schoenberg 1950).

Note that

$$\frac{Y_j - a}{\sqrt{2\pi}g_u} \exp\{-0.5(\frac{Y_j - a}{g_u})^2\}$$

can be rewritten in convolution form as

$$\begin{aligned} & \frac{Y_j - a}{\sqrt{2\pi}g_u} \exp\{-0.5(\frac{Y_j - a}{g_u})^2\} \\ = & \int_{-\infty}^{\infty} \frac{Y_j - \theta}{\sqrt{2\pi(g_u^2 - g_l^2)}} \exp\{-0.5\frac{(a-\theta)^2}{g_u^2 - g_l^2}\} \frac{1}{\sqrt{2\pi}g_l} \exp\{-0.5(\frac{\theta - Y_j}{g_l})^2\}d\theta, \end{aligned} \tag{C.2}$$

provided $0 < g_l < g_u$ (see Silverman (1981)).

We now can now substitute the convolution representation of (C.2) into the formula for $\frac{\partial}{\partial a}S(a, x_i|g_u)$ and have

$$\begin{aligned} \frac{\partial}{\partial a}S(a, x_i|g_u) &= \sum_{j=1}^{n} \frac{-1}{2\pi h g_u^3} \exp\{-0.5(\frac{x - x_j}{h})^2\} \exp\{-0.5(\frac{Y_j - a}{g_u})^2\}(Y_j - a) \\ &= \sum_{j=1}^{n} \frac{-1}{\sqrt{2\pi}h g_u^2} \exp\{-0.5(\frac{x - x_j}{h})^2\} \\ &\cdot \int_{-\infty}^{\infty} \frac{Y_j - \theta}{\sqrt{2\pi(g_u^2 - g_l^2)}} \exp\{-0.5\frac{(a-\theta)^2}{g_u^2 - g_l^2}\} \frac{1}{\sqrt{2\pi}g_l} \exp\{-0.5(\frac{\theta - Y_j}{g_l})^2\}d\theta. \end{aligned} \tag{C.3}$$

Interchanging the integral with the summand yields

$$\begin{aligned}\frac{\partial}{\partial a}S(a,x_i|g_u) &= \int_{-\infty}^{\infty}\sum_{j=1}^{n}\frac{-1}{\sqrt{2\pi}hg_u^2}\exp\{-0.5(\frac{x-x_j}{h})^2\}\frac{Y_j-\theta}{\sqrt{2\pi}g_l}\exp\{-0.5(\frac{\theta-Y_j}{g_l})^2\} \\ &\cdot \quad \frac{1}{\sqrt{2\pi(g_u^2-g_l^2)}}\exp\{-0.5\frac{(a-\theta)^2}{g_u^2-g_l^2}\}d\theta \\ &= \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}S(\theta,x_i|g_l)\frac{1}{\sqrt{2\pi(g_u^2-g_l^2)}}\exp\{-0.5\frac{(a-\theta)^2}{g_u^2-g_l^2}\}d\theta. \qquad \text{(C.4)}\end{aligned}$$

Since $\frac{\partial}{\partial\theta}S(\theta,x_i|g_l)$ is continuous and bounded, and the normal density $\frac{1}{\sqrt{2\pi(g_u^2-g_l^2)}}\exp\{-0.5\frac{(a-\theta)^2}{g_u^2-g_l^2}\}$ is a Pólya frequency function we can invoke Theorem (C.1). Thus we have

$$\xi\left(\frac{\partial}{\partial a}S(a,x_i|g_u)\right) \leq \xi\left(\frac{\partial}{\partial a}S(a,x_i|g_l)\right),$$

and the proof of Theorem (4.1) is complete.

Vita

David Allan Burt was born on July 14, 1969 in Daytona Florida. In 1973 he moved with his mother Frances Lee Salyer to Kingsport Tennessee where he spent the remainder of his childhood. At the age of 8 David dedicated his life to Christ and was Baptized into the church of Morrison City Mission, Kingsport Tennessee.

In 1987 he graduated high school and went on to attend college at King College, Bristol Tennessee where he earned his Bachelors degree in Mathematics. He then went on to complete his Masters degree in Mathematics at East Tennessee State University, Johnson City Tennessee. From 1992 to 1995 he worked as an instructor and math lab coordinator at North East State Technical Community College, Blountville Tennessee. On November 6, 1999 he was marred to Rebecca Lee Saunders. David Allan Burt received his Doctor of Philosophy degree in Statistics on March 23, 2000 and is currently working as a clinical statistician at Abbott Labs, Lake Bluff IL.