A Comparison of Two Linear Nonparametric Regression Techniques

Sylvain Sardy

Utah State University
A COMPARISON OF TWO LINEAR NONPARAMETRIC
REGRESSION TECHNIQUES

by

Sylvain Sardy

A thesis submitted in partial fulfillment
of the requirements for the degree
of
MASTER OF SCIENCE
in
Mathematics

UTAH STATE UNIVERSITY
Logan, Utah
1992
ACKNOWLEDGEMENTS

I would like to express my appreciation to my major professor, Dr. Michael Windham, for his rigorous but friendly guidance throughout the completion of my thesis.

Thank you Dr. Adele Cutler for your enthusiastic support! Exchanging ideas in the development of Chapter 5's iterative smoother has been an enjoyable and valuable experience.

Dr. Richard Cutler has been a great teacher and taught me helpful materials in statistics for this thesis.

I am thankful to Drs. M. Windham, A. Cutler and R. Cutler for nominating me for Graduate Student Researcher of the Year. Their guidance and their support have been important factors besides my work on this thesis and on a report.

I am grateful to Dr. J. Koebbe for being a member of my committee.

I also want to thank Dr. Kenneth Bosworth, my previous advisor and major professor, for introducing me into the topic of nonparametric regression.

Sylvain Sardy
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS .......................................................... ii

LIST OF FIGURES .............................................................. iv

ABSTRACT .............................................................................. v

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

2 THE STATISTICAL MODEL .................................................... 3

3 THE CLASS OF LINEAR SMOOTHERS ..................................... 6
   3.1 Definition ................................................................. 6
   3.2 How to choose the smoothing parameter ......................... 8
      3.2.1 Types of parameters ............................................ 8
      3.2.2 Global criterion .................................................. 10
      3.2.3 Local criterion ................................................... 16
      3.2.4 Robust criterion .................................................. 17
   3.3 How to obtain a confidence interval ............................... 19

4 KERNEL SMOOTHER .......................................................... 21
   4.1 Definition ................................................................. 21
      4.1.1 The kernel .......................................................... 21
      4.1.2 The bandwidth .................................................... 25
   4.2 Properties ................................................................. 27
   4.3 Confidence interval .................................................... 29

5 ITERATIVE SMOOTHER ...................................................... 31
   5.1 Definition ................................................................. 31
   5.2 Properties ................................................................. 33
   5.3 Confidence Interval .................................................... 42
   5.4 Example .................................................................. 42

REFERENCES ....................................................................... 42

APPENDIX .......................................................................... 46
LIST OF FIGURES

5.1 Example of a local smoothing iterative procedure .................. 43
ABSTRACT

A Comparison of Two Linear Nonparametric Regression Techniques

by

Sylvain Sardy

Utah State University, 1992

Major Professor: Dr. Michael Windham
Department: Mathematics

This thesis presented a useful tool in regression. Nonparametric linear regression techniques were described in the general context of regression. A comparison of two of these techniques, kernel regression and iterative regression, showed various aspects of nonparametric linear regressors. (54 pages)
Chapter 1

INTRODUCTION

The objective of this study is to present the class of nonparametric linear regression (or smoothing) techniques, to introduce the related concepts, and to illustrate the general idea by the study of two elements of this nonparametric linear class.

Applied sciences deal with data. These data have often been recorded with some error due to the measurement device. Therefore they need to be treated in order to extract information. Statisticians and mathematicians have developed appropriate techniques according to the type of data encountered. For example measurements, $Y$, have been taken on different locations, $x$, so that a set of data of the type $(x,Y)$ is available to the applied scientist. Unfortunately the data are noisy so that plotting $Y$ (the response variable) versus $x$ (the predictor variable) does not give a clear picture of the underlying phenomenon that links the response variable to the predictor variable. According to a model, a specific smoother operates on the set of noisy data to give a decomposition

$$Y = s(x) + r$$

where $r$ is the residual, so that $s$ is close (in some sense) to the true underlying function $f$ that describes the phenomenon. If the scientist, based on the plotting of $Y$ versus $x$ and on prior information, believes that $f$ belongs to some class of parametric functions, then he will use parametric regression techniques. However, in some cases, the scientist has no clue to the phenomenon. This is the case where the nonparametric regression techniques will provide him with a powerful tool.
Recently multivariate \((\dim(x, Y) > 2)\) smoothing techniques handle the high dimensionality by decomposing the problem into bivariate \((\dim(x, Y) = 2)\) smooths so that an efficient and fast automatic bivariate smoother is needed. Examples of such techniques are projection pursuit \([13]\), nonlinear additive regression \([14]\), and generalized additive models \([15]\). The automatic bivariate smoother is iteratively applied until criterion of convergence is reached. Therefore bivariate smoothing techniques constitute an important subclass.
Chapter 2

THE STATISTICAL MODEL

1. In order to develop a regression technique, one needs assumptions about:

   (a) The underlying function: for instance, believing the underlying function to be
discontinuous or continuous will lead to two different techniques.

   (b) The noise: the noise can be correlated, biased, heteroscedastic or have a non
symmetric distribution. In such a case, developing an efficient regressor will be
difficult. Usually, nicer properties of the noise will be assumed.

   (c) The data: because of the measurement technique, the data can have outliers, so
that a robust regressor will be helpful to decrease the influence of these points
on the rest of the data.

2. One also needs to state what kind of information to get from the data:

   (a) Estimation: estimation of the underlying function at the $x_i$'s.

   (b) Interpolation: estimation between the $x_i$'s.

   (c) Extrapolation: estimation out of the range of the $x_i$'s.

   (d) Confidence Interval: interval in which the underlying function is believed to lie
with a certain probability.

In the following section, we address the problem of estimating the function at the $x_i$'s
(2.a), and developing techniques to get a confidence interval (2.d). We will also look for a
robust regressor, in the case where the data contain outliers (1.c). Finally, for (1.a) and (1.b), we use the usual regression model

\[ Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n \]

where

- \( n \) is the number of data available.
- \( Y \) is the response random vector: \( \dim(Y) = n \times 1 \).
- \( X \) is the predictor random or fixed vector: \( \dim(X) = n \times d \).
- \( f \) is the underlying function, and is assumed to be in \( C^m \).
- \( \varepsilon \) is the noise random vector: \( \dim(\varepsilon) = n \times 1 \), and
  \[
  \begin{align*}
  \mathbb{E}(\varepsilon) &= 0 \\
  \text{Var}(\varepsilon) &= \sigma^2 I_{n \times n}.
  \end{align*}
  \]

We observe that the function belongs to \( C^m \). In the next section, the degree of smoothness, \( m \), will be assumed to be 2, in seeking an asymptotic result.

The noise has mean 0, which means that the data is a collection of unbiased estimates of the underlying function at the \( x_i \)'s. The noises are also assumed to be homoscedastic and, important, independent (not only uncorrelated). If this last assumption is not satisfied, then an ARIMA time series model is more appropriate, and the regressors described in the next sections will no longer be consistent.

We will also assume that the dimension of the predictor vector is \( n \times 1 \). In this case, \( d = 1 \), the smoother is the so-called bivariate smoother. For the kernel method in higher dimensions, we refer to [1] and [2]. However, this method suffers from the so-called “curse of dimensionality,” which means that, to go up to dimension \( d > 1 \) and keep the same efficiency as with \( n \) data in dimension 1, one will need \( n^d \) data. Methods such as the II method [3] or the ones using a decomposition into bivariate smooths [13] [14] [15] are preferable when the dimension \( d \) is high (i.e., 3, 4, ...).

Often, for convenience, we will consider the case where the predictor vector is made of equally spaced values. However, the techniques presented here can handle unequally spaced data (we will show how). The unequally spaced design is often referred to as the
random predictor design, but we will always assume the vector $x$ is a deterministic vector.
Chapter 3

THE CLASS OF LINEAR SMOOTHERS

3.1 Definition

A wide class of estimators, the class of linear estimators, is commonly used because properties of linear estimators are easy to derive mathematically. Any estimator of this class has a corresponding real matrix $H$ called the hat matrix because of the notation: $\hat{Y} = HY$.

**Parametric regression** assumes a parametric shape for the underlying function. Therefore, the hat matrix is fixed. For instance, if the underlying function is believed to be quadratic, i.e., $Y_i = a + bx_i + cx_i^2 + \varepsilon_i$, and if the measure of closeness to the underlying function is the regular least squares, the estimate of $\theta = (a, b, c)'$ is found by minimizing $\varepsilon'\varepsilon$. The parameter estimate is $\hat{\theta} = (X'X)^{-1}X'Y$, which makes the hat matrix be $H = X(X'X)^{-1}X'$, where

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}$$

So parametric estimators which are defined given $p = \dim(\Theta) \geq 1$ and $X \in \mathbb{R}^{n\times p}$ as

$$\hat{f}_p : \mathbb{R}^n \mapsto \mathbb{R}^n$$

$$(Y; X, p) \mapsto \hat{f}_p(Y; X, p) = H(X, p)Y$$

are linear functions of one variable, $Y$. If the least squares criterion is used, it is interesting to note that such an estimator gives the best linear unbiased estimates (BLUE). However, due to linearity among the columns of $X$, the variance of the estimates might be huge, since $\text{Var}(\hat{f}) = \sigma^2 H$. 
An attempt to reduce the instability when solving $X'X\beta = X'Y$ is to slightly perturb the matrix $X'X$ with the matrix $\lambda I_{p\times p}$ that is full rank. The estimator then becomes $\hat{f} = X(X'X + \lambda I)^{-1}X'Y$, and is now biased but has a smaller variance. Such an estimator is called a *ridge regression estimator* and falls into the class of nonparametric estimators.

**Non-parametric regression** does not assume any parametric form for the underlying function $f$, but only assumes that $f$ belongs to some set of functions such as $C^2$. It uses the fact that the data $Y_i$ at $x_i$ as well as the data $Y_{i\pm k(\lambda)}$ in the neighborhood $N_i(\lambda)$ of $x_i$ contains some information about $f(x_i)$. By taking a weighted average of these data in some fashion, one can achieve a good estimation of $f$ at $x_i$ — or a bad one if the weighted average is not appropriate.

$$ f(x_i) = \text{wavg}(Y_{j} : j \in N_i(\lambda)) $$

$$ = \langle W_{\lambda,i}, Y \rangle_{\mathbb{R}^n} $$

where the selection of the weights sequence, $W_{\lambda,i} = (W_{\lambda,ij})_{j=1,n}$, with $\sum_{j=1}^{n} W_{\lambda,ij} = 1$, is driven by the selection of $\lambda$ that is chosen to optimize a selected goodness of fit criterion. Data to the right or to the left of $x_i$ have equal importance in predicting $f(x_i)$; therefore the weight sequence will be symmetric around $W_{\lambda,ii}$.

For any $\lambda$, called the smoothing parameter, corresponds a hat matrix $H(\lambda)$, whose $i^{th}$ row contains the weight sequence to estimate $f$ at $x_i$, $W_{\lambda,i}$. So the hat matrix is a symmetric matrix. The hat matrix is free to vary until the criterion of goodness of fit is reached. Since the hat matrix is the matrix of weight sequences, we will use the notation $W(\lambda)$ when referring to the hat matrix in nonparametric regression.
It is interesting to note that

$$\hat{\theta}_i = \arg \min_{\theta_i} \frac{1}{n} \sum_{j=1}^{n} W_{\lambda,ij}(Y_j - \theta_i)^2$$

(3.1)

$$\iff \hat{\theta}_i = \sum_{j=1}^{n} W_{\lambda,ij} Y_j,$$

so that choosing a particular weight sequence is equivalent to minimizing a particular local weighted least squares.

In a general framework, a nonparametric estimator is a function of two variables, $\lambda$ and $Y$, but are only linear in $Y$ since, given $x \in \mathbb{R}^{n \times d}$, they are defined as

$$w : \Theta \rightarrow \mathbb{R}^{n \times n}$$

$$\lambda; x \rightarrow w(\lambda; x) = W(\lambda; x) = W_{\lambda}$$

and

$$\hat{f}_{NP} : \mathbb{R}^{n} \times \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n}$$

$$Y, w(\lambda; x) \rightarrow \hat{f}_{NP}(Y, w(\lambda; x)) = W_{\lambda}Y$$

where $\dim(\Theta) = 1$ if the criterion of goodness of fit is global.

The choice of $\lambda$ is crucial since it corresponds to the choice of a model in parametric regression.

### 3.2 How to choose the smoothing parameter

#### 3.2.1 Types of parameters

The smoothing parameter $\lambda$ governs the smoothness of the estimated curve. Choosing it in a wrong way will result either in undersmoothing the underlying function (the extreme case is to pass through the noisy data), or in oversmoothing it (the extreme case is to get the plane of $\mathbb{R}^d$ that is the "closest" to the data). So, the selection of the smoothing parameter is a big issue.
The form of the smoothing parameter can be anything. Depending on the type of criterion of goodness of fit, the smoothing parameter takes the form of a single variable or a vector.

A simple and fast way to smooth is to optimize a *global* criterion of goodness of fit. This means that a single variable (dim(\(\Theta\)) = 1) will drive the goodness of fit criterion. However, the flexibility of such an estimator is poor. Indeed, a curve can be first linear and then wiggly, and the single variable estimator will not fit either one of the two parts well.

On the latter example, one would like to have two criteria (a criterion for the linear part and a second one for the wiggly part) so that the fit on both parts will be more appropriate. This amounts to selecting a vector of two smoothing parameters (dim(\(\Theta\)) = 2). In a more general way, one will look for a smoothing parameter for each estimate, and its selection will come from optimizing a *local* criterion of goodness of fit. Hence, a local linear operator will be driven by the selection of a vector of \(n\) smoothing parameters, where \(n\) is the number of data.

Different types of linear estimators with a specific smoothing parameter have been developed. When smoothing with a kernel technique, the smoothing parameter (called \(h\) then) is in \(\mathbb{R}^+\) if the criterion is global ((\(\mathbb{R}^+\))^n if local). When smoothing with an iterative scheme, the smoothing parameter (called \(j\) then) is in \(\mathbb{N}\) (the set of positive integer numbers) if the criterion is global (\(\mathbb{N}^n\) if local).

The role of the smoothing parameter is particularly obvious in the context of smoothing using splines. Indeed, this technique looks for a smooth function (in a Sobolev space \(W_2^n[a, b]\)) made of piecewise polynomials that minimizes, over all the functions in the
Sobolev space, the criterion

\[ \sum_{i=1}^{n} (Y_i - \tilde{f}_\lambda(x_i))^2 / n - \lambda \int_a^b \tilde{f}_\lambda^{(m)}(x)dx, \lambda \in \mathbb{R}^+ . \]

This notation shows clearly that \( \lambda \) governs the tradeoff between the goodness-of-fit (first term) and the smoothness (second term). For instance, \( \lambda \) equal to zero gives the worst case of undersmoothing, whereas \( \lambda \) equal to infinity gives the worst case of oversmoothing.

3.2.2 Global criterion

In this section, we will see how one can define and then estimate a global criterion of goodness-of-fit.

For two functions \( f_1 \) and \( f_2 \), the distance between \( f_1 \) and \( f_2 \) can be defined in many ways. For positive functions (e.g., density functions), the distance defined as

\[ d_H^2(f_1, f_2) = \| f_1^{1/2} - f_2^{1/2} \|^2_2 = \int (f_1^{1/2} - f_2^{1/2})^2 \]

is called the Hellinger distance and has nice robust properties.

Another possible measure of "closeness" is the one related to the \( L_1 \) norm

\[ d_1^2(f_1, f_2) = \| f_1 - f_2 \|^2_1 = (\int |f_1 - f_2|)^2. \]

But the most common distance in regression is the distance related to the \( L_2 \) norm

\[ d_2^2(f_1, f_2) = \| f_1 - f_2 \|^2_2 = \int (f_1 - f_2)^2. \]

This distance between two functions is the one we are going to use, because the square loss function is differentiable at zero, which is not the case for the absolute value loss function. Moreover the expected value of the square of a random variable is related to its variance, whereas the expected value of the absolute value of a random variable is not a statistical entity.
In the context of regression, $f_1 = f$ is the unknown underlying function, and $f_2 = \hat{f}(Y)$ is the fitted linear function that is the closest to $f$ for our selected definition of closeness

$$ds(\hat{f}) = d^2(f, \hat{f}(Y)) = |f - \hat{f}(Y)|^2.$$  

Because of the $Y$ vector, this distance is a stochastic quantity. For theoretical results, deterministic quantities are more tractable, and

$$d_D(\hat{f}) = E_Y(d^2(f, \hat{f}(Y))) = E_Y(|f - \hat{f}(Y)|^2)$$

will be the deterministic measure of closeness since it does not rely on a particular set of data.

The integral is not computable since the only information about $f$ is a finite set of noisy data $(x_i, Y_i)_{i=1,n}$. But $d_D$ can be estimated by

$$\hat{d}_D(\hat{f}) = \sum_{i=1}^n E_Y((f(x_i) - \hat{f}(x_i, Y))^2)s(x_i)$$

where $s$ is some weight function that can be used to reflect the spacing of the data. The distance function, $\hat{d}_D(\cdot)$, is often called the Risk function, $R_n(\cdot)$. For simplicity, we will assume from now on that the $x$'s are equally spaced (i.e., $s(\cdot) = 1/n$); the Risk then becomes

$$R_n(\hat{f}) = \frac{1}{n} \sum_{i=1}^n E_Y((f(x_i) - \hat{f}(x_i, Y))^2)$$  

(3.2)

It is important to note that the Risk function can be split into two parts: the square bias function $b_n^2(\cdot)$ and the variance function $v_n(\cdot)$ since

$$R_n(\hat{f}) = \frac{1}{n} \sum_{i=1}^n (f(x_i) - E_Y(\hat{f}(x_i, Y))^2 + \frac{1}{n} \sum_{i=1}^n \text{Var}(\hat{f}(x_i, Y)) = b_n^2(\hat{f}) + v_n(\hat{f}).$$

(3.3)

Based on this measure, an estimator $\hat{f}$ is said to be consistent if $R_n(\hat{f}) \xrightarrow{n \to \infty} 0$. The rate of decay to zero provides information about how effective $\hat{f}$ is an estimator of the
underlying function $f$. It also is a criterion of comparison between two estimators: an estimator $\hat{f}$ is said to be asymptotically optimal if, uniformly in $f$ that belongs to some nice set of functions, $R_n(\hat{f})$ tends to zero as $n$ tends to infinity at the rate $\frac{1}{n^\alpha}$ for some positive $\alpha$, and no other estimator achieves a better rate.

Parametric estimators are asymptotically optimal over the nonparametric ones, in general. Indeed, the rate of convergence of $R_n(\cdot)$ to zero of a parametric estimator is like $\frac{1}{n}$ or $\alpha = 1$; for nonparametric estimators, however, $\alpha$ is typically less than 1. This is the price to pay to make fewer assumptions about the shape of the underlying function. But, if the parametric model is not correct, the Risk does not tend to zero anymore and the estimator is no longer consistent. This shows that nonparametric estimators are superior to the parametric ones, in the only case where the parametric model might not be the correct one.

A nonparametric linear estimator is entirely determined by the selection of $\lambda$. So the Risk function can be written equivalently as

$$R_n(\hat{f}_\lambda) = R_n(W_\lambda) = R_n(\lambda)$$

The Risk is our criterion of goodness of fit; so the smoothing parameter $\lambda$ will be selected by minimizing the Risk function, or by minimizing any function biased to the Risk by a constant.

However, in practice, $R_n(\lambda)$ is unknown since (of course) $f$ is unknown. The only information available about $f$ is the set of unbiased noisy data $(Y_i)_{i=1,n}$ and the assumption that $f$ is in some set of functions.

If one had a second set of data $(Y_i^*)_{i=1,n}$ at the same given locations $(x_i)_{i=1,n}$, one could think of estimating the Risk with the information brought by the second set. This
leads to the Prediction Risk function,

\[ P_n(\lambda) = \frac{1}{n} \sum_{i=1}^{n} E_{Y, Y^*} (Y_i^* - \hat{f}_\lambda(Y, x_i))^2 \]

\[ = \sigma^2 + R_n(\lambda) \]

that is biased to the Risk by the constant \( \sigma^2 \). So minimizing \( R_n \) over \( \lambda \) is equivalent to minimizing \( P_n \). However, in practice, a second set of data is rarely available.

A first idea to estimate \( R_n(\hat{f}) \) would be to use the Mean Squared Error

\[ MSE_n(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{f}_\lambda(x_i))^2. \]

However, by noting that

\[ MSE_n(\lambda) = \frac{1}{n} Y'(I - W_\lambda)^2 Y' \]

where \( W_\lambda \) is a symmetric matrix, we can show that MSE is a biased estimate of the Risk since

\[ E(MSE_n(\lambda)) = R_n(\lambda) + \sigma^2 - \frac{2\sigma^2}{n} tr(W_\lambda). \]

The problem with the bias is that it is, in general, not constant and hard to estimate, because \( \sigma^2 \) can not be estimated accurately (see appendix). Moreover the trace of the matrix \( W_\lambda \) might be expensive to compute or estimate. There are many ways to go around the problem.

1. If the bias were constant, then minimizing \( MSE_n \) would amount to minimizing \( R_n \).
   - One possibility to achieve this property is simply to use a nil-trace estimator, i.e., a linear estimator which corresponding matrix’s trace is zero.
   - The other possibility is to build a hat matrix \( W_\lambda \) such that its trace is constant for any selection of \( \lambda \).
A particular case of nil-trace estimator is an estimator for which the corresponding hat matrix $W_\lambda$ has a zero diagonal. This corresponds to estimate the function at a given location using all data in a neighborhood of the location except the data at the location itself since such an estimator can be written as

\[
\begin{pmatrix}
\hat{Y}_1 \\
\hat{Y}_2 \\
\vdots \\
\hat{Y}_n
\end{pmatrix} = \begin{pmatrix}
0 & W_{\lambda,ij} \\
0 & W_{\lambda,ij} \\
\vdots & 0 \\
W_{\lambda,ij} & 0 \\
\end{pmatrix} \begin{pmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_n
\end{pmatrix}
\]

(3.4)

Obviously, $tr(W_\lambda) = 0$ for every selection of $\lambda$.

This leads to the so called Cross Validation estimator, $CV_n$. Originally, such an estimator was based on the idea of the Prediction Risk. Indeed, the $CV_n$ estimator simulates artificially a second set of data $(Y^*_i)_{i=1,n}$, and one can show that

\[
CV_n(\lambda) = P_n(\lambda) = \frac{1}{n} \sum_{i=1}^n E_Y (Y_i - \hat{f}_{\lambda,-i}(Y_i, x_i))^2
\]

where $\hat{f}_{\lambda,-i}$ is the notation for the estimator in (3.4). The possible drawback of this method is that the estimation of $f$ at $x_i$ is done with a weight sequence that puts no weight on the most trusted data $Y_i$. Another way to look at it is to see that, by the equivalence (3.1), using such an estimator is equivalent to minimizing the sum of local weighted least squares with no penalizing weight on the squared residuals $(Y_i - \hat{f}(x_i))^2$, i.e., on the residual of where we want to estimate at $x_i$.

2. By modifying $W_\lambda$ a little, one can get a nil-trace estimator $\tilde{W}_\lambda$, so that $MSE_n$ becomes an unbiased estimate of $\tilde{P}_n$ that is hopefully close enough to $P_n$ to give a good estimate, $\tilde{\lambda}$, of $\lambda$. This kind of estimator is referred to as Generalized Cross Validation estimator, $GCV_n$, and a theorem gives an upper bound on the relative
error made by approximating $P_n$ by $GCV_n$. But first, let's look at the transformation that makes the hat matrix $W_\lambda$ a nil-trace matrix.

If $W_\lambda = (1 + \alpha_\lambda)W_\lambda - \alpha_\lambda I$, where $\alpha_\lambda = \frac{tr(W_\lambda)}{tr(I - W_\lambda)}$, then $tr(W_\lambda) = 0$ for every selection of $\lambda$.

It follows that

\[
MSE_n(\lambda) = MSE_n(W_\lambda)
\]

\[= \frac{1}{n}Y'(I - W_\lambda)^2Y'
\]

\[= \frac{1}{n}Y'(I - W_\lambda)^2Y/\left(\frac{1}{n}tr(I - W_\lambda)\right)^2
\]

\[= \frac{1}{n} \frac{MSE_n(\lambda)}{(\frac{1}{n}tr(I - W_\lambda))^2}
\]

\[= GCV_n(\lambda),
\]

and $MSE_n$ is an unbiased estimate of $P_n$.

Now, the following GCV theorem gives an upper bound to the relative error between $\hat{P}$ and our criterion of goodness of fit $P$.

GCV Theorem (see [1], p. 31): Let $\tau^{(j)}_\lambda = \frac{1}{n}tr(W_\lambda^j)$, $j = 1$ and $j = 2$, and assume that $\tau^{(1)}_\lambda \leq 1$. Then

\[
\frac{|E(GCV_n(\lambda)) - P_n(\lambda)|}{R_n(\lambda)} \leq g_n(\lambda)
\]

where

\[g_n(\lambda) = \frac{2\tau^{(1)}_\lambda + (\tau^{(1)}_\lambda)^2/\tau^{(2)}_\lambda}{(1 - \tau^{(1)}_\lambda)^2}
\]

A consequence of this theorem is that the $GCV$ is an asymptotically unbiased estimator of the Prediction Risk function.
3.2.3 Local criterion

The drawback of using a single variable that drives the global aspect of the estimator is a lack of flexibility. In this section, we will see how one can select and then estimate local selection criteria for the smoothing parameter vector.

Based on the global measure of closeness, we can derive the local deterministic measure of closeness at \( x_i \)

\[
d_d(\hat{f}(x_i)) = d_d(\lambda_i) = E_Y((f - \hat{f}_{\lambda_i})^2 s_i)
\]

where \( s_i \) is some symmetric density function centered around \( x_i \), and \( \lambda_i \) is the local smoothing parameter at \( x_i \) (see [4]).

With a finite set of (equally spaced) data, the local deterministic measure of closeness at \( x_i \) is the local Risk function

\[
R_n(\lambda_i) = \frac{1}{n} \sum_{j=1}^{n} E_Y((f(x_j) - \hat{f}_{\lambda_i}(x_j))^2 s_i(x_j))
\]

where now \( s_i = (s_i(x_1), \ldots, s_i(x_i), \ldots, s_i(x_n)) \) is a normalized weight sequence.

The natural but biased estimator of the local Risk is the weighted MSE

\[
LMSE_n(\lambda_i) = \frac{1}{n} \sum_{j=1}^{n} (Y_j - \hat{f}_{\lambda_i}(x_j))^2 s_i(x_j).
\]

And by noting that

\[
LMSE_n(\lambda_i) = \frac{1}{n} Y'(I - W_{\lambda_i}) S_i^2(I - W_{\lambda_i}) Y
\]

where \( S_i^2 = \text{diag}(s_i(x_1), \ldots, s_i(x_i), \ldots, s_i(x_n)) \), one can show that

\[
E(LMSE_n(\lambda_i)) = R_n(\lambda) + \sigma^2 - \frac{2\sigma^2}{n} \text{tr}(W_{\lambda_i} S_i^2).
\]

The idea of a matrix with a constant trace property and the GCV idea will not work anymore because of the matrix \( S_i^2 \). The CV is the only alternative to achieve a constant
biased estimate of $R$, since $tr(W, S_\lambda^2) = 0$ for any weight sequence $s_i$. Philippe Vieu [4], for a particular nonparametric linear regressor (kernel), proves that this local Cross Validation method is asymptotically optimal with respect to the local Risk criterion. It is believed that the proof can be extended to other regressors. Particular attention has been given to the kernel estimator for local smoothing parameters estimation. These techniques will be described in chapter 5.

3.2.4 Robust criterion

For the definition of the distance between two functions, we chose the $L^2$ norm, leading to a Risk function that can be estimated by using the $MSE$, the Mean Squared Error. This function penalizes an estimate for being far from the data with the square loss function. Let's now imagine that a bad measurement $Y_i^\#$ has been collected at $x_i$. Due to the square loss function that diverges quickly (quadratically) to infinity, the estimator $\hat{f}_i$ can not be far from $Y_i^\#$ at $x_i$ when minimizing the global $MSE$. Therefore, the data $Y_i^\#$ will attract the curve to itself. This behavior is unfortunate because $Y_i^\#$ was not representative of the value of the underlying function $f_i$. This kind of data is called an outlier, and due to the square loss function its influence on the behavior of the estimator is too great.

Developing estimators resistant to this kind of data leads to the so called robust estimators that are no longer linear functions of the data. We discuss here the main idea of the different robust techniques used, and we refer the reader to [2] for more information.

An interesting smoothing technique is to use instead of a moving average (i.e., weight sequence) a moving median. The estimator in this case is

$$\hat{f}(x_i) = \text{med}(Y_j : j \in N_i(\lambda))$$
where

\[ N_i(\lambda) = \{ k : x_k \text{ is one of the } \lambda \text{-Nearest Neighbors of } x_i, \lambda = 2j + 1 \} \]

Median smoothing is highly robust. However it gives a non smooth estimate of the underlying function, and (due to sorting) is computationally expensive.

The **L-smoothing** technique does local *trimmed averages* of \( \{ Y(k) \}_{k \in N_i(\lambda)} \) which means that no weights are put on the \( \alpha \) smallest and biggest order statistics in a neighborhood of \( x_i \)

\[ \hat{f}(x_i) = \sum_{l=i-j+\alpha}^{i+j-\alpha} Y(l) \]

The **R-smoothing** technique is based on a local nonparametric rank test. The idea is that the estimate \( \hat{f}(x_i) \) should be such that the two-sample test statistic based on the sample \( \{ Y_j - \hat{f}(x_i) \}_{j \in N} \) and \( \{ \hat{f}(x_i) - Y_j \}_{j \in N} \), is roughly zero. Again an expensive sorting is needed.

Since the square loss function does not give robust estimators, an **M-smoothing** technique downweights extreme residuals by modifying the loss function. A well known robust loss function is

\[
l(r) = \begin{cases} 
(1/2)r^2 & \text{if } |r| \leq c \\
 c|r| - (1/2)r^2 & \text{if } |r| > c 
\end{cases}
\]

where \( c \) is typically one or two times the standard deviation \( \sigma \). The appendix gives an estimator of \( \sigma^2 \) before smoothing (i.e., not based on the residual sum of squares). So the square loss function is extended in a continuous way by the absolute value loss function. So why not use the absolute value loss function on the all range of errors? The absolute value loss function is known to be robust but computationally expensive for parametric estimators known as LAD (Least Absolute Deviations) estimators [12].
estimation of the smoothing parameter $\lambda$ of a nonparametric LAD estimators is however not expensive: one can minimize the $MAE(\lambda)$ (Mean Absolute Error) function like the $MSE(\lambda)$ is minimized. The difficulty lies in the statistical properties of the MAE. It is indeed difficult to derive its bias to the LAD. Therefore the robust loss function (3.8) is preferred in the sense that it tends to give robustness to the estimator without disturbing its statistical properties.

Consistency and asymptotic normality have been derived for these robust linear nonparametric regression techniques.

3.3 How to obtain a confidence interval

Starting with a set of unbiased high variance data, a smoothing procedure will give a set of slightly biased smaller variance data using a local averaging. To improve the fit, local smoothing techniques and bias correction by bootstrapping procedures tend to achieve a better fit. Therefore, the assumption of a negligible bias compared to the variance is often reasonable.

This is the assumption made to get a confidence interval for the estimate. Then under the hypothesis of the Liapunov theorem [1], the local average estimator converges in distribution to a normal random variable with mean the underlying function and variance is $\sigma^2$ times the sum of the square of the weights, that is

$$\frac{\hat{f}_\lambda(x_i) - f(x_i)}{\sigma \sqrt{\sum_{j=1}^{n} w_{ij}^2}} \rightarrow N(0,1).$$

This result provides a quick way to get a confidence interval. However, for a small sample, the normal asymptotic distribution might not be appropriate. Moreover an estimate of the variance is required. Finally the assumption of a negligible bias compared to the variance
might be violated.

Many bootstrapping techniques have been developed to get a confidence interval. Härdle and Bowman [5] propose a bootstrapping procedure by resampling from the residuals \( \{e_i^* = Y_i - \hat{f}_\lambda(x_i) - m\}_{i=1,n} \) after smoothing a first time to get \( \lambda \) (\( m \) is the constant such that the residuals are centered around 0). By smoothing the \( \{Y_i^* = \hat{f}_\lambda(x_i) + e_i^*\}_{i=1,n} \), one gets \( b \) bootstrapped estimates of the underlying function. And a pointwise ranking of the \( bn \) bootstrapped estimates gives a pointwise variable confidence interval. This technique will adapt better to singularities in the underlying function (e.g., discontinuity) than the normal approximation approach.
Chapter 4
KERNEL SMOOTHER

4.1 Definition

Different forms of kernel estimators have been proposed (see Eubank [1]). For its properties and also for its simplicity, we will present the estimator suggested first by Nadaraya and Watson who were working on density estimation. The form of a kernel estimator is

\[
\hat{f}_h(Y; x_i) = \sum_{j=1}^{n} Y_j K\left(\frac{x_i - x_j}{h}\right) / \sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h}\right)
\]

\[
= \sum_{j=1}^{n} Y_j \frac{K((x_i - x_j)/h)}{\sum_{j=1}^{n} K((x_i - x_j)/h)}
\]

\[
= < Y, W_i(h; x, K) >_{\mathbb{R}^n}
\]

where \( W \) is a weight sequence determined by the smoothing parameter \( h \), the function \( K \), and the explanatory variables \( x \). Usually \( K \) is fixed and \( \lambda = h \) is the variable to select. The smoothing parameter, \( h \), is called the bandwidth because the weighted average will take place in a neighborhood of width \( h \). The function \( K \) is called the kernel function and gives different weight sequences for different \( h \)'s.

4.1.1 The kernel

The kernel is the function that determines the weight sequence and therefore has to satisfy some conditions. A kernel \( K \) is said to be of order \( p \) if

\[
\int_{\mathbb{R}} x^j K(x) dx = \begin{cases} 
1 & \text{if } j = 0 \\
0 & \text{if } j \in [1, p-1] \\
C \neq 0 & \text{if } j = p
\end{cases}
\]

(1)  
(2)  
(3)
Asymptotic results show that the higher the order, the smaller the bias of the estimate but the bigger the variance. When dealing with a finite set of data, the kernels of degree 2 perform as well as kernels of higher order. Therefore we will consider the case $p = 2$.

Condition (1) implies that the weight sequence will be roughly normalized. However, when using the Nadaraya and Watson kernel estimator, this condition is not necessary since the weight sequence is automatically normalized. This condition is necessary for the simpler estimator $\sum^n Y_j K\left(\frac{x_i - z_j}{h}\right)$ to assure a weighted average of the data.

Condition (2) forces the weight sequence to be symmetric.

The constant $C$ in condition (3) plays a role in the asymptotic bias.

A fourth condition that stems from asymptotic results assures that the asymptotic variance of the estimator is finite

$$\int_{\mathbb{R}} K^2(x)dx < \infty. \quad (4)$$

It is also sensible for the weight sequence to be nonnegative, which leads to the fifth condition

$$K(x) \geq 0, \ \forall x \in \mathbb{R}. \quad (5)$$

Finally, the weight sequence should be maximum at zero in order to put the most weight on the data where the estimation is taking place, i.e.,

$$K(x) \leq K(0), \ \forall x \in \mathbb{R}. \quad (6)$$

All these conditions impose some restrictions on the type of kernel function that can be used. A commonly used kernel is the rectangular or uniform kernel. The kernel technique becomes then equivalent to the k-Nearest Neighbor technique. This kernel is

$$K(x) = \begin{cases} 1/2 & |x| \leq 1 \\ 0 & |x| > 1. \end{cases}$$
Its support is compact and obviously it is not a continuous function. The estimator can be seen as a weighted sum of kernel functions (the weights being the $Y_i$'s). So the estimator is not a continuous function of the bandwidth $h$. For the class of kernels of order 2, it has the property of minimum variance [1].

The latter kernel gives equal weights. The Epanechnikov kernel [6] has a parabolic shape so that condition (6) is satisfied with a strict inequality. This kernel is the most used in practice because of its simplicity and asymptotic optimality. It is defined by

$$K(x) = \begin{cases} \frac{3(1 - x^2)}{4} & |x| \leq 1 \\ 0 & |x| > 1. \end{cases}$$

This kernel is a continuous function. So the estimator is a continuous function of the bandwidth $h$. Note, however, that this kernel is not differentiable at $x = \pm 1$, the boundaries of its compact support.

The estimate at $x_i$ uses a weighted average of $Y_j$'s in a neighborhood $N_i(h)$ of $x_i$. The latter kernel will create a sequence of weights inversely proportional to the distance from $x_i$ so that data far from $x_i$ contribute relatively less to the weighted average. It is reasonable to put no weight on $Y_j$ when $|x_i - x_j|$ is too big. This is what a compact support kernel estimator does. It also has the advantage of being computationally cheaper than infinite support kernel estimator. Infinite support kernel are yet used in some instances.

The Gaussian density function

$$K(x) = \exp\left(-\frac{x^2}{2}\right) / \sqrt{2\pi}$$

is often used for density estimation when the underlying density function is believed to have infinite support. Indeed an infinite compact kernel will reproduce better the tail behavior in the density function.
Regression is not concerned with the problem of estimating the tail of a density function. However regression techniques suffer from a lack of data at the boundaries of the domain. A point $x_i$ is said to be in the left boundary of the domain $[x_1, x_n] = [a, b]$ if $x_i - h < a$ (resp. $x_i + h > b$ for the right boundary). Therefore the weighted average close to the boundaries will not be symmetric. Three methods have been proposed to improve the goodness of fit of the estimator at the boundaries.

The first method is due to Gasser and Müller [10] (see also [7]). For each estimation in the boundary, the technique modifies the kernel function with the aim of preserving its order $p$. Indeed, a compact support kernel is no longer defined on $[-1, 1]$ in the left boundary but on $[-q, 1]$. Therefore, the kernel is no longer of order $p$ since

$$\int_{-q}^{1} K(x)dx \neq 1 \quad \text{and} \quad \int_{-q}^{1} xK(x)dx \neq 0.$$ 

Gasser and Müller propose to use a kernel $K_q(x)$ that stays of order $p$ for any $q$. For instance, the left boundary kernel corresponding to kernels of order 2 is

$$K_q(x) = \frac{4(q^3 + 1) - 6(1 - q^2)x}{(q + 1)^4}.$$ 

The advantage is that the estimator has the same bias order at the boundaries as inside the domain. However the variance of the estimator becomes big as $q$ tends to 0, even though its order is the same as inside the domain. Therefore this method does not improve the goodness of fit (bias square + variance) at the boundaries.

The second method is due to Rice [9] and is based on Richardson extrapolation. His boundary estimator

$$\tilde{f}(x_i) = \tilde{f}_{h}(x_i) + \beta(q)[\tilde{f}_{h}(x_i) - \tilde{f}_{\alpha h}(x_i)]$$

is a linear combination of estimates with bandwidth $h$ and $\alpha h$. $\beta(q)$ is chosen such that
the bias of the estimate in the boundaries is of the same order as inside the domain. Rice derives a simple formula for $\beta(q)$ for any kernel $K$ of order $p = 2$. The value of $\alpha$ is chosen such that the variance in the boundaries is of the same order as in the inside of the domain which leads to $\alpha = 1 - q$. Hence this method is driven by a single variable $h$ ($\beta$ and $\alpha$ are defined for any $q$). The method is simple and computationally inexpensive.

The third method is due to Peter Hall and Thomas E. Wehrly [8]. Their idea is to create pseudo data outside the range of the data $[x_{(1)}, x_{(n)}]$. These pseudo data come from a geometrical construction and differ from the true curve by an amount of $O(1/n^2)$ in the neighborhood of the the boundary points $x_{(1)}$ and $x_{(n)}$. So the boundaries are pushed further which allows good fitting in the real boundaries.

After all the restrictions that a kernel function has to satisfy, a wide class of kernels still remains. Asymptotic results will allow us to determine an "optimal" asymptotic kernel. However the notion of optimal is subjective. It can be a minimum variance, a minimum bias or a minimum Risk optimality criterion. With a finite set of data, the asymptotically optimal (relative to the Risk function) kernel estimator does not seem to give better results than any other sensible kernel estimator. Higher order kernels that asymptotically enhance the bias of the estimator do not obtain a better fit with a finite set of data. Therefore the choice of the kernel does not seem to be a big issue and the Epanechnikov kernel [6] is the most used in practice. The selection of the bandwidth is a more crucial aspect of kernel estimators.

4.1.2 The bandwidth

The bandwidth $h$ is the smoothing parameter.
If $h$ is smaller than the smallest distance from any point to its nearest neighbor, then

$$
\hat{f}_h(x_i) = \frac{Y_i K(0)}{K(0)} = Y_i \quad \forall i \in [1, n].
$$

This is the extreme case of undersmoothing since no smoothing takes place ($\hat{f} = I Y$).

The other extreme case occurs when $h$ is chosen so big that

$$
\hat{f}_h(x_i) = \frac{1}{n} \sum_{i=1}^{n} Y_i \quad \forall i \in [1, n].
$$

This will result in oversmoothing the underlying function by its average on the range of the data since then

$$
\hat{f}(x_i) \approx \int_{x(0)}^{x(n)} f(x) dx \quad \forall i \in [1, n].
$$

Two different types of smoothing parameter exist according to the type of the goodness-of-fit criterion: local or global.

When a unique variable drives the goodness-of-fit of the estimator, the criterion is said to be global. The Cross Validation technique is one technique to get an unbiased estimate of the Risk function for the class of nil-trace operators. The CV technique has the advantage of being relatively inexpensive, but a data point is dropped in the weighted average. The General Cross Validation requires the estimation of the trace of the matrix $W_h$ in order to achieve a nearly unbiased estimate of the Risk function. The trace of $W_h$ can be approximated by $K(0)/h$ since

$$\text{tr}(W_h) = \frac{K(0)}{\sum_{i=1}^{n} \sum_{j=1}^{n} K((x_i - x_j)/h)} \approx \frac{K(0)}{h}.$$

Therefore $GCV_n(h) = n \frac{\text{MSE}_n(h)}{(n-K(0)/h)^2}$ is as expensive as CV to compute. Obviously the trace of the kernel linear estimator is not a constant function of the bandwidth $h$ so that the regular $MSE$ is a biased estimate of the Risk.
For local criteria of goodness-of-fit, many techniques have been proposed. W. Härdle and A.W. Bowman [5] proposed a bootstrap technique. For this method the second derivative of the underlying function, $f''$, needs to be estimated with efficiency. J.G. Staniswalis [11] assumes that the variance of the noise, $\sigma^2$, is estimated with efficiency. Both of these techniques require the estimation of a quantity that cannot be estimated accurately. The technique that assume no estimability condition on either the variance, $\sigma^2$, or the second derivative, $f''$, is the one based on local weighted cross validation proposed by P. Vieu [4] to estimate the local Risk function. For further information and proofs, we refer to their publications [4], [5], [11].

4.2 Properties

The asymptotic properties will lead us to the selection of an “optimal” kernel. The bias of a compact support kernel of order $p = 2$ is of the order of $h^2$ since

$$E(\hat{f}(x_i)) = f(x_i) + h^2 \frac{f''(x_i)}{2} \int_{-1}^{1} u^2 K(u) du + O(h^4)$$

for $x_i$'s which are not in the boundaries. A proof of this result is stated in [7]. So the bigger the bandwidth, the bigger the bias where the underlying function has a large second derivative (i.e., a peak). On the other hand, when the underlying function is nearly flat, a bigger bandwidth will not damage the bias but improve the variance of the estimator since

$$\text{Var}(\hat{f}(x_i)) = \frac{\sigma^2}{nh} \int_{-1}^{1} K^2(u) du + O(h^5).$$

These two results shed some light on why a local bandwidth estimator should be used. They also give a clue to which type of numerical minimization technique will give a good estimate of the optimal bandwidth. Indeed the Risk function, $R_n(h)$, is asymptotically
convex since

\[ R_n(h) \approx \frac{1}{n} \sum_{i=1}^{n} \left[ h^2 f''(x_i) \int_{-1}^{1} u^2 K(u)du \right]^2 + \frac{\sigma^2}{nh} \int_{-1}^{1} K^2(u)du, \]

so that a golden section search is appropriate to locate the minimum of the Risk function. However the Risk function for a given finite set of data might have local minima. A way to check roughly if a correct bandwidth has been selected is to compare the pre- and post-variance of the noise (see appendix).

Based on these two results, one can derive the optimal bandwidth by minimizing over the bandwidth \(h\) the Risk function. Eubank states the result in [1] that under some regularity conditions (assuming that the data have been rescaled on [0,1]) the asymptotic minimizer of \(R_n(h)\) is

\[ h_{\text{opt}} = \left\{ \frac{\sigma^2}{n} \int_{-1}^{1} K^2(u)du \left[ \int_{-1}^{1} f''(u)^2 du \right] \int_{-1}^{1} u^2 K(u)du \right\}^{1/5} \]

and the asymptotic optimal Risk is

\[ R_n(h_{\text{opt}}) \approx \frac{1.25}{n^{4/5}} \left\{ \int_{0}^{1} f''(u)^2 du \right\}^{1/5} \left\{ \sigma^4 \int_{-1}^{1} K^2(u)du \int_{-1}^{1} u^2 K(u)du \right\}^{2/5} \]

so that the rate of convergence is of the order of \(n^{-4/5}\).

Based on this asymptotic result, one can determine the optimal kernel by minimizing the asymptotic Risk over \(K\). The solution to this minimization problem for kernel of order \(p = 2\) is the Epanachnikov kernel, which is such that

(4.1) \[ \left\{ \left[ \int_{-1}^{1} K^2(u)du \right]^2 \left[ \int_{-1}^{1} u^2 K(u)du \right] \right\}^{2/5} \approx .3491 \]

whereas for the rectangular kernel

(4.2) \[ \left\{ \left[ \int_{-1}^{1} K^2(u)du \right]^2 \left[ \int_{-1}^{1} u^2 K(u)du \right] \right\}^{2/5} \approx .3701 . \]
So the constant of the rate of convergence $n^{-4/5}$ is $(1.25)(.3491)$ for the Epanachnikov kernel, and $(1.25)(.3701)$ for the rectangular kernel.

In terms of the variance the rectangular kernel is optimal since

$$\int_{-1}^{1} K^2(u)du = .5$$

and for the Epanachnikov kernel

$$\int_{-1}^{1} K^2(u)du = .6 .$$

The definition of optimal is subjective but both the rectangular kernel and the Epanachnikov kernel are strong candidates for kernel function. The estimation of the bandwidth is however much more determinate than the selection of a kernel to achieve a good smoothing. The smooth estimates still carry some uncertainty in terms of bias and variance. Basically the operation of smoothing has decreased considerably the variance damaging the bias. It is therefore interesting to get a confidence interval in which the underlying function lies.

### 4.3 Confidence interval

Assuming that the magnitude of the bias of the estimates is negligible relative to the variance (i.e., $n \to \infty$, $h \to 0$ in such a way that $nh \to \infty$ and $nh^5 \to 0$), then under regularity conditions

$$\sqrt{nh} \frac{\hat{f}_h(x_i) - f(x_i)}{\sigma[\int_{-1}^{1} K^2(u)du]^{1/2}}$$

has an asymptotic standard normal distribution. So an approximate $100(1 - \alpha)\%$ confidence interval for $f$ is

$$\hat{f}_h(x_i) \pm z_{\alpha/2} \sigma \sqrt{\frac{1}{nh} \int_{-1}^{1} K^2(u)du}. $$
Note that $h_{opt}$ does not satisfy the condition $nh_{opt}^5 \to 0$ as $n \to \infty$, $h \to 0$ since $nh_{opt}^5 = O(n^{-1/5})^5n = O(1)$. This represents a drawback of the normal approximation method.

Bootstrapping procedures are more flexible but are computationally much more expensive for getting a confidence interval for $f$. For bootstrap techniques, we refer to [1] and [5].
Chapter 5
ITERATIVE SMOOTHER

5.1 Definition

An iterative nonparametric linear smoother is an estimator of the form

$$\hat{f}^{(j)} = W_j Y$$

where $W_j$ is a matrix of weight sequences, $\lambda = j$ is the smoothing parameter, $Y$ is the vector of noisy data, and $\hat{f}^{(j)}$ is the vector containing an estimation of the underlying function $f$ (at the $x_i$s) after the $j^{th}$ iteration.

This smoother is called "iterative" because successive estimates of $f$ are achieved by iterating the weighted sequence until the criterion of goodness of fit is optimum. Therefore, the smoothing parameter is the power of exponentiation of the matrix $W$ since

$$(5.1) \quad \hat{f}^{(j)} = W \hat{f}^{(j-1)} = W^j \hat{f}^{(0)} = W^j Y.$$ 

Here, the smoothing parameter, $j$, is a nonnegative integer. Choosing $j$ equal to zero amounts to performing no smoothing at all, since $\hat{f}^{(0)} = W^{(0)}Y = IY = Y$. Choosing $j$ on the other end of the domain, namely infinity, corresponds to a limit of the smoothing procedure, and it will be interesting to look at the smoothing result of such a limit, if it exists.

The other important element of the iterative smoother is the definition of the weight sequence matrix, $W$. It should have some sensible smoothing characteristics, as well as some nice statistical and computational properties. The weight sequence matrix we are
proposing meets these objectives in some ways; other sensible weight sequence matrices will have other properties. A comparison between two different iterative procedures will be based on the rate of convergence to zero of their Risk function (3.2), their ability to estimate the Risk function with a finite set of data, and their computational expense.

Without loss of generality, we will assume until the end of the section that the data are equally spaced, i.e. \( x_{i+1} - x_i = h = 1/(n - 1), \ i = 1, \ldots, n - 1. \)

For \( 2 \leq i \leq n - 1, \) we note that

\[
E\left( \frac{Y_{i-1} + Y_{i+1}}{2} \right) = f(x_i) + \frac{h^2}{2} f''(x_i) + O(h^4) \quad \text{and} \quad Var\left( \frac{Y_{i-1} + Y_{i+1}}{2} \right) = \sigma^2/2,
\]

so that the two weights of 1/2 are symmetric around \( x_i \) on \( Y_{i-1} \) and \( Y_{i+1}. \) If the data are not equally spaced then the weights are no longer 1/2 but are such that the bias of the same order.

For the two extreme points of the range of the data, namely at \( x_1 \) and \( x_n, \) the same weighted average can not be used and

\[
E\left( \frac{Y_1 + Y_2}{2} \right) = f(x_1) + \frac{h^2}{2} f''(x_1) + O(h^4) \quad \text{and} \quad Var\left( \frac{Y_1 + Y_2}{2} \right) = \sigma^2/2
\]
\[
E\left( \frac{Y_{n-1} + Y_n}{2} \right) = f(x_n) - \frac{h^2}{2} f''(x_n) + O(h^4) \quad \text{and} \quad Var\left( \frac{Y_{n-1} + Y_n}{2} \right) = \sigma^2/2,
\]

so these two estimates do not have a symmetric weight sequence. A point is said to be in the boundary if the weight sequence to estimate at this point is not symmetric around the point.

This estimate of the function is now slightly biased but the variance has been reduced by half. This estimate will represent the first step of an iterative scheme

\[
\hat{f}^{(1)}(x_i) = \begin{cases} 
(Y_1 + Y_2)/2, & i = 1 \\
(Y_{i-1} + Y_{i+1})/2, & 2 \leq i \leq n - 1 \\
(Y_n + Y_{n-1})/2, & i = n.
\end{cases}
\]
This particular design gives the weight matrix

\[
W = \frac{1}{2} \begin{pmatrix}
1 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
. & . & . & . \\
0 & 1 & 0 & 1 \\
1 & 1 \\
\end{pmatrix},
\]

and the iterative estimate is defined as (5.1).

In the remainder of this section, we will look at the statistical properties of this iterative smoother in terms of the discussion of Chapter 4.

5.2 Properties

The trace of the matrix \( W \) is the most interesting feature of this design because it enhances statistical and computational properties.

Theorem 5.1 Under the assumption that the data are equally spaced, the matrix \( W \) has the property that

\[\text{tr}(W^{2j+1}) = 1, \forall j \in \mathbb{N}.\]  

Proof: Let \( T = 2W \). \( W \)'s eigenvalues are \( T \)'s eigenvalues divided by 2. Solving \( Tv = \lambda v \) is equivalent to solving the system

\[v_{i-1} - \lambda v_i + v_{i+1} = 0, \ i = 1, n\]

with the boundary conditions

\[
\begin{cases}
v_0 &= v_1 \\
v_{n+1} &= v_n
\end{cases}
\]

Solving the characteristic equation

\[r^2 - \lambda r + 1 = 0\]
and the boundary conditions leads to the eigenvalues of $T$, namely,

$$\lambda_k(T) = 2\cos\left(\frac{k\pi}{n}\right), \ k = 0, n - 1$$

or

$$\lambda_k(W) = \cos\left(\frac{k\pi}{n}\right), \ k = 0, n - 1.$$

And by noting that

$$\cos\left(\frac{(n-k)\pi}{n}\right) = -\cos\left(\frac{k\pi}{n}\right), \ k = 1, n$$

the trace is

$$tr(W^{2j+1}) = \cos(0) + \sum \lambda_k^{2j+1} + (-\lambda_k)^{2j+1}
= 1$$

for every natural number $j$.

\[
\square
\]

Due to this property, the Mean Square Error is biased to the Risk function by a constant. So we can get a cheap (no trace has to be estimated) and unbiased (minimizing $MSE(\cdot)$ is equivalent to minimizing $R(\cdot)$) estimate of the optimal smoothing parameter. The matrix to the power an even number does not have a trace equal to one, but a trace tending to one as the number of iterations tends to infinity. It would be expensive to estimate, and therefore, we are redefining the iterative estimator by

$$\hat{f}^{(j)} = W^{2j+1}Y, \ j \in \mathcal{N}.$$

And for computational efficiency, we can write this new estimate in the form

$$\hat{f}^{(j)} = [W^{2j}]^j(WY), \ j \in \mathcal{N}.$$
We note that when the data are not equally spaced this property is no longer true. However $\lambda_1 = 1$ is still the largest eigenvalue and the other ones are almost opposite of each other, so that if the data are unequally spaced, the selection of the smoothing parameter will be almost unbiased. An open question is by how much the unequally spaced design will affect the selection of the smoothing parameter.

The limit of the smoothing procedure, as the smoothing parameter tends to infinity, is important statistical information.

**Theorem 5.2** For equally spaced or unequally spaced design, the smoothing matrix, $W$, is such that

\begin{equation}
W_j \xrightarrow{j \to \infty} \frac{1}{n} J_{n \times n},
\end{equation}

where $J$ is the matrix full of 1’s.

**Proof:** $W$ is a transition matrix of a regular Markov chain. Therefore $W_j \xrightarrow{j \to \infty} L$ such that each row of $L$ is the same probability vector $l = (l_1, \ldots, l_n)$ (see [17]). The vector $l$ is unique and such that $Ll = l$. Note that the probability (eigen) vector (associated to the eigenvalue 1) $l = (1, \ldots, 1)/n$ satisfies $Ll = l$.

\[ \square \]

The average of the function between $x(1)$ and $x(n)$, the constant function of equation $y = \int_{x(1)}^{x(n)} f(x)dx/(x(n) - x(1))$, is the approximate smoothing limit. For the kernel smoother of Chapter 5, this corresponds to an infinite bandwidth $h$. For most of the underlying functions, reaching this limit would give an oversmoothing estimate. But, if the underlying function is the constant function, then the best regression, in terms of minimizing the Risk, is the limit of the smoothing procedure. This might represent a
drawback of the method since, in practice, one does not want to wait an infinite time to get the regression estimate. However, one rarely wants to smooth data from a constant function, or, if yes, then a parametric model will be obvious.

The rate of convergence to zero of the Risk function is used for comparing the efficiency of nonparametric linear smoothers. The formula (3.3) is useful for this purpose. The bias and variance of the estimates at the \( x_i \)'s are needed. The weights come from a Taylor series argument, so the rows of \( W^j \) always represent a weighted average with coefficients the binomial coefficients \( \left( \frac{1}{2^j} \sum_{i=1}^{j} \binom{j}{i} \right) = \frac{1}{2^j} (1+1)^j = 1 \). Hence, the variance of any estimate is

\[
Var(f^{(j)}(x_i)) = \sigma^2 / (2^j)^2 \sum_{k=0}^{j} \binom{j}{k}^2, \quad j = 0, 1, 2, \ldots \quad i = 1, \ldots, n
\]

which makes the variance function \( v \) in (3.3)

\[
v(j) = \frac{\sigma^2}{4^j} \sum_{k=0}^{j} \binom{j}{k}^2 \quad i.e. \quad v(j) = \frac{\sigma^2}{4^j} \binom{2j}{j}
\]

So the variance function in (3.3) after \( j \) iterations is

\[
(5.4) \quad v(j) = \frac{\sigma^2}{4^j} \binom{2j}{j}
\]

The bias function \( b \) in (3.3) is not as easy to estimate. Indeed, the bias of the estimates, after one iteration, is of order \( h^2 \), except for the two estimates at the boundary points, for which it is of order \( h \). And the more iterations, the more estimates with bias of order \( h \). However, we are looking for the rate of convergence of the Risk function as the number of points, \( n \), tends to infinity. So, the bias of order \( h \) of the \( 2(2j - 1) \) boundary estimates after \( j \) iterations is negligible compared to the other \( n - 2(2j - 1) \) "inside" estimates that
have not been contaminated by the boundary modification as long as \( n/j \to \infty \). The squared bias function at any location \( x_i \) is defined by

\[
b_i(j) = jh^2 f''(x_i)/2.\]

Indeed for an "inside" point \( x_i \), the bias of the estimate is

\[
b_i(j) = \frac{h^2}{2^j} \left[ \sum_{i=0}^{(j-1)/2} \binom{j}{i} (j - 2i)^2 f''(x_i) \right]
= jh^2 f''(x_i)/2.\]

So the square of the bias function in (3.3) after \( j \) iterations is

\[
b^2(j) = j^2 h^4 \sum_{i=1}^{n} [f''(x_i)]^2 / (4n).\]

Assuming that the second derivative of \( f \) exists and is in \( L_2 \), the asymptotic bias function is

\[
(5.5) \quad b^2(j) = j^2 h^4 I / 4
\]

where \( I = \int [f''(x)]^2 dx \).

**Theorem 5.3** Under the assumption that \( n/j \to \infty \) and that \( f'' \in L_2 \), the asymptotic Risk function for this estimator is

\[
(5.6) \quad R(j) = \frac{j^2}{4n^4} I + \frac{\sigma^2 h^2}{4^3} \binom{2j}{j},
\]

the minimum is obtained for

\[
(5.7) \quad j_{\text{opt}} = \left( \frac{\sigma^2 n^4}{\sqrt{\pi I}} \right)^{2/5},
\]

and the asymptotic optimal Risk is

\[
(5.8) \quad R(j_{\text{opt}}) = \frac{1.25}{n^{4/5} \pi^{2/5}} \sigma^{8/5} I^{1/5}.\]
**Proof:** The first part of the proof, (5.6), stems from the identity (3.3), and from the results (5.4) and (5.5).

To get the optimal asymptotic smoothing parameter, (5.7), one needs to minimize the asymptotic Risk function (5.6) and hence solve

\[
\frac{R(j+1) - R(j)}{\sigma^2} = 0
\]

\[\iff \quad \frac{2j+1}{4\pi^4} I = \sigma^2 \frac{4}{4j+1(j+1)} \binom{2j}{j}\]

Using the approximation that

\[j! \approx \sqrt{2\pi} e^{-j} j^{j+1/2}\]

we can show that

\[\frac{1}{4j} \binom{2j}{j} \approx 1/\sqrt{j\pi}.
\]

Therefore,

\[
\frac{(2j+1)(j+1)}{2\pi^4} I = \frac{\sigma^2}{\sqrt{\pi} j} \iff \frac{(2j+1)(j+1)}{\sqrt{\pi} I} = \frac{2\sigma^2 n^4}{\sqrt{\pi} I}
\]

which, for \(j\) not significantly small, is approximately equivalent to solving

\[j^2 \sqrt{j} = \frac{\sigma^2 n^4}{\sqrt{\pi} I} \iff j_{\text{opt}} \approx \left(\frac{\sigma^2 n^4}{\sqrt{\pi} I}\right)^{2/5}.
\]

Then by plugging \(j_{\text{opt}}\), (5.7), into \(R(j)\), (5.6), one gets the result (5.8).

\[\square\]

However the condition \(n/j_{\text{opt}} \xrightarrow{n \to \infty} \infty\) is not satisfied unless \(\frac{1}{\sigma^2} = O(n^{\frac{3}{2} + \alpha})\), \(\alpha > 0\).

This later condition means that the ratio of the wigglyness of the data, \(I\), over the variance of the noise, \(\sigma^2\), has to be high to a certain order and that as the number of data tends to infinity the wigglyness of the underlying function must be great relatively to the variance of the noise. This is a meaningful assumption to have on a smoother. If the condition is
not satisfied, the performance of the smoother will be poorer due to the first order bias introduced from the boundaries into the inside.

For this particular optimal smoothing parameter, the Risk function is (5.8). So the estimator is consistent and the rate of convergence is \( n^{-4/5} \) like the kernel smoother and the constant, \((1.25)(.6326)\), is bigger than the constant of the kernel smoother with the Epanachnikov kernel (4.1) and the rectangular one (4.2).

**The Risk function is asymptotically convex.** We have seen that the Risk function of a kernel smoother is asymptotically convex, so that a golden section procedure can locate the minimum of \( R_n(h; Y, x) \). For an iterative smoother, the Risk function is estimated at each step so that the descent to its minimum is progressive. The convexity result gives a rule for stopping the iterations when the estimated Risk function (i.e., the MSE) stops decreasing. The iterative search represents an alternative to the golden section search. We might expect the iterative search to be faster than the golden section search when the underlying function is wiggly and slower when the underlying function is highly linear.

**Theorem 5.4** *The Risk function* \( R_n(j; Y, x) \) *is asymptotically a convex function of the smoothing parameter* \( j \).

**Proof:** The asymptotic Risk function is given by (5.6). Obviously the bias term is an increasing function of \( j \) and is a convex function since \( b^2(j) = Cj^2 \).

The variance term is a decreasing function of \( j \) since

\[
\frac{(v(j+1) - v(j)) \sigma^2}{\sigma^2} = \left( \frac{2(j+1)}{j+1} \right) / 4^{j+1} - \left( \frac{2j}{j} \right) / 4^j
\]

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

\[
= -\frac{2}{4^{j+1}(j+1)} \left( \frac{2j}{j} \right) < 0.
\]
One can also show that

\[
\frac{(v(j) - v(j - 1))}{\sigma^2} = -\frac{1}{4^j(2j - 1)} \binom{2j}{j}.
\]

So its second derivative is positive since

\[
\frac{(v(j + 1) - 2v(j) + v(j - 1))}{\sigma^2} = -\frac{2}{4^j+1(j + 1)} \binom{2j}{j} + \frac{1}{4^j(2j - 1)} \binom{2j}{j}
= \binom{2j}{j} \frac{1}{4^j(j-1)(j+1)} > 0.
\]

The local smoothing procedure tends to locally oversmooth points in the boundaries since \(tr(W_\lambda S_i^2) > 0\) in (3.7). However, if the condition \(\frac{1}{\sigma^2} = O(n^{\frac{3}{4}+\alpha})\), \(\alpha > 0\) holds, the unbiasedness of the local estimate will improve the local fitting since, for “inside” points, \(tr(W_\lambda S_i^2) = 0\) in (3.7). The local procedure is, however, expensive unless a computation trick is used. Recall that \(LMSE_{ij}^{(j)}\), (3.6), the local MSE at \(x_i\) after \(j\) iterations, is the local criterion of goodness-of-fit

\[
LMSE_{ij}^{(j)} = \sum_{k=i-K_i}^{i+L_i} (Y_k - \hat{f}_k^{(j)})^2 s_{ik},
\]

where

\[
\begin{align*}
    s_{ik} &= (1 - (\frac{x_i-x_k}{h})^2) \text{ is any weight sequence} \\
    &= \text{chosen here to be the Epanachnikov one.} \\
    h &= \text{the (fixed) span of the local compact support,} \\
    &= \text{typically } h = 5\% (x(n) - x(1)). \\
    \hat{f}_k^{(j)} &= \text{the estimate at } x_k \text{ after } j \text{ iterations.} \\
    Y_k &= \text{the data at } x_k. \\
    K_i, L_i &= \text{are such that the weights } s_{ij} \text{ in the sum (5.9) are not equal to 0.}
\end{align*}
\]

Let \(r_k^{(j)} = (\hat{f}_k^{(j)} - Y_k)^2\) and \(\epsilon_{ik} = \frac{x_i-x_k}{h}\). Then (5.9) becomes

\[
LMSE_{ij}^{(j)} = \sum_{K_i,L_i} r_k^{(j)} (1 - \epsilon_{ik}^2)
= \sum_{K_i,L_i} r_k^{(j)} - \sum_{K_i,L_i} r_k^{(j)} \epsilon_{ik}^2
= R_i^{(j)} - W_i^{(j)}
\]
Let also \( P_i^{(j)} = \sum_{K, L} r_k^{(j)} e_{ik} \).

It is expensive to estimate the local \( \text{LMSE}_i^{(j)} \)'s for each \( j = 1, \ldots, j_{\text{opt}} \) and for each \( i = 1, \ldots, n \). The global number of iterations, \( j_{\text{opt}} \), is equal to the maximum of the local number of iterations \( j_{\text{opt}}(i), i = 1, \ldots, n \) (i.e., \( j_{\text{opt}} = \max_i j_{\text{opt}}(i) \)). For each \( j \) the number of computations is \( O(n^2) \).

An updating method decreases the computation time. The number of computations of the following updating technique is \( O(n) \) for each \( j \). Hence if the total number of iterations \( j_{\text{opt}} \) is small the smoothing will be efficient. Simulations tend to show that sometimes \( j_{\text{opt}} \) is large because \( j_{\text{opt}}(i) \) is large for very few \( i \)'s only. Using two different spans, \( h_1 = 5\% \) and \( h_2 = 10\% \), will improve this bad behavior. A subjective stopping rule can be used to prevent too many iterations, such as the one used when the assumption \( J_{\text{opt}} = O(n^{3+\alpha}), \alpha > 0 \) is not satisfied. Indeed in this case the local iterative smoothing technique tends to locally oversmooth so that stopping the iterations for \( j \) too great is a reasonable thing to do. A reasonable stopping rule is \( \max_i |j_i^{(j)} - j_i^{(j-1)}| < \epsilon \).

Knowing \( \text{LMSE}_{i+1}^{(j)} \), one can get \( \text{LMSE}_{i+1}^{(j)} \) by updating. Indeed, with the Epanachnikov weights for the \( s_{iks} \),

\[
\text{LMSE}_{i+1}^{(j)} = R_{i+1}^{(j)} - W_{i+1}^{(j)}
\]

and \( R_{i+1}^{(j)}, W_{i+1}^{(j)} \) are found to be

1. \( R_{i+1}^{(j)} = R_i^{(j)} + \Delta_i^r \)
2. \( W_{i+1}^{(j)} = W_i^{(j)} + d_i^2 R_i^{(j)} + 2d_i P_i^{(j)} + \Delta_i^w \)
3. \( P_{i+1}^{(j)} = P_i^{(j)} + d_i R_i^{(j)} + \Delta_i^p \)
where

\[
\begin{align*}
    d_i &= \frac{x_{i+1} - x_i}{h} \\
    \Delta_i^r &= - \sum_{k=i-K_i}^{i+L_i+1} r_k + \sum_{k=i+1+L_i}^{i+1+L_i+1} r_k \\
    \Delta_i^p &= - \sum_{k=i-K_i}^{i+1+L_i+1} r_k (e_{ik} + d_i)^2 + \sum_{k=i+1+L_i+1}^{i+1+L_i+1} r_k (e_{ik} + d_i)^2 \\
    \Delta_i^p &= - \sum_{k=i-K_i}^{i+1+L_i+1} r_k (e_{ik} + d_i) + \sum_{k=i+1+L_i}^{i+1+L_i+1} r_k (e_{ik} + d_i).
\end{align*}
\]

Thanks to this technique the computation time is linear in the number of data \( n \).

5.3 Confidence Interval

The variance of the estimate at \( x_i \) after \( j_i = j_{\text{opt}}(i) \) iterations is given by \( \frac{s^2}{n_i} \left( \frac{2j_i}{j_i} \right) \).

By normal approximation, one can get a local confidence band for the underlying function \( f \).

As with the kernel smoother bootstrap to get a confidence interval, one can bootstrap on the residuals to get a confidence interval.

5.4 Example

The following data come from a simulation: the sawtooth function is the underlying function

\[
f(x) = \begin{cases} 
    2x, & 0 \leq x < 1/2 \\
    2x - 2, & 1/2 \leq x \leq 1 
\end{cases}
\]

and the noise is normal \( N(0, 1/3) \).
Figure 5.1: Example of a local smoothing iterative procedure
REFERENCES


APPENDIX
THE VARIANCE OF THE NOISE $\sigma^2$

In nonparametric statistics the variance of the noise is a statistical quantity that can be estimated, and is useful in many ways.

The variance of the noise can be estimated before smoothing.

Let’s suppose for simplicity that the $x_i$'s are random variables (i.e., dimension 1) and equally spaced (i.e., $x_{i+1} - x_i = h = 1/(n - 1)$). With the raw data, $\sigma^2$ can be estimated before smoothing by

$$\hat{\sigma}^2 = \frac{Y' A' A Y}{6(n - 2)}$$

where

$$A_{(n-2) \times n} = \begin{pmatrix} 1 & -2 & 1 & 1 & -2 & 1 & \cdots & 1 & -2 & 1 \\ 1 & -2 & 1 & 1 & -2 & 1 & \cdots & 1 & -2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & -2 & 1 & 1 & -2 & 1 & \cdots & 1 & -2 & 1 \end{pmatrix}, \quad E(Y) = \mu = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}, \quad D(Y) = \sigma^2 I_{n \times n}.$$

The expected value of $\hat{\sigma}^2$ is

$$E(\hat{\sigma}^2) = \frac{\sigma^2 tr(A'A) - \mu'A'A \mu}{6(n - 2)}$$

$$= \sigma^2 + \frac{1}{6(n - 2)} \sum_{i=2}^{n-1} (f(x_{i-1}) - 2f(x_i) + f(x_{i+1}))^2$$

$$= \sigma^2 + \frac{h^4 \sum_{i=2}^{n-1} (f''(x_i))^2}{6} + O\left(\frac{1}{n^2}\right).$$

So, assuming that $f''$ is in $L^2$, $bias(\hat{\sigma}^2) \approx h^4/6 \int (f''(x))^2 dx = O(1/n^4)$.

The variance of $\hat{\sigma}^2$ is

$$Var(\hat{\sigma}^2) = \frac{1}{36(n - 2)^2} [(\mu_4 - 3\mu_2^2)a'a + 2\mu_2^2 tr((A'A)^2) + 4\mu_2 \mu'(A'A)^2 \mu + 4\mu_3 \mu'A'A \mu]$$

$$= \frac{1}{36(n - 2)^2} [(\mu_4 - 3\mu_2^2)(36n - 92) + 2\mu_2^2(70n - 176)$$

$$+ 4\mu_2 \mu'(A'A)^2 \mu + 4\mu_3 \mu'A'A \mu]$$
where $\mu_i$ is the $i^{th}$ moment of $Y$ and $a$ is the diagonal vector of $A'A$.

So assuming that the noise has a finite $4^{th}$ moment

$$Var(\hat{\sigma}^2) = O\left(\frac{1}{n}\right).$$

Assuming the symmetry in the distribution of the noise is reasonable in many practical cases. So the term with $\mu_3$ is zero. However, none of the higher moments are estimable.

In order to get a confidence interval for $\sigma^2$, we will make the stronger assumption that the noise is normally distributed. In this case, $\mu_4 = 3\mu_2^2$ and the variance of the estimate becomes simply

$$Var(\hat{\sigma}^2) = \frac{1}{36(n-2)^2} \left(2\sigma^4(70n - 176) + 4\sigma^2 \mu'(A'A)^2\mu\right).$$

Using $\mu'(A'A)^2\mu = (A'A\mu)'(A'A\mu)$ and a Taylor series expansion of $f$ to its $4^{th}$ order, we can see that $\mu'(A'A)^2\mu = O\left(\frac{1}{n^4}\right)$. Hence

$$Var(\hat{\sigma}^2) = \frac{2\sigma^4(70n - 176)}{36(n-2)^2} + O\left(\frac{1}{n^6}\right).$$

is the variance of the nearly unbiased estimate $\hat{\sigma}^2 = \sum_{i=2}^{n-1}(Y_{i-1} - 2Y_i + Y_{i+1})^2/6(n-2)$.

**Theorem** Under the assumption that $f''$ exists and that the noise is i.i.d. normally distributed $N(0, \sigma^2)$, an approximate $(1 - \alpha)$% confidence interval for $\sigma^2$ is

$$\left[\frac{\hat{\sigma}^2}{1 + z_{\alpha/2}^2/3\sqrt{\frac{35}{n-2}}}, \frac{\hat{\sigma}^2}{1 - z_{\alpha/2}^2/3\sqrt{\frac{35}{n-2}}}\right].$$

**Proof:** Let $Z_i = Y_i - 2Y_{i+1} + Y_{i+2}$, $i = 1, \ldots, n - 2$.

They are distributed $N(\mu_i - 2\mu_{i+1} + \mu_{i+2}, 6\sigma^2)$. If the underlying function is assumed to be twice differentiable, then they are nearly distributed $N(0, 6\sigma^2)$. So, the $Z_i^2$'s form a stationary sequence of dependent random variables.
Let $V_i = \mathbb{Z}_i^2 - 6\sigma^2$. The $\hat{\sigma}^2$ is equal to $\sum_{i=1}^{n-2} V_i / 6(n-2) + \sigma^2$. To apply the usual central limit theorem, we need the $V_i$'s to be independent, which is not the case here. In [16], a slightly different version of the theorem is proposed, where the variables are $m$-dependent. By construction, the $V_i$'s form a 2-dependent and stationary sequence. Each variable has expected value zero and a finite $12^{th}$ moment since the noise is assumed to be normally distributed. If $S_{n-2} = V_1 + \ldots + V_{n-2}$, Theorem 27.5 ([16], p. 316) assures that

$$Var(S_{n-2})/n \xrightarrow[n \to \infty]{} \rho^2 = E(V_i^2) + 2 \sum_{i=1}^{2} E(X_1 X_{1+i})$$

$$= 72\sigma^4 + 32\sigma^4 + 2\sigma^4$$

$$= 140\sigma^4$$

where the series converges absolutely.

Moreover $\rho = \sqrt{140}\sigma > 0$. So $S_{n-2}/\rho \sqrt{n-2} \xrightarrow[n \to \infty]{} N(0,1)$. Therefore

$$P(-z_{\alpha/2} < 6\sqrt{\frac{n-2}{140}} \frac{\hat{\sigma}^2 - \sigma^2}{\sigma^2} < z_{\alpha/2}) = 1 - \alpha$$

$$\iff P(-z_{\alpha/2} \sqrt{\frac{35}{(n-2)}} + 1 < \frac{\hat{\sigma}^2}{\sigma^2} < z_{\alpha/2} \sqrt{\frac{35}{(n-2)}} + 1) = 1 - \alpha$$

$$\iff P(\frac{\hat{\sigma}^2}{1+z_{\alpha/2} \sqrt{\frac{35}{(n-2)}}} < \sigma^2 < \frac{\hat{\sigma}^2}{1-z_{\alpha/2} \sqrt{\frac{35}{(n-2)}}}) = 1 - \alpha.$$ 

\(\square\)

After smoothing, the variance of the noise can be estimated by the usual Residual Sum of Squares

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (Y_i - \hat{f}(x_i))^2}{n}$$

$$= \frac{1}{n} Y'(I - W_{\lambda_{opt}})^2 Y.$$ 

This is a nearly unbiased estimate of $\sigma^2$ under the null hypothesis that the optimal smoothing parameter has indeed been selected. There is unfortunately no parametric test for

$$H_0 : \sigma^2 = \hat{\sigma}^2$$

$$vs$$

$$H_1 : \sigma^2 \neq \hat{\sigma}^2.$$ 

But looking at both the pre- and post- estimate of $\sigma^2$ can identify a bad selection of the smoothing parameter.