INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI
A CORPUS-LINGUISTIC ANALYSIS OF FIFTEENTH- AND SIXTEENTH-CENTURY RJAZANIAN LEGAL DOCUMENTS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate
School of The Ohio State University

By
Syeng-Mann Yoo, M.A.

*****

The Ohio State University
2001

Dissertation Committee:
Professor Daniel E. Collins, Adviser
Professor Charles E. Gribble
Professor Predrag Matejic

Approved by

Daniel E. Collins
Adviser
Department of Slavic and East European Languages and Literatures
ABSTRACT

In this dissertation, we propose a function estimation procedure using adaptive free-knot splines (AFKS) as well as an associated algorithm for implementation in nonparametric regression.

Knot selection is extremely important for estimating a function with complicated structure by spline functions. The existing knot selection schemes such as stepwise forward and/or backward selection are restricted to select knots from distinct design points and suffer the difficulty of knot confounding and are thus unsuitable for our purpose. The adaptive free-knot splines allow for variable knot numbers and replacements of knots at any locations and multiple knots at the same location. This flexibility leads to an adaptive spline estimator that adapts to any function with inhomogeneous smoothness, including a sharp kink or discontinuity, which substantially improves the representation power of splines.

A new knot selection scheme is proposed, by utilizing an adaptive model selection criterion and an Evolutionary Algorithm (EA). In this dissertation, two versions of the adaptive model selection strategy, one using the generalized degree of freedom (GDF) introduced by Ye (1998) and the other using the adaptive model selection criterion proposed by Shen and Ye (2001), are investigated. The adaptive model selection guards against the selection error in searching through a large candidate
knot space, while the Evolutionary Algorithm, a stochastic optimization procedure, enables us to search the global minima of an adaptive model selection criterion.

The performance of the proposed procedure is examined via simulations, which demonstrates the superiority of the proposed methodology over the existing spline methods and wavelet shrinkage estimators. Its usefulness is illustrated by application to a real data set.
To my soon to be wife Mayumi.
ACKNOWLEDGMENTS

I wish to express gratitude to my advisor, Professor Xiaotong Shen. He supervised all through my research. Obviously, this work would be impossible without his suggestive advice, warm encouragement and patience.

I am grateful to my parents, Masuei and Atsuko Miyata, for their understanding and support through my graduate study.

Finally, this dissertation is dedicated to my fiancée, Mayumi. Her love and support are beyond words.
VITA

October 10, 1968 ................................................. Born - Tokyo, Japan

March 1991 ............................................................ B.A. Economics, Hitotsubashi University, Tokyo, Japan
March 1993 ............................................................ M.A. Economics, Hitotsubashi University, Tokyo, Japan

PUBLICATIONS

Research Publications


FIELDS OF STUDY

Major Field: Statistics
TABLE OF CONTENTS

Abstract ............................................................................................................................. ii
Dedication .............................................................................................................................. iv
Acknowledgments ............................................................................................................. v
Vita ....................................................................................................................................... vi
List of Tables ................................................................................................................... lx
List of Figures ................................................................................................................ x
Chapters:

1. INTRODUCTION ................................................................................................ 1

2. REGRESSION SPLINE ...................................................................................... 4
   2.1 Spline Representation ................................................................................ 4
   2.2 Estimation .................................................................................................... 6
   2.3 Conventional Model Selection Criterion ............................................... 7
   2.4 Conventional Knot Selection Scheme ..................................................... 8

3. ADAPTIVE FREE-KNOT SPLINE ........................................................................ 11
   3.1 Adaptive Model Selection Criterion ...................................................... 11
   3.2 Evolutionary Algorithm ............................................................................. 13
   3.3 Adaptive Free-knot Splines ...................................................................... 16
   3.4 Simulation Study ......................................................................................... 23
      3.4.1 Effects of Multiple Knots. ............................................................... 24
      3.4.2 Numerical Examples ....................................................................... 26

vii
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.</td>
<td>FURTHER ENHANCEMENT</td>
<td>30</td>
</tr>
<tr>
<td>4.1</td>
<td>New Adaptive Model Selection Criterion</td>
<td>30</td>
</tr>
<tr>
<td>4.2</td>
<td>Implementation and Evolutionary Algorithm</td>
<td>32</td>
</tr>
<tr>
<td>4.3</td>
<td>Numerical Examples</td>
<td>34</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Simulation</td>
<td>35</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Real Data Example</td>
<td>37</td>
</tr>
<tr>
<td>5.</td>
<td>INFEERENCE ON REGRESSION SPLINE</td>
<td>40</td>
</tr>
<tr>
<td>5.1</td>
<td>Inference</td>
<td>40</td>
</tr>
<tr>
<td>5.2</td>
<td>Initial Estimation of Local Bias and Variance</td>
<td>44</td>
</tr>
<tr>
<td>5.3</td>
<td>Bootstrap Estimation of Local Bias and Variance</td>
<td>47</td>
</tr>
<tr>
<td>5.4</td>
<td>Simulation Study</td>
<td>50</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Empirical error probability</td>
<td>51</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Cross-Validation</td>
<td>60</td>
</tr>
<tr>
<td>6.</td>
<td>FURTHER RESEARCH</td>
<td>63</td>
</tr>
</tbody>
</table>

Appendices:

| A. | TECHNICAL PROOFS                          | 66   |
| A.1| Proofs of Theorems in Chapters 3 and 4    | 66   |
| A.2| Proofs of Theorems in Chapter 5           | 68   |
|    | Bibliography                              | 84   |
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Average MSEs for different procedures.</td>
</tr>
<tr>
<td>3.2</td>
<td>Average MSEs and the standard deviations of MSEs (in parentheses) for different procedures.</td>
</tr>
<tr>
<td>4.1</td>
<td>Average MSEs and the standard deviations of the MSEs (in parentheses) for different procedures.</td>
</tr>
<tr>
<td>5.1</td>
<td>Empirical Error Probability in Example 1</td>
</tr>
<tr>
<td>5.2</td>
<td>Empirical Error Probability in Example 2</td>
</tr>
<tr>
<td>5.3</td>
<td>Empirical Error Probability in Example 3</td>
</tr>
</tbody>
</table>
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The EEG recording of the epilepsy brain wave.</td>
<td>2</td>
</tr>
<tr>
<td>3.1</td>
<td>Flow chart of EA for knot selection.</td>
<td>18</td>
</tr>
<tr>
<td>3.2</td>
<td>Probability density functions for knot mutation: (a) knot number, (b) knot location.</td>
<td>20</td>
</tr>
<tr>
<td>3.3</td>
<td>Parallel tempering sequences.</td>
<td>22</td>
</tr>
<tr>
<td>3.4</td>
<td>Test function $\theta$ (solid line) and estimated function (dotted line) from noisy data with $n = 1000$.</td>
<td>25</td>
</tr>
<tr>
<td>3.5</td>
<td>Test function $\theta$ (solid line) and estimated function (dotted line).</td>
<td>25</td>
</tr>
<tr>
<td>3.6</td>
<td>Test function $\theta$ (solid line) and the fitted function (dotted line) from noisy data with $n = 101$.</td>
<td>27</td>
</tr>
<tr>
<td>3.7</td>
<td>Four benchmark examples (Block, Bump, HeaviSine and Doppler). The true test function (solid line) and the fitted curve (dotted line) for AFKS and the GDF.</td>
<td>28</td>
</tr>
<tr>
<td>4.1</td>
<td>Four benchmark examples (Block, Bump, HeaviSine and Doppler). The true test function (solid line) and the fitted curve with the AFKS and the adaptive model selection criterion (dotted line).</td>
<td>36</td>
</tr>
<tr>
<td>4.2</td>
<td>Fitted curve for the epilepsy brain wave data.</td>
<td>38</td>
</tr>
<tr>
<td>4.3</td>
<td>The plot of the epilepsy brain wave residuals.</td>
<td>39</td>
</tr>
<tr>
<td>4.4</td>
<td>QQ-plot of the residuals.</td>
<td>39</td>
</tr>
</tbody>
</table>
5.1 Fitted curves and confidence bands in Example 1 ............... 52
5.2 Local bias estimate in Example 1 ........................................ 53
5.3 Fitted curves and confidence bands in Example 2 ............... 55
5.4 Local bias estimate in Example 2 ........................................ 56
5.5 Local s.d. estimate in Example 3 ........................................ 60
5.6 Local bias estimate in Example 3 ........................................ 61
5.7 Cross-Validation in Example 2, $h$ : number of observations in a local window ......................................................... 62
Consider a regression model:

\[ Y_i = \theta(x_i) + \epsilon_i, \quad i = 1, \ldots, n, \]  

where the design points \( \{x_i\}_{i=1}^n \) are in \([0, 1]\) and the random errors \( \{\epsilon_i\}_{i=1}^n \) are independently and identically distributed with \( E(\epsilon_i) = 0 \) and \( E(\epsilon_i^2) = \sigma^2 > 0 \). Our goal is to estimate the regression function \( \theta \) from data \( Z = \{Z_i = (x_i, Y_i)\}_{i=1}^n \). In practice, the degree and type of smoothness of \( \theta \) are unknown and spatial adaptation is necessary to estimate \( \theta \). Such situations with complicated structured data often occur in signal processing. For example, in medical signal processing, a brain wave of an epileptic patient, measured by ElectroEncephaloGraph (EEG), typically exhibits high non-linearity with periodic and sharp spikes between slowly changing waves, as displayed in Figure 1.1. The details regarding the EEG recording in Figure 1.1 were described in Miwakeichi, Ramirez-Padron, Valdes-Sosa and Ozaki (2001). For other types of signals such as speech signals, radar signals, and stock price volatility, spatial adaptation is clearly essential to recover the signal and to extract, analyze and interpret relevant characteristics of data.

In the past two decades, many estimators for the nonparametric regression have been investigated intensively, including kernel smoothers (Gasser and Müller, 1984),
regression splines (Agarwal and Studden, 1980), smoothing splines (Whittaker, 1923), loess smoothers (Cleveland, 1979) and the wavelet thresholding estimator (Donoho and Johnstone, 1995).

In this dissertation, we concentrate on two aspects of regression spline estimators. First, we investigate optimal knot selection schemes for splines. Although the existing spline methods with stepwise knot selection have proven effective to fit smooth functions, they suffer from the knot confounding and perform poorly for non-stationary inhomogeneous functions. To fully utilize the potential of splines, more flexible representations of splines together with more accurate knot selection is necessary. For this purpose, we introduce free knot spline estimators and two versions of an adaptive
model selection criterion to guard against selection error and enhance the performance, and utilize an Evolutionary Algorithm, a stochastic optimization procedure, to seek the global optimal knot sequence. A new knot selection scheme for free knot splines, based on the adaptive penalty and the EA, is proposed and is shown to yield the optimal knots for splines. The performance of the proposed procedure is examined via simulations. It is revealed that the proposed method captures the inhomogeneous smooth structure of regression functions and it provides better fits than alternative adaptive spline and wavelet shrinkage estimators.

Second, we study inference of spline estimators. It is well known that prediction using nonparametric regression suffers from bias-variance trade-off and it is necessary to estimate the local behavior of spline estimators to make accurate inferences. For this purpose, consistent estimators of the local bias and the variance of regression splines with and without Bootstrapping are introduced and a confidence interval based on the regression spline estimator with bias-correction is constructed.

This dissertation is organized as follows: In Chapter 2, a regression spline estimator and a conventional knot selection scheme are discussed. In Chapter 3, an adaptive model selection criterion using the generalized degree of freedom and an Evolutionary Algorithm, for knot selection, are investigated, and a new knot selection scheme is proposed. In Chapter 4, an enhanced version of the adaptive model selection criterion is introduced and an improved version of the knot selection scheme is presented. In Chapter 5, the local bias and the variance estimators of a regression spline estimator are defined and a bias-corrected confidence interval is constructed. Further research is discussed in Chapter 6. Technical proofs are in the Appendix.
2.1 Spline Representation

For motivation, let us first consider the case in which \( \theta \) is differentiable. Taylor's expansion of \( \theta(x) \) at any \( a \in (0, 1) \) yields that

\[
\theta(x) = \theta(a) + \sum_{j=1}^{m-1} \frac{\theta^{(j)}(a)}{j!} (x_i - a)^j + R_m(x_i),
\]

where \( R_m(x_i) = \theta^{(m)}(\xi_i)(x_i - a)^m/m! \) for some \( \xi_i = a + c_i(x_i - a) \) and \( 0 < c_i < 1 \) is a remainder in the above expansion. This Taylor's expansion suggests polynomial regression for (1.1) can be performed after \( R_m(x_i) \) is ignored. However, in practice, the polynomial regression estimator is numerically unstable, especially near boundaries of the design points and the global shape of a polynomial is very sensitive to some local features of data. Therefore, to approximate the regression function \( \theta \), one may use an estimator that is more flexible to local structures of \( \theta \). One candidate is a spline function, which is defined as a piecewise polynomial with smoothness at joints, called knots. More precisely, a polynomial spline \( s(x) \) of order \( m \) \( (m \geq 1) \) with a knot sequence \( t = (t_1, \ldots, t_k) \) for \( 0 = t_0 < t_1 \leq \cdots \leq t_k < t_{k+1} = 1 \), including \( k_i \) \( (k_i \leq m) \) repeated knots at each location \( t_i \), is an \( m \)-th order polynomial in each
interval \([t_i, t_{i+1}]\). The spline has the \(p_i\)-th continuous derivative at each knot location \(t_i\), where \(p_i\) is \((m - 1)\) minus the number \(k_i\) of repeated knots at \(t_i\). The total number \(k\) of knots is \(\sum_{i=1}^{k_0} k_i < (n - m)\), where \(k_0\) is the number of distinct knots. A popular choice of the order \(m\) is 4, equivalent to the degree 3, which yields a cubic spline. The order \(m\) of the spline will be fixed in what follows.

Given a fixed order \(m\) and a knot sequence \((t, k)\), a spline function \(s(x)\) can be expressed in terms of a truncated power basis:

\[
\{x^j, \ j = 0, \ldots, m - 1; \ (x - t_i)^{m-j}, \ j = 1, \ldots, k_i, \ l = 1, \ldots, k\},
\]

which spans a linear space of spline functions with \((t, k)\), where \((\cdot)_+ = \max\{\cdot, 0\}\). The truncated power basis representation is of the form:

\[
s(x) = \sum_{i=0}^{m-1} \alpha_i x^i + \sum_{l=1}^{k_0} \sum_{j=1}^{k_l} \alpha_{l,j} (x - t_l)^{m-j}.
\]

As a special case, if there are no multiple knots, that is \(k_l = 1\) for \(l = 1, \ldots, k\), the truncated power representation of the spline can be simplified as:

\[
s(x) = \sum_{i=0}^{m-1} \alpha_i x^i + \sum_{l=1}^{k} \alpha_{l+m-1} (x - t_l)^{m-1}.
\]

This representation is simple and intuitive, although it may be computationally unstable. In computation, the corresponding equivalent B-spline basis is often used. The normalized B-spline basis \(\{B_l(x; t), \ l = 1, \ldots, k + m\}\), defined by a knot sequence \(t = (0 = t_0 < t_1 \leq \ldots \leq t_k < t_{k+1} = 1)\), is:

\[
B_l(x; t) = (t_l - t_{l-m})[t_{l-m}, \ldots, t_l][x - \cdot]^{m-1}, \ l = 1, \ldots, k + m,
\]

where \([t_{l-m}, \ldots, t_l]g\) is the \(m\)-th order divided difference of the function \(g\), and \(t_l = t_{\min(\max(l,0), k+1)}\), for \(l = 1 - m, \ldots, k + m\). By the Curry-Schoenberg Theorem (c.f. de
Boor (1978), p.113), the (normalized) B-spline functions form a basis of the collection of splines with \((t, k)\). The B-spline basis can be constructed inductively, which allows fast implementation, c.f., de Boor (1978).

2.2 Estimation

Let \( h = (t, k) \) be a tuning parameter for the B-spline representation. For any spline function \( s(x) \) with \( h \), there exists \( a \in \mathbb{R}^{k+m} \) such that \( s(x) = B(x, t)'a \), where \( B(x, t) = (B_1(x; t), \ldots, B_{k+m}(x; t)) \). For convenience, in the sequel, \( B(\cdot, t) \) will be abbreviated as \( B(\cdot) \) or \( B \). To estimate \( \theta \) in (1.1), \( \theta \) is approximated by a spline \( s(x) = \sum_{l=1}^{k+m} \alpha_l B_l(x, t) \). The regression spline estimator \( \hat{\theta}_n(x, h) \) is defined as \( \sum_{l=1}^{k+m} \hat{\alpha}_l B_l(x, t) \), where \( \hat{\alpha}_l \) is the least squares estimator of \( \alpha_l \), obtained by minimizing the least squares criterion:

\[
\min_{\alpha_1, \ldots, \alpha_{k+m}} \sum_{i=1}^{n} \left( Y_i - \sum_{l=1}^{k+m} \alpha_l B_l(x_i, t) \right)^2.
\]

The least squares estimates \( \hat{a} \) and \( \hat{\theta}_n(x, h) \) can then be expressed as

\[
\hat{a} = (B(x)'B(x))^{-1}B(x)'y,
\]

\[
\hat{\theta}_n(x, h) = B(x)\hat{a} = B(x)(B(x)'B(x))^{-1}B(x)'y = My,
\]

where \( [B(x)]_i = B_i(x_i; t) \), \( i = 1, \ldots, n \), \( l = 1, \ldots, k + m \) and \( M = B(B'B)^{-1}B' \).

In the past two decades, the nonparametric regression has been extensively investigated. Stone (1980, 1982) obtained the optimal rate of convergence for general nonparametric regression. Agarwal and Studden (1980) studied the rate of convergence of the regression spline in the one-dimensional case. Stone and Koo (1986), Friedman and Silverman (1989), Friedman (1991), Stone (1994) and Luo and Wahba (1997) investigated the knot selection procedure for splines in the multidimensional...

2.3 Conventional Model Selection Criterion

As we discussed in Sections 2.2 and 2.3, the regression spline estimator \( \hat{\theta}_n(x, h) \) depends not only on the observations \( \{(x_i, Y_i)\}_{i=1}^n \) but also on the choice of knots. In contrast to the conventional spline methodology with knots confined to distinct design points, the free-knot spline methodology allows for replacements of knots at any locations, repeated knots at the same location, and variable knot numbers. This flexibility leads to an adaptive spline estimator with variable multiple knots, which enables free-knot splines to adapt to any function with non-stationary and inhomogeneous smoothness.

Usually, the optimal locations and the number of knots of the adaptive spline estimator are selected by minimizing a model selector. Most existing model selection criteria are of the form:

\[
\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\theta}_{M(h)}(x_i))^2 + \lambda \text{Tr}(M(h))\sigma^2, \quad \lambda > 0. \tag{2.6}
\]

Many model selection criteria are only different by the value of \( \lambda \); for example, \( \lambda = 2 \) gives AIC (Akaike (1973)) and \( C_p \) (Mallows (1973)), \( \lambda = \log(n) \) gives BIC (Schwarz (1978)) and \( \lambda = 2 \log(k + m) \) yields Risk Inflation Criterion (RIC) (George and Foster (1994)). In the literature of spline estimation, other choices of \( \lambda \) have been proposed for knot selection. For example, \( \lambda = 3 \) and \( \lambda = 1.2 \) are used for MARS (Friedman
(1991)) and HAS (Luo and Wahba (1997)), respectively. In addition, the other well-known model selection criteria such as CV and GCV are asymptotically equivalent to AIC and $C_p$, although they are in a form different from (2.6); see Stone (1977) and Hastie and Tibshirani (1990).

Among these conventional model selection criteria, AIC, $C_p$, BIC and RIC are derived using the following different principles: the Kullback-Leibler information, the unbiased risk estimation, an asymptotic Bayes factor and the minimax principle, respectively. In practice, however, they suffer from selection bias. For example, an inflated selection criterion such as BIC is suitable only in a situation in which a small number of knots is appropriate, and vice versa. For accurate knot selection, the penalty $\lambda$ in (2.6) needs to be estimated adaptively from data $\{(x_i, Y_i)\}_{i=1}^n$ to reduce the selection bias. Discussion of data-adaptive penalty will be deferred to Chapters 3 and 4.

If $\sigma^2$ is unknown, any good $n^{1/2}$-consistent estimator such as $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2 / 2$ (Rice (1984)) or $\hat{\sigma} = \text{Median}\{|Y_{2i} - Y_{2i-1}|/(.6745\sqrt{2}) : i = 1, \ldots, n/2\}$ (Donoho and Johnstone (1995)) can be used, where $\{Y_1, \ldots, Y_n\}$ are sorted according to the ordered $x_i$'s.

### 2.4 Conventional Knot Selection Scheme

In the literature of spline estimation, Friedman and Silverman (1989, TURBO), Friedman (1991, MARS) and Luo and Wahba (1997, HAS) use different knot selection schemes. The knot selection for these splines is performed in a fashion similar to variable selection in the stepwise regression. Given the order $m$ of a spline, stepwise knot selection is typically performed as follows:
Initialization: Choose an initial knot sequence. For conventional knot selection schemes, candidate knots are chosen only from the distinct design points \( \{x_i\}_{i=1}^n \).

Forward selection: Add one knot at a time, yielding the smallest value of the model selector among the current knots and the candidate knots. Repeat the process until no improvement can be made.

Backward elimination: Eliminate one knot at a time, yielding the largest value of the model selector when it is removed from the current knots. Repeat the process until no further improvement can be obtained.

Forward selection and backward elimination can be performed alternately.

The conventional spline methods have proven efficient via simulations to estimate relatively smooth regression functions. However, these methods suffer from some drawbacks in estimating spatially inhomogeneous smooth functions and perform poorly in estimating piecewise smooth functions with discontinuities. The difficulties are 1) the conventional spline methods typically confine the set of candidate knots to the distinct design points and do not allow for multiple knots at the same location. This restriction results in knot confounding, and enforces smoothness at knots leading to over-smoothing around a discontinuity or sharp kink of the regression function, 2) Stepwise knot selection often fails to locate the global minimizer of the model selector and is essentially suboptimal.

The need for a more intelligent knot search scheme has driven a recent work of Zhou and Shen (2001). Still, more flexible spline representations and more accurate knot search schemes are necessary to fully develop the potential of the spline
methodology. This motivates the subsequent development of a stochastic optimization technique.
CHAPTER 3

ADAPTIVE FREE-KNOT SPLINE

In this chapter, we propose an adaptive free-knot spline (AFKS) estimator which allows for variable multiple knots selected by a new knot selection scheme with an Evolutionary Algorithm (EA) and an adaptive model selection criterion.

3.1 Adaptive Model Selection Criterion

In this section, we propose an adaptive model selection criterion for selecting $h = (t, k) \in H = [0, 1]^{(n-m)} \times \{0, \ldots, n-m\}$ comprising all possible knot locations $t = (t_1, \ldots, t_k)$ including repeated knots as well as the number $k$ of knots. As discussed in Chapter 2, performance of the regression spline estimator $\hat{\theta}_n(x, h) = \hat{\theta}_n(h) = \hat{\theta}_{M(h)} = M(h)Y$ depends highly on the accuracy of the estimated tuning parameter $h = (t, k)$. An adaptive model selection criterion with data-driven penalty is necessary to reduce the selection bias for optimal knot selection.

To measure the discrepancy between the regression function $\theta$ and the fitted spline $\hat{\theta}_n$, we use a comparative quadratic loss,

$$l(\theta, \hat{\theta}_n(h)) = n^{-1}(\theta - \hat{\theta}_n(h))'(\theta - \hat{\theta}_n(h)) + \sigma^2.$$  \hfill (3.1)

For simplicity, we first consider a fixed $\sigma^2$, and then discuss how to estimate $\sigma^2$. Since $l(\theta, \hat{\theta}_n(h))$ depends on unknown $\theta$ in (3.1), $l(\theta, \hat{\theta}_n(h))$ is estimated by a model
selection criterion that is a generic form of (2.6), namely,
\[ n^{-1}[(Y - \hat{\theta}_n(h))'(Y - \hat{\theta}_n(h)) + 2D_n\sigma^2]. \quad (3.2) \]

In view of (2.6), \( D_n \) here can be interpreted as a measure of the “degrees of freedom” used in estimation for a model. The quantity \( D_n \) in (3.2) is obtained by minimizing the \( L_2 \)-distance between the comparative quadratic loss \( l(\theta, \hat{\theta}_n(h)) \) and (3.2); that is,
\[
E[l(\theta, \hat{\theta}_n(h)) - n^{-1}[(Y - \hat{\theta}_n(h))'(Y - \hat{\theta}_n(h)) + 2D_n\sigma^2]]^2. \quad (3.3)
\]
The optimal \( D_n \) is \( g_n = \sum_{i=1}^{n} E(\varepsilon_i \hat{\theta}_n(x_i, h)) \), which is called the generalized degree of freedom in Ye (1998). The optimal prediction of the comparative loss \( l(\theta, \hat{\theta}_n(h)) \) in the sense of (3.3) is
\[ n^{-1}[(Y - \hat{\theta}_n(h))'(Y - \hat{\theta}_n(h)) + 2g_n\sigma^2]. \quad (3.4) \]

By (A.1) of Ye (1998), \( g_n \) is estimated by an integral
\[
\frac{1}{\tau^2} \int \delta' \hat{\theta}_n(h)[Y + \delta]\phi(\delta)d\delta, \quad (3.5)
\]
where \( \phi(\cdot) \) is the probability density function of \( N(0, \tau^2 I) \), \( \hat{\theta}_n(h)[Y + \delta], \cdots, \hat{\theta}_n(x_n, h)[Y + \delta]' \) is the spline estimate based on pseudo data \( \{(x_i, Y_i + \delta_i)\}_{i=1}^{n} \), \( Y = (Y_1, \cdots, Y_n)' \) and \( \delta = (\delta_1, \cdots, \delta_n)' \). Here \( \tau \) controls the bias of the estimator, and (3.5) is asymptotically unbiased as \( \tau \to 0 \), as shown in Proposition 2 of Ye (1998). A practical choice of \( \tau \) is between 0.5\( \sigma \) and \( \sigma \), as argued in Ye (1998). Here we use a Monte Carlo regression method suggested by Ye (1998) to approximate \( g_n \). The procedure is as follows:

**Algorithm 1**

For a given knot sequence \( h = (t, k) \) and the hat matrix \( M(h) \),
1. Generate \( \delta_j = (\delta_{j1}, \ldots, \delta_{jn}) \sim N(0, \tau^2 I), \ j = 1, \ldots, T. \)

2. Evaluate

\[
\hat{\theta}_{M(h)}(Y_j^*) = B(x)(B(x)'B(x))^{-1}B(x)'Y_j^* = M(h)Y_j^*
\]

where \( Y_j^* = (Y_{j1}^*, \ldots, Y_{jn}^*)' \), \( Y_{ji}^* = Y_i + \delta_{ji}, \ i = 1, \ldots, n, \ j = 1, \ldots, T. \)

3. For each \( i = 1, \ldots, n \), compute the regression slope:

\[
\hat{\theta}_{M(h),i}(Y + \delta_j) = \alpha + \hat{\lambda}_i \delta_{ji}, \quad j = 1, \ldots, T.
\]

Finally, \( g_n \) is approximated by \( \hat{g}_n = \sum_{i=1}^{n} \hat{\lambda}_i. \)

Substituting \( \hat{g}_n \) for \( g_n \) in (3.4), our adaptive model selection criterion for \( h \in H \) is defined as:

\[
S(\hat{g}_n(h)) = n^{-1}[(Y - \hat{\theta}_{M(h)})'(Y - \hat{\theta}_{M(h)}) + 2\hat{g}_n\sigma^2]. \quad (3.6)
\]

When \( \sigma^2 \) is unknown, a consistent estimator such as \( \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n-1}(Y_{i+1} - Y_i)^2/2 \) (Rice, 1984) or \( \hat{\sigma} = \text{Median}\{[y_{2i} - y_{2i-1}]/(.6745\sqrt{2}) : i = 1, \ldots, n/2\} \) (Donoho and Johnstone, 1995) may be used.

### 3.2 Evolutionary Algorithm

In the literature of spline estimation, knot selection for the conventional splines is known to be a difficult nonlinear problem. Despite the successes of knot selection schemes such as forward, backward and stepwise selection for smooth functions, these schemes are much less satisfactory because they often fail to locate the optimal knots. The problem of knot selection is very critical for functions with complicated structures such as inhomogeneous smooth functions, although this problem may be less so for
homogeneous smooth functions where the performance is less sensitive to exact knot locations. In many situations, poor performance of a spline procedure is due to inexact knot locations as well as the number of knots.

In our situation, selecting an optimal \( h = (t, k) \) is more difficult and critical than that of conventional splines, because 1) the free-knot splines allow for not only the placement of knots at different locations but also multiple knots at the same point, and 2) the objective function, defined in (3.6), is not smooth with respect to the parameter \( h \).

The aforementioned discussion leads us to seek a stochastic searching strategy as opposed to a conventional deterministic one. The main advantage of using a stochastic searching strategy is that the search yields a global optimizer, or at least a good local optimizer whose value is close to the optimum, when appropriately implemented in practice. In contrast, none of the aforementioned deterministic searches guarantee this. The disadvantage is perhaps that the algorithm converges more slowly than its deterministic counterpart. With recent advances in computing, using such a stochastic searching strategy becomes more feasible.

In the following, we develop a knot selection scheme based on an Evolutionary Algorithm (EA). An EA, a stochastic-based strategy, imitates the adaptation process in biological evolution and provides a near optimal solution for an optimization problem. EAs have been used by Rogers (1991) for MARS and recently by Liang, Truong and Wong (2000) for Bayesian least squares spline fitting. Here, we use an evolutionary-based search algorithm, which incorporates the original genetic algorithm of Holland (1975) with the moves of simulated annealing, for a continuum population.
Let $f$, mapping $S$ to $\mathbb{R}$, be a real-valued function whose value represents the fitness to an environment, where $S$ is a search space of $f$. In our context, $f(h) = S(\hat{\theta}_n(h))$ is the model selection criterion with $S = H = [0, 1]^{(n-m)} \times \{0, 1, \ldots, n-2m\}$, and an element $h \in H$ is called an individual. Here we seek the global minimizer of $f$ such that

$$f^* = \min_{h \in S} f(h) = \min_{h \in H} S(\hat{\theta}_n(h)).$$

An EA begins with a set of $\xi$ initial values (individuals), called the population, followed by $\lambda$ offspring that are generated by applying random operations, called crossover and mutation, to the population. Then $\xi$ individuals are selected from $\xi$ parents and $\lambda$ offspring according to the value $f$. The set of selected individuals becomes a population in the next generation. In our construction, $\xi = \lambda$ is chosen to be 10.

**Crossover**: The crossover operation randomly selects $\kappa \in \{2, 3, \ldots, \xi\}$ individuals from the present population and recombines at random crossover points to produce a new offspring. The crossover operation mimics the fusion of chromosomes in generation of an embryo. For example, if $\kappa = 2$ with $x^1, x^2 \in \mathcal{P}$, and a randomly selected crossover point is at the $i$-th position, then

$$\begin{align*}
  \{ x^1 &= (x^1_1, \ldots, x^1_i, x^1_{i+1}, \ldots, x^1_{\kappa}) \\
  x^2 &= (x^2_1, \ldots, x^2_i, x^2_{i+1}, \ldots, x^2_{\kappa}) \} \xrightarrow{\text{crossover}} \\
  \{ y^1 &= (x^1_1, \ldots, x^1_i, x^2_{i+1}, \ldots, x^2_{\kappa}) \\
  y^2 &= (x^2_1, \ldots, x^2_i, x^1_{i+1}, \ldots, x^1_{\kappa}) \}
\end{align*}$$

If $\kappa = 1$, then the crossover operation is not applied.

**Mutation**: The mutation operation is applied with probability $p^m > 0$ and randomly modifies an individual $x$ in the present population by adding a random variable $Z$ to create a new offspring, i.e.,

$$x \xrightarrow{\text{mutation}} x + Z.$$
where $Z$ follows a mutation distribution. The distribution of $Z$ may differ from individual to individual. If the mutation probability $p^m = 1$, then the mutation operation is applied to all individuals.

**Selection:** The selection operation selects $\xi$ individuals from the $\xi$ individuals in the present (parent) population and $\lambda$ offspring obtained by the crossover and the mutation operations. There are mainly two types of selections: random and deterministic ones. The most popular random selection is the *proportional selection* (c.f., Holland (1975)), which selects individuals from $\{x^1, \ldots, x^{\xi+\lambda}\}$ with a probability distribution defined by $p^i = f(x^i) / \sum_{i=1}^{\xi+\lambda} f(x^i), \ i = 1, \ldots, (\xi + \lambda)$. The deterministic selection chooses the best $\xi$ individuals according to the value of $f$. This selection is also known as the *top-$\xi$ selection* rule and will be used in our EA. Other selection rules are available; see Rudolph (1997). The evolution process proceeds until a termination criterion is satisfied.

### 3.3 Adaptive Free-knot Splines

Knot selection is controlled by our EA that interactively updates knot locations and knot numbers in order to improve the objective function value. To describe our EA, let the number of the generation of a population be $\nu$, and let each individual in a population be ordered and indexed by an integer $i$. Define $h^{(i,\nu)} = (k^{(i,\nu)}, \xi^{(i,\nu)})$ by the $i$-th individual in the $\nu$-th generation, which is called a parent of an offspring $h^{(\xi+i,\nu)}, i = 1, \ldots, \lambda$. Evolution for knots proceeds as follows. First, the mutation operation is applied to parents to generate offspring, including a mutation of $k$ and that
of $t$. Then, the selection operation is applied after the mutation. The process alternates until a stopping criterion is met. Here the crossover operation is not performed, since it is not helpful in knot selection based on our limited numerical experience.

For knot mutation, the mutation distribution is critical and needs to be carefully developed in order for our EA to possess certain features.

The first feature is that the variance of the mutation distribution decays as time evolves. This feature assures that regions that may contain global optimizers are checked first in the beginning of the evolutionary process and the exact location of the global optimizer is examined thoroughly at a later stage of the process. The process of decreasing mutation variance resembles the "cooling" scheme of a sequence of temperature ladders, which is the key feature of Simulated Annealing. The temperature ladders, which indicate the levels of variability of the mutation distributions, are indexed by an integer $\zeta$ in the sequel. Each temperature ladder is used for a number of generations. That is, the mutation with the same extent of variability in one particular temperature ladder is applied to a number of generations until the termination criterion for the temperature ladder is met. The number of generations within the same temperature ladder is random and the rule to move to the next temperature ladder will be described later.

The second feature is that the mutations are inhomogeneous among the individuals in a population. This feature improves search efficiency and accuracy. The third feature is that the speed of decay of the mutation variance is controlled by a mixture of sequences of various temperature ladders, known as parallel tempering to be discussed below. Figure 3.1 provides a flow chart of the evolution process of our EA for knot selection.
 EA is terminated as the termination condition is satisfied.

Figure 3.1: Flow chart of EA for knot selection
For convenience in what is to follow, let $k$ denote the total number of knots including the exterior knots and change the notation of the knot sequence $t$ as $0 = t_1 = \cdots = t_m < t_{m+1} \leq \cdots \leq t_{k-m} < t_{k-m+1} = \cdots = t_k = 1$, where $t_1 = \cdots = t_m = 0$ and $t_{k-m+1} = \cdots = t_k = 1$ represent external knots.

For the $\nu$-th generation within the $\zeta$-th temperature ladder, $\nu = 1, 2, 3, \ldots$ and $\zeta = 1, 2, \ldots, n_{\text{temp}}$, the build-in components of our EA are described as follows:

**Knot Number Mutation:** To take into account the fact that the knot number $k$ is an integer between $2m$ and $n$, the mutation distribution is chosen to be an integer part of a truncated normal distribution because the mean and variance of a normal distribution are determined separately and the shape of a normal distribution is easily controlled. For knot number $k^{(\xi+i, \nu)}$ of the $(\xi+i)$-th offspring in the $\nu$-th generation, with the $\zeta$-th temperature ladder, we first sample $w$ from $N(k^{(i, \nu)}, c^{(\nu, \zeta)}k^{(i, \nu)})$, where the parameter $c^{(\nu, \zeta)}$ controls the cooling scheme of the mutation distribution, to be discussed below. Define $k^{(\xi+i, \nu)}$ to be $[w + 0.5]$ if $w \in (2m + 0.5, n)$; otherwise resample until this requirement is met, where $[w]$ denotes the integer part of $w$. This choice of distribution allows us to search around the current knot number, and with high probability, as suggested by part (a) of Figure 3.2, provides a probability density function of the knot number mutation distribution for $k^{(\xi+i, \nu)}$ with $k^{(i, \nu)} = 20, c^{(\nu, \zeta)} = 1.0$ and $n = 128$.

**Intermediate Offspring:** After $k^{(\xi+i, \nu)}$ is computed, an intermediate offspring $t^{(\xi+i, \nu)}$ is generated from the parent knot location $t^{(i, \nu)}$. The knot location mutation will be applied to the intermediate offspring. For $j = m + 1, \ldots, k^{(\xi+i, \nu)} - m$, sample $t^{(\xi+i, \nu)}_j$ from the discrete uniform distribution on $\{t^{(i, \nu)}_{m+1}, \ldots, t^{(i, \nu)}_{k^{(i, \nu)}-m}\}$. Define $t^{(\xi+i, \nu)}$ to be $(t^{(\xi+i, \nu)}_{m+1}, \ldots, t^{(\xi+i, \nu)}_{k^{(\xi+i, \nu)}-m})$. 

19
Figure 3.2: Probability density functions for knot mutation: (a) knot number, (b) knot location.

**Knot Location Mutation:** The mutation distributions for $t_{j}^{(ζ+1,ν)}$, $j = m + 1, \ldots, k^{(ζ+ν)} - m$, in the $ν$-th generation and the $ζ$-th temperature ladder are restricted to the range of the design points. The mutations are inhomogeneous among the knots to satisfy properties: (1) the mutation has high probability to search in an interval $(t_{j-1}^{(ζ+ν)}, t_{j+1}^{(ζ+ν)})$ determined by adjacent knots, because we do not need to search for $t_{j}^{(ζ+1,ν)}$ in places outside the neighborhood. Consequently, the variance of the knot location mutation should be large if the distance from the neighborhood is large and vice versa. (2) The standard deviation is greater than some positive constant, say $inc = 1/n$ or the minimum increment of the design points. If the standard deviation is zero then the knot location mutation is not performed to $t_{j}^{(ζ+1,ν)}$. 
To implement this inhomogeneous mutation, the variance is defined as follows: For 
\( j = m + 1, \ldots, k(\xi + i, \nu) - m + 1 \), let 
\[ \text{dif}^{(\xi + i, \nu)}_j = t^{(\xi + i, \nu)}_j - t^{(\xi + i, \nu)}_{j-1} \]
and
\[ d^{(\xi + i, \nu)}_j = \begin{cases} 
2\text{inc} & \text{if } \text{dif}^{(\xi + i, \nu)}_j < 2\text{inc}, \\
\text{dif}^{(\xi + i, \nu)}_j + 2\text{inc} & \text{if } 2\text{inc} < \text{dif}^{(\xi + i, \nu)}_j < 4\text{inc}, \\
2\text{dif}^{(\xi + i, \nu)}_j & \text{if } 4\text{inc} < \text{dif}^{(\xi + i, \nu)}_j.
\end{cases} \]
Then replace \( d^{(\xi + i, \nu)}_j \) by \( \max\{\text{inc}, \min\{s^{(\nu, \zeta)}, d^{(\xi + i, \nu)}_j\}\} \), where \( s^{(\nu, \zeta)} \) controls the cooling scheme. Now generate \( Z \sim N(t^{(\xi + i, \nu)}_j, 1) \) and \( v \) from \( \{d^{(\xi + i, \nu)}_j, -d^{(\xi + i, \nu)}_{j-1}\} \) with probability 1/2, and let \( t = v|Z| \). Replace \( t^{(\xi + i, \nu)}_j \) by \( t \) if \( t \in (0, 1) \); otherwise repeat the process. Part (b) of Figure 3.2, displays a probability density function of the knot location mutation for \( t^{(\xi + i, \nu)}_j \) with \( t^{(\xi + i, \nu)}_{j+1} = 0.25, t^{(\xi + i, \nu)}_j = 0.5, t^{(\xi + i, \nu)}_j = 0.6, s^{(\nu, \zeta)} = 0.5 \) and \( n = 128 \).

**Multiple Knots:** The offspring knot sequence in \( t^{(\xi + i, \nu)} \) is sorted. Then, for \( j = m + 1, \ldots, k(\xi + i, \nu) - m - 1 \), if \( |t^{(\xi + i, \nu)}_j - t^{(\xi + i, \nu)}_{j+1}| < \text{inc} \) and the multiplicity of \( t^{(\xi + i, \nu)}_j \) is less than \( m \), the order of the spline, then let \( t^{(\xi + i, \nu)}_{j+1} \) be \( t^{(\xi + i, \nu)}_j \). Otherwise remove \( t^{(\xi + i, \nu)}_{j+1} \) from \( t^{(\xi + i, \nu)} \).

**Selection:** The next generation is selected by the top-\( \xi \) selection rule. That is, the \( \xi \) individuals that give the \( \xi \)-smallest values are selected from the \( \xi \) parents and the \( \lambda \) offspring, to form a population in the next generation.

**Termination Criterion:** If \( f^*_\nu = \min\{f(h^{(1, \nu)}), \ldots, f(h^{(\xi + \lambda, \nu)})\} \) decreases in the \( n \)over generations or does not change in the \( n \)limit successive generations, then the next temperature ladder defined by the parallel tempering is employed. Otherwise, the same procedure is repeated. When \( n \)temp steps of the parallel tempering are examined, the EA is terminated.
Parallel Tempering: Different temperature ladders are mixed to control the cooling scheme. For each knot number mutation in the $\nu$-th generation within the $\zeta$-th temperature ladder, $c^{(\nu,\zeta)}$ is randomly selected from three sequences $\{c_1^{(\zeta)}, c_2^{(\zeta)}, c_3^{(\zeta)}\}$ with equal probability for $\zeta = 1, \ldots, n_{\text{temp}}$. The initial values are set to be $c_1^{(1)} = c_2^{(1)} = c_3^{(1)} = 1.0$. At the end of each temperature ladder, $c_i^{(\zeta)} (i = 1, 2, 3)$ are updated using the rule: $c_1^{(\zeta+1)} = c_1^{(\zeta)} * T$, $c_2^{(\zeta+1)} = c_2^{(\zeta)} - 1/n_{\text{temp}}$, $zz^{(\zeta+1)} = zz^{(\zeta)} / T$ and $c_3^{(\zeta)} = 1.0 - zz$, where $T (= 0.95)$ is a rate of decay and $zz^{(1)} = T^{n_{\text{temp}}}$. In other words, $c_1^{(\zeta)}$ and $c_2^{(\zeta)}$ decay exponentially and linearly, respectively, while $c_3^{(\zeta)}$ decreases slowly. Figure 3.3 displays $c_1^{(\zeta)}, c_2^{(\zeta)}$ and $c_3^{(\zeta)}$ as a function of the temperature ladder $\zeta$. Fast decay of $c_1^{(\zeta)}$ assures a sampler to stay around the global optimum, while slow decay of $c_3^{(\zeta)}$ is suitable for a sampler to escape from a local minimum. A mixture of different speeds of decay determined by stochastic choice enhances the
performance of our EA. For the parallel tempering in the knot location mutation, \( s^{(c)} \) is selected from \( \{s_1^{(c)}, s_2^{(c)}, s_3^{(c)}\} \) with equal probability. The initial values are set to be \( s_1^{(1)} = s_2^{(1)} = s_3^{(1)} = 1.0 \), which is equal to the range of the design points, and \( s_i^{(c)}, i = 1, 2, 3 \) are updated in the same way as before.

For the minimization problem (3.7) the EA described above is applied. Let \( h^{(\nu)} \) be the optimal knot sequence in the \( \nu \)-th generation such that \( f(h^{(\nu)}) = f^{*} \) = \( \min\{f(h^{(1,\nu)}), \ldots, f(h^{(\nu+\lambda,\nu)})\} \). Then the following convergence property can be established.

**Theorem 3.3.1.**

\[
\lim_{\nu \to \infty} f(h^{(\nu)}) = \lim_{\nu \to \infty} S(\hat{\theta}_n(h^{(\nu)})) = \min_{h \in H} f(h) \quad \text{a.s.}
\]

The proof is given in the Appendix. The convergence properties for the conventional deterministic knot selection schemes have not been theoretically justified. In contrast, Theorem 3.3.1 provides theoretical justification of the convergence of the knot selection with the Evolutionary Algorithm.

### 3.4 Simulation Study

In this section, we conduct simulations to examine the performance of our adaptive free-knot spline (AFKS). Here, the cubic splines are used \( (m = 4) \) and the mean squared error (MSE) \( n^{-1} \sum_{i=1}^{n} (\theta(x_i) - \hat{\theta}(x_i))^2 \) is used to measure the goodness of fit. The number and the location of knots are chosen by AFKS.

The tuning parameters for our EA are specified as follows: For the GDF estimation, the size of \( \tau \) is set to be \( 0.5\hat{\sigma} \), where \( \hat{\sigma} = \text{Median}\{|y_{2i} - y_{2i-1}|/(0.6745\sqrt{2}) : i = 1, \cdots, n/2\} \). The sizes of the population and the offspring in each generation are
\[ \xi = \lambda = 10. \] The initial population begins with one knot at the center of the design points. The parameters for termination are set to be \( n_{\text{temp}} = n_{\text{limit}} = n_{\text{over}} = 100, \ T = 0.85. \) To save computational time, the process is terminated when the value of the objective function has no improvement after 5 successive temperature steps.

### 3.4.1 Effects of Multiple Knots.

To see the benefit of splines with multiple knots, we examine visual quality of the estimated functions in a simulated example that has been previously studied by Wang (1995) who uses a wavelet shrinkage approach, by Koo (1997) who uses a linear regression spline that allows for discontinuity, and by Liang, Truong and Wong (2000) who employed a Bayesian curve fitting approach. The test function is following:

\[
\theta(x) = 2\sin(4\pi x) - 6|x - 0.4|^{0.3} - 0.5\text{sign}(0.7 - x), \ x \in [0, 1]. \tag{3.8}
\]

This test function has a kink at 0.4 and a discontinuity point at 0.7. The design points \( \{x_i\}_{i=1}^n \) are uniformly spaced in \([0, 1]\). The sample size is 1000 and the standard deviation of the normal error \( \varepsilon \) is 0.2. For the GDF estimation, the simulation size is chosen as \( B = 200. \)

From Figures 3.4 and 3.5, we see that the function estimated by AFKS captures the kink as well as the jump of the test function. In contrast, the conventional splines with non-multiple knots are unable to capture either the kink or the jump. Furthermore, the estimated function appears to be smoother than that of Koo (1997) and similar to that of Liang, Truong and Wong (2000). The MSE of the fit is 0.00209.
Figure 3.4: Test function $\theta$ (solid line) and estimated function (dotted line) from noisy data with $n = 1000$.

Figure 3.5: Test function $\theta$ (solid line) and estimated function (dotted line).
3.4.2 Numerical Examples

In this section we compare the proposed procedure with various existing spline procedures.

Simulation 1

First we examine the performance of the procedure for relatively smooth functions. Consider Test Function 2 of DiMatteo, Genovese and Kass (2000). The test function is of the form:

\[ \theta(x) = \sin(x) + 2 \exp(-30x^2), \quad x \in [-2, 2]. \]

The sample size is \( n = 101 \) and the design points \( \{x_i\}_{i=1}^{n} \) are uniformly spaced on \([-2, 2]\). The error \( \varepsilon \) is normally distributed with \( \sigma = 0.3 \). For estimation of the GDF, \( B = 400 \).

AFKS is compared with SARS by Zhou and Shen (2001), the DMS method by Denison, Mallick and Smith (1996) and Modified-DMS (abbreviated M-DMS) and the Bayesian MARS (abbreviated BARS) by DiMatteo, Genovese and Kass (2000). The average MSEs of these methods are summarized in Table 3.1. Clearly, AFKS with GDF performs well. Interestingly, BARS equipped with BIC performs well for relatively smooth functions, while SARS equipped with AIC performs well for non-smooth or less smooth functions.

<table>
<thead>
<tr>
<th></th>
<th>SARS</th>
<th>DMS</th>
<th>M-DMS</th>
<th>BARS</th>
<th>AFKS(AIC)</th>
<th>AFKS(BIC)</th>
<th>AFKS(GDF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>0.015</td>
<td>0.025</td>
<td>0.012</td>
<td>0.008</td>
<td>0.0245</td>
<td>0.0112</td>
<td>0.0127</td>
</tr>
<tr>
<td>SD</td>
<td>0.001</td>
<td>0.002</td>
<td>0.001</td>
<td>0.001</td>
<td>0.0097</td>
<td>0.0041</td>
<td>0.0022</td>
</tr>
</tbody>
</table>

Table 3.1: Average MSEs for different procedures.
The test function, noisy data, and the estimated function based on a randomly chosen realization are given in Figure 3.6. The corresponding MSE is 0.01513. Overall, AFKS with GDF does well.

![Figure 3.6: Test function \( \theta \) (solid line) and the fitted function (dotted line) from noisy data with \( n = 101 \).](image)

**Simulation 2**

To investigate the performance of AFKS for inhomogeneous smooth functions, including discontinuity, we consider four benchmark examples (Block, Bump, HeaviSine and Doppler) in Donoho and Johnstone (1995) with sample size \( n = 128 \), \( \sigma = 1 \) and the signal-to-noise ratio \( SD(f)/\sigma = 7 \). In this simulation, AFKS equipped with the GDF is compared with SARS in Zhou and Shen (2001), which compared favorably with a number of alternatives for these four examples. The average MSEs based on 100 runs are reported in Table 3.2.

27
Figure 3.7: Four benchmark examples (Block, Bump, HeaviSine and Doppler). The true test function (solid line) and the fitted curve (dotted line) for AFKS and the GDF.

<table>
<thead>
<tr>
<th></th>
<th>SARS</th>
<th>AFKS (AIC)</th>
<th>AFKS (BIC)</th>
<th>AFKS (GDF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block</td>
<td>0.7683</td>
<td>0.6486 (0.1140)</td>
<td>0.7629 (0.2082)</td>
<td>0.5887 (0.1212)</td>
</tr>
<tr>
<td>HeaviSine</td>
<td>0.3356</td>
<td>0.2810 (0.0993)</td>
<td>0.3499 (0.1148)</td>
<td>0.2548 (0.0899)</td>
</tr>
<tr>
<td>Bump</td>
<td>1.189</td>
<td>0.7522 (0.2051)</td>
<td>1.4560 (0.7732)</td>
<td>0.6162 (0.1116)</td>
</tr>
<tr>
<td>Doppler</td>
<td>0.4517</td>
<td>0.5688 (0.1568)</td>
<td>0.7954 (0.3984)</td>
<td>0.4780 (0.1102)</td>
</tr>
</tbody>
</table>

Table 3.2: Average MSEs and the standard deviations of MSEs (in parentheses) for different procedures.
The test and the estimated functions based on a randomly selected realization are displayed in Figure 3.7. The results of SARS are reproduced from Table 1 of Zhou and Shen (2001). Clearly, AFKS with the GDF outperforms AFKS with AIC and BIC. Indeed, the GDF achieves the desirable goal of adaptive selection, which is in contrast to AIC or BIC. This supports the utility of the GDF in model selection. Furthermore, AFKS with GDF yields significantly smaller MSEs on average as compared to SARS in three of four examples. For the Doppler example, AFKS with the GDF does slightly worse but insignificantly relative to the corresponding standard deviations.
In this chapter, a different adaptive model selector is introduced in an attempt to enhance the performance of adaptive knot selection. In the previous chapter, the degrees of freedom of the model is estimated from the data using the GDF. As an alternative, the new adaptive model selector estimates the optimal data adaptive penalty $\lambda$ together with the degrees of freedom in the model selector (2.6).

4.1 New Adaptive Model Selection Criterion

For any fixed $\lambda > 0$, the optimal knot sequence $\hat{h}(\lambda) = (\hat{t}, \hat{k})$ is estimated by minimizing (2.6) with respect to $h$, which yields $\hat{\theta}_{\hat{M}(\lambda)} = (\hat{\theta}_{\hat{M}(\lambda,1)}, \ldots, \hat{\theta}_{\hat{M}(\lambda,n)})$ with $\hat{M}(\lambda) = M(\hat{h}(\lambda))$. To select the optimal $\lambda$, we use a comparative quadratic loss $l(\theta, \hat{\theta}_{\hat{M}(\lambda)}) = n^{-1}(\theta - \hat{\theta}_{\hat{M}(\lambda)})'(\theta - \hat{\theta}_{\hat{M}(\lambda)}) + \sigma^2$. For simplicity, we first assume that $\sigma^2$ is fixed, and then discuss how to estimate $\sigma^2$. Since $l(\theta, \hat{\theta}_{\hat{M}(\lambda)})$ depends on the unknown $\theta$, we estimate $l(\theta, \hat{\theta}_{\hat{M}(\lambda)})$ using a model selection criterion of the form

$$n^{-1}[(Y - \hat{\theta}_{\hat{M}(\lambda)})'(Y - \hat{\theta}_{\hat{M}(\lambda)}) + \eta\sigma^2],$$

where the optimal $\eta(\lambda) \equiv g_0(\lambda) = (2/\sigma^2)E(\epsilon'\hat{\theta}_{\hat{M}(\lambda)})$ is obtained by minimizing

$$E\{l(\theta, \hat{\theta}_{\hat{M}(\lambda)}) - n^{-1}[(Y - \hat{\theta}_{\hat{M}(\lambda)})'(Y - \hat{\theta}_{\hat{M}(\lambda)}) + \eta(\lambda)\sigma^2]^2]\}.$$
with respect to $\eta \in (0, \infty)$. This leads to the optimal loss estimator:

$$n^{-1}[(Y - \hat{\theta}_{M(\lambda)})'(Y - \hat{\theta}_{M(\lambda)}) + \tilde{g}_o(\lambda)\sigma^2]. \quad (4.2)$$

Here the quantity $\tilde{g}_o(\lambda)/2$ in (4.2) is in fact the generalized degree of freedom (GDF) in Ye (1998). As suggested in Ye (1998), $\tilde{g}_o(\lambda)$ can be generally estimated by

$$\tilde{g}_o(\lambda) = \frac{2}{\tau^2} \int \delta' \hat{\theta}_{M(\lambda)}(Y+\delta) \phi_\tau(\delta) d\delta, \quad (4.3)$$

where $\hat{\theta}_{M(\lambda)}(Y+\delta) = (\hat{\theta}_{M(\lambda)}(Y+\delta,1), \ldots, \hat{\theta}_{M(\lambda)}(Y+\delta,n))$ is the estimated $\theta$ based on $\{X_i, (Y_i + \delta_i)\}_{i=1}^n$ for any $\lambda$, and $\phi_\tau$ is the $n$-dimensional probability density of $N(0, \tau^2 I)$. For practical implementation, $\tau$ is set equal to $0.5\sigma$, which controls the bias of $\tilde{g}_o(\lambda)$. As before, $\tau \in [0.5\sigma, \sigma]$ is generally a good choice. A Monte Carlo regression method suggested by Ye (1998) that can be used to approximate (4.3) proceeds in three steps:

**Algorithm 2**

1. Generate $\delta_j = (\delta_{j1}, \ldots, \delta_{jn}) \sim N(0, \tau^2 I)$ for $j = 1, \ldots, T$.

2. For each fixed $\lambda$ and $\delta_j$, $j = 1, \ldots, T$, compute the optimal $\hat{h} = \hat{h}(\lambda)$ by minimizing

$$(Y_j^* - \hat{\theta}_{M(h)})'(Y_j^* - \hat{\theta}_{M(h)}) + \lambda \text{Tr}(M(h))(\sigma^2 + \tau^2) \quad (4.4)$$

with respect to $h$ and evaluate $\hat{\theta}_{M(\lambda)}(Y_j^*)$, where $Y_j^* = Y + \delta_j$, $j = 1, \ldots, T$.

3. For each $i = 1, \ldots, n$, compute the regression slope:

$$\hat{\theta}_{M(\lambda),i}(Y + \delta_j) = \alpha + \lambda_i \delta_{ji}, \quad j = 1, \ldots, T.$$ 

Finally, $\tilde{g}_o(\lambda)$ is approximated by $2 \sum_{i=1}^n \lambda_i$. 

31
As suggested in Ye (1998), \( T \) should be no less than \( n \). Based on the above approximation, the proposed new adaptive model selection criterion for selecting \( \lambda \) is

\[
S(\hat{\theta}_{M(\lambda)}, Y) = (Y - \hat{\theta}_{M(\lambda)})'(Y - \hat{\theta}_{M(\lambda)}) + \hat{g}_0(\lambda)\sigma^2. \tag{4.5}
\]

Minimizing (4.5) over \( \lambda \in (0, \infty) \), we obtain the optimal \( \lambda^* \), yielding the adaptive free-knot spline estimator \( \hat{\theta}_{M(\lambda^*)} = \hat{\theta}(x, \hat{h}(\lambda^*)) \). In practice, the minimization in (4.5) is performed in a restricted set \((0, \lambda_{\text{max}})\) with a sufficiently large \( \lambda_{\text{max}} \).

When \( \sigma^2 \) is unknown, a consistent estimator such as \( \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2 / 2 \) (Rice (1984)) or \( \hat{\sigma} = \text{Median}\{|Y_{2i} - Y_{2i-1}|/(1.6745\sqrt{2}) : i = 1, \cdots, n/2\} \) (Donoho and Johnstone (1995)) can be used, where \( Y \) is sorted according to the ordered \( x \).

### 4.2 Implementation and Evolutionary Algorithm

To guard against selection error due to the knot search, we formulate an optimization problem via an EA as follows: Let \( f \), mapping \( S \) to \( \mathbb{R} \), be a real-valued function whose value represents the fitness to an environment, where \( S \) is a search space of \( f \). In our context, \( f(h, \lambda) = S(\hat{\theta}_{M(h(\lambda))}, Y) \) is the model selection criterion (4.5) with \( S = H \times (0, \lambda_{\text{max}}) \), \( h \in H = [0, 1]^{(n-m)} \times \{0, 1, \ldots, n-m\} \) and \( \lambda \in (0, \lambda_{\text{max}}) \). Our goal is to find the minimizer of \( f \) such that

\[
f(h^*, \lambda^*) = \min_{h \in H, \lambda \in (0, \lambda_{\text{max}})} f(h, \lambda). \tag{4.6}
\]

The algorithm for selecting the optimal parameter \((h, \lambda)\) comprises three steps:

**Step 1.** For any fixed \( \lambda \), minimize (2.6) with respect to \( h = (t, k) \) to obtain the optimal \( \hat{h} \).

**Step 2.** Estimate \( \hat{g}_0(\lambda) \) and evaluate (4.5).
Step 3. Minimize (4.5) with respect to $\lambda \in (0, \lambda_{\text{max}})$.

The optimization involved in Steps 1-3 is performed using an EA, to be described below.

For any fixed $\lambda$, the EA described in Section 3.3 is applied in a similar way to minimize (2.6) with respect to $h$. For fixed $\lambda$, to estimate $\tilde{g}_\nu(\lambda)$ and evaluate (4.5), we again need to apply our EA in (4.4) to obtain the optimal $\hat{h}$ for the pseudo data $\{(x, Y_j^*)\}$ for $j = 1, \ldots, T$. To save computational cost, we use the optimal $h$ obtained in Step 1 as the initial population of this EA and set $n_{\text{limit}} = n_{\text{over}} = 1$ for the termination criterion.

To optimize the adaptive penalty (4.5) with respect to $\lambda$, we again apply a version of EA as described above. For this version, the population size $\xi$ is set to be 2. Let $\lambda^{(i,l)}$, $i = 1, 2$, be the $i$-th penalty in the $l$-th generation, with initial values $\lambda^{(1,0)} = 2$ and $\lambda^{(2,0)} = \log(n)$. Some additional $\lambda$'s are also examined for $\lambda^{(j,0)} = (j + 15)(\log(n) - 2.0)/37$, $j = 3, \ldots, 20$. This “training” step is effective for saving the computational cost. Then two $\lambda^{(j,0)}$'s corresponding to the two smallest values for the adaptive penalty (4.5) will be used for the initial population of EA. The mutation distribution for the $(\xi + i)$-th offspring is chosen to be the truncated normal distribution $N(\lambda^{(i,l)}, s^{(l)}\lambda^{(i,l)})$ restricted to $(0, \lambda_{\text{max}})$. The controlling parameter $s^{(l)}$ decreases as the generation index $l$ increases, which is determined by the parallel tempering as before. Minimizing (4.5) over $\lambda \in (0, \lambda_{\text{max}})$ via the EA yields the adaptive free-knot spline estimator $\hat{\theta} = \hat{\theta}_{\lambda_{\text{opt}}}^{\lambda_{\text{opt}}}.$

For a fixed penalty $\lambda$, let $h^{(\nu)}$ be the optimal knot sequence in the $\nu$-th generation such that $f(h^{(\nu)}) = f^{(*)}_{\nu} = \min\{f(h^{(1,\nu)}), \ldots, f(h^{\xi+\lambda, \nu})\}$. Similarly, let $\lambda^{(l)}$ be the
optimal penalty in the $l$-th generation such that

$$S(\hat{\theta}_{\mathcal{M}(\lambda(l))}) = \min \{S(\hat{\theta}_{\mathcal{M}(\lambda(1))}), S(\hat{\theta}_{\mathcal{M}(\lambda(2))})\}.$$ 

For the minimization (4.6), two versions of EA are applied to simultaneously select the optimal knots with a fixed penalty as well as the optimal penalty.

**Theorem 4.2.1.**

$$\lim_{\nu \to \infty, l \to \infty} f(h^{(\nu)}, \lambda^{(l)}) = \lim_{\nu \to \infty} \lim_{l \to \infty} S(\hat{\theta}_{\mathcal{M}(h^{(\nu)}(\lambda(l)))}), Y = \min_{h \in H, \lambda \in (0, \lambda_{\text{max}})} f(h, \lambda) \quad a.s.$$ 

The proof is given in the Appendix.

### 4.3 Numerical Examples

In this section we examine the performance of our proposed procedure via simulations. In these simulations, the cubic spline with $m = 4$ is used and performance is measured by the mean squared error (MSE) $n^{-1} \sum_{i=1}^{n} (\theta(x_i) - \hat{\theta}(x_i))^2$. The number and the location of knots are selected by the proposed procedure.

The parameters of the EAs are as follows. The sizes of the population and the offspring for knot selection are set to be $\xi = \lambda = 10$. For knot selection with a fixed penalty $\lambda$, the initial population of the EA is a single knot at the center of the range of the design points. To have multiple knots, two neighboring knots will be collapsed if the distance between them is less than $\text{inc} =$ the minimum spacing of the design points. The parameters that control the stopping rule of our EA are set as follows: $\text{ntemp} = \text{nlimit} = \text{nover} = 100$ and $T = 0.95$.

To estimate the adaptive model selector $\hat{g}_a(\lambda)$, again we apply an EA for the knot search with $Y_j^* = Y + \delta_j, \ j = 1, \ldots, T$. To save computational cost, the fitted knot
sequence \( h = (t, k) \) obtained above is used as the initial population. The termination condition parameters of the EA for \( g_o(\lambda) \) estimation are \( ntemp = 100 \), \( nlimit = nover = 1 \) and \( T = 0.95 \).

The version of the EA for the optimal \( \lambda \) selection is applied as discussed in Section 4.2. To save computational time, the EA is applied with 10 generations.

### 4.3.1 Simulation

In this simulation, four benchmark examples (Block, Bump, HeaviSine and Doppler) in Donoho and Johnstone (1995) are examined with sample size \( n = 128 \), \( \sigma = 1 \) and the signal-to-noise ratio \( SD(f)/\sigma = 7 \). For estimation of \( g_o(\lambda) \) in (4.3), the simulation size \( T \) is set to be 300. Here, \( \hat{\sigma} \) is estimated by \( \hat{\sigma} = \text{Median}\{ |Y_{2i} - Y_{2i-1}|/(.6745\sqrt{2}) : i = 1, \cdots, n/2 \} \), as opposed to \( \sigma^2 = 1 \) in Donoho and Johnstone (1995). The plots of the true mean functions by solid curves and the fitted lines by dotted curves are given in Figure 4.1. To compare the proposed procedure to an existing adaptive spline estimation procedure, the results of SARS (Zhou and Shen, 2001) are reproduced from Table 1 of Zhou and Shen (2001). To understand the benefit of using the adaptive model selection criterion, we compare AFKS to free-knot splines with knots chosen by AIC, BIC and the original GDF in Ye (1998) for the same setting. The average MSEs (AMSEs) of the spline estimator equipped with the proposed adaptive model selection criterion is abbreviated as AMS. The average MSEs as well as the corresponding standard deviations based on 50 runs are reported in Table 4.1, where the results for AIC, BIC and GDF are reproduced from Table 2 of Miyata and Shen (2001).
Figure 4.1: Four benchmark examples (Block, Bump, HeaviSine and Doppler). The true test function (solid line) and the fitted curve with the AFKS and the adaptive model selection criterion (dotted line).

<table>
<thead>
<tr>
<th></th>
<th>SARS</th>
<th>AFKS (AIC)</th>
<th>AFKS (BIC)</th>
<th>AFKS (GDF)</th>
<th>AFKS (AMS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block</td>
<td>0.768</td>
<td>0.649 (0.114)</td>
<td>0.763 (0.208)</td>
<td>0.589 (0.121)</td>
<td>0.549 (0.115)</td>
</tr>
<tr>
<td>HeaviSine</td>
<td>0.337</td>
<td>0.281 (0.099)</td>
<td>0.350 (0.115)</td>
<td>0.255 (0.090)</td>
<td>0.236 (0.079)</td>
</tr>
<tr>
<td>Bump</td>
<td>1.189</td>
<td>0.752 (0.205)</td>
<td>1.456 (0.773)</td>
<td>0.616 (0.112)</td>
<td>0.595 (0.087)</td>
</tr>
<tr>
<td>Doppler</td>
<td>0.452</td>
<td>0.569 (0.157)</td>
<td>0.795 (0.398)</td>
<td>0.478 (0.110)</td>
<td>0.495 (0.128)</td>
</tr>
</tbody>
</table>

Table 4.1: Average MSEs and the standard deviations of the MSEs (in parentheses) for different procedures.
Except for the Doppler model, AFKS yields smaller MSEs than those of SARS on average. In particular, in the Block and Bump models, AFKS performs significantly better than SARS. For the Doppler model, the AMSE of AFKS with GDF is worse than that of SARS although the difference is small. In comparing the three model selectors AIC, BIC and GDF, we see that the adaptive model selector provides the best fit for all four examples. This demonstrates the superiority of the adaptive model selector for model selection.

4.3.2 Real Data Example

We now apply the AFKS to the brain wave signal data presented in Figure 1.1. This EEG recording of the epilepsy brain wave was obtained from an 11 year old female patient in the Cuban Neuroscience Center; see Miwakeichi, Ramirez-Padron, Valdes-Sosa and Ozaki (2001). The noisy signal was observed at \( n = 631 \) uniformly spaced time points. Our goal is to recover the true signal from its noisy background. As observed in Miwakeichi, Ramirez-Padron, Valdes-Sosa and Ozaki (2001), this brain wave signal data set contains relatively low noise, so the adaptive free-knot spline methodology seems appropriate and powerful.

The estimated signal by the AFKS is displayed in Figure 4.2. The order \( m \) of spline is chosen to be 4 with about 150 knots selected by our procedure. Clearly, the AFKS performs well and captures the main features of the true brain wave while it successfully wipes out noisy vibration in the slow wave area. To check the model assumption and assess the spline fit, the residuals, defined as the differences between the observed and the estimated signals, are displayed in Figure 4.3. The QQ-plot of
the residual is in Figure 4.4. The residual plot does not reveal any unusual patterns, which suggests that the spline fit is good.

Figure 4.2: Fitted curve for the epilepsy brain wave data.
Figure 4.3: The plot of the epilepsy brain wave residuals.

Figure 4.4: QQ-plot of the residuals.
CHAPTER 5

INFERENCES ON REGRESSION SPLINES

In the previous chapters, we investigated fitting algorithms for the regression splines. In this chapter, we develop an inference theory for prediction using the regression spline estimator. Here, we assume that the knot sequence \( h = (t, k) \) is fixed and independent of the data \( \{(x_i, Y_i)\}_{i=1}^n \) and that the regression function \( \theta \) is sufficiently smooth. In the case of adaptive spline estimation, the knot locations are functions of \( \{Y_i\}_{i=1}^n \), and the elements of the hat matrix of the regression spline are correlated with the response. Furthermore, the dimension of the hat matrix is the number of knots \( k \), which also depends on \( \{Y_i\}_{i=1}^n \). This imposes a challenge for developing the distribution theory for adaptive splines. Here we will not study this aspect.

5.1 Inference

For given \( h = (t, k) \), as suggested by (2.5),

\[
E(\hat{\theta}_n(x, h)) = E(M(\theta(x) + \epsilon)) = M\theta(x) \neq \theta(x),
\]

40
so that $\hat{\theta}_n(x, h)$ may be biased. On the other hand, for independently and identically distributed error, $\{\epsilon_i\}_{i=1}^n$, we have

$$\sum_{i=1}^n \text{Var}(\hat{\theta}_n(x_i, h)) = \sigma^2 \text{Tr}(M) = \sigma^2 (k + m). \quad (5.1)$$

When (5.1) is used to measure the total variability of $\hat{\theta}_n(x_i, h)$, the variance is an increasing function of the number $k$ of knots. However, as will be seen later, the bias $(\theta(x) - M\theta(x))$ increases as the maximum increment of the knots increases. In other words, the bias is inversely proportional to the number of knots. This is the so-called bias-variance trade-off phenomenon; that is, we can not choose the smoothing parameter $h = (t, k)$ to make the bias and the variance small at the same time, c.f. Hastie and Tibshirani (1990).

Zhou, Shen and Wolfe (1998) provided explicit expressions for the asymptotic pointwise bias and the variance of regression splines. We state their results under the following two assumptions.

(A1) For some constant $M > 0$, $\max_{1 \leq l \leq k} |u_{l+1} - u_l| = o(k^{-1})$, $u/\min_{1 \leq l \leq k+1} u_l \leq M$,

where $u_l = t_l - t_{l-1}$, $l = 1, \ldots, k + 1$, $u = \max_{1 \leq l \leq k+1} u_l$.

Under this condition, we have no multiple knot and $M^{-1} < ku < M$. See also Agarwal and Studden (1980).

(A2) Let $Q_n(x)$ and $Q(x)$ be the empirical distribution function of $\{x_i\}_{i=1}^n$ and an unknown distribution with a positive density $q(x)$, respectively. Assume that

$$\sup_{x \in [0,1]} |Q_n(x) - Q(x)| = o(k^{-1}).$$

41
Theorem 5.1.1 (Theorem 2.1 of Zhou, Shen and Wolfe (1998)). If \((A_1)\) and \((A_2)\) hold with \(k = o(n)\), \(\theta \in C^m[0, 1]\), then for any \(x \in (t_i, t_{i+1}) (i = 0, \ldots, k)\),

\[
\theta(x) - E(\hat{\theta}_n(x, h)) = b(x) + o(u^m),
\]

\[
Var(\hat{\theta}_n(x, h)) = \frac{\sigma^2}{n}B(x)G^{-1}(q)B(x)' + o((nu)^{-1}),
\]

where

\[
b(x) = \frac{\theta^{(m)}(x) u_t^m}{m!} B_m \left( \frac{x - t_i}{u_t} \right), \quad G(q) = \int_0^1 B(x)'B(x)q(x)dx.
\]

and \(B_m\) is the \(m\)-th Bernoulli polynomial.

Theorem 5.1.2 (Theorem 3.1 of Zhou, Shen and Wolfe (1998)). In addition to the assumptions in Theorem 5.1.1, suppose that \(k < Cn^{1/(2m+1)}\) for some constant \(C\). Then, for any \(x \in [0, 1]\),

\[
\frac{\hat{\theta}_n(x, h) + b(x) - \theta(x)}{\sqrt{Var(\hat{\theta}_n(x, h))}} \xrightarrow{d} N(0, 1).
\]

The asymptotic bias and variance of the regression splines provided in Theorems 5.1.1 and 5.1.2 depend on the higher order derivatives of the unknown regression function \(\theta\) and the covariance structure of the error term. To make proper inference, we need to estimate the local bias and variance of the regression spline. To do this, we use an idea based on moving blocks with the \(m_n\)-nearest neighborhood. The idea for utilizing the blocking technique is summarized as follows. From (1.1) and (2.5), we have

\[
\text{Residual} = y - \hat{y} = y - \hat{\theta}
\]

\[
= (I - M)\theta(x) + (I - M)\epsilon
\]

\[
= (\text{true bias}) + (I - M)\epsilon. \tag{5.2}
\]
This suggests that the bias at $x$ can be estimated by averaging this residual in the neighborhood of a specific point of interest $x_i$. The second term on the right hand side of (5.2) is the so called "residual" in the conventional linear regression. The average of the second term, $[(I - M)e]_i$, in the neighborhood of $x_i$ is expected to converge to zero almost surely as the number of observations in the neighborhood tends to infinity. In view of Theorem 5.1.1, the true bias at $x_i$ depends on the higher order derivatives of the regression function $\theta$. If the regression function is smooth enough, i.e., the higher order derivatives of $\theta$ are continuous, then the local true bias term at $x$ in the neighborhood of $x_i$ would be very close to the bias of the regression estimate at $x_i$. Then the average of the residual term in the block will converge to the true bias of $\hat{\theta}$ at $x_i$ as the sizes of the neighborhood blocks tend to zero.

Next, we construct the "centered" residual, defined as the difference between the residual and the local bias estimate. It is approximately equal to $(I - M)e$ and can be used to estimate the variance $\sigma^2$.

Let
\[ d_n = \max_{2 \leq i \leq n} (x_i - x_{i-1}) = o(n^{-\alpha}), \quad 0 < \alpha \leq 1, \quad (5.3) \]
and let $\{m_n\}$ be a sequence of positive integers such that $m_n = O(n^\beta)$ for $0 < \beta \leq \alpha$. Let $[M_n^L(x_i), M_n^U(x_i)]$ be the smallest interval containing the $m_n$ $x$'s closest to $x_i$ and let $M_n(x_i)$ be an index set corresponding to the $x$'s in $[M_n^L(x_i), M_n^U(x_i)]$. By (5.3), the assumptions that $m_n \to \infty$ and
\[ \forall x \in [0, 1], \quad (M_n^U(x) - M_n^L(x)) \leq (m_n - 1)d_n = o(n^{-\alpha+\beta}) \to 0 \text{ as } n \to \infty, \quad (5.4) \]
it follows that the number of observations in each block tends to infinity but the window width shrinks to zero. The assumption that the number of observations in
each $m_n$ neighborhood tends to infinity is necessary to assure convergence to zero of the average of the $(I - M)e$ terms. The assumption (5.4) says that the window size shrinks to zero, which assures that the average of the $(I - M)\theta(x)$ terms converges to the true bias at $x$.

One practical issue is how to select the optimal window size. We will address this issue separately for finite and large samples. To determine a window size for a given sample, we can use some model selector such as CV (Cross-Validation). We will discuss how to select the optimal rate for increasing the window size as the total number of observations goes to infinity after Theorem 5.2.1.

5.2 Initial Estimation of Local Bias and Variance

The proposed procedure for the initial estimation of the local bias and the standard deviation of a fitted regression spline can be described as follows:

**Step 11** Fit a regression spline curve to obtain an estimate

$$
\hat{\theta}_n(x, h) = B(x)\hat{\theta} = B(x)(B(x)'B(x))^{-1}B(x)'y = My.
$$

**Step 12** Define the local bias estimate at $x_i$:

$$
\hat{b}_i = \text{Ave}_{k\in M_n(x_i)}(y_k - \hat{\theta}_n(x_k)) = \frac{1}{m_n} \sum_{k\in M_n(x_i)} (y_k - \hat{\theta}_n(x_k)).
$$

**Step 13** Homoscedastic case. (Constant variance) Estimate the variance by the bias corrected residuals as follows:

$$
\hat{\sigma}^2 = \text{Var}_{1\leq k\leq n}(y_k - \hat{\theta}_n(x_k) - \hat{b}_k) = \frac{1}{n} \sum_{k=1}^{n} (y_k - \hat{\theta}_n(x_k) - \hat{b}_k)^2.
$$
Step 14 The bias of the fitted curve $\hat{\theta}_n$ is corrected as follows:

$$\tilde{\theta}(t) = \hat{\theta}_n(t) + \hat{b}_i + (t - x_i) \frac{(\hat{\theta}_n(x_{i+1}) + \hat{b}_{i+1}) - (\hat{\theta}_n(x_i) + \hat{b}_i)}{x_{i+1} - x_i}, \text{ for } x_i \leq t < x_{i+1}.$$  

Remark. The local window size $(M_n^U(x) - M_n^L(x))$ is of order of $o(n^{-\alpha+\beta})$. For a given sample size, the number $m_n$ of observations in a block is fixed. In practice, the value of $m_n$ may be chosen by a cross-validation criterion; See § 5.4.2 for more details.

Remark. The result can be generalized to cover the case where the variance of the $\epsilon_i$'s is not constant.

$$Y_i = \theta(x_i) + \epsilon_i, \quad i = 1, \ldots, n, \quad (5.5)$$

where

$$E(\epsilon_i) = 0 \text{ and } \text{Var}(\epsilon_i) = \sigma^2 w(x_i),$$

with $\sigma^2 > 0$, $w(\cdot) > 0$, $w(\cdot) \in C^1[0,1]$ and the $\epsilon_i$'s are uncorrelated.

In (5.5), the local bias estimate is the same as defined above. The local variance of a fitted regression spline for this setting can be estimated as follows:

**Step I3' Heteroscedastic case. (non-constant variance)** Estimate the local variance using $m_n$-neighborhoods as follows:

$$\tilde{\sigma}^2_i = \text{Var}_{k \in M_n(x_i)}(y_k - \hat{\theta}_n(x_k) - \hat{b}_k) = \frac{1}{m_n} \sum_{k \in M_n(x_i)} (y_k - \hat{\theta}_n(x_k) - \hat{b}_k)^2.$$  

Then the following asymptotic results on the initial local estimates can be established.
Theorem 5.2.1 (Consistency of $\hat{b}_i$). Suppose $(A_1)$ and $(A_2)$ are satisfied. If $\theta \in C^{(m+1)}[0,1]$ and $k = o(n)$, then

$$\hat{b}_i \xrightarrow{p} J_{i1}'(I - M)\theta(x), \text{ as } n \to \infty,$$

where $J_{i1}$ is a column vector whose elements are 0 except for the $i$-th element which is 1.

Remark. In view of Theorem 5.2.1, if $o(h^d) = o(n^{-\kappa'})$, $0 \leq \kappa' \leq \kappa$, where $\kappa = \min\{\alpha - \beta, \nu\beta\}$ and $0 < \nu < 1/2$, then

$$|\hat{b}_i - J_{i1}'(I - M)\theta(x)| = o_p(n^{-\kappa}).$$

Since $0 < \beta \leq \alpha \leq 1$, $\kappa$ is maximized when $\alpha = 1$ and $\beta = 2/3$ and the upper bound for the maximum value of $\kappa$ is $1/3$. That is, the optimal rate of convergence of $\hat{b}_i$ is $o_p(n^{-1/3-\delta})$ for any $\delta > 0$.

Theorem 5.2.2 (Consistency of $\hat{v}_i$). Under the assumptions of Theorem 5.2.1, if $m_n = o(nu)$ and $E(e_i^4) \leq C < \infty$ for $i = 1, \ldots, n$, then

$$\hat{v}_i^2 \xrightarrow{p} \sigma^2 w(x_i), \text{ as } n \to \infty.$$

Theorem 5.2.3 (Uniform asymptotic unbiasedness of $\tilde{\theta}$). Under the assumptions of Theorem 5.2.1, we have for any $1/2 > \eta > 0$,

$$\sup_{\eta < x < 1 - \eta} |\tilde{\theta}(x) - \theta(x)| \xrightarrow{p} 0, \text{ as } n \to \infty.$$

Theorem 5.2.4 (Asymptotic normality). Under the assumptions of Theorem 5.1.2, if $m/(2m+1) \leq \kappa = \min\{\alpha - \beta, \nu\beta\}$ and $0 < \nu < 1/2$, then

$$\sqrt{n}\frac{\hat{\theta}_n(x_i) + \hat{b}_i - \theta(x_i)}{\hat{v}_i} \xrightarrow{d} N(0, B(x)G^{-1}(q)B(x)'), \quad i = 1, \ldots, n.$$
5.3 Bootstrap Estimation of Local Bias and Variance

The bootstrap methodology was introduced by Efron (1979) for two purposes: 1) estimating the sampling distribution of statistics, and 2) estimating the accuracy of statistics, such as the standard deviation. The bootstrap methodology was theoretically studied independently by Singh (1981) and Bickel and Freedman (1981). Singh (1981) proved the validity of the bootstrap method in the case of \( \theta = E(X_i) = \mu, \) \( \hat{\theta} = \bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i, \) and \( m = n \) and gave the rate of convergence of the bootstrap distribution. This result has been generalized using the Edgeworth expansion of the bootstrap distribution, c.f., Hall (1992). Bickel and Freedman (1981) used the Mallows metric \( d_p(F, G) \) defined on a set of distribution functions to measure the discrepancy between two distribution functions \( F \) and \( G. \) The definition of the Mallows metric is \( d_p(F, G) = \inf E(\|X - Y\|^p)^{1/p}, \) where the infimum is taken over pairs of random variables \( X \) and \( Y, \) and the marginal distribution of \( X \) and \( Y \) are \( F \) and \( G, \) respectively. Bickel and Freedman (1981) proved the consistency of the bootstrap estimates and the weak convergence of the bootstrap distribution for the sample mean, von Mises functionals, the empirical process, and the quantile process. Freedman (1981) applied the bootstrap method to linear regression. This procedure is applied to regression splines in this section.

We now estimate the local bias and the standard error using the bootstrap method. In view of Theorem 5.2.4, we know that

\[
\frac{\hat{\theta}_n(x_i) + \hat{\beta}_i - \theta(x_i)}{\hat{v}_i}
\]

is asymptotically pivotal. The bootstrap version of this quantity is constructed as follows.
Step B1 Let \( e_i \) be the bias corrected and centered residual, i.e.,

\[
e_i = \frac{y_i - \hat{\theta}_n(x_i) - \hat{b}_i}{\hat{v}_i} - \bar{e}, \quad i = 1, \ldots, n,
\]

where \( \bar{e} = n^{-1} \sum_{i=1}^n (y_i - \hat{\theta}_n(x_i) - \hat{b}_i)/\hat{v}_i \), and the \( \hat{v}_i \)'s are 1 when the model is homoscedastic.

Step B2 Let \( B \) represent the bootstrap resampling number. Let \( \{e_{b,1}^*, \ldots, e_{b,n}^*, \ b = 1, \ldots, B\} \) be resamples drawn from \( \{e_1, \ldots, e_n\} \) randomly and with replacement as \( b \)-th replicate. Let

\[
Y_{b,i}^* = \hat{\theta}_n(x_i) + \hat{b}_i + \hat{v}_i e_{b,i}^*, \quad i = 1, \ldots, n, \ b = 1, \ldots, B,
\]

\[
\hat{\theta}_{nb}^* = B(x)(B(x)'B(x))^{-1}B(x)'Y_b^* = MY_b^*, \quad Y_b^* = (Y_{b,1}^*, \ldots, Y_{b,n}^*)'.
\]

Step B3 Let

\[
\hat{b}_{b,i}^* = \text{ave}_{k \in M_n(x_i)} (Y_{b,k}^* - \hat{\theta}_{nb}^*(x_k)) = \frac{1}{m_n} \sum_{k \in M_n(x_i)} (Y_{b,k}^* - \hat{\theta}_{nb}^*(x_k)).
\]

Homoscedastic case.

\[
\hat{v}_{b}^2 = \text{Var}_{1 \leq k \leq n} (Y_{b,k}^* - \hat{\theta}_{nb}^*(x_k) - \hat{b}_{b,k}^*) = \frac{1}{n} \sum_{k=1}^n (Y_{b,k}^* - \hat{\theta}_{nb}^*(x_k) - \hat{b}_{b,k}^*)^2.
\]

Heteroscedastic case.

\[
\hat{v}_{b,i}^2 = \text{Var}_{k \in M_n(x_i)} (Y_{b,k}^* - \hat{\theta}_{nb}^*(x_k) - \hat{b}_{b,k}^*) = \frac{1}{m_n} \sum_{k \in M_n(x_i)} (Y_{b,k}^* - \hat{\theta}_{nb}^*(x_k) - \hat{b}_{b,k}^*)^2,
\]

for \( x_i, i = (h + 1)/2, \ldots, n - (h - 1)/2, \ b = 1, \ldots, B \). Then the bootstrap estimate of the local bias and standard deviation at \( x_i \) are given by

\[
\hat{b}_i^* = \frac{1}{B} \sum_{b=1}^B \hat{b}_{b,i}^*, \quad \text{and} \quad \hat{v}_i^* = \frac{1}{B} \sum_{b=1}^B \hat{v}_{b,i}^*.
\]
The proposed bootstrap method is justified by the following theorem.

**Theorem 5.3.1.** Assume that the random error \( \varepsilon \)'s are independently and identically distributed with \( E(\varepsilon_i) = 0 \) and \( \text{Var}(\varepsilon_i) = \sigma^2 > 0 \) for \( i = 1, \ldots, n \). Then, under the conditions of Theorem 5.2.1, the bootstrap principle holds in the following sense:

\[
\lim_{n \to \infty} d_2 \left( \left\{ \left( \hat{\theta}_n^*(x_i) + \hat{\varepsilon}_i^* \right) - \left( \hat{\theta}_n(x_i) + \hat{\varepsilon}_i \right) \right\}, \left\{ \left( \hat{\theta}_n(x_i) + \hat{\varepsilon}_i \right) - \theta(x_i) \right\} \right) = 0,
\]

where \( d_2(F,G) \) is the Mallows metric between the distributions \( F \) and \( G \). The convention is that the random variables which appear in the arguments of the Mallows metric represent the corresponding distributions.

Theorem 5.3.1, together with Lemma 8.3 of Bickel and Freedman (1981), implies that the (conditionally known) bootstrap distribution of \( \left\{ \left( \hat{\theta}_n^*(x_i) + \hat{\varepsilon}_i^* \right) - \left( \hat{\theta}_n(x_i) + \hat{\varepsilon}_i \right) \right\} \) given the sample \( \{(x_i, Y_i)\}_{i=1}^n \) and the (unknown) distribution of \( \left\{ \left( \hat{\theta}_n(x_i) + \hat{\varepsilon}_i \right) - \theta(x_i) \right\} \) converge to the same distribution in probability as \( n \to \infty \). The bootstrap principle is applied to construct pointwise confidence intervals and a simultaneous confidence band for the true regression function \( \theta \) as follows.

**Step B4: Confidence Interval** Let \( z_{1,2}^l \) and \( z_{1,2}^u \) be the lower and upper percentiles, respectively, such that

\[
P \left( z_{1,2}^l < \frac{\hat{\theta}_n^*(x_i) + \hat{\varepsilon}_i^* - \hat{\theta}_n(x_i) + \hat{\varepsilon}_i}{\hat{\varepsilon}_i^*} < z_{1,2}^u \mid \mathcal{X} \right) = \alpha,
\]

where \( P( \cdot \mid \mathcal{X} ) \) is the bootstrap distribution. A confidence interval for \( \theta \) at \( x = x_i \) is given by

\[
\hat{I}_2 = (\hat{\theta}_n(x_i) + \hat{\varepsilon}_i - z_{1,2}^u \hat{\varepsilon}_i^*, \hat{\theta}_n(x_i) + \hat{\varepsilon}_i - z_{1,2}^l \hat{\varepsilon}_i^*).
\]
Step B5: Simultaneous Confidence Band Let $w_2$ be a solution of

$$P\left(\sup_{0<\chi<1}\left|\frac{\hat{\theta}_n(x_i) + \hat{b}_i}{\hat{\theta}_n(x_i)} - \{\hat{\theta}_n(x_i) + \hat{b}_i\}\right| < w_2 \chi\right) = \alpha.$$  

A simultaneous confidence band for $\theta$ is given by

$$\hat{J}_2 = \left\{(x, y) : x \in (0, 1), y = \hat{\theta}_n(x_i) + \hat{b}_i \pm w_2 \hat{\theta}_n\right\}.$$  

5.4 Simulation Study

To examine the performance of the proposed procedures for local estimation of the bias and the standard deviation of a regression spline and for construction of a confidence interval and band for $\theta(x)$, we conduct some simulation studies. In Section 5.4.1, three test functions are used and the corresponding empirical error probabilities of the confidence interval and band are calculated. In Section 5.4.2, the Cross-Validation criterion is employed to determine the optimal window size for the local estimation of the bias and standard error. In Example 1, the test function is smooth and the random error has a constant variance. Hence the bias of the regression spline estimator is very small relative to the total amplitude of variation of the test function. In Examples 2 and 3, the test function is based on “A Difficult Smoothing Problem” of Friedman and Silverman (1989). The error in Example 2 is homoscedastic, while that in Example 3 is heteroscedastic. In these examples, the continuity of the derivative of the test function is not satisfied. As will be seen, the error probability is reasonably acceptable if the continuous derivative assumption is satisfied. (See Example 1.)
5.4.1 Empirical error probability

To demonstrate the effect of the local bias correction, we examine the bootstrapping with and without the bias correction.

In the following models, cubic polynomial splines are fitted as initial estimates. The knots of the splines are placed at the 25%, 50% and 75% quantiles of \( \{x_i\}_{i=1}^n \). The significance level is \( \alpha = 0.05 \). The pointwise confidence intervals at \( x = 0.6 \) and simultaneous confidence bands are estimated. In Tables 5.1, 5.2, 5.3, \( pzc \) and \( pz_{nc} \) represent the empirical error probabilities of the confidence interval with and without bias correction, respectively, and \( pw_c \) and \( pw_{nc} \) stand for the empirical error probabilities of the simultaneous confidence band with and without bias correction, respectively. Here \( h \) is the percent of observations in each window for the local estimation of the bias and standard deviation with respect to the total number of the observations. For example, if \( h = 10\% \) and \( N = 100 \), then the number of observations in the local window is 10. Let \( B \) be the bootstrap resampling number and \( n \) be the number of observations. The empirical error probability is calculated based on 1000 runs.

Example 1. A random sample \( (x_j, y_j) \) of size \( n \) is generated from the following model:

\[
y_j = \sin(6x_j) + \epsilon_j, \quad j = 1, \ldots, n,
\]

where \( x_j \overset{iid}{\sim} U(0,1) \), \( \epsilon_j \overset{iid}{\sim} N(0, 1/9) \)

and the \( x_j \)'s and \( \epsilon_j \)'s are independent. Figure 5.1 displays the test function \( \sin(6x) \) and simultaneous confidence bands with and without bias correction for \( n = 300, B = 100 \) and \( h = 10\% \). Figure 5.2 displays the true bias of the regression spline and its bootstrap estimate. Table 5.1 summarizes the empirical error probabilities of the
Figure 5.1: Fitted curves and confidence bands in Example 1

confidence intervals and simultaneous confidence bands for different \((h, n, B)\) combinations.

The model in Example 1 satisfies the required regularity conditions. The amplitude of variability in the true bias is very small relative to the total variation of the test function. However, from Figure 5.2, we note that the bootstrap estimate of the local bias captures the oscillatory pattern in the true bias, although we could not place the initial estimate of the local bias on the same scale because it was too unstable and its variation was far greater than those of either the true bias or the bootstrap estimate.

In almost all cases the error probability of the bias corrected procedure is improved and closer to the target value 0.05 as suggested by Table 5.1. In general, the error probability gets closer to 0.05 as \(B\) and \(n\) increase. The error probability is improved
as the window size \( h \) increases from 10% to 20%, but it becomes worse as \( h \) is increased from 20% to 50%. This indicates that the local window is too wide to capture the local information about the relationship between the response variable and the predictor variable if it contain 50% of the observations.

**Example 2.** The second model is a constant variance version of the example given in Friedman and Silverman (1989), whereas the example of Friedman and Silverman (1989) itself is treated in Example 3. Here we have

\[
y_j = \begin{cases} 
  \epsilon_j & x_j \leq 0 \\
  \sin[2\pi(1 - x_j)^2] + \epsilon_j, & 0 < x_k \leq 1 
\end{cases}, \quad j = 1, \ldots, n,
\]

where \( x_j \overset{iid}{\sim} U(-0.2, 1), \epsilon_j \overset{iid}{\sim} N(0, 1/3^2) \)

and the \( x_j \)'s and \( \epsilon_j \)'s are independent. Figure 5.3 displays the true curve, the fitted curves and the confidence bands with and without the bias correction for
Table 5.1: Empirical Error Probability: Example 1.

<table>
<thead>
<tr>
<th>$h$</th>
<th>$B$</th>
<th>$n = 100$</th>
<th>$n = 300$</th>
<th>$n = 500$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>$p_{zc}$ : 0.082 $p_{zc} : 0.088$</td>
<td>$p_{zc} : 0.076$ $p_{zc} : 0.073$</td>
<td>$p_{zc} : 0.069$ $p_{zc} : 0.060$</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>$p_{zc} : 0.075$ $p_{zc} : 0.071$</td>
<td>$p_{zc} : 0.076$ $p_{zc} : 0.073$</td>
<td>$p_{zc} : 0.087$ $p_{zc} : 0.081$</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>$p_{zc} : 0.075$ $p_{zc} : 0.066$</td>
<td>$p_{zc} : 0.071$ $p_{zc} : 0.063$</td>
<td>$p_{zc} : 0.080$ $p_{zc} : 0.061$</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>$p_{zc} : 0.107$ $p_{zc} : 0.088$</td>
<td>$p_{zc} : 0.109$ $p_{zc} : 0.097$</td>
<td>$p_{zc} : 0.122$ $p_{zc} : 0.089$</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>$p_{zc} : 0.068$ $p_{zc} : 0.065$</td>
<td>$p_{zc} : 0.076$ $p_{zc} : 0.071$</td>
<td>$p_{zc} : 0.082$ $p_{zc} : 0.068$</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>$p_{zc} : 0.067$ $p_{zc} : 0.049$</td>
<td>$p_{zc} : 0.061$ $p_{zc} : 0.049$</td>
<td>$p_{zc} : 0.077$ $p_{zc} : 0.044$</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>$p_{zc} : 0.117$ $p_{zc} : 0.107$</td>
<td>$p_{zc} : 0.103$ $p_{zc} : 0.083$</td>
<td>$p_{zc} : 0.115$ $p_{zc} : 0.119$</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>$p_{zc} : 0.075$ $p_{zc} : 0.075$</td>
<td>$p_{zc} : 0.064$ $p_{zc} : 0.078$</td>
<td>$p_{zc} : 0.085$ $p_{zc} : 0.097$</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>$p_{zc} : 0.068$ $p_{zc} : 0.068$</td>
<td>$p_{zc} : 0.065$ $p_{zc} : 0.065$</td>
<td>$p_{zc} : 0.082$ $p_{zc} : 0.085$</td>
<td></td>
</tr>
</tbody>
</table>

$p_{zc}$ : error probability of C.I. with bias correction at $x = 0.6$

$p_{wz}$ : error probability of confidence bound with bias correction

$h$ : percentage of samples in a local window relative to the sample size

$B$ : number of bootstrap iteration; $n$ : number of observations.
Figure 5.3: Fitted curves and confidence bands in Example 2

\( n = 300, B = 100, h = 10\% \). The test function has a kink at \( x = 0 \). This means that the model assumption that \( f \in C^m[0,1], m \geq 1 \), is not satisfied at this point. This violation of the assumption causes a large bias of the fitted spline curve around the origin.

Figure 5.4 displays the initial and bootstrap estimates of the local bias. Both estimates are good fits to the true bias although the bootstrapped estimate provides a smoother curve. In Figure 5.3, we see that the bias of the original fitted curve is corrected, especially in the range of \((-0.2, 0.2)\).

Table 5.2 shows the empirical error probability of the confidence intervals and bands for each \((h, n, B)\) combination. For the confidence interval, we observe a pattern similar to that in Model I. That is, the case of \( h = 20\% \) is better than \( h = 10\% \) or
50%. The larger $n$ and $B$ yield an empirical error that is closer to 0.05. For the confidence bands, the empirical error probabilities are quite high because of the large bias that occurs in a neighborhood of the origin. For the error probabilities of the confidence bands with $h = 10\%$, we observe a large improvement in the bias-corrected version relative to the uncorrected versions. This implies that the smaller band width is better to see the local information about the relationship between the response and the independent variables.
<table>
<thead>
<tr>
<th>$h$</th>
<th>$B$</th>
<th>$n = 100$</th>
<th>$n = 300$</th>
<th>$n = 500$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>0.141 $p_z_{nc}$ : 0.267</td>
<td>0.139 $p_z_{nc}$ : 0.825</td>
<td>0.124 $p_z_{nc}$ : 0.994</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.099 $p_z_{c}$ : 0.160</td>
<td>0.076 $p_z_{c}$ : 0.178</td>
<td>0.066 $p_z_{c}$ : 0.269</td>
<td></td>
</tr>
<tr>
<td>100%</td>
<td>0.123 $p_z_{nc}$ : 0.202</td>
<td>0.106 $p_z_{nc}$ : 0.804</td>
<td>0.091 $p_z_{nc}$ : 0.992</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.073 $p_z_{c}$ : 0.144</td>
<td>0.055 $p_z_{c}$ : 0.131</td>
<td>0.058 $p_z_{c}$ : 0.242</td>
<td></td>
</tr>
<tr>
<td>200%</td>
<td>0.102 $p_z_{nc}$ : 0.174</td>
<td>0.091 $p_z_{nc}$ : 0.792</td>
<td>0.087 $p_z_{nc}$ : 0.988</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.066 $p_z_{c}$ : 0.131</td>
<td>0.045 $p_z_{c}$ : 0.125</td>
<td>0.060 $p_z_{c}$ : 0.238</td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td>0.179 $p_z_{nc}$ : 0.239</td>
<td>0.132 $p_z_{nc}$ : 0.825</td>
<td>0.131 $p_z_{nc}$ : 0.988</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.101 $p_z_{c}$ : 0.194</td>
<td>0.089 $p_z_{c}$ : 0.753</td>
<td>0.092 $p_z_{c}$ : 0.971</td>
<td></td>
</tr>
<tr>
<td>100%</td>
<td>0.132 $p_z_{nc}$ : 0.169</td>
<td>0.103 $p_z_{nc}$ : 0.743</td>
<td>0.090 $p_z_{nc}$ : 0.985</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.065 $p_z_{c}$ : 0.127</td>
<td>0.059 $p_z_{c}$ : 0.668</td>
<td>0.060 $p_z_{c}$ : 0.958</td>
<td></td>
</tr>
<tr>
<td>200%</td>
<td>0.106 $p_z_{nc}$ : 0.143</td>
<td>0.083 $p_z_{nc}$ : 0.730</td>
<td>0.063 $p_z_{nc}$ : 0.979</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.064 $p_z_{c}$ : 0.123</td>
<td>0.060 $p_z_{c}$ : 0.656</td>
<td>0.047 $p_z_{c}$ : 0.965</td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td>0.175 $p_z_{nc}$ : 0.237</td>
<td>0.132 $p_z_{nc}$ : 0.794</td>
<td>0.126 $p_z_{nc}$ : 1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.166 $p_z_{c}$ : 0.225</td>
<td>0.144 $p_z_{c}$ : 0.779</td>
<td>0.154 $p_z_{c}$ : 0.997</td>
<td></td>
</tr>
<tr>
<td>100%</td>
<td>0.129 $p_z_{nc}$ : 0.174</td>
<td>0.107 $p_z_{nc}$ : 0.767</td>
<td>0.118 $p_z_{nc}$ : 0.999</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.121 $p_z_{c}$ : 0.168</td>
<td>0.103 $p_z_{c}$ : 0.769</td>
<td>0.111 $p_z_{c}$ : 0.981</td>
<td></td>
</tr>
<tr>
<td>200%</td>
<td>0.114 $p_z_{nc}$ : 0.170</td>
<td>0.088 $p_z_{nc}$ : 0.765</td>
<td>0.077 $p_z_{nc}$ : 0.992</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.102 $p_z_{c}$ : 0.168</td>
<td>0.120 $p_z_{c}$ : 0.767</td>
<td>0.108 $p_z_{c}$ : 0.978</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2: Empirical Error Probability : Example 2.

$p_z_{nc}$: error probability of C.I. with no bias correction at $x = 0.6$

$p_z_{c}$: error probability of C.I. with bias correction at $x = 0.6$

$p_w_{nc}$: error probability of confidence bound with no bias correction

$p_w_{c}$: error probability of confidence bound with bias correction

$h$: percentage of samples in a local window relative to the sample size

$B$: number of bootstrap iteration; $n$: number of observations.
**Example 3.** The third model is the same as Example 2 except that a heteroscedastic error is assumed. Here we have

\[
y_j = \begin{cases} 
\epsilon_j & x_j \leq 0 \\
\sin[2\pi(1 - x_j)^2] + \epsilon_j & 0 < x_k \leq 1, \\
\end{cases} \quad j = 1, \ldots, n,
\]

where \( x_j \sim U(-0.2, 1), \epsilon_j \sim N(0, \max\{0.05, x_j\}^2) \)

and the \( x_j \)'s and \( \epsilon_j \)'s are independent. In this example the standard deviation is constant for \( x \leq 0 \) and increases linearly for \( x > 0 \). The model assumption, \( f \in C^m[0, 1], m \geq 1 \), is again not satisfied at \( x = 0 \) because of the kink of the test function.

Figure 5.5 displays the initial and bootstrap estimates of the local standard deviation. In Figure 5.5 we note that the local standard deviation is reasonably well estimated for small positive \( x \) around \( 0 < x < 0.4 \). However, there is considerable fluctuation in the estimates for \( x \) greater than 0.4. We believe that this is because of the increased error variation. The departure of the estimates from the true standard deviation for negative \( x \)'s is due to the boundary effects.

Figure 5.6 displays the initial and bootstrap estimates of the local bias. For small \( x < 0.2 \), the estimates capture the pattern of the true bias. For \( x > 0.2 \), the poor performances are due to relatively small true bias and the increased variability in the error term.

The empirical error probabilities are summarized in Table 5.3. For the confidence intervals, we observe a pattern similar to that in Example 2 although we do not see a large improvement when \( n \) is increased from 300 to 500. The error probabilities of the confidence bands are quite high. The simultaneous confidence band is very sensitive to the point of the true function which causes a large bias at certain \( x \) values.
Table 5.3: Empirical Error Probability in Example 3.

- \( p_{z_{nc}} \): error probability of C.I. with no bias correction at \( x = 0.6 \)
- \( p_z \): error probability of C.I. with bias correction at \( x = 0.6 \)
- \( p_{w_{nc}} \): error probability of confidence bound with no bias correction
- \( p_w \): error probability of confidence bound with bias correction
- \( h \): percentage of samples in a local window relative to the sample size
- \( B \): number of bootstrap iteration; \( n \): number of observations.

<table>
<thead>
<tr>
<th>( h )</th>
<th>( B )</th>
<th>( n = 100 )</th>
<th>( n = 300 )</th>
<th>( n = 500 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>( p_{z_{nc}} ): 0.089 ( p_w_{nc} ): 0.990</td>
<td>( p_{z_{nc}} ): 0.034 ( p_w_{nc} ): 1.000</td>
<td>( p_{z_{nc}} ): 0.036 ( p_w_{nc} ): 1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( p_z ): 0.069 ( p_w ): 0.668</td>
<td>( p_z ): 0.074 ( p_w ): 0.984</td>
<td>( p_z ): 0.060 ( p_w ): 0.999</td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>( p_{z_{nc}} ): 0.056 ( p_w_{nc} ): 0.979</td>
<td>( p_{z_{nc}} ): 0.023 ( p_w_{nc} ): 1.000</td>
<td>( p_{z_{nc}} ): 0.024 ( p_w_{nc} ): 1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( p_z ): 0.065 ( p_w ): 0.590</td>
<td>( p_z ): 0.053 ( p_w ): 0.973</td>
<td>( p_z ): 0.045 ( p_w ): 0.998</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>( p_{z_{nc}} ): 0.038 ( p_w_{nc} ): 0.974</td>
<td>( p_{z_{nc}} ): 0.015 ( p_w_{nc} ): 1.000</td>
<td>( p_{z_{nc}} ): 0.026 ( p_w_{nc} ): 1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( p_z ): 0.051 ( p_w ): 0.529</td>
<td>( p_z ): 0.049 ( p_w ): 0.976</td>
<td>( p_z ): 0.048 ( p_w ): 0.997</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>( p_{z_{nc}} ): 0.107 ( p_w_{nc} ): 0.990</td>
<td>( p_{z_{nc}} ): 0.101 ( p_w_{nc} ): 1.000</td>
<td>( p_{z_{nc}} ): 0.073 ( p_w_{nc} ): 1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( p_z ): 0.096 ( p_w ): 0.789</td>
<td>( p_z ): 0.086 ( p_w ): 1.000</td>
<td>( p_z ): 0.075 ( p_w ): 1.000</td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>( p_{z_{nc}} ): 0.068 ( p_w_{nc} ): 0.984</td>
<td>( p_{z_{nc}} ): 0.071 ( p_w_{nc} ): 1.000</td>
<td>( p_{z_{nc}} ): 0.079 ( p_w_{nc} ): 1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( p_z ): 0.070 ( p_w ): 0.707</td>
<td>( p_z ): 0.064 ( p_w ): 1.000</td>
<td>( p_z ): 0.075 ( p_w ): 1.000</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>( p_{z_{nc}} ): 0.067 ( p_w_{nc} ): 0.980</td>
<td>( p_{z_{nc}} ): 0.076 ( p_w_{nc} ): 1.000</td>
<td>( p_{z_{nc}} ): 0.060 ( p_w_{nc} ): 1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( p_z ): 0.051 ( p_w ): 0.698</td>
<td>( p_z ): 0.060 ( p_w ): 0.999</td>
<td>( p_z ): 0.057 ( p_w ): 1.000</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>( p_{z_{nc}} ): 0.117 ( p_w_{nc} ): 1.000</td>
<td>( p_{z_{nc}} ): 0.115 ( p_w_{nc} ): 0.100</td>
<td>( p_{z_{nc}} ): 0.120 ( p_w_{nc} ): 1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( p_z ): 0.146 ( p_w ): 0.733</td>
<td>( p_z ): 0.138 ( p_w ): 0.993</td>
<td>( p_z ): 0.136 ( p_w ): 1.000</td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>( p_{z_{nc}} ): 0.075 ( p_w_{nc} ): 0.969</td>
<td>( p_{z_{nc}} ): 0.106 ( p_w_{nc} ): 1.000</td>
<td>( p_{z_{nc}} ): 0.095 ( p_w_{nc} ): 1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( p_z ): 0.115 ( p_w ): 0.724</td>
<td>( p_z ): 0.111 ( p_w ): 0.977</td>
<td>( p_z ): 0.107 ( p_w ): 1.000</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>( p_{z_{nc}} ): 0.068 ( p_w_{nc} ): 0.960</td>
<td>( p_{z_{nc}} ): 0.081 ( p_w_{nc} ): 1.000</td>
<td>( p_{z_{nc}} ): 0.080 ( p_w_{nc} ): 1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( p_z ): 0.095 ( p_w ): 0.721</td>
<td>( p_z ): 0.089 ( p_w ): 0.968</td>
<td>( p_z ): 0.081 ( p_w ): 0.993</td>
<td></td>
</tr>
</tbody>
</table>
From Tables 5.1, 5.2 and 5.3, we note that the bootstrap procedure provides better empirical error probabilities and smoother estimates of the bias and standard deviation. The bias corrected procedure presents the empirical error probability smaller and closer to the target value $\alpha = 0.05$.

### 5.4.2 Cross-Validation

In this subsection, we apply the Cross-Validation criterion to Example 2 to demonstrate the selection method for choosing a better window size for the local estimation of bias and standard deviation.

Let $\hat{\theta}^{(-i)}$ be the initial (not bias-corrected) regression spline fit computed by removing $(x_i, y_i)$ from the data. Let $\hat{b}_{h,k}^{(-i)}$ be the initial bias estimate at $x_k$ with the
window size $h$ based on $\hat{\theta}^{(-i)}$. Since the proposed bias-corrected version on the interval $[x_{i-1}, x_{i+1}]$ is a linear interpolation between the bias-corrected fit at $x_{i-1}$ and $x_{i+1}$, the bias corrected fit on $x_i$ without using $(x_i, y_i)$ is represented as follows:

$$\tilde{\theta}_h^{(-i)}(x_i) = (\hat{\theta}^{(-i)}(x_{i-1}) + \hat{\theta}_h^{(-i)}(x_{i+1})) + \frac{(\hat{\theta}^{(-i)}(x_{i+1}) + \hat{\theta}_h^{(-i)}(x_{i+1}) - \hat{\theta}^{(-i)}(x_{i-1}) - \hat{\theta}_h^{(-i)}(x_{i-1}))}{x_{i+1} - x_{i-1}}.$$

Then the Cross-Validation for the initial bias-corrected fit with the window size $h$ is given by

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{\theta}_h^{(-i)}(x_i))^2.$$

Figure 5.7 displays a curve of the Cross-validation values for Example 2 with $n = 500$ and various $h$ values.
Figure 5.7: Cross-Validation in Example 2, $h$: number of observations in a local window

Figure 5.7 shows that the Cross-validation curve has a valley at $h = 40$. Then the CV value goes up as $h$ is increased to $h = 125$ and the curve stays flat for larger values of $h$. This result suggests using 40 (or about 10% of the total number of observations) nearest neighborhoods in a local window to estimate the local bias and standard deviation. As we can see, this observation coincides with our finding for the bootstrap version of Example 2 in the previous subsection, where the empirical error probabilities for 10% or 20% of the observations in a window are better than those for 50% of the observations.
CHAPTER 6

FURTHER RESEARCH

In this dissertation, we have proposed a new methodology for knot selection of free knot splines using an Evolutionary Algorithm and the adaptive model selection criterion in nonparametric regression.

The methodology can be generalized to a multivariate regression model represented as:

\[ Y_i = \theta(x_i) + \epsilon_i, \quad i = 1, \ldots, n, \]

where \( x = (x_1, \ldots, x_p) \). A natural candidate to estimate the regression function \( \theta \) is the tensor product spline estimator. For example, a tensor product spline with \( p = 2 \) is written as follows:

\[
s(x^1, x^2) = \sum_{i, j \in S} \alpha_{ij} B^1_i(x^1, t^1) B^2_j(x^2, t^2),
\]

with \( t^i = (0 = t^i_0 < t^i_1 \leq \cdots \leq t^i_{k^i} < t^i_{k^i+1} = 1) \), \( i = 1, 2 \) and

\[ S \subset \{(l, l') : 0 \leq l \leq k^1, 0 \leq l' \leq k^2\} \]

is a set of knots. Then to select knots an Evolutionary Algorithm is applied to \( S \). In this case, however, the covariance structure in the predictor vector \( x \) should be taken into account in designing the mutation distribution of the EA.
The methodology can also be applied to the additive model:

$$Y_i = \sum_{j=1}^{p} \theta_j(x_j) + \epsilon_i, \quad i = 1, \ldots, n. \quad (6.2)$$

See Hastie and Tibshirani (1990). The spline representation for the Generalized Additive Model (GAM) is as follows:

$$s(x) = \alpha + \sum_{j=1}^{p} s_j(x_j),$$

where $s_j(x), j = 1, \ldots, p$, are univariate spline functions. Then AFKS can be alternately applied to the individual $s_j$'s. It is of great interest to estimate the local behavior of these multivariate regression estimators, such as the local bias and variance, and to construct a confidence interval, although the asymptotic theory for these candidate estimators has not yet been established.

In many practical situations, the error variance is not constant. In this case, the regression model (1.1) is generalized to (5.5) and the regression spline estimator

$$\hat{\theta}_n(x, h) = \sum_{l=1}^{k+m} \hat{\alpha}_l B_l(x, t)$$

is estimated by minimizing the weighted least squares criterion:

$$\min_{\alpha_1, \ldots, \alpha_{k+m}} \sum_{i=1}^{n} w^{-1}(x_i) \left( Y_i - \sum_{l=1}^{k+m} \alpha_l B_l(x_i, t) \right)^2, \quad (6.3)$$

where $w(\cdot)$ is defined in (5.5). The consistent estimators, $\hat{\sigma}_i^2$ and $\hat{\sigma}_i^2$ with and without bootstrapping, of the local variance $\text{Var}(\epsilon_i) = \sigma^2 w(x_i)$ are defined in Chapter 5. If $\sigma^2 w(x_i)$ can be effectively estimated by $\hat{\sigma}_i^2$ or $\hat{\sigma}_i^2$, then the estimate of the local variance is substituted in (6.3) to fit the heterogeneous model.

In Chapter 5, we developed the theory for prediction using the fixed knot regression spline estimator. As seen in previous chapters, the adaptive spline with variable multiple knots adapts to inhomogeneous functions. Furthermore, the fitted value of
multiple knots with a sharp kink or discontinuity can be used as a point estimator of the change point of the regression function. Hence the distribution theory of the prediction and the knot location for an adaptive spline is of vital importance. On the other hand, with variable multiple knots the elements and the dimension of the hat matrix $M(h)$ are functions of $\{(x_i, Y_i)\}_{i=1}^n$ and are correlated with the response variable $Y$. Hence the usual regression model is not applicable and the distribution theory for the adaptive spline estimator has not yet been established. Gong (1986) and Breiman (1992) studied an application of the bootstrap technique to variable selection for a linear regression model. This approach may also help for the distribution theory of the adaptive spline estimator.
APPENDIX A

TECHNICAL PROOFS

A.1 Proofs of Theorems in Chapters 3 and 4

The proofs of Theorem 3.3.1 and Theorem 4.2.1 follow from a slightly more general result in Theorem A.1.1. Consider the minimization problem $f^* = \min_{x \in S} f(x)$ as defined in (3.7) and (4.6) via a general EA, where the mutation distributions can be arbitrary. Let $\mathcal{P}^{(\nu)} = \{x^{(1,\nu)}, \ldots, x^{(\xi,\nu)}\}$ be the population of an EA in the $\nu$-th generation and $f_{(\nu)}^* = \min\{f(x^{(1,\nu)}), \ldots, f(x^{(\xi,\nu)})\}$. Let $y^{(\nu)}$ be a solution such that $f(y^{(\nu)}) = \max_{1 \leq i \leq \xi} f(x^{(i,\nu)})$. In other words, $y^{(\nu)}$ is the worst individual in the $\nu$-th generation. Define $P_{(\nu)}(y, A)$ to be a transition probability of $y^{(\nu)}$ in the $\nu$-th generation from $y \in S$ to $A \subset S$ determined by the mutation distribution, i.e., $P_{(\nu)}(y, A) = P(y^{(\nu)} \in A | y^{(\nu-1)} = y)$. Denote by $P_{(\nu)}(y, A)$ the $\nu$-step transition probability. Let $A_\epsilon = \{x \in S : f(x) < f^* + \epsilon\}$.

Clearly, $\{f_{(\nu)}^*\}_{\nu=0}^\infty$ is a stochastic process indexed by $\nu$. Since the state of $f_{(\nu)}^*$ only depends on $\mathcal{P}^{(\nu-1)}$ and the mutation distribution for $(\nu-1)$-th generation, $\{f_{(\nu)}^*\}_{\nu=0}^\infty$ forms an inhomogeneous Markov chain. Then we have the following convergence property.
Theorem A.1.1. Let $Z^{(i,\nu)}$ be the mutation variable for $x^{(i,\nu)}$. Assume that for any $\nu \geq 0$ and $x^{(i,\nu)} \in S$, $i = 1, \ldots, \xi$, 

$$\forall \epsilon > 0, \exists \delta > 0, P(X^{(\xi+i,\nu)} \in A_\epsilon) = P(x^{(i,\nu)} + Z^{(i,\nu)} \in A_\epsilon) \geq \delta. \quad (A.1)$$

Then

$$f^{(\nu)} \rightarrow f^* = \min_{x \in S} f(x) \text{ a.s. as } \nu \rightarrow \infty.$$

Before we prove Theorem A.1.1, we need the following lemma.

Lemma 1. Under the assumption of Theorem A.1.1, for $\nu \geq 1$, and $\epsilon > 0,$

$$P^{(\nu)}(y, A_\epsilon) \geq 1 - (1 - \delta)^\nu.$$

Proof. If $y^{(\nu-1)} = y \in A_\epsilon$, then $f(y^{(\nu-1)}) < f^* + \epsilon$, which implies, by the top-$\xi$ selection rule, that $f(y^{(\nu)}) \leq f(y^{(\nu-1)})$. Hence $y^{(\nu)} \in A_\epsilon$ and $P^{(\nu)}(y, A_\epsilon) = 1$. On the other hand, for any $y^{(\nu-1)} = y \in S \setminus A_\epsilon$, by (A.1),

$$P^{(\nu)}(y, A_\epsilon) = P(y^{(\nu)} \in A_\epsilon | y^{(\nu-1)} = y) \geq P(y^{(\nu-1)} + Z^{(\nu-1)} \in A_\epsilon) \geq \delta$$

where $Z^{(\nu-1)}$ is the mutation variable for $y^{(\nu-1)}$. The conclusion follows by induction; see Rudolph (1997).

Proof of Theorem A.1.1. It suffices to show that $\sum_{\nu=0}^{\infty} P(f^{(\nu)} - f^* > \epsilon) < \infty$ for any $\epsilon > 0$. By Lemma 1,

$$P(f^{(\nu)} - f^* > \epsilon) \leq P(f(Y^{(\nu)}) > f^* + \epsilon) = P(Y^{(\nu)} \notin A_\epsilon) \leq (1 - \delta)^\nu.$$

Therefore $\sum_{\nu=0}^{\infty} P(f^{(\nu)} - f^* > \epsilon) \leq 1 + \sum_{\nu=1}^{\infty} (1 - \delta)^\nu = 1/\delta < \infty$. This completes the proof. \qed
Proof of Theorem 3.3.1. It suffices to verify that our EA satisfies (A.1). This is so because the variances of the mutation distributions are set to be greater than a prefixed constant that is independent of \( \nu \). □

Proof of Theorem 4.2.1. The proof is similar to that of Theorem 3.3.1. □

A.2 Proofs of Theorems in Chapter 5

Proof of Theorem 5.2.1. Since the size of the window for the local estimation shrinks to 0, for each \( x_i \in (t_{I^*}, t_{I^*+1}] \), \( l^* = 0, \ldots, k \),

\[ \exists N \in \mathbb{N}, \forall n \geq N, \forall x_k \in M_n(x_i), x_k \in (t_{I^*}, t_{I^*+1}] \]

From Theorem 5.1.1,

\[ \theta(x_k) - E(\hat{\theta}(x_k)) = b(x_k) + o(u^m), \tag{A.2} \]

where \( b(x_k) = \frac{\theta^{(d)}(x_k) u_{I^*}^d}{d!} B_d \left( \frac{x_k - t_{I^*}}{u_{I^*}} \right) \), for \( x_k \in (t_{I^*}, t_{I^*+1}] \).

Let \( J_{k1} \) be a column vector whose elements are 0 except for \( k \)-th element which equals to 1. Then (A.2) is equivalent to

\[ J_{k1}'(I - M)\theta(X) = b(x_k) + o(u^m). \tag{A.3} \]

where \( M = B(x)(B(x)'B(x))^{-1}B(x)' \), \( [B(x)]_{il} = B_l(x_i; t), i = 1, \ldots, n, l = 1, \ldots, k + m \). Since

\[ |\hat{b}_i - J_{i1}'(I - M)\theta(X)| \leq |\hat{b}_i - b(x_i)| + |J_{i1}'(I - M)\theta(X) - b(x_i)| \]

\[ = |\hat{b}_i - b(x_i)| + o(u^m), \quad \text{by (A.3)} \]

we may want to show that

\[ |\hat{b}_i - b(x_i)| = o_p(n^{-\kappa}), \text{ as } n \to \infty, \]

68
where \( \kappa = \min\{\alpha - \beta, \nu\beta\}, 0 < \nu < 1/2 \). By the definition of the local bias estimator \( \hat{b}_i \), we have

\[
\hat{b}_i = \frac{1}{m_n} \sum_{k \in M_n(x_i)} (y_k - \hat{\theta}(x_k))
\]

\[
= \frac{1}{m_n} \sum_{k \in M_n(x_i)} J'_{k_1}(I - M)(\theta(X) + \epsilon)
\]

\[
= \frac{1}{m_n} \sum_{k \in M_n(x_i)} (J'_{k_1}(I - M)\theta(X) + \epsilon_k - J'_{k_1}M\epsilon)
\]

\[
= \frac{1}{m_n} \sum_{k \in M_n(x_i)} (b(x_k) + o(u^m) + \epsilon_k - J'_{k_1}M\epsilon)
\]

by (A.3)

\[
= \frac{1}{m_n} \sum_{k \in M_n(x_i)} b(x_k) + \frac{1}{m_n} \sum_{k \in M_n(x_i)} \epsilon_k - \frac{1}{m_n} J'_{k_1}(x_i)M\epsilon + o(u^m)
\]

\[
= I + II + III + o(u^m)
\]

(A.4)

where \( J_n(x_i) = \sum_{k \in M_n(x_i)} J_{k_1} \). Since \( \theta^{(m)} \) and the Bernoulli polynomial are differentiable on \( (M^L_n, M^U_n) \), by the mean value theorem,

\[
\forall k \in M_n(x_i), \exists c_k \in (M^L_n, M^U_n), b(x_k) = b(x_i) + b'(c_k)|x_k - x_i|.
\]

(A.5)

Then

\[
I = \frac{1}{m_n} \sum_{k \in M_n(x_i)} b(x_k) = \frac{1}{m_n} \sum_{k \in M_n(x_i)} \{b(x_i) + b'(c_k)|x_k - x_i|\}
\]

\[
= b(x_i) + R_{n,i},
\]

\[
|R_{n,i}| \leq \frac{1}{m_n} \sum_{k \in M_n(x_i)} |b'(c_k)||x_k - x_i| \leq \sup_{z \in (M^L_n, M^U_n)} |b'(z)|(M^U_n - M^L_n) = o(n^{-\alpha + \beta})
\]

(A.6)

and \( |I - b(x_i)| = o(n^{-\alpha + \beta}) \).

Since \( \text{Var}(\epsilon_i) = \mathbb{E}(\epsilon_i^2) = \sigma w(x_i) \) and \( w \) is continuous on \([0, 1]\), \( \text{Var}(\epsilon_i)'s \) are uniformly bounded. Then, by the weak law of large numbers, we have , for \( 0 < \nu < 1/2 \),

\[
II = o_p(m_n^{-\nu}) = o_p(n^{-\beta \nu}).
\]

(A.7)
By the Chebyshev's inequality, for any $\eta > 0$,

$$P(\| III \| > \eta) \leq (m_n \eta)^{-2} E(J'_n(x_i) M \epsilon' M J_n(x_i))$$

$$= \frac{\sigma^2}{(m_n \eta)^2} J'_n(x_i) M W M J_n(x_i) \quad W = \text{diag}(w(x_1), \ldots, w(x_n))$$

$$\leq \frac{\sigma^2 w_{\text{max}}}{(m_n \eta)^2} J'_n(x_i) M J_n(x_i) \quad w_{\text{max}} = \max_{x \in [0,1]} w(x).$$

Let $G_{n,k} = n^{-1} B(x)' B(x)$. Since $G_{n,k}$ is non-negative definite, there exists an orthogonal matrix $C$ such that $G_{n,k} = C \Lambda C'$ where $\Lambda$ is a diagonal matrix with eigenvalues of $G_{n,k}$. Let $\lambda_{\text{min}}$ and $\lambda_{\text{max}}$ be the minimum and maximum eigenvalues of $G_{n,k}$, respectively. Then, by Lemma 6.2 of Zhou, Shen and Wolfe (1998)

$$J'_n(x_i) P J_n(x_i) = J'_n(x_i) B(x) (B(x)' B(x))^{-1} B(x)' J_n(x_i)$$

$$= \frac{1}{n} J'_n(x_i) B(x) G_{n,k}^{-1} B(x)' J_n(x_i)$$

$$\leq \frac{1}{\lambda_{\text{min}}} J'_n(x_i) (n^{-1} B(x) B(x)') J_n(x_i)$$

$$\leq \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} J'_n(x_i) J_n(x_i) = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} m_n$$

for some positive constants $c_1, c_2$. Hence

$$P(\| III \| > \eta) \leq \frac{\sigma^2 w_{\text{max}} c_2 + o(1)}{m_n \lambda^2} \frac{c_3}{c_1 + o(1) m_n} = \frac{c_3}{m_n \eta^2}$$

where $c_3$ is a constant. Hence, for $0 < \nu < 1/2$,

$$III = o_p(n^{-\nu}) = o_p(n^{-\beta \nu}). \quad (A.8)$$

Therefore, by (A.4), (A.6), (A.7), (A.8), we have

$$|\hat{b}_i - J'_i(I - M)\theta(X)| \leq |\hat{b}_i - b(x_i)| + o(u^m) = o(n^{-(\alpha - \beta)}) + o_p(n^{-\beta \nu}) + o(u^m)$$

$$= o(n^{-\kappa}) + o(u^m)$$

70
where $\kappa = \min\{(\alpha - \beta), \beta \nu\}$. Hence $\hat{b}_i$ converges to the true bias of $\hat{\theta}$ at $x_i$, $J'_{k1}(I - M)\theta(X)$, in probability, as $n \to \infty$. \hfill \square

**Lemma 2.** If $m_n = o(nu)$, then

$$\lim_{n \to \infty} \sum_{k \in \mathcal{M}_n(x_i)} J'_{k1}MJ_{k1} = 0.$$  

**Proof.** Let $G_{n,k} = n^{-1}B(x)'B(x)$. Then

$$n^2(J'_{k1}MJ_{k1})^2 = (J'_{k1}B(x)G_{n,k}^{-1}B(x)'J_{k1})^2 \leq \lambda_{\min}^{-2}(J'_{k1}B(x)B(x)'J_{k1})^2 \leq \lambda_{\min}^{-2} \sum_{l=1}^{k+m} B_l(x_k; t)^2.$$  

Using the property of the normalized B-splines, for any $x \in [0, 1]$,

$$0 \leq B_l(x; t) \leq 1, \quad \sum_{l=1}^{k+m} B_l(x; t) = 1, \quad l = 1, \ldots, k + m.$$  

See, for example, de Boor (1978), p.110, we obtain that $\sum_{l=1}^{k+m} B_l(x_k; t)^2 = \sum_{l=1}^{k+m} B_l(x_k; t) = 1$, which implies

$$n^2(J'_{k1}MJ_{k1})^2 \leq \lambda_{\min}^{-2},$$

and

$$|J'_{k1}MJ_{k1}| \leq \frac{1}{\lambda_{\min}n}.$$  

By Lemma 6.2 of Zhou, Shen and Wolfe (1998)

$$\left| \sum_{k \in \mathcal{M}_n(x_i)} J'_{k1}MJ_{k1} \right| \leq \frac{m_n}{\lambda_{\min}n} \leq \frac{m_n}{(c_1 + o(1))nh} = o(1).$$  

\hfill \square
Proof of Theorem 5.2.2. Note that

\[ \hat{v}_i^2 = \frac{1}{m_n} \sum_{k \in M_n(x_i)} (y_k - \hat{\theta}(x_k) - \hat{\epsilon}_k)^2 \]

\[ = \frac{1}{m_n} \sum_{k \in M_n(x_i)} (J'_{k_1}(\theta(X) + \epsilon) - J'_{k_1}M(\theta(X) + \epsilon) - \hat{\epsilon}_k)^2 \]

\[ = \frac{1}{m_n} \sum_{k \in M_n(x_i)} ((J'_{k_1}(I - M)\theta(X) - \hat{\epsilon}_k) + J'_{k_1}(I - M)\epsilon)^2 \]

\[ = \frac{1}{m_n} \sum_{k \in M_n(x_i)} \left\{ (J'_{k_1}(I - M)\theta(X) - \hat{\epsilon}_k)^2 + 2(J'_{k_1}(I - M)\theta(X) - \hat{\epsilon}_k)(J'_{k_1}(I - M)\epsilon) \right. \]

\[ + \left. J'_{k_1}(I - M)\epsilon\epsilon'(I - M)J_{k_1} \right\} \]

\[ = I + II + III. \quad (A.9) \]

By (A.3), (A.4) and Theorem 5.2.1, we have

\[ |J'_{k_1}(I - M)\theta(X) - \hat{\epsilon}_k| \leq |J'_{k_1}(I - M)\theta(X) - b(x_k)| + |b(x_k) - \hat{\epsilon}_k| \]

\[ = o(u^m) + o_p(n^{-\kappa}). \quad (A.10) \]

Hence \( J'_{k_1}(I - M)\theta(X) - \hat{\epsilon}_k \) converges to zero in probability and its convergence is independent of \( k \in M_n(x_i) \). Thus

\[ I = o_p(n^{-\kappa}) + o(u^m). \quad (A.11) \]

and converges to zero in probability as \( n \to \infty \). Next,

\[ |II| \leq \frac{2}{m_n} \sum_{k \in M_n(x_i)} |J'_{k_1}(I - M)\theta(X) - \hat{\epsilon}_k||J'_{k_1}(I - M)\epsilon| \]

\[ \leq (o_p(n^{-\kappa}) + o(u^m)) \frac{2}{m_n} \sum_{k \in M_n(x_i)} J'_{k_1}|(I - M)\epsilon| \quad \text{by (A.10)} \]

\[ = (o_p(n^{-\kappa}) + o(u^m)) \frac{2}{m_n} J'_{n}(x_i)|(I - M)\epsilon|. \]
For any \( \lambda > 0 \), by Chebyshev's inequality and the fact that \( M \) is idempotent and non-negative definite,

\[
P(2m_n^{-1} J'_n(x_i)|(I - M)\epsilon| > \lambda) \leq 4(m_n \lambda)^{-2} E(J'_n(x_i)(I - M)\epsilon\epsilon'(I - M)J_n(x_i))
\]

\[
= 4(m_n \lambda)^{-2} J'_n(x_i)(I - M)\sigma^2 W(I - M)J_n(x_i)
\]

\[
\leq \frac{4\sigma^2 w_{\text{max}}}{(m_n \lambda)^2} J'_n(x_i)(I - M)J_n(x_i)
\]

\[
= \frac{4\sigma^2 w_{\text{max}}}{(m_n \lambda)^2} (m_n - J'_n(x_i)M J_n(x_i))
\]

\[
\leq \frac{4\sigma^2 w_{\text{max}}}{\lambda^2 m_n}
\]

where \( w_{\text{max}} = \max_{x \in [0,1]} w(x) \). Therefore, for \( 0 < \nu < 1/2 \),

\[
\Pi = (o_p(n^{-\kappa}) + o(p_n^m)) o_p(m_n^{-\nu}) = (o_p(n^{-\kappa}) + o(u^m)) o_p(n^{-\beta\nu})
\]  

(A.12)

and converges to zero in probability as \( n \to \infty \). Now let

\[
\text{III} = \frac{1}{m_n} \sum_{k \in M_n(x_i)} J'_{k1}(I - M)\epsilon\epsilon'(I - M)J_{k1}
\]

\[
= \frac{1}{m_n} \sum_{k \in M_n(x_i)} \{ J'_{k1}\epsilon\epsilon'J_{k1} - 2J'_{k1}M\epsilon\epsilon'J_{k1} + J'_{k1}M\epsilon\epsilon'MJ_{k1} \}
\]

\[
= (i) + (ii) + (iii).
\]  

(A.13)

By the weak law of large numbers, as \( n \to \infty \),

\[
\frac{1}{m_n} \sum_{k \in M_n(x_i)} \epsilon_k^2 - \frac{1}{m_n} \sum_{k \in M_n(x_i)} \sigma^2 w(x_k) \xrightarrow{p} 0.
\]

Since \( w(\cdot) \) is continuous,

\[
\frac{1}{m_n} \sum_{k \in M_n(x_i)} \sigma^2 w(x_k) - \sigma^2 w(x_i) \to 0,
\]

we have

\[
(i) \xrightarrow{p} \sigma^2 w(x_i).
\]  

(A.14)
By Chebyshev's inequality, $\forall \eta > 0$,

$$P(|(ii)| > \eta) \leq \frac{4}{(m_n \eta)^2} \mathbb{E} \left( \sum_{k \in M_n(x_i)} J'_{k1} M \epsilon \epsilon' J_{k1} \right)^2$$

$$\leq \frac{4}{(m_n \eta)^2} \mathbb{E} J'_n(x_i) M \epsilon \epsilon' J'_n(x_i) J'_n(x_i) \epsilon \epsilon' M J_n(x_i)$$

$$= \frac{4}{(m_n \eta)^2} J'_n(x_i) M \mathbb{E} \left( \sum_{k \in M_n(x_i)} \epsilon_k \right)^2 \epsilon \epsilon' M J_n(x_i),$$

where $J'_n(x_i) \epsilon = \sum_{k \in M_n(x_i)} \epsilon_k$. Since $\{\epsilon_k\}_{k=1}^n$ are independent each other, the $(i', j')$-element of the expected value in the above expression is non-zero only if $i' = j' \in M_n(x_i)$, which is equal to $\mathbb{E}(\epsilon_i^4) \leq C < \infty$. Thus

$$P(|(ii)| > \eta) \leq \frac{4C}{(m_n \eta)^2} J'_n(x_i) M J_n(x_i) J'_n M J_n(x_i)$$

$$\leq \frac{4C}{(m_n \eta)^2} \text{tr}[J'_n(x_i) M J_n(x_i)]^2$$

$$= \frac{4C}{(m_n \eta)^2} \lambda_{\text{max}}^2 (k + m)^2.$$

Then

$$(ii) = O_p(m_n^{-1}) = O_p(n^{-\beta}) = o_p(n^{-\kappa}), \quad (A.15)$$

and (ii) converges to 0 in probability. Finally, for any $\lambda > 0$, by the Markov's inequality,

$$P((iii) > \lambda) \leq \frac{\sigma^2}{m_n \lambda} \sum_{k \in M_n(x_i)} J'_{k1} M W M J_{k1}$$

$$= \frac{\sigma^2 \omega_{\text{max}}}{m_n \lambda} \sum_{k \in M_n(x_i)} J'_{k1} M J_{k1}$$

$$= \frac{\sigma^2 \omega_{\text{max}}}{\lambda m_n} o(1) \quad \text{by Lemma 2.}$$

Thus

$$(iii) = o_p(m_n^{-1}) = o_p(n^{-2\kappa}) \quad (A.16)$$

74
and (iii) converges to 0 in probability. Therefore

$$\lim N \to \infty, \quad N \to \infty.$$

By (A.9), (A.11), (A.12) and (A.17), we have

$$\hat{v}_i^2 \to \sigma^2 w(x_i), \quad \text{as } N \to \infty.$$

Lemma 3 (de Boor (1978), p.40). Let $[g]$ be a piecewise linear interpolation of $g$ at points $\tau_1 < \cdots < \tau_n$ which is defined by

$$[f](x) = \theta(\tau_i) + (x - \tau_i) \frac{\theta(\tau_{i+1}) - \theta(\tau_i)}{\tau_{i+1} - \tau_i}, \quad \tau_i \leq x \leq \tau_{i+1}, \; i = 1, \ldots, n - 1.$$

If $g \in C^2[0,1]$, then we have

$$|g - [g]| \leq \frac{1}{8} |\tau|^2 \max_{0 \leq \xi \leq 1} |g''(\xi)|,$$

where $\tau = \max_i (\tau_{i+1} - \tau_i)$.

Proof. By Theorem I.2, de Boor (1978), p.8, Osculatory interpolation, we have

$$g(x) = g(\tau_i) + (x - \tau_i)[\tau_i, \tau_{i+1}]g + (x - \tau_i)(x - \tau_{i+1})[\tau_i, \tau_{i+1}, x]g$$

where $[\xi_i, \ldots, \xi_{i+k}]g$ is the $k$-th order divided difference of $g$. Then

$$g - [g] = g(x) - (g(\tau_i) + (x - \tau_i)[\tau_i, \tau_{i+1}]g + (x - \tau_i)(x - \tau_{i+1})[\tau_i, \tau_{i+1}, x]g.$$}

The maximum value of $(x - \tau_i)(x - \tau_{i+1})$ occurs at $(\tau_{i+1} - \tau_i)/2$ and is $(\Delta \tau_i/2)^2$ where $\Delta \tau_i = |\tau_i - \tau_{i+1}|$. On the other hand, by (vii) of de Boor (1978), p.8,

$$\exists \xi \in [\tau_i, \tau_{i+1}], \; [\tau_i, \tau_{i+1}, x]g = \frac{1}{2!} g^{(2)}(\xi).$$
Then we have

$$|g - [g]| \leq \frac{1}{8} \tau^2 \max_{0 \leq x \leq 1} |g''(x)|$$

where $\tau = \max_{i=1, \ldots, n-1} \Delta t_i$. \hfill $\square$

**Proof of Theorem 5.2.3.** Let $[\theta]$ be a piecewise linear interpolation of $\theta$ at points $x_1, \ldots, x_n$. Then

$$\sup_{\eta < x < 1} |\tilde{\theta}(x) - \theta(x)| \leq \sup_{\eta < x < 1} |\tilde{\theta}(x) - [\theta](x)| + \sup_{\eta < x < 1} |[\theta](x) - \theta(x)|$$

$$= I + II. \quad (A.18)$$

By Lemma 3 and (5.3),

$$II \leq \frac{1}{8} \left( \max_{2 \leq i \leq n} (x_i - x_{i-1}) \right)^2 \sup_{x \in [0,1]} |\theta''(x)| = \frac{1}{8} d_n^2 \sup_{x \in [0,1]} |\theta''(x)| = o(n^{-2\alpha}), \ 0 < \alpha \leq 1. \quad (A.19)$$

Since $\tilde{\theta}$ and $[\theta]$ are both piecewise linear at points $x_1, \ldots, x_n$, the maximum difference between these two occurs at one of $x_1, \ldots, x_n$. Then

$$I = \max_{\eta \leq x_i \leq 1 - \eta} |\tilde{\theta}(x_i) + \hat{b}_i - \theta(x_i)|$$

$$\leq \max_{\eta \leq x_i \leq 1 - \eta} |\tilde{\theta}(x_i) - E(\tilde{\theta}(x_i))| + \max_{\eta \leq x_i \leq 1 - \eta} |E(\tilde{\theta}(x_i)) + b(x_i) - \theta(x_i)| + \max_{\eta \leq x_i \leq 1 - \eta} |\hat{b}_i - b(x_i)|$$

$$= (i) + (ii) + (iii).$$

By Theorem 3.1 of Zhou, Shen and Wolfe (1998), p.1777, for any $x \in [0,1],$

$$\frac{\tilde{\theta}(x) - E(\tilde{\theta}(x))}{\sqrt{\text{Var}(\tilde{\theta}(x))}} \overset{d}{\to} N(0, 1).$$

By lemma 6.6 of Zhou, Shen and Wolfe (1998), $\text{Var}(\tilde{\theta}(x)) = o(1)$. Thus $\tilde{\theta}(x) - E(\tilde{\theta}(x))$ converges to zero in distribution. This implies the convergence in probability. Thus
we have that (i) converges to 0 in probability. By Theorem 5.1.1, (ii) = o(u^m).
Finally, by Theorem 5.2.1, (iii) = o_p(n^{-\kappa}). Thus
\[ I \overset{p}{\to} 0. \]  
(A.20)

By (A.18), (A.19) and (A.20), we have
\[ \sup_{\eta < x < 1 - \eta} |\tilde{\theta}(x) - \theta(x)| \overset{p}{\to} 0, \text{ as } n \to \infty. \]

□

Proof of Theorem 5.2.4. The condition \( m/(2m + 1) \leq \kappa = \min\{\alpha - \beta, \nu \beta\} \) is equiva-
lent to the following:
\[ \frac{1}{2} - (\alpha - \beta) - \frac{1}{2(2m + 1)} \leq 0, \quad \text{and} \quad \frac{1}{2} - \beta \nu - \frac{1}{2(2m + 1)} \leq 0. \]  
(A.21)

It is easy to see that the set of \((\alpha, \beta)\)s which satisfy (A.21) and \(0 < \beta \leq \alpha \leq 1, 0 < \nu < 1/2\) is not empty. Hence
\[ \sqrt{n} \frac{\hat{\theta}(x_i) + \hat{b}_i - \theta(x_i)}{\hat{\sigma}_i} = \sqrt{n \text{Var}(\hat{\theta}(x_i))} \left( \frac{\hat{\theta}(x_i) + b(x_i) - \theta(x_i)}{\sqrt{\text{Var}(\hat{\theta}(x_i))}} + \frac{\hat{b}_i - b(x_i)}{\sqrt{\text{Var}(\hat{\theta}(x_i))}} \right). \]  
(A.22)

By Theorem 5.1.2,
\[ \frac{\hat{\theta}(x_i) + b(x_i) - \theta(x_i)}{\sqrt{\text{Var}(\hat{\theta}(x_i))}} \overset{d}{\to} N(0, 1). \]  
(A.23)

By Theorems 5.1.1 and 5.2.2,
\[ \frac{\sqrt{n \text{Var}(\hat{\theta}(x_i))}}{\hat{\sigma}_i} \overset{p}{\to} (B(x)G^{-1}(q)B(x))^{1/2}. \]  
(A.24)

The condition \( k \leq Cn^{1/(2m+1)} \) and the fact \( M^{-1} < ku < M \) implies \( u = O(n^{-1/(2m+1)}) \).

By lemma 6.6 of Zhou, Shen and Wolfe (1998), \( \text{Var}(\hat{\theta}(x_i)) = O((nu)^{-1}) \). Combining
these facts, (A.21) and Theorem 5.2.1, we have

\[
\frac{\hat{\theta}_i - b(x_i)}{\sqrt{\text{Var}(\hat{\theta}(x_i))}} = O(n^{1/2-1/(2(2m+1))})(o(n^{-(\alpha - \beta)}) + o_p(n^{-\beta \nu}) + o(u^m))
\]

\[
= o(n^{1/2-1/(2(2m+1))}-(\alpha - \beta)) + o_p(n^{1/2-1/(2(2m+1))} - \beta \nu)
\]

\[
+ o(n^{1/2-1/(2(2m+1))} - d/(2m+1))
\]

\[= o_p(1). \quad (A.26)
\]

The desired result follows from (A.22), (A.23), (A.24), (A.26). \(\square\)

Let \(F\) and \(F_n\) be the distribution function of \(\epsilon\) and the empirical distribution function of \(\{\epsilon_i\}_{i=1}^n\), respectively. Let \(\tilde{F}_n\) and \(\tilde{F}_n\) be the empirical distribution function of the bias-corrected residuals \(\{\epsilon_i\}_{i=1}^n\) and the empirical distribution of the bias corrected and centred residuals. As defined before, the Mallows metric \(d_p(F, G)\) between two distribution functions \(F\) and \(G\) is \(\inf E(\|X - Y\|^p)^{1/p}\), where the infimum is taken over pairs of random variables \(X\) and \(Y\), and the marginal distribution of \(X\) is \(F\) and the one of \(Y\) is \(G\).

Before proving Theorem 5.3.1, we state two lemmas corresponding to Lemmas 2.1 and 2.2 of Freedman (1981).

**Lemma 4.**

\[E(d_2(\tilde{F}_n, F_n)^2) \leq o(1) + o^2 \left( \frac{1}{m_n} + \frac{k + m}{n} \right).\]

**Proof.** By the definition of the Mallows metric,

\[d_2(\tilde{F}_n, F_n)^2 \leq \frac{1}{n} \sum_{i=1}^n (e_i - \epsilon_i)^2 = \frac{1}{n} ||e - \epsilon||^2.\]
Letting $J_n(x_i) = \sum_{k \in M_n(x_i)} J_{k1}$ as before, the bias-corrected residual $e$ can be expressed as $e = (I - m^{-1}_n L_n)(I - M)Y$ where $L_n = (J_n(x_1)', \ldots, J_n(x_n)')'$. Then

$$E\|e - e\|^2 = E\|(I - m^{-1}_n L_n)(I - M)Y - e\|^2$$

$$\leq \|(I - m^{-1}_n L_n)(I - M)\theta(X)\|^2 + E\|(I - m^{-1}_n L_n)(I - M) - I\|e\|^2.$$ 

The $i$-th element of $(I - m^{-1}_n L_n)(I - M)\theta(X)$ is represented as follows:

$$J_{i1}'(I - M)\theta(X) - \frac{1}{m_n} J_n(x_i)(I - M)\theta(X)$$

$$= b(x_i) + o(u^m) - \frac{1}{m_n} \sum_{k \in M_n(x_i)} J_{k1}'(I - M)\theta(X) \quad \text{by (A.3)}$$

$$= b(x_i) - \frac{1}{m_n} \sum_{k \in M_n(x_i)} b(x_k) + o(u^m)$$

$$= o(n^{-\alpha+\beta}) + o(u^m) \quad \text{by (A.6)}$$

Thus $J_{i1}'(I - M)\theta(X) - \frac{1}{m_n} J_n(x_i)(I - M)\theta(X)$ converges to zero and

$$\|(I - m^{-1}_n L_n)(I - M)\theta(X)\|^2 = o(n).$$

On the other hand,

$$E\|(I - m^{-1}_n L_n)(I - M) - I\|e\|^2$$

$$= \text{tr}(((I - m^{-1}_n L_n)(I - M) - I)E(e'e')((I - m^{-1}_n L_n)(I - M) - I)')$$

$$= \sigma^2 \text{tr}((m^{-1}_n L_n + P - m^{-1}_n L_n P)(m^{-1}_n L_n + P - m^{-1}_n L_n P)')$$

$$= \sigma^2 \{ m^{-2}_n \text{tr}[L_n L_n'] + m^{-1}_n \text{tr}[L_n P] + m^{-2}_n \text{tr}[L_n P L_n']$$

$$+ m^{-1}_n \text{tr}[PL_n'] + \text{tr}[P] - m^{-1}_n \text{tr}[PL_n']$$

$$- m^{-2}_n \text{tr}[L_n P L_n'] - m^{-1}_n \text{tr}[L_n P] + m^{-2}_n \text{tr}[L_n P L_n'] \}$$

$$= \sigma^2 \left( m^{-2}_n \text{tr}[L_n L_n'] + \text{tr}[P] - m^{-2}_n \text{tr}[L_n P L_n'] \right) \leq \sigma^2 \left( \frac{n(m_n)}{m^2_n} + (k + m) \right).$$
Therefore
\[ E[|e - \bar{e}|^2] \leq o(n) + \sigma^2 \left( \frac{n(m_n)}{m_n^2} + (k + m) \right) \]
and
\[ E(d_2(\tilde{F}_n, F_n)^2) \leq o(1) + \sigma^2 \left( \frac{1}{m_n} + \frac{1 + k + m}{n} \right). \]

\begin{proof}
By Lemma 8.8 of Bickel and Freedman (1981),
\[
d_2(\hat{F}_n, F_n)^2 = d_2(\hat{F}_n - E(\hat{F}_n), F_n - E(F_n))^2 + |E(\hat{F}_n) - E(F_n)|^2
\]
\[ = d_2(\hat{F}_n, F_n - n^{-1} \sum_{i=1}^{n} \epsilon_i)^2 + (n^{-1} \sum_{i=1}^{n} \epsilon_i)^2, \]
\[
d_2(\tilde{F}_n, F_n)^2 = d_2(\tilde{F}_n - E(\tilde{F}_n), F_n - E(F_n))^2 + |E(\tilde{F}_n) - E(F_n)|^2
\]
\[ = d_2 \left( \left( \hat{F}_n + n^{-1} \sum_{i=1}^{n} \hat{\epsilon}_i \right) - n^{-1} \sum_{i=1}^{n} \hat{\epsilon}_i, F_n - n^{-1} \sum_{i=1}^{n} \epsilon_i \right)^2 + \left( n^{-1} \sum_{i=1}^{n} (\hat{\epsilon}_i - \epsilon_i) \right)^2
\]
\[ = d_2(\hat{F}_n, F_n - n^{-1} \sum_{i=1}^{n} \epsilon_i)^2 + \left( n^{-1} \sum_{i=1}^{n} (\hat{\epsilon}_i - \epsilon_i) \right)^2. \]

Thus
\[
d_2(\hat{F}_n, F_n - n^{-1} \sum_{i=1}^{n} \epsilon_i)^2 = d_2(\hat{F}_n, F_n)^2 - \left( n^{-1} \sum_{i=1}^{n} (\hat{\epsilon}_i - \epsilon_i) \right)^2,
\]
and
\[
d_2(\tilde{F}_n, F_n)^2 = (n^{-1} \sum_{i=1}^{n} \epsilon_i)^2 - \left( n^{-1} \sum_{i=1}^{n} (\hat{\epsilon}_i - \epsilon_i) \right)^2 + d_2(\tilde{F}_n, F_n)^2
\]
\[ \leq (n^{-1} \sum_{i=1}^{n} \epsilon_i)^2 + d_2(\tilde{F}_n, F_n)^2. \]
\end{proof}
Therefore
\[
\mathbb{E}(d_2(\hat{F}_n, F_n)^2) \leq o(1) + \sigma^2 \left( \frac{1}{mn} + \frac{1 + k + m}{n} \right).
\]

\[\square\]

**Proof of Theorem 5.3.1.** The proof is parallel to the one of Freedman (1981); See also Härdle and Bowman (1988).

First, we have
\[
\hat{\theta}(x_i) + \hat{b}_i = J_{i1}^T P Y + mn^{-1} J_n'(x_i)(I - M) Y
\]
\[
= \left( \left( J_{i1}^T - mn^{-1} J_n'(x_i) \right) P + mn^{-1} J_n'(x_i) \right) Y
\]
\[
= \left( \left( J_{i1}^T - mn^{-1} J_n'(x_i) \right) P + mn^{-1} J_n'(x_i) \right) (\theta(X) + \epsilon)
\]
and
\[
(\hat{\theta}(x_i) + \hat{b}_i) - \mathbb{E}(\hat{\theta}(x_i) + \hat{b}_i) = \left( \left( J_{i1}^T - mn^{-1} J_n'(x_i) \right) P + mn^{-1} J_n'(x_i) \right) \epsilon.
\]

Similarly,
\[
(\hat{\theta}^*(x_i) + \hat{b}_i^*) - \mathbb{E}^*(\hat{\theta}^*(x_i) + \hat{b}_i^*) = \left( \left( J_{i1}^T - mn^{-1} J_n'(x_i) \right) P + mn^{-1} J_n'(x_i) \right) \epsilon^*
\]
where \(\mathbb{E}^*\) represents the expectation under a bootstrap distribution and \(\epsilon^*\) is a bootstrap sample from a bias corrected and centred residuals, \(e_1, \ldots, e_n\). By Lemma 8.8 of Bickel and Freedman (1981), the Mallows metric is decomposed, using (A.27) and
(A.22), as follows:

\[
d_2 \left( \left\{ (\hat{\theta}(x_i) + \hat{\theta}_i^*) - (\hat{\theta}(x_i) + \hat{\theta}_i) \right\}, \left\{ (\hat{\theta}(x_i) + \theta(x_i)) \right\} \right)^2 \\
= d_2 \left( \left\{ (\hat{\theta}(x_i) + \hat{\theta}_i^*) - (\hat{\theta}(x_i) + \hat{\theta}_i) \right\} - E^* \left\{ (\hat{\theta}(x_i) + \hat{\theta}_i^*) - (\hat{\theta}(x_i) + \hat{\theta}_i) \right\} \\
+ E^* \left\{ (\hat{\theta}(x_i) + \hat{\theta}_i^*) - (\hat{\theta}(x_i) + \hat{\theta}_i) \right\} \right)^2 \\
= d_2 (A\epsilon^*, A\epsilon)^2 + \| (A\theta(X) - (\hat{\theta}(x_i) + \hat{\theta}_i)) - (A\theta(X) - \theta(x_i)) \|^2 \\
= d_2 (A\epsilon^*, A\epsilon)^2 + \| (\hat{\theta}(x_i) + \hat{\theta}_i) - \theta(x_i) \|^2 \\
= I + II \tag{A.29}
\]

where

\[
A = \left( (J_{i1} - m_n^{-1} J_n'(x_i)) M + m_n^{-1} J_n'(x_i) \right).
\]

By Theorem 5.2.3,

\[
|\hat{\theta}(x_i) - \theta(x_i)| = |(\hat{\theta}(x_i) + \hat{\theta}_i) - \theta(x_i)| \xrightarrow{P} 0, \text{ as } n \to \infty.
\]

Thus

\[
II \xrightarrow{P} 0, \text{ as } n \to \infty. \tag{A.30}
\]

By Lemma 8.9 of Bickel and Freedman (1981),

\[
d_2 (A\epsilon^*, A\epsilon)^2 \leq \text{tr}[AA^T]d_2 (\bar{F}_n, F)^2,
\]

82
Now
\[ \text{tr}[AA'] = \text{tr}[(J'_{i1} - m_n^{-1} J'_n(x_i)) M + m_n^{-1} J'_n(x_i)) (J'_{i1} - m_n^{-1} J'_n(x_i)) M + m_n^{-1} J'_n(x_i)] \]
\[ = \text{tr}[(J'_{i1} - m_n^{-1} J'_n(x_i)) M J_{i1} - m_n^{-1} J'_n(x_i)] + 2 \text{tr}[m_n^{-1} J'_n(x_i) M J_{i1}] + \text{tr}[m_n^{-2} J'_n(x_i) J_n(x_i)] \]
\[ = \text{tr}[J'_{i1} M J_{i1}] - 2m_n^{-1} \text{tr}[J'_n(x_i) M J_{i1}] + m_n^{-2} \text{tr}[J'_n(x_i) M J_n(x_i)] \]
\[ + 2m_n^{-1} \text{tr}[J'_n(x_i) M J_{i1}] - 2m_n^{-2} \text{tr}[J'_n(x_i) M J_n(x_i)] + m_n^{-2} \text{tr}[J'_n(x_i) J_n(x_i)] \]
\[ = J'_{i1} n^{-1} B(x) G_n^{-1} B(x) J_{i1} - m_n^{-2} J'_n(x_i) M J_n(x_i) + m_n^{-1} \]
\[ \leq \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} + m_n^{-1}. \]

Thus we have
\[ d_2(Ae^*, Ae)^2 \leq (\lambda_{\text{max}}/\lambda_{\text{min}} + m_n^{-1}) d_2(\hat{F}_n, F)^2 \]
\[ \leq (\lambda_{\text{max}}/\lambda_{\text{min}} + m_n^{-1}) 2(d_2(\hat{F}_n, F_n)^2 + d_2(F_n, F)^2). \]

By Lemma 8.4 of Bickel and Freedman (1981), \( d_2(F_n, F)^2 \) converges to zero in probability. By Lemma 5, \( E(d_2(\hat{F}_n, F_n)^2) \) converges to zero. Using the Markov’s inequality, this implies \( d_2(\hat{F}_n, F)^2 \) converges to zero in probability. Thus
\[ I \xrightarrow{P} 0, \text{ as } n \to \infty. \quad (A.31) \]

Combining (A.29), (A.30) and (A.31), we have
\[ d_2 \left( \left\{ (\theta^*(x_i) + \delta_i^*) - (\hat{\theta}(x_i) + \delta_i) \right\}, \left\{ (\hat{\theta}(x_i) + \delta_i) - \theta(x_i) \right\} \right)^2 \xrightarrow{P} 0, \text{ as } n \to \infty. \]
BIBLIOGRAPHY


84


