University of Kentucky
UKnowledge

University of Kentucky Doctoral Dissertations
Graduate School

2010

NONPARAMETRIC ESTIMATION OF DERIVATIVES WITH APPLICATIONS

Benjamin Hall
University of Kentucky, Benjamin.Hall@uky.edu

Right click to open a feedback form in a new tab to let us know how this document benefits you.

Recommended Citation

Hall, Benjamin, "NONPARAMETRIC ESTIMATION OF DERIVATIVES WITH APPLICATIONS" (2010). *University of Kentucky Doctoral Dissertations*. 114.
https://uknowledge.uky.edu/gradschool_diss/114

This Dissertation is brought to you for free and open access by the Graduate School at UKnowledge. It has been accepted for inclusion in University of Kentucky Doctoral Dissertations by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

ABSTRACT OF DISSERTATION

Benjamin Hall

The Graduate School
University of Kentucky
2010

NONPARAMETRIC ESTIMATION OF DERIVATIVES WITH APPLICATIONS

ABSTRACT OF DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Arts and Sciences
at the University of Kentucky

By
Benjamin Hall
Lexington, Kentucky

Director: Dr. Richard Charnigo, Professor of Statistics
Lexington, Kentucky 2010

Copyright© Benjamin Hall 2010

ABSTRACT OF DISSERTATION

NONPARAMETRIC ESTIMATION OF DERIVATIVES WITH APPLICATIONS

We review several nonparametric regression techniques and discuss their various strengths and weaknesses with an emphasis on derivative estimation and confidence band creation. We develop a generalized C(p) criterion for tuning parameter selection when interest lies in estimating one or more derivatives and the estimator is both linear in the observed responses and self-consistent. We propose a method for constructing simultaneous confidence bands for the mean response and one or more derivatives, where simultaneous now refers both to values of the covariate and to all derivatives under consideration. In addition we generalize the simultaneous confidence bands to account for heteroscedastic noise. Finally, we consider the characterization of nanoparticles and propose a method for identifying a proper subset of the covariate space that is most useful for characterization purposes.

KEYWORDS: compound estimator, bias correction, tuning parameter, confidence bands, heteroscedastic noise

Author's signature: Benjamin Hall

Date: May 3, 2010

NONPARAMETRIC ESTIMATION OF DERIVATIVES WITH APPLICATIONS

By
Benjamin Hall

Director of Dissertation: Richard Charnigo

Director of Graduate Studies: William Griffith

Date: May 3, 2010

RULES FOR THE USE OF DISSERTATIONS

Unpublished dissertations submitted for the Doctor's degree and deposited in the University of Kentucky Library are as a rule open for inspection, but are to be used only with due regard to the rights of the authors. Bibliographical references may be noted, but quotations or summaries of parts may be published only with the permission of the author, and with the usual scholarly acknowledgments.

Extensive copying or publication of the dissertation in whole or in part also requires the consent of the Dean of the Graduate School of the University of Kentucky.

A library that borrows this dissertation for use by its patrons is expected to secure the signature of each user.

Name	Date

DISSERTATION

Benjamin Hall

The Graduate School
University of Kentucky
2010

NONPARAMETRIC ESTIMATION OF DERIVATIVES WITH APPLICATIONS

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Arts and Sciences
at the University of Kentucky

By
Benjamin Hall
Lexington, Kentucky

Director: Dr. Richard Charnigo, Professor of Statistics
Lexington, Kentucky 2010

Copyright© Benjamin Hall 2010

TABLE OF CONTENTS

Table of Contents iii

List of Figures v

List of Tables vi

Chapter 1 A Review of Nonparametric Regression 1
  1.1 Introduction 1
  1.2 Kernel regression 1
  1.3 Local regression 3
  1.4 Orthogonal basis functions 5
  1.5 Smoothing splines 6
  1.6 Compound estimation 7
  1.7 Confidence bands in nonparametric regression 10
  1.8 Applications of nonparametric derivative estimation 12

Chapter 2 Generalized C(p) 14
  2.1 Previous work by Charnigo and Srinivasan 14
  2.2 Properties of empirical derivatives 16
  2.3 Simulation studies of generalized C(p) 38

Chapter 3 Simultaneous Confidence Bands for a Function and Its Derivatives 47
  3.1 Previous work by Knafl, Sacks, and Ylvisaker 47
  3.2 Confidence bands over a function and its derivatives 49
  3.3 Accounting for the bias 56
  3.4 Interpolating between the grid points 62
  3.5 Modifications for unknown variance 67
  3.6 Simulation studies 68
  3.7 Ethanol example 72

Chapter 4 Confidence Bands in the Presence of Heteroscedastic Noise 77
  4.1 Heteroscedastic noise 77
  4.2 Simulation studies 86
  4.3 Fetal growth example 90

Chapter 5 Limited Angle Recovery of Nanoparticle Characteristics 93
  5.1 Introduction 93
  5.2 Limited angles for the reference curves 94
  5.3 Limited angles for particles of unknown size 107
  5.4 Simulation studies 115

Chapter 6 Conclusions and future research 120

Bibliography 125

Vita 129

LIST OF FIGURES

1.1 Characterization of nanoparticles by scattering profiles 13

2.1 Comparison of weights for empirical derivatives 21
2.2 The largest weights 22
2.3 Empirical first derivatives 24
2.4 Empirical second derivatives 27
2.5 Empirical third derivatives 30
2.6 Contour Plot of the Target (1st Derivative) 44
2.7 Contour Plot of the Generalized C(p) (1st Derivative) 44
2.8 Contour Plot of the Target (2nd Derivative) 45
2.9 Contour Plot of the Generalized C(p) (2nd Derivative) 45
2.10 Contour Plot of the Target (3rd Derivative) 46
2.11 Contour Plot of the Generalized C(p) (3rd Derivative) 46

3.1 Simultaneous confidence bands for the mean response and first derivative 69
3.2 Simultaneous confidence bands for the mean response and two derivatives 70
3.3 Mean response and 1st derivative of ethanol data 74
3.4 Mean response and first two derivatives of ethanol data 75
3.5 Mean response and 1st derivative of ethanol data (80% confidence) . . . 76

4.1 Variance functions used in simulations 87
4.2 Comparison of confidence bands across different variance functions . . . 89
4.3 Mean response and first derivative of fetal growth data 91
4.4 Mean response and first derivative of fetal growth up to 28 weeks 92

5.1 Scattering profiles for M11 96
5.2 Evaluating characterization ability of M11 97
5.3 Example of potential limited angle choice pitfall 98
5.4 Hypothetical scattering profiles and derivatives 99
5.5 Threshold for selecting limited angles 102
5.6 M11 first derivatives of scattering profiles 103
5.7 M33 scattering profiles 104
5.8 M33 first derivatives of scattering profiles 105

LIST OF TABLES

2.1 Weights for empirical derivatives 20
2.2 Results from simulation studies assessing generalized C(p) 41

3.1 Cutoff approximations for the mean response and first derivative 51
3.2 Cutoff approximations for the mean response and first two derivatives . 54
3.3 Simulation results for simultaneous confidence bands 71

4.1 Simulation results for simultaneous confidence bands 88
4.2 More simulation results for simultaneous confidence bands 89

5.1 Determining the threshold for M11 106
5.2 Evaluation of inverse problem limited angles for M_{11} 110
5.3 Further evaluation of inverse problem limited angles for M_{11} 111
5.4 Using the mean response and two derivatives together for M_{11} 111
5.5 Evaluation of inverse problem limited angles for M_{33} 113
5.6 Using the mean response and two derivatives together for M_{33} 113
5.7 Further evaluation of inverse problem limited angles for M_{33} 114
5.8 Simulation 1: results for evaluating $\mathcal{T}_M$ 118
5.9 Simulation 2: results for evaluating $\mathcal{T}_M^*$ 119

Chapter 1 A Review of Nonparametric Regression

1.1 Introduction

Consider a situation in which we observe noisy data

$$Y_i = \mu(x_i) + \epsilon_i \quad \text{for } i \in \{1, \dots, n\}, \tag{1.1}$$

where $x_i \in \mathcal{X}$, $\mathcal{X}$ is a compact interval, the Y_i are observed responses, $\mu(x)$ is the true mean response function, and the ϵ_i are zero-mean random errors. If the functional form of $\mu(x)$ were specified up to a set of unknown parameters, such knowledge could be exploited to produce an estimate of the mean response function by estimating the desired model parameters. Such is the case in linear regression, logistic regression, and Poisson regression, among other models. However, in many situations we are unwilling to specify the functional form of the mean response, $\mu(x)$, but still seek to estimate it. In this case we enter the realm of nonparametric regression.

Many methods have been devised to estimate the mean response function in this situation, among them kernel regression, smoothing splines, orthogonal series estimators, local regression, and compound estimation. Techniques have been established for constructing confidence bands and extensions have been made to estimate derivatives of the mean function. In this section we discuss the success of some of these methods as well as their shortcomings.

1.2 Kernel regression

Kernel regression (or kernel smoothing) is a conceptually simple method that grew naturally out of kernel density estimation. Nadaraya (1964) and Watson (1964) applied the Rosenblatt-Parzen kernel density estimator (Rosenblatt 1956, Parzen 1962) to regression and proposed the following estimator:

$$\widehat{\mu}_{nw}(x) := \frac{n^{-1}\sum_{i=1}^{n} K_h(x - x_i)Y_i}{n^{-1}\sum_{i=1}^{n} K_h(x - x_i)} \tag{1.2}$$

with

$$K_h(u) = h^{-1}K(u/h)$$

where K is a symmetric real function that integrates to one and h is a nonnegative scale factor. Similar alternative kernel estimators are proposed by Priestley and Chao (1972):

$$\widehat{\mu}_{pc}(x) := \sum_{i=1}^{n}(x_i - x_{i-1})K_h(x - x_i)Y_i,$$

and Gasser and Muller (1979):

$$\widehat{\mu}_{gm}(x) := \sum_{i=1}^{n}\left[\int_{s_{i-1}}^{s_i} K_h(x-u)du\right] Y_i \text{ with } s_i := (x_{i-1} + x_i)/2.$$

In any case, we are forced to choose inputs for K and, more importantly, h.

With regard to K, the Epanechnikov kernel,

$$K(u) = 0.75(1 - u^2)1_{|u|\leq 1}, \tag{1.3}$$

is optimal under reasonable assumptions in the sense that it minimizes asymptotic mean integrated square error (Epanechnikov 1969). However, other commonly used kernels have been shown to be nearly as efficient (Hardle 1990, Loader 1999). Thus, preference is often given to a smoother or simpler kernel since little is sacrificed in terms of efficiency.

The choice of h (called bandwidth) is given much more attention due to its critical role in driving a bias-variance trade-off. That is, large values of h produce smooth curves and eliminate much of the probable error associated with the data, but with a price of inducing systematic error. Small values of h lead to less smoothing and less systematic error but result in a noisy fit with large variance. The choice of such a parameter (called a tuning parameter), is a common one in nonparametric regression. Hardle (1990) notes that automated methods for determining bandwidth selection in kernel regression include cross validation, penalizing functions, and the plug-in method.

A feature of kernel regression that is common to all of the techniques discussed in this chapter is that the estimator is linear in the observed responses. That is, the estimator can be written in the form:

$$\widehat{\mu}(x) := \sum_{i=1}^{n} l_i(x) Y_i, \tag{1.4}$$

for functions $l_1, ..., l_n$ which do not depend on $Y_1, ..., Y_n$. This property plays an important role in both the development of generalized C(p) and the construction of confidence bands in the following chapters.

For kernel regression, estimation of derivatives of the mean response can be accomplished by differentiating the $l_i(x)$ from (1.4) with respect to x. For example, the Priestley-Chao estimator of the q^{th} derivative of the mean response is

$$\widehat{\frac{d^q}{dx^q}\mu_{pc}}(x_i) := h^{-(q+1)} \sum_{i=1}^{n} (x_i - x_{i-1}) K^{(q)}(x - x_i) Y_i,$$

Importantly, this means that estimates of the derivatives are the derivatives of the estimates, a property referred to as “self-consistency”.

The primary advantage of kernel regression lies in its simplicity. The choice of the bandwidth parameter h is generally the only option the user must consider. A great deal of asymptotic theory has also developed around kernel regression. This theory is useful and was, for instance, exploited by Eubank and Speckman (1993) to create confidence bands around the estimate. The self-consistency of kernel smoothing derivative estimation is another desirable quality, although derivative estimation is only possible when the kernel is sufficiently smooth. A downside to kernel regression is that the estimates it produces may have bias that depends on the first derivative of the mean function. Bias that depends on low-ordered derivatives can be eliminated by techniques such as local regression.

1.3 Local regression

Local regression is a more sophisticated technique that can be viewed as an extension of kernel regression. The local regression approach is to estimate the mean function at a given value of the covariate with a polynomial within the so-called smoothing window. To be more specific, if (for a given j) we minimize

$$\sum_{i=1}^{n} K_h(x_i - x)\left(Y_i - \left[a_0 + a_1(x_i - x) + ... + \frac{1}{j!}a_j(x_i - x)^j\right]\right)^2 \tag{1.5}$$

with respect to the a_i, then the local regression estimate of $\mu(x)$ is $\hat{a}_0$. The smoothing window is determined by K and h.

As in kernel regression, there is a bias-variance tradeoff that is primarily driven by h. The polynomial degree is commonly chosen to be either one (local linear) or two (local quadratic). Local constant regression is equivalent to the Nadaraya-Watson estimator of (1.2). Loader (1999) mentions cross validation, generalized cross validation, and the CP criterion as methods for tuning parameter selection in local regression. The CP criterion (Mallows 1973), which we extend in the next chapter, is defined in such a way that it equals the sum of the squared error over the design points in expected value. Specifically,

$$CP(\widehat{\mu}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \widehat{\mu}(x_i))^2 - n + 2\sum_{i=1}^{n}||l(x_i)||^2. \tag{1.6}$$

The local regression estimate of $\mu^{(q)}(x)$ is $\hat{a}_q$ from (1.5) for $q \leq j$. Importantly, this implies that local regression is not self-consistent, meaning derivatives of an estimate are not estimates of the derivatives. For instance, Loader (1999) gives the first derivative of the estimate to be

$$\frac{d\widehat{\mu}(x)}{dx} = \hat{a}_1 + e_1^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}'\hat{\epsilon},$$

where $e_1^T = (1, 0, ..., 0)^T$, $\mathbf{X}$ is the design matrix, $\mathbf{W}$ is a diagonal matrix with entries $K_h(x_i - x)$, and $\hat{\epsilon}$ is the vector of residuals.

A strength of local regression is that bias cannot depend on derivatives of the mean function up to order j from (1.5). For instance, the bias of a local quadratic fit can only depend on derivatives of the mean response of order three and higher. However, local regression does not have the self-consistency property and estimates of derivatives of order higher than j are not even possible. If $\mu(x)$ is estimated using $j = 1$ and an estimate of the second derivative is then needed, one must estimate $\mu(x)$ again, this time using $j \geq 2$.

1.4 Orthogonal basis functions

Estimation using orthogonal basis functions involves the assumption that for some known basis functions $\{\phi_k\}_{k=0}^{\infty}$, $\mu(x) = \sum_{k=0}^{\infty} \beta_k \phi_k(x)$. The problem of estimation can then be translated to simply estimation of the β_k. Of course only finitely many of the β_k can be estimated, so for this method to work, it must be the case that $\sum_{k=0}^{J} \beta_k \phi_k(x)$ provides a good approximation of $\mu(x)$ for modest J. Then if we partition the covariate space $\mathcal{X}$ into $\{A_i\}_{i=1}^{n}$ such that each $x_i \in A_i$, the β_k can be estimated by

$$\hat{\beta}_k = \sum_{i=1}^{n} Y_i \int_{A_i} \phi_k(x) dx,$$

and the estimation of the mean response function is

$$\widehat{\mu}_{os}(x) = \sum_{k=0}^{J} \hat{\beta}_k \phi_k(x). \tag{1.7}$$

The estimator, known as an orthogonal series estimator, is linear in the observed responses with the $l_i(x)$ from (1.4) defined by Hardle (1990) as

$$l_i(x) = \sum_{k=0}^{J} \phi_k(x) \int_{A_i} \phi_k(u) du.$$

Estimation of derivatives using orthogonal basis functions is accomplished by differentiating the estimate of the mean function, i.e. self-consistently. However, this requires that the basis functions be sufficiently differentiable, which is not always the case. Wavelets, which fall into the category of orthogonal series estimators, excel when estimating functions that are discontinuous (see Charnigo et al 2006 and references therein). However, they are ill-suited to the estimation of derivatives. In addition, Charnigo and Srinivasan (2010a) note that misspecification of the basis functions can make estimation of derivatives difficult. They further note that obtaining an estimator from (1.7) that estimates $\mu(x)$ well is not sufficient to ensure that its derivative will also estimate $\mu'(x)$ well. The estimate of the derivative may need a larger J than the estimate of the mean response function. Thus, how to choose the basis functions and how large to choose J are questions without clear answers when we are seeking to recover one or more derivatives with an orthogonal series estimator.

1.5 Smoothing splines

Smoothing splines are yet another approach to the nonparametric regression problem (Schoenberg 1964, Reinsch 1967). Splines, in general, are piecewise polynomials. The polynomials of order J are pieced together at “knots” in such a way that the spline and its first $J-1$ derivatives of the spline are continuous at the knots, where J is the order of the spline.

A useful approach to employing splines for smoothing purposes begins by considering the residual sum of squares, a commonly used “goodness of fit” measure:

$$\sum_{i=1}^{n}(Y_i - \widehat{\mu}(x_i))^2. \tag{1.8}$$

The smaller this quantity is, the better the fit. However, simply minimizing (1.8) will not work as an estimator of μ since such a problem has infinitely many solutions. In fact any function which interpolates the data will reduce (1.8) to zero. And aside from issues of nonuniqueness, such an estimator would be unacceptably volatile.

The problems of nonuniqueness and rapid variation can be solved all at once by additionally imposing a “roughness penalty” before minimization. Spline smoothing defines roughness in terms of derivatives. For instance, if roughness is defined in terms of the second derivative, the following quantity must be minimized:

$$\sum_{i=1}^{n}(Y_i - \widehat{\mu}(x_i))^2 + \lambda \int (\widehat{\mu}''(x))^2 dx, \tag{1.9}$$

where $\lambda > 0$ is a tuning parameter. Methods for choosing λ exist and include generalized cross-validation (Wahba 1977).

The minimizer of (1.9), which we denote $\widehat{\mu}_s(x)$, is unique over the class of all twice differentiable functions and is referred to as the cubic smoothing spline. The roughness penalty could instead be defined in terms of higher-order derivatives which lead to higher-order splines.

In minimizing (1.9) the knots correspond to the design points. Since smoothing splines estimate derivatives self-consistently, the order of the spline must be larger than the order of the derivative being estimated. A rule of thumb is that if q derivatives are to be estimated, J should be at least $2q+3$. This has a couple of implications. One is that a smoothing spline estimator is not analytic. Second, since J is chosen based on the number of derivatives, say q, that one desires to estimate, the smoothing spline estimate of $\mu(x)$ will depend on q. That is, smoothing splines lack invariance with respect to the number of derivatives one is interested in estimating. This is bothersome not only conceptually but also computationally. For example, suppose one estimates the first 2 derivatives and, following the rule of thumb, sets $J = 7$. If, subsequently, an estimate of the fifth derivative is also requested, $J = 7$ will be deemed inadequate. We will need to reset J to 13 and entirely new estimates of the first two derivatives will have to be produced.

A smoothing spline estimator is linear in the observed responses. However, writing such an estimator in the form of (1.4) is very difficult, though Silverman (1984) was able to derive a tractable asymptotic approximation for the l_i.

1.6 Compound estimation

Compound estimation is a recent development by Charnigo and Srinivasan (2010a). The method is designed especially for the case where one is interested in estimating the mean response and one or more derivatives simultaneously. The technique involves first defining "pointwise estimators" of $\mu^{(j)}(a)/j!$ for $0 \leq j \leq J$ and $a \in I_n$ where $I_n \subset \mathcal{X}$ is a set of "centering points". With these pointwise estimators denoted by $\widetilde{c}_{j;a}$, a polynomial

$$\widetilde{\mu}_{J;a}(x) := \sum_{j=0}^{J} \widetilde{c}_{j;a}(x-a)^j$$

is then defined for each centering point. Then the compound estimator is defined to be

$$\mu^*(x) := \sum_{a \in I_n} W_{a,n}(x)\widetilde{\mu}_{J;a}(x), \tag{1.10}$$

where

$$W_{a,n}(x) := \frac{\exp[-\beta(x-a)^2]}{\sum_{c \in I_n} \exp[-\beta(x-c)^2]}$$

and $\beta > 0$ is a tuning parameter.

The pointwise estimators may be defined in a number of ways. The restrictions are only that they satisfy

$$\sup_{a\in I_n} MSE\left[\widetilde{c}_{j;a}\right] \leq C\ n^{-2\alpha_j} \quad \text{and} \quad \sup_{a\in I_n} |\widetilde{c}_{j;a}| \leq C \tag{1.11}$$

for some positive constants $C, \alpha_0, \ldots, \alpha_J$. Options for obtaining them include many of the methods described in previous sections. Stone (1980) shows that if $\mu(x)$ has $J+1$ continuous derivatives, the x_i are equispaced, and the ϵ_i are independent and identically normally distributed, then (1.11) is satisfied by local regression estimators with rectangular weights, that is with $K(u) = I_{[-1,1]}(u)$, and $h := \xi n^{-1/(2J+3)}$ for a positive constant ξ.

Charnigo and Srinivasan (2010b) offer pointwise estimators that are inductive. That is, given $\widetilde{c}_{0;a}$ and $\widetilde{c}_{1;a}$ for every $a \in I_n$, the formula for all subsequent $\widetilde{c}_{j;a}$ depends on $\widetilde{c}_{0;a}, ..., \widetilde{c}_{j-1;a}$, but not $\widetilde{c}_{j+1;a}, \widetilde{c}_{j+2;a}$, etc. Such estimators are also invariant: If one computes the compound estimator and then decides that J should be increased, the previously calculated pointwise estimators do not need to be re-calculated. For example if $\widetilde{c}_{0;a}, ..., \widetilde{c}_{5;a}$ are calculated with $J = 5$ and it is then determined that J should instead be 7, $\widetilde{c}_{0;a}, ..., \widetilde{c}_{5;a}$ remain the same. One simply needs to inductively calculate the additional pointwise estimators $\widetilde{c}_{6;a}, \widetilde{c}_{7;1}$.

Compound estimation recovers derivatives self-consistently. If the pointwise estimators used to define the compound estimator satisfy (1.11), then the compound estimator recovers the mean response and its first $\lfloor J/2 \rfloor$ derivatives consistently (in a probabilistic sense). In fact, Charnigo and Srinivasan (2010a) show that if the pointwise estimators satisfy (1.11) with $\alpha_j := (J+1-j)/(2J+3)$ for $0 \leq j \leq J$ (such is the case for the estimators from local regression with rectangular weights described above) and ν is an infinitesimally small positive number, then

$$\sup_{x\in I} \left| \frac{d^j}{dx^j}\mu^*(x) - \mu^{(j)}(x) \right| = o_p\left(n^{(2j-J-1/2)/(2J+3)+\nu}\right) \qquad \text{for} \ \ 0 \leq j \leq \lfloor J/2 \rfloor,$$

where I is a compact interval contained in the interior of $\mathcal{X}$.

The compound estimator as defined by (1.10) has the unfortunate feature shared by all of nonparametric methods we have mentioned that the choice of J will depend on how many derivatives one is interested in estimating. Thus, the estimate of $\mu^{(j)}(x)$ depends on if one is also interested in estimating $\mu^{(j+k)}(x)$, for positive integers j and k. Thus the compound estimator of (1.10) lacks invariance with respect to the number of derivatives being estimated.

Charnigo and Srinivasan (2010b) thus propose the following "extended" compound estimator:

$$\mu_{\infty}^{*}(x) := \frac{\sum_{a \in I_n} \exp\left[-\beta_n (x-a)^2\right] \sum_{j=0}^{\infty} \widetilde{c}_{j;a} (x-a)^j}{\sum_{a \in I_n} \exp\left[-\beta_n (x-a)^2\right]}. \tag{1.12}$$

Assume we use the inductive pointwise estimators with $\widetilde{c}_{0;a}$ and $\widetilde{c}_{1;a}$ chosen so that they consistently estimate $\mu(a)$ and $\mu'(a)$, respectively. Then the extended compound estimator (1.12) is invariant with respect to how many derivatives one is interested in estimating.

At first one may object to the ∞ used in (1.12) as impractical. However, a clever modification is to simply replace $\widetilde{c}_{j;a}$ with

$$\widetilde{c}_{j;a} 1_{n \geq N_j},$$

where $\{N_k\}_{k=0}^{\infty}$ is a strictly increasing sequence of positive integers. Charnigo and Srinivasan (2010b) do this and show that if there exist a positive integer $n_0 > 1$, a nonincreasing sequence of positive numbers $\{\alpha_j\}_{j=0}^{\infty}$, and a sequence of positive numbers $\{K_j\}_{j=0}^{\infty}$ such that, for each $j \geq 0$ and $n \geq n_0$,

$$\sup_{x \in I} \left|\mu^{(j)}(x)\right| / j! \leq K_0 2^{-j},$$

$$\sup_{a \in I_n} MSE[\widetilde{c}_{j;a}] \leq K_j n^{-2\alpha_j}, \text{ and}$$

$$\sup_{a \in I_n} |\widetilde{c}_{j;a}| \leq K_0 2^{-j},$$

the resulting extended compound estimator consistently estimates every derivative of $\mu(x)$.

However, compound estimation does pay a price for its self-consistency and extended compound estimation pays a further price for its invariance with respect to the number of derivatives being estimated. That price is slower convergence. The optimal (pointwise) rate of convergence for a mean response that has $J+1$ continuous derivatives is $O_p(n^{(j-J-1)/(2J+3)})$ for $j \leq J$ (Stone 1980). As noted above, the self-consistent compound estimator has a convergence rate of $o_p(n^{(2j-J-1/2)/(2J+3)+\nu})$ for $j \leq \lfloor J/2 \rfloor$. The self-consistent and invariant extended compound estimator has an even slower convergence rate of $O_p\left(n^{-(2j+1)(\log n)^{\xi-1}}\right)$, where $\xi \in (0,1)$ is arbitrary but fixed.

1.7 Confidence bands in nonparametric regression

Just as with parametric techniques, nonparametric regression is not strictly concerned with simply estimating the mean response function. There may be interest in constructing a region which contains the mean response function with a desired level of confidence. Such a region is usually defined by upper and lower bounds referred to as confidence bands. To be precise, $L(x)$ and $U(x)$ form $100(1-\alpha)\%$ confidence bands for the mean response function $\mu(x)$ if

$$P(L(x) \leq \mu(x) \leq U(x), \forall x) = 1 - \alpha. \tag{1.13}$$

To emphasize that the bands are valid for all values of the covariate at the same time, bands satisfying (1.13) are sometimes referred to as simultaneous confidence bands. If equality in (1.13) is replaced by $\geq$, the bands are said to be conservative.

Efforts to place confidence bands around the estimated mean response rely on, among other considerations, a method to account for the bias, i.e. to adjust for the discrepancy between the expected value of the nonparametric technique and the true mean response. One such method, in the general case of nonparametric estimators that are linear in the observed response, utilizes bounds on derivatives which in turn are used to bound the bias (Hall and Titterington 1988). In the special case of kernel regression, Eubank and Speckman (1993) use asymptotic results to motivate an estimate of the bias. In particular, since under mild conditions

$$\mathbb{E}\left[\hat{\mu}_h(x)\right] \sim \mu(x) + h^2 B \mu''(x)$$

where $\hat{\mu}_h(x)$ is a kernel density estimator with bandwidth h and

$$B = \frac{1}{2}\int u^2 K(u)du,$$

it is proposed that the bias be estimated by

$$\hat{b}_{h,\lambda}(x) = h^2 B \hat{\mu}''_\lambda(x),$$

where

$$\hat{\mu}''_\lambda(x) = \frac{1}{n\lambda}\sum_{i=1}^{n} Y_i K^*\left(\frac{x - x_i}{\lambda}\right),$$

and K^* is a square-integrable kernel that satisfies $\int u^j K^*(u)du = 0$ for $j = 0, 1$ and $\int u^2 K^*(u)du = length(\mathcal{X})$. The bands they propose are then defined as:

$$\hat{\mu}_h(x) - \hat{b}_h(x) \pm \frac{V}{\sqrt{nh}}\left[\sqrt{-2\ \log h} + \frac{1}{\sqrt{-2\ \log h}}(C + x_\alpha)\right]\sigma$$

where

$$V^2 = \int K(u)^2 du,$$

$$C = \log\left(\frac{1}{2\pi}\left[\int K'(u)^2 du / \int K(u)^2 du\right]^{1/2}\right),$$

and

$$x_\alpha = -\log\left(\frac{-\log(1-\alpha)}{2}\right).$$

If the variance is unknown, they estimate it by any $\sqrt{n}$-consistent estimator.

They prove that bands produced by this method achieve the nominal coverage probabilities asymptotically. They also demonstrate empirically that such bands provide coverage probabilities nearer to the nominal level than bands from a Bonferroni-type approach that does not account for bias correction. The Bonferroni approach is asymptotically conservative.

To our knowledge, Claeskens and Van Keilegom (2003) are the only authors to attempt confidence bands for the first derivative of $\mu(x)$. They do so using local polynomial estimation in the maximum likelihood setting. Importantly, their estimators are not self-consistent. In Chapter 3, we extend the work of Knafl et. al (1985) to provide confidence bands for derivative estimates in the case where the estimator is linear in the observed responses and self-consistent.

An issue when using confidence bands in practice is that the variance must be estimated. When the ϵ_i's are assumed to be independent normal with common variance σ^2, many such estimates are available. In the following chapters, we use the variance estimator from Loader (1999):

$$\widehat{\sigma}^2 := \frac{\sum_{i=1}^{n} \left[Y_i - \widehat{\mu}(x_i)\right]^2}{n - \sum_{m=1}^{n} l_m(x_m)}, \tag{1.14}$$

where denominator of (1.14) can be interpreted as the degrees of freedom.

1.8 Applications of nonparametric derivative estimation

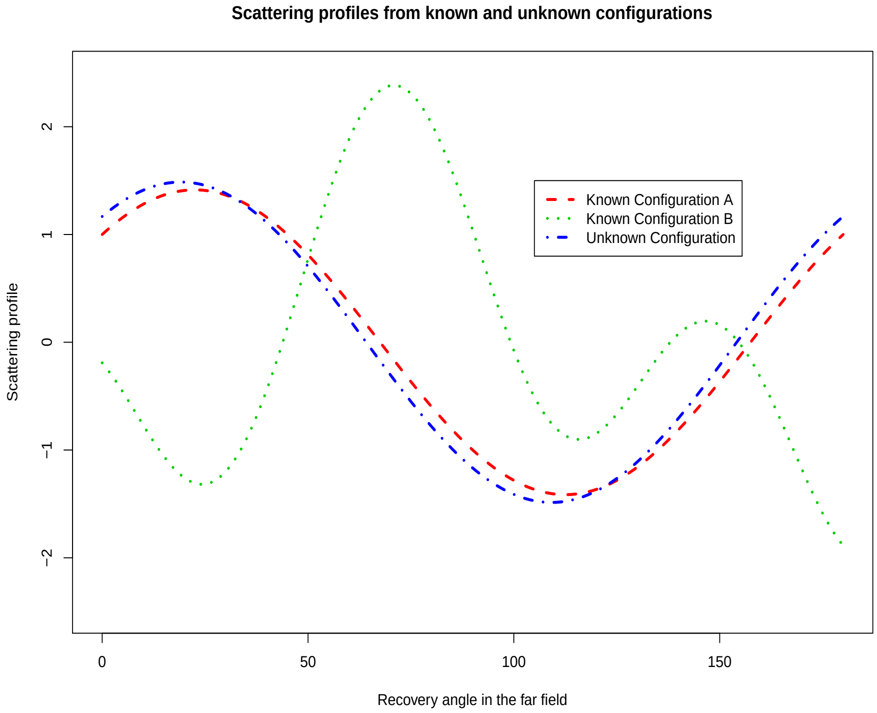
Of course, estimating derivatives nonparametrically is more than a mere intellectual exercise. The applications are important and wide-ranging; we describe two of them below.

Ramsay and Silverman (2002) analyze human growth data by estimating height, velocity, and acceleration curves. This application demonstrates the importance of self-consistency (Charnigo and Srinivasan 2009a): If one is trying to locate the peak of a growth spurt, an estimator that is not self-consistent will give different answers depending on whether one uses the estimated height, velocity, or acceleration curve to locate the peak.

Another application is the characterization of nanoparticles by scattering profiles (Francoeur et al 2007), where the scattering profile is a mean response function with far field recovery angle as the covariate. Nanoparticles with different configurations will produce different scattering profiles. (Configuration refers to a specific value of a characteristic such as size or agglomeration level.) Thus, we can obtain an estimate of the scattering profile for a given unknown configuration of nanoparticles and compare it to scattering profiles for known configurations. We would then classify the unknown configuration based on the proximity of its estimated scattering profile to the scattering profiles for known configurations. This idea is illustrated in Figure 1.1, taken from Charnigo et al (2010). In actuality, we do not observe single elements of scattering profiles at each θ, but rather a 4×4 matrix of scattering elements, each of which can be viewed as a scattering profile and used for characterization purposes.

Previous research suggests that the M_{11}, M_{12}, M_{33}, and M_{34} profiles are most useful (Francouer et al 2007, Manickavasagam and Menguc 1997).

Figure 1.1: Characterization of nanoparticles by scattering profiles



The scattering profile for the nanoparticles with unknown configuration (blue dot-dash) more closely resembles the scattering profile for the nanoparticles with known configuration "A" (red dash) than that for the nanoparticles with known configuration "B" (green dot). Hence, one concludes that the unknown configuration is much closer to the known configuration A than to the known configuration B.

Derivatives of scattering profiles can be even more effective for characterization purposes (Charnigo et al 2007), and consideration of the scattering profiles and one or more derivatives simultaneously can be more effective still (Charnigo et al 2010).

However, not all ranges of the covariate are equally valuable for characterization. In Chapter 5 we explore methods for identifying a proper subset of $\mathcal{X}$ most useful for characterization purposes.

Copyright© Benjamin Hall, 2010.

Chapter 2 Generalized C(p)

2.1 Previous work by Charnigo and Srinivasan

Suppose that we observe noisy data

$$Y_i = \mu(x_i) + \epsilon_i \quad \text{for } i \in \{1, \ldots, n\}, \tag{2.1}$$

where the design points x_i are equispaced on a compact interval $\mathcal{X} \subset \mathbb{R}$, $\mu(x)$ is a real-valued function defined on $\mathcal{X}$ that has $(q+1)$ continuous derivatives for some positive integer q, and the ϵ_i are independent zero-mean random errors with common variance $\sigma^2 \in (0, \infty)$.

To provide a method for tuning parameter selection when interest lies in estimating $\frac{d^q}{dx^q}\mu(x)$ with an estimator that is both linear in the observed responses and with which derivatives are estimated self-consistently, Charnigo and Srinivasan (2008) have proposed generalized C(p) as a data-based surrogate for $\sum_{i=1}^n \left[\widehat{\frac{d^q}{dx^q}\mu(x_i)} - \frac{d^q}{dx^q}\mu(x_i)\right]^2$. Such a quantity is important because simply controlling data-based estimates of $\sum_{i=1}^n [\widehat{\mu}(x_i) - \mu(x_i)]^2$ will not guarantee that $\frac{d^q}{dx^q}\mu(x)$ is well estimated.

Generalized C(p) is a penalized residual sum of squares type quantity. The key ingredient in the ordinary C(p) (1.6) criterion is the Y_i, which are noise-corrupted versions of the true mean response $\mu(x)$. Therefore, an attempt to extend the C(p) criterion to estimation of derivatives will require inputs that resemble noise-corrupted versions of the appropriate derivative. Hence the development of generalized C(p) begins with the definition of empirical derivatives which will serve as such inputs. For a positive integer k, empirical first derivatives are defined as

$$Y_{i;k}^{(1)} := \sum_{j=1}^{k} w_j \left(\frac{Y_{i+j} - Y_{i-j}}{x_{i+j} - x_{i-j}}\right) \text{ with } w_j = j^2 / \sum_{l=1}^{k} l^2 \text{ for } j \in \{1, \ldots, k\}, \tag{2.2}$$

and can be viewed as noise corrupted versions of $\mu'(x_i)$. The motivation underlying this definition is that the ordinary first order difference quotient $(Y_i - Y_{i-1})/(x_i - x_{i-1})$ will have a prohibitively large variance. Empirical derivatives are appealing because with appropriate weights (i.e. w_j) they are variance-reducing. To this end, the w_j in the above definition are chosen to minimize the variance.

Higher order empirical derivatives are similarly defined as

$$Y_{i;k}^{(q)} := \sum_{j=1}^{k} w_j \left(\frac{Z_{i+j;j}^{(q-1)} - Z_{i-j;j}^{(q-1)}}{x_{i+j} - x_{i-j}} \right), \tag{2.3}$$

where $Z_{i;j}^{(0)} := Y_i$ and $Z_{i;j}^{(p)} := \left(Z_{i+j;j}^{(p-1)} - Z_{i-j;j}^{(p-1)} \right) / (x_{i+j} - x_{i-j})$ for any positive integer $p < q$ and (2.3) can be viewed as noise corrupted versions of $\mu^{(q)}(x_i)$.

Defining the weights for higher-order empirical derivatives is more complicated. Charnigo and Srinivasan (2008) again propose using the variance-minimizing weights. Unfortunately, these weights for empirical second derivatives are shown to be the solution to

$$\mathbf{A}_k \begin{bmatrix} w_2 \\ w_3 \\ \vdots \\ w_k \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

where $\mathbf{A}_k$ is the $(k-1) \times (k-1)$ matrix whose $(r-1, j-1)$ element is

$$\left[2\left(1 - \frac{1}{r^2}\right)\left(1 - \frac{1}{j^2}\right) + 1 + 1_{r=j}\frac{1}{r^4} \right] / \left[1 + 2\left(1 - \frac{1}{r^2}\right) \right]$$

for $r, j \in \{2, \ldots, k\}$ and $w_1 = 1 - \sum_{j=2}^{k} w_j$. The weights are unavailable in a simple closed-form. The complications in the calculations of the variance-minimizing weights for higher order derivatives arise from the fact that the summands of (2.3) may be correlated, whereas the summands of (2.2) are not. In contrast to the variance-minimizing weights of empirical first derivatives, the variance-minimizing weights of empirical second derivatives can be both positive and negative. We further discuss the choice of the w_j for second and higher order empirical derivatives in the subsequent section.

The ability of empirical derivatives to mimic noise-corrupted versions of the true derivatives depends on the choice of k. A small k may not reduce the variance sufficiently, while a large k may introduce excessive bias. In addition, definition (2.3) does not work when $i < qk + 1$ or $i > n - qk$; some kind of modification must be made. A large k will exacerbate these boundary issues. Also, as the order of the derivative increases, the size of k necessary to control the variance increases as well. The choice of k is discussed further in the subsequent section.

Empirical derivatives are employed as inputs for the generalized C(p) criterion. For each $i, m \in \{1, \dots, n\}$, let $c_{i,m}$ be defined so that $\sum_{m=1}^{n} c_{i,m} Y_m = Y_{i;k}^{(q)}$ and let $l_m(x_i)$ be defined as in (1.4). Then the generalized C(p) criterion is defined as

$$\begin{aligned} & GCP(\mathbf{Y}, \widehat{\mu}) \qquad (2.4) \\ := \; & \sum_{i=1}^{n} s_i \left(\sum_{m=1}^{n} c_{i,m} Y_m - \widehat{\frac{d^q}{dx^q} \mu(x_i)} \right)^2 + \sigma^2 \sum_{i=1}^{n} s_i \sum_{m=1}^{n} \left(2 c_{i,m} \frac{d^q}{dx^q} l_m(x_i) - c_{i,m}^2 \right), \end{aligned}$$

where $\mathbf{Y}$ is shorthand for $Y_1, \dots, Y_n$, and for each i, $s_i \geq 0$, perhaps defined as $s_i := 1_{qk+1 \leq i \leq n-qk}$ to alleviate boundary issues. Since empirical derivatives are proxies for noise-corrupted observations of the true derivatives, generalized C(p) can be viewed as a penalized residual sum of squares for the fitted derivative of order q. The penalty is included in order to match generalized C(p) with its target, $\sum_{i=1}^{n} s_i \left(\frac{d^q}{dx^q} \mu(x_i) - \widehat{\frac{d^q}{dx^q} \mu(x_i)} \right)^2$, in expected value up to a remainder term. Thus if $\{\widehat{\mu}_\lambda : \lambda \in \Lambda\}$ is a family of estimators indexed by a scalar or vector tuning parameter λ belonging to a finite set Λ, $\widehat{\lambda} := \arg\min_{\lambda \in \Lambda} GCP(\mathbf{Y}, \widehat{\mu}_\lambda)$. is expected to be a good choice for λ (i.e. $\sum_{i=1}^{n} s_i \left(\frac{d^q}{dx^q} \mu(x_i) - \widehat{\frac{d^q}{dx^q} \mu_{\widehat{\lambda}}(x_i)} \right)^2$ will be close to $\min_{\lambda \in \Lambda} \sum_{i=1}^{n} s_i \left(\frac{d^q}{dx^q} \mu(x_i) - \widehat{\frac{d^q}{dx^q} \mu_\lambda(x_i)} \right)^2$).

The success of generalized C(p) is investigated using simulation studies below.

2.2 Properties of empirical derivatives

As we have noted, the choice of k when defining empirical derivatives drives a bias-variance tradeoff. In this section we investigate this relationship and demonstrate both empirically and asymptotically that a balance between the bias and variance is achievable, with the resultant empirical derivatives appearing as noise corrupted versions of the true derivative. We begin with a discussion of the variance-minimizing weights for higher order empirical derivatives.

Similar to those for empirical second derivatives, the variance minimizing weights for empirical third derivatives and empirical fourth derivatives are the solutions to sets of linear equations and are unavailable in simple closed-form. These weights are defined in the following propositions.

Proposition 2.2.1 *Assume that model (2.1) holds. Then the variance of* $Y_{i;k}^{(3)}$ *from expression (2.3), where* $3k+1 \leq i \leq n-3k$, *is minimized when* $w_1 = 1 - \sum_{j=2}^{k} w_j$ *and*

$$\boldsymbol{A}_k \begin{bmatrix} w_2 \\ w_3 \\ \vdots \\ w_k \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

where $\boldsymbol{A}_k$ *is the* $(k-1) \times (k-1)$ *matrix whose* $(r-1, j-1)$ *element is*

$$\left[20\left(1 + 1_{r=j}\frac{1}{r^6}\right) + \frac{6}{3^3}\left(1_{r=3} + 1_{j=3} - 1_{j=3r}\frac{1}{r^6} - 1_{r=3j}\frac{1}{j^6}\right)\right] \Big/ \left[20 + \frac{6}{3^3}1_{r=3}\right]$$

for $r, j \in \{2, \ldots, k\}$.

Proof:. Let $g_n := 2n^{-1} \times length(\mathcal{X})$. Then $x_{i+j} - x_{i-j} = g_n j$ and the variance of $Y_{i;k}^{(3)}$ is

$$2\sigma^2 g_n^{-6} \sum_{j=1}^{k} \left[\frac{10w_j^2}{j^6} - \frac{6w_j w_{3j}}{3^3 j^6}\, 1_{3j \leq k}\right], \tag{2.5}$$

Substituting $1 - \sum_{j=2}^{k} w_j$ for w_1 in (2.5) and setting the partial derivative with respect to w_r, $r \in \{2, \ldots, k\}$, equal to 0 yields

$$\begin{aligned} & 40\sigma^2 g_n^{-6}\left(\sum_{j=2}^{k} w_j - 1 + \frac{w_r}{r^6}\right) \\ - & \frac{12}{3^3}\sigma^2 g_n^{-6}\left(-w_3 1_{3\leq k} + \left\{1 - \sum_{j=2}^{k} w_j\right\} 1_{r=3} + \sum_{j=2}^{k}\left\{\frac{w_{3j}}{j^6}1_{3j\leq k, r=j} + \frac{w_j}{j^6}1_{r=3j}\right\}\right) \\ = & \ 0 \end{aligned}$$

from which we obtain

$$\sum_{j=2}^{k} w_j \left[20\left(1 + \frac{1_{r=j}}{r^6}\right) + \frac{6}{3^3}\left(1_{r=3} + 1_{j=3} - \frac{1_{j=3r}}{r^6} - \frac{1_{r=3j}}{j^6}\right)\right] \Big/ \left[20 + \frac{6}{3^3}1_{r=3}\right] = 1. \tag{2.6}$$

Writing out the $k-1$ equations of the form (2.6) as r ranges over $\{2, \ldots, k\}$, we acquire the matrix equation

$$\mathbf{A}_k \begin{bmatrix} w_2 \\ w_3 \\ \vdots \\ w_k \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

This completes the proof since the fact that (2.5) approaches ∞ as $\max_{j \in \{1,\ldots,k\}} |w_j| \to \infty$ implies that a global minimizer exists and is found by setting partial derivatives equal to 0. ■

Proposition 2.2.2 *Assume that model (2.1) holds. Then the variance of $Y_{i;k}^{(4)}$ from expression (2.3), where $4k+1 \leq i \leq n-4k$, is minimized when $w_1 = 1 - \sum_{j=2}^{k} w_j$ and*

$$\boldsymbol{A}_k \begin{bmatrix} w_2 \\ w_3 \\ \vdots \\ w_k \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

where $\boldsymbol{A}_k$ is the $(k-1) \times (k-1)$ matrix whose $(r-1, j-1)$ element is

$$\frac{\left[\left(140 - \frac{72}{r^4} - \frac{72}{j^4} + \frac{72}{(rj)^4}\right) + 68\frac{1_{r=j}}{r^8} + \left(1_{j=2} + 1_{r=2} - \frac{1_{j=2r}}{r^8} - \frac{1_{r=2j}}{j^8}\right)\right]}{\left[140 - \frac{72}{r^4} + 1_{r=2}\right]}$$

for $r, j \in \{2, \ldots, k\}$.

Proof:. Let $g_n := 2n^{-1} \times \mathrm{l}ength(\mathcal{X})$. Then $x_{i+j} - x_{i-j} = g_n j$ and the variance of $Y_{i;k}^{(4)}$ is

$$\sigma^2 g_n^{-8} \left[36 \left(\sum_{j=1}^{k} \frac{w_j}{j^4} \right)^2 + 2 \sum_{j=1}^{k} \left(\frac{17 w_j^2}{j^8} - \frac{8 w_j w_{2j}}{2^4 j^8} \, 1_{2j \leq k} \right) \right], \tag{2.7}$$

Substituting $1-\sum_{j=2}^{k} w_j$ for w_1 in (2.7) and setting the partial derivative with respect to w_r, $r \in \{2,\ldots,k\}$, equal to 0 yields

$$\begin{aligned} & 72\sigma^2 g_n^{-8}\left(1-\sum_{j=2}^{k} w_j + \sum_{j=2}^{k}\frac{w_j}{j^4}\right)\left(\frac{1}{r^4}-1\right) \\ + \;& 68\sigma^2 g_n^{-8}\left(\sum_{j=2}^{k} w_j - 1 + \frac{w_r}{r^8}\right) \\ - \;& \frac{16}{2^4}\sigma^2 g_n^{-8}\left(-w_2 + \left\{1-\sum_{j=2}^{k} w_j\right\}1_{r=2} + \sum_{j=2}^{k}\left\{\frac{w_{2j}}{j^8}1_{2j\le k, r=j} + \frac{w_j}{j^8}1_{r=2j}\right\}\right) \\ = \;& 0 \end{aligned}$$

from which we obtain

$$\frac{\sum_{j=2}^{k} w_j\left[\left(140-\frac{72}{r^4}-\frac{72}{j^4}+\frac{72}{(rj)^4}\right)+68\frac{1_{r=j}}{r^8}+\frac{16}{2^4}\left(1_{j=2}+1_{r=2}-\frac{1_{j=2r}}{r^8}-\frac{1_{r=2j}}{j^8}\right)\right]}{\left[140-\frac{72}{r^4}+\frac{16}{2^4}1_{r=2}\right]} = 1. \tag{2.8}$$

Writing out the $k-1$ equations of the form (2.8) as r ranges over $\{2,\ldots,k\}$, we acquire the matrix equation

$$\mathbf{A}_k \begin{bmatrix} w_2 \\ w_3 \\ \vdots \\ w_k \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

This completes the proof since the fact that (2.7) approaches ∞ as $\max_{j\in\{1,\ldots,k\}}|w_j| \to \infty$ implies that a global minimizer exists and is found by setting partial derivatives equal to 0. ■

The calculations of the variance-minimizing weights become increasingly difficult as the order of the empirical derivative and as k increase. Thus, rather than defining the weights to be variance-minimizing, we propose defining empirical derivatives of order q to have the form (2.3) with weights

$$w_j := \begin{cases} j^{2q}/\sum_{l=1}^{k} l^{2q} & : \; q \textit{ odd} \\ \frac{\left(j^{2q-1}\sum_{l=1}^{k} l^{q} - j^{2q}\sum_{l=1}^{k} l^{q-1}\right)}{\left(\sum_{l=1}^{k} l^{2q-1}\sum_{l=1}^{k} l^{q} - \sum_{l=1}^{k} l^{2q}\sum_{l=1}^{k} l^{q-1}\right)} & : \; q \textit{ even} \end{cases} \tag{2.9}$$

for $j \in \{1, \ldots, k\}$. Formula (2.9) with $q = 1$ is compatible with the definition of empirical first derivatives previously given.

The weights of definition (2.9) are much easier to compute than the variance-minimizing weights. As we will see, their simple closed form also allows us to demonstrate some nice asymptotic properties of empirical derivatives. Importantly, the weights of definition (2.9) are fairly similar to the variance minimizing weights. Table 2.1 compares, for $k \in \{2, 3, 4\}$ and $q \in \{1, 2, 3, 4\}$, the weights we use to define empirical derivatives to the weights that minimize the variance of $Y_{i;k}^{(q)}$.

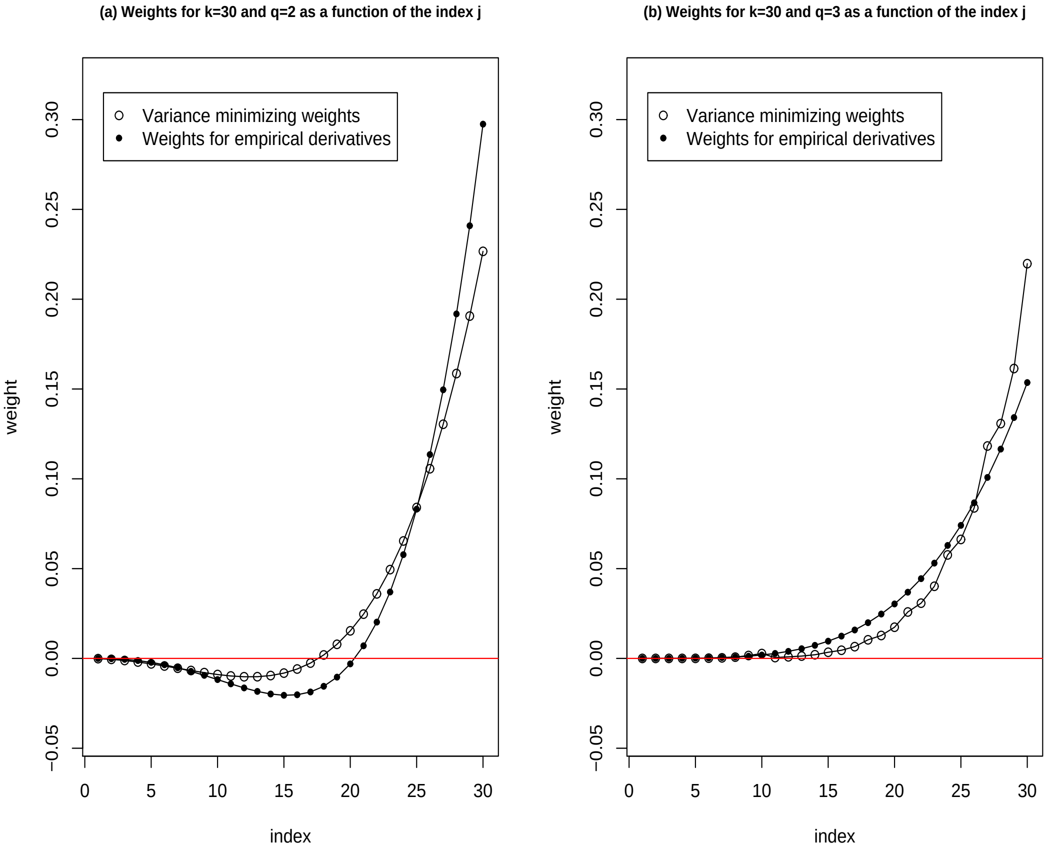
Table 2.1: Weights for empirical derivatives

	$k = 2$		$k = 3$			$k = 4$			
q, Method	$j = 1$	$j = 2$	$j = 1$	$j = 2$	$j = 3$	$j = 1$	$j = 2$	$j = 3$	$j = 4$
1, Minimize	.200	.800	.071	.286	.643	.033	.133	.300	.533
1, Empirical	.200	.800	.071	.286	.643	.033	.133	.300	.533
2, Minimize	-.143	1.143	-.071	.000	1.071	-.037	-.069	.136	.970
2, Empirical	-.333	1.333	-.095	-.190	1.286	-.037	-.148	0	1.185
3, Minimize	.015	.985	.011	.072	.916	.002	.013	.163	.822
3, Empirical	.015	.985	.001	.081	.918	.000	.013	.149	.838
4, Minimize	-.022	1.022	-.005	-.044	1.050	-.002	-.012	.015	.999
4, Empirical	-.067	1.067	-.003	-.180	1.183	.000	-.032	-.192	1.224

Method Minimize refers to choosing $w_1, \ldots, w_k$ so that the variance of $Y_{i;k}^{(q)}$ in expression (2.3) is minimized, which entails solving a system of linear equations. Method Empirical refers to choosing $w_1, \ldots, w_k$ according to prescription (2.9) for empirical derivatives.

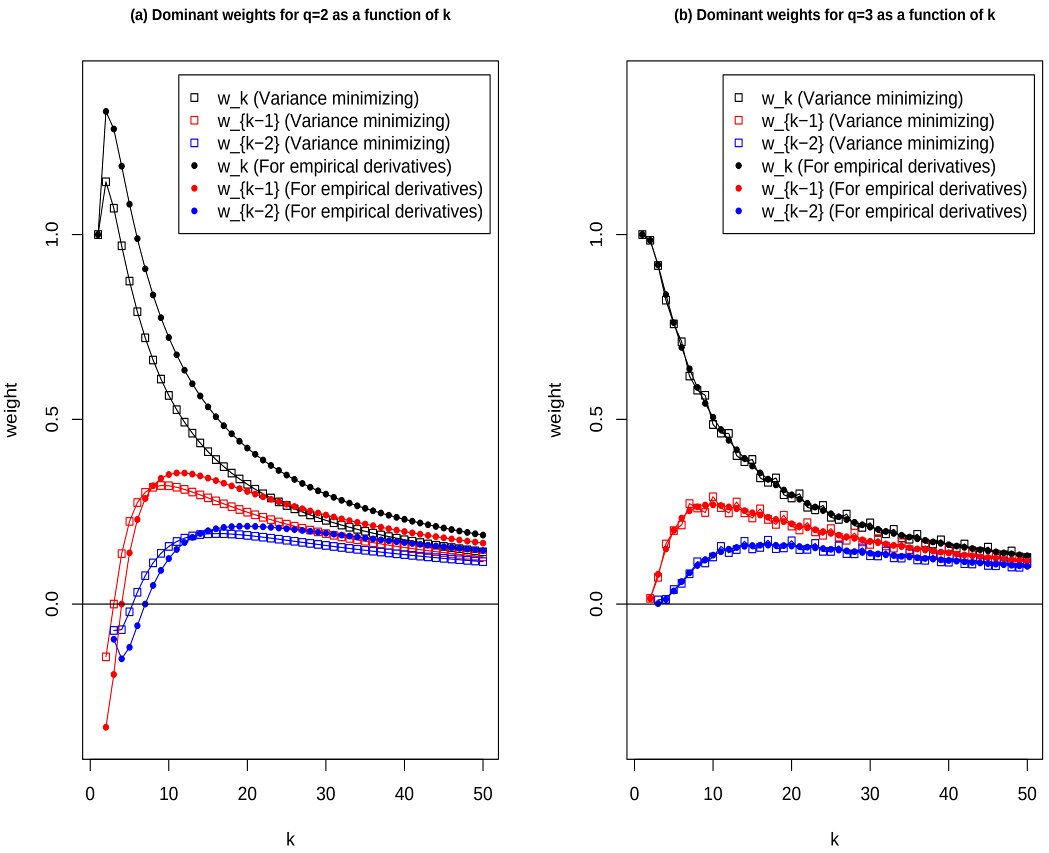
Some noteworthy features are shared by both the weights of Definition 2.9 and the variance minimizing weights. First, the weights are both positive and negative when q is even. The weights are all positive when q is odd. As a further illustration beyond Table 2.1, panel a of figure 2.1 shows that the first 20 weights from formula (2.9) are negative when $q = 2$ and $k = 30$. Likewise, the first 17 variance-minimizing weights are negative. In contrast, panel b of Figure 2.1 illustrates that the weights for empirical derivatives are all positive when q is odd. Second, the largest weights are the last few (i.e., those for which the index j is closest to k). However, even these become small as k increases. This is shown for $q = 2$ in panel a of Figure 2.2, which reveals that w_k, w_{k-1}, and w_{k-2} from formula (2.9) are eventually monotone decreasing in k. The variance-minimizing weights exhibit similar behavior. Panel b of Figure 2.2 demonstrates that w_k, w_{k-1}, and w_{k-2} from formula (2.9) become small with large k when $q = 3$.

Figure 2.1: Comparison of weights for empirical derivatives



Panel a shows $w_1, \ldots, w_{30}$ that minimize the variance of $Y_{i;30}^{(2)}$ in expression (2.3) as well as $w_1, \ldots, w_{30}$ that define second order empirical derivatives based on formula (2.9). Panel b shows $w_1, \ldots, w_{30}$ that minimize the variance of $Y_{i;30}^{(3)}$ in expression (2.3) as well as $w_1, \ldots, w_{30}$ that define third order empirical derivatives based on formula (2.9).

Figure 2.2: The largest weights



Panel a shows, as a function of $k \in \{1, \ldots, 50\}$, the w_k, w_{k-1}, and w_{k-2} with which the variance of $Y^{(2)}_{i;k}$ in expression (2.3) is minimized as well as the w_k, w_{k-1}, and w_{k-2} that define second order empirical derivatives based on formula (2.9). Panel b shows, as a function of $k \in \{1, \ldots, 50\}$, the w_k, w_{k-1}, and w_{k-2} with which the variance of $Y^{(3)}_{i;k}$ in expression (2.3) is minimized as well as the w_k, w_{k-1}, and w_{k-2} that define third order empirical derivatives based on formula (2.9).

We now examine some asymptotic properties of empirical derivatives. The following proposition and corollary examine the variance of $Y_{i;k}^{(1)}$.

Proposition 2.2.3 *Assume that model (2.1) holds. Suppose that there exist* $\alpha \in (0,1)$ *and* $c \in (0,\infty)$ *such that* $kn^{-\alpha} \to c$ *as* $n \to \infty$. *Then the empirical first derivatives with* $k+1 \leq i \leq n-k$ *have* $O\left(n^{2-3\alpha}\right)$ *variances.*

Proof: The variance of $Y_{i;k}^{(1)}$ is

$$\frac{\sigma^2 n^2}{2 \times length(\mathcal{X})^2} \left\{ \left(1 - \sum_{j=2}^{k} w_j\right)^2 + \sum_{j=2}^{k} \frac{w_j^2}{j^2} \right\}. \tag{2.10}$$

Substituting $j^2 / \sum_{l=1}^{k} l^2$ for w_j and using the identity $\sum_{j=1}^{k} j^2 = k^3/3 + k^2/2 + k/6$, we simplify expression (2.10) to

$$\frac{\sigma^2 n^2}{2 \times length(\mathcal{X})^2} \sum_{j=1}^{k} \frac{j^4}{j^2 (\sum_{l=1}^{k} l^2)^2} = \frac{\sigma^2 n^2}{2 \times length(\mathcal{X})^2} \frac{1}{k^3/3 + k^2/2 + k/6} = O\left(n^2 k^{-3}\right), \tag{2.11}$$

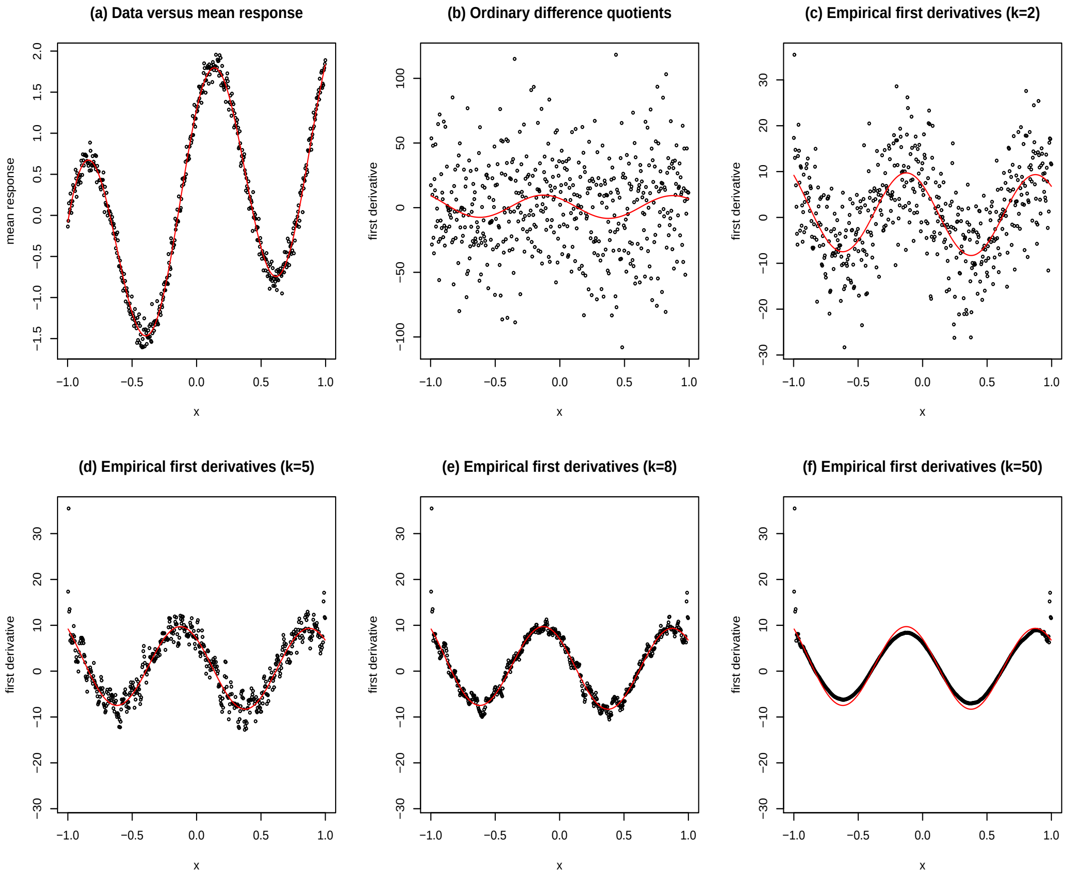
The last equality yields the desired result. ■

Proposition 2.2.3 would also hold if variance-minimizing weights were employed since such weights by definition minimize expression (2.10). Equality (2.11) yields an immediate corollary.

Corollary 2.2.1 *Suppose that there exists* $c \in (0,\infty)$ *such that* $kn^{-2/3} \to c$ *as* $n \to \infty$. *Then, under the conditions of Proposition 2.2.3, the variance of* $Y_{i;k}^{(1)}$ *converges to* $3\sigma^2/(2c^3 \times length(\mathcal{X})^2)$.

Charnigo and Srinivasan (2008) show that under the conditions of Proposition 2.2.3 empirical first derivatives have $O\left(n^{\alpha-1}\right)$ biases. Thus, both the variances and the biases of the empirical first derivatives tend to 0 when $\alpha \in (2/3, 1)$. The balancing of the variance and the bias for empirical first derivatives is demonstrated visually in the following figure.

Figure 2.3: Empirical first derivatives



Panel a shows a simulated data set of size 500 from model (2.1) with $\mu(x) := \cos(2\pi x) + \sin(2\pi x) + \log(4/3 + x)$, equispaced $x_i \in [-1, 1]$, and $\epsilon_i \overset{iid}{\sim} N(0, 0.1^2)$. Panel b displays ordinary first order difference quotients, which are barely distinguishable from noise; as a reference, $\mu'(x)$ is plotted against x for $x \in [-1, 1]$. Panels c through f depict empirical first derivatives for $k \in \{2, 5, 8, 50\}$. A careful look at panel f reveals that with $k = 50$ the negligible variance has come at the price of clear bias, at least for some values of the covariate x.

We now proceed with similar results for empirical second derivatives. The following proposition and corollary examine the variance of $Y_{i;k}^{(2)}$.

Proposition 2.2.4 *Assume that model (2.1) holds. Suppose that there exist* $\alpha \in (0,1)$ *and* $c \in (0,\infty)$ *such that* $kn^{-\alpha} \to c$ *as* $n \to \infty$. *Then the empirical second derivatives with* $2k+1 \leq i \leq n-2k$ *have* $O\left(n^{4-5\alpha}\right)$ *variances.*

Proof. Let $g_n := 2n^{-1} \times length(\mathcal{X})$. The variance of $Y_{i;k}^{(2)}$ is

$$\sigma^2 g_n^{-4} \left[\sum_{j=1}^{k} \frac{2w_j^2}{j^4} + 4 \left(\sum_{j=1}^{k} \frac{w_j}{j^2} \right)^2 \right]. \tag{2.12}$$

Substituting into (2.12) the weights prescribed by (2.9) yields

$$\begin{aligned}
& O(n^4) \sum_{j=1}^{k} \frac{1}{j^4} \left(\frac{j^3 \sum_{l=1}^{k} l^2 - j^4 \sum_{l=1}^{k} l}{\sum_{l=1}^{k} l^2 \sum_{l=1}^{k} l^3 - \sum_{l=1}^{k} l^4 \sum_{l=1}^{k} l} \right)^2 \\
& + O(n^4) \left[\sum_{j=1}^{k} \frac{1}{j^2} \left(\frac{j^3 \sum_{l=1}^{k} l^2 - j^4 \sum_{l=1}^{k} l}{\sum_{l=1}^{k} l^2 \sum_{l=1}^{k} l^3 - \sum_{l=1}^{k} l^4 \sum_{l=1}^{k} l} \right) \right]^2 \\
= \; & O(n^4) \sum_{j=1}^{k} \frac{1}{j^4} \left(j^3 O(k^{-4}) + j^4 O(k^{-5}) \right)^2 \\
& + O(n^4) \left[\; 0 \; \right]^2 \\
= \; & O(n^4) \sum_{j=1}^{k} \left(j^2 O(k^{-8}) + j^3 O(k^{-9}) + j^4 O(k^{-10}) \right) \\
= \; & O(n^4 k^{-5}),
\end{aligned} \tag{2.13}$$

which implies the desired result. ■

Careful bookkeeping of the coefficients involved in (2.13) yields the following corollary.

Corollary 2.2.2 *Suppose that there exists* $c \in (0,\infty)$ *such that* $kn^{-4/5} \to c$ *as* $n \to \infty$. *Then, under the conditions of Proposition 2.2.4, the variance of* $Y_{i;k}^{(2)}$ *converges to* $5\sigma^2/(3c^5 \times length(\mathcal{X})^4)$.

The following proposition examines the bias of $Y^{(2)}_{i;k}$.

Proposition 2.2.5 *Under the conditions of Proposition 2.2.4, the empirical second derivatives have* $O\left(n^{\alpha-1}\right)$ *biases.*

Proof. Let $B := \sup_{x\in\mathcal{X}}\left|\frac{d^3}{dx^3}\mu(x)\right|$. The compactness of $\mathcal{X}$ implies that B is finite. Noting that

$$\mu(x_{i+2j}) = \mu(x_i) + (g_n j)\mu'(x_i) + (g_n j)^2\mu''(x_i)/2 + (g_n j)^3\frac{d^3}{dx^3}\mu(\xi_{i,i+2j})/6$$

and

$$\mu(x_{i-2j}) = \mu(x_i) - (g_n j)\mu'(x_i) + (g_n j)^2\mu''(x_i)/2 - (g_n j)^3\frac{d^3}{dx^3}\mu(\xi_{i,i-2j})/6$$

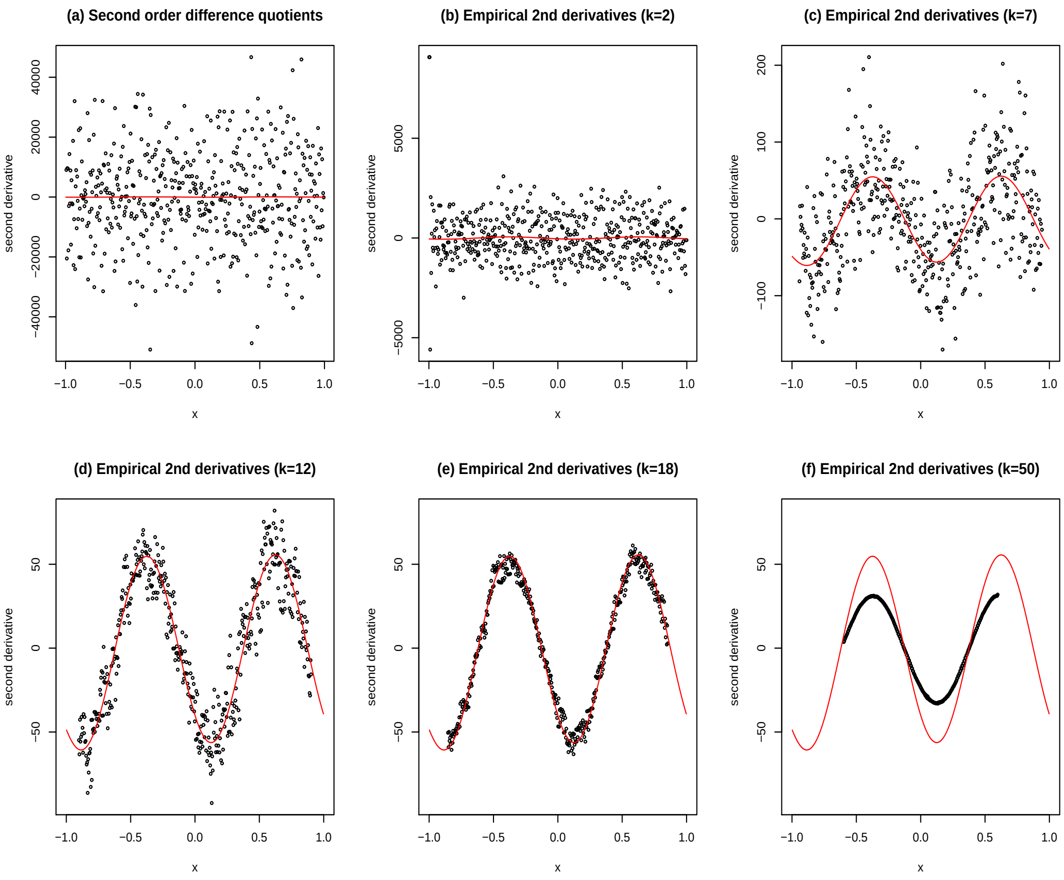
for some $\xi_{i,i+2j} \in [x_i, x_{i+2j}]$ and $\xi_{i,i-2j} \in [x_{i-2j}, x_i]$, we find that the absolute value of the bias of $Y^{(2)}_{i;k}$ is

$$\begin{aligned}
&\left|\sum_{j=1}^{k} w_j\left\{\frac{\mu(x_{i+2j}) - 2\mu(x_i) + \mu(x_{i-2j})}{(g_n j)^2} - \mu''(x_i)\right\}\right| \\
\leq\ & B\sum_{j=1}^{k}|w_j|\frac{g_n j}{3} \\
\leq\ & O(n^{-1})\sum_{j=1}^{k} j\frac{j^3\sum_{l=1}^{k} l^2 + j^4\sum_{l=1}^{k} l}{\left|\sum_{l=1}^{k} l^3\sum_{l=1}^{k} l^2 - \sum_{l=1}^{k} l^4\sum_{l=1}^{k} l\right|} \\
=\ & O(n^{-1})\sum_{j=1}^{k}\left(j^4 O(k^{-4}) + j^5 O(k^{-5})\right) \\
=\ & O(kn^{-1}).
\end{aligned}$$

The last equality completes the proof. ■

Thus, both the variances and the biases of the empirical second derivatives tend to 0 when $\alpha \in (4/5, 1)$. The balancing of the variance and the bias for empirical second derivatives is demonstrated visually in the following figure.

Figure 2.4: Empirical second derivatives



Panel a shows ordinary second order difference quotients based on the simulated data set from Figure 2.3; as a reference, $\mu''(x)$ is plotted against x for $x \in [-1, 1]$. Panel b shows empirical second derivatives with $k = 2$. Panels c through f show empirical second derivatives with $k \in \{7, 12, 18, 50\}$ and attention restricted to $2k + 1 \leq i \leq n - 2k$. Taking $k = 12$ yields empirical second derivatives of comparable visual quality to the empirical first derivatives in Figure 2.3 with $k = 5$, demonstrating that a larger q warrants a larger k. Taking $k = 50$ goes too far, however, yielding empirical second derivatives that seriously understate the local extrema of $\mu''(x)$. Another problem with taking k as large as 50 is that the restriction $2k + 1 \leq i \leq n - 2k$ then wipes out 40% of the values of the index i.

The following propositions for empirical third derivatives are similar to those for empirical first and second derivatives.

Proposition 2.2.6 *Assume that model (2.1) holds. Suppose that there exist* $\alpha \in (0,1)$ *and* $c \in (0,\infty)$ *such that* $kn^{-\alpha} \to c$ *as* $n \to \infty$*. Then the empirical third derivatives with* $3k+1 \le i \le n-3k$ *have* $O\left(n^{6-7\alpha}\right)$ *variances.*

Proof. Let $g_n := 2n^{-1} \times \mathrm{l}ength(\mathcal{X})$. The variance of $Y_{i;k}^{(3)}$ is

$$2^7\sigma^2 g_n^{-6} \sum_{j=1}^{k} \left[\frac{10w_j^2}{(2j)^6} - \frac{6w_{3j}w_j}{12^3 j^6} \, 1_{3j \le k} \right],$$

which is bounded above by

$$2^7\sigma^2 g_n^{-6} \sum_{j=1}^{k} \frac{10w_j^2}{(2j)^6}. \tag{2.14}$$

Substituting into (2.14) the weights prescribed by (2.9) yields

$$\begin{aligned} & O(n^6) \sum_{j=1}^{k} \frac{1}{j^6} \left(\frac{j^6}{\sum_{l=1}^{k} l^6} \right)^2 \\ = \; & O(n^6) \sum_{j=1}^{k} \frac{1}{j^6} \, j^{12} O(k^{-14}) \\ = \; & O(n^6 k^{-7}), \end{aligned} \tag{2.15}$$

which implies the desired result. ■

Careful bookkeeping of the coefficients involved in (2.15) yields the following corollary.

Corollary 2.2.3 *Suppose that there exists* $c \in (0,\infty)$ *such that* $kn^{-6/7} \to c$ *as* $n \to \infty$*. Then, under the conditions of Proposition 2.2.6, the variance of* $Y_{i;k}^{(3)}$ *is asymptotically bounded above by* $35\sigma^2/(16c^7 \times length(\mathcal{X})^6)$*.*

Proposition 2.2.7 *Under the conditions of Proposition 2.2.6, the empirical third derivatives have* $O\left(n^{\alpha-1}\right)$ *biases.*

Proof . Let $B := \sup_{x\in\mathcal{X}} \left|\frac{d^4}{dx^4}\mu(x)\right|$. The compactness of $\mathcal{X}$ implies that B is finite. We have

$$\begin{aligned}
\mu(x_{i+3j}) &= \mu(x_i) + (1.5g_nj)\mu'(x_i) + (1.5g_nj)^2\mu''(x_i)/2 \\
&\quad +(1.5g_nj)^3\frac{d^3}{dx^3}\mu(x_i)/6 + (1.5g_nj)^4\frac{d^4}{dx^4}\mu(\xi_{i,i+3j})/24, \\
\mu(x_{i+j}) &= \mu(x_i) + (0.5g_nj)\mu'(x_i) + (0.5g_nj)^2\mu''(x_i)/2 \\
&\quad +(0.5g_nj)^3\frac{d^3}{dx^3}\mu(x_i)/6 + (0.5g_nj)^4\frac{d^4}{dx^4}\mu(\xi_{i,i+j})/24, \\
\mu(x_{i-j}) &= \mu(x_i) - (0.5g_nj)\mu'(x_i) + (0.5g_nj)^2\mu''(x_i)/2 \\
&\quad -(0.5g_nj)^3\frac{d^3}{dx^3}\mu(x_i)/6 + (0.5g_nj)^4\frac{d^4}{dx^4}\mu(\xi_{i,i-j})/24,
\end{aligned}$$

and

$$\begin{aligned}
\mu(x_{i-3j}) &= \mu(x_i) - (1.5g_nj)\mu'(x_i) + (1.5g_nj)^2\mu''(x_i)/2 \\
&\quad -(1.5g_nj)^3\frac{d^3}{dx^3}\mu(x_i)/6 + (1.5g_nj)^4\frac{d^4}{dx^4}\mu(\xi_{i,i-3j})/24
\end{aligned}$$

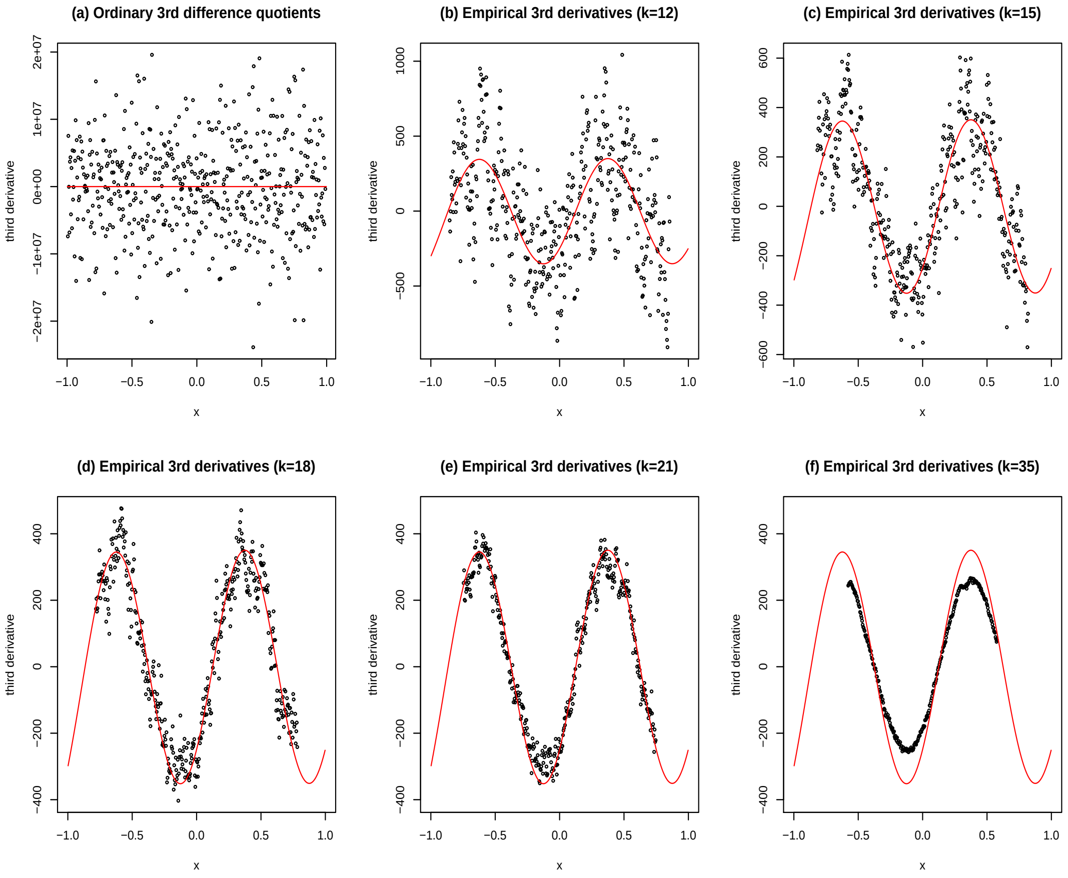
for some $\xi_{i,i+3j} \in [x_i, x_{i+3j}]$, $\xi_{i,i+j} \in [x_i, x_{i+j}]$, $\xi_{i,i-j} \in [x_{i-j}, x_i]$, and $\xi_{i,i-3j} \in [x_{i-3j}, x_i]$. As such, we find that the absolute value of the bias of $Y^{(3)}_{i;k}$ is

$$\begin{aligned}
&\left|\sum_{j=1}^{k} w_j \left\{\frac{\mu(x_{i+3j}) - 3\mu(x_{i+j}) + 3\mu(x_{i-j}) - \mu(x_{i-3j})}{(g_nj)^3} - \frac{d^3}{dx^3}\mu(x_i)\right\}\right| \\
\leq\ & B\sum_{j=1}^{k} w_j \frac{7g_nj}{16} \\
=\ & O(n^{-1})\sum_{j=1}^{k} j\frac{j^6}{\sum_{l=1}^{k} l^6} \\
=\ & O(n^{-1})\sum_{j=1}^{k} j^7 O(k^{-7}) \\
=\ & O(kn^{-1}).
\end{aligned}$$

The last equality completes the proof. ■

Thus, both the variances and the biases of the empirical third derivatives tend to 0 when $\alpha \in (6/7, 1)$. The balancing of the variance and the bias for empirical third derivatives is demonstrated visually in the following figure.

Figure 2.5: Empirical third derivatives



Panel a shows ordinary third order difference quotients based on the simulated data set from Figure 2.3; as a reference, $\mu'''(x)$ is plotted against x for $x \in [-1, 1]$. Panels b through f show empirical third derivatives with $k \in \{12, 15, 18, 21, 35\}$ and attention restricted to $3k + 1 \leq i \leq n - 3k$. Taking $k = 18$ yields empirical third derivatives of comparable visual quality to the empirical second derivatives in Figure 2.4 with $k = 12$, demonstrating once again that a larger q warrants a larger k. On the other hand, even taking $k = 35$ goes too far and bias becomes a serious problem. This indicates that the choice of k becomes more delicate as q increases.

The following propositions generalize the previous results for the variance of empirical derivatives.

Proposition 2.2.8 *Assume that model (2.1) holds. Suppose that there exist* $\alpha \in (0, 1)$ *and* $c \in (0, \infty)$ *such that* $kn^{-\alpha} \to c$ *as* $n \to \infty$*. Then if* q *is odd and* $3 \leq q \leq 7$*, the empirical* q*th derivatives with* $qk + 1 \leq i \leq n - qk$ *have* $O\left(n^{2q-(2q+1)\alpha}\right)$ *variances.*

Proof. Let $g_n := 2n^{-1} \times length(\mathcal{X})$. If q is odd and $3 \le q \le 7$, then the variance of $Y_{i;k}^{(q)}$ is

$$
\begin{aligned}
& 2\sigma^2 g_n^{-2q} \sum_{j=1}^{k} \left\{ \frac{\left[\sum_{a=0}^{\frac{q}{2}-\frac{1}{2}} \binom{q}{a}^2\right] w_j^2}{j^{2q}} \right. \\
& \left. +2 \sum_{a=0}^{\frac{q}{2}-\frac{3}{2}} \sum_{b=a+1}^{\frac{q}{2}-\frac{1}{2}} \left[\frac{(-1)^{a+b}\binom{q}{a}\binom{q}{b} w_{(q-2a)j} w_{(q-2b)j}}{(q-2a)^q (q-2b)^q j^{2q}} \right] 1_{[(q-2a)j \le k]} \right\} \\
= \; & 2\sigma^2 g_n^{-2q} \sum_{j=1}^{k} \left[\frac{\left[\sum_{a=0}^{\frac{q}{2}-\frac{1}{2}} \binom{q}{a}^2\right] w_j^2}{j^{2q}} \right] \\
& + \frac{4\sigma^2 g_n^{-2q}}{(\sum_{l=1}^{k} l^{2q})^2} \sum_{j=1}^{k} j^{2q} \sum_{a=0}^{\frac{q}{2}-\frac{3}{2}} \sum_{b=a+1}^{\frac{q}{2}-\frac{1}{2}} (-1)^{a+b} \binom{q}{a} \binom{q}{b} (q-2a)^q (q-2b)^q 1_{[(q-2a)j \le k]} \\
\le \; & 2\sigma^2 g_n^{-2q} \sum_{i=1}^{k} \left[\frac{\left[\sum_{a=0}^{\frac{q}{2}-\frac{1}{2}} \binom{q}{a}^2\right] w_i^2}{i^{2q}} \right] \qquad (2.16) \\
= \; & \frac{2\sigma^2 g_n^{-2q} \left[\sum_{a=0}^{\frac{q}{2}-\frac{1}{2}} \binom{q}{a}^2\right]}{\sum_{j=1}^{k} j^{2q}} \\
= \; & \frac{2\sigma^2 g_n^{-2q} \left[\sum_{a=0}^{\frac{q}{2}-\frac{1}{2}} \binom{q}{a}^2\right]}{(k^{2q+1})/(2q+1)} [1 + o(1)] \qquad (2.17) \\
= \; & O(n^{2q} k^{-(2q+1)})
\end{aligned}
$$

which yields the desired result. Line (2.16) is established in Lemma 2.2.1. Line (2.17) uses the fact that $\sum_{j=1}^{k} j^m = k^{m+1}/(m+1) + O(k^m)$ where m is a nonnegative integer.

■

Lemma 2.2.1 *For all odd $q \ge 3$,*

$$
\sum_{a=0}^{\frac{q}{2}-\frac{3}{2}} \sum_{b=a+1}^{\frac{q}{2}-\frac{1}{2}} \left[(-1)^{a+b} \binom{q}{a} \binom{q}{b} (q-2a)^q (q-2b)^q \right] 1_{[(q-2a)j \le k]} \le 0.
$$

Proof. First, note that

$$
\frac{\binom{q}{x}(q-2x)^q}{\binom{q}{x+1}(q-2(x+1))^q} = \frac{x+1}{q-x} \left(\frac{q-2x}{q-2x-2} \right)^q \qquad (2.18)
$$

is positive and increasing for all $x < (q-1)/2$ where x is a nonnegative integer. Hence, for such x, the quantity $\binom{q}{x}(q-2x)^q$ is decreasing in x if and only if (2.18) is greater than 1. This implies that once $\binom{q}{x}(q-2x)^q$ becomes decreasing, it remains decreasing for all subsequent $x < (q-1)/2$. Thus the quantity $\binom{q}{x}(q-2x)^q$ must satisfy one of the following three cases:

Case 1: $\binom{q}{x}(q-2x)^q$ is increasing for $0 \le x \le (q-1)/2$,

Case 2: $\binom{q}{x}(q-2x)^q$ is decreasing for $0 \le x \le (q-1)/2$, or

Case 3: $\binom{q}{x}(q-2x)^q$ is increasing for $0 \le x \le m$ where m is a nonnegative integer and $\binom{q}{x}(q-2x)^q$ is decreasing for $m \le x \le (q-1)/2$.

Consider case 1. Consider also that

$$\binom{q}{\frac{q-3}{2}}\left[q-2\left(\frac{q-3}{2}\right)\right]^q - \binom{q}{\frac{q-1}{2}}\left[q-2\left(\frac{q-1}{2}\right)\right]^q = \frac{2[q(3^q-1)-(3^q+3)](q!)}{(\frac{q+1}{2})!(\frac{q-3}{2})!(q+3)(q-1)}$$

which is positive, implying that $\binom{q}{x}(q-2x)^q$ is decreasing for $(q-3)/2 \le x \le (q-1)/2$. Thus we have a contradiction and are left with cases 2 and 3.

Consider case 2. The fact that $\binom{q}{x}(q-2x)^q$ is positive and decreasing for $0 \le x \le (q-1)/2$ implies that for $0 \le a \le (q-3)/2$,

$$\sum_{b=a+1}^{(q-1)/2} (-1)^{a+b}\binom{q}{b}(q-2b)^q \le 0.$$

This in turn implies that

$$\sum_{a=0}^{(q-3)/2}\binom{q}{a}(q-2a)^q \sum_{b=a+1}^{(q-1)/2}\left[(-1)^{a+b}\binom{q}{b}(q-2b)^q\right]1_{(q-2a)j\le k} \le 0,$$

which implies the desired result.

Finally, consider case 3. Similar to case 2, the fact that $\binom{q}{x}(q-2x)^q$ is positive and decreasing for $m \le x \le (q-1)/2$ implies that

$$\sum_{a=m}^{(q-3)/2}\binom{q}{a}(q-2a)^q \sum_{b=a+1}^{(q-1)/2}(-1)^{a+b}\binom{q}{b}(q-2b)^q 1_{(q-2a)j\le k} \le 0. \tag{2.19}$$

For $0 \le a \le m-1$,

$$\binom{q}{a}(q-2a)^q \sum_{b=a+1}^{(q-1)/2}(-1)^{a+b}\binom{q}{b}(q-2b)^q = f_0(a) - f_1(a) + f_2(a) - f_3(a),$$

where

$$\begin{aligned}
f_0(a) \quad &:= \quad \binom{q}{a}(q-2a)^q 1_{a\le m-2}\sum_{b=a+1}^{m-1}\binom{q}{b}(q-2b)^q 1_{a+b\in\mathcal{E}}, \\
&= \quad \binom{q}{a}(q-2a)^q 1_{a\le m-2}\sum_{b=a+2}^{m-1}\binom{q}{b}(q-2b)^q 1_{a+b\in\mathcal{E}}
\end{aligned}$$

$$f_1(a) := \binom{q}{a}(q-2a)^q 1_{a\le m-2}\sum_{b=a+1}^{m-1}\binom{q}{b}(q-2b)^q 1_{a+b\in\mathcal{O}},$$

$$f_2(a) := \binom{q}{a}(q-2a)^q 1_{a+m\in\mathcal{E}}\sum_{b=m}^{(q-1)/2}(-1)^{a+b}\binom{q}{b}(q-2b)^q,$$

and

$$f_3(a) := -\binom{q}{a}(q-2a)^q 1_{a+m\in\mathcal{O}}\sum_{b=m}^{(q-1)/2}(-1)^{a+b}\binom{q}{b}(q-2b)^q,$$

with $\mathcal{O}$ denoting the set of odd positive integers and $\mathcal{E}$ denoting the set of even positive integers. Note that $f_0(a)$, $f_1(a)$, $f_2(a)$, $f_3(a)\ge 0$. Note also that for $0\le a<m-1$, $f_0(a)\le f_1(a+1)$ and $f_2(a)\le f_3(a+1)$. In addition, $f_0(m-1)=f_2(m-1)=0$. These facts together imply that

$$\begin{aligned}
&\sum_{a=0}^{m-1}\left[f_0(a)-f_1(a)+f_2(a)-f_3(a)\right]1_{(q-2a)j\le k} \\
= \quad &\sum_{a=0}^{m-1}\binom{q}{a}(q-2a)^q\sum_{b=a+1}^{(q-1)/2}\left[(-1)^{a+b}\binom{q}{b}(q-2b)^q\right]1_{(q-2a)j\le k} \\
\le \quad &0
\end{aligned}$$

This, combined with (2.19), yields the desired result. ■

Proposition 2.2.9 *Assume that model (2.1) holds. Suppose that there exist* $\alpha\in(0,1)$ *and* $c\in(0,\infty)$ *such that* $kn^{-\alpha}\to c$ *as* $n\to\infty$. *Then if* q *is even and* $2\le q\le 6$, *the empirical* q^{th} *derivatives with* $qk+1\le i\le n-qk$ *have* $O\left(n^{2q-(2q+1)\alpha}\right)$ *variances.*

Proof. Let $g_n := 2n^{-1}\times length(\mathcal{X})$. Let $I(a,j) := 1_{[(\frac{q}{2}-a)j\le k]}$. Let $h(c) := \binom{q}{c}\left(\frac{q}{2}-c\right)^q$. Assume q is even and $2\le q\le 6$. Take $S_1=\sum_{l=1}^{k} l^q$, $S_2=\sum_{l=1}^{k} l^{q-1}$, and $S_3 = (\sum_{l=1}^k l^{2q-1})(\sum_{l=1}^k l^q) - (\sum_{l=1}^k l^{2q})(\sum_{l=1}^k l^{q-1})$. Then the variance of $Y_{i;k}^{(q)}$ is

$$\begin{aligned}
& 2\sigma^2 g_n^{-2q} \sum_{j=1}^k \left\{ \frac{w_j^2 \sum_{a=0}^{\frac{q}{2}-1} \binom{q}{a}^2}{j^{2q}} + 2\sum_{a=0}^{\frac{q}{2}-2}\sum_{b=a+1}^{\frac{q}{2}-1} \frac{(-1)^{a+b}\binom{q}{a}\binom{q}{b} w_{(\frac{q}{2}-a)j} w_{(\frac{q}{2}-b)j}}{(\frac{q}{2}-a)^q(\frac{q}{2}-b)^q j^{2q}} I(a,j) \right\} \\
& +\sigma^2 g_n^{-2q}\left[\binom{q}{q/2}\sum_{j=1}^k \frac{w_j}{j^q}\right]^2 \\
= \; & 2\sigma^2 g_n^{-2q} \sum_{j=1}^k \left\{ \frac{w_j^2 \sum_{a=0}^{\frac{q}{2}-1} \binom{q}{a}^2}{j^{2q}} + 2\sum_{a=0}^{\frac{q}{2}-2}\sum_{b=a+1}^{\frac{q}{2}-1} \frac{(-1)^{a+b}\binom{q}{a}\binom{q}{b} w_{(\frac{q}{2}-a)j} w_{(\frac{q}{2}-b)j}}{(\frac{q}{2}-a)^q(\frac{q}{2}-b)^q j^{2q}} I(a,j) \right\} \\
& +\sigma^2 g_n^{-2q}\left[0\right]^2 \\
= \; & 2\sigma^2 g_n^{-2q} \sum_{j=1}^k \left\{ \frac{w_j^2 \sum_{a=0}^{\frac{q}{2}-1} \binom{q}{a}^2}{j^{2q}} \right. \\
& \left. +\frac{2j^{2q}}{{S_3}^2}\sum_{a=0}^{\frac{q}{2}-2}\sum_{b=a+1}^{\frac{q}{2}-1} (-1)^{a+b}h(a)h(b)\left[\frac{S_1}{(\frac{q}{2}-a)j} - S_2\right]\left[\frac{S_1}{(\frac{q}{2}-b)j} - S_2\right] I(a,j) \right\} \\
= \; & 2\sigma^2 g_n^{-2q} \sum_{j=1}^k \left[\frac{w_j^2 \sum_{a=0}^{\frac{q}{2}-1} \binom{q}{a}^2}{j^{2q}} \right] + O(n^{2q}k^{-(2q+1)}) \qquad (2.20) \\
= \; & \frac{2\sigma^2 \sum_{a=0}^{\frac{q}{2}-1} \binom{q}{a}^2}{g_n^{2q}} \cdot \frac{(\sum_{j=1}^k j^{2q-2})S_1^2 - 2(\sum_{j=1}^k j^{2q-1})S_1S_2 + (\sum_{j=1}^k j^{2q})S_2^2}{S_3^2} \\
& +O(n^{2q}k^{-(2q+1)}) \\
= \; & \frac{8\sigma^2 g_n^{-2q}\left\{\sum_{a=0}^{\frac{q}{2}-1}\binom{q}{a}^2\right\} q(2q+1)k^{-(2q+1)}}{(2q-1)}[1+o(1)] + O(n^{2q}k^{-(2q+1)}) \\
= \; & O(n^{2q}k^{-(2q+1)})
\end{aligned}$$

which yields the desired result. Line (2.20) is established in Lemma (2.2.2). The proof again uses the fact that $\sum_{j=1}^k j^m = k^{m+1}/(m+1) + O(k^m)$ where m is a nonnegative integer. ■

Lemma 2.2.2 *Under the conditions of Proposition (2.2.9),*

$$\sum_{j=1}^k \frac{2j^{2q}}{{S_3}^2}\sum_{a=0}^{\frac{q}{2}-2}\sum_{b=a+1}^{\frac{q}{2}-1} (-1)^{a+b}h(a)h(b)\left[\frac{S_1}{(\frac{q}{2}-a)j} - S_2\right]\left[\frac{S_1}{(\frac{q}{2}-b)j} - S_2\right] I(a,j)$$

where $h(c) := \binom{q}{c}\left(\frac{q}{2}-c\right)^q$ *and* $I(a,j) := 1_{[(\frac{q}{2}-a)j\le k]}$ *is* $O(k^{-(2q+1)})$.

Proof. First, note that the indicator function, $I(a, j) := 1_{[(\frac{q}{2}-a)j \leq k]}$ is only on for $1 \leq (q/2 - b)j \leq (q/2 - a)j \leq k$. This means that

$$\frac{h(a)j^q}{(q/2-a)j} = \binom{q}{a}\left[(q/2-a)j\right]^{q-1} \leq \binom{q}{a}k^{q-1}$$

and

$$\frac{h(b)j^q}{(q/2-b)j} = \binom{q}{b}\left[(q/2-b)j\right]^{q-1} \leq \binom{q}{b}k^{q-1}.$$

Also, since $\sum_{j=1}^{k} j^m \leq k^{m+1}$ for any positive integer m, $S_1 \leq k^{q+1}$ and $S_2 \leq k^q$. This implies that

$$|S_1 - S_2(q/2-a)j| \leq k^{q+1}$$

and

$$|S_1 - S_2(q/2-b)j| \leq k^{q+1}.$$

Together these imply that

$$\begin{aligned}
&\sum_{a=0}^{\frac{q}{2}-2}\sum_{b=a+1}^{\frac{q}{2}-1}(-1)^{a+b}\frac{h(a)h(b)j^{2q-2}}{(\frac{q}{2}-a)(\frac{q}{2}-b)}\left[S_1 - S_2\left(\frac{q}{2}-a\right)j\right]\left[S_1 - S_2\left(\frac{q}{2}-b\right)j\right]I(a,j) \\
\leq\ &\sum_{a=0}^{\frac{q}{2}-2}\sum_{b=a+1}^{\frac{q}{2}-1}\left\{\frac{h(a)j^q}{(\frac{q}{2}-a)j}\right\}\left|S_1 - S_2\left(\frac{q}{2}-a\right)j\right|\left\{\frac{h(b)j^q}{(\frac{q}{2}-b)j}\right\}\left|S_1 - S_2\left(\frac{q}{2}-b\right)j\right|I(a,j) \\
\leq\ &\sum_{a=0}^{\frac{q}{2}-2}\sum_{b=a+1}^{\frac{q}{2}-1}\binom{q}{a}\binom{q}{b}k^{4q}.
\end{aligned}$$

This implies that

$$\sum_{j=1}^{k}\sum_{a=0}^{\frac{q}{2}-2}\sum_{b=a+1}^{\frac{q}{2}-1}(-1)^{a+b}\frac{h(a)h(b)j^{2q-2}}{(\frac{q}{2}-a)(\frac{q}{2}-b)}\left[S_1 - S_2\left(\frac{q}{2}-a\right)j\right]\left[S_1 - S_2\left(\frac{q}{2}-b\right)j\right]I(a,j)$$

is $O(k^{4q+1})$. Division by S_3^2, which is $O(k^{6q+2})$, then yields the desired result. ■

Proposition 2.2.10 *Assume that model (2.1) holds. Suppose that there exist* $\alpha \in (0, 1]$ *and* $c \in (0, \infty)$ *such that* $kn^{-\alpha} \to c$ *as* $n \to \infty$*. Then empirical* q^{th} *derivatives have* $O(n^{\alpha-1})$ *biases.*

Proof. Let $B := \sup_{x\in\mathcal{X}} \left|\frac{d^{q+1}}{dx^{q+1}}\mu(x)\right|$. The compactness of $\mathcal{X}$ implies that B is finite.

Let $g_n := 2n^{-1} \times \mathrm{l}ength(\mathcal{X})$. Applying Taylor's Theorem, there exist $\xi_{i+j} \in [x_i, x_{i+j}]$ and $\xi_{i-j} \in [x_{i-j}, x_i]$, such that for $d \in \{0, 1, ..., q\}$, $i \in \{qk + 1, ..., n - qk\}$ and $j \in \{i, ..., k\}$,

$$\begin{aligned}
& \frac{\mu^{(d)}(x_{i+j}) - \mu^{(d)}(x_{i-j})}{g_n j} \\
= \quad & \mu^{(d+1)}(x_i)1_{d<q} + \sum_{b=2}^{q-d} \frac{\mu^{(d+b)}(x_i)}{b!}\left(\frac{g_n j}{2}\right)^{b-1} 1_{[b\in\mathbb{O}]} \\
+ \quad & \frac{(\frac{j}{2}g_n)^{q-d}\mu^{(q+1)}(\xi_{i+j})}{2(q-d+1)!} + \frac{(-\frac{j}{2}g_n)^{q-d}\mu^{(q+1)}(\xi_{i-j})}{2(q-d+1)!} \\
= \quad & \mu^{(d+1)}(x_i)1_{d<q} + \sum_{b=2}^{q-d} c_b\mu^{(d+b)}(x_i)(g_n j)^{b-1} \\
+ \quad & c_{q-d+1}(g_n j)^{q-d}\mu^{(q+1)}(\xi_{i+j}) + c_{q-d+1}(g_n j)^{q-d}\mu^{(q+1)}(\xi_{i-j})(-1)^{q-d}
\end{aligned}$$

where c_b depends only on b for $b \in \{1, ..., q - d + 1\}$ and $\mathbb{O}$ denotes the set of positive odd integers. Importantly, c_b does not depend on j. This implies that there exist $\xi_{i+j,p} \in [x_i, x_{i+j}]$ and $\xi_{i-j,p} \in [x_{i-j}, x_i]$ for $p \in \{1, ..., q\}$ such that

$$\begin{aligned}
& \Bigg\{\Bigg(\mu^{(d)}(x_{i+j}) + \sum_{b=1}^{q-d} c_b\mu^{(d+b)}(x_{i+j})(g_n j)^b + \sum_{p=1}^{d} k_{p,1}\mu^{(q+1)}(\xi_{i+j,p})(g_n j)^{(q-d+1)}\Bigg) - \\
& \Bigg(\mu^{(d)}(x_{i-j}) + \sum_{b=1}^{q-d} c_b\mu^{(d+b)}(x_{i-j})(g_n j)^b + \sum_{p=1}^{d} k_{p,2}\mu^{(q+1)}(\xi_{i-j,p})(g_n j)^{(q-d+1)}\Bigg)\Bigg\} / g_n j \\
= \quad & \mu^{(d+1)}(x_i)1_{d<q} + \sum_{b=1}^{q-d-1} k_b\mu^{(d+b+1)}(x_i)(g_n j)^b \\
+ \quad & \sum_{p=1}^{d+1} \left\{k_{p,1}\mu^{(q+1)}(\xi_{i+j,p})(g_n j)^{q-d} + k_{p,2}\mu^{(q+1)}(\xi_{i-j,p})(g_n j)^{q-d}\right\} \qquad (2.21)
\end{aligned}$$

where k_b depends only on b, and $k_{p,1}$ and $k_{p,2}$ depend only on p. Note that both terms involved in the subtraction in the numerator of the left-hand side of (2.21) are in the form of the right-hand side of (2.21). In fact, (2.21) shows that subtracting one term of this form from another and dividing by $g_n j$ results in a term of the same form with d incremented by 1, the covariate value re-centered, and different coefficients. Now define

$$Z_{i,j}^{(1)} = \frac{Y_{i+j} - Y_{i-j}}{g_n j} = \left[\frac{\mu(x_{i+j}) - \mu(x_{i-j})}{g_n j} + \frac{\epsilon_{i+j} - \epsilon_{i-j}}{g_n j}\right] \tag{2.22}$$

and

$$Z_{i,j}^{(d)} = \frac{Z_{i+j,j}^{(d-1)} - Z_{i-j,j}^{(d-1)}}{g_n j}.$$

Note that when we take expectations the random component (the term in (2.22) involving the ϵ's) is eliminated. Therefore, by (2.21),

$$\begin{aligned}\mathbb{E}[Z_{i,j}^{(1)}] &= \mu'(x_i) + \sum_{b=1}^{q-1} k_b \mu^{(b+1)}(x_i)(g_n j)^b \\ &+ \sum_{p=1}^{q} \left\{k_{p,1}\mu^{(q+1)}(\xi_{i+j,p})(g_n j)^q + k_{p,2}\mu^{(q+1)}(\xi_{i-j,p})(g_n j)^q\right\}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[Z_{i,j}^{(d)}] &= \mu^{(d)}(x_i) + \sum_{b=1}^{q-d} k_b \mu^{(b+d)}(x_i)(g_n j)^b \\ &+ \sum_{p=1}^{q} \left\{k_{p,1}\mu^{(q+1)}(\xi_{i+j,p})(g_n j)^{q-d+1} + k_{p,2}\mu^{(q+1)}(\xi_{i-j,p})(g_n j)^{q-d+1}\right\}.\end{aligned}$$

Now consider that for $q > 1$,

$$Y_{i;k}^{(q)} = \sum_{j=1}^{k} w_j Z_{i,j}^{(q)}.$$

So

$$\mathbb{E}\left[Y_{i;k}^{(q)}\right] = \sum_{j=1}^{k} w_j \left[\mu^{(q)}(x_i) + \sum_{p=1}^{q} k_{p,1}\mu^{(q+1)}(\xi_{i+j,p}) g_n j + \sum_{p=1}^{q} k_{p,2}\mu^{(q+1)}(\xi_{i-j,p}) g_n j\right]$$

So if q is odd, the bias of $Y_{i;k}^{(q)}$ is bounded in absolute value by

$$\sum_{j=1}^{k} \left(\frac{j^{2q+1} g_n}{\sum_{l=1}^{k} l^{2q}}\right) \left(\sum_{p=1}^{q} (k_{p,1} + k_{p,2}) B\right) = O(kn^{-1})$$

and if q is even, the bias of $Y_{i;k}^{(q)}$ is bounded in absolute value by

$$\sum_{j=1}^{k} \left(\frac{(j^{2q} S_1 + j^{2q+1} S_2) g_n}{S_3}\right) \left(\sum_{p=1}^{q} (k_{p,1} + k_{p,2}) B\right) = O(kn^{-1}).$$

This implies the desired result. ■

2.3 Simulation studies of generalized C(p)

Recall that generalized C(p) was defined so that it matched its target in expected value up to a remainder term. To be more specific, Charnigo and Srinivasan (2008) provide the following theorem:

Theorem 2.3.1 *Assume that model (2.1) holds. Then*

$$\mathbb{E}\left[GCP(\boldsymbol{Y}, \widehat{\mu})\right] = \mathbb{E}\left[\sum_{i=1}^{n} s_i \left(\frac{d^q}{dx^q}\mu(x_i) - \widehat{\frac{d^q}{dx^q}\mu(x_i)}\right)^2\right] + \sum_{i=1}^{n}\left\{s_i r_i^2 - 2 s_i r_i b_i\right\},$$

where

$$b_i := \mathbb{E}\left[\widehat{\frac{d^q}{dx^q}\mu(x_i)}\right] - \frac{d^q}{dx^q}\mu(x_i) = \sum_{m=1}^{n} \mu(x_m)\frac{d^q}{dx^q} l_m(x_i) - \frac{d^q}{dx^q}\mu(x_i)$$

and

$$r_i := \mathbb{E}\left[\sum_{m=1}^{n} c_{i,m} Y_m\right] - \frac{d^q}{dx^q}\mu(x_i) = \sum_{m=1}^{n} c_{i,m}\mu(x_m) - \frac{d^q}{dx^q}\mu(x_i).$$

This result is important, but it also raises a couple of important issues. The first is the importance of the remainder term. Under certain assumptions, the remainder term is asymptotically irrelevant.

The second issue is that for generalized C(p) to be an effective method for tuning parameter selection it really needs to do more than match its target in expected value. The key is, after all, whether or not generalized C(p) can find tuning parameters that minimize (or nearly minimize) the target for a given realization of data. So what we really need is for the dependence of generalized C(p) upon the tuning parameters to be similar to the dependence of the target upon the tuning parameters. That is, a contour plot of generalized C(p) over the vector of tuning parameters needs to look like a contour plot of the target over the vector of tuning parameters. We address this issue with simulation studies, both by examining the success of generalized C(p) in finding the minimizing (or nearly minimizing) values of the tuning parameters and by comparing contour plots of generalized C(p) and the target.

For our simulation study to assess the performance of generalized C(p) we set the mean response function to be $\mu(x) := \cos(2\pi x) + \sin(2\pi x) + \log(4/3 + x)$ on

$\mathcal{X} := [-1, 1]$ and estimated three derivatives using compound estimation with $J := 3$ and 27 centering points (Charnigo and Srinivasan, 2010a) with the inductive pointwise estimators (Charnigo and Srinivasan, 2010b). The tuning parameters under consideration, h (where $h = h_2 = h_3$) and β, were put into a vector λ that ranged through $\Lambda := \{(2^{-a/2}, 2^{b/2})' : a, b \in \{6, 7, \ldots, 14\}\}$. We wanted to see whether generalized C(p) could identify a λ that minimized or nearly minimized the target $\sum_{i=1}^{n} s_i \left(\frac{d^q}{dx^q}\mu(x_i) - \widehat{\frac{d^q}{dx^q}\mu(x_i)} \right)^2$ at $q \in \{1, 2, 3\}$. For simplicity and fairness, and to mitigate boundary issues, we set $s_i := 1_{13 \le i \le 488}$ at $q = 1$, $s_i := 1_{49 \le i \le 452}$ at $q = 2$, and $s_i := 1_{109 \le i \le 392}$ at $q = 3$.

Twenty-five data sets of size $n = 500$ were generated from model (2.1) with the mean response as indicated above and $\epsilon_i \overset{iid}{\sim} N(0, 0.1^2)$. For each data set and each $\lambda \in \Lambda$, we calculated the generalized C(p) criterion (2.4) at $q = 1$ four times using four sets of empirical derivatives, one for each $k \in \{3, 6, 9, 12\}$. We also computed generalized C(p) at $q = 2$ four times, once for each $k \in \{6, 12, 18, 24\}$, and at $q = 3$ four times, once for each $k \in \{9, 18, 27, 36\}$. To avoid ambiguity in presenting our results, we hereafter attach subscripts of "1", "2", or "3" to k according to whether $q = 1$, 2, or 3.

We examined six quantities in our assessment. The first is

$$Q_1 := \sum_{i=1}^{n} s_i \left(\frac{d}{dx}\mu(x_i) - \widehat{\frac{d}{dx}\mu_{\widehat{\lambda}}(x_i)} \right)^2 / \min_{\lambda \in \Lambda} \sum_{i=1}^{n} s_i \left(\frac{d}{dx}\mu(x_i) - \widehat{\frac{d}{dx}\mu_{\lambda}(x_i)} \right)^2 .$$

The second is

$$Q_2 := \sum_{i=1}^{n} s_i \left(\frac{d}{dx}\mu(x_i) - \widehat{\frac{d}{dx}\mu_{\widehat{\lambda}}(x_i)} \right)^2 / \sum_{\lambda \in \Lambda} |\Lambda|^{-1} \sum_{i=1}^{n} s_i \left(\frac{d}{dx}\mu(x_i) - \widehat{\frac{d}{dx}\mu_{\lambda}(x_i)} \right)^2 ,$$

where $|\Lambda|$ is the number of elements in Λ.

It should be noted that Q_1 cannot be less than 1 and represents the degree to which $\widehat{\lambda}$ chosen by generalized C(p) inflates the target beyond its minimum value over $\lambda \in \Lambda$ at $q = 1$. On the other hand, Q_2 can be less than 1 and reflects the extent to which $\widehat{\lambda}$ deflates or inflates the target relative to a haphazard choice of the tuning parameter from Λ at $q = 1$.

The third and fourth quantities (Q_3 and Q_4) are analogously defined for the second derivative, while the fifth and sixth quantities (Q_5 and Q_6) correspond to the third derivative.

The results are summarized in Table 2.2. Regarding estimation of the first derivative, generalized C(p) is highly successful in that, for any $k_1 \in \{3, 6, 9, 12\}$,

$$\sum_{i=1}^{n} s_i \left(\frac{d}{dx}\mu(x_i) - \widehat{\frac{d}{dx}\mu_{\widehat{\lambda}}(x_i)} \right)^2$$

is usually very close to

$$\min_{\lambda \in \Lambda} \sum_{i=1}^{n} s_i \left(\frac{d}{dx}\mu(x_i) - \widehat{\frac{d}{dx}\mu_{\lambda}(x_i)} \right)^2 .$$

The median excess of the former over the latter is 4.5% or less, while the upper quartile of the excess is 17.4% or less. Likewise, the quantity $\sum_{i=1}^{n} s_i \left(\frac{d}{dx}\mu(x_i) - \widehat{\frac{d}{dx}\mu_{\widehat{\lambda}}(x_i)} \right)^2$ is much lower than the average that such a quantity takes over Λ. The median reduction as compared to the average over Λ is 92.7% or more.

Table 2.2: Results from simulation studies assessing generalized C(p)

	Q_1			Q_2		
k_1	25^{th}	50^{th}	75^{th}	25^{th}	50^{th}	75^{th}
3	1	1.00120	1.17415	.04888	.07303	.09364
6	1	1.04514	1.15421	.05172	.06625	.08074
9	1	1.04537	1.12771	.04344	.06897	.08209
12	1	1	1.06717	.05123	.05774	.07064
	Q_3			Q_4		
k_2	25^{th}	50^{th}	75^{th}	25^{th}	50^{th}	75^{th}
6	1.27065	2.01808	2.96714	.34445	.44733	.68571
12	1.14121	1.38769	1.84555	.23505	.29373	.34414
18	1	1.13218	1.56073	.19304	.22598	.34013
24	1.07415	1.13939	1.32915	.20455	.26645	.33587
	Q_5			Q_6		
k_3	25^{th}	50^{th}	75^{th}	25^{th}	50^{th}	75^{th}
9	9.08715	18.65524	27.22242	.42120	.76122	1.14737
18	1	1	1	.03713	.04834	.07309
27	1	1	27.50306	.04030	.07232	1.07667
36	35.74749	42.52426	51.01249	1.39332	1.71011	1.87213

The entries in columns 25^{th}, 50^{th}, and 75^{th} show the respective percentiles of Q_1 through Q_6 based on 25 simulated data sets of size $n = 500$ from model (2.1) with $\mu(x) := \sin(2\pi x) + \cos(2\pi x) + \log(4/3 + x)$, x_i equispaced on $[-1, 1]$, and $\epsilon_i \overset{iid}{\sim} N(0, 0.1^2)$. The numbers k_1, k_2, and k_3 identify how many summands are in the empirical first, second, and third derivatives defining generalized C(p).

With respect to the second derivative, generalized C(p) does fairly well for $k_2 = 18$ and $k_2 = 24$ (median excess 13.9% or less, median reduction 73.4% or more) but somewhat less well for $k_2 = 12$ and much less well for $k_2 = 6$. As for the third derivative, generalized C(p) does very well with $k_3 = 18$ (median and upper quartile excess 0%) but is hit-or-miss with $k_3 = 27$ (median excess 0%, upper quartile excess 2650%) and fares quite badly with $k_3 = 9$ and $k_3 = 36$.

Choosing a tuning parameter to optimize estimation for a higher order derivative is more difficult than doing so for a lower order derivative, and — especially for a higher order derivative — the performance of generalized C(p) can be sensitive to the number of summands in the empirical derivatives.

In the following graphs, we present contour plots for both the true value of the target and generalized C(p) when estimating the first three derivatives of the mean response with the compound estimator using $J = 3$, 27 centering points, and the inductive pointwise estimators. These plots come from a *single* data set using the mean response function given above with $\epsilon_i \overset{iid}{\sim} N(0, 0.1^2)$. For calculating generalized C(p) we used $k_1 = 9$, $k_2 = 18$, and $k_3 = 27$.

Figures 2.6 and 2.7 are contour plots for the target and generalized C(p), respectively, at the first derivative evaluated over the grid of possible tuning parameters. The striking similarity between the plots is an indication that generalized C(p) would be an excellent criterion for tuning parameter selection. Tuning parameters that result in a small generalized C(p) (the purple region of Figure 2.7) also result in a small target. Similarly, tuning parameters that generalized C(p) indicate should be avoided (the yellow region of Figure 2.7) are indeed poor choices based on the target.

Figures 2.8 and 2.9 are contour plots for the target and generalized C(p), respectively, at the second derivative evaluated over the grid of possible tuning parameters. The similarities between the plots are again strong. It is again evident that tuning parameters chosen to make generalized C(p) small will lead to small values of the target.

Figures 2.10 and 2.11 are contour plots for the target and generalized C(p), respectively, at the third derivative evaluated over the grid of possible tuning parameters.

Although the similarities between the contour plots are diminished slightly, they are still strong enough to indicate that utilizing generalized C(p) as a guide is a great improvement over haphazardly choosing tuning parameters.

Figure 2.6: Contour Plot of the Target (1st Derivative)

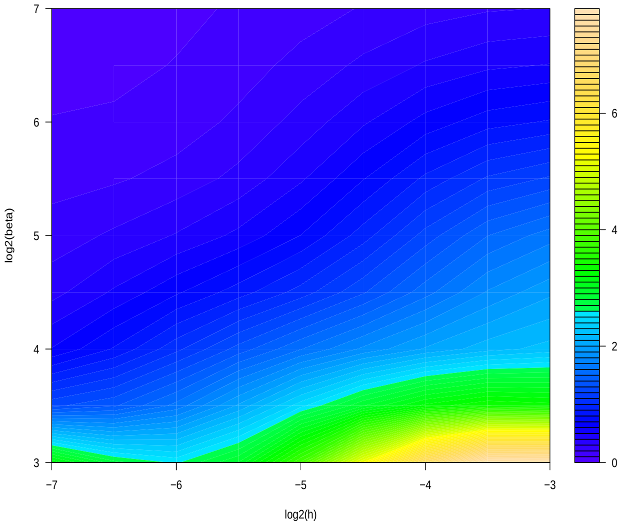


Figure 2.7: Contour Plot of the Generalized C(p) (1st Derivative)

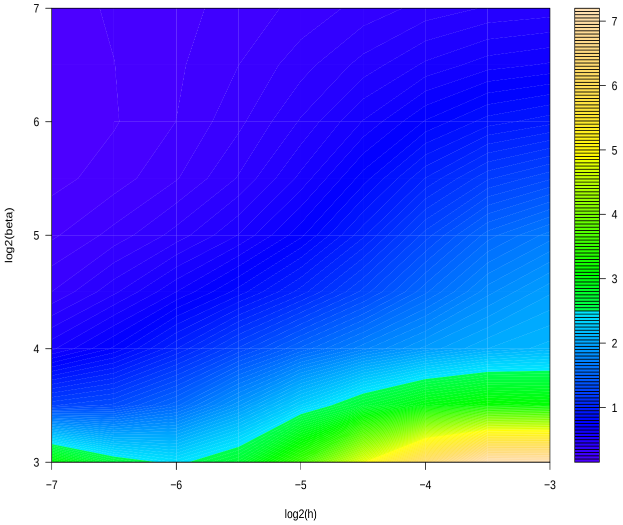


Figure 2.8: Contour Plot of the Target (2nd Derivative)

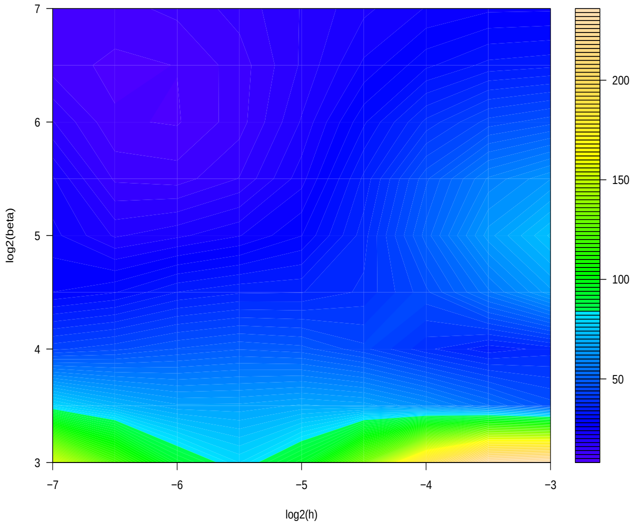


Figure 2.9: Contour Plot of the Generalized C(p) (2nd Derivative)

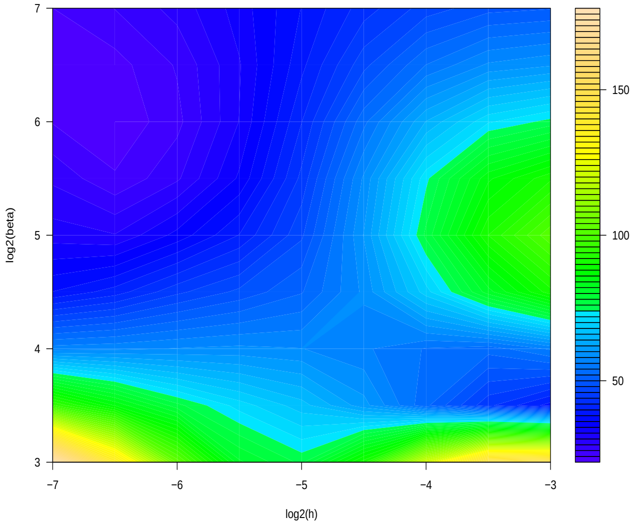


Figure 2.10: Contour Plot of the Target (3rd Derivative)

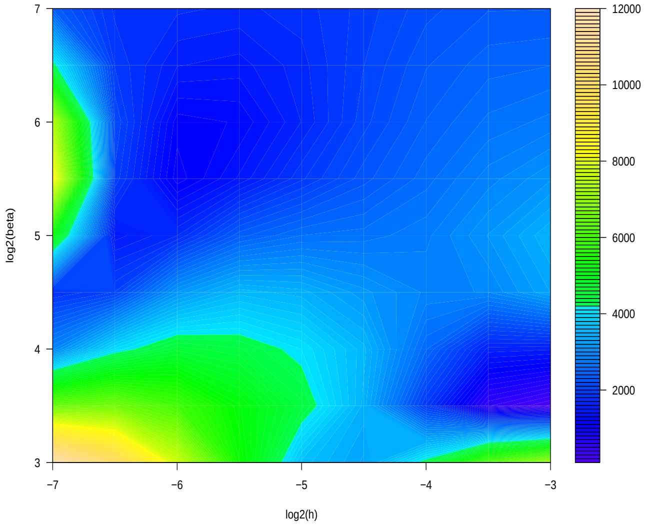
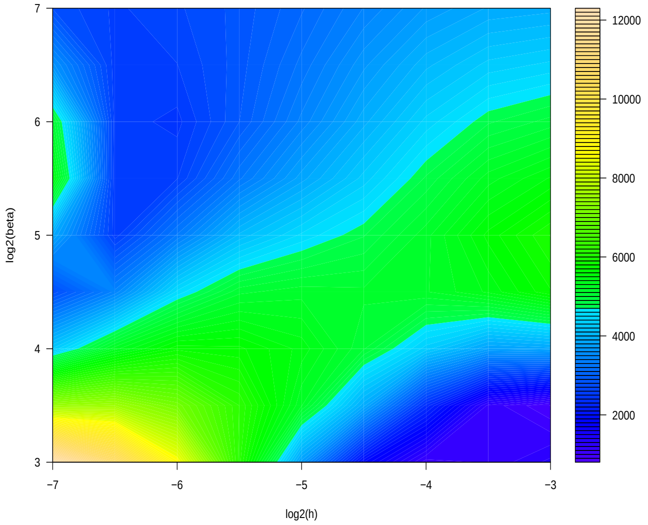


Figure 2.11: Contour Plot of the Generalized C(p) (3rd Derivative)



Copyright© Benjamin Hall, 2010.

Chapter 3 Simultaneous Confidence Bands for a Function and Its Derivatives

3.1 Previous work by Knafl, Sacks, and Ylvisaker

Suppose we observe data from model (1.1) and the ϵ_i are independently normally distributed with common variance σ^2. Also consider estimating the mean response function with any nonparametric estimator that is linear in the observed responses (i.e. satisfying (1.4). Knafl et al. (1985) exploit both the normality and the linearity in this situation to create conservative confidence bands for $\mu(x)$.

They begin by considering the situation in which $\widehat{\mu}(x)$ is unbiased. This situation is unrealistic in nonparametric regression but is true for simple linear regression. In such a case, $\widehat{\mu}(x) - \mu(x) = \sum_{i=1}^{n} l_i(x)\epsilon_i$. Now if $\mathbf{G} = \{\xi_1, ..., \xi_G\}$ is a grid of points from the covariate space and we define

$$Z(x) := \frac{\sum_{i=1}^{n} l_i(x)\epsilon_i}{\sigma D(x)}, \text{where } D(x) = \sqrt{\sum_{i=1}^{n} l_i(x)^2},$$

then

$$\left[\begin{array}{cccc} Z(\xi_1) & Z(\xi_2) & \cdots & Z(\xi_G) \end{array}\right]^t \sim MVN\left(\mathbf{0}, \Sigma\right)$$

where Σ has diagonal entries of 1 and off-diagonal entries of

$$\Sigma_{kj} = \frac{\sum_{i=1}^{n} l_i(\xi_k) l_i(\xi_j)}{D(\xi_k) D(\xi_j)}.$$

The next step is to note that

$$\begin{aligned} & P\left(\max_{x \in \mathbf{G}} |Z(x)| > z_\alpha\right) \\ \leq \quad & P\left(|Z(\xi_1)| > z_\alpha\right) + \sum_{j=1}^{G-1} P\left(|Z(\xi_j)| \leq z_\alpha, |Z(\xi_{j+1})| > z_\alpha\right). \end{aligned} \tag{3.1}$$

If G is small and the correlations between the $Z(\xi_j)$ are large, then the conservativeness of (3.1) will be small.

Now if z_α is chosen to make (3.1) equal to α, then

$$P\left(|\widehat{\mu}(x)-\mu(x)| \leq z_\alpha \sigma D(x), x \in \mathbf{G}\right) \geq 1-\alpha. \tag{3.2}$$

Inequality (3.2) is only valid for unbiased estimators of $\mu(x)$. In the case of nonparametric regression, the bias will need to be considered. If $B(x)$ is the absolute value of the bias, then by the triangle inequality

$$\begin{aligned} & P\left(|\widehat{\mu}(x)-\mu(x)| \leq B(x)+z_\alpha \sigma D(x), x \in \mathbf{G}\right) \\ \geq\ & P\left(\left|\sum_{i=1}^{n} l_i(x)\epsilon_i\right| \leq z_\alpha \sigma D(x), x \in \mathbf{G}\right) \\ \geq\ & 1-\alpha. \end{aligned} \tag{3.3}$$

The final step is to make the bands valid over the entire covariate space rather than just at the grid points. This can be accomplished through linear interpolation. From (3.3),

$$P\left(|\widehat{\mu}_I(x)-\mu_I(x)| \leq B_I(x)+z_\alpha \sigma D_I(x), x \in \mathcal{X}\right) \geq 1-\alpha,$$

where the subscript I indicates linear interpolation between the grid points. One more application of the triangle inequality yields

$$P\left(|\widehat{\mu}_I(x)-\mu(x)| \leq |\mu(x)-\mu_I(x)|+B_I(x)+z_\alpha \sigma D_I(x), x \in \mathcal{X}\right) \geq 1-\alpha.$$

Unfortunately, the quantity $\mu(x)-\mu_I(x)$ is unknown. Knafl et al solve this problem by restricting attention to a class of functions which will allow this difference to be bounded. For example, if $\mathbf{G}$ is a uniform grid and

$$|\mu'(x)| \leq m, \forall x \in \mathcal{X}$$

then

$$\sup_{x \in \mathcal{X}} |\mu(x)-\mu_I(x)| \leq m\gamma/2,$$

where γ is the mesh size of the grid. In the end, a $100(1-\alpha)\%$ confidence band would be

$$\widehat{\mu}_I(x) \pm \left(m\gamma/2+B_I(x)+z_\alpha \sigma D_I(x)\right).$$

In what follows we extend this confidence bands approach to the situation where the bands are not only simultaneous over the covariate, but also over the mean function and one or more derivatives. Quantities like bias, interpolation error, and noise variance are generally unknown but are, for convenience, often treated as if they were known or as if upper bounds for them were available. We explicitly address the estimation of bias, interpolation error, and noise variance. While our methodology for simultaneous confidence bands can work with upper bounds for such quantities, it does not rely on the availability of these upper bounds.

3.2 Confidence bands over a function and its derivatives

In this section we define confidence bands that are simultaneous over the covariate and over the mean response and one or more derivatives. This requires multiple regions each defined by an upper and a lower boundary. To be precise, $L_0(x), ..., L_J(x)$ and $U_0(x), ..., U_J(x)$ form conservative $100(1-\alpha)\%$ confidence bands for the mean response function and the first J derivatives if

$$P(L_0(x) \leq \mu(x) \leq U_0(x), ..., L_J(x) \leq \mu^{(J)}(x) \leq U_J(x), \forall x \in \mathcal{X}) \geq 1 - \alpha. \tag{3.4}$$

To proceed, in addition to the requirement of the previous section that the nonparametric regression estimator be linear in the observed responses, in what follows we require that it also be self-consistent. Thus, for $q \in \{1, 2, ..., J\}$,

$$\widehat{\mu^{(q)}(x)} = \sum_{i=1}^{n} l_i^{(q)}(x) Y_i. \tag{3.5}$$

We begin by considering the situation where simultaneous confidence bands are to be placed around the mean response and its first derivative. If we define

$$Z_1(x) := \frac{\sum_{i=1}^{n} l_i'(x)\epsilon_i}{\sigma D_1(x)}, \text{where } D_1(x) = \sqrt{\sum_{i=1}^{n} l_i'(x)^2}.$$

then

$$\begin{bmatrix} Z(\xi_1) & \cdots & Z(\xi_G) & Z_1(\xi_1) & \cdots & Z_1(\xi_G) \end{bmatrix}^t \sim MVN\left(\mathbf{0}, \Sigma\right)$$

where Σ has 1's on the diagonal with

$$\mathrm{Cov}\left(Z(\xi_j), Z(\xi_k)\right) = \frac{\sum_{i=1}^{n} l_i(\xi_k) l_i(\xi_j)}{D(\xi_k)D(\xi_j)},$$

$$\mathrm{Cov}\left(Z_1(\xi_j), Z_1(\xi_k)\right) = \frac{\sum_{i=1}^{n} l_i'(\xi_k) l_i'(\xi_j)}{D_1(\xi_k)D_1(\xi_j)},$$

and

$$\mathrm{Cov}\left(Z(\xi_j), Z_1(\xi_k)\right) = \frac{\sum_{i=1}^{n} l_i(\xi_j) l_i'(\xi_k)}{D(\xi_j)D_1(\xi_k)}.$$

Now consider that

$$\begin{aligned} & P\left(\max_{x\in\mathbf{G}} |Z(x)| > z_\alpha \text{ or } \max_{x\in\mathbf{G}} |Z_1(x)| > z_\alpha\right) \\ \leq \; & P\left(|Z(\xi_1)| > z_\alpha \text{ or } |Z_1(\xi_1)| > z_\alpha\right) \qquad (3.6) \\ + \; & \sum_{j=1}^{G-1} P\left(|Z(\xi_j)| \leq z_\alpha, |Z_1(\xi_j)| \leq z_\alpha, \{|Z(\xi_{j+1})| > z_\alpha \text{ or } |Z_1(\xi_{j+1})| > z_\alpha\}\right). \end{aligned}$$

We could then choose z_α so that (3.6) is equal to α. The benefit of (3.6) is that probabilities involving multivariate normal vectors of dimension 4 are numerically much easier to evaluate than probabilities involving multivariate normal vectors of dimension $2G$. However, we note that this choice of z_α can be refined.

To obtain a less conservative approximation for z_α, we could note that

$$\begin{aligned} & P\left(\max_{x\in\mathbf{G}} |Z(x)| > z_\alpha \text{ or } \max_{x\in\mathbf{G}} |Z_1(x)| > z_\alpha\right) \\ \leq \; & P\left(|Z(\xi_1)| > z_\alpha \text{ or } |Z_1(\xi_1)| > z_\alpha\right) \qquad (3.7) \\ + \; & P\left(|Z(\xi_1)| \leq z_\alpha, |Z_1(\xi_1)| \leq z_\alpha, \{|Z(\xi_2)| > z_\alpha \text{ or } |Z_1(\xi_2)| > z_\alpha\}\right) \\ + \; & \sum_{j=1}^{G-2} P\left(|Z(\xi_j)| \leq z_\alpha, |Z_1(\xi_j)| \leq z_\alpha,\right. \\ & \left.|Z(\xi_{j+1})| \leq z_\alpha, |Z_1(\xi_{j+1})| \leq z_\alpha, \{|Z(\xi_{j+2})| > z_\alpha \text{ or } |Z_1(\xi_{j+2})| > z_\alpha\}\right). \end{aligned}$$

We then choose z_α so that (3.7) is equal to α. Since

$$\begin{aligned} & \sum_{j=1}^{G-1} P\left(|Z(\xi_j)| \leq z_\alpha, |Z_1(\xi_j)| \leq z_\alpha, \{|Z(\xi_{j+1})| > z_\alpha \text{ or } |Z_1(\xi_{j+1})| > z_\alpha\}\right) \\ \geq \; & P\left(|Z(\xi_1)| \leq z_\alpha, |Z_1(\xi_1)| \leq z_\alpha, \{|Z(\xi_2)| > z_\alpha \text{ or } |Z_1(\xi_2)| > z_\alpha\}\right) \\ + \; & \sum_{j=1}^{G-2} P\left(|Z(\xi_j)| \leq z_\alpha, |Z_1(\xi_j)| \leq z_\alpha,\right. \\ & \left.|Z(\xi_{j+1})| \leq z_\alpha, |Z_1(\xi_{j+1})| \leq z_\alpha, \{|Z(\xi_{j+2})| > z_\alpha \text{ or } |Z_1(\xi_{j+2})| > z_\alpha\}\right), \end{aligned}$$

the z_α determined by (3.7) is less than that determined by (3.6). We could continue to refine the approximation by choosing z_α so that:

$$\begin{aligned}
& P\left(|Z(\xi_1)| > z_\alpha \text{ or } |Z_1(\xi_1)| > z_\alpha\right) \\
+ \; & P\left(|Z(\xi_1)| \leq z_\alpha, |Z_1(\xi_1)| \leq z_\alpha, \{|Z(\xi_2)| > z_\alpha \text{ or } |Z_1(\xi_2)| > z_\alpha\}\right) \\
+ \; & P\left(|Z(\xi_1)| \leq z_\alpha, |Z_1(\xi_1)| \leq z_\alpha, |Z(\xi_2)| \leq z_\alpha, |Z_1(\xi_2)| \leq z_\alpha,\right. \\
& \left.\{|Z(\xi_3)| > z_\alpha \text{ or } |Z_1(\xi_3)| > z_\alpha\}\right) \\
+ \; & \sum_{j=1}^{G-3} P\left(|Z(\xi_j)| \leq z_\alpha, |Z_1(\xi_j)| \leq z_\alpha, |Z(\xi_{j+1})| \leq z_\alpha, |Z_1(\xi_{j+1})| \leq z_\alpha,\right. \\
& \left.|Z(\xi_{j+2})| \leq z_\alpha, |Z_1(\xi_{j+2})| \leq z_\alpha, \{|Z(\xi_{j+3})| > z_\alpha \text{ or } |Z_1(\xi_{j+3})| > z_\alpha\}\right) \\
= \; & \alpha
\end{aligned}$$

In fact, such refinements can continue and eventually yield an exact value for z_α. However, further refinements require more coding and computational resources and yield decreasing returns. The following Table 3.1 illustrates how successive refinements of the approximation result in smaller (less conservative) z_α.

Table 3.1: Cutoff approximations for the mean response and first derivative

Refinement	1	2	3	4
n=100, G=25	3.218	3.200	3.194	3.192
n=100, G=50	3.314	3.294	3.280	3.276
n=500, G=25	3.217	3.198	3.192	3.190
n=500, G=50	3.313	3.293	3.279	3.275

For sample sizes of 100 and 500 equally spaced values from $\mathcal{X} = [-1, 1]$ and uniform grid sizes of 25 and 50, the table shows $z_{.05}$ approximations if the compound estimator is used with $J = 3$, 27 centering points, $\beta = 100$, and local regression pointwise estimators with nearest neighbor fractions of .3. Refinement 1 corresponds to (3.6). Refinement 2 corresponds to (3.7). Refinements 3 and 4 are the next two successive refinements of the approximation.

Now assume that upper bounds on the absolute value of the bias for the mean function, $B(x)$, and on the absolute value of the bias for the first derivative, $B_1(x)$, are available. (The bias of the estimator of the mean response is $\sum_{i=1}^{n} l_i(x)\mu(x_i) - \mu(x)$

and the bias of the estimator of the first derivative is $\sum_{i=1}^{n} l_i'(x)\mu(x_i) - \mu'(x)$.) Then

$$\begin{aligned}
& P(|\widehat{\mu}(x) - \mu(x)| \leq B(x) + z_\alpha \sigma D(x), |\widehat{\mu}'(x) - \mu'(x)| \leq B_1(x) + z_\alpha \sigma D_1(x), x \in \mathbf{G}) \\
= \; & P\left(\left| \sum_{i=1}^{n} l_i(x)[\mu(x_i) + \epsilon_i] - \mu(x) \right| \leq B(x) + z_\alpha \sigma D(x), \right. \\
& \left. \left| \sum_{i=1}^{n} l_i'(x)[\mu(x_i) + \epsilon_i] - \mu'(x) \right| \leq B_1(x) + z_\alpha \sigma D_1(x), x \in \mathbf{G} \right) \\
\geq \; & P\left(\left| \sum_{i=1}^{n} l_i(x)\epsilon_i \right| \leq z_\alpha \sigma D(x), \left| \sum_{i=1}^{n} l_i'(x)\epsilon_i \right| \leq z_\alpha \sigma D_1(x), x \in \mathbf{G} \right) \qquad (3.8) \\
\geq \; & 1 - \alpha, \qquad (3.9)
\end{aligned}$$

where line (3.8) follows from the triangle inequality and line (3.9) follows from the determination of z_α above.

This means that

$$\widehat{\mu}(x) \pm (B(x) + z_\alpha \sigma D(x))$$

and

$$\widehat{\mu}'(x) \pm (B_1(x) + z_\alpha \sigma D_1(x))$$

provide at least $100(1-\alpha)\%$ coverage for the mean response and the first derivative at the grid points.

To obtain confidence bands that are valid over $\mathcal{X}$ we further assume that upper bounds on $\sup_{x \in \mathcal{X}} |\mu(x) - \mu_I(x)|$, call it M_0, and on $\sup_{x \in \mathcal{X}} |\mu'(x) - \mu_I'(x)|$, call it M_1, where $\mu_I'(x)$ is the linear interpolant of $\mu'(x)$, are available. Now

$$\begin{aligned}
& 1 - \alpha \\
\leq \; & P(|\widehat{\mu}(x) - \mu(x)| \leq B(x) + z_\alpha \sigma D(x), |\widehat{\mu}'(x) - \mu'(x)| \leq B_1(x) + z_\alpha \sigma D_1(x), x \in \mathbf{G}) \\
= \; & P(|\widehat{\mu}_I(x) - \mu_I(x)| \leq B_I(x) + z_\alpha \sigma D_I(x), \\
& |\widehat{\mu}_I'(x) - \mu_I'(x)| \leq B_{1I}(x) + z_\alpha \sigma D_{1I}(x), x \in \mathcal{X}) \\
\leq \; & P(|\widehat{\mu}_I(x) - \mu(x)| \leq M_0 + B_I(x) + z_\alpha \sigma D_I(x), \\
& |\widehat{\mu}_I'(x) - \mu'(x)| \leq M_1 + B_{1I}(x) + z_\alpha \sigma D_{1I}(x), x \in \mathcal{X}), \qquad (3.10)
\end{aligned}$$

where line (3.10) follows from the assumptions above and the triangle inequality.

The preceding paragraphs constitute a proof of the following theorem:

Theorem 3.2.1 *Assume that model (1.1) holds where the ϵ_i are independent and identically normally distributed. Let $\widehat{\mu}(x)$ be self-consistent and linear in the observed responses. Let $\boldsymbol{G}$, $B(x)$, $B_1(x)$, M_0, M_1, $D(x)$ and $D_1(x)$ be as defined above and let the subscript I denote linear interpolation between grid points. Then*

$$P(L_0(x) \leq \mu(x) \leq U_0(x), L_1(x) \leq \mu'(x) \leq U_1(x), \forall x \in \mathcal{X}) \geq 1 - \alpha,$$

where

$$U_0(x), L_0(x) := \widehat{\mu}_I(x) \pm (M_0 + B_I(x) + z_\alpha \sigma D_I(x))$$

and

$$U_1(x), L_1(x) := \widehat{\mu}'_I(x) \pm (M_1 + B_{1I}(x) + z_\alpha \sigma D_{1I}(x)).$$

If one is interested in higher-order derivatives, the approach is similar. More than that, this methodology can be used to create simultaneous confidence bands for any combination of derivatives. For instance, simultaneous confidence bands can be created for the first and second derivatives even if no confidence bands are created for the mean response. Furthermore, this methodology can be applied to any subinterval of $\mathcal{X}$. In fact, the subintervals do not have to be the same for each derivative under consideration. For instance, confidence bands could be constructed that are simultaneous over the entire covariate space for the mean response and over a subinterval for the first derivative. This generalization is described explicitly below in Theorem 3.2.3. Such flexibility is attractive since it may reflect regions over which the researcher is comfortable making assumptions about bias and interpolation error.

If confidence bands around derivatives of order $p_1, p_2, ..., p_J$ are desired, define

$$Z_p(x) := \frac{\sum_{i=1}^n l_i^{(p)}(x)\epsilon_i}{\sigma D_p(x)}, \text{where } D_p(x) = \sqrt{\sum_{i=1}^n l_i^{(p)}(x)^2}$$

for each $p \in \{p_1, ..., p_J\}$. Then

$$\begin{bmatrix} Z_{p_1}(\xi_1) & \cdots & Z_{p_1}(\xi_G) & \cdots & Z_{p_J}(\xi_1) & \cdots & Z_{p_J}(\xi_G) \end{bmatrix}^t \sim MVN(\mathbf{0}, \Sigma) \qquad (3.11)$$

where Σ has 1's on the diagonal with

$$\text{Cov}(Z_a(\xi_k), Z_b(\xi_j)) = \frac{\sum_{i=1}^n l_i^{(a)}(\xi_k) l_i^{(b)}(\xi_j)}{D_a(\xi_k) D_b(\xi_j)}$$

for $a, b \in \{p_1, p_2, \cdots, p_J\}$.

Then

$$\begin{aligned} & P\left(\cup_{r=1}^{J} \max_{x \in \mathbf{G}} |Z_{p_r}(x)| > z_\alpha\right) \\ \leq \; & P\left(\cup_{r=1}^{J} |Z_{p_r}(\xi_1)| > z_\alpha\right) \\ + \; & \sum_{j=1}^{G-1} P\left(\{\cap_{r=1}^{J} |Z_{p_r}(\xi_j)| \leq z_\alpha\} \cap \{\cup_{r=1}^{J} |Z_{p_r}(\xi_{j+1})| > z_\alpha\}\right). \end{aligned} \tag{3.12}$$

We choose z_α so that (3.12) is equal to α. Again, this approximation can be refined. To obtain a less conservative approximation for z_α, note that

$$\begin{aligned} & P\left(\cup_{r=1}^{J} \max_{x \in \mathbf{G}} |Z_{p_r}(x)| > z_\alpha\right) \\ \leq \; & P\left(\cup_{r=1}^{J} |Z_{p_r}(\xi_1)| > z_\alpha\right) \\ + \; & P\left(\{\cap_{r=1}^{J} |Z_{p_r}(\xi_1)| \leq z_\alpha\} \cap \{\cup_{r=1}^{J} |Z_{p_r}(\xi_2)| > z_\alpha\}\right) \\ + \; & \sum_{j=1}^{G-2} P\left(\{\cap_{r=1}^{J} |Z_{p_r}(\xi_j)| \leq z_\alpha\} \cap \{\cap_{r=1}^{J} |Z_{p_r}(\xi_{j+1})| \leq z_\alpha\}\right. \\ & \left. \cap \{\cup_{r=1}^{J} |Z_{p_r}(\xi_{j+2})| > z_\alpha\}\right). \end{aligned} \tag{3.13}$$

Setting (3.13) equal to α will result in a less conservative z_α. In fact, such refinements could continue until one arrived at an exact value of z_α The following Table 3.2 illustrates how successive refinements of the approximation result in smaller (less conservative) z_α when bands are to be simultaneous over the mean response and the first two derivatives.

Table 3.2: Cutoff approximations for the mean response and first two derivatives

Refinement	1	2	3
n=100, G=25	3.305	3.290	3.287
n=100, G=50	3.441	3.408	3.395
n=500, G=25	3.303	3.287	3.285
n=500, G=50	3.438	3.405	3.392

For sample sizes of 100 and 500 equally spaced values from $\mathcal{X} = [-1, 1]$ and uniform grid sizes of 25 and 50, the table shows $z_{.05}$ approximations if the compound estimator is used with $J = 3$, 27 centering points, $\beta = 100$, and local regression pointwise estimators with nearest neighbor fractions of .3.

Thus for higher order derivatives we obtain the following theorem:

Theorem 3.2.2 *Assume that model (1.1) holds where the* ϵ_i *are independent and identically normally distributed. Let* $\widehat{\mu}(x)$ *be self-consistent and linear in the observed responses. Let* $\boldsymbol{G}$ *be a grid of points from* $\mathcal{X}$*. Let* $B_a(x)$ *be an upper bound on the absolute value of the bias of* $\widehat{\mu}^{(a)}(x)$ *and* $M_a \geq \sup_{x\in\mathcal{X}} |\mu^{(a)}(x) - \mu_I^{(a)}(x)|$ *for* $a \in \{p_1, ..., p_J\}$*. Let* $D_p(x)$ *be as defined above and let the subscript* I *denote linear interpolation between grid points. Then*

$$P(L_{p_1}(x) \leq \mu^{(p_1)}(x) \leq U_{p_1}(x), ..., L_{p_J}(x) \leq \mu^{(p_J)}(x) \leq U_{p_J}(x), \forall x \in \mathcal{X}) \geq 1 - \alpha,$$

where

$$U_a(x), L_a(x) := \widehat{\mu}_I^{(a)}(x) \pm (M_a + B_{aI}(x) + z_\alpha \sigma D_{aI}(x)).$$

The following theorem applies when confidence bands are placed over different subintervals of the covariate space for different derivatives under consideration:

Theorem 3.2.3 *Assume that model (1.1) holds where the* ϵ_i *are independent and identically normally distributed. Let* $\widehat{\mu}(x)$ *be self-consistent and linear in the observed responses. Let* $\boldsymbol{G}_a$ *be grids of points from* $E_a \subset \mathcal{X}$ *for* $a \in \{p_1, ..., p_J\}$*. Let* $B_a(x)$ *be an upper bound on the absolute value of the bias of* $\widehat{\mu}^{(a)}(x)$ *on* E_a *and* $M_a \geq \sup_{x\in E_a \subset\mathcal{X}} |\mu^{(a)}(x) - \mu_{I_a}^{(a)}(x)|$ *for* $a \in \{p_1, ..., p_J\}$*. Let* $D_p(x)$ *be as defined above and let the subscripts* I_a *denote linear interpolation between the grid points in* E_a*. Then*

$$P\left(\cap_{a=p_1}^{p_J}\{L_a(x) \leq \mu^{(a)}(x) \leq U_a(x), \forall x \in E_a\}\right) \geq 1 - \alpha,$$

where

$$U_a(x), L_a(x) := \widehat{\mu}_{I_a}^{(a)}(x) \pm (M_a + B_{aI_a}(x) + z_\alpha \sigma D_{aI_a}(x)).$$

Note that the definition of z_α will need to be modified to accommodate confidence bands placed over different subintervals of the covariate space for different derivatives under consideration. The easiest way to do this is to choose the grids so that they have the same cardinality. Then $\mathbf{G}_a = \{\xi_{a1}, ..., \xi_{aG}\}$ for $a \in \{p_1, ..., p_J\}$ and the definition of z_α can be obtained from (3.12) or (3.13) with the Z_a evaluated at $\xi_{aj}, j \in \{1, ..., G\}$ rather than at $\xi_j, j \in \{1, ..., G\}$.

If the grids not only cover different subintervals but also have different cardinality, then obtaining z_α is slightly more difficult. If we let the ordering of $\{p_1, ..., p_J\}$ be by increasing cardinality of the associated grid size (i.e. $\mathbf{G}_1 \leq \mathbf{G}_2 \leq ... \leq \mathbf{G}_J$, then

$$\begin{aligned} & P\left(\cup_{r=1}^{J} \max_{x \in \mathbf{G}_r} |Z_{p_r}(x)| > z_\alpha \right) \\ \leq \; & P\left(\cup_{r=1}^{J} |Z_{p_r}(\xi_{r1})| > z_\alpha \right) \qquad (3.14) \\ + \; & \sum_{s=1}^{J} \sum_{j=1_{s=1}+\mathbf{G}_{s-1}1_{s>1}}^{\mathbf{G}_s - 1} P\left(\{\cap_{r=s}^{m} |Z_{p_r}(\xi_j)| \leq z_\alpha\} \cap \{\cup_{r=s}^{J} |Z_{p_r}(\xi_{j+1})| > z_\alpha\} \right) \end{aligned}$$

and z_α can be obtained by setting the right side of (3.14) equal to α, where $\mathbf{G}_0 = 0$.

3.3 Accounting for the bias

How to account for the bias is a major factor in nonparametric regression confidence band construction. As mentioned in Section 1.7, several strategies have been proposed for dealing with the bias when constructing confidence bands for the mean response. The method of Eubank and Speckman (1993), which employs asymptotic-based corrections for the bias, is useful only in the special case of kernel regression. Knafl et. al (1985) assume that a bound on the bias is known and proceed from there. Hall and Titterington (1990) assume only a bound on one or more derivatives of the mean response and use such an assumption to bound the bias. One may or may not have a good rationale for making such an assumption. However, to obtain a useful (i.e. relatively tight) bound on the first derivative (in order to bound the bias of the estimator of the mean response) while at the same time seeking to place a confidence band around the first derivative seems to necessitate an iterative procedure. Therefore, we pursue a different approach to bias estimation.

Since

$$Bias_\mu[\widehat{\mu(x)}] = \sum_{i=1}^{n} l_i(x)\mu(x_i) - \mu(x),$$

a proposed estimate of the bias in estimating the mean response is (Loader 1999):

$$\widehat{Bias}(x) := \sum_{i=1}^{n} l_i(x)\widehat{\mu}(x_i) - \widehat{\mu}(x).$$

However, Loader cautions that one should not simply shift both upper and lower bands for the mean response by subtracting the estimate of the bias. Such efforts merely result in bands centered around undersmoothed estimates of the mean response. This strategy does not solve the bias problem. It reduces the bias problem but at the price of a more volatile mean response estimate.

We note that (3.3) can be extended to provide bias estimates for derivatives. Since

$$Bias_{\mu}[\widehat{\mu^{(p)}(x)}] = \sum_{i=1}^{n} l_i^{(p)}(x)\mu(x_i) - \mu^{(p)}(x),$$

a proposed estimate of the bias is:

$$\widehat{Bias_p(x)} := \sum_{i=1}^{n} l_i^{(p)}(x)\widehat{\mu}(x_i) - \widehat{\mu}^{(p)}(x). \tag{3.15}$$

As when constructing confidence bands for mean responses, we should not simply use (3.15) to shift both the upper and lower bands for $\mu^{(p)}(x)$. However, we can take the absolute value of (3.15) as an estimate of $B_p(x)$. For convenience this estimate is hereafter denoted $\widehat{B}_p(x)$. Then substituting $\widehat{B}_p(x)$ for $B_p(x)$ in the $U_p(x), L_p(x)$ formulas of the previous section's theorems addresses the bias issue without assuming that a bound for the bias is known. If there is concern about the disparity between $\widehat{B}_p(x)$ and $B_p(x)$, in that the former may underestimate the latter, then that disparity itself can be estimated and incorporated into the confidence bands. We will elaborate on that idea later in this section.

Supposing for now that the disparity between $\widehat{B}_p(x)$ and $B_p(x)$ is not worrisome, some conservatism in the confidence bands can be eliminated. This is because the use of $B_p(x)$ in previous theorems did not exploit any information about the sign of the bias. If the bias were known to be positive, then reducing the lower confidence band by $B_p(x)$ would be reasonable, but there would be no need to raise the upper confidence band by $B_p(x)$. Likewise, if the bias were known to be negative, then raising the upper confidence band by $B_p(x)$ would be reasonable, but there would be no need to reduce the lower confidence band by $B_p(x)$. Of course, if one is assuming rather than estimating an upper bound for $B_p(x)$, then typically one does not know the sign of the bias. On the other hand, if one is estimating an upper bound for $B_p(x)$ via the absolute value of (3.15), then one does have information about the sign of the bias.

Therefore we propose the following approach: At each grid point, determine if the estimate of the bias is positive or negative. If positive, then the lower band is shifted down by $\widehat{Bias_p(x)}$ and the upper band is left unchanged. If negative, then the upper band is shifted up by $|\widehat{Bias_p(x)}|$ and the lower band is left unchanged. We justify this approach by the following where we consider bands for the mean response and note that this justification is easily extended to bands for derivatives:

$$\begin{aligned} & P(\widehat{\mu}(x) - z_\alpha \sigma D(x) - Bias_\mu[\widehat{\mu}(x)] I_{Bias_\mu[\widehat{\mu}(x)]>0} \leq \mu(x) \leq \widehat{\mu}(x) + z_\alpha \sigma D(x) \\ & -Bias_\mu[\widehat{\mu}(x)] I_{Bias_\mu[\widehat{\mu}(x)]<0}, x \in \mathbf{G}) \\ \geq \; & P(\widehat{\mu}(x) - z_\alpha \sigma D(x) - Bias_\mu[\widehat{\mu}(x)] \leq \mu(x) \leq \widehat{\mu}(x) + z_\alpha \sigma D(x) \\ & -Bias_\mu[\widehat{\mu}(x)], x \in \mathbf{G}) \\ \geq \; & P\left(\left| \sum_{i=1}^{n} l_i(x) \epsilon_i \right| \leq z_\alpha \sigma D(x), x \in \mathbf{G} \right) \\ \geq \; & 1 - \alpha. \end{aligned}$$

We now return to the question of handling the disparity between $\widehat{B}_p(x)$ and $B_p(x)$. Since the difference between $\widehat{B}_p(x)$ and $B_p(x)$ depends in large part on the difference between $\sum_{i=1}^n l_i^{(p)}(x)\widehat{\mu}(x_i)$ and $\sum_{i=1}^n l_i^{(p)}(x)\mu(x_i)$, we note that for $a_p(x) > 0$,

$$
\begin{aligned}
& P\left(\left|\sum_{i=1}^n l_i^{(p)}(x)\widehat{\mu}(x_i) - \sum_{i=1}^n l_i^{(p)}(x)\mu(x_i)\right| \geq a_p(x)\right) \\
\leq\; & (4/9)a_p(x)^{-2}\mathbb{E}\left[\left(\sum_{i=1}^n l_i^{(p)}(x)\widehat{\mu}(x_i) - \sum_{i=1}^n l_i^{(p)}(x)\mu(x_i)\right)^2\right] \qquad (3.16) \\
=\; & (4/9)a_p(x)^{-2}\mathbb{E}\left[\left(\sum_{i=1}^n l_i^{(p)}(x)\sum_{j=1}^n l_j(x_i)Y_j - \sum_{i=1}^n l_i^{(p)}(x)\mu(x_i)\right)^2\right] \\
=\; & (4/9)a_p(x)^{-2}\mathbb{E}\left[\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n\sum_{m=1}^n l_i^{(p)}(x)l_j(x_i)l_k^{(p)}(x)l_m(x_k)Y_jY_m\right. \\
& \left.-2\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n l_i^{(p)}(x)l_j(x_i)l_k^{(p)}(x)Y_j\mu(x_k) + \sum_{i=1}^n\sum_{j=1}^n l_i^{(p)}(x)l_j^{(p)}(x)\mu(x_i)\mu(x_j)\right] \\
=\; & (4/9)a_p(x)^{-2}\left[\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n\sum_{m=1}^n l_i^{(p)}(x)l_j(x_i)l_k^{(p)}(x)l_m(x_k)\{\mu(x_j)\mu(x_m) + \sigma^2 1_{j=m}\}\right. \\
& \left.-2\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n l_i^{(p)}(x)l_j(x_i)l_k^{(p)}(x)\mu(x_j)\mu(x_k) + \sum_{i=1}^n\sum_{j=1}^n l_i^{(p)}(x)l_j^{(p)}(x)\mu(x_i)\mu(x_j)\right],
\end{aligned}
$$

where line (3.16) follows from the Vysochanskii-Petunin Inequality (Vysochanskii and Petunin 1980). By substituting $\widehat{\mu}(x)$ for $\mu(x)$, this latter quantity can be set equal to $P_0 \in (0, 1)$ and the equation can then be solved for $a_p(x)$. The value of $a_p(x)$ will need to be evaluated at each of the grid points and the upper band is then shifted up by $a_p(x)$ while the lower band is shifted down by $a_p(x)$. In practice the conservativeness of (3.16) is large enough that $P_0 \leq \alpha$ preserves the nominal confidence level.

The rate at which $a_p(x)$ goes to zero will depend on the nonparametric regression estimator being employed and the order of the derivative p. The following proposition demonstrates this for kernel regression with a compactly supported kernel.

Proposition 3.3.1 *Assume the conditions of Theorem 3.2.3 hold. Assume that the nonparametric regression estimator is the kernel estimator*

$$\widehat{\mu}(x) = (nh)^{-1}\sum_{i=1}^n K((x - x_i)/h)Y_i$$

with the kernel function supported on $[-1,1]$. *Further suppose that*

$$\sup_x |\widehat{\mu}^{(p)}(x) - \mu^{(p)}(x)| = O_p\left(\left(\frac{n}{\log n}\right)^{(p-J-1)/(2J+3)}\right)$$

with bandwidth $h \propto n^{-1/(2J+3)}$. *Then* $a_p(x) = O_p((\frac{n}{\log n})^{(p-J-1)/(2J+3)})$.

Proof. Let K_p denote the maximum absolute value of the kernel function's p^{th} derivative. Then

$$\begin{aligned}
&\left|\sum_{i=1}^{n} l_i^{(p)}(x)\{\widehat{\mu}(x_i) - \mu(x_i)\}\right| \\
\leq\ & \sum_{i=1}^{n} |l_i^{(p)}(x)||\{\widehat{\mu}(x_i) - \mu(x_i)\}| \\
\leq\ & \sum_{i=1}^{n} |l_i^{(p)}(x)| O_p\left(\left(\frac{n}{\log n}\right)^{-(J+1)/(2J+3)}\right) \\
=\ & \sum_{i:|x_i-x|\leq h} |l_i^{(p)}(x)| O_p\left(\left(\frac{n}{\log n}\right)^{-(J+1)/(2J+3)}\right) \qquad (3.17) \\
\leq\ & \sum_{i:|x_i-x|\leq h} K_p n^{-1} h^{-(p+1)} O_p\left(\left(\frac{n}{\log n}\right)^{-(J+1)/(2J+3)}\right) \\
\leq\ & 2K_p h^{-p} O_p\left(\left(\frac{n}{\log n}\right)^{-(J+1)/(2J+3)}\right) \\
\leq\ & 2K_p O_p\left(\left(\frac{n}{\log n}\right)^{(p-J-1)/(2J+3)}\right)
\end{aligned}$$

where (3.17) follows because the kernel is supported on $[-1,1]$. Then noting that (3.16) is set equal to a constant implies the desired result. ■

We can then investigate how well the estimated bias performs through simulation. We generated 100 data sets of size 50 from model (1.1) using the mean response function $\mu(x) := \sin(2\pi x) + \cos(2\pi x) + \log(4/3 + x)$ with $x_1, ..., x_n$ equispaced on $\mathcal{X} = [-1,1]$ and normally distributed error terms with variance $\sigma^2 = .01$. We sought to recover the mean response and the first derivative using compound estimation with filtration and extrapolation using $J = 4$, 27 centering points, $\beta = 15$ (150 during filtration and extrapolation), $\kappa = 1.1$, and local regression pointwise estimators obtained with nearest neighbor fraction .30 (.15 during filtration and extrapolation). We then placed simultaneous 95% confidence intervals around the estimates of the mean response and the first derivative at each of 25 grid points equispaced on $[-1, 1]$ using the approach described in the previous paragraphs. We use the language 'simultaneous confidence intervals' in this instance to emphasize that here the intervals are constructed only at the grid points and not over the entire covariate space. The simultaneous confidence intervals for the mean response and first derivative contained the true values ar all of the grid points 100% of the time. The simultaneous confidence intervals for the mean response and the first two derivatives contained the true values at all of the grid points 97% of the time.

If one wishes to *guarantee* that, asymptotically, the actual coverage probability does not fall short of the nominal coverage probability, then one may proceed as follows. Let r_p be the convergence rate of the nonparametric regression technique being employed (i.e. $n^{r_p}(\widehat{\mu}^{(p)}(x) - \mu^{(p)}(x)) = O_p(1)$) and define

$$\begin{aligned}\widehat{\beta}_{pL}(x) &:= \max\left\{\left|\widehat{Bias}_p(x)\right| 1_{\widehat{Bias}_p(x)>0} + a_p(x), g(n)\right\}\\ \widehat{\beta}_{pU}(x) &:= \max\left\{\left|\widehat{Bias}_p(x)\right| 1_{\widehat{Bias}_p(x)<0} + a_p(x), g(n)\right\},\end{aligned}$$

where $g(n) \to 0$ and $n^{r_p} g(n) \to \infty$. For a practical implementation we could take $g(n) := Cn^{-r_p} \log\log n$ where C is a constant chosen so that $|\widehat{Bias}_p(x)| + a_p(x) > g(n)$ with high probability for small n.

We can then make the following modification of Theorem 3.2.2 (and we can analogously modify Theorem 3.2.3):

Theorem 3.3.1 *Assume that model (1.1) holds where the* ϵ_i *are independent and identically normally distributed. Let* $\widehat{\mu}(x)$ *be self-consistent and linear in the observed responses. Let* $\boldsymbol{G}$ *be a grid of points from* $\mathcal{X}$*. Let* $\widehat{\beta}_{pL}(x)$ *and* $\widehat{\beta}_{pU}(x)$ *be as defined above and* $M_p \geq \sup_{x\in\mathcal{X}} |\mu^{(p)}(x) - \mu_I^{(p)}(x)|$ *for* $p \in \{p_1, ..., p_J\}$*. Let* $D_p(x)$ *be as defined above and let the subscript* I *denote linear interpolation between grid points. Then*

$$\liminf_{n\to\infty} P(L_{p_1}(x) \leq \mu^{(p_1)}(x) \leq U_{p_1}(x), ..., L_{p_J}(x) \leq \mu^{(p_J)}(x) \leq U_{p_J}(x), \forall x \in \mathcal{X}) \geq 1-\alpha,$$

where

$$L_p(x) := \widehat{\mu}_I^{(p)}(x) - (M_p + \widehat{\beta}_{pL}(x)_I + z_\alpha \sigma D_{pI}(x))$$

and

$$U_p(x) := \widehat{\mu}_I^{(p)}(x) + (M_p + \widehat{\beta}_{pU}(x)_I + z_\alpha \sigma D_{pI}(x)).$$

Proof. Let

$$\begin{aligned} A &:= \Big\{\widehat{\mu}_I^{(a)}(x) - z_\alpha \sigma D_{aI}(x) - M_a - \Big|Bias_\mu[\widehat{\mu^{(a)}(x)}]\Big|_I \le \mu^{(a)}(x) \\ &\quad \le \widehat{\mu}_I^{(a)}(x) + z_\alpha \sigma D_{aI}(x) + M_a + \Big|Bias_\mu[\widehat{\mu^{(a)}(x)}]\Big|_I, \forall x \in \mathcal{X}, \\ &\quad \forall a \in \{p_1, ..., p_J\}\Big\}, \\ B &:= \Big\{\widehat{\mu}_I^{(a)}(x) - z_\alpha \sigma D_{aI}(x) - M_a - \widehat{\beta}_{aL}(x)_I \le \mu^{(a)}(x) \\ &\quad \le \widehat{\mu}_I^{(a)}(x) + z_\alpha \sigma D_{aI}(x) + M_a + \widehat{\beta}_{aU}(x)_I, \forall x \in \mathcal{X}, \forall a \in \{p_1, ..., p_J\}\Big\}, \end{aligned}$$

and

$$\begin{aligned} C &:= \Big\{\widehat{\beta}_{aU}(x)_I \ge \Big|Bias_\mu[\widehat{\mu^{(a)}(x)}]\Big|_I \text{ and } \widehat{\beta}_{aL}(x)_I \ge \Big|Bias_\mu[\widehat{\mu^{(a)}(x)}]\Big|_I, \forall x \in \mathcal{X}, \\ &\quad \forall a \in \{p_1, ..., p_J\}\Big\}. \end{aligned}$$

Now note that $(A \cap C) \subset B$ and so $P(B) \ge P(A \cap C) = P(A) - P(A \cap C^c)$. By Theorem 3.2.2, $P(A) \ge 1 - \alpha$ and since $\widehat{\mu}^{(a)}(x)$ converges to $\mu^{(a)}(x)$ at the rate r_a, $P(C^c) \to 0$ which implies the desired result. ■

3.4 Interpolating between the grid points

To obtain confidence bands that are simultaneous over the entire covariate space rather than simply over a finite grid of points we must interpolate between the grid points. As mentioned previously, this can be accomplished if we assume that upper bounds on $\sup_{x \in E_p} |\mu^{(p)}(x) - \mu_I^{(p)}(x)|$, call them M_p, for $p \in \{p_1, ..., p_J\}$, where the subscript I denotes linear interpolation between the grid points, are available.

If such bounds are unavailable, then we propose using the following estimates:

$$\widehat{M}_p := \begin{cases} \max_{x_i \in E_p} |\widehat{\mu}^{(p)}(x_i) - \widehat{\mu}_I^{(p)}(x_i)| & : \quad n < N_{m_0} \\ \max_{\gamma_i \in E_p} |\widehat{\mu}^{(p)}(\gamma_i) - \widehat{\mu}_I^{(p)}(\gamma_i)| & : \quad N\left(m, \frac{1}{m}\right) \le N_m \le n < N_{m+1} \end{cases}$$

for $m \geq m_0$, where m_0 is an arbitrary positive integer, the γ_i constitute a grid of m points from $E_p \subset \mathcal{X}$ that become dense as $m \to \infty$ and $\{N_m : m \in \{1, 2, \ldots\}\}$ is a strictly increasing sequence of positive integers.

The large-sample justification for using these estimates is shown in Theorem 3.4.1 whose proof relies on the following lemmas:

Lemma 3.4.1 *If* $\widehat{\mu}^{(p)}(x) \to^P \mu^{(p)}(x), \forall x \in E_p \subset \mathcal{X}$, *then* $|\widehat{\mu}^{(p)}(x) - \widehat{\mu}_I^{(p)}(x)| \to^P |\mu^{(p)}(x) - \mu_I^{(p)}(x)|, \forall x \in E_p$.

Proof. Let $\epsilon > 0$ and $t \in E_p \subset \mathcal{X}$. If $t \in \mathbf{G}_p$, then there is nothing to show. Otherwise, let a and b be the grid points from $\mathbf{G}_p$ immediately below and immediately above t, respectively. Then

$$\begin{aligned}
& P\left(\left||\widehat{\mu}^{(p)}(t) - \widehat{\mu}_I^{(p)}(t)| - |\mu^{(p)}(t) - \mu_I^{(p)}(t)|\right| > \epsilon\right) \\
\leq\ & P\left(|\widehat{\mu}^{(p)}(t) - \widehat{\mu}_I^{(p)}(t) - \mu^{(p)}(t) + \mu_I^{(p)}(t)| > \epsilon\right) \\
\leq\ & P\left(\left||\widehat{\mu}^{(p)}(t) - \mu^{(p)}(t)| + |\widehat{\mu}_I^{(p)}(t) - \mu_I^{(p)}(t)|\right| > \epsilon\right) \\
\leq\ & P\left(|\widehat{\mu}^{(p)}(t) - \mu^{(p)}(t)| > \epsilon/2\right) + P\left(|\widehat{\mu}_I^{(p)}(t) - \mu_I^{(p)}(t)| > \epsilon/2\right) \\
=\ & P\left(|\widehat{\mu}^{(p)}(t) - \mu^{(p)}(t)| > \epsilon/2\right) \\
+\ & P\left(\left|\widehat{\mu}^{(p)}(a) + \frac{t-a}{b-a}\left[\widehat{\mu}^{(p)}(b) - \widehat{\mu}^{(p)}(a)\right] - \mu^{(p)}(a) - \frac{t-a}{b-a}\left[\mu^{(p)}(b) - \mu^{(p)}(a)\right]\right|\right. \\
& \left. > \epsilon/2\right) \\
\leq\ & P\left(|\widehat{\mu}^{(p)}(t) - \mu^{(p)}(t)| > \epsilon/2\right) \\
+\ & P\left(|\widehat{\mu}^{(p)}(a) - \mu^{(p)}(a)| + \frac{t-a}{b-a}\left|\widehat{\mu}^{(p)}(b) - \mu^{(p)}(b)\right| + \frac{t-a}{b-a}\left|\widehat{\mu}^{(p)}(a) - \mu^{(p)}(a)\right| > \epsilon\right) \\
\leq\ & P\left(|\widehat{\mu}^{(p)}(t) - \mu^{(p)}(t)| > \epsilon/2\right) \\
+\ & 2P\left(|\widehat{\mu}^{(p)}(a) - \mu^{(p)}(a)| > \epsilon/6\right) + P\left(|\widehat{\mu}^{(p)}(b) - \mu^{(p)}(b)| > \epsilon/6\right) \\
\to\ & 0,
\end{aligned}$$

which implies the desired result. ■

Lemma 3.4.2 *If* $\widehat{f}(\gamma_i) \to^P f(\gamma_i), \forall i \in \{1, ..., m\}$, *then* $\max_{i \in \{1,...,m\}} \widehat{f}(\gamma_i) \to^P \max_{i \in \{1,...,m\}} f(\gamma_i)$.

Proof. Let $\epsilon > 0$ and $i^* := \text{argmax}_{i\in\{1,...,m\}} f(\gamma_i)$. Also define $\epsilon_j := \frac{f(\gamma_{i^*})-f(\gamma_j)}{2}, \forall j \in \{1, ..., m\}$. Then

$$\begin{aligned}
& P\left(\left|\max_{i\in\{1,\ldots,m\}} \widehat{f}(\gamma_i) - f(\gamma_{i^*})\right| > \epsilon\right) \\
\leq\ & \sum_{j=1}^{m} P\left(\left|\widehat{f}(\gamma_j) - f(\gamma_{i^*})\right| > \epsilon, \widehat{f}(\gamma_j) = \max_{i\in\{1,\ldots,m\}} \widehat{f}(\gamma_i)\right) \\
=\ & \sum_{j=1}^{m} \left\{P\left(\left|\widehat{f}(\gamma_j) - f(\gamma_{i^*})\right| > \epsilon, \widehat{f}(\gamma_j) = \max_{i\in\{1,\ldots,m\}} \widehat{f}(\gamma_i), f(\gamma_j) = f(\gamma_{i^*})\right)\right. \\
& \left. + P\left(\left|\widehat{f}(\gamma_j) - f(\gamma_{i^*})\right| > \epsilon, \widehat{f}(\gamma_j) = \max_{i\in\{1,\ldots,m\}} \widehat{f}(\gamma_i), f(\gamma_j) \neq f(\gamma_{i^*})\right)\right\} \\
\leq\ & \sum_{j=1}^{m} \left\{P\left(\left|\widehat{f}(\gamma_j) - f(\gamma_j)\right| > \epsilon\right) + P\left(\widehat{f}(\gamma_j) = \max_{i\in\{1,\ldots,m\}} \widehat{f}(\gamma_i), f(\gamma_j) \neq f(\gamma_{i^*})\right)\right\} \\
\leq\ & \sum_{j=1}^{m} \left\{P\left(\left|\widehat{f}(\gamma_j) - f(\gamma_j)\right| > \epsilon\right) + P\left(\widehat{f}(\gamma_j) \geq \widehat{f}(\gamma_{i^*}), f(\gamma_j) < f(\gamma_{i^*})\right)\right\} \\
\leq\ & \sum_{j=1}^{m} \left\{P\left(\left|\widehat{f}(\gamma_j) - f(\gamma_j)\right| > \epsilon\right)\right. \\
& \left. + \sum_{j:f(\gamma_j)<f(\gamma_{i^*})} P\left(\left|\widehat{f}(\gamma_j) - f(\gamma_j) + f(\gamma_{i^*}) - \widehat{f}(\gamma_{i^*})\right| > \epsilon_j\right)\right\} \\
\leq\ & \sum_{j=1}^{m} \left\{P\left(\left|\widehat{f}(\gamma_j) - f(\gamma_j)\right| > \epsilon\right)\right. \\
& \left. + \sum_{j:f(\gamma_j)<f(\gamma_{i^*})} P\left(\left|\widehat{f}(\gamma_j) - f(\gamma_j)\right| + \left|f(\gamma_{i^*}) - \widehat{f}(\gamma_{i^*})\right| > \epsilon_j\right)\right\} \\
\leq\ & \sum_{j=1}^{m} \left\{P\left(\left|\widehat{f}(\gamma_j) - f(\gamma_j)\right| > \epsilon\right)\right. \\
& + \sum_{j:f(\gamma_j)<f(\gamma_{i^*})} P\left(\left|\widehat{f}(\gamma_j) - f(\gamma_j)\right| > \epsilon_j/2\right) \\
& \left. + P\left(\left|f(\gamma_{i^*}) - \widehat{f}(\gamma_{i^*})\right| > \epsilon_j/2\right)\right\} \\
\to\ & 0,
\end{aligned}$$

which implies the desired result. ■

Lemma 3.4.3 *If $f(x)$ is continuous on $E_p \subset \mathcal{X}$, where E_p is compact and $\{\gamma_i\}_{i=1}^m$ constitute a grid of points from E_p that become dense as $m \to \infty$, then* $\max_{i\in\{1,\ldots,m\}} f(\gamma_i) \to \sup_{x\in E_p} f(x)$ *as $m \to \infty$.*

Proof. Let $f(t_0) = \sup_{x\in E_p} f(x)$. Now, $\forall \delta > 0, \exists M \in \mathbb{N}$ such that $\forall m \geq M, \exists i^* \in \{1, ..., m\}$ such that $|\gamma_{i^*} - t_0| < \delta$. Then by the continuity of f, $\forall \epsilon > 0, \exists M$ such that $\forall m \geq M, \exists i^* \in \{1, ..., m\}$ such that $\left|f(\gamma_{i^*}) - \sup_{x\in E_p} f(x)\right| < \epsilon$. This implies that $\forall \epsilon > 0, \exists M$ such that $\forall m \geq M$, $\left|\max_{i\in\{1,\ldots,m\}} f(\gamma_i) - \sup_{x\in E_p} f(x)\right| < \epsilon$ since $f(\gamma_{i^*}) \leq \max_{i\in\{1,\ldots,m\}} f(\gamma_i) \leq \sup_{x\in E_p} f(x)$, which implies the desired result. ■

Theorem 3.4.1 *If $\widehat{\mu}^{(p)}(x) \to^P \mu^{(p)}(x), \forall x \in E_p$ and $\mu^{(p)}(x)$ is continuous on E_p, where E_p is compact, then $\widehat{M}_p \to^P \sup_{x\in E_p} |\mu^{(p)}(x) - \mu_I^{(p)}(x)|$.*

Proof. Let $\widehat{f}(x) := |\widehat{\mu}^{(p)}(x) - \widehat{\mu}_I^{(p)}(x)|$ and $f(x) := |\mu^{(p)}(x) - \mu_I^{(p)}(x)|$. Define $\widehat{M}_p(m) := \max_{i\in\{1,\ldots,m\}} \widehat{f}(\gamma_i)$ and $M_p(m) := \max_{i\in\{1,\ldots,m\}} f(\gamma_i)$. By Lemmas 3.4.1 and 3.4.2, $\forall m \in \mathbb{N}, \widehat{M}_p(m) \to^P M_p(m)$. So $\forall m \in \mathbb{N}, \exists N_0(m)$ such that $\forall n \geq N_0(m)$,

$$P\left(\left|\widehat{M}_p(m) - M_p(m)\right| \geq 1/m\right) \leq 1/m.$$

Now for $m > 1$ define $N_m := \max\left(N_0(m), N_{m-1} + 1\right)$. Define

$$\widetilde{M}_p := \begin{cases} M_p(n) & : \quad n < N_{m_0} \\ M_p(m) & : \quad N_m \leq n < N_{m+1} \end{cases}$$

for $m \geq m_0$, where m_0 is an arbitrary positive integer.

Let $\epsilon > 0$. Then $\exists \widetilde{m}_1 \in \mathbb{N}$ such that $1/\widetilde{m}_1 \leq \epsilon/2$. Also by Lemma 3.4.3, $\exists \widetilde{m}_2$ such that $\forall m \geq \widetilde{m}_2$,

$$P\left(\left|\max_{i\in\{1,\ldots,m\}} f(\gamma_i) - \sup_{x\in E_p} f(x)\right| \geq \epsilon/2\right) \leq \epsilon.$$

Let $m_\epsilon = \max\{\widetilde{m}_1, \widetilde{m}_2\}$. For $N_{m_\epsilon+1} \leq n$, $\left|\widehat{M}_p - \widetilde{M}_p\right| = \left|\widehat{M}_p(m_n) - M_p(m_n)\right|$ where

$N_{m_n} \leq n < N_{m_\epsilon+1}$ and so

$$\begin{aligned}
& P\left(\left|\widehat{M}_p - \sup_{x\in E_p} f(x)\right| \geq \epsilon\right) \\
\leq\ & P\left(\left|\widehat{M}_p - \widetilde{M}_p\right| + \left|\widetilde{M}_p - \sup_{x\in E_p} f(x)\right| \geq \epsilon\right) \\
\leq\ & P\left(\left|\widehat{M}_p(m_n) - M_p(m_n)\right| \geq \epsilon/2\right) + P\left(\left|M_p(m_n) - \sup_{x\in E_p} f(x)\right| \geq \epsilon/2\right) \\
\leq\ & P\left(\left|\widehat{M}_p(m_n) - M_p(m_n)\right| \geq \frac{1}{m_\epsilon}\right) + \epsilon/2 \\
\leq\ & P\left(\left|\widehat{M}_p(m_n) - M_p(m_n)\right| \geq \frac{1}{m_n}\right) + \epsilon/2 \\
\leq\ & \frac{1}{m_n} + \epsilon/2 \\
\leq\ & \frac{1}{m_\epsilon} + \epsilon/2 \\
\leq\ & \epsilon,
\end{aligned}$$

which implies the desired result. ■

The effectiveness of these estimates can be investigated through simulation. If the mean response function is $\mu(x) := \sin(2\pi x) + \cos(2\pi x) + \log(4/3 + x)$ on $\mathcal{X} = [-1, 1]$ and there are 50 equispaced grid points, then the true value of M_0 is .01257. For the 500 data sets generated above with a variance of $\sigma^2 = .01$, the median $\widehat{M}$ is .01457 and $\widehat{M}$ is greater than M (that is, it is conservative) 86.2% of the time. The true value of M_1 is .07358. The median $\widehat{M}$ is .14075 and $\widehat{M}_1$ is greater than M_1 (that is, it is conservative) 100% of the time.

Finally, we demonstrate in the following proposition that interpolation error can be made to vanish asymptotically.

Proposition 3.4.1 *If $\mu^{(p)}(x)$ is continuous on $E_p \subset \mathcal{X}$, where E_p is a compact interval, and if $\boldsymbol{G}_p$ becomes dense in E_p as $|\boldsymbol{G}_p| \to \infty$, then $\sup_{x\in E_p} \left|\mu^{(p)}(x) - \mu_I^{(p)}(x)\right| \to 0$ as $|\boldsymbol{G}_p| \to \infty$.*

Proof. Let $\epsilon > 0$ and $x \in \text{int}(E_p)$. There exists $\delta_0 > 0$ depending on ϵ but not x such that if $y \in E_p$ and $|x - y| < \delta_0$, then $\left|\mu^{(p)}(x) - \mu^{(p)}(y)\right| < \epsilon/2$ by the continuity of $\mu^{(p)}$ and the compactness of E_p. Also, since $\mathbf{G}_p$ becomes dense in E_p as $|\mathbf{G}_p| \to \infty$, there exists G_0 such that $\forall |\mathbf{G}_p| \geq G_0$, there exist $a, b \in \mathbf{G}_p$ such that $a \leq x \leq b, |b-a| < \delta_0$. So there exists G_0 such that $\forall |\mathbf{G}_p| \geq G_0$, there exist $a, b \in \mathbf{G}_p$ such that $a \leq x \leq b, |\mu^{(p)}(a) - \mu^{(p)}(x)| < \epsilon/2$ and $|\mu^{(p)}(b) - \mu^{(p)}(a)| < \epsilon/2$. If $a = x = b$, then $|\mu^{(p)}(x) - \mu_I^{(p)}(x)| = 0$. Otherwise,

$$\begin{aligned}
& \left|\mu^{(p)}(x) - \mu_I^{(p)}(x)\right| \\
= & \left|\mu^{(p)}(x) - \frac{x-a}{b-a}\left[\mu^{(p)}(b) - \mu^{(p)}(a)\right] - \mu^{(p)}(a)\right| \\
\leq & \left|\mu^{(p)}(x) - \mu^{(p)}(a)\right| + \left|\mu^{(p)}(b) - \mu^{(p)}(a)\right| \\
\leq & \epsilon/2 + \epsilon/2 \\
= & \epsilon.
\end{aligned}$$

Now since $\left|\mu^{(p)}(x) - \mu_I^{(p)}(x)\right| \leq \epsilon, \forall x \in \operatorname{int}(E_p)$, then $\sup_{x \in E_p} \left|\mu^{(p)}(x) - \mu_I^{(p)}(x)\right| \leq \epsilon$, which implies the desired result. ■

3.5 Modifications for unknown variance

Theorems 3.2.1, 3.2.2, and 3.2.3 assume that σ^2 is known. We now consider the situation in which σ^2 is unknown. In this case we estimate σ^2 by (1.14), which is a consistent estimator of σ^2.

Assume that

$$\mathbf{Z} := \left[\begin{array}{ccccccc} Z_{p_1}(\xi_1) & \cdots & Z_{p_1}(\xi_G) & \cdots & Z_{p_J}(\xi_1) & \cdots & Z_{p_J}(\xi_G) \end{array}\right]^t \to^L \mathbf{Z}^*$$

where $\mathbf{Z}^* \sim MVN(\mathbf{0}, \Sigma^*)$ for some symmetric positive definite matrix Σ^*.

Then consider that by Slutsky's Theorem,

$$\widehat{Z} := \left[\begin{array}{ccccccc} \widehat{Z}_{p_1}(\xi_1) & \cdots & \widehat{Z}_{p_1}(\xi_G) & \cdots & \widehat{Z}_{p_J}(\xi_1) & \cdots & \widehat{Z}_{p_J}(\xi_G) \end{array}\right]^t \to^L \mathbf{Z}^*,$$

where $\widehat{Z}_a(\xi_j) := (\sigma/\widehat{\sigma})\, Z_a(\xi_j)$.

Then for any Borel set $A \in \mathbb{R}^{G*m}$, $P(\mathbf{Z} \in A) \to P(\mathbf{Z}^* \in A)$ and $P(\widehat{Z} \in A) \to P(\mathbf{Z}^* \in A)$. Thus $P(\mathbf{Z} \in A) - P(\widehat{Z} \in A) \to 0$ justifying $P(\widehat{Z} \in A)$ as an approximation to $P(\mathbf{Z} \in A)$.

For small samples, we propose using the Sattherthwaite degrees of freedom estimate, $n - \sum_{m=1}^{n} l_m(x_m)$, and employing a multivariate t-distribution in place of a multivariate normal distribution in (3.11) with Σ as the scale matrix of the multivariate t-distribution.

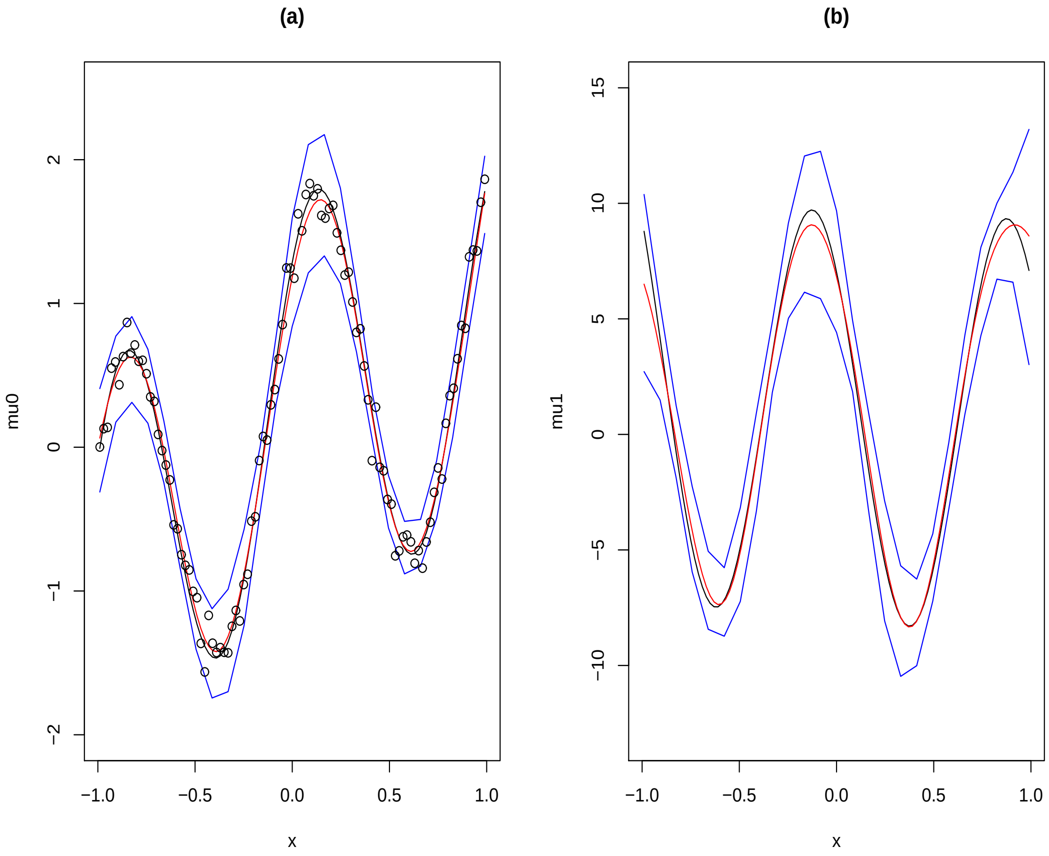
3.6 Simulation studies

We investigated this methodology which we have described for constructing simultaneous confidence bands for a mean response and its derivatives through simulation. To do this we generated 1000 data sets from (1.1) with the true underlying mean response set to be $\mu(x) := \sin(2\pi x) + \cos(2\pi x) + \log(4/3 + x)$ with $x_1, ..., x_n$ equispaced on $\mathcal{X} = [-1, 1]$ and $\sigma = .1$. We did this for $n \in \{50, 100\}$ and $\alpha \in \{.05, .20\}$. We estimated the mean response and its derivatives using compound estimation with filtration and extrapolation with $J = 4$, 27 centering points, $\beta = 15$ (150 during filtration and extrapolation), local regression pointwise estimates using nearest neighbor fraction .30 (.15 during filtration and extrapolation), and $\kappa = 0.1$. We then placed simultaneous confidence bands around the estimates of the mean response and first derivative and simultaneous confidence bands around the estimates of the mean response and first two derivatives, in each case using $G = 25$.

The results of this simulation study are recorded below in Table 3.3. In each case, the confidence bands are conservative, achieving at least the nominal coverage level. Note the bands constructed using $\alpha = .20$ for the mean response and first two derivatives successfully capture the true mean response and first two derivatives at all values of the covariate only 86.3 and 84.0 percent of the time for sample sizes of 50 and 100, respectively. This is reassuring because it indicates that while the bands are indeed conservative, they are not unreasonably so.

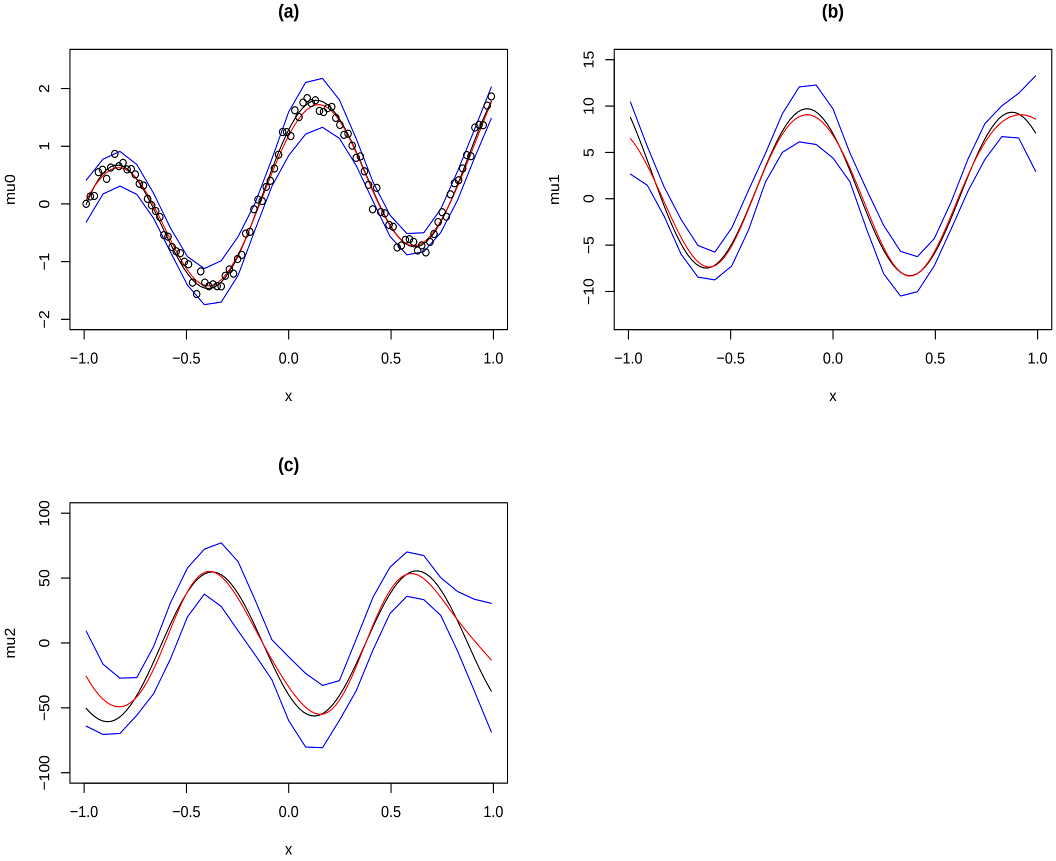
Below we include plots of some simulated data sets and their accompanying simultaneous confidence bands.

Figure 3.1: Simultaneous confidence bands for the mean response and first derivative



The black curves indicate the true mean response and first derivative in panels (a) and (b), respectively. The simulated data for this sample of size 100 is also displayed in panel (a). The red curves represent the estimated mean response and first derivative and the blue curves represent the confidence bands with $\alpha = .05$. In this case, the confidence bands successfully contain the true mean response and first derivative at all values of the covariate.

Figure 3.2: Simultaneous confidence bands for the mean response and two derivatives



The black curves indicate the true mean response, first derivative, and second derivative in panels (a), (b), and (c), respectively. The simulated data for this sample of size 100 is also displayed in panel (a). The red curves represent the estimated mean response and derivatives and the blue curves represent the confidence bands with $\alpha = .05$. In this case, the confidence bands successfully contain the true mean response and first two derivatives at all values of the covariate.

Table 3.3: Simulation results for simultaneous confidence bands

	α=.05		α=.20	
	0,1	0,1,2	0,1	0,1,2
n=50	100%	99.3%	99.6%	86.3%
n=100	100%	96.7%	98.0%	84.0%

The columns labeled $0, 1$ indicate that simultaneous confidence bands were constructed for the mean response and its first derivative. The columns labeled $0, 1, 2$ indicate that simultaneous confidence bands were constructed for the mean response and its first two derivatives. The entries represent the percentages of simultaneous confidence bands that contained the true mean response and the true derivative(s) at each value of the covariate based on the 1000 simulated data sets.

3.7 Ethanol example

In this section we apply our methodology to a data set from Brinkman (1981) involving exhaust emissions. This data set has been examined elsewhere using nonparametric regression techniques (Cleveland 1979, Loader 1999). Loader discusses how to estimate the mean response and first two derivatives of the concentration of certain pollutants (NOx) with respect to the equivalence ratio (E) using local regression. Loader does not, however, discuss how to place confidence bands around the estimates.

In the graphs below we estimate the mean response and first two derivatives using compound estimation with filtration and extrapolation. In Figure 3.3 we obtain simultaneous 95% confidence bands for the mean response and the first derivative. In Figure 3.4 we obtain simultaneous 95% confidence bands for the mean response and the first two derivatives.

An initial glance at these figures may lead to the perception that the bands seem wide. This perception is due to a couple of factors. The first is that this data has a relatively low signal-to-noise ratio. The estimated mean response has a range of only 3.4 while the estimated standard deviation is .9. The wide bands reflect the uncertainty inherent in this relatively large standard deviation. The bands may also seem wide because intuition about confidence intervals around the estimate of the mean response at a given point does not translate easily to confidence bands that are simultaneous over the mean response and one or more derivatives.

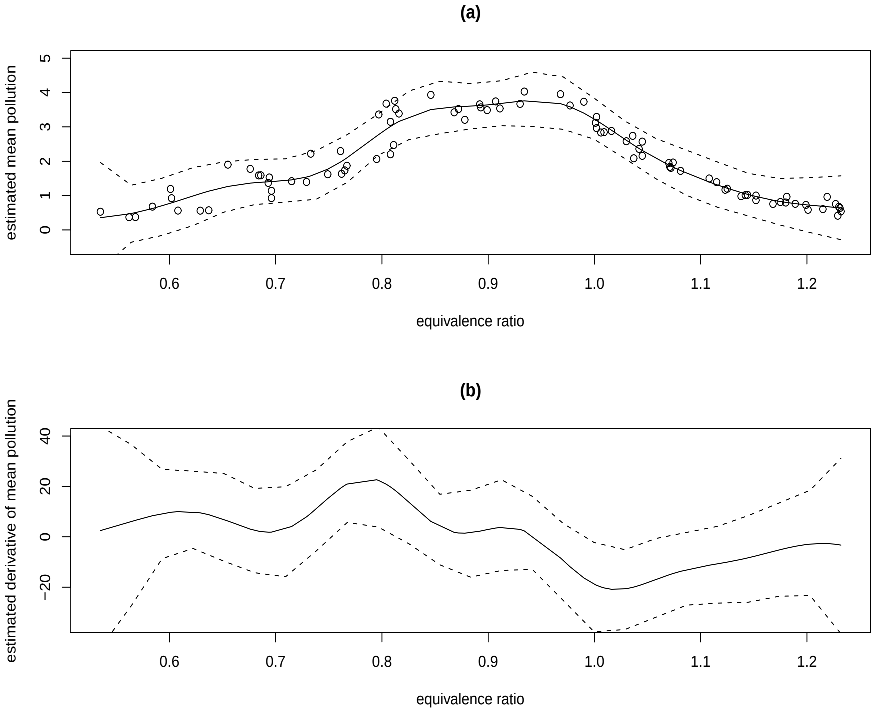
Despite the low signal-to-noise ratio, the confidence bands do allow us to determine ranges of E over which pollution is clearly increasing and decreasing. Pollution is clearly increasing as E ranges from .755 to .812 and is clearly decreasing as E ranges from .99 to 1.06 based on Figure 3.3.

The bands in Figure 3.4 are simultaneous over the mean response and two derivatives. The price of having bands that are simultaneous over the mean response and two derivatives as opposed to bands that are simultaneous over the mean response and just one derivative is wider bands. However, in this case that price is small. The band for the mean response in Figure 3.4 is .8% wider on average than the band for the mean response in Figure 3.3. The band for the first derivative of the mean response in Figure 3.4 is also .8% wider on average than the band for the first derivative of the mean response in Figure 3.3.

The assertion that the confidence bands we have described are simultaneous over both the covariate space and one or more derivatives is obviously much stronger than an assertion that two points form a confidence interval for an estimate of the mean response at a given point. It seems reasonable then, that a researcher who requires 95% confidence in the latter situation may consider a lesser confidence level in the former situation. Reducing the confidence level will make the bands narrower and perhaps more useful to the researcher. To provide such an example, in Figure 3.5 we display 80% simultaneous confidence bands for the mean response and the first derivative.

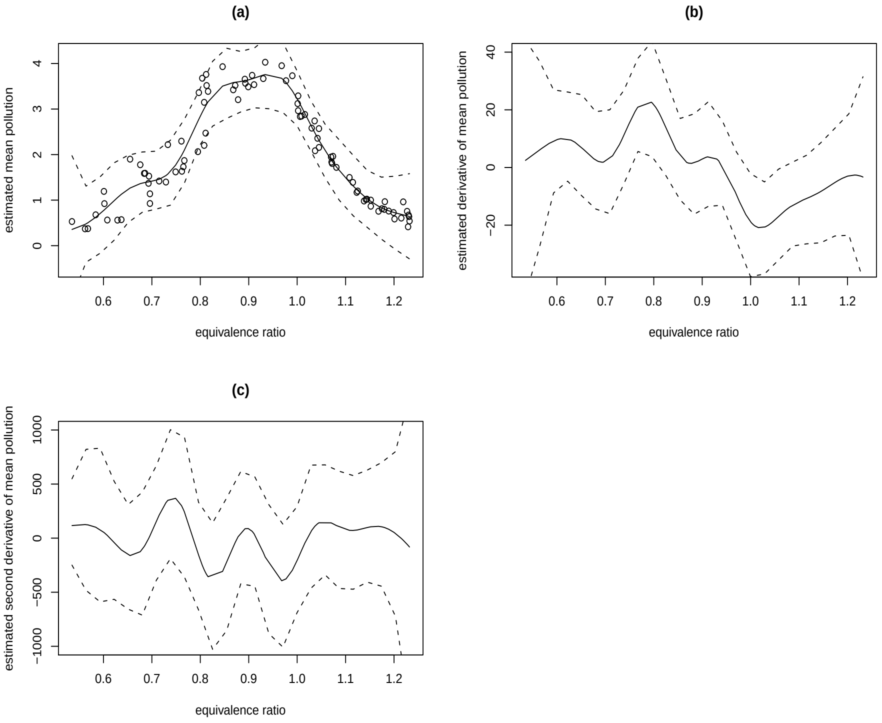
Copyright© Benjamin Hall, 2010.

Figure 3.3: Mean response and 1st derivative of ethanol data



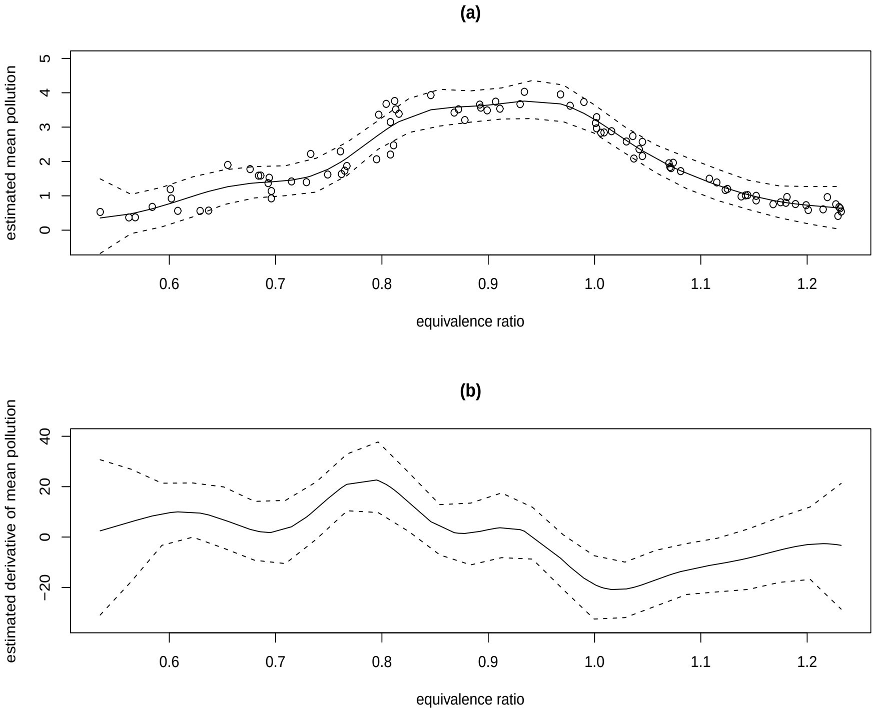
The solid curves in panels (a) and (b) indicate the estimated mean response and first derivative, respectively. The circles represent the observed data. The dashed curves indicate the 95% confidence bands. We obtained the estimates using compound estimation with filtration and extrapolation with $J = 4$, 27 centering points, $\beta = 15$ (150 during filtration and extrapolation), nearest neighbor local regression pointwise estimates using nnfrac $= .30$ (nnfrac$_0 = .15$ during filtration and extrapolation), and $\kappa = 0.1$. The confidence bands were constructed with 25 grid points.

Figure 3.4: Mean response and first two derivatives of ethanol data



The solid curves in panels (a), (b), and (c) indicate the estimated mean response, first derivative, and second derivative, respectively. The circles represent the observed data. The dashed curves indicate the 95% confidence bands. We obtained the estimates using compound estimation with filtration and extrapolation with $J = 4$, 27 centering points, $\beta = 15$ (150 during filtration and extrapolation), local regression pointwise estimates using nearest neighbor fraction .30 (.15 during filtration and extrapolation), and $\kappa = 0.1$. The confidence bands were constructed with 25 grid points.

Figure 3.5: Mean response and 1st derivative of ethanol data (80% confidence)



The solid curves in panels (a) and (b) indicate the estimated mean response and first derivative, respectively. The circles represent the observed data. The dashed curves indicate the 80% confidence bands. We obtained the estimates using compound estimation with filtration and extrapolation with $J = 4$, 27 centering points, $\beta = 15$ (150 during filtration and extrapolation), nearest neighbor local regression pointwise estimates using nnfrac = .30 (nnfrac$_0$ = .15 during filtration and extrapolation), and $\kappa = 0.1$. The confidence bands were constructed with 25 grid points.

Chapter 4 Confidence Bands in the Presence of Heteroscedastic Noise

4.1 Heteroscedastic noise

In the previous chapter we constructed simultaneous confidence bands for a mean response and its derivatives around nonparametric estimators that are linear in the observed responses. We assumed that the observed responses arose from model (1.1) with the error terms independently and identically normally distributed. In this chapter we relax this assumption to allow for heteroscedastic noise.

In what follows, we assume that model (1.1) holds where the ϵ_i are independent and

$$\epsilon_i \sim N(0, \sigma^2(x_i)),$$

where $\sigma^2(x) > 0$ is Lipschitz continuous of order 1 on $\mathcal{X}$. (A function is Lipschitz continuous of order a on $\mathcal{X}$ if there exists a positive constant m such that $|f(x_1) - f(x_2)| \leq m|x_1 - x_2|^a$ for all $x_1, x_2 \in \mathcal{X}$.)

If we are to proceed in a manner at all like the method of the previous chapter, we will need an estimator of $\sigma^2(x)$. For reasons we will later make clear, it is sufficient to estimate $\sigma^2(x)$ only at the grid of points $\mathbf{G}$ used in the construction of the confidence bands.

To estimate $\sigma^2(x)$ at a grid point $\xi \in \mathbf{G}$ we first define

$$S_{\xi,j} := \left(Y_{a+m+1-2j} - Y_{a+m-2j}\right)/\sqrt{2}, \tag{4.1}$$

for $j \in \{1, 2, ..., m\}$ where $m \in \mathbb{N}$ is chosen such that $m \to \infty$ and $m^{3/2}n^{-1}\log m \to 0$ as $n \to \infty$ and $a := \operatorname{argmin}_{i\in\{1,\ldots,n\}}\{x_i - \xi : x_i - \xi \geq 0\}$. Then, as we show in Corollary 4.1.1, $S^2_{\xi,j} \to^L \sigma^2(\xi)\chi^2_1$ under mild conditions. Definition (4.1) can easily be modified for grid points on the boundaries. For ξ on the left boundary, define

$$S_{\xi,j} := \left(Y_{2j} - Y_{2j-1}\right)/\sqrt{2},$$

for $j \in \{1, 2, ..., m\}$. For ξ on the right boundary, define

$$S_{\xi,j} := \left(Y_{n-2j+2} - Y_{n-2j+1}\right)/\sqrt{2},$$

for $j \in \{1, 2, ..., m\}$. The results of Lemma 4.1.1 and Corollary 4.1.1 still apply.

The next step is to apply a nonparametric regression estimator to the set of $S^2_{\xi,j}$. Any nonparametric technique is acceptable for this purpose as long as it is linear in the observed responses (where in this case the observed responses are the $S^2_{\xi,j}$) such that $\sum_{j=1}^{m} l_j(x) = 1$ and $\sum_{j=1}^{m} l_j^2(x) \to 0$ as $m \to \infty$. All of the methods discussed in Chapter 1 meet these criteria. Note that the estimator will need to be applied $|\mathbf{G}|$ times, once to the $S^2_{\xi,j}$ for each ξ. The resulting estimator, $\widetilde{\sigma}^2(\xi)$ is a consistent estimator of $\sigma^2(\xi)$ as demonstrated in Proposition 4.1.1, which relies on Lemma 4.1.1.

Lemma 4.1.1 *Assume that model (1.1) holds where the ϵ_i are independent and normally distributed. Assume that the variance of ϵ_i is $\sigma^2(x_i)$, where $\sigma^2(x) > 0$ and $\mu(x)$ are Lipschitz continuous functions of order 1 on $\mathcal{X}$, and the x_i are design points for which the mesh is $O(n^{-1})$. Then $\forall \xi \in \boldsymbol{G}$, $\max_j \left|S^2_{\xi,j} - X_j\right| = O_p(mn^{-1} \log m)$, where*

$$X_j = \sigma^2(\xi) \left(\epsilon_{a+m+1-2j}/\sigma(x_{a+m+1-2j}) - \epsilon_{a+m-2j}/\sigma(x_{a+m-2j})\right)^2 /2.$$

Proof. First, consider a random variable having a half normal distribution which is the result of taking the absolute value of a standard normal random variable. The half normal random variable has distribution function $F(x) = 2\Phi(x) - 1$ and density function $f(x) = 2\phi(x)$ for $x \geq 0$. Then

$$[1 - F(x)]/f(x) = [2 - 2\Phi(x)]/[2\phi(x)] = [1 - \Phi(x)]/\phi(x) \sim 1/x.$$

(Note that $f(x) \sim g(x)$ if and only if $f(x)/g(x) \to 1$ as $n \to \infty$.) So if we take b_n to be the $1 - 1/n$ quantile of the half normal distribution, then

$$1/n = 2 - 2\Phi(b_n) \sim \frac{2e^{-b_n^2/2}}{b_n\sqrt{2\pi}}$$

and, by Cramer (1946),

$$b_n = \sqrt{2 \log n} - \frac{\log(\pi \log n)/2}{\sqrt{2 \log n}} + O(1/\log n).$$

Also $2n\phi(b_n) \sim b_n$ and so by David and Nagaraja (1980), $b_n(X_{n:n} - b_n)$ or equivalently $\sqrt{2 \log n}(X_{n:n} - b_n)$, where $X_{n:n}$ denotes the largest in a random sample of size n

from the half normal distribution, converges in law to the Gumbel distribution. This implies that $X_{n:n} = O_p(\sqrt{\log n})$ which implies that

$$X_{n:n}^2 = O_p(\log n). \tag{4.2}$$

Now let $\xi \in \mathbf{G}$. Let $j_1 := a + m - 2j$ and $j_2 := a + m + 1 - 2j$. Then

$$\begin{aligned}
& \max_j \left| S_{\xi,j}^2 - X_j \right| \\
= & \max_j \left| \epsilon_{j_1}^2 \left(1 - \frac{\sigma^2(\xi)}{\sigma^2(x_{j_1})} \right) /2 + \epsilon_{j_2}^2 \left(1 - \frac{\sigma^2(\xi)}{\sigma^2(x_{j_2})} \right) /2 + \left[\mu(x_{j_1}) - \mu(x_{j_2}) \right]^2 /2 \right. \\
& \left. - \epsilon_{j_1}\epsilon_{j_2} \left(1 - \frac{\sigma^2(\xi)}{\sigma(x_{j_1})\sigma(x_{j_2})} \right) + \epsilon_{j_1} \left[\mu(x_{j_1}) - \mu(x_{j_2}) \right] - \epsilon_{j_2} \left[\mu(x_{j_1}) - \mu(x_{j_2}) \right] \right| \\
= & \max_j \left| (\epsilon_{j_1}^2 + \epsilon_{j_2}^2) O(mn^{-1}) + O(m^2 n^{-2}) - \epsilon_{j_1}\epsilon_{j_2} O(mn^{-1}) + (\epsilon_{j_1} - \epsilon_{j_2}) O(mn^{-1}) \right| \\
= & O(mn^{-1}) \max_j \left| \epsilon_{j_1}^2 + \epsilon_{j_2}^2 \right| \\
= & O(mn^{-1}) O_p(\log m), \qquad (4.3)
\end{aligned}$$

where line (4.3) is obtained from (4.2). ■

Note that Lipschitz continuity of $\mu(x)$ is ensured if $\mu(x)$ is differentiable, which is tacit if a positive integer belongs to $\{p_1, ..., p_J\}$.

Corollary 4.1.1 *Assume that the conditions of Lemma 4.1.1 hold. Let $\xi \in \boldsymbol{G}$. Then $S_{\xi,j}^2 \to^L \sigma^2(\xi)\chi_1^2$ as $n \to \infty$.*

Proof. Since $X_j \sim \sigma^2(\xi)\chi_1^2$ implies that $X_j \to^L \sigma^2(\xi)\chi_1^2$ and $S_{\xi,j}^2 - X_j \to^P 0$, we have

$$\begin{aligned}
S_{\xi,j}^2 &= S_{\xi,j}^2 - X_j + X_j \\
&\to^L 0 + \sigma^2(\xi)\chi_1^2.
\end{aligned}$$

■

Proposition 4.1.1 *Assume that the conditions of Lemma 4.1.1 hold. Let $\xi \in \boldsymbol{G}$ be a grid point. Assume that $\widetilde{\sigma}^2(\xi)$ represents a nonparametric regression estimator* *applied to the set of* $S^2_{\xi,j}$. *Also assume that the nonparametric regression estimator is linear in the observed responses (where in this case the observed responses are the* $S^2_{\xi,j}$*) and is defined such that* $\sum_{j=1}^m l_j(x) = 1$ *and* $\sum_{j=1}^m l_j^2(x) \to 0$ *as* $m \to \infty$. *Then* $\widetilde{\sigma}^2(\xi) \to^P \sigma^2(\xi)$ *as* $m \to \infty$.

Proof. Since $\sum_{j=1}^m l_j^2(x) \to 0$, then $\sum_{j=1}^m |l_j(x)| = o(m^{1/2})$ by Holder's inequality. So for any $\epsilon > 0$

$$\begin{aligned}
& P\left(\left|\widetilde{\sigma}^2(\xi) - \sigma^2(\xi)\right| \geq \epsilon\right) \\
\leq\; & P\left(\left|\widetilde{\sigma}^2(\xi) - \sum_{j=1}^m l_j(\xi)X_j\right| \geq \epsilon/2\right) + P\left(\left|\sum_{j=1}^m l_j(\xi)X_j - \sigma^2(\xi)\right| \geq \epsilon/2\right) \\
=\; & P\left(\left|\sum_{j=1}^m l_j(\xi)S^2_{\xi,j} - \sum_{j=1}^m l_j(\xi)X_j\right| \geq \epsilon/2\right) + P\left(\left|\sum_{j=1}^m l_j(\xi)X_j - \sigma^2(\xi)\right| \geq \epsilon/2\right) \\
\leq\; & P\left(\sum_{j=1}^m \left|S^2_{\xi,j} - X_j\right| |l_j(\xi)| \geq \epsilon/2\right) + Var\left[\sum_{j=1}^m l_j(\xi)X_j\right] / (\epsilon/2)^2 \\
\leq\; & P\left(\max_j \left|S^2_{\xi,j} - X_j\right| \sum_{j=1}^m |l_j(\xi)| \geq \epsilon/2\right) + 4\left(\sum_{j=1}^m l_j^2(\xi)\right) Var\left[X_1\right]/\epsilon^2 \\
=\; & P\left(O_p(mn^{-1}\log m)o(m^{1/2}) \geq \epsilon/2\right) + 8(\sigma^2(\xi))^2\left(\sum_{j=1}^m l_j^2(\xi)\right)/\epsilon^2 \qquad (4.4) \\
\to\; & 0,
\end{aligned}$$

which implies the desired result. Line (4.4) follows from Lemma 4.1.1. ■

We are now in a position to use this estimator of the heteroscedastic variance to construct simultaneous confidence bands for the mean response and its derivatives. Our approach is similar to that of Chapter 3.

We first define

$$Z_p(\xi) := \frac{\sum_{i=1}^n l_i^{(p)}(\xi)\epsilon_i}{\sqrt{\sum_{i=1}^n \sigma^2(x_i) l_i^{(p)}(\xi)^2}}. \qquad (4.5)$$

We can then use (4.5) to determine z_α from (3.12) or (3.13). By applying the ideas of chapter 3, if $B_p(x)$ is a bound on the absolute value of the bias of the estimator of $\mu^{(p)}(x)$ (or estimates such a value; see section 3.3), and if

$M_p \geq \sup_{x \in \mathcal{X}} |\mu^{(p)}(x) - \mu_I^{(p)}(x)|$ (or estimates such a value; see section 3.4), and if $D_p(x) := \sqrt{\sum_{i=1}^n l_i^{(p)}(x)^2 \sigma^2(x_i)}$, then

$$U_p(x), L_p(x) := \widehat{\mu}_I^{(p)}(x) \pm \left(M_p + B_{pI}(x) + z_\alpha D_{pI}(x)\right),$$

where $p \in \{p_1, ..., p_J\}$ and the subscript I denotes linear interpolation between the grid points, form $100(1-\alpha)\%$ simultaneous confidence bands for derivatives $p_1, ..., p_J$.

Of course, the problem is that in practice, $\sigma^2(x)$ is unknown. However since, as we demonstrate in Lemma 4.1.2 below, at the grid points

$$\frac{\sum_{i=1}^n l_i^{(p)}(\xi)^2 \sigma^2(x_i)}{\sigma^2(\xi) \sum_{i=1}^n l_i^{(p)}(\xi)^2} \rightarrow^P 1,$$

we can justify using $\widetilde{\sigma}(\xi)\sqrt{\sum_{i=1}^n l_i^{(p)}(\xi)^2}$ in place of $D_p(\xi)$ when constructing the confidence bands.

Now define $\widehat{\Sigma}$ to be a variance-covariance matrix which estimates the covariance of $Z_{p_a}(\xi_j)$ and $Z_{p_b}(\xi_k)$ to be 0 if $\xi_j \neq \xi_k$ and

$$\frac{\sum l_i^{(p_a)}(\xi_j) l_i^{(p_b)}(\xi_j)}{\sqrt{\sum l_i^{(p_a)}(\xi_j)^2 \sum l_i^{(p_b)}(\xi_j)^2}}$$

otherwise.

Now let $\mathbf{Z}_s$ be a multivariate normal random variable with mean zero and variance-covariance matrix I. Suppose that

$$\mathbf{Z} := \begin{bmatrix} Z_{p_1}(\xi_1) & \cdots & Z_{p_J}(\xi_1) & \cdots & Z_{p_1}(\xi_G) & \cdots & Z_{p_J}(\xi_G) \end{bmatrix}^t \rightarrow^L MVN\left(\mathbf{0}, \Sigma^*\right)$$

for some positive definite symmetric matrix Σ^* and that $\widehat{\Sigma} \rightarrow \Sigma^*$. (These suppositions can be checked for the particular nonparametric regression estimator being employed. We show below that they are true for kernel regression.) Then for any Borel set $A \in \mathbb{R}^{G*J}$, we have both $P(\mathbf{Z} \in A) \rightarrow P(\Sigma^{*1/2}\mathbf{Z}_s \in A)$ and $P(\widehat{\Sigma}^{1/2}\mathbf{Z}_s \in A) \rightarrow P(\Sigma^{*1/2}\mathbf{Z}_s \in A)$. Hence $P(\mathbf{Z} \in A) - P(\widehat{\Sigma}^{1/2}\mathbf{Z}_s \in A) \rightarrow 0$ which justifies treating $\mathbf{Z}$ as if its variance/covariance matrix were $\widehat{\Sigma}$. In particular, with $A := \{\mathbf{x} \in \mathbb{R}^{G*J} : \max_{1 \leq j \leq G*j} |x_j| > z_\alpha\}$, we see that z_α may be estimated using $\widehat{\Sigma}$ rather than the incalculable (since $\sigma^2(x)$ is unknown) variance/covariance matrix of $\mathbf{Z}$.

For small samples, we propose using the Satterthwaite approximation degrees of freedom estimate and employing a multivariate t-distribution in place of a multivariate normal distribution when estimating z_α.

Lemma 4.1.2 *Assume that the conditions of Lemma 4.1.1 hold. Assume the there exists* $C > 0$ *such that*

$$\sum_{i:|x_i-\xi|>Cm/n} l_i^{(p)}(\xi)^2 = o\left(\sum_{i:|x_i-\xi|\leq Cm/n} l_i^{(p)}(\xi)^2 \right). \tag{4.6}$$

Then $\forall \xi \in \boldsymbol{G}$,

$$\frac{\sum_{i=1}^n l_i^{(p)}(\xi)^2 \sigma^2(x_i)}{\sigma^2(\xi) \sum_{i=1}^n l_i^{(p)}(\xi)^2} \to^P 1.$$

Proof. Let $\xi \in \mathbf{G}$ be a grid point. Let $\epsilon > 0$. Then

$$\begin{aligned}
& P\left(\left| \frac{\sum_{i=1}^n l_i^{(p)}(\xi)^2 \sigma^2(x_i)}{\sigma^2(\xi) \sum_{i=1}^n l_i^{(p)}(\xi)^2} - 1 \right| \geq \epsilon \right) \\
= \quad & P\left(\left| \frac{\sum_{i:|x_i-\xi|\leq Cm/n} l_i^{(p)}(\xi)^2 \sigma^2(x_i)}{\sigma^2(\xi) \sum_{i:|x_i-\xi|\leq Cm/n} l_i^{(p)}(\xi)^2} [1 + o(1)] - 1 \right| \geq \epsilon \right) \\
= \quad & P\left(\left| \frac{\sum_{i:|x_i-\xi|\leq Cm/n} l_i^{(p)}(\xi)^2 \sigma^2(\xi)}{\sigma^2(\xi) \sum_{i:|x_i-\xi|\leq Cm/n} l_i^{(p)}(\xi)^2} [1 + o(1)] - 1 \right| \geq \epsilon \right) \\
= \quad & P\left(|[1 + o(1)] - 1| \geq \epsilon \right) \\
\to \quad & 0
\end{aligned}$$

which implies the desired result. ■

The validity of assumption (4.6) in Lemma 4.1.2 depends on the nonparametric regression technique being employed, but note that this assumption is true for kernel regression with a compactly supported kernel and bandwidth $h \leq Cm/n$ since in this case the left side of (4.6) is 0.

We now demonstrate that our suppositions about $\mathbf{Z}$ and $\widehat{\Sigma}$ are reasonable by illustrating their satisfaction.

Consider kernel regression with a compactly supported kernel $K(u)$ that has p_J+1 continuous derivatives and such that the $J \times J$ matrix R with (a, b) entry

$$\int_{\mathbb{R}} K^{(p_a)}(u) K^{(p_b)}(u) du \Big/ \sqrt{\int_{\mathbb{R}} K^{(p_a)}(u)^2 du \int_{\mathbb{R}} K^{(p_b)}(u)^2 du}$$

is positive definite. Let Σ^* be a $GJ \times GJ$ block diagonal matrix containing G copies of R. Note that Σ^* is positive definite and symmetric.

Lemma 4.1.3 *Let $K(u)$ and Σ^* be as described above. Assume that the bandwidth parameter $h \leq Cm/n \to 0$ as $n \to \infty$, that $nh \to 0$, and that the design points are uniform on $\mathcal{X}$. Then, under the conditions of Lemma 4.1.1, with kernel regression we have $Var(\boldsymbol{Z}) \to \Sigma^*$ and $\widehat{\Sigma} \to \Sigma^*$ as $n \to \infty$.*

Proof. Without loss of generality, take $\mathcal{X} := [-1, 1]$. First consider the elements of $Var(\mathbf{Z})$ corresponding to $\xi_j \neq \xi_k$. Consider n large enough so that $|\xi_j - \xi_k| > 2h$. Note that $K^{(p_a)}((\xi_j - x_i)/h) \neq 0$ only if $|\xi_j - x_i| \leq h$. But if $|\xi_j - x_i| \leq h$ then

$$|\xi_k - x_i| = |\xi_k - \xi_j + \xi_j - x_i| \geq ||\xi_k - \xi_j| - |\xi_j - x_i|| > h,$$

which means that $K^{(p_b)}((\xi_k - x_i)/h) = 0$. So for n large enough that $|\xi_j - \xi_k| > 2h$, the elements of $Var(\mathbf{Z})$ corresponding to $\xi_j \neq \xi_k$ are 0.

Now consider the elements of $Var(\mathbf{Z})$ corresponding to $\xi_j = \xi_k$. Then

$$\begin{aligned}
& \frac{\sum_{i=1}^n l_i^{(p_a)}(\xi_j) l_i^{(p_b)}(\xi_j)\sigma^2(x_i)}{\sigma^2(\xi_j)\sqrt{\sum_{i=1}^n l_i^{(p_a)}(\xi_j)^2 \sum_{i=1}^n l_i^{(p_b)}(\xi_j)^2}} \\
= \; & \frac{\sum_{i:|x_i-\xi_j|\le Cm/n} l_i^{(p_a)}(\xi_j) l_i^{(p_b)}(\xi_j)\sigma^2(x_i)}{\sigma^2(\xi_j)\sqrt{\sum_{i=1}^n l_i^{(p_a)}(\xi_j)^2 \sum_{i=1}^n l_i^{(p_b)}(\xi_j)^2}} \\
= \; & \frac{\sigma^2(\xi_j)\sum_{i:|x_i-\xi_j|\le Cm/n} l_i^{(p_a)}(\xi_j) l_i^{(p_b)}(\xi_j)}{\sigma^2(\xi_j)\sqrt{\sum_{i=1}^n l_i^{(p_a)}(\xi_j)^2 \sum_{i=1}^n l_i^{(p_b)}(\xi_j)^2}}[1+o(1)] \\
= \; & \frac{\sum_{i:|x_i-\xi_j|\le 2m/n} K^{(p_a)}(\frac{\xi_j-x}{h}) K^{(p_b)}(\frac{\xi_j-x}{h})}{\sqrt{\sum_{i=1}^n K^{(p_a)}(\frac{\xi_j-x}{h})^2 \sum_{i=1}^n K^{(p_b)}(\frac{\xi_j-x}{h})^2}}[1+o(1)] \\
= \; & \frac{\int_{-1}^1 K^{(p_a)}(\frac{\xi_j-x}{h}) K^{(p_b)}(\frac{\xi_j-x}{h})dx}{\sqrt{\int_{-1}^1 K^{(p_a)}(\frac{\xi_j-x}{h})^2 dx \int_{-1}^1 K^{(p_b)}(\frac{\xi_j-x}{h})^2 dx}}[1+o(1)], \qquad (4.7) \\
= \; & \frac{\int_{(\xi_j-1)/h}^{(\xi_j+1)/h} K^{(p_a)}(u) K^{(p_b)}(u)du}{\sqrt{\int_{(\xi_j-1)/h}^{(\xi_j+1)/h} K^{(p_a)}(u)^2 du \int_{(\xi_j-1)/h}^{(\xi_j+1)/h} K^{(p_b)}(u)^2 du}}[1+o(1)], \\
= \; & \frac{\int_{\mathbb{R}} K^{(p_a)}(u) K^{(p_b)}(u)du}{\sqrt{\int_{\mathbb{R}} K^{(p_a)}(u)^2 du \int_{\mathbb{R}} K^{(p_b)}(u)^2 du}}[1+o(1)],
\end{aligned}$$

where line (4.7) and line (4.8) below come from the fact that for any function $f(x)$ which is bounded in n and has a continuous derivative such that $|f'(x)| \le M(n) =$

$o(n)$ we have that

$$\begin{aligned}
&\left|\frac{2}{n}\sum_{i=1}^{n} f(x_i) - \int_{-1}^{1} f(x)dx\right| \\
= &\left|\frac{2f(x_1)}{n} + \sum_{i=2}^{n} f(x_i)(x_i - x_{i-1}) - \sum_{i=2}^{n}\int_{x_{i-1}}^{x_i} f(x)dx\right| \\
= &\left|\frac{2f(x_1)}{n} + \sum_{i=2}^{n} f(x_i)(x_i - x_{i-1}) - \sum_{i=2}^{n} f(\gamma_i)(x_i - x_{i-1})\right| \\
= &\left|\frac{2f(x_1)}{n} + \sum_{i=2}^{n}[f(x_i) - f(\gamma_i)](x_i - x_{i-1})\right| \\
\leq &\left|\frac{2f(x_1)}{n}\right| + \left|\sum_{i=2}^{n} M(n)(x_i - x_{i-1})^2\right| \\
= &\left|\frac{2f(x_1)}{n}\right| + \left|\frac{4(n-1)M(n)}{n^2}\right| \\
\to &\ 0.
\end{aligned}$$

Note that the elements of $\widehat{\Sigma}$ and Σ^* corresponding to $\xi_j \neq \xi_k$ or $p_a = p_b$ are equal by definition. For $\xi_j = \xi_k$ and $p_a \neq p_b$ we have

$$\begin{aligned}
&\frac{\sum_{i=1}^{n} l_i^{(p_a)}(\xi_j) l_i^{(p_b)}(\xi_j)}{\sqrt{\sum_{i=1}^{n} l_i^{(p_a)}(\xi_j)^2 \sum_{i=1}^{n} l_i^{(p_b)}(\xi_j)^2}} \\
= &\frac{\sum_{i=1}^{n} K^{(p_a)}(\frac{\xi_j - x_i}{h}) K^{(p_b)}(\frac{\xi_j - x_i}{h})}{\sqrt{\sum_{i=1}^{n} K^{(p_a)}(\frac{\xi_j - x_i}{h})^2 \sum_{i=1}^{n} K^{(p_b)}(\frac{\xi_j - x_i}{h})^2}} \\
= &\frac{\int_{-1}^{1} K^{(p_a)}(\frac{\xi_j - x}{h}) K^{(p_b)}(\frac{\xi_j - x}{h})dx}{\sqrt{\int_{-1}^{1} K^{(p_a)}(\frac{\xi_j - x}{h})^2 dx \int_{-1}^{1} K^{(p_b)}(\frac{\xi_j - x}{h})^2 dx}}[1 + o(1)] \qquad (4.8) \\
= &\frac{\int_{(\xi_j - 1)/h}^{(\xi_j + 1)/h} K^{(p_a)}(u) K^{(p_b)}(u)du}{\sqrt{\int_{(\xi_j - 1)/h}^{(\xi_j + 1)/h} K^{(p_a)}(u)^2 du \int_{(\xi_j - 1)/h}^{(\xi_j + 1)/h} K^{(b)}(u)^2 du}}[1 + o(1)] \\
\to &\frac{\int_{\mathbb{R}} K^{(p_a)}(u) K^{(p_b)}(u)du}{\sqrt{\int_{\mathbb{R}} K^{(p_a)}(u)^2 du \int_{\mathbb{R}} K^{(p_b)}(u)^2 du}},
\end{aligned}$$

which implies the desired result. ∎

4.2 Simulation studies

We investigated this methodology for constructing $100(1-\alpha)\%$ simultaneous confidence bands for a mean response and its derivatives in the presence of heteroscedastic noise through simulation. To do this we generated 500 data sets each of size 50 from (1.1) with the true underlying mean response set to be $\mu(x) := \sin(2\pi x) + \cos(2\pi x) + \log(4/3 + x)$ with $x_1, ..., x_{50}$ equispaced on $\mathcal{X} = [-1, 1]$ and four different variance functions:

$$\sigma_1^2(x) = \frac{(x+2)^3}{40},$$

$$\sigma_2^2(x) = \frac{3x^2}{4} + .01,$$

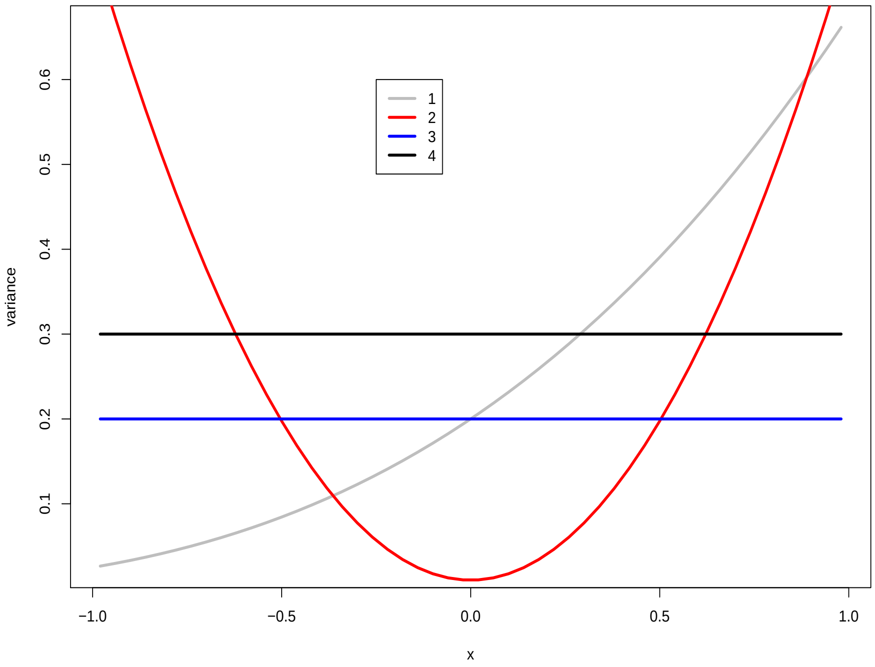
$$\sigma_3^2(x) = .2,$$

$$\sigma_4^2(x) = .3.$$

Note that $\sigma_1^2(x)$ is monotone increasing, while $\sigma_2^2(x)$ is monotone decreasing over $[-1, 0]$ and monotone increasing over $[0, 1]$. Figure 4.1 depicts these variance functions graphically. We performed simulations for $m \in \{12, 16, 20, 24\}$ (recall that m is the number of $S_{\xi,j}^2$'s used to determine $\widetilde{\sigma}^2(\xi)$) and $\alpha \in \{.05, .20\}$. We estimated the mean response and its derivatives using compound estimation with filtration and extrapolation with $J = 4$, 27 centering points, $\beta = 15$ (150 during filtration and extrapolation), local regression pointwise estimates using nearest neighbor fraction .30 (.15 during filtration and extrapolation), and $\kappa = 0.1$. We then placed simultaneous confidence bands around the estimates of the mean response and first derivative and simultaneous confidence bands around the estimates of the mean response and first two derivatives, in each case constructing the bands by interpolating (see Section 3.4) over a grid size of $G = 25$.

The results of this simulation study are recorded below in Table 4.1. In most cases, the confidence bands are conservative, achieving at least the nominal coverage level. In the case of σ_2^2, where the coverage level is not achieved for $\alpha = .20$, the problem is likely that the variance is excessively large near the boundaries, where estimation is already extremely difficult. In a follow-up simulation with $\sigma_5^2 = \sigma_2^2/2$

Figure 4.1: Variance functions used in simulations



The gray line is the variance function σ_1^2. The red line is σ_2^2. The blue line is σ_3^2. The black line is σ_4^2.

we obtained the results displayed in Table 4.2. The success of the bands appears to depend on both the shape and magnitude of the variance function. Notice that the bands are most conservative when the variance is constant (σ_3^2 and σ_4^2). A researcher who is able to recognize that the variance is indeed constant will be able to obtain narrower bands by employing the methodology of the previous chapter. A researcher who makes the more conservative assumption allowing for heteroscedastic variance will get more conservative simultaneous confidence bands.

Figure 4.2 illustrates how the width of the bands changes to accommodate noise generated with different variance functions. Notice that the width of the bands corresponding to data with variance function $\sigma_1^2(x)$ increases as the variance increases, i.e. as x increases. The bands corresponding to data with variance function $\sigma_2^2(x)$ are narrowest where the variance is smallest (where $x \approx 0$). These features are somewhat distorted by the fact that the bands naturally enlarge near the boundaries even with constant variance. Nevertheless, the bands obviously respond to the heteroscedasticity.

Table 4.1: Simulation results for simultaneous confidence bands

	α=.05		α=.20	
	0,1	0,1,2	0,1	0,1,2
$\sigma_1^2, m = 12$	100.0%	100.0%	99.8%	99.8%
$\sigma_1^2, m = 16$	99.4%	98.4%	84.0%	79.0%
$\sigma_1^2, m = 20$	98.8%	98.2%	79.4%	74.6%
$\sigma_1^2, m = 24$	97.8%	97.0%	76.2%	72.0%
$\sigma_2^2, m = 12$	95.8%	93.8%	66.2%	60.0%
$\sigma_2^2, m = 16$	95.8%	94.2%	67.8%	61.0%
$\sigma_2^2, m = 20$	95.8%	94.2%	68.4%	61.2%
$\sigma_2^2, m = 24$	96.4%	94.4%	70.0%	64.0%
$\sigma_3^2, m = 12$	100.0%	100.0%	100.0%	100.0%
$\sigma_3^2, m = 16$	100.0%	100.0%	100.0%	99.8%
$\sigma_3^2, m = 20$	100.0%	100.0%	100.0%	99.6%
$\sigma_3^2, m = 24$	100.0%	100.0%	100.0%	99.6%
$\sigma_4^2, m = 12$	100.0%	100.0%	97.6%	96.4%
$\sigma_4^2, m = 16$	100.0%	100.0%	97.8%	97.2%
$\sigma_4^2, m = 20$	100.0%	100.0%	99.4%	98.8%
$\sigma_4^2, m = 24$	100.0%	100.0%	96.2%	95.4%

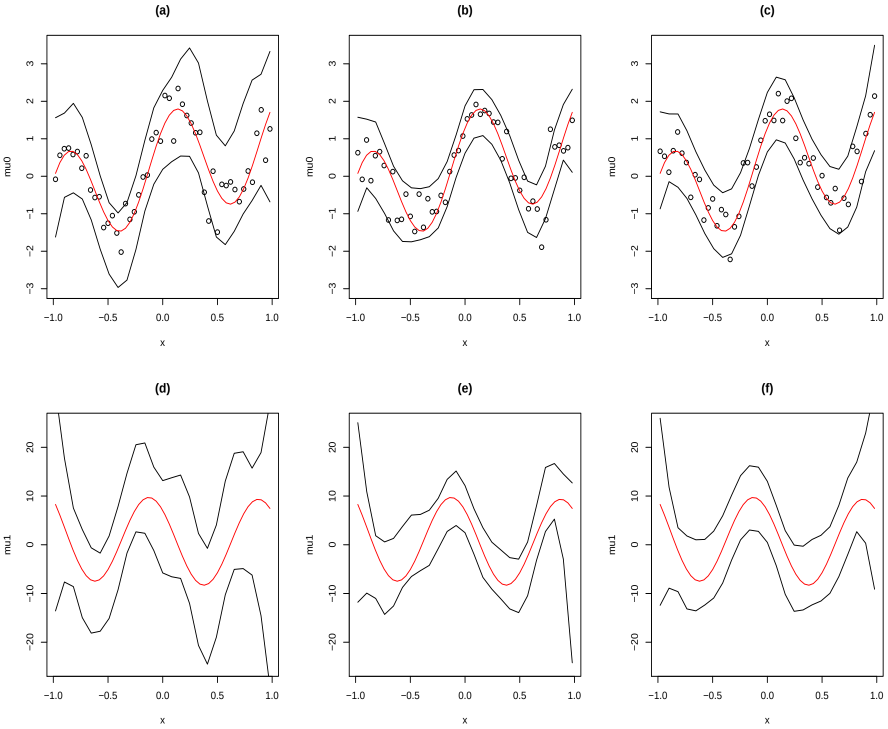
The columns labeled $0, 1$ indicate that simultaneous confidence bands were constructed for the mean response and its first derivative. The columns labeled $0, 1, 2$ indicate that simultaneous confidence bands were constructed for the mean response and its first two derivatives. The entries represent the percentages of simultaneous confidence bands that contained the true mean response and the true derivative(s) at each value of the covariate based on the 500 simulated data sets.

Table 4.2: More simulation results for simultaneous confidence bands

	α=.05		α=.20	
	0,1	0,1,2	0,1	0,1,2
$\sigma_5^2, m = 12$	100.0%	100.0%	100.0%	100.0%
$\sigma_5^2, m = 24$	100.0%	100.0%	97.2%	95.0%

The columns labeled $0, 1$ indicate that simultaneous confidence bands were constructed for the mean response and its first derivative. The columns labeled $0, 1, 2$ indicate that simultaneous confidence bands were constructed for the mean response and its first two derivatives. The entries represent the percentages of simultaneous confidence bands that contained the true mean response and the true derivative(s) at each value of the covariate based on the 500 simulated data sets.

Figure 4.2: Comparison of confidence bands across different variance functions



The circles in panels (a), (b), and (c) represent simulated data with variance functions $\sigma_1^2(x)$, $\sigma_2^2(x)$, and $\sigma_3^2(x)$, respectively. In each case the mean response is $\mu(x) := \sin(2\pi x) + \cos(2\pi x) + \log(4/3 + x)$. The red lines in the top row of panels are plots of this mean response. The red lines in the bottom row of panels are plots of the first derivative of this mean response. The black lines are 80% simultaneous confidence bands with $m = 12$.

4.3 Fetal growth example

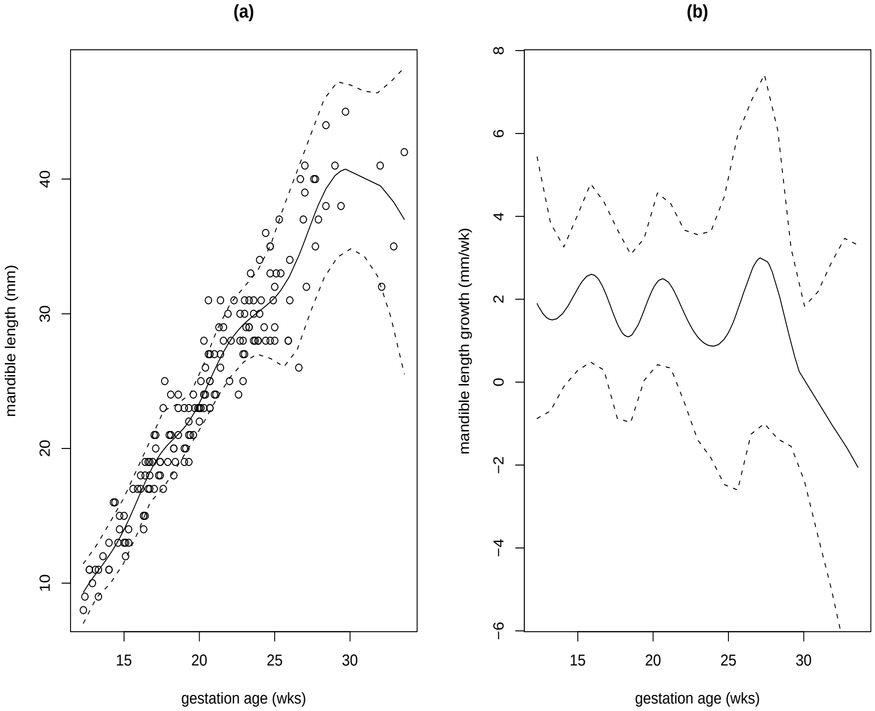
In this section we apply our methodology to a data set of size 167 from Royston and Altman (1994) involving fetal growth. The variables under consideration are the age of the fetus (in weeks) and the length of the mandible (in mm). The data set provides a clear example of heteroscedastic noise, with the variance increasing with age.

This data set is interesting because while we would expect mandible length to be a monotone increasing function, nonparametric regression estimates of the mean response and its derivative indicate a downward trend in mandible length during weeks 28 to 34 (see Figure 4.3). But is this trend based on reality or is it due to the increased variance and sparsity of data points at larger values of age? We can examine this question using our methodology for confidence band construction.

The confidence bands for the mean response in panel (a) of Figure 4.3 become wider as age increases, reflecting the heteroscedastic noise. While no convex function fits inside the confidence bands for the mean response, the bands do indicate that the true mean response could still be a monotone increasing function. This can also be seen from the confidence bands for the first derivative in panel (b). These bands identify 14 to 17 weeks and 19 to 22 weeks as regions over which growth is clearly positive. And while the estimate of the first derivative is negative after week 28, the confidence bands over this region are wide that we cannot rule out *growth* as high as 2 mm/week.

A couple of points are worth noting in this analysis. The first is that Royston and Altman (1994) chose to exclude the nine observations for which age was greater than 28 weeks. They did this for two reasons: these observations had 'excessive measurement error' and represented a 'highly selected group'. We note that if the latter reason is indeed valid, these observations should be excluded from our analysis. However, the first reason should not necessitate their exclusion for our purposes. Increasing measurement error is simply heteroscedastic noise which is reflected in our method by the increasingly wide bands. The second point is that just as in Chapter 3, a researcher who normally employs 95% confidence intervals may be willing to

Figure 4.3: Mean response and first derivative of fetal growth data

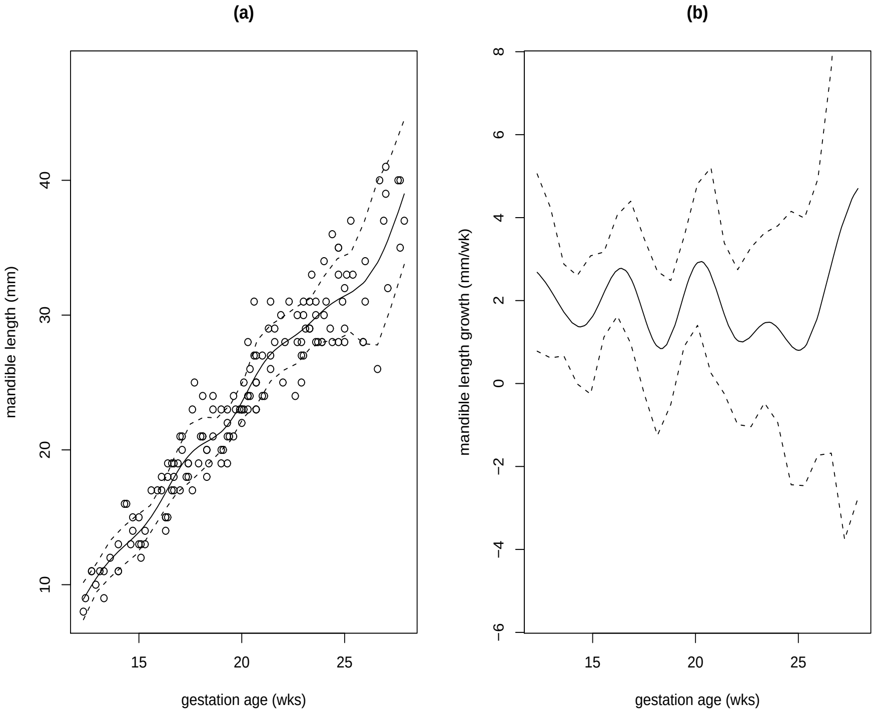


The solid curves in panels (a) and (b) indicate the estimated mean response and first derivative, respectively. The circles represent the observed data. The dashed curves indicate the 95% confidence bands. We obtained the estimates using compound estimation with filtration and extrapolation with $J = 4$, 27 centering points, $\beta = 15$ (150 during filtration and extrapolation), local regression pointwise estimates using nearest neighbor fraction .30 (.15 during filtration and extrapolation), and $\kappa = 0.1$. The confidence bands were constructed with $G = 25$ grid points and $m = 20$.

consider simultaneous confidence bands with a reduced confidence level. In Figure 4.4 we display 80% confidence intervals applied to the 158 observations remaining when we exclude measurements with a gestational age greater than 28 weeks.

Copyright© Benjamin Hall, 2010.

Figure 4.4: Mean response and first derivative of fetal growth up to 28 weeks



The solid curves in panels (a) and (b) indicate the estimated mean response and first derivative, respectively. The circles represent the observed data. The dashed curves indicate the 80% confidence bands. We obtained the estimates using compound estimation with filtration and extrapolation with $J = 4$, 27 centering points, $\beta = 15$ (150 during filtration and extrapolation), local regression pointwise estimates using nearest neighbor fraction .30 (.15 during filtration and extrapolation), and $\kappa = 0.1$. The confidence bands were constructed with $G = 25$ grid points and $m = 20$.

Chapter 5 Limited Angle Recovery of Nanoparticle Characteristics

5.1 Introduction

As mentioned in Chapter 1, one of the many areas in which nonparametric derivative estimation is useful is the characterization of nanoparticles. In this chapter we offer an enhancement to the characterization methods of Francoeur et al (2007), Charnigo et al (2007), and Charnigo et al (2010) by presenting methodology which identifies subsets of the covariate space (i.e. the possible values of the far-field recovery angle θ; in this case $\mathcal{X} = [0, 180]$) which are most useful for characterization purposes. Identification of such subsets is desirable because it reduces the time and resources required to perform the characterization.

The nanoparticle characterization process entails two major steps which involve nonparametric estimation: the forward problem and the inverse problem. The forward problem is that of obtaining the reference curves, i.e. the scattering profiles (and their derivatives) for known configurations. We assume that construction of the reference curves entails negligible random error since, for example, an experimenter may process many samples with S known and calculate an average. A nonparametric regression estimator is still necessary, however, since observations of the scattering profile are only available on a discrete subset of $\mathcal{X}$. We refer to this subset as $\mathcal{T}$. The set $\mathcal{T}$ needs to be dense in order for the error involved in estimating the entire curve from the finite grid to be small.

The inverse problem is that of characterizing unknown configurations and involves obtaining estimates of the scattering profiles (and their derivatives) for these configurations. These estimates are obtained by applying a nonparametric regression estimator to observed data which arises from the following model:

$$Y_i = M(\theta_i; S) + \epsilon_i \tag{5.1}$$

where $M(\theta_i; S)$ represents the scattering profile for configuration S and the ϵ_i are distributed independently and identically with mean 0 and variance σ^2. For the inverse problem S is unknown to the experimenter. Again the observations are obtained over a discrete subset of $\mathcal{X}$. We refer to this set as $\mathcal{T}^*$. It is not necessarily the case that $\mathcal{T} = \mathcal{T}^*$.

This chapter describes how to choose $\mathcal{T}$ for the forward problem and $\mathcal{T}^*$ for the inverse problem. Importantly, we offer methodology for saving effort in solving both cases.

5.2 Limited angles for the reference curves

In this section we propose a resource-saving method for identifying an optimal subset of far-field recovery angles to be used in the construction of the reference curves.

Consider that for a given characteristic many reference curves will need to be constructed. Exactly how many will depend on the domain (call it $\mathcal{C}_0$) of possible characteristic values and the fineness of the grid (call it $\mathcal{C}_2$) which forms a discrete approximation of $\mathcal{C}_0$. For example, if size is the characteristic, more reference curves will be required if the domain of possible sizes (in nm) is $\mathcal{C}_0 = [5, 100]$ as opposed to $\mathcal{C}_0 = [5, 50]$. Also more reference curves will be required if $\mathcal{C}_2 = \{5, 10, 15, ..., 50\}$ as opposed to $\mathcal{C}_2 = \{5, 14, 23, ..., 50\}$. To perform characterization more precisely, we can take $\mathcal{C}_1$ to be a grid which is denser than $\mathcal{C}_2$ and define, for any c in $\mathcal{C}_1$ which is not also in $\mathcal{C}_2$,

$$M(\theta; c) := M(\theta, a) + \frac{c-a}{b-a}[M(\theta, b) - M(\theta, a)],$$

where a is the largest element of $\mathcal{C}_2$ less than c and b is the smallest element of $\mathcal{C}_2$ greater than c.

Since the construction of each reference curve involves obtaining data at each $\theta \in \mathcal{T}$, the cardinality of $\mathcal{T}$ will clearly impact the effort required to construct the reference curves. A $\mathcal{T}$ with a small cardinality which nevertheless covers the regions of $\mathcal{X}$ most useful for characterization would be expedient. In what follows we demonstrate how to identify such a $\mathcal{T}$.

The idea is that we will choose an initial $\mathcal{T}_I$ which is sparse. We then evaluate the $\theta \in \mathcal{T}_I$ and determine which are 'good' choices. Finally we choose $\mathcal{T}$ to be a grid which is dense around these ‘good’ choices and excludes the intervals around the ‘bad’ choices. Of course, to do this we must define what makes a given θ ‘good’ for characterization purposes.

Consider the simplest case, depicted previously in Figure 1.1, in which we have scattering profiles available for two known configurations (call them A and B) and wish to characterize a third unknown configuration as closer to one or the other known configuration. Then it is intuitively reasonable that a good choice of θ will be one for which the quantity

$$|M(\theta; A) - M(\theta; B)| \tag{5.2}$$

is large. In Figure 1.1, this implies that $\theta = \{25, 75\}$ are good choices while $\theta = \{50, 150\}$ are poor choices.

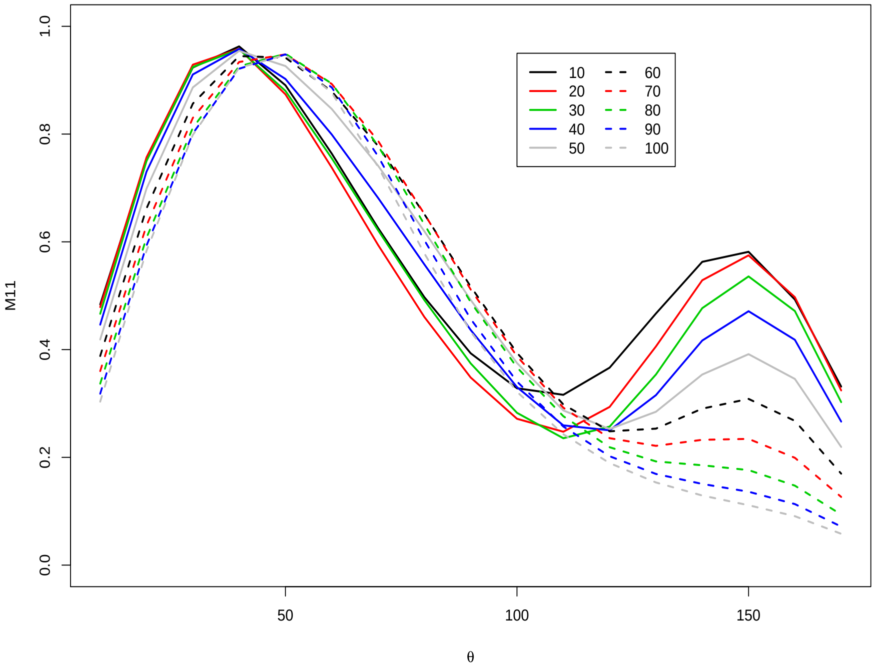
The situation is much more complicated when we involve the scattering profiles for multiple known configurations (i.e. the cardinality of $\mathcal{C}_2$ is greater than 2). In this case there does not exist a single value such as (5.2) which describes the characterization ability of a given θ. This can be seen below in Figure 5.1, which depicts the scattering profiles of 10 different nanoparticles of known sizes. Examining $\theta \approx 150$ we see that $|M(150; 50) - M(150; 60)|$ is relatively large (indicating that $\theta \approx 150$ is a good choice for distinguishing between 50 nm and 60 nm configurations), while $|M(150; 10) - M(150; 20)|$ is relatively small (indicating that $\theta \approx 150$ is a poor choice for distinguishing between 10 nm and 20 nm configurations).

We need a quantity that measures, for a given θ, the magnitude of the change in $M(\theta; S)$ as S changes. This quantity is

$$\delta(\theta, S) := \left| \frac{\partial}{\partial S} M(\theta; S) \right| \tag{5.3}$$

and can be estimated using nonparametric regression methods by viewing S as the covariate. When $\delta(\theta, S)$ is large for given θ and S, the implication is that the θ under consideration is a good choice for characterizing nanoparticles of approximate size S. Figure 5.2 displays $\delta(\theta, S)$ as a function of S. The graph indicates, for example, that $\theta = 160$ is a good choice for characterizing particles whose size is between 40 nm and

Figure 5.1: Scattering profiles for M11

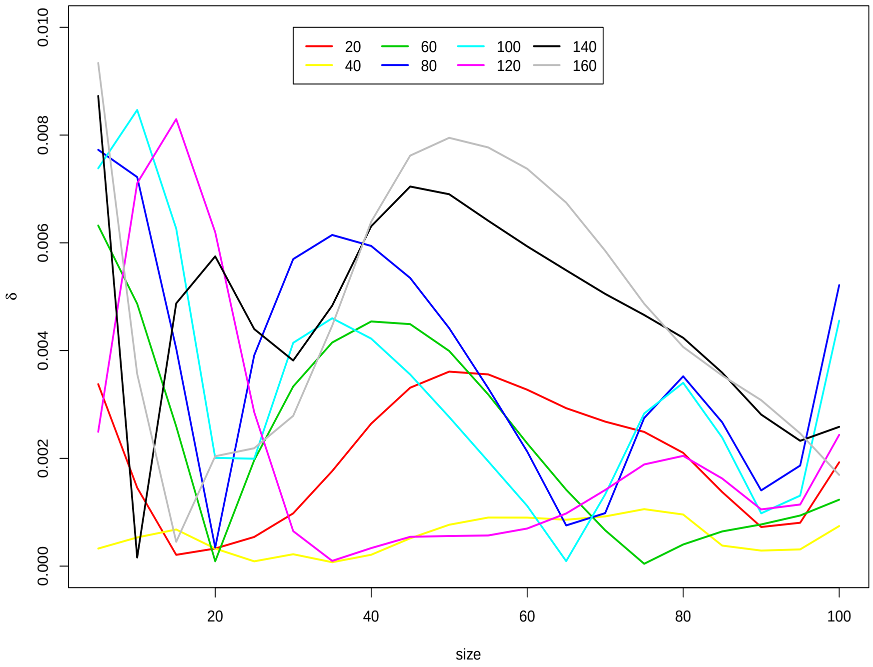


The graph above presents M_{11} scattering profiles as a function of θ. Each curve represents a different size (in nm). The curves were constructed with $\mathcal{T}_I = \{10, 20, ..., 170\}$ using compound estimation with filtration and extrapolation and $J = 7$, 27 centering points, $\beta = 12$ (120 for the initial run), and $\kappa = 0.2$. Local constant and slope estimates were obtained using smoothing splines, while coefficients of higher-order local fits were obtained using inductive estimators with $h_2 = h_2 = 1/150$, $h_4 = h_5 = h_6 = h_7 = 1/30$. Agglomeration level is set at 50%.

80 nm, but not a good choice for characterizing particles whose size is about 15 nm. For particles of approximate size 15 nm, $\theta = 120$ is a better choice.

The quantity $\delta(\theta, S)$ is the primary ingredient in our determination of $\mathcal{T}$. However, we need to consider two additional issues before making this determination. The first is that nanoparticles of different sizes are not all characterized with equal levels of difficulty. From Figure 5.2 we see that there are several angles at which the characterization of size 15 nm particles is more easily performed than is the characterization of size 95 nm particles at *any* angle. This is potentially problematic for a couple of reasons. We want to avoid selecting too many angles merely on the basis

Figure 5.2: Evaluating characterization ability of M11

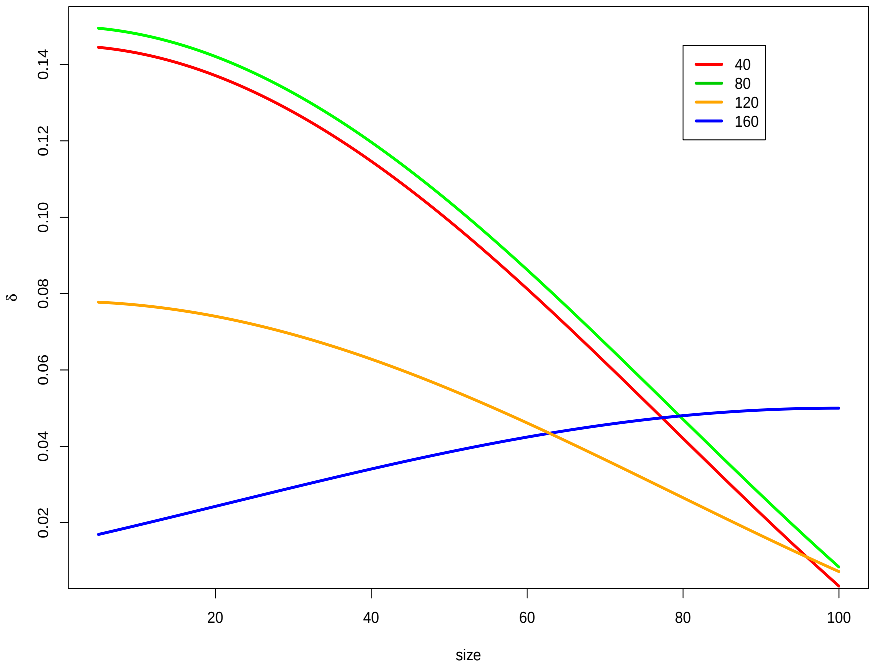


The graph above presents the quantity $\delta(\theta, S)$ for the M_{11} scattering profile as a function of S. Each curve represents a different θ. Estimation was performed using compound estimation with filtration and extrapolation and $J = 7$, 27 centering points, $\beta = 12$ (120 during filtration and extrapolation), and $\kappa = 0.2$. Local constant and slope estimates were obtained using smoothing splines with splines of order 7, while coefficients of higher-order local fits were obtained using inductive estimators with $h_2 = h_3 = 1/150$, $h_4 = h_5 = h_6 = h_7 = 1/30$. Agglomeration level is set at 50%.

that they characterize the same, easily characterized size particles well. That is, we do not want to increase the cardinality of $\mathcal{T}$ for redundant characterization ability at sizes which are easily characterized. At the same time, we want to avoid selecting too few (or zero!) angles at which characterization of difficult-sized particles is performed relatively well on the basis of small absolute characterization ability.

As an example consider the hypothetical situation presented in Figure 5.3. If $\theta = 40$ (represented by the red curve) and $\theta = 80$ (represented by the green curve) are included in $\mathcal{T}$, there is clearly no need to include $\theta = 120$ (represented by the orange curve). Therefore, the researcher may be tempted to define $\mathcal{T}$ to include only values of θ for which $\delta(\theta, S)$ exceeds .08 for some S. This accomplishes the goal of excluding $\theta = 120$. However, this is problematic in that such a threshold also excludes $\theta = 160$ (represented by the blue curve), which is the only value of θ with substantial ability to characterize particles larger than 90 nm. Indeed, there is no threshold which will exclude $\theta = 120$ and include $\theta = 160$ at the same time.

Figure 5.3: Example of potential limited angle choice pitfall



The graph above presents hypothetical $\delta(\theta, S)$ quantities as a function of S. Each curve represents a different θ. Note that if the threshold for inclusion of angles is set at .06 or above, the blue curve ($\theta = 160$) is not selected and it will then be extremely difficult nanoparticles larger than 90 nm. If the threshold is set at .06 or below, the orange curve ($\theta = 120$), which supplies only redundant characterization ability, will be selected.

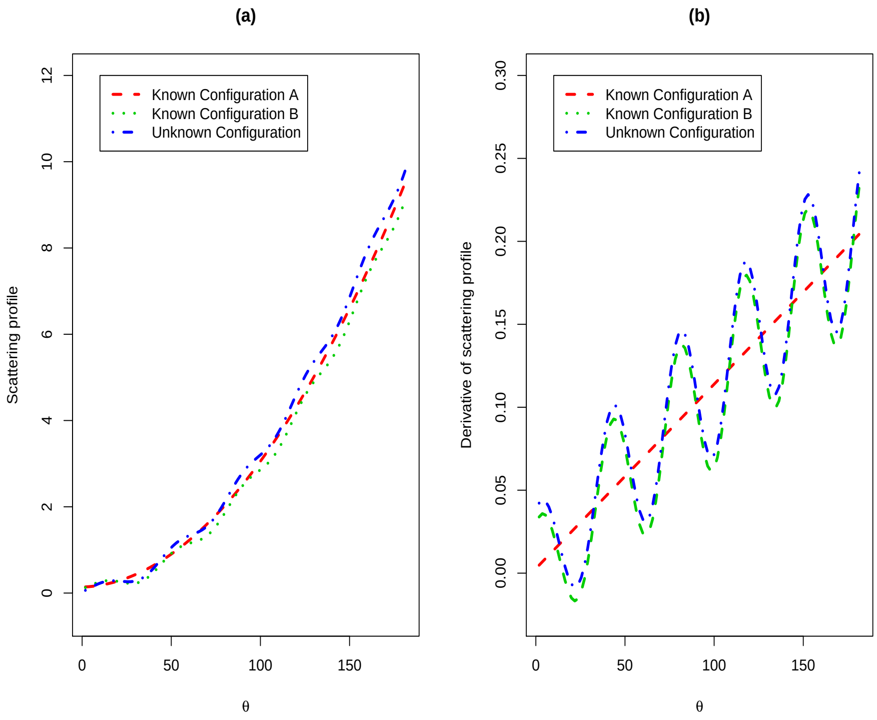
To avoid this dilemma we propose the following normalization:

$$\Delta(\theta, S) := \frac{\delta(\theta, S)}{\sum_{t \in \mathcal{T_I}} \delta(t, S)/|\mathcal{T}_I|}, \tag{5.4}$$

where $|\mathcal{T}_I|$ indicates the cardinality of $\mathcal{T}_I$.

The second issue is that we need to be able to consider the characterization ability at a given θ not only for the scattering profile, but also for its derivative(s) with respect to θ. To illustrate, consider the hypothetical situation depicted in Figure 5.4. In this case characterization by the scattering profile is ambiguous while the first derivative provides valuable information for characterization. Charnigo et al (2007) have illustrated that such situations, where derivatives provide additional characterization ability, exist in nanoparticle characterization.

Figure 5.4: Hypothetical scattering profiles and derivatives



Panel (a) depicts scattering profiles while Panel (b) depicts their derivatives. Characterization by the scattering profile is ambiguous. The first derivative of the scattering profile provides valuable information, namely that known configuration B is a better guess for the unknown configuration than known configuration A.

To address this issue we return to our assertion that a good choice of θ is one for which $\delta(\theta, S)$ is large. More specifically, a good choice of θ is one for which $\delta(\theta, S)$ is large relative to the error involved in the computation of $M^*(\theta; S)$ for the unknown configuration, where the $*$ superscript indicates the result of a nonparametric regression estimator.

In particular, if we assume that the bias of the nonparametric regression estimator is negligible, then the size of the errors can be expressed in terms of the variance. The difficulty is that while, analogous to (5.3),

$$\delta_k(\theta, S) := \left| \frac{\partial^{k+1}}{\partial \theta^k \partial S} M(\theta; S) \right| \tag{5.5}$$

can be computed for both the scattering profile ($k = 0$) and its derivative(s) ($k > 0$), the variances of $M^*(\theta; S)$ and $\frac{\partial^k}{\partial \theta^k} M^*(\theta; S)$ are different. However, if the nonparametric regression estimator used to compute $M^*(\theta; S)$ for the unknown configuration is linear in the observations (i.e. satisfies (1.4)) and self-consistent (i.e. satisfies (3.5)) then we can determine the variance of the estimator $M^*(\theta; S)$ relative to the variance of the estimator $\frac{\partial^k}{\partial \theta^k} M^*(\theta; S)$. For example, with $k = 1$ the variances will be

$$Var[M^*(\theta; S)] = \sum_{i=1}^{T} l_i^2(\theta) \sigma^2, \tag{5.6}$$

and

$$Var\left[\frac{\partial}{\partial \theta} M^*(\theta; S)\right] = \sum_{i=1}^{T} l_i'^2(\theta) \sigma^2, \tag{5.7}$$

where σ^2 is the (constant) variance of the data from (5.1), T is the cardinality of $\mathcal{T}$, and the l_i and l_i' are defined by the nonparametric regression technique being employed and depend on $\mathcal{T}$. A useful fact about these variances is that if we divide $\delta_0(\theta, S)$ by the square root of $\sum_{i=1}^{T} l_i^2(\theta)$ and divide $\delta_1(\theta, S)$ by the square root of $\sum_{i=1}^{T} l_i'^2(\theta)$ then the results, call them $\widetilde{\delta}_0$ and $\widetilde{\delta}_1$, are on the same scale and are both measures of absolute characterization ability. Therefore we propose the following modification of (5.4):

$$\Delta_k(\theta, S) := \frac{\widetilde{\delta}_k(\theta, S)}{\sum_{k \le K} \sum_{t \in \mathcal{T_I}} \widetilde{\delta}_k(t, S) / (|\mathcal{T}_I| \times K)}, \tag{5.8}$$

where $|\mathcal{T}_I|$ indicates the cardinality of $\mathcal{T}_I$. The obstacle now is that (5.6) and (5.7) depend on $\mathcal{T}$ which is what we are trying to determine.

To get around this obstacle we propose an iterative procedure. We first define $\mathcal{T}_0$ to be a grid which covers $\mathcal{X}$ and is as dense as we ultimately want $\mathcal{T}$ to be. Begin by using $\mathcal{T}_0$ in (5.6) and (5.7). After normalization, we can then determine $\mathcal{T}_1 \subset \mathcal{T}_0$. (The mechanism for determining $\mathcal{T}_1$ from $\mathcal{T}_0$ is spelled out below). We then use $\mathcal{T}_1$ in (5.6) and (5.7) and determine $\mathcal{T}_2$. A few iterations usually results in convergence.

We have addressed how to account for differing levels of characterization difficulty between sizes and how to handle both the scattering profile and its derivative(s). We now describe how at each iteration we make the next choice for $\mathcal{T}$. Our approach is to define the following threshold:

$$M := \min_{S \in \mathcal{C}_2} \max_{\theta \in \mathcal{T}_I} \max_{k \leq K} \Delta_k(\theta, S), \tag{5.9}$$

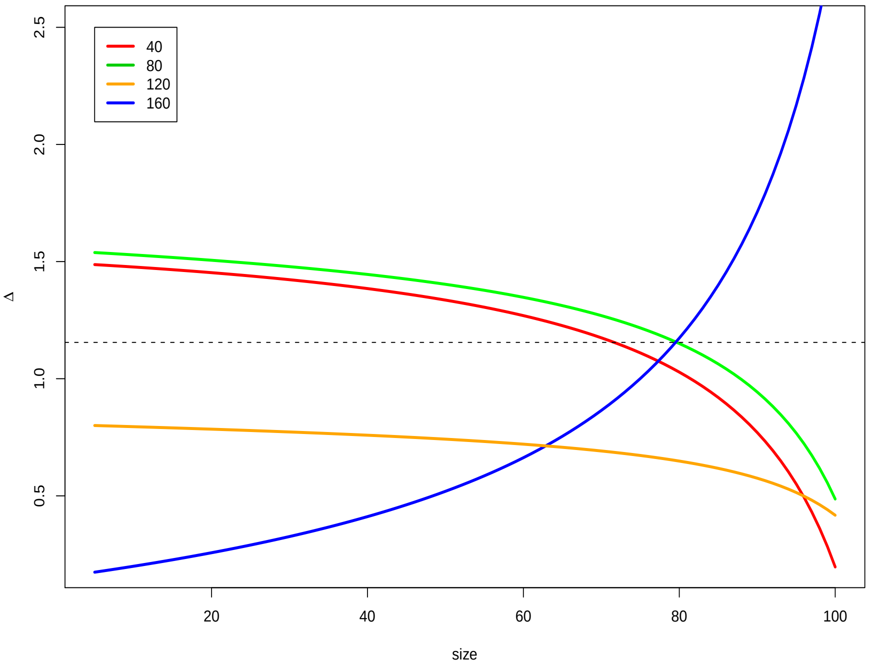
where K represents the largest derivative under consideration. (See Figure 5.5 for a graphical illustration of the choice of M.) As mentioned above, Δ_k, and therefore M, depend on $\mathcal{T}$ which is why the iterative approach is necessary. At iteration i we include in $\mathcal{T}_i$ any $\theta \in \mathcal{T}_I$ for which $\Delta_k(\theta, S)$ reaches or exceeds the threshold M for some $S \in \mathcal{C}_2$ and $k \in \{0, 1, ..., K\}$. In addition, for every $\theta \in \mathcal{T}_0$ which is not in $\mathcal{T}_I$, we determine the closest $\theta_I \in \mathcal{T}_I$ to θ. If θ_I is included in $\mathcal{T}_i$ then we include θ as well. For example, if $\mathcal{T}_I = \{10, 20, ..., 170\}$, $\mathcal{T}_0 = \{1, 2, ..., 179\}$, and we include $\theta_I = 20$, then we would also include $\theta = \{16, 17, 18, 19, 21, 22, 23, 24, 25\}$.

The researcher may wish to perform characterization using several derivatives. When selecting $\mathcal{T}$, however, we recommend using at most one derivative. Due to the fact that $\mathcal{T}_I$ is by definition 'sparse', we should not expect to obtain good estimates of high order derivatives at this stage. This does not mean that high order derivatives cannot be used to save resources. In fact we do this in the next section for the inverse problem.

We now illustrate the selection of $\mathcal{T}$ with a couple of practical examples. We apply our methodology to characterize based on size using the M_{11} and M_{33} profiles where the agglomeration level is set at 50% and we take $\mathcal{T}_I = \{10, 20, ..., 170\}$, $\mathcal{T}_0 = \{1, 2, 3, ..., 179\}$ and $K = 1$. Figure 5.6 displays the first derivatives of the M_{11} scattering profiles depicted in Figure 5.1. Figures 5.7 and 5.8 display the scattering profiles and first derivatives for M_{33}.

For M_{11}, Table 5.1 displays how we arrive at a threshold of $M = 3.47$ for the first

Figure 5.5: Threshold for selecting limited angles



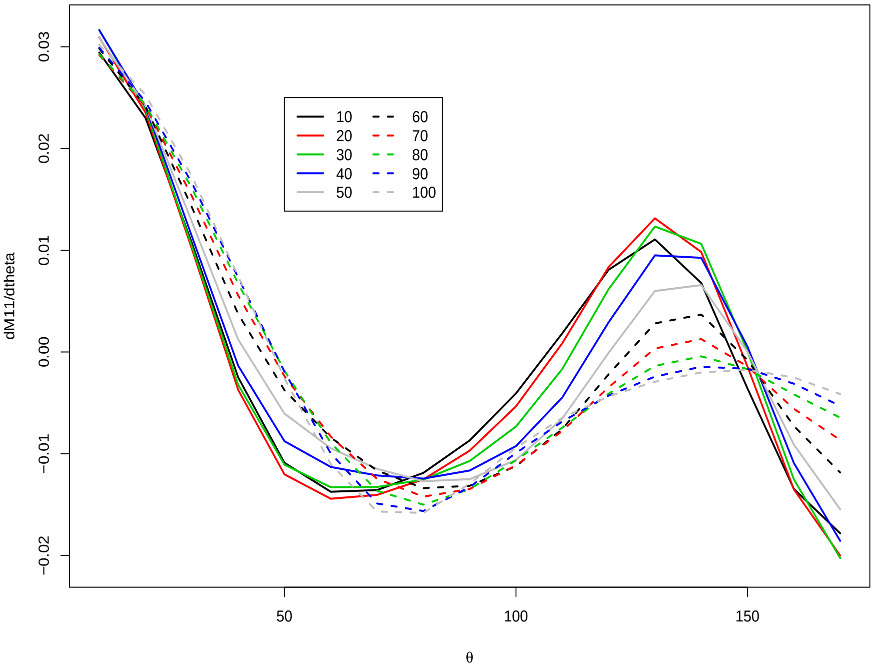
The solid curves represent hypothetical $\Delta(\theta, S)$ quantities as a function of S. Each curve represents a different θ. The dashed black curve represents the threshold. In this situation we would include 40, 80, and 160 and exclude 120 from $\mathcal{T}$.

iteration. This implies that we should include

$$\{70, 80, 90, 100, 110, 120, 130, 140, 150, 160\}$$

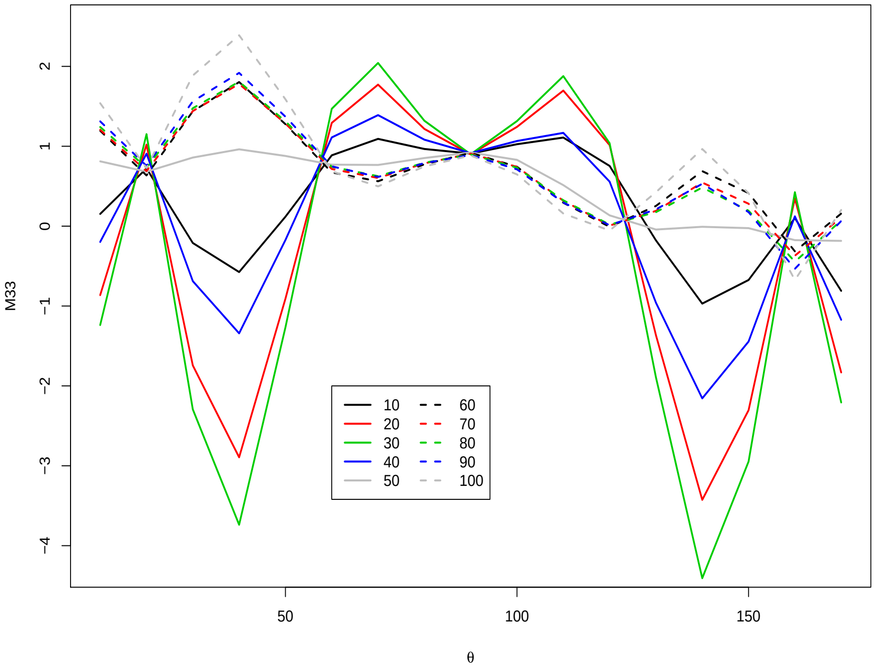
from $\mathcal{T}_I$ in $\mathcal{T}_1$ and hence we obtain $\mathcal{T}_1 = \{66, 67, ..., 165\}$. We achieve convergence after the second iteration with $\mathcal{T}_2 = \{76, 77, ..., 165\} = \mathcal{T}$. So rather than collecting data at 179 points, we need only examine the 17 points of $\mathcal{T}_I$ and the 90 points of $\mathcal{T}$. This is a reduction by more than 40%. For M_{33} we obtain $\mathcal{T}_1 = \{26, 27, ..., 55\} \cup \{136, 137, ..., 165\} = \mathcal{T}_2 = \mathcal{T}$ which represents a reduction in the number of data points by more than 55%.

Figure 5.6: M11 first derivatives of scattering profiles



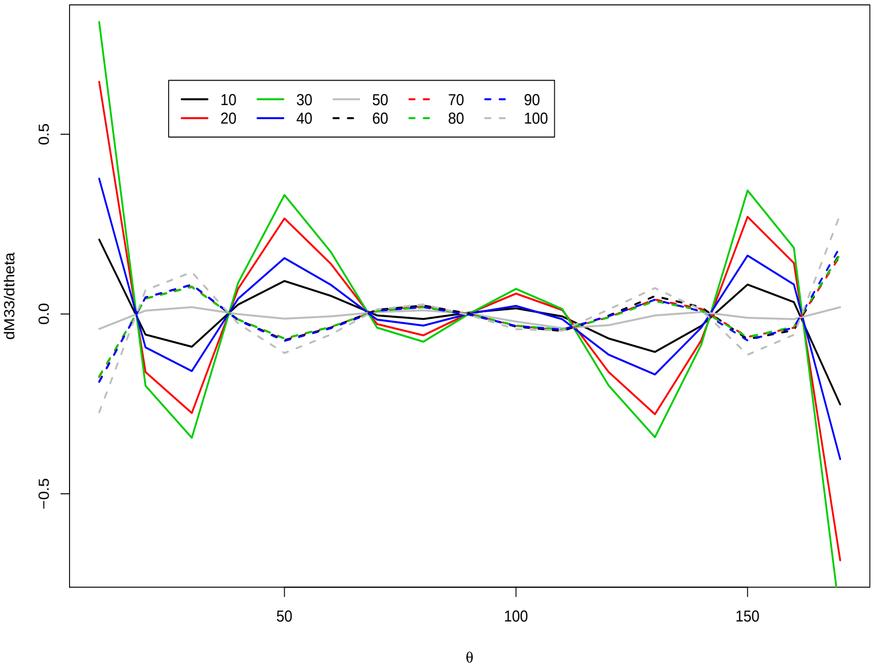
The figure displays the first derivative of the M_{11} scattering profiles. Each curve represents a different size (in nm). The curves were constructed with $\mathcal{T}_I = \{10, 20, ..., 170\}$ using compound estimation with filtration and extrapolation and $J = 7$, 27 centering points, $\beta = 12$ (120 during filtration and extrapolation), and $\kappa = 0.2$. Local constant and slope estimates were obtained using smoothing splines with splines of order 7, while coefficients of higher-order local fits were obtained using inductive estimators with $h_2 = h_3 = 1/150$, $h_4 = h_5 = h_6 = h_7 = 1/30$. Agglomeration level is set at 50%.

Figure 5.7: M33 scattering profiles



The graph above presents M_{33} scattering profiles as a function of θ. Each curve represents a different size (in nm). The curves were constructed with $\mathcal{T}_I = \{10, 20, ..., 170\}$ using compound estimation with filtration and extrapolation and $J = 7$, 27 centering points, $\beta = 12$ (120 during filtration and extrapolation), and $\kappa = 0.2$. Local constant and slope estimates were obtained using smoothing splines with splines of order 7, while coefficients of higher-order local fits were obtained using inductive estimators with $h_2 = h_3 = 1/150$, $h_4 = h_5 = h_6 = h_7 = 1/30$. Agglomeration level is set at 50%.

Figure 5.8: M33 first derivatives of scattering profiles



The figure displays the first derivative of the M_{33} scattering profiles. Each curve represents a different size (in nm). The curves were constructed with $\mathcal{T}_I = \{10, 20, ..., 170\}$ using compound estimation with filtration and extrapolation and $J = 7$, 27 centering points, $\beta = 12$ (120 during filtration and extrapolation), and $\kappa = 0.2$. Local constant and slope estimates were obtained using smoothing splines with splines of order 7, while coefficients of higher-order local fits were obtained using inductive estimators with $h_2 = h_3 = 1/150$, $h_4 = h_5 = h_6 = h_7 = 1/30$. Agglomeration level is set at 50%.

Table 5.1: Determining the threshold for M11

	θ																
S	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170
5	.65	1.04	.82	.03	1.39	2.47	3.01	3.37	**3.50**	3.27	2.51	1.20	.85	2.49	3.15	2.90	1.17
10	.14	.58	.41	.23	1.37	2.35	3.13	3.88	4.40	**4.56**	4.24	3.53	2.05	.05	1.14	1.42	.34
15	.34	.12	.07	.33	1.00	1.54	2.06	2.72	3.46	4.18	4.70	**4.98**	4.42	2.39	.87	.17	.49
20	.84	.27	.12	.27	.36	.03	.36	.34	.53	2.28	4.38	6.30	**6.98**	4.83	2.82	1.69	1.38
25	.74	.45	.17	.07	.76	1.96	3.26	4.14	3.80	2.13	.30	2.69	**4.21**	3.52	2.47	1.80	1.31
30	.67	.63	.33	.14	1.20	2.54	3.81	**4.71**	4.62	3.43	1.53	.47	2.13	2.41	2.13	1.83	1.22
35	.83	.93	.60	.05	1.32	2.61	3.60	**4.22**	4.07	3.15	1.65	.06	1.80	2.51	2.58	2.39	1.47
40	1.02	1.23	.85	.08	1.29	2.51	3.24	**3.57**	3.32	2.52	1.31	.20	1.96	2.87	3.12	2.99	1.76
45	1.18	1.49	1.06	.22	1.21	2.39	2.96	3.09	2.75	2.04	1.02	.31	2.06	3.10	**3.50**	3.44	1.98
50	1.32	1.73	1.26	.35	1.11	2.26	2.73	2.73	2.34	1.70	.85	.34	2.07	3.23	3.79	**3.81**	2.18
55	1.46	1.94	1.41	.47	.96	2.06	2.44	2.36	1.94	1.39	.69	.39	2.13	3.42	4.13	**4.24**	2.38
60	1.62	2.12	1.52	.57	.78	1.76	2.03	1.84	1.41	.97	.44	.56	2.36	3.77	4.61	**4.79**	2.64
65	1.89	2.41	1.71	.69	.61	1.38	1.33	.86	.37	.11	.14	1.00	2.89	4.42	5.37	**5.55**	3.03
70	1.83	2.23	1.64	.73	.35	.65	.03	1.02	1.58	1.45	1.14	1.48	2.94	4.17	4.89	**4.93**	2.75
75	1.53	1.77	1.40	.68	.14	.03	1.15	2.54	3.11	2.67	1.88	1.70	2.56	3.32	**3.70**	3.57	2.09
80	1.38	1.45	1.16	.58	.03	.34	1.73	3.22	**3.75**	3.16	2.16	1.77	2.38	2.93	3.13	2.91	1.76
85	1.45	1.21	.74	.28	.06	.71	1.98	3.17	**3.48**	2.87	2.00	1.77	2.52	3.14	3.41	3.22	1.83
90	1.72	.98	.03	.38	.18	1.31	2.17	2.54	2.33	1.80	1.40	1.69	2.87	3.74	**4.28**	4.26	2.14
95	1.69	1.04	.04	.45	.20	1.59	2.71	3.18	2.87	2.19	1.67	1.80	2.61	3.12	**3.47**	3.46	1.73
100	1.25	1.28	.91	.35	.09	1.10	2.99	4.65	**4.97**	4.04	2.76	2.02	1.93	1.77	1.56	1.27	.92

The table presents values of $\max_{k \leq 1} \Delta_k(\theta, S)$ for $\theta \in \mathcal{T}_I = \{10, 20, ..., 170\}$ and $S \in \mathcal{C}_2$ for the M_{11} scattering profile. The bolded values represent $\max_{\theta \in \mathcal{T}_I} \max_{k \leq 1} \Delta_k(\theta, S)$ for $S \in \mathcal{C}_2$. The threshold is $M = \min_{S \in \mathcal{C}_2} \max_{\theta \in \mathcal{T}_I} \max_{k \leq 1} \Delta_k(\theta, S) = 3.47$.

5.3 Limited angles for particles of unknown size

The previous section describes how to solve the forward problem of nanoparticle characterization using fewer observations than previous methods by choosing the sample strategically. In this section we shift to the inverse problem and describe how fewer observations can be used at this stage as well.

We first note that if $\mathcal{T} \subset \mathcal{X}$ is identified for the construction of the reference curves, then attention can also be confined to $\mathcal{T}$ for the construction of scattering profiles for unknown configurations. Since $\mathcal{T}$ excludes regions of $\mathcal{X}$ irrelevant for characterization, there is no need to examine observations outside of $\mathcal{T}$ to solve the inverse problem. Thus, it may be tempting to simply accept $\mathcal{T}$ as the limited subset of $\mathcal{X}$ for the inverse problem in addition to the forward problem. There are, however, several reasons why a different method is desirable for choosing a limited subset for the inverse problem.

First, the scientist may have constructed the reference curves using a $\mathcal{T}$ not selected by the method of the previous section. The total number of observations required to construct the reference curves is $|\mathcal{T}| \times |\mathcal{C}_2|$, but even if this number is very large, the reference curves only have to be constructed once. On the other hand, a scattering profile has to be constructed for every unknown configuration. Since this task will be performed repeatedly, a scientist who did not restrict $\mathcal{T}$ for the forward problem may wish to impose a restriction for the inverse problem. Second, rather than having no knowledge about the configuration of particles of unknown size, the researcher may have limited knowledge about the configuration. For example, a scientist may be confident that nanoparticles are smaller than 50 nm and wish to characterize the particles more precisely. In this case, we would like to exploit the researcher's a priori knowledge to further restrict the angles to those relevant at sizes below 50 nm. Third, aside from the savings in time and resources, the identification of a limited subset of $\mathcal{X}$ for the construction of scattering profiles can be used to overcome boundary issues, which we demonstrate below.

Our approach is to partition $\mathcal{T} \subset \mathcal{X}$ into intervals and evaluate the characterization ability of each interval for the sizes under consideration. The intervals with high characterization ability are then combined to form $\mathcal{T}^*$, the set from which observations are collected for the unknown configuration. To do this we define 'high characterization ability' in terms of $\widetilde{\delta}_k(\theta, S)$.

There are several reasons we use $\widetilde{\delta}_k$ rather than Δ_k from (5.8): the definition of Δ_k would have to be modified to even be possible in this situation, $\widetilde{\delta}_k$ can be viewed as a measure of absolute characterization ability which allows for easier interpretation and comparison, and if we assume a priori that the unknown size of the nanoparticle is restricted to some $\mathcal{C}_S \subset \mathcal{C}_0$ then the pitfall situations which Δ_k was designed to overcome (see Figure 5.3) become less likely. Finally, it is possible to perform a normalization analogous to (5.8) on the results of this section even after using $\widetilde{\delta}_k$.

Since $\widetilde{\delta}_k(\theta, S)$ is a measure of the ability to perform characterization at angle θ for particles of approximate size S using the k^{th} derivative of the scattering profile,

$$B_k := \left(\int_{S_1}^{S_2} \int_{\theta_1}^{\theta_2} \widetilde{\delta}_k(\theta, S) d\theta dS \right) / (S_2 - S_1) \tag{5.10}$$

measures the ability to characterize particles whose size is between S_1 and S_2 with the k^{th} derivative of the scattering profile over $[\theta_1, \theta_2]$. Regions of $\mathcal{X}$ for which B_k is large correspond to regions over which characterization via derivative k is best performed. In practice, (5.10) will be approximated by a Riemann sum since $\widetilde{\delta}_k(\theta, S)$ is only available on the discrete set $\mathcal{T}$. Note that since $\widetilde{\delta}_k(\theta, S)$ is scaled by the standard deviation of $\frac{\partial^{k+1}}{\partial\theta^k \partial S} M^*(\theta; S)$, it is possible to compare the characterization ability between derivative(s) as well as between regions of $\mathcal{T}$.

Once we have evaluated the characterization ability of each subset of $\mathcal{T}$ we must decide which subsets to include in $\mathcal{T}^*$. We recommend taking $\mathcal{T}^*$ to be the union of the $p\%$ of intervals which have the highest B_k. The choice of p will depend on how much the researcher wants to narrow $\mathcal{T}$. If $\mathcal{T}$ is already substantially limited, p should be large. On the other hand, if the researcher wants $\mathcal{T}^*$ to be considerably smaller than $\mathcal{T}$, p should be small.

We now show how the choice of $\mathcal{T}^*$ can be used to overcome boundary issues. Characterization entails minimizing an integrated squared discrepancy on $\mathcal{T}^*$. The estimate of the unknown configuration's scattering profile and its derivative(s) on $\mathcal{T}^*$ are inputs for the discrepancy. If, as with previous methods (Charnigo et al 2007, Charnigo et al 2010), the choice of $\mathcal{T}^*$ contains values of θ near the boundaries of $\mathcal{X}$, then it is inherently difficult to estimate the scattering profile, and especially its derivatives, near the boundaries of $\mathcal{T}^*$. Poor estimates can then contribute to poor characterization. However, if $\mathcal{T}^* \subset \mathcal{T}$ does not contain values of θ near the boundaries of $\mathcal{X}$ then we could define $\overline{\mathcal{T}}^*$ to be the union of $\mathcal{T}^*$ and j points extending past each boundary of $\mathcal{T}^*$. (In our experience setting $j = 7$ has worked well.) We then recommend collecting observations from $\overline{\mathcal{T}}^*$ but performing characterization only over $\mathcal{T}^*$. Hence the boundary problem of estimation is pushed outside of the range over which characterization is performed.

Of course it is possible that a boundary of $\mathcal{T}^*$ will coincide with that of $\mathcal{X}$. If the scientist is unwilling to accept the boundary problem in this case he/she should refine $\mathcal{T}^*$ by excluding the j points nearest to the boundary of $\mathcal{X}$. These points are then used in estimation but not characterization.

We now illustrate with a couple of examples. Recall that for the M_{11} scattering profiles we determined $\mathcal{T} = \{76, 77, ..., 165\}$. Tables 5.2 and 5.3 display values of B_k for various subsets of $\mathcal{T}$ and various ranges of S. The generally high magnitude of B_0 compared to B_1 in Table 5.2 indicates that for the M_{11} scattering profile the mean response is much better at characterization than the first derivative. Similarly, the first derivative appears to be much more useful than the second derivative. Also note that the generally smaller entries in the last row for each B_k indicate that particles of size 75 to 100 nm are more difficult to characterize than smaller nanoparticles.

If we take p to be 78% and consider the mean response, then we would take $\mathcal{T}^* = \{76, 77, ..., 145\}$ to characterize particles of size 5 to 25 nm, $\mathcal{T}^* = \{76, 77, ..., 105\} \cup \{126, 127, ..., 165\}$ to characterize particles of size 25 to 75 nm, and
$\mathcal{T}^* = \{86, 87, ..., 105\} \cup \{116, 117, ..., 165\}$ to characterize particles of size 75 to 100 nm. Interestingly, if we assume less prior knowledge (i.e. consider wider $[S_1, S_2]$ intervals), we would take $\mathcal{T}^* = \{76, 77, ..., 105\} \cup \{126, 127, ..., 165\}$ to characterize particles of size 5 to 50 nm and to characterize particles of size 50 to 100 nm, again

only considering the characterization ability of the mean response.

Table 5.2: Evaluation of inverse problem limited angles for M_{11}

	B_0								
	θ_1 representing $\{\theta_1, \theta_1+1, ..., \theta_1+9\}$								
$[S_1, S_2]$	76	86	96	106	116	126	136	146	156
[5,25]	.131	.149	.199	.212	.191	.162	.146	.124	.100
[25,50]	.145	.134	.116	.058	.045	.116	.158	.141	.121
[50,75]	.067	.060	.054	.029	.047	.119	.183	.187	.172
[75,100]	.056	.058	.062	.057	.064	.079	.098	.095	.085

	B_1								
	θ_1 representing $\{\theta_1, \theta_1+1, ..., \theta_1+9\}$								
$[S_1, S_2]$	76	86	96	106	116	126	136	146	156
[5,25]	.010	.015	.022	.038	.050	.057	.048	.027	.016
[25,50]	.003	.017	.033	.053	.059	.045	.021	.009	.012
[50,75]	.005	.008	.013	.028	.045	.052	.036	.011	.016
[75,100]	.004	.003	.006	.005	.008	.016	.014	.004	.009

	B_2								
	θ_1 representing $\{\theta_1, \theta_1+1, ..., \theta_1+9\}$								
$[S_1, S_2]$	76	86	96	106	116	126	136	146	156
[5,25]	.002	.005	.007	.011	.013	.008	.009	.015	.010
[25,50]	.004	.007	.007	.005	.004	.008	.011	.008	.004
[50,75]	.001	.002	.003	.007	.007	.002	.008	.014	.008
[75,100]	.002	.003	.001	.004	.005	.003	.003	.007	.004

The columns represent values of θ_1, θ_2 is set to $\theta_1 + 9$, and the rows represent values of size over which B_k is computed. Values are tabulated for the mean response (B_0), the first derivative (B_1), and the second derivative (B_2).

While the characterization ability of the mean response appears to be greater than that of the derivatives, the characterization ability of the derivatives need not be disregarded. Charnigo et al (2010) demonstrate how characterization can be performed using the mean response and its derivatives together. In this case, the characterization ability of a an interval $[\theta_1, \theta_2]$ could be measured by $B_0 + B_1 + B_2$. Table 5.4 displays this quantity for the M_{11} scattering profiles. If we take p to be 78% then we would take $\mathcal{T}^* = \{86, 87, ..., 155\}$ to characterize particles of size 5 to 25 nm using the mean response and its first two derivatives simultaneously, $\mathcal{T}^* = \{76, 77, ..., 105\} \cup \{126, 127, ..., 165\}$ to characterize particles of size 25 to 50 nm, $\mathcal{T}^* = \{76, 77, ..., 95\} \cup \{116, 117, ..., 165\}$ to characterize particles of size 50

Table 5.3: Further evaluation of inverse problem limited angles for M_{11}

	B_0								
	θ_1								
$[S_1, S_2]$	76	86	96	106	116	126	136	146	156
[5,50]	.137	.138	.147	.119	.103	.137	.158	.141	.120
[50,100]	.061	.061	.062	.050	.061	.092	.125	.124	.113

	B_1								
	θ_1								
$[S_1, S_2]$	76	86	96	106	116	126	136	146	156
[5,50]	.006	.016	.028	.046	.055	.052	.034	.016	.014
[50,100]	.005	.005	.009	.014	.022	.028	.021	.007	.011

	B_1								
	θ_1								
$[S_1, S_2]$	76	86	96	106	116	126	136	146	156
[5,50]	.003	.006	.007	.008	.008	.008	.010	.012	.007
[50,100]	.002	.003	.002	.005	.006	.003	.004	.009	.005

The columns represent values of θ and the rows represent values of size over which B_k is computed. Values are tabulated for the mean response (B_0), the first derivative (B_1), and the second derivative (B_2).

to 75 nm, and $\mathcal{T}^* = \{96, 97, ..., 165\}$ to characterize particles of size 75 to 100 nm. If we assume less prior knowledge (i.e. consider wider $[S_1, S_2]$ intervals), we would take $\mathcal{T}^* = \{86, 87, ..., 155\}$ to characterize particles of size 5 to 50 nm and $\mathcal{T}^* = \{86, 87, ..., 105\} \cup \{116, 117, ..., 165\}$ to characterize particles of size 50 to 100 nm.

Table 5.4: Using the mean response and two derivatives together for M_{11}

	$B_0 + B_1 + B_2$								
	θ_1								
$[S_1, S_2]$	76	86	96	106	116	126	136	146	156
[5,25]	.144	.168	.228	.261	.253	.228	.204	.165	.126
[25,50]	.152	.158	.155	.116	.108	.169	.190	.159	.138
[50,75]	.073	.070	.070	.064	.099	.173	.226	.213	.197
[75,100]	.062	.065	.069	.065	.078	.097	.115	.106	.098

The columns represent values of θ_1, θ_2 is set to $\theta_1 + 9$, and the rows represent values of size over which B_k is computed. The entries are the sums $B_0 + B_1 + B_2$ for the corresponding values of size and angular intervals.

As a second example which illustrates that the methodology of this section can be applied even if the scientist did not employ the methodology of Section 5.2, consider taking $\mathcal{T} = \{1, 2, ..., 179\}$ for the M_{33} scattering profiles. To prevent potential boundary issues, exclude the 7 points closest to each boundary. This leaves the set $\{8, 9, ..., 172\}$ under consideration for $\mathcal{T}^*$. Tables 5.5 and 5.7 display values of B_k for various subsets of $\mathcal{T}$ and size intervals. Derivatives appear much more important for the M_{33} scattering profiles than they did for the M_{11} profiles. If we consider the mean response and take $p = 64\%$ we get $\mathcal{T}^* = \{23, 24, ..., 52\} \cup \{98, 99, ..., 172\}$ for nanoparticles of size 5 to 25 nm, $\mathcal{T}^* = \{23, 24, ..., 67\} \cup \{98, 99, ..., 157\}$ for nanoparticles of size 25 to 50 nm, and $\mathcal{T}^* = \{8, 9, ..., 67\} \cup \{128, 129, ..., 172\}$ for nanoparticles of size 50 to 100 nm. If the mean response and first two derivatives are considered together as in Table 5.6 we obtain $\mathcal{T}^* = \{38, 39, ..., 52\} \cup \{83, 84, ..., 172\}$ for nanoparticles of size 5 to 25 nm, $\mathcal{T}^* = \{38, 39, ..., 67\} \cup \{98, 99, ..., 172\}$ for nanoparticles of size 25 to 50 nm, $\mathcal{T}^* = \{8, 9, ..., 37\} \cup \{53, 54, ..., 67\} \cup \{113, 114, ..., 172\}$ for nanoparticles of size 50 to 75 nm, and $\mathcal{T}^* = \{7, 8, ..., 37\} \cup \{68, 69, ..., 82\} \cup \{113, 114, ..., 172\}$ for nanoparticles of size 75 to 100 nm.

Table 5.5: Evaluation of inverse problem limited angles for M_{33}

	B_0										
	θ_1										
$[S_1, S_2]$	8	23	38	53	68	83	98	113	128	143	158
[5,25]	.063	.108	.119	.099	.075	.055	.398	.849	.765	.475	.153
[25,50]	.144	.215	.238	.240	.128	.030	.164	.545	.628	.395	.105
[50,75]	.228	.251	.227	.219	.105	.028	.017	.106	.351	.431	.238
[75,100]	.168	.153	.126	.125	.067	.035	.030	.041	.133	.365	.352
	B_1										
	θ_1										
$[S_1, S_2]$	8	23	38	53	68	83	98	113	128	143	158
[5,25]	.052	.022	.024	.039	.017	.170	.422	.233	.132	.215	.197
[25,50]	.060	.023	.012	.062	.105	.086	.239	.287	.110	.208	.164
[50,75]	.034	.020	.029	.056	.124	.056	.049	.174	.195	.085	.148
[75,100]	.012	.026	.020	.024	.075	.058	.060	.063	.205	.124	.088
	B_2										
	θ_1										
$[S_1, S_2]$	8	23	38	53	68	83	98	113	128	143	158
[5,25]	.015	.026	.021	.016	.034	.171	.102	.196	.143	.042	.008
[25,50]	.023	.025	.019	.036	.023	.070	.136	.128	.194	.043	.019
[50,75]	.031	.022	.006	.032	.029	.089	.077	.135	.105	.133	.027
[75,100]	.015	.008	.009	.016	.025	.056	.055	.119	.080	.120	.081

The columns represent values of θ_1, θ_2 is set to $\theta_1 + 14$, and the rows represent values of size over which B_k is computed. Values are tabulated for the mean response (B_0), the first derivative (B_1), and the second derivative (B_2).

Table 5.6: Using the mean response and two derivatives together for M_{33}

	$B_0 + B_1 + B_2$										
	θ_1										
$[S_1, S_2]$	8	23	38	53	68	83	98	113	128	143	158
[5,25]	.13	.16	.16	.06	.13	.39	.92	1.28	1.04	.73	.36
[25,50]	.23	.26	.27	.34	.26	.19	.54	.96	.93	.65	.29
[50,75]	.29	.29	.26	.31	.26	.17	.14	.42	.65	.65	.41
[75,100]	.20	.19	.16	.17	.17	.15	.15	.22	.42	.61	.52

The columns represent values of θ_1, θ_2 is set to $\theta_1 + 9$, and the rows represent values of size over which B_k is computed. The entries are the sums $B_0 + B_1 + B_2$ for the corresponding values of size and angular intervals.

Table 5.7: Further evaluation of inverse problem limited angles for M_{33}

	B_0										
	θ_1										
$[S_1, S_2]$	8	23	38	53	68	83	98	113	128	143	158
[5,50]	.12	.18	.19	.18	.11	.04	.24	.63	.66	.43	.13
[50,100]	.18	.18	.15	.15	.08	.03	.03	.06	.21	.37	.30
	B_1										
	θ_1										
$[S_1, S_2]$	8	23	38	53	68	83	98	113	128	143	158
[5,50]	.06	.02	.02	.05	.07	.12	.29	.26	.12	.20	.18
[50,100]	.02	.02	.02	.03	.09	.06	.06	.11	.20	.11	.11
	B_2										
	θ_1										
$[S_1, S_2]$	8	23	38	53	68	83	98	113	128	143	158
[5,50]	.02	.03	.02	.03	.03	.11	.12	.15	.17	.05	.01
[50,100]	.02	.01	.01	.02	.03	.07	.07	.12	.09	.12	.06

The columns represent values of θ_1, θ_2 is set to $\theta_1 + 14$, and the rows represent values of size over which B_k is computed. Values are tabulated for the mean response (B_0), the first derivative (B_1), and the second derivative (B_2).

5.4 Simulation studies

In the previous sections we described methods to determine $\mathcal{T}_M$ and $\mathcal{T}^*_M$, limited subsets of $\mathcal{X}$ for constructing the reference curves and the scattering profiles for unknown configurations, respectively. (The subscript M indicates that the methodology of the previous sections was employed.) In this section we evaluate the characterization ability of these regions empirically. We conducted two simulation studies: one to evaluate the limited angles for the reference curves and one to evaluate the limited angles for the scattering profiles of unknown configurations. Note that the latter can incorporate prior knowledge about the unknown configuration which the former does not.

Simulation 1 was designed to evaluate our methodology's choice of $\mathcal{T}_M$. We examined the M_{11} and M_{33} scattering profiles with agglomeration level set at 50% with the goal of inferring nanoparticle size. Recall our method's suggestions of ${}_{11}\mathcal{T}_M = \{76, 77, ..., 165\}$ for the former and ${}_{33}\mathcal{T}_M = \{26, 27, ..., 55\} \cup \{136, 137, ..., 165\}$ for the latter. To mitigate boundary issues we used observations from ${}_{11}\overline{\mathcal{T}}_M = \{69, 70, ..., 172\}$ and ${}_{33}\overline{\mathcal{T}}_M = \{19, 20, ..., 62\} \cup \{129, 130, ..., 172\}$. We then compared how well characterization was performed over $\mathcal{T}_M$ as compared to over $\mathcal{T}_0 = \{1, 2, ..., 179\}$ where in each case we set $\mathcal{T}^* := \mathcal{T}$. For each of M_{11} and M_{33} we performed 4 sets of experiments: using the mean response only, using the first derivative only, using the second derivative only, and using the mean response and the first two derivatives simultaneously. Each set of experiments was performed twice: once with low-noise and once with high noise. We performed characterization by defining the estimate of the nanoparticle's size, c^*, to be the $c \in \mathcal{C}_1$ for which

$$D[M^*(\theta; u), M(\theta; c), K] := \frac{\int_{\mathcal{T}^*} [\frac{\partial^K}{\partial^K \theta} M^*(\theta; u) - \frac{\partial^K}{\partial^K \theta} M(\theta; c)]^2 d\theta}{\int_{\mathcal{T}^*} [\frac{\partial^K}{\partial^K \theta} M^*(\theta; u)]^2 d\theta}, \tag{5.11}$$

was minimal, where $M^*(\theta; u)$ denotes the estimate of the scattering profile for the particle with unknown size. To perform characterization using the mean response and first two derivatives simultaneously we sought to minimize

$$\sum_{K=0}^{2} D[M^*(\theta; u), M(\theta; c), K].$$

To carry out the simulations we used previously validated Fortran code (Venkata et al) to generate $M_{11}^{(d)}(\theta; s)$ and $M_{33}^{(d)}(\theta; s)$ for $d \in \{1, 2, ..., 14\}$, $s \in \mathcal{S} = \{10, 20, ..., 90\}$ and $\theta \in \{1, 2, ..., 179\}$ where d refers to systematic disturbances to the physical parameters described below and s refers to size. The systematic disturbances reflect the nature of the laboratory setting. We then added random noise as in model (5.1) to obtain $M_{11}^{(d,r_d)}(\theta; s)$ and $M_{33}^{(d,r_d)}(\theta; s)$ for $d \in \{1, 2, ..., 14\}$, $r_d \in \{1, 2, ..., 15\}$, $s \in \{10, 20, ..., 90\}$, and $\theta \in \{1, 2, ..., 179\}$. The random errors were assumed to be independently and identically normally distributed with mean zero. For the low noise setting and the M_{11} scattering profile, the standard deviation was set to 3% of the standard deviation of $M_{11}^{(d)}(\theta; s)$ over θ. For the high noise we instead used 6%. Noise was added analogously for M_{33}. For each simulated scattering profile we then classified the size of the nanoparticles once using data from $\mathcal{T}_0 := \{1, 2, ..., 179\}$ and once using data only from the recommended $\overline{\mathcal{T}}_M$ and compared the results. Recall that the goal of $\mathcal{T}_M$ was to save resources by having a reduced cardinality while at the same time sacrificing little in terms of characterization ability. To provide a comparison we also performed characterization over three naive choices with reduced cardinality as described in Table 5.8.

Comparisons were made on the basis of Root Mean Square Error defined as

$$\sqrt{\frac{1}{1890} \sum_{s \in \mathcal{S}} \sum_{d=1}^{14} \sum_{r_d=1}^{15} [c^*(s, d, r_d) - s]^2}, \tag{5.12}$$

where $c^*(s, d, r_d)$ is the estimate of the nanoparticle size when the true size is s with disturbance d and noise pattern r_d imposed. The results are displayed in Table 5.8.

Physical Parameters. The radiation beam is assumed to have a wavelength of 514.5 nm (argon-ion laser) with an angle of incidence of 23 degrees, the prism and substrate are both made of sapphire with a refractive index of 1.77304, the substrate is coated with a 20 nm gold thin film with a complex refractive index of $0.50 + 1.86i$, and the scatterers are gold spherical particles ($0.50 + 1.86i$) with 50% agglomeration.

Disturbances. The following is a list of the disturbances in our simulation study.

1. angle of incidence of 24 degrees
2. angle of incidence of 22 degrees

3. angle of incidence of 25 degrees

4. angle of incidence of 21 degrees

5. +1 degree offset in measurement angle

6. -1 degree offset in measurement angle

7. solid angle of 3 degrees in far field measurement

8. angle of incidence of 21 degrees, +1 degree offset in measurement angle

9. angle of incidence of 25 degrees, -1 degree offset in measurement angle

10. incident beam spread over a solid angle (23 and 24 degrees)

11. incident beam spread over a solid angle (23 and 24 degrees), solid angle of 3 degrees in far field measurement

12. incident beam spread over a solid angle (22, 23, and 24 degrees), solid angle of 2 degrees in far field measurement

13. incident beam spread over a solid angle (22 and 23 degrees), +1 degree offset in measurement angle

14. incident beam spread over a solid angle (22 and 23 degrees) -1 degree offset in measurement angle

The results of Simulation 1 indicate that our choice of $\mathcal{T}_M$ accomplished the task of selecting a limited range of angles over which characterization can be performed effectively. Inflation of the RMSE is expected when we use fewer data points, but we tolerate small inflation in exchange for the time and resource savings. The inflation of the RMSE in using $\mathcal{T}_M$ rather than $\mathcal{T}_0$ is indeed small. For example, the first row indicates that our choice of ${}_{11}\mathcal{T}_M$ enabled us to use 42% fewer data points than $\mathcal{T}_0$ with only a 0.7% increase in the RMSE.

Some of the rows in Table 5.8 are more important than others. For example, if the researcher were characterizing based on the M_{33} scattering profile it is clear that only the mean response should be consulted. Importantly, our choice of ${}_{33}\mathcal{T}_M$ in this case results in very little inflation of RMSE (3.4% and 6.0% for the low and high noise settings, respectively), whereas the best naive choice, $\mathcal{T}_C$, inflates the RMSE by 154% and 160% for the high and low noise settings, respectively.

Simulation 2 was designed to evaluate our methodology's choice of $\mathcal{T}_M^*$. This sim-

Table 5.8: Simulation 1: results for evaluating $\mathcal{T}_M$

Noise	Profile	Derivative(s)	$\mathcal{T}_0$	$\mathcal{T}_M$	$\mathcal{T}_L$	$\mathcal{T}_C$	$\mathcal{T}_R$
Low	M_{11}	0	3.83	3.85	11.22	7.54	3.68
Low	M_{11}	1	4.77	5.11	9.08	6.47	4.84
Low	M_{11}	2	3.73	4.72	9.42	10.96	4.63
Low	M_{11}	0,1,2	3.89	4.34	8.95	5.78	3.90
Low	M_{33}	0	2.70	2.79	8.22	6.86	7.08
Low	M_{33}	1	5.21	8.36	9.61	7.56	9.31
Low	M_{33}	2	10.01	15.25	34.47	16.86	13.56
Low	M_{33}	0,1,2	6.99	8.69	14.22	11.47	7.32
High	M_{11}	0	3.86	3.95	11.66	7.15	3.88
High	M_{11}	1	4.92	5.76	9.97	6.43	5.86
High	M_{11}	2	4.06	6.14	11.77	13.59	7.55
High	M_{11}	0,1,2	4.13	4.98	9.73	7.46	4.93
High	M_{33}	0	2.73	2.90	8.34	7.11	7.15
High	M_{33}	1	5.35	8.88	10.63	8.13	9.39
High	M_{33}	2	10.24	16.60	35.68	22.13	14.21
High	M_{33}	0,1,2	7.13	9.35	14.31	12.55	7.41

The table compares the RMSE when characterization is performed over various subsets of $\mathcal{X}$. Recall that $\mathcal{T}_0 = \{1, 2, ..., 179\}$. For M_{11} we used ${}_{11}\mathcal{T}_L = \{1, 2, ..., 104\}$, ${}_{11}\mathcal{T}_C = \{38, 39, ..., 141\}$, and ${}_{11}\mathcal{T}_R = \{76, 77, ..., 179\}$, all of which have the same cardinality (104) as ${}_{11}\overline{\mathcal{T}}_M = \{69, 70, ..., 172\}$. For M_{33} we used ${}_{33}\mathcal{T}_L = \{1, 2, ..., 88\}$, ${}_{33}\mathcal{T}_C = \{46, 47, ..., 133\}$, and ${}_{33}\mathcal{T}_R = \{92, 93, ..., 179\}$, all of which have the same cardinality (88) as ${}_{33}\overline{\mathcal{T}}_M = \{19, 20, ..., 62\} \cup \{129, 130, ..., 172\}$. Recall while observations were collected over $\overline{\mathcal{T}}_M$, the characterization of (5.11) was performed with $\mathcal{T}^* = \mathcal{T}_M$ to mitigate boundary issues. Column 1 indicates whether the noise level was low (3%) or high (6%). Column 2 indicates whether the M_{11} profile or the M_{33} profile was used. Column 3 indicates which derivatives were consulted. Columns 4 through 8 display the values of RMSE.

ulation was similar to Simulation 1 except that this time we assumed prior knowledge that the size was restricted to either $[5, 50]$ (with $S = \{10, 15, 20, ..., 45\}$ for the simulation) or $[50, 100]$ (with $S = \{55, 60, 65, ..., 95\}$ for the simulation). We also only examined the 'low' noise setting. The results appear in Table 5.9.

The choice of $\mathcal{T}_M^*$ did appear to accomplish the goal of using fewer observations without sacrificing a great deal of characterization ability. In fact characterization ability was *improved* in several cases by using $\mathcal{T}_M^*$ as opposed to using the larger $\mathcal{T}$. If we exclude the irrelevant final row (since we would clearly not use the second derivative of the M_{33} scattering profile for characterization), the worst inflation of RMSE for $\mathcal{T}_M^*$ over $\mathcal{T}$ was 10.3%. While the naive choices do occasionally beat our methodology's choice of $\mathcal{T}_M^*$, each naive choice is also frequently disastrous, with RMSE inflation up to 190%. It should also be noted that in the previous section our methodology indicated that nanoparticles with sizes between 50 and 100 nm would be more difficult to characterize than particles smaller than 50 nm. This prediction is validated by the results of Simulation 2.

Table 5.9: Simulation 2: results for evaluating $\mathcal{T}_M^*$

$[S_1, S_2]$	Profile	Derivative	$\mathcal{T}$	$\mathcal{T}_M^*$	$\mathcal{T}_L^*$	$\mathcal{T}_C^*$	$\mathcal{T}_R^*$
$[5, 50]$	M_{11}	0	3.58	3.82	4.16	3.28	2.98
$[5, 50]$	M_{11}	1	5.20	5.36	5.28	5.15	6.61
$[5, 50]$	M_{11}	2	4.91	5.05	13.36	14.22	7.62
$[5, 50]$	M_{33}	0	2.82	2.96	2.90	3.05	3.66
$[5, 50]$	M_{33}	1	2.23	2.46	2.69	3.56	4.21
$[5, 50]$	M_{33}	2	4.32	3.97	4.78	4.41	5.44
$[50, 100]$	M_{11}	0	4.54	4.32	7.19	4.41	2.74
$[50, 100]$	M_{11}	1	4.90	4.55	5.45	5.75	5.22
$[50, 100]$	M_{11}	2	4.20	3.66	12.83	10.06	5.32
$[50, 100]$	M_{33}	0	2.81	2.86	2.82	3.73	5.15
$[50, 100]$	M_{33}	1	7.08	7.98	11.40	10.54	8.08
$[50, 100]$	M_{33}	2	12.24	11.84	16.73	12.52	12.42

The table compares the RMSE when characterization is performed over various subsets of $\mathcal{X}$. The range $[S_1, S_2]$ indicates the prior knowledge of the researcher. We took $\mathcal{T} = \mathcal{T}_M = \{76, 77, ..., 165\}$ for M_{11} and $\mathcal{T} = \mathcal{T}_0 = \{1, 2, ..., 179\}$ for M_{33}. See the previous section for our method's suggestions for $\mathcal{T}_M^*$. For comparison, for M_{11} we also used ${}_{11}\mathcal{T}_L^* = \{76, 77, ..., 145\}$, ${}_{11}\mathcal{T}_C^* = \{86, 87, ..., 155\}$, and ${}_{11}\mathcal{T}_R^* = \{96, 97, ..., 165\}$, all of which have the same cardinality (70) as ${}_{11}\mathcal{T}_M^*$. For M_{33} we used ${}_{33}\mathcal{T}_L^* = \{1, 2, ..., 142\}$, ${}_{33}\mathcal{T}_C^* = \{19, 20, ..., 161\}$, and ${}_{33}\mathcal{T}_R^* = \{37, 38, ..., 179\}$, all of which have similar cardinality to ${}_{33}\mathcal{T}_M^*$. Column 1 indicates the range of possible sizes under consideration. Column 2 indicates whether the M_{11} profile or the M_{33} profile was used. Column 3 indicates which derivatives were consulted. Columns 4 through 8 display the values of RMSE.

Copyright© Benjamin Hall, 2010.

Chapter 6 Conclusions and future research

In this dissertation we have studied four problems which involve the nonparametric estimation of derivatives. In Chapter 2 we described a generalized C(p) criterion for selecting the tuning parameters of a nonparametric estimator of derivatives. This was important because tuning parameters chosen by the ordinary C(p) criterion are not guaranteed to lead to good estimates of derivatives. Generalized C(p) faced challenges not encountered by ordinary C(p). For example, both ordinary C(p) and generalized C(p) require as ingredients noise-corrupted versions of the function which is being estimated. With ordinary C(p) this problem is trivial since the observations themselves are noise-corrupted versions of the mean response. With generalized C(p) this required the development of empirical derivatives to serve this purpose.

In Chapter 3 we outlined a technique for constructing simultaneous confidence bands for a nonparametric regression estimator of a mean response and its derivatives. Such techniques were previously available only for local regression. Our method is more flexible in that it can be utilized with any self-consistent nonparametric regression estimator which is linear in the observed responses. Many previous methods for confidence bands around nonparametric estimates require assumptions that the bias, interpolation error, and variance be either known or bounded. We proposed data-based techniques for accommodating situations where these quantities are unknown.

In Chapter 4 we generalized the simultaneous confidence bands to account for heteroscedastic noise. This required that we define an appropriate estimator of the variance function. Obtaining asymptotic justifications for the bands in the presence of heteroscedastic noise also required stronger assumptions on the nonparametric estimator being employed. We demonstrated that these necessary assumptions are satisfied, for example, by kernel regression.

Finally, in Chapter 5 we proposed a method for choosing a limited subset of angles over which to perform nanoparticle characterization. This method enables the scientist to save time and resources for both the forward problem and the inverse problem while at the same time overcoming boundary issues inherent in previous methods.

While each of these chapters represent steps forward in their respective efforts, each also contains avenues for future research. The generalized C(p) criterion performed very well in selecting tuning parameters in simulations. However, a choice must be made for the inputs $k_1, ..., k_q$. It would be helpful to develop a data-based method of choosing these inputs. Generalized C(p)also requires that the nonparametric regression estimator be self-consistent and linear in the observed responses. This rules out, for example, local polynomial estimation. However, the assumptions could be relaxed so that generalized C(p) could be applied to any nonparametric regression estimator for which the estimates of the mean response and its derivative(s) are linear in the observed responses, allowing the self-consistency assumption to be dropped. This would enable generalized C(p) to be compared to the IRSC method (Fan and Gijbels 1995), which is a data-based method for tuning parameter selection when estimating derivatives using local polynomial estimation. Other comparisons could also be made to naive methods for choosing tuning parameters such as CV, GCV, or AIC, which are used when interest lies only in $\mu(x)$. Further studies could be conducted to examine how generalized C(p) performs in the presence of very large noise. Also, while Charnigo and Srinivasan (2010) demonstrate that generalized C(p) is asymptotically unbiased, it remains to be shown that the chosen parameter is optimal in theory. If λ^* represents the minimizer of $\mathbb{E}\sum_{i=1}^{n}\left[\widehat{\frac{d^q}{dx^q}\mu(x_i)} - \frac{d^q}{dx^q}\mu(x_i)\right]^2$ and $\widehat{\lambda}$ represents the minimizer of generalized C(p), then showing, for example, that $\widehat{\lambda}/\lambda^* \to^P 1$ would strengthen the theoretical justification for generalized C(p).

The simultaneous confidence bands which we proposed in Chapter 3 also required that the nonparametric regression estimator be self-consistent and linear in the observed responses. Again, this assumption could be relaxed to accommodate nonparametric regression estimators, such as local polynomial estimation, for which the estimates of the mean response and its derivative(s) are linear in the observed responses. It would then be possible to compare our method to that of Claeskens and

Van Keilegom (2003) which works for local polynomial estimation. The inequality in (3.16) enables the confidence bands to account for the bias through data-based estimates. A tighter inequality would result in narrower, and thus less conservative bands.

Chapter 4 provides an important generalization of Chapter 3 by allowing for heteroscedastic noise. However, there are other generalizations which would make the simultaneous confidence bands even more widely applicable. The methods described in Chapters 3 and 4 require that errors be both normal and independent. Simultaneous confidence bands which allow for correlated and/or non-normal errors would be helpful. To accommodate correlated errors, a correlation structure would have to be assumed, and this problem would be more difficult for some structures than for others. Dealing with non-normal errors is interesting because some nonparametric regression techniques do not require that errors be normally distributed. However, our confidence bands rely heavily on properties of the multivariate normal and multivariate t distributions. This makes the extension of our confidence bands to non-normal errors difficult.

Our method for determining limited angles for nanoparticle characterization allows for comparisons to be made between the characterization ability of the mean response and its derivative(s). The characterization method of Charnigo et al (2010), which characterizes based on the mean response and its derivative(s) simultaneously, requires a weighting scheme for the mean response and its derivative(s). Specifically, they define the discrepancy function

$$\begin{aligned}
& D[M^*(\theta;u), M(\theta;c), K, \mathbf{w}] \\
:= \quad & \frac{e^{w_0}}{e^{w_0}+e^{w_1}+...+e^{w_K}} \times \frac{\int_{\mathcal{X}}[M^*(\theta;u)-M(\theta;c)]^2 d\theta}{\int_{\mathcal{X}}[M^*(\theta;u)]^2 d\theta} \\
+ \quad & \frac{e^{w_1}}{e^{w_0}+e^{w_1}+...+e^{w_K}} \times \frac{\int_{\mathcal{X}}[\frac{\partial}{\partial\theta}M^*(\theta;u)-\frac{\partial}{\partial\theta}M(\theta;c)]^2 d\theta}{\int_{\mathcal{X}}[\frac{\partial}{\partial\theta}M^*(\theta;u)]^2 d\theta} + ... \\
+ \quad & \frac{e^{w_K}}{e^{w_0}+e^{w_1}+...+e^{w_K}} \times \frac{\int_{\mathcal{X}}[\frac{\partial^K}{\partial^K\theta}M^*(\theta;u)-\frac{\partial^K}{\partial^K\theta}M(\theta;c)]^2 d\theta}{\int_{\mathcal{X}}[\frac{\partial^K}{\partial^K\theta}M^*(\theta;u)]^2 d\theta}
\end{aligned}$$

where u denotes the unknown configuration, $c \in \mathcal{C}_1$ refers configurations for which reference curves have been constructed and $\mathbf{w} := (w_0, w_1, ..., w_K)'$ is the weight vector.

Note that our method replaces $\mathcal{X}$ with $\mathcal{T}^*$. The unknown configuration is then characterized by the $c \in \mathcal{C}_1$ for which $D[M^*(\theta;u), M(\theta;c), K, \mathbf{w}]$ is minimized. It would be advantageous if the evaluations of the characterization abilities of the mean response and its derivative(s) could be translated into recommendations for the weight inputs. An initial idea is to measure the characterization ability of, for example, the mean response and its first two derivatives using weights (w_0, w_1, w_2) by the quantity:

$$\frac{e^{w_0}B_0 + e^{w_1}B_1 + e^{w_2}B_2}{e^{w_0} + e^{w_1} + e^{w_2}}. \tag{6.1}$$

Note that (6.1) is a generalization of our recommendation to use $B_0 + B_1 + B_2$ to evaluate the combined characterization ability of the mean response and the first two derivatives using equal weights. However, note that the recommendation for weights cannot be simply based on maximizing (6.1) since such a scheme would place all of the weight on the derivative corresponding to $\max\{B_0, B_1, B_2\}$. In reality the situation is more complicated than (6.1) indicates. There is correlation between the characterization abilities represented by B_0, B_1, and B_2. One interesting alternative may be to define the weights as functions of θ. This would allow a nanoparticle to be classified, for example, based primarily on the mean response over some ranges of θ and based primarily on the 1st or 2nd derivative over other ranges.

Determining limited angles for nanoparticle characterization is beneficial because it saves resources by requiring fewer observations. However, another way to reduce the number of observations is to sample over a grid which is less dense. Suppose the researcher wants to conduct characterization while using only T observations. Our method gives the researcher the ability to choose these T points so that they are located in a range of $\mathcal{X}$ which is optimal for characterization with the denseness of the grid fixed. However, the researcher could place the T points so that they are equispaced over $\mathcal{X}$. That is, the range could be fixed and the denseness varied according to the choice of T. Future research could explore how to determine optimal locations to place the T points while letting the range and the denseness vary simultaneously. Yet another interesting avenue for future research is to place prior knowledge about a nanoparticle's configuration into a Bayesian framework.

Copyright© Benjamin Hall, 2010.

Bibliography

[1] Bickel, P.J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Annals of Statistics*, **1**, 1071–1095.

[2] Brinkman, B.N. (1981). Ethanol fuel - a single-cylinder engine study of efficiency and exhaust emissions. *SAE Transactions*. **9**0, 1410-1424.

[3] Charnigo, R., Francoeur, M., Menguc, M.P., Brock, A., Leichter, M., and Srinivasan, C. (2007). Derivatives of scattering profiles: tools for nanoparticle characterization. *Journal of the Optical Society of America A*. **2**4, 2578-2589.

[4] Charnigo, R., Francoeur, M., Kenkel, P., Menguc, M.P., Hall, B., and Srinivasan, C. (2010). On estimating quantitative features of nanoparticles. Submitted for publication.

[5] Charnigo, R. and Srinivasan, C. (2010a). Self-consistent estimation of mean response functions and their derivatives. Submitted for publication.

[6] Charnigo, R. and Srinivasan, C. (2010b). On simultaneous estimation of a mean response and its derivatives. Submitted for publication.

[7] Charnigo, R. and Srinivasan, C. (2008). A generalized C(p) criterion for derivative estimation. Tech report.

[8] Charnigo, R., Sun, J., and Muzic, R. (2006). A semi-local paradigm for wavelet denoising. *IEEE Transactions on Image Processing*, **15**, 666-677.

[9] Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Annals of Statistics*, **31**, 1852–1884.

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

[11] Cramer, H. (1946). *Mathematical Methods of Statistics.* Princeton University Press.

[12] David, H.A. and Nagaraja, H.N. (2003). *Order Statistics.* John Wiley and Sons, Inc, Hoboken, New Jersey.

[13] Eubank, R.L. and Speckman, P.L. (1993). Confidence bands in nonparametric regression. *Journal of the American Statistical Association*, **88**, 1287-1301.

[14] Fan, J., and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society, Series B*, **57**, 371-394.

[15] Francoeur, M., Venkata, P.G., and Menguc, M.P. (2007). Sensitivity analysis for characterization of gold nanoparticles and 2D agglomerates via surface plasmon scattering patterns. *Journal of Quantitative Spectroscopy and Radiative Transfer*, **106**, 44-55.

[16] Gasser, T. and Muller, H.G. (1979). Kernel estimation of regression functions. In: *Smoothing Techniques for Curve Estimations*, eds. Gasser and Rosenblatt. Heidelberg: Springer.

[17] Hall, P., and Titterington, D.M. (1988). On confidence bands in nonparametric density estimation and regression. *Journal of Multivariate Analysis*, **27**, 228-254.

[18] Hall, P., Kay, J.W., and Titterington, D.M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77**, 415-419.

[19] Hardle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, Cambridge.

[20] Knafl, G., Sacks, J., and Ylvisaker, D. (1985). Confidence bands for regression functions. *Journal of the American Statistical Association*, **80**, 683-691.

[21] Loader, C. (1999). *Local Regression and Likelihood.* Springer, New York.

[22] Manickavasagam, S., and Menguc, M. P. (1997). Scattering matrix elements of fractal-like soot agglomerates. *Applied Optics*, **36**, 1337-1351.

[23] Mallows, C.L. (1973). Some comments on C_p. *Technometrics.* **1**5, 661-675.

[24] Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and Its Applications*, **9**, 141-142.

[25] Naiman, D. (1986). Conservative confidence bands in curvilinear regression. *Annals of Statistics*, **14**, 896-906.

[26] Parzen, E. (1962). On estimation of a probability density and mode. *Annals of Mathematical Statistics*, **35**, 1065-1076.

[27] Priestley, M.B. and Chao, M.T. (1972). Nonparametric function fitting. *Journal of the Royal Statistical Society, Series B*, **34**, 385-392.

[28] Ramsay, J.O. and Silverman, B.W. (2002). *Applied Functional Data Analysis.* Springer-Verlag, New York.

[29] Reinsch, H. (1967). Smoothing by spline functions. *Numerische Mathematik*, **10**, 177-183.

[30] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, 642-649.

[31] Royston, P. and Altman, D. (1994). Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. *Journal of the Royal Statistical Society, Series C*, **43**, 429-467.

[32] Schoenberg, I.J. (1964). Spline functions and the problem of graduation. *Mathematics*, **52**, 947-950.

[33] Silverman, B.W. (1984). Spline smoothing: the equivalent variable kernel method. *Annals of Statistics*, **12**, 898-916.

[34] Stone, C.J. (1980). Optimal rates of convergence for nonparametric estimators. *Annals of Statistics*, **8**, 1348-1360.

[35] Sun, J. and Loader, C. (1994). Confidence bands for linear regression smoothing. *Annals of Statistics*, **22**, 1328-1345.

[36] Vysochankii, D.F. and Petunin, Y.I. (1980). Justification of the 3σ rule for unimodal distributions. *Theory of Probability and Mathematical Statistics*, **21**, 2536.

[37] Wahba, G. (1977). A survey of some smoothing problems and the method of generalized cross-validation for solving them. In: *Applications of Statistics*, ed. P.R. Krishnaiah. Amsterdam: North Holland.

[38] Watson, G.S. (1964). Smooth regression analysis. *Sankya, Series A*, **26**, 359-372.

[39] Venkata, P.G., Aslan, M.M, Menguc, M.P., and Videen, G. (2007). Surface plasmon scattering by gold nanoparticles and two-dimensional agglomerations. *ASME Journal of Heat Transfer*, **129**, 60-70.

Vita

Benjamin Hall

Birth Place and Date: Frankfort, Kentucky, August 8, 1983.

Education

University of Kentucky, Lexington, KY
M.S. in Statistics, 2008

Georgetown College, Georgetown, KY
B.S. in Mathematics, 2006