Chapter 1

Local polynomial fitting

J. FAN
University of North Carolina
Chapel Hill, USA
& The Chinese University of Hong Kong

I. GIJBELS
Catholic University of Louvain
Institute of Statistics
Louvain-la-Neuve, Belgium

1.1 Introduction

Nonparametric regression is on data-analytic approaches to regression problems arising from many statistical applications. The aim of these model-free nonparametric approaches is to release the restrictive assumptions imposed on the form of a regression function and to let data search for a suitable function that describes well the given data. These approaches are powerful in detecting possible nonlinearity and in exploring fine functional relationships, and provide useful diagnostic tools to parametric regression models. Local polynomial fitting is a powerful nonparametric method and can easily be understood and implemented.

1.1.1 Nonlinear regression

There is a surge of interest in nonlinear regression in the literature. With advancement of modern sciences and technologies, there are many situations in which a linear relationship can not describe well the structural relationship among a few variables. Consider for example the data presented in Figure 1.1, where the relationship between the concentration of nitric oxides in engine exhaust (taken as dependent variable) and the equivalence ratio (taken as independent variable), a measure of the richness of the air/ethanol mix, is depicted for a burning of ethanol in a single-cylinder automobile test engine. From visualization, one can see that the relationship between the concentration of nitric oxides and the equivalence ratio is highly nonlinear. What can we do with such a nonlinear data set? Traditionally, trial-and-error on polynomial regression are often employed. Figure 1.1 presents the resulting fits by using four different degrees of polynomials. One can easily see that all resulting

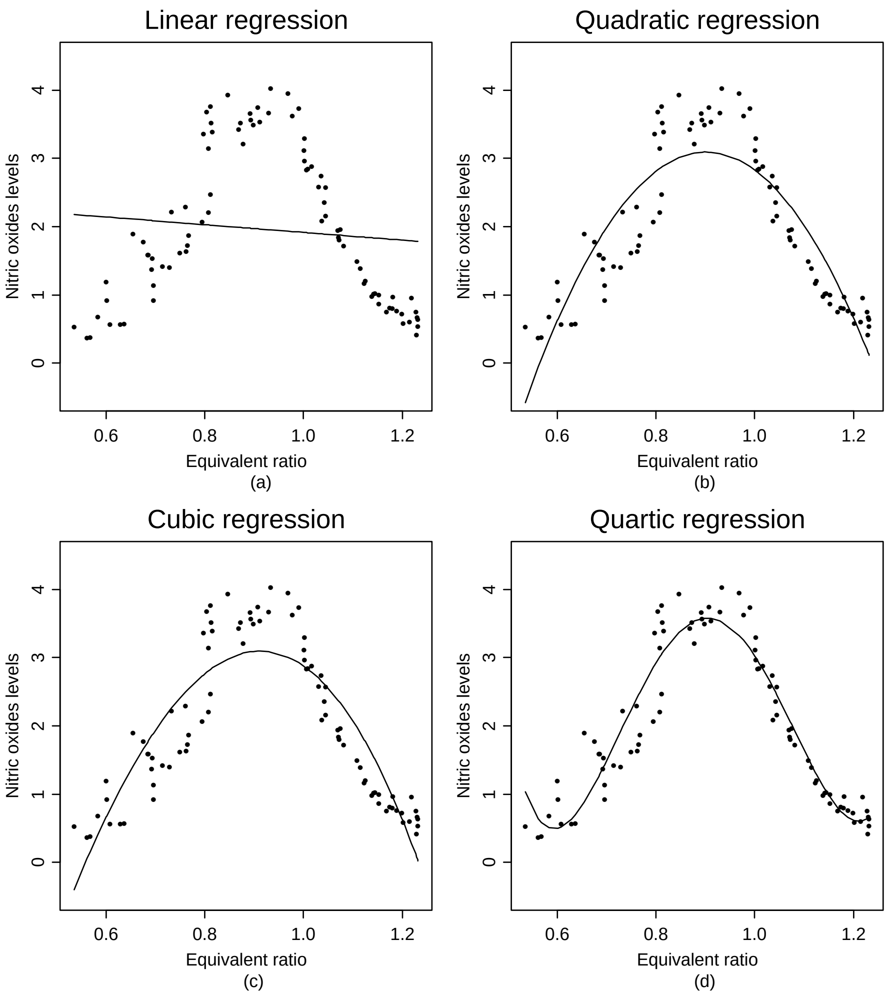

Figure 1.1: Polynomial fits to the ethanol data. Presented are the scatterplots of the concentration of nitric oxides against the equivalence ratio along with the fitted polynomial regression functions.

fits have large biases. Having homogeneous degrees of smoothness, polynomial functions are not effective in modeling many nonlinear functions. Witness of this includes the ethanol data in Figure 1.1 and an example on a motorcycle data set presented in Figure 1.2 of Fan and Gijbels (1996).

1.1.2 Why smoothing is needed?

The nonlinearity of the ethanol data is detected with the assistance of scatterplot visualization. This approach, while simple, has serious limitations. For one thing, it can not process data beyond the domain of our visualization.

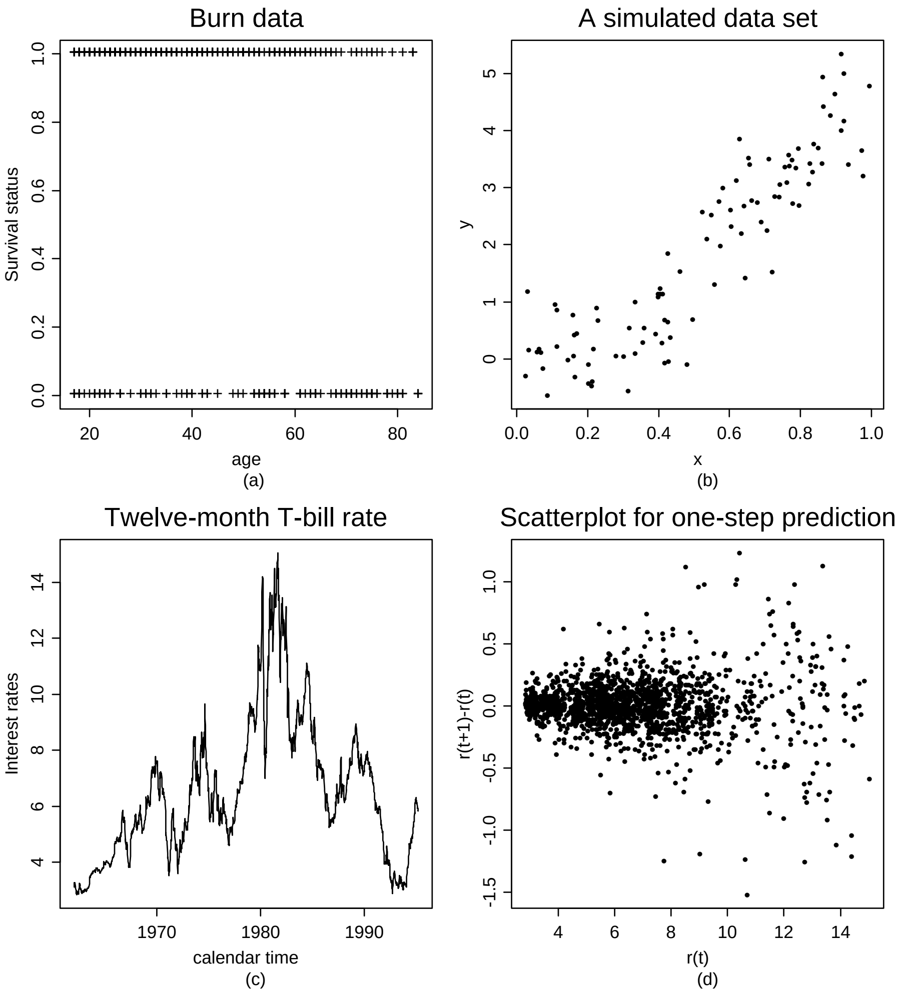

Figure 1.2: Limitations of scatterplot visualization. (a) Binary response data whose mean regression function is hard to image without a nonparametric smoother; (b) A simulated data set for which the discontinuity of the mean regression function at $x = 0.5$ is hard to be detected without a derivative estimator. (c) The yields $\{r_t\}$ of twelve-month Treasury bill. (d) The scatterplot of $\{r_t\}$ versus $\{r_{t+1} - r_t\}$ whose conditional variance function is hard to picture by visual inspection.

Consider the burn data presented in Figure 1.2 (a) which consist of 435 adults between ages 17 and 85 suffering from burn injuries. The data were collected at the University of Southern California General Hospital Burn Center and contain information on several variables. The binary response variable is taken to be 1 for those who survived their burn injuries and zero otherwise, and the age of the victims is taken as the covariate. One has a difficult time to trace the mean response curve via using scatterplot visualization. Figure 1.2 (b) presents a simulated set of data of size 100 from the model $Y = 4X^2 + I(X > 0.5) + \varepsilon$, where $\varepsilon \sim N(0; 0.5^2)$ is independent of $X \sim$ Uniform[0, 1]. From scatterplot visualization, one can hardly detect a jump at the point $x = 0.5$. For this purpose, one needs to visualize the derivative curve of the regression function, which is beyond the domain of our visualization. Figure 1.2 (c) concerns the yields of the twelve-month U.S. Treasury bill from the second market rates (on Fridays), which are annualized using a 360-day year of bank interest and quoted on a discount basis. The data consist of 1,735 weakly observations, from January 5, 1962 to March 31, 1995. Let r_t denote the interest rate on week t. Figure 1.2 (d) gives the scatterplot for the one-step ahead prediction where the points $\{(r_t, r_{t+1} - r_t)\}$ are presented. It is difficult to picture the volatility function $\sqrt{\mathrm{Var}(r_{t+1}|r_t)} = \sqrt{\mathrm{Var}(r_{t+1} - r_t|r_t)}$ without the aid of a smoothing technique.

The scatterplot visualization method loses its appeal when dealing with high-dimensional data. It is difficult to visualize data beyond two-dimensions. Further, because of the intrinsic difficulty called "curse of dimensionality" in analyzing high-dimensional data, high-dimensional regression surfaces typically are to be approximated by lower-dimensional functions and lower-dimensional smoothers serve as building blocks and are iteratively used for high-dimensional data analyses. On this behalf, the scatterplot visualization method is not useful because it can hardly be automated.

1.1.3 Local polynomial regression

While data presented in Figures 1.1 and 1.2 show different kinds of nonlinearity, they admit one common feature: the underlying curves to be estimated are smooth (except for a jump in Figure 1.2 (d)). What this means is that the underlying function can locally be approximated by a linear function. Let h be the size of the local neighborhood. Then for a given point x_0, we can locally model the data as

$$Y_i = a(x_0) + b(x_0)X_i + \text{error}, \text{ for } X_i \in x_0 \pm h,$$

where $a(x_0)$ and $b(x_0)$ are two local parameters. This naturally leads to the local regression problem: find $\widehat{a}(x_0)$ and $\widehat{b}(x_0)$ to minimize

$$\sum_{i=1}^{n} \{Y_i - a(x_0) - b(x_0)X_i\}^2 I\left(\frac{|X_i - x_0|}{h} \le 1\right). \tag{1.1}$$

The resulting estimator of the regression function at the point x_0 would naturally be $\widehat{m}(x_0) = \widehat{a}(x_0) + \widehat{b}(x_0)x_0$. As x_0 runs through a grid of points in a region of interest, we obtain the estimated curve.

An improved version of the above idea is to include a smooth weight function, which weighs down the contribution from an observation X_i away from x_0. Let K be a nonnegative unimodal function with support in $[-1, 1]$ so that (1.2) below uses only the data in the strip $x_0 \pm h$. The weighted local regression now becomes minimizing

$$\sum_{i=1}^{n} \{Y_i - a(x_0) - b(x_0)X_i\}^2 K\left(\frac{X_i - x_0}{h}\right). \tag{1.2}$$

When K is the uniform density function on $[-1, 1]$, (1.2) reduces to (1.1). In general, the weight function K does not have to be compactly supported, as long as it decays fast enough to eliminate the impact of remote data points.

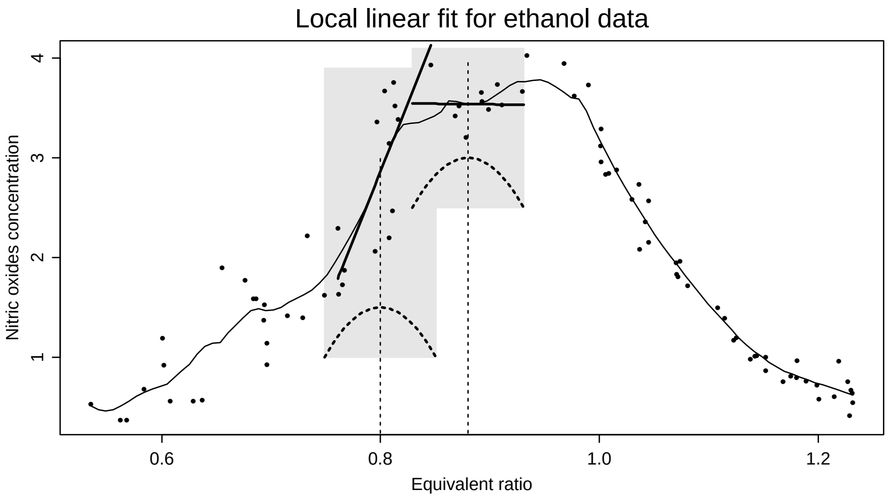

Figure 1.3: Illustration of the local linear fit. For each given point x_0, a line is fitted through the data contained in the strip $x_0 \pm 0.051$, using the weight function indicated at the bottom of the strip.

As an illustration of the above local linear regression approach, consider the ethanol data presented in Figure 1.1. The window size h is taken to be 0.051. To estimate the regression curve at the point $x_0 = 0.8$, we fit a line through the data in the neighborhood $x_0 \pm 0.051$ with a weight function $K(t) = (1 - t^2)_+$, where the subscript "+" denotes the positive part of the function so that K has support $[-1, 1]$, and obtain the estimated value $\widehat{m}(0.8) = 2.86$. See Figure 1.3. Suppose now that we wish to estimate the regression function at another point $x_0 = 0.88$, we then fit another line using the data in the slided window 0.88 ± 0.051. The whole curve is obtained by estimating the regression function in a grid of points. For example, the curve in Figure 1.3 was obtained by running 101 local linear regressions, associated with 101 grid points from 0.535 to 1.232.

To put the above basic idea in a more general statistical framework, let us assume that the observed data $\{(X_i, Y_i), i = 1, \cdots, n\}$ are a random sample from a certain population (X, Y) with conditional mean function $m(x) = E(Y|X = x)$ and conditional variance function $\sigma^2(x) = \mathrm{Var}(Y|X = x)$. For the ease of comprehensation, we write

$$Y = m(X) + \sigma(X)\varepsilon, \tag{1.3}$$

with $E(\varepsilon|X=x)=0,$ and $\mathrm{Var}(\varepsilon|X=x)=1$. For a given point x_0, if $m(x_0)$ has a p^{th} order derivative, then by Taylor's expansion,

$$\begin{aligned} m(x) &\approx m(x_0)+m'(x_0)(x-x_0)+\cdots+\frac{m^{(p)}(x_0)}{p!}(x-x_0)^p \\ &= \beta_0+\beta_1(x-x_0)+\cdots+\beta_p(x-x_0)^p, \end{aligned} \tag{1.4}$$

for x in a neighborhood of x_0, where $\beta_j=m^{(j)}(x_0)/j!$. Let h be the size of the neighborhood. Since the local model (1.4) is valid for $X_i \in x_0 \pm h$, we consider the following *local polynomial regression* problem: minimize with respect to $\beta_0,\cdots,\beta_p$

$$\sum_{i=1}^{n}\{Y_i-\beta_0-\beta_1(X_i-x_0)-\cdots-\beta_p(X_i-x_0)^p\}^2K\left(\frac{X_i-x_0}{h}\right). \tag{1.5}$$

Let $\widehat{\beta}_0,\cdots,\widehat{\beta}_p$ be the resulting fitted coefficients. Then, it is clear that $\widehat{\beta}_\nu$ is an estimator for $\beta_\nu=m^{(\nu)}(x_0)/\nu!$ or equivalently,

$$\widehat{m}_\nu(x_0)=\nu!\widehat{\beta}_\nu, \quad \nu=0,\cdots,p, \tag{1.6}$$

is an estimator for $m^{(\nu)}(x_0)$. Note that when $p=1$, the above local polynomial regression problem reduces to the *local linear regression* procedure and β_0 and β_1 are just a reparametrization of the local parameters a and b in (1.2). We prefer the parametrization in (1.4) because these parameters represent a multiple of the local derivatives.

1.1.4 Kernel function and bandwidth parameter

The function weight K in (1.5) is called a *kernel function*. In the regression context, it does not have to integrate up to one: any multiple of K results in the same weighting scheme. Commonly-used kernel functions include the

Gaussian kernel:	$K(t)=(\sqrt{2\pi})^{-1}\exp(-t^2/2)$
Symmetric Beta family:	$K(t)=\frac{1}{\mathrm{Beta}(1/2,\gamma+1)}(1-t^2)_+^\gamma, \quad \gamma=0,1,2,\cdots$

Note that the symmetric Beta kernels have support $[-1,1]$. When $\gamma=0,1,2$ and 3, the resulting kernels are called respectively the uniform, the Epanechnikov, the biweight and the triweight kernel function. Figure 1.4 depicts these kernel functions.

The parameter h in (1.5) is called a *bandwidth*, or a *smoothing parameter*. It can be vaguely understood as the size of a local neighborhood. However, this interpretation can be difficult when the kernel K has unbounded support such as for example a Gaussian kernel. The bandwidth parameter is confounded with the effective support of the kernel K. For example, using h_0 with a uniform kernel function is effectively about the same as using $2h_0$ with a triweight kernel function, because the effective support of the latter is about half as wide as that of the uniform kernel function. Because of this confounding effect, we take the kernel function K fixed when we discuss bandwidth selection. The

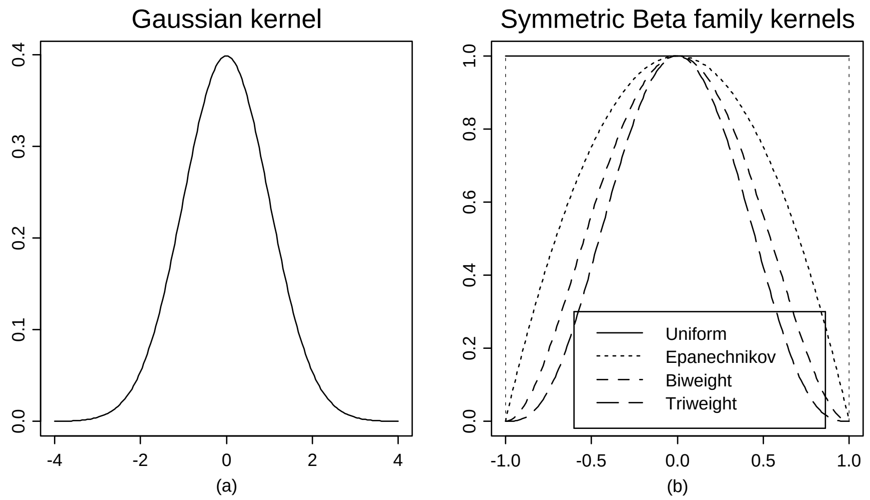

Figure 1.4: Commonly-used kernels. (a) Gaussian kernel; (b) Symmetric Beta family of kernels that are renormalized to have maximum height 1.

normalization of a kernel can be defined via the concept of *canonical kernel* as in Marron and Nolan (1988).

To demonstrate the effect of bandwidth selection, Figure 1.5 summarizes three estimated curves using local linear fits with the Epanechnikov kernel and bandwidths $h = 0$, 0.051 and ∞. When $h = 0$, in the local neighborhood of a design points X_i, there is only one data point in the neighborhood and the least-squares estimate (while degenerate) gives the fitted value to be the observed data. Hence, the resulting estimate is basically the linear interpolation. Such an estimate which appears wiggly is an example of an undersmoothed curve. When h increases, the size of the local neighborhood increases and so does the potential modeling bias. On the other hand, the number of local data points increases and this reduces the variance of the estimated coefficients. When $h = \infty$, the resulting local model becomes a global model and hence the resulting estimate is a straight line, which has large modeling bias for the ethanol data set. This kind of estimate, which ignores the features of the underlying curve, is called an oversmoothed estimate. As h ranges from 0 to ∞, we can obtain a family of estimated curves, ranging from the most complex curve (interpolation) to the simplest curve (line) and there are many interesting functions lying between these two extreme situations. Thus, the family of potential nonlinear functions produced by a local polynomial fit is rich.

A satisfactory estimated function should trade off the balance between the bias (requiring a small bandwidth) and the variance (requiring a large bandwidth). An example of this is the curve in Figure 1.5 with $h = 0.051$. There are two possible methods for choosing an appropriate amount of smoothing. A data-driven approach consists of letting the data choose a bandwidth that minimizes the risk of an estimated curve. Such an automatic bandwidth selection

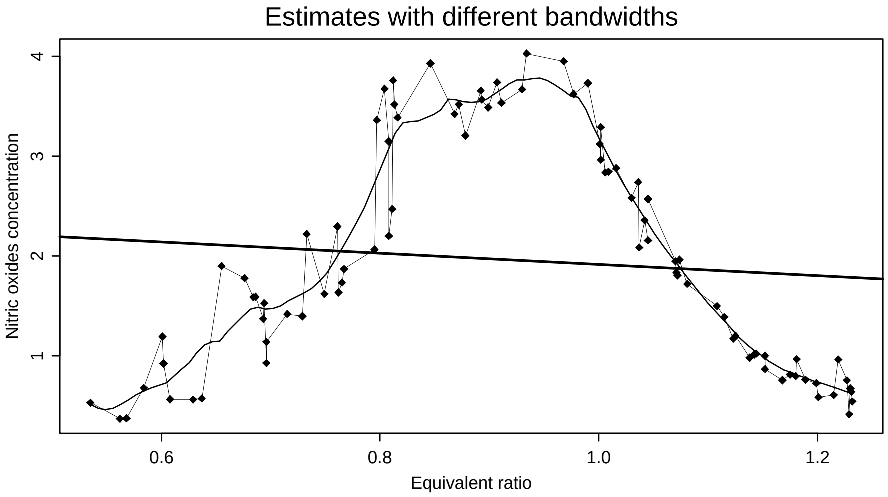

Figure 1.5: Local linear fit with bandwidth $h = 0$ (thin curve), 0.051 (solid curve) and ∞ (thick line). Different h gives different amount of smoothing. As h varies from 0 to ∞, the resulting estimate runs from the most complicated curve to the simplest curve.

method avoids subjectivity and is very useful for multivariate regression analysis. A subjective choice of the bandwidth can be thought in a similar matter, except that the trade-off now relies on the visualization. The drawback of this approach is that it can not be automated and can be difficult to use when the data are out of the domain of visualization as explained in Section 1.1.2.

Local modeling techniques provide parsimony to parametric modeling in the sense that when the bandwidth is large, the local model becomes a global model. Therefore, the boundary between parametric and nonparametric modeling is not clear particularly when one takes the bandwidth into consideration. The key difference is that in a parametric method, one insists using bandwidth $h = \infty$, while in nonparametric modeling, the global modeling is only a possible candidate.

1.2 Properties of local polynomial fitting

The basic idea of local polynomial fitting is very simple: in a neighborhood of x_0 a polynomial of degree p is fitted to the data, and the $p+1$ coefficients of the fit lead to estimates for $\beta_\nu = m^{(\nu)}(x_0)/\nu!$, $\nu = 0, \cdots, p$, where $m^{(\nu)}(x_0)$ denotes the ν^{th} derivative of the regression function $m(\cdot)$ evaluated at the point x_0. The neighborhood of the point x_0 is determined by two parameters, a symmetric kernel function K and a bandwidth parameter h.

In this section we summarize some properties of the local polynomial regression estimator $\widehat{m}_\nu(x_0)$ of $m^{(\nu)}(x_0)$, defined by (1.6). When the design density $f_X(\cdot)$ has a finite support, one needs to study separately the behavior of the local polynomial regression estimator for points x_0 in the interior of the support of the design density, and at the boundary of the support. It is well known that for kernel type estimators the behavior in the interior and at the boundary can be quite different. In the first two subsections we focus on estimation in the interior while the behavior at the boundary is discussed in the third subsection.

1.2.1 Bias and variance

Since model (1.3) involves the random design $X_1, \cdots, X_n$ one can consider conditional bias and variance expressions, given $X_1, \cdots, X_n$, or unconditional bias and variance. The question whether to study the conditional mean squared error $E\left[\{\widehat{m}_\nu(x_0) - m^{(\nu)}(x_0)\}^2|\mathcal{X}\right]$, with $\mathcal{X} = (X_1, \cdots, X_n)$, or the unconditional mean squared error $E\left[\{\widehat{m}_\nu(x_0) - m^{(\nu)}(x_0)\}^2\right]$ is of a phylosophical nature and relates to the question whether one focuses more on the behavior of the estimators for the design points at hand, or more on the general behavior of the estimators for all possible design realizations from the design density f_X. We do not enter into this discussion here, and emphasize on the conditional expressions for bias and variance, given $(X_1, \cdots, X_n)$, and complement this by making some remarks on unconditional results.

Let x_0 be a point at the interior of the support of the design density f_X. Recall that the estimators $\widehat{\beta}_\nu$, $\nu = 0, \cdots, p$, for $\beta_\nu = m^{(\nu)}(x_0)/\nu!$, $\nu = 0, \cdots, p$, are obtained from minimizing (1.5) with respect to $\beta_0, \cdots, \beta_p$. For the derivations which follow it is more convenient to use matrix notation. Denote the design matrix of the least-squares problem (1.5) by

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - x_0) & \cdots & (X_1 - x_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x_0) & \cdots & (X_n - x_0)^p \end{pmatrix},$$

and let

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \qquad \text{and} \qquad \widehat{\beta} = \begin{pmatrix} \widehat{\beta}_0 \\ \vdots \\ \widehat{\beta}_p \end{pmatrix}.$$

Further, denote by

$$\mathbf{W} = \text{diag}\{K_h(X_i - \mathbf{X}\beta)\}$$

the diagonal matrix containing the weights, where $K_h(\cdot) = K(\cdot/h)/h$ is a rescaled version of the kernel K. The weighted least-squares problem in (1.5) can then be expressed as

$$\min_\beta (\mathbf{y} - \mathbf{X}\beta)^T \mathbf{W} (\mathbf{y} - \mathbf{X}\beta),$$

where A^T denotes the transpose of a vector or a matrix A. The solution to this minimization problem follows from simple least-squares theory and is given by

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}.$$

The exact conditional bias and variance expressions for the estimated vector $\widehat{\beta}$ of $\beta = (\beta_0, \cdots, \beta_p)^T = (m(x_0), \cdots, m^{(p)}(x_0)/p!)^T$ are easily obtained from the above explicit expression for the estimator. Introducing the following notation:

$\mathbf{m} = (m(X_1), \cdots, m(X_n))^T$	$n \times 1$ vector
$\mathbf{r} = \mathbf{m} - \mathbf{X}\beta$	$n \times 1$ vector of residuals
$S_n = \mathbf{X}^T\mathbf{W}\mathbf{X}$	$(p+1) \times (p+1)$ matrix
$\Sigma = \mathrm{diag}\{K_h^2(X_i - x_0)\sigma^2(X_i)\}$	$n \times n$ diagonal matrix
$S_n^* = \mathbf{X}^T\Sigma\mathbf{X}$	$(p+1) \times (p+1)$ matrix,

the expressions for the *conditional bias* and *conditional variance* are as follows:

$$\begin{aligned} E(\widehat{\beta}|\mathcal{X}) &= S_n^{-1}\mathbf{X}^T\mathbf{W}\mathbf{m} = \beta + S_n^{-1}\mathbf{X}^T\mathbf{W}\mathbf{r} \\ \mathrm{Var}(\widehat{\beta}|\mathcal{X}) &= S_n^{-1}S_n^*S_n^{-1}. \end{aligned} \tag{2.1}$$

Seifert and Gasser (1996) investigated in detail the finite-sample variance of local polynomial regression estimators. They found that no upper bound can be given for the conditional variance, and that the unconditional variance can be infinite when a kernel with compact support is used. The difficulties apparently arise in regions where the design is sparse, and when the design realization has a small variance and/or a skew empirical distribution. Seifert and Gasser (1996) proposed two small-sample modifications of local polynomial regression estimators which they show have a good finite-sample behavior. When kernels with infinite support are used, such as a Gaussian kernel, then there is not such a problematic finite-sample behavior.

Asymptotic expressions for the conditional bias and variance will be useful for, for example, the bandwidth selection problem. We first derive an asymptotic expression for the conditional variance, which can be obtained by establishing asymptotic expressions for the matrices S_n and S_n^* appearing in the exact variance expression. Note that the matrix $S_n = (S_{n,j+\ell})_{0\le j,\ell\le p}$ where

$$S_{n,j} = \sum_{i=1}^{n} K_h(X_i - x_0)(X_i - x_0)^j, \tag{2.2}$$

and that the matrix S_n^* equals $\left(S_{n,j+\ell}^*\right)_{0\le j,\ell\le p}$ with

$$S_{n,j}^* = \sum_{i=1}^{n} K_h^2(X_i - x_0)(X_i - x_0)^j\sigma^2(X_i). \tag{2.3}$$

Therefore one of the tasks is to find asymptotic expressions for both type of elements $S_{n,j}$ and $S_{n,j}^*$. Before doing so we introduce some more notation:

$$\mu_j = \int u^j K(u)du \qquad \nu_j = \int u^j K^2(u)du$$

$$\begin{aligned} &S = (\mu_{j+\ell})_{0\le j,\ell\le p} && c_p = (\mu_{p+1}, \cdots, \mu_{2p+1})^T \\ &\widetilde{S} = (\mu_{j+\ell+1})_{0\le j,\ell\le p} && \widetilde{c}_p = (\mu_{p+2}, \cdots, \mu_{2p+2})^T \\ &S^* = (\nu_{j+\ell})_{0\le j,\ell\le p}\ . && \end{aligned}$$

For the quantity defined in (2.2) we find

$$\begin{aligned}
S_{n,j} &= ES_{n,j} + O_P\left\{\sqrt{\mathrm{Var}(S_{n,j})}\right\} \\
&= nh^j \int u^j K(u) f_X(x_0 + hu) du \\
&\quad + O_P\left(\sqrt{nE\{K_h^2(X_1 - x_0)(X_1 - x_0)^{2j}\}}\right) \\
&= nh^j\{f_X(x_0)\mu_j + o(1) + O_P\left((nh)^{-1/2}\right)\} \\
&= nh^j f_X(x_0)\mu_j\{1 + o_P(1)\}, \qquad (2.4)
\end{aligned}$$

where we assumed that $h \to 0$, $nh \to \infty$ and $f_X(\cdot)$ is continuous in a neigborhood of x_0. For the matrix S_n we then obtain

$$S_n = nf_X(x_0)HSH\{1 + o_P(1)\}, \qquad (2.5)$$

where $H = \mathrm{diag}(1, h, \cdots, h^p)$.

Using similar arguments the asymptotic behavior of S_n^* is derived. From definition (2.3) and expression (2.4) it is already clear that

$$S_{n,j}^* = nh^{j-1} f_X(x_0)\sigma^2(x_0)\nu_j\{1 + o_P(1)\},$$

provided that $\sigma^2(\cdot)$ is continuous in a neighborhood of x_0. This yields

$$S_n^* = nh^{-1} f_X(x_0)\sigma^2(x_0)HS^*H\{1 + o_P(1)\}. \qquad (2.6)$$

Expressions (2.5) and (2.6) lead to the following asymptotic expression for the conditional variance

$$\mathrm{Var}(\widehat{\beta}|\mathcal{X}) = \frac{\sigma^2(x_0)}{f_X(x_0)nh} H^{-1}S^{-1}S^*S^{-1}H^{-1}\{1 + o_P(1)\}, \qquad (2.7)$$

provided that $f_X(x_0) > 0$.

Assuming that $m^{(p+1)}(\cdot)$ is continuous in a neighborhood of x_0 and using Taylor expansion for the vector of residuals $\mathbf{r}$ the conditional bias can be expressed as

$$\begin{aligned}
S_n^{-1}\mathbf{X}^T\mathbf{W}\mathbf{r} &= S_n^{-1}\mathbf{X}^T\mathbf{W}\left[\beta_{p+1}(X_i - x_0)^{p+1} + o_P\{(X_i - x_0)^{p+1}\}\right]_{1 \le i \le n} \\
&= S_n^{-1}\left\{\beta_{p+1}c_n + o_P(nh^{p+1})\right\}, \qquad (2.8)
\end{aligned}$$

with $\beta_{p+1} = m^{(p+1)}(x_0)/(p+1)!$ and $c_n = (S_{n,p+1}, \cdots, S_{n,2p+1})^T$. Using (2.4) and (2.5) this leads to the asymptotic expression for the conditional bias

$$\mathrm{Bias}\left(\widehat{\beta}|\mathcal{X}\right) = H^{-1}S^{-1}c_p\beta_{p+1}h^{p+1}\{1 + o_P(1)\}. \qquad (2.9)$$

The asymptotic conditional variance of $\widehat{\beta}_\nu = \widehat{m}_\nu(x_0)/\nu!$ is obtained by taking the $(\nu+1)^{\mathrm{th}}$ diagonal element of the right-hand side of (2.7). The asymptotic conditional bias of $\widehat{m}_\nu(x_0)/\nu!$ is given by the $(\nu+1)^{\text{th}}$ element of the vector provided in the right-hand side of (2.9). Here one needs to be careful however. When K is symmetric, the $(\nu+1)^{\text{th}}$ element of $S^{-1}c_p$ is zero in case that $p-\nu$ is even, and therefore the leading bias term is zero. This is valid, but not informative. In order to find the asymptotic conditional bias for $\widehat{\beta}_\nu$ when $p-\nu$ is even, it suffices to use Taylor expansions up to a higher order in the above derivation for the asymptotic bias. A higher order Taylor expansion will of course require more stringent conditions on the functions $f_X(\cdot)$ and $m(\cdot)$ and on the bandwidth h. We do not elaborate further on this point. More details can be found in Fan and Gijbels (1996).

The asymptotic expressions for conditional bias and variance of $\widehat{m}_\nu(x_0)$ are summarized in the following theorem, which was obtained by Fan (1992) and Ruppert and Wand (1994). For simplicity of the statement, we assume that the kernel function is symmetric, having a bounded support.

Theorem 1. Assume that $f_X(x_0) > 0$ and that $f_X(\cdot)$, $m^{(p+1)}(\cdot)$ and $\sigma^2(\cdot)$ are continuous in a neighborhood of x_0. If $h \to 0$ such that $nh \to \infty$ then the asymptotic conditional variance of $\widehat{m}_\nu(x_0)$ is given by

$$\mathrm{Var}\left\{\widehat{m}_\nu(x_0)|\mathcal{X}\right\} = e_{\nu+1}^T S^{-1} S^* S^{-1} e_{\nu+1} \frac{(\nu!)^2\sigma^2(x_0)}{f_X(x_0) n h^{1+2\nu}} + o_P\left(n^{-1}h^{-(1+2\nu)}\right), \tag{2.10}$$

where $e_{\nu+1} = (0,0,\cdots,0,1,0,\ldots,0)^T$ denotes the unit vector of dimension $p+1$ with 1 on position $\nu+1$.

For $p-\nu$ odd the asymptotic conditional bias of $\widehat{m}_\nu(x_0)$ is

$$\mathrm{Bias}\left\{\widehat{m}_\nu(x_0)|\mathcal{X}\right\} = e_{\nu+1}^T S^{-1} c_p \frac{\nu!}{(p+1)!} m^{(p+1)}(x_0) h^{p+1-\nu} + o_P\left(h^{p+1-\nu}\right). \tag{2.11}$$

For $p-\nu$ even, and under the additional assumptions that $f_X'(\cdot)$ and $m^{(p+2)}(\cdot)$ are continuous in a neighborhood of x_0 and $nh^3 \to \infty$, the asymptotic conditional bias is given by

$$\begin{aligned}\mathrm{Bias}\left\{\widehat{m}_\nu(x_0)|\mathcal{X}\right\} \quad = \quad & e_{\nu+1}^T S^{-1} \widetilde{c}_p \frac{\nu!}{(p+2)!} \Big\{ m^{(p+2)}(x_0) \\ & + (p+2) m^{(p+1)}(x_0) \frac{f_X'(x_0)}{f_X(x_0)} \Big\} h^{p+2-\nu} + o_P\left(h^{p+2-\nu}\right).\end{aligned} \tag{2.12}$$

Further it can be shown under certain conditions that the estimator $\widehat{m}_\nu(x_0)$ is asymptotically normal:

$$\left\{\mathrm{AV}_{\nu,p}(x_0)\right\}^{-1/2} \left\{\widehat{m}_\nu(x_0) - m^{(\nu)}(x_0) - \mathrm{Ab}_{\nu,p}(x_0)\right\} \xrightarrow{D} \mathrm{N}(0;1), \tag{2.13}$$

where $\mathrm{AV}_{\nu,p}(x_0)$ and $\mathrm{Ab}_{\nu,p}(x_0)$ are the leading terms in respectively the asymptotic conditional variance and bias given above.

1.2.2 Equivalent kernels

The estimator $\widehat{\beta}_\nu$ of $m^{(\nu)}(x_0)/\nu!$ can be rewritten as a type of kernel estimator as follows

$$\widehat{\beta}_\nu = e_{\nu+1}^T \widehat{\beta} = e_{\nu+1}^T S_n^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y} = \sum_{i=1}^n W_\nu^n \left(\frac{X_i - x_0}{h} \right) Y_i, \tag{2.14}$$

with $W_\nu^n(t) = e_{\nu+1}^T S_n^{-1} \{1, th, \cdots, (th)^p\}^T K(t)/h$. Note that the definition of this weight function $W_\nu^n(\cdot)$ depends on the design points and on x_0. This dependence reflects the fact that local polynomial regression estimators adapt automatically to the design and to the location x_0 (either in the interior or near the boundary of the support).

It is easily seen that the kernel weights W_ν^n satisfy some discrete moment conditions. Indeed, for $0 \le \nu, q \le p$, we find

$$\begin{aligned}
&\sum_{i=1}^n (X_i - x_0)^q W_\nu^n \left(\frac{X_i - x_0}{h} \right) \\
&= e_{\nu+1}^T S_n^{-1} \sum_{i=1}^n (X_i - x_0)^q \begin{pmatrix} 1 \\ X_i - x_0 \\ \vdots \\ (X_i - x_0)^p \end{pmatrix} K_h(X_i - x_0) \\
&= e_{\nu+1}^T S_n^{-1} S_n e_{q+1} = \delta_{\nu,q}.
\end{aligned}$$

Consequently the finite-sample bias is zero when using local polynomial fitting of degree p to estimate polynomials of order up to p.

The above representation of the local polynomial regression estimator as a conventional kernel estimator has an asymptotic counterpart. This can easily be found by substituting the asymptotic expression (2.5) for the matrix S_n into the definition of $W_\nu^n(t)$, yielding an asymptotic expression for the kernel function in (2.14)

$$W_\nu^n(t) = \frac{1}{nh^{\nu+1} f_X(x_0)} e_{\nu+1}^T S^{-1} (1, t, \cdots, t^p)^T K(t) \{1 + o_P(1)\},$$

which in turn leads to the following asymptotic expression for the estimator

$$\widehat{\beta}_\nu = \frac{1}{nh^{\nu+1} f_X(x_0)} \sum_{i=1}^n K_\nu^* \left(\frac{X_i - x_0}{h} \right) Y_i \{1 + o_P(1)\},$$

where we denoted

$$K_\nu^*(t) = e_{\nu+1}^T S^{-1} (1, t, \cdots, t^p)^T K(t), \tag{2.15}$$

which is called the *equivalent kernel.* It is easily seen, using similar arguments as before, that this equivalent kernel satisfies the moment conditions

$$\int u^q K_\nu^*(u) du = \delta_{\nu,q} \qquad 0 \le \nu, q \le p,$$

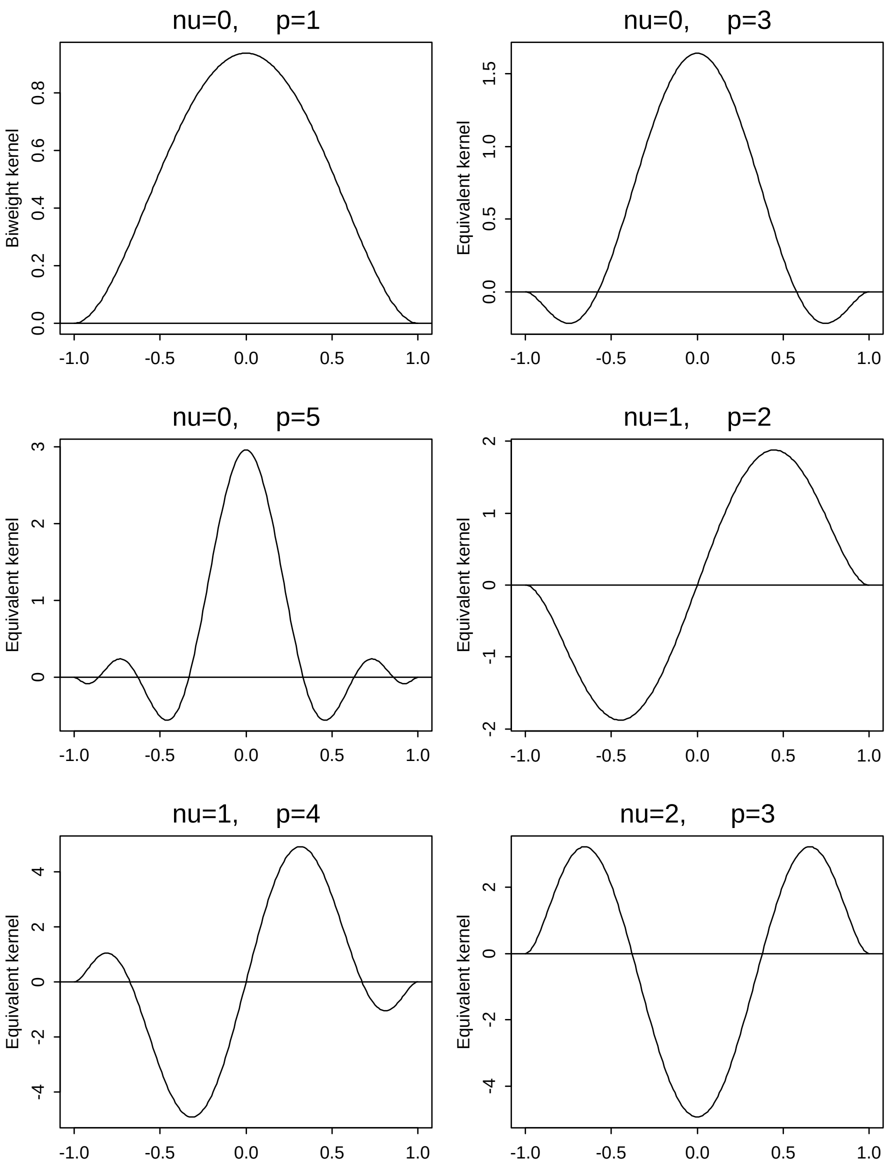

Figure 1.6: Biweight kernel and corresponding equivalent kernels for some values of p and ν.

which is the asymptotic counterpart of the discrete moment conditions on $W_\nu^n(\cdot)$ discussed above. A picture of the biweight kernel and some corresponding equivalent kernels is provided in Figure 1.6. See Fan and Gijbels (1996) for more details on equivalent kernels and their relationship with other kernels discussed in the literature.

The asymptotic conditional bias and variance of the local polynomial regression estimator provided in Theorem 1 can be rewritten in terms of the equivalent kernel K_ν^*. For the asymptotic conditional variance this gives

$$\operatorname{Var}\{\widehat{m}_\nu(x_0)|\mathcal{X}\} = \int K_\nu^{*2}(t)dt \frac{(\nu!)^2\sigma^2(x_0)}{f_X(x_0)nh^{1+2\nu}} + o_P\left(n^{-1}h^{-(1+2\nu)}\right). \quad (2.16)$$

The asymptotic conditional bias of $\widehat{m}_\nu(x_0)$ can be re-expressed as

$$\operatorname{Bias}\{\widehat{m}_\nu(x_0)|\mathcal{X}\} = \left\{\int t^{p+1}K_\nu^*(t)dt\right\}\frac{\nu!}{(p+1)!}m^{(p+1)}(x_0)h^{p+1-\nu} + o_P\left(h^{p+1-\nu}\right), \quad (2.17)$$

for $p-\nu$ odd and as

$$\begin{aligned}\operatorname{Bias}\{\widehat{m}_\nu(x_0)|\mathcal{X}\} &= \left\{\int t^{p+2}K_\nu^*(t)dt\right\}\frac{\nu!}{(p+2)!}\left\{m^{(p+2)}(x_0)\right.\\ &\quad \left. +(p+2)m^{(p+1)}(x_0)\frac{f_X'(x_0)}{f_X(x_0)}\right\}h^{p+2-\nu} + o_P\left(h^{p+2-\nu}\right),\end{aligned} \quad (2.18)$$

for $p-\nu$ even.

Note that the leading term in the asymptotic conditional bias has a more complex structure for $p-\nu$ even. Furthermore, the term depends on the design density $f_X(x_0)$ and its derivative function $f_X'(x_0)$. This rather complex structure of the leading term is unpleasant to work with when, for example, doing bandwidth selection based on L_2-type measures of performance. See also Section 1.3.2.

As an illustration we look at four special cases: local constant ($p=0$) and local linear ($p=1$) fitting for estimating the regression function ($\nu=0$) and local linear ($p=1$) and local quadratic ($p=2$) fitting for estimating the first derivative of the regression function ($\nu=1$). Table 2.1 lists, in the second column, the expression for the equivalent kernel for each of these cases. In the last column the equivalent kernels for boundary points are provided. These will be discussed in the next subsection, where one can also find the expressions for the $\mu_{j,c}$'s (see (2.19)). Table 2.2 presents the leading terms in the asymptotic conditional bias and variance.

Table 2.1: *Some examples of equivalent kernels.*

ν	p	$K_\nu^*(t)$	$K_{\nu,c}^*(t)$
0	0	$K(t)$	$\mu_{0,c}^{-1}K(t)$
0	1	$K(t)$	$(\mu_{2,c}-\mu_{1,c}t)K(t)/(\mu_{0,c}\mu_{2,c}-\mu_{1,c}^2)$
1	1	$\mu_2^{-1}tK(t)$	$(-\mu_{1,c}+\mu_{0,c}t)K(t)/(\mu_{0,c}\mu_{2,c}-\mu_{1,c}^2)$
1	2	$\mu_2^{-1}tK(t)$	a

$$a = \frac{-\mu_{1,c}\mu_{4,c}+\mu_{2,c}\mu_{3,c}+t(\mu_{0,c}\mu_{4,c}-\mu_{2,c}^2)-t^2(\mu_{0,c}\mu_{3,c}-\mu_{1,c}\mu_{2,c})}{(\mu_{0,c}\mu_{2,c}-\mu_{1,c}^2)\mu_{4,c}-\mu_{2,c}^3-\mu_{0,c}\mu_{3,c}+2\mu_{1,c}\mu_{2,c}\mu_{3,c}}.$$

Table 2.2: *Leading terms in the asymptotic conditional bias and variance.*

ν	p	leading term in bias	leading term in variance
0	0	$\mu_2 \frac{1}{2} \left\{ m''(x_0) + 2m'(x_0) \frac{f'_X(x_0)}{f_X(x_0)} \right\} h^2$	$\nu_0 \frac{\sigma^2(x_0)}{f_X(x_0)nh} \equiv V_0$
0	1	$\mu_2 \frac{1}{2} m''(x_0) h^2$	V_0
1	1	$\mu_2^{-1} \mu_4 \frac{1}{6} \left\{ m'''(x_0) + 3m''(x_0) \frac{f'_X(x_0)}{f_X(x_0)} \right\} h^2$	$\mu_2^{-2} \nu_2 \frac{\sigma^2(x_0)}{f_X(x_0)nh^3} \equiv V_1$
1	2	$\mu_2^{-1} \mu_4 \frac{1}{6} m'''(x_0) h^2$	V_1

We close this section by remarking that local polynomial fits possess various optimality properties in the decision theoretical sense. See Stone (1982), Fan (1992) and Sections 3.4 and 3.5 of Fan and Gijbels (1996) for discussions.

1.2.3 Estimation at the boundary

In this section we briefly discuss the behavior of the local polynomial regression estimator $\widehat{m}_\nu(x_0)$ for x_0 at the boundary of the support of the design density $f_X(\cdot)$. Assume for simplicity that this support is $[0, 1]$ and that one wants to estimate the regression function or its derivatives at left *boundary points* of the form $x_0 = ch$ or at right boundary points $x_0 = 1 - ch$, with $c > 0$. If the kernel function K has support $[-1, 1]$, then real boundary points are those for which $c < 1$, whereas for $c > 1$ we have interior points.

The asymptotic conditional bias and variance expressions for the estimator when $x_0 = 1 - ch$ is a right boundary point, are derived in an analogous way to that for interior points. The only essential difference is in the asymptotic moment calculations. This results for example in a different asymptotic behavior of the finite-sample moments $S_{n,j}$, defined in (2.2). For these moments we now obtain

$$S_{n,j} = nh^j f_X(1-) \mu_{j,c} \{1 + o_P(1)\}, \tag{2.19}$$

with $\mu_{j,c} = \int_{-\infty}^{c} u^j K(u) du$, a truncated version of the moment μ_j. Compare with (2.4). This difference in the asymptotic behavior of $S_{n,j}$ for boundary points carries through to the definition of the equivalent kernel which is now given by

$$K^*_{\nu,c}(t) = e^T_{\nu+1} S_c^{-1} (1, t, \cdots, t^p)^T K(t), \tag{2.20}$$

with $S_c = (\mu_{j+\ell,c})_{0 \le j,\ell \le p}$. Note that the only difference with the definition of the equivalent kernel for interior points is the replacement of the matrix S by the matrix S_c.

The asymptotic conditional variance for a right boundary point $x_0 = 1 - ch$ is given by

$$\mathrm{Var}\left\{\widehat{m}_\nu(x_0) | \mathcal{X}\right\} = \int_{-\infty}^{c} K^{*2}_{\nu,c}(t) dt \frac{(\nu!)^2 \sigma^2(1-)}{f_X(1-) nh^{1+2\nu}} + o_P\left(n^{-1} h^{-(1+2\nu)}\right), \tag{2.21}$$

and the asymptotic conditional bias

$$\begin{aligned} \mathrm{Bias}\left\{\widehat{m}_\nu(x_0) | \mathcal{X}\right\} &= \left\{ \int_{-\infty}^{c} t^{p+1} K^*_{\nu,c}(t) dt \right\} \frac{\nu!}{(p+1)!} m^{(p+1)}(1-) h^{p+1-\nu} \\ &\quad + o_P\left(h^{p+1-\nu}\right), \end{aligned} \tag{2.22}$$

is informative now for both $p-\nu$ odd and even, since now the leading term in (2.22) is not necessarily zero for $p-\nu$ even.

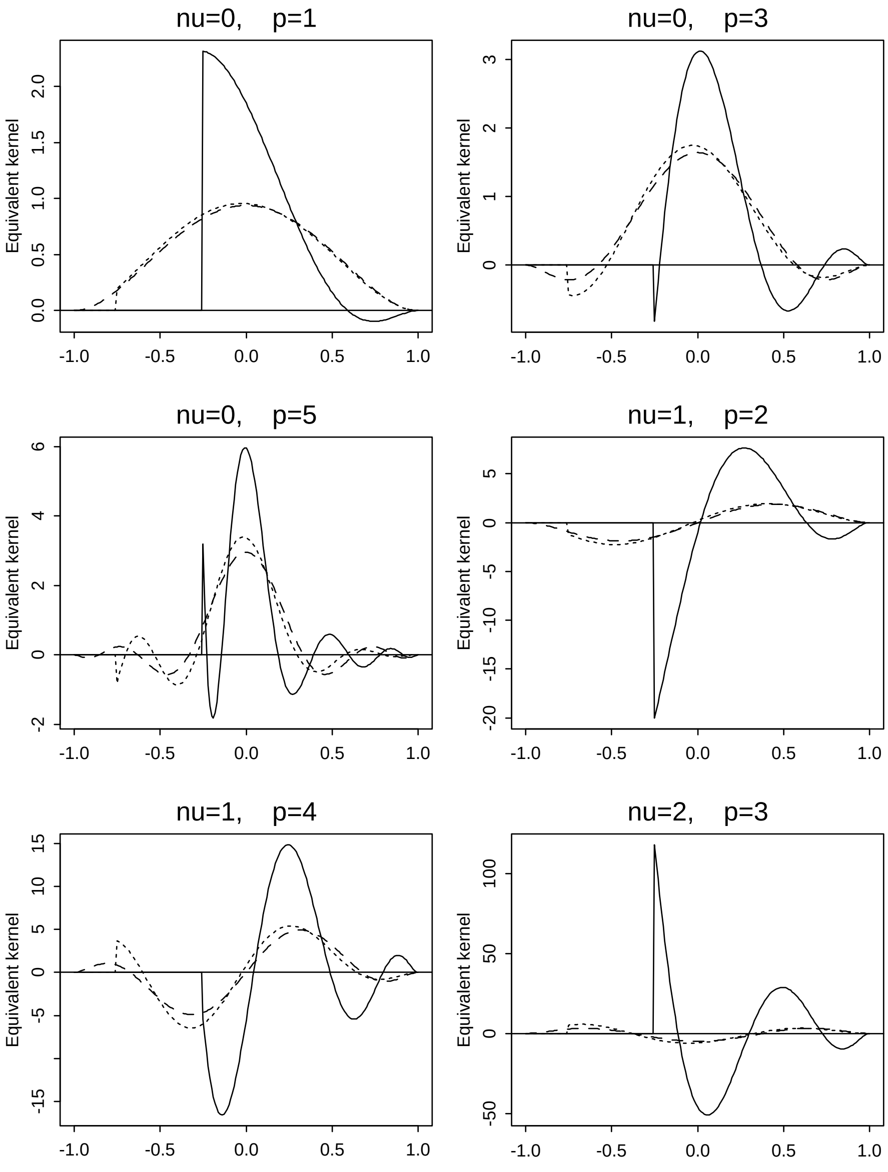

Figure 1.7: Biweight kernel and corresponding equivalent kernels for right boundary points $x_0 = 1 - ch$ with $c = 0.25$ and $c = 0.75$ and for interior points $c \geq 1$.

For left boundary points $x_0 = ch$ similar derivations can be done. Now the truncated moments $\mu_{j,c} = \int_{-c}^{+\infty} u^j K(u) du$ appear in expression (2.19) and with these truncated moments the equivalent kernel is defined as in (2.20). The asympotic conditional bias and variance expressions are similar to these provided in (2.21) and (2.22) but with the appropriate definition of the equivalent kernel, and with the right-hand limits $m^{(p+1)}(0+)$, $\sigma^2(0+)$ and $f_X(0+)$ instead of the left-hand limits at the point 1.

Table 2.1 provides expressions for equivalent kernels at the boundary for some values of p and ν. Figure 1.7 presents the biweight kernel and some of the corresponding equivalent kernels at two right boundary points $x_0 = 1 - ch$, with $c = 0.25$ and 0.75 and at interior points ($c \geq 1$).

Note that for $p - \nu$ odd the asymptotic conditional bias and variance are of the same order at the boundary and in the interior. Compare expressions (2.16) and (2.17) with expressions (2.21) and (2.22) respectively. This implies that, for $p - \nu$ odd, local polynomial regression estimators have the same L_2-rate of convergence at the boundary and in the interior. In this sense this estimation procedure for $p - \nu$ odd adapts automatically to the boundary of the support of the design density. Note however that the constant factors, determined by the kernel function K, p and ν, in the leading terms of the asymptotic conditional bias and variance are different for the boundary region than for the interior, and so there are certain boundary effects which will be noticed. A discussion on these constant factors can be found in Fan and Gijbels (1992).

For $p - \nu$ even the rate of convergence of the local polynomial regression estimator is different at the boundary and in the interior. Compare expressions (2.18) and (2.22). For a local constant fit for example the conditional bias term is of order h^2 for points in the interior and is of order h for points at the boundary.

1.3 Choice of bandwidth

Local polynomial fitting requires a choice for the bandwidth parameter h, for the degree p of the polynomial and for the kernel function K. It is well known that the choice of the kernel function is of less importance in kernel smoothing. This holds true for local polynomial fitting. For a thorough discussion on optimal choices for the kernel function see Fan and Gijbels (1996). We discuss only the choice of the bandwidth and the degree of the polynomial p. This section is devoted to the problem of bandwidth selection.

1.3.1 Different types of bandwidths

As depicted already in the introduction, the choice of the bandwidth parameter h can have a big influence on the final smoothing result, and hence is quite important. One can disinguish between different types of bandwidths.

The bandwidth h could be chosen to be the same constant all over the range of x in $m(x)$. Such a bandwidth is called a *constant* or also *global bandwidth*, and is a reasonable choice when the unknown regression function shows approximately the same smoothness all over the estimation range.

When the regression function shows a different behavior in different regions, for example is nearly linear in one region and changes rather rapidly in some other region, then it may be advisable to use different bandwidths in these different regions. In the latter case the bandwidth h changes with the location x and is denoted by $h(x)$. This bandwidth is referred to as *local variable bandwidth*. A bandwidth can also change with the data point X_i, resulting in $h(X_i)$, which is called a *global variable bandwidth*.

1.3.2 Optimal bandwidths

When talking about *optimal bandwidths* one first needs to specify the criterion which is used for evaluating the performance of an estimator. Such an evaluation can be based on any appropriate distance function between the unknown regression function and the estimated function. Examples include the L_1, L_2 and L_∞ distances as well as the Hellinger distance. For mathematical convenience one often focuses on the L_2-loss function. A measure of local performance of the estimator is then the conditional mean squared error

$$\begin{aligned} \mathrm{MSE}_{\nu,p}(x_0;h) &= E\left[\{\widehat{m}_\nu(x_0) - m^{(\nu)}(x_0)\}^2|\mathcal{X}\right] \\ &= \left[\mathrm{Bias}\left\{\widehat{m}_\nu(x_0)|\mathcal{X}\right\}\right]^2 + \mathrm{Var}\left\{\widehat{m}_\nu(x_0)|\mathcal{X}\right\}, \end{aligned} \tag{3.1}$$

and a measure of global performance is the conditional mean integrated squared error

$$\begin{aligned} \mathrm{MISE}_{\nu,p}(h) &= \int E\left[\{\widehat{m}_\nu(x) - m^{(\nu)}(x)\}^2|\mathcal{X}\right] w(x)dx \\ &= \int \left[\mathrm{Bias}\left\{\widehat{m}_\nu(x)|\mathcal{X}\right\}\right]^2 w(x)dx + \int \mathrm{Var}\left\{\widehat{m}_\nu(x)|\mathcal{X}\right\} w(x)dx, \end{aligned} \tag{3.2}$$

with $w(\cdot)$ a positive weight function.

The ideal optimal global bandwidth is the one minimizing the conditional mean integrated squared error:

$$h^o_{\nu,p} = \mathrm{argmin}_h \mathrm{MISE}_{\nu,p}(h). \tag{3.3}$$

The ideal optimal local variable bandwidth is the bandwidth minimizing the conditional mean squared error:

$$h^o_{\nu,p}(x_0) = \mathrm{argmin}_h \mathrm{MSE}_{\nu,p}(x_0;h). \tag{3.4}$$

Approximations of these ideal optimal bandwidths are obtained by minimizing the asymptotic expressions for the MSE and the MISE, denoted by AMSE and AMISE respectively, which follow directly from the asymptotic expressions for bias and variance provided in Section 1.2.1. Since the case $p-\nu$ even leads to more complex expressions for the asymptotic conditional bias it is convenient to focus on the case $p-\nu$ odd. In Section 1.4 we discuss in somewhat more detail the choice between $p-\nu$ odd or even, and argue in favor of the case $p-\nu$ odd.

Minimization of the asymptotic conditional mean squared error for $p-\nu$ odd leads to the theoretical optimal local variable bandwidth given by

$$\begin{aligned} h_{\nu,p}^{\text{opt}}(x_0) &= \text{argmin}_h \text{AMSE}_{\nu,p}(x_0;h) \\ &= C_{\nu,p}(K)\left[\frac{\sigma^2(x_0)}{\{m^{(p+1)}(x_0)\}^2 f_X(x_0)}\right]^{1/(2p+3)} n^{-1/(2p+3)}, \quad (3.5) \end{aligned}$$

where the constant factor $C_{\nu,p}(K)$ depends only on K, p and ν, and is given by

$$C_{\nu,p}(K) = \left[\frac{\{(p+1)!\}^2(2\nu+1)\int K_\nu^{*2}(t)dt}{2(p+1-\nu)\{\int t^{p+1}K_\nu^*(t)dt\}^2}\right]^{1/(2p+3)}. \quad (3.6)$$

These constant factors are easy to calculate and are tabulated for certain choices of K, p and ν in Table 3.2 of Fan and Gijbels (1996).

The theoretical optimal constant or global bandwidth is obtained by minimizing the asymptotic conditional mean integrated squared error, and is given by

$$\begin{aligned} h_{\nu,p}^{\text{opt}} &= \text{argmin}_h \text{AMISE}_{\nu,p}(h) \\ &= C_{\nu,p}(K)\left[\frac{\int \sigma^2(x)w(x)/f_X(x)dx}{\int\{m^{(p+1)}(x)\}^2 w(x)dx}\right]^{1/(2p+3)} n^{-1/(2p+3)}. \quad (3.7) \end{aligned}$$

These theoretical optimal bandwidths can not be directly used in practice since they depend on unknown quantities $m^{(p+1)}(\cdot)$, $\sigma^2(\cdot)$ and $f_X(\cdot)$, and hence can not be calculated. In the next section we briefly discuss some practical data-driven bandwidth choices, which consistently estimate $h_{\nu,p}^{\text{opt}}$ or $h_{\nu,p}^{\text{opt}}(x_0)$.

1.3.3 Data-driven bandwidths

Rule of thumb global bandwidth selector

A simple data-driven global bandwidth selector is obtained by using (3.7) and considering a reference parametric model for the unknown regression function $m(\cdot)$. One can for example take $m(\cdot)$ to be a polynomial. For the degree of the polynomial one could, for example, take $p+3$ which would imply that $m^{(p+1)}(\cdot)$ is a quadratic function, leaving a certain flexibility for fitting functions with a curvature. More precisely, consider the parametric model

$$m(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_{p+3}x^{p+3}, \quad (3.8)$$

and take $w(\cdot)$ in (3.7) equal to $f_X(\cdot)w_0(\cdot)$ for a given positive weight function $w_0(\cdot)$. Fit the parametric model (3.8) to the data and obtain the fit

$$\check{m}(x) = \check{\alpha}_0 + \check{\alpha}_1 x + \cdots + \check{\alpha}_{p+3}x^{p+3},$$

as well as $\check{\sigma}^2$, the standardized residual sum of squares from this parametric fit. Substituting the estimated quantities

$$\check{m}^{(p+1)}(x) = (p+1)!\check{\alpha}_{p+1} + (p+2)!\,\check{\alpha}_{p+2}x + \frac{(p+3)!}{2}\,\check{\alpha}_{p+3}x^2,$$

and $\check{\sigma}^2$ into (3.7) leads to the estimated expression

$$C_{\nu,p}(K)\left[\frac{\check{\sigma}^2\int w_0(x)dx}{n\int\{\check{m}^{(p+1)}(x)\}^2 w_0(x)f_X(x)dx}\right]^{1/(2p+3)}.$$

Note that the denominator can be written as $nE[\{\check{m}^{(p+1)}(X)\}^2 w_0(X)]$ which can be estimated by $\sum_{i=1}^n\{\check{m}^{(p+1)}(X_i)\}^2 w_0(X_i)$. Hence an estimator for $h_{\nu,p}^{\text{opt}}$ in (3.7) is given by

$$\check{h}_{\text{ROT}} = C_{\nu,p}(K)\left[\frac{\check{\sigma}^2\int w_0(x)dx}{\sum_{i=1}^n\{\check{m}^{(p+1)}(X_i)\}^2 w_0(X_i)}\right]^{1/(2p+3)}. \tag{3.9}$$

The above data-driven bandwidth selector is of the same nature as the rule of thumb bandwidth selector in density estimation using a normal density as a reference (see for example Silverman (1986)). It is a rather crude bandwidth selector, but it gives a first idea of an appropriate magnitude for the bandwidth parameter. We will illustrate the use of this bandwidth selector later on in Section 1.4.

Plug-in global bandwidth selector

The theoretical global bandwidth in (3.7) contains two unknown quantities, namely $\int\sigma^2(x)w(x)/f_X(x)dx$ and $\int\{m^{(p+1)}(x)\}^2 w(x)dx$, and hence the basic task is to estimate these unknown quantities. Ruppert, Sheather and Wand (1995) propose nonparametric estimators for these quantities and substitute them into (3.7) leading to direct plug-in type of bandwidth selectors. They focus on local linear regression estimation ($p=1$ and $\nu=0$) under homoscedastic errors ($\sigma^2(x)=\sigma^2$ for all x). We briefly discuss their procedure.

Taking $w(x)=f_X(\cdot)I\{x\in[a,b]\}$, with $[a,b]\subset I\!R$ the support of $f_X(\cdot)$, the theoretical optimal global bandwidth is given by

$$h_{0,1}^{\text{opt}} = C_{0,1}(K)\left[\frac{\sigma^2(b-a)}{n\int_a^b\{m''(x)\}^2 f_X(x)dx}\right]^{1/5}, \tag{3.10}$$

where the unknown quantities to be estimated are

$$\theta_{22}\equiv\int_a^b\{m''(x)\}^2 f_X(x)dx \qquad \text{and} \qquad \sigma^2.$$

Note that $\theta_{22}=E[\{m''(X)\}\}^2 I\{X\in[a,b]\}]$ and hence an appropriate estimator is

$$\widehat{\theta}_{22}=\frac{1}{n}\sum_{i=1}^n\{\widehat{m}_2(X_i;g)\}\}^2\, I\{X_i\in[a,b]\},$$

where $\widehat{m}_2(\cdot;g)$ is obtained from a local cubic fit using a pilot bandwidth g. Ruppert, Sheather and Wand (1995) derive the theoretical optimal choice for the pilot bandwidth g, which is obtained from minimizing the asymptotic mean squared error. They find that

$$g_{\mathrm{AMSE}} \simeq C(K)\left[\frac{\sigma^2(b-a)}{n|\theta_{24}|}\right]^{1/7}, \tag{3.11}$$

where $\theta_{24} = \int m''(x)m^{(4)}(x)dx$ and where the definition of the constant factor $C(K)$ depends on the sign of the quantity θ_{24}, but its definitions involve only the kernel function K. Note that the theoretical optimal bandwidth g_{AMSE} is of order $n^{-1/7}$, compared to the order $n^{-1/5}$ for the theoretical optimal bandwidth $h_{0,1}^{\mathrm{opt}}$, and involves the unknown quantity θ_{24}.

The unknown variance σ^2 is estimated by

$$\widehat{\sigma}^2 = (\mathrm{DF})^{-1}\sum_{i=1}^{n}\{Y_i - \widehat{m}(X_i;\lambda)\}^2, \tag{3.12}$$

where DF is the degrees of freedom (see Ruppert, Sheather and Wand (1995) for the exact expression), and $\widehat{m}(\cdot;\lambda)$ is a local linear regression estimator using a pilot bandwidth λ. The theoretical optimal value for λ can be shown to be

$$\lambda_{\mathrm{AMSE}} = C^*(K)\left[\frac{\sigma^4(b-a)}{n^2\theta_{22}^2}\right]^{1/9}, \tag{3.13}$$

where $C^*(K)$ is a constant factor, depending only on the kernel function K.

Since both g_{AMSE} and λ_{AMSE} depend on unknown quantities themselves, one needs to think of some strategies to deal with this. Ruppert, Sheather and Wand (1995) propose three plug-in bandwidth selection strategies. For brevity we only mention one of them:

1. Divide the sample $X_1, \cdots, X_n$ into N subsamples, and obtain least-squares quartic fits of $m(\cdot)$ and σ^2 using the subsamples. Use these least-squares quartic fits to get preliminary estimators for θ_{24} and σ^2;

2. Use the preliminary estimators from Step 1 to estimate g_{AMSE} and θ_{22};

3. Estimate λ_{AMSE} by using the preliminary estimator for σ^2 from Step 1 and the estimator for θ_{22} from Step 2. Obtain the final estimator of σ^2 from (3.12);

4. Substitute the estimators for θ_{22} and σ^2 into (3.10) and obtain the estimator for $h_{0,1}^{\mathrm{opt}}$.

For a detailed description of this and other plug-in global bandwidth selection procedures see Ruppert, Sheather and Wand (1995).

Residual squares and pre-asymptotic substitution bandwidth selectors

The rule of thumb bandwidth selector and the plug-in bandwidth selection procedures are *post-asymptotic substitution methods* since they both focus on the theoretical optimal global bandwidth (3.7) which minimizes the asymptotic conditional mean integrated squared error. This theoretical optimal bandwidth is a first order approximation of the ideal optimal global bandwidth (3.3) which minimizes the finite-sample conditional mean integrated squared error.

An attempt to stay closer to the finite-sample level, and hence to avoid going to asymptotics immediately, consists of trying to estimate the mean squared error and minimize this estimated mean squared error or an integrated version of it. This would result in bandwidth selectors which rely less on asymptotics and hopefully would overall perform better. The bandwidth selectors of this subsection are based on this idea and are called *pre-asymptotic substitution bandwidth selectors*.

We first describe how to obtain an estimator $\widehat{\mathrm{MSE}}_{\nu,p}(x_0;h)$ of $\mathrm{MSE}_{\nu,p}(x_0;h)$. We only outline the estimation procedure. For more details we refer to Fan and Gijbels (1995a, 1996). As is clear from (3.1) we need to provide estimates for the conditional bias and variance given in (2.1). Note that the only unknown quantity in the conditional bias is the vector of residuals $\mathbf{r}$, and that the conditional variance depends on the unknown variance function $\sigma^2(\cdot)$. So the essential task is to estimate these two unknown quantities.

Recall that the vector of residuals in the conditional bias is equal to

$$\mathbf{r} = \mathbf{m} - \mathbf{X}\beta = \begin{pmatrix} m(X_1) - \sum_{j=1}^{n} \beta_j (X_1 - x_0)^j \\ \vdots \\ m(X_n) - \sum_{j=1}^{n} \beta_j (X_n - x_0)^j \end{pmatrix}.$$

Using a Taylor expansion of order a, for some given positive integer a, the i^{th} element of the above vector can be approximated by

$$\beta_{p+1}(X_i - x_0)^{p+1} + \cdots + \beta_{p+a}(X_i - x_0)^{p+a}. \tag{3.14}$$

The rate of convergence of the final bandwidth selector will depend on the choice of the integer a. See Huang (1995) or Fan and Gijbels (1996) for a discussion on this issue. For simplicity of presentation we take $a = 2$ in what follows. Using the approximation in (3.14) the conditional bias can be approximated as follows

$$S_n^{-1}\mathbf{X}^T\mathbf{W}\mathbf{r} \approx S_n^{-1} \begin{pmatrix} \beta_{p+1}S_{n,p+1} + \beta_{p+2}S_{n,p+2} \\ \vdots \\ \beta_{p+1}S_{n,2p+1} + \beta_{p+2}S_{n,2p+2} \end{pmatrix},$$

where the elements $S_{n,j}$ have been defined in (2.2). An estimator for the right-hand side of the above expression is

$$S_n^{-1} \begin{pmatrix} \widehat{\beta}_{p+1}S_{n,p+1} + \widehat{\beta}_{p+2}S_{n,p+2} \\ \vdots \\ \widehat{\beta}_{p+1}S_{n,2p+1} + \widehat{\beta}_{p+2}S_{n,2p+2} \end{pmatrix}, \tag{3.15}$$

with $\widehat{\beta}_{p+1}$ and $\widehat{\beta}_{p+2}$ the estimated regression coefficients from a local polynomial fit of degree $p+2$. This fitting step requires a pilot bandwidth h^*, and we suggest an adaptive choice of this pilot bandwidth later on.

The conditional variance in (2.1) depends on the unknown function $\sigma^2(\cdot)$. Assuming local homoscedasticity the conditional variance can be approximated by

$$S_n^{-1}(\mathbf{X}\mathbf{W}^2\mathbf{X})S_n^{-1}\sigma^2(x_0).$$

An estimator for the quantity $\sigma^2(x_0)$ is the normalized weighted residual sum of squares resulting from a local polynomial fit of order $p+2$ using a pilot bandwidth h^*. More explicitly, $\sigma^2(x_0)$ can be estimated by

$$\widehat{\sigma}^2(x_0) = \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 K_{h^*}(X_i - x_0)}{\operatorname{tr}\left\{\mathbf{W}^* - \mathbf{W}^*\mathbf{X}^* S_n^{*-1}\mathbf{X}^{*T}\mathbf{W}^*\right\}}, \tag{3.16}$$

where $\mathbf{X}^*$, $\mathbf{W}^*$ and S_n^* are defined similarly to $\mathbf{X}$, $\mathbf{W}$ and S_n respectively, but now for a local polynomial fit of order $p+2$ using a pilot bandwidth h^*. Note that the estimator (3.12) is similar in nature to the estimator (3.16), but the former one is a global type of estimator, which is intuitively explained by the assumption of homoscedasticity.

The estimator for the conditional variance is then given by

$$S_n^{-1}(\mathbf{X}\mathbf{W}^2\mathbf{X})S_n^{-1}\widehat{\sigma}^2(x_0). \tag{3.17}$$

An estimator for the conditional bias, respectively variance, of the estimator $\widehat{m}_\nu(x_0) = \nu!\widehat{\beta}_\nu$ is then given by

$$\begin{aligned}
\widehat{b}_{\nu,p}(x_0) &= \nu! \times \text{ the } (\nu+1)^{\text{th}} \text{ element of the estimated vector in (3.15)} \\
\widehat{V}_{\nu,p}(x_0) &= (\nu!)^2 \times \text{ the } (\nu+1)^{\text{th}} \text{ diagonal element} \\
&\qquad \text{of the estimated matrix in (3.17).}
\end{aligned} \tag{3.18}$$

This leads to the following estimator for the conditional mean squared error of $\widehat{m}_\nu(x_0)$

$$\widehat{\text{MSE}}_{\nu,p}(x_0) = \widehat{b}^2_{\nu,p}(x_0) + \widehat{V}_{\nu,p}(x_0). \tag{3.19}$$

Note that, in marked contrast with the post-asymptotic substitution methods, (3.18) provides estimates of conditional bias and variance which are the raw materials for constructing confidence intervals. See Section 4.4 of Fan and Gijbels (1996) and Fan, Farmen and Gijbels (1996).

We next describe a procedure for an adaptive choice of the pilot bandwidth h^*. Consider the *Residual Sum of Squares* criterion (RSC)

$$\text{RSC}(x_0; h) = \widehat{\sigma}(x_0)\{1 + (p+1)V\}, \tag{3.20}$$

with $\widehat{\sigma}^2(\cdot)$ defined similarly to (3.16) but for a local polynomial fit of degree p, and where V is the first diagonal element of the matrix $S_n^{-1}(\mathbf{X}\mathbf{W}^2\mathbf{X})S_n^{-1}$.

It can be shown, under certain conditions, that the conditional expectation of the statistic $\mathrm{RSC}(x_0; h)$ can be approximated up to the first order as follows

$$E\left\{\mathrm{RSC}(x_0; h)|\mathcal{X}\right\} \approx \sigma^2(x_0) + C_p\beta_{p+1}^2 h^{2p+2} + (p+1)a_0\frac{\sigma^2(x_0)}{f_X(x_0)nh},$$

where $a_0 = \int K_0^{*2}(t)dt$, and $C_p = \mu_{2p+2} - c_p^T S^{-1} c_p$. From this it follows, heuristically, that the minimizer of $E\{\mathrm{RSC}(x_0; h)|\mathcal{X}\}$ is approximately equal to the minimizer of the right-hand side of the above expression which is

$$h_0(x_0) = \left\{\frac{a_0\sigma^2(x_0)}{2C_p\beta_{p+1}^2 f_X(x_0)n}\right\}^{1/(2p+3)}. \tag{3.21}$$

Comparing this quantity $h_0(x_0)$ with the theoretical optimal variable bandwidth in (3.5) one finds that they relate as follows

$$h_{\nu,p}^{\mathrm{opt}}(x_0) = \mathrm{adj}_{\nu,p}\ h_0(x_0), \tag{3.22}$$

where the adjusting constant is given by

$$\mathrm{adj}_{\nu,p} = \left[\frac{(2\nu+1)C_p\int K_\nu^{*2}(t)dt}{(p+1-\nu)\{\int t^{p+1}K_\nu^*(t)dt\}^2\int K_0^{*2}(t)dt}\right]^{1/(2p+3)}. \tag{3.23}$$

See Fan and Gijbels (1996) for some tabulated values of these adjusting constants.

From the above considerations it is easily seen that an appropriate data-driven global bandwidth selector for estimating $m^{(\nu)}(\cdot)$ on an interval $[a, b]$ by local polynomial fitting of degree p is

$$\widehat{h}_{\nu,p}^{\mathrm{RSC}} = \mathrm{adj}_{\nu,p}\mathrm{argmin}_h \int_{[a,b]} \mathrm{RSC}(y; h)dy. \tag{3.24}$$

The performance of this data-driven global bandwidth estimator has been studied in Fan and Gijbels (1995a). It is this data-driven bandwidth selector which is used as a pilot bandwidth in the refined bandwidth selection procedure which we describe next.

We only discuss the data-driven global bandwidth selection procedure. An analogous procedure, using similar ingredients, exists for local variable bandwidth selection. For details on the data-driven bandwidth selectors based on the residual squares criterion we refer to Fan and Gijbels (1995a).

The algorithm for calculating the *refined global bandwidth selector* reads as follow:

Pilot estimation step. A polynomial of degree $p+2$ is fitted, using the pilot bandwidth $h^* = \widehat{h}_{p+1,p+2}^{\mathrm{RSC}}$, defined in (3.24). This local polynomial fit provides the estimates $\widehat{\beta}_{p+1}$, $\widehat{\beta}_{p+2}$ and $\widehat{\sigma}^2(x_0)$;

Bandwidth selection. The estimator $\widehat{\mathrm{MSE}}_{\nu,p}(\cdot; h)$ is calculated via (3.19) and then the refined global bandwidth selector is given by

$$\widehat{h}_{\nu,p}^{\mathrm{R}} = \mathrm{argmin}_h \int_{[a,b]} \widehat{\mathrm{MSE}}_{\nu,p}(y; h)dy; \tag{3.25}$$

Estimation of $m^{(\nu)}(\cdot)$. The bandwidth $\widehat{h}^{\mathrm{R}}_{\nu,p}$ is used to fit a polynomial of degree p and to obtain $\widehat{m}_\nu(\cdot)$.

We also refer to bandwidth (3.25) as the global pre-asymptotic substitution bandwidth selector.

The local linear fit for the ethanol data presented in Figure 1.3 was obtained using the refined global bandwidth selector (3.25).

Empirical-bias bandwidth selector

Ruppert (1995) proposed a data-driven local variable bandwidth selector which is based on the idea of modeling the bias term. The basic principle is the same as that used in the previous paragraphs: estimate the conditional mean squared error and then minimize this estimate to obtain the bandwidth selector.

The main difference with the pre-asymptotic substitution method discussed above is in the estimation of the conditional bias of $\widehat{m}_\nu(x_0)$. In Ruppert's approach the bias is estimated empirically by calculating $\widehat{m}_\nu(x_0;h) = \widehat{m}_\nu(x_0)$ on a grid of h values, and then modeling in an appropriate way the behavior of $\widehat{m}_\nu(x_0;h)$ as a function of h. The appropriate model is dictated by the asymptotic expression for the conditional bias.

Ruppert (1995) proposed the following procedure for estimating the bias of $\widehat{m}_\nu(x_0;h_0)$. Let $J_b > 1$ be an integer and let $h_0^1, h_0^2, \cdots, h_0^{J_b}$ be in a neighborhood of h_0. Calculate $\widehat{m}_\nu(x_0;h_0^\ell)$ for $\ell = 1, \cdots, J_b$. Then for some integer $a \geq 1$ fit the model

$$\mathrm{d}_0(x_0) + \mathrm{d}_{p+1-\nu}(x_0)h^{p+1-\nu} + \cdots + \mathrm{d}_{p+a-\nu}(x_0)h^{p+a-\nu}, \tag{3.26}$$

to the 'data' $\{(h_0^\ell, \widehat{m}_\nu(x_0;h_0^\ell) : \ell = 1, \cdots, J_b\}$ using ordinary least-squares. This results in the fit

$$\widehat{\mathrm{d}}_0(x_0) + \widehat{\mathrm{d}}_{p+1-\nu}(x_0)h^{p+1-\nu} + \cdots + \widehat{\mathrm{d}}_{p+a-\nu}(x_0)h^{p+a-\nu}.$$

The form of the parametric model (3.26) is inspired by the asymptotic expressions. See for example (2.17) and (2.18) and higher order asymptotic expressions in for example Huang (1995). An estimator for the bias is then

$$\widehat{\mathrm{d}}_{p+1-\nu}(x_0)h^{p+1-\nu} + \cdots + \widehat{\mathrm{d}}_{p+a-\nu}(x_0)h^{p+a-\nu}. \tag{3.27}$$

Other bandwidth selectors

A *nearest neighbor* type of bandwidth is a useful 'reparametrization' of the constant bandwidth h which is such that there are about $2k$, with $k > 0$ a given integer, data points in each neighborhood. Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ denote the ordered observations $X_1, \cdots, X_n$, and let $X_{(\ell)}$ be the design point which is closest to the point x_0. Consider the bandwidth

$$\widehat{h}(x_0) = \frac{X_{(\ell+k)} - X_{(\ell-k)}}{2}. \tag{3.28}$$

The discrete parameter k could be selected using for example cross-validation. See for example Fan and Gijbels (1996).

The bandwidth used in the *loess* procedure, provided in S-Plus is also of the nearest neighbor type. See for example Cleveland (1979) and Cleveland, Grosse and Shyu (1992).

1.4 Choice of the degree

What is the effect of the choice of the degree of the polynomial on the estimation procedure? To get an idea about this we can first look at the asymptotic expressions for the conditional bias and variance provided in (2.16) – (2.18). As already mentioned in Section 1.2.2 the asymptotic conditional bias has a simpler structure when $p-\nu$ is odd than when $p-\nu$ is even. This is an important fact for the bandwidth selection problem. Looking at the orders of the terms, we can conclude that the order of the asymptotic variance does not depend on p. The order of the asymptotic conditional bias however does increase with p, and hence this bias reduces when p is larger. But what about the constant factors in the asymptotic conditional variance? How do they change with p? An investigation of these constant factors revealed that they remain the same when passing from an even degree ($p-\nu$ even) fit to the consecutive odd order fit (fit of odd degree $p-\nu+1$). See Ruppert and Wand (1994) and Fan and Gijbels (1995b). In other words, when $p-\nu$ is even and K is symmetric the local polynomial fit of degree p uses one parameter less than the local polynomial fit of degree $p+1$ without saving on the asymptotic variance. But the extra parameter enables the local polynomial fit of order $p+1$ to reduce the bias of the estimate significantly, particularly at the boundary. Further the constant factors in the asymptotic conditional variance increase with the degree, since more parameters have to be fitted. See Table 3.3 in Fan and Gijbels (1996).

Based on the above purely asymptotic facts one could conclude that a local polynomial fit with $p-\nu$ odd is preferable above a local polynomial fit with $p-\nu$ even. Keeping in mind that a higher order local polynomial fit involves more parameters, it is recommended to use $p=\nu+1$ or $p=\nu+3$ in practice.

Of course one should also consider the influence of the degree p on the finite-sample performance of the estimators, but this is very case-dependent, and no precise information is available on this finite-sample behavior.

So far it has been assumed that in whole the estimation region local polynomial fits of the same order are carried through. When the unknown curve has different degrees of smoothness in different regions, for example a flat part in a certain region and a rapidly changing part in another region, it might be more reasonable to adapt the degree of the local polynomial fit to the point x_0 under consideration. The bandwidth h is assumed to be fixed here.

An adaptive procedure for choosing the degree of the polynomial fit locally is easy in principle: in a fixed neighborhood of x_0 choose that order of fit such that the estimated mean squared error $\widehat{\mathrm{MSE}}_{\nu,p}(x_0;h)$ is the smallest. We describe the procedure more precisely. Denote by

$$x_j = x_L + j\Delta, \qquad j = 0, \cdots, n_{\text{grid}},$$

the grid points at which the estimated curve is to be evaluated, with x_L the first grid point. Suppose that we only consider local polynomial fits up to order

R. For a given bandwidth h the algorithm for an adaptive choice of the order of fit reads as follows:

Step 1. For all orders p ($\nu < p \leq R$) and for all grid points obtain $\widehat{\text{MSE}}_{\nu,p}(x_j;h)$;

Step 2. For each order and for each grid point calculate the smoothed estimated MSE by taking the weighted local average of the estimated MSE in the neighboring $2[h/\Delta]+1$ grid points;

Step 3. For each grid point x_j choose the order p_j which has the smallest smoothed estimated MSE, fit a polynomial of degree p_j locally and obtain the estimator for $m^{(\nu)}(x_j)$.

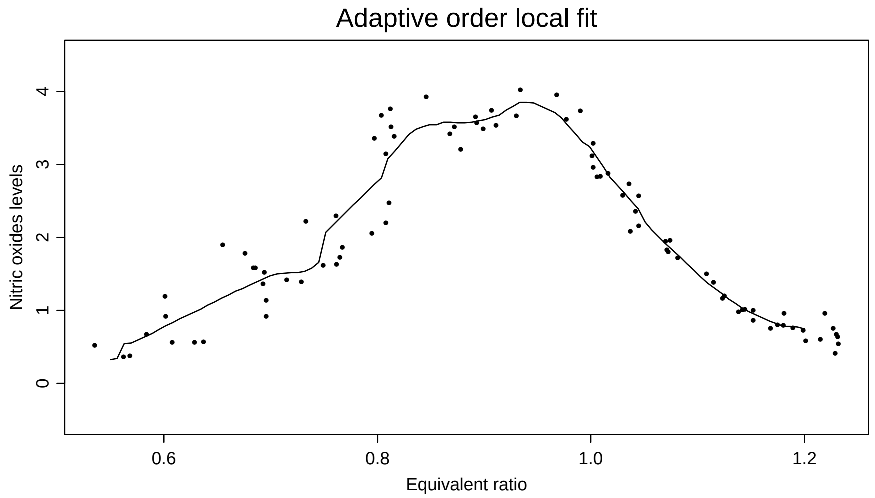

Figure 1.8: Adaptive order local fit for the ethanol data using the rule of thumb bandwidth selector.

Simulation studies in Fan and Gijbels (1995b) revealed that with such an adaptive choice of the degree of the polynomial, the problem of bandwidth selection is less crucial. Hence a crude rule of thumb data-driven bandwidth selector such as the one described in Section 1.3.3 could be sufficient.

As an illustration of the adaptive order estimation procedure we depict in Figure 1.8 the adaptive order local fit for the ethanol data, using the bandwidth $1.5 \times \check{h}_{\text{ROT}}$ defined in (3.9) with $p = 1$ and $\nu = 0$.

1.5 Local modeling

The idea of local polynomial regression extends readily to other settings where the loss function can be non-quadratic. Further, the family of local models does not have to be the family of polynomial functions. It can be any favored family of parametric models. We firstly give a general framework for the local modeling and then illustrate the framework by a few examples. More details can be found in Chapter 5 of Fan and Gijbels (1996).

1.5.1 A general framework

Suppose that we are interested in estimating a population function $\theta(\cdot)$ based on the observed data $\{(X_i, Y_i), i = 1, \cdots, n\}$ from a certain population (X, Y). The function $\theta(\cdot)$ can be, for example, the logit transform of the conditional probability function for the bivariate data given in Figure 1.2 (a) in the context of generalized linear models, or the hazard regression function in the proportional hazards model, or a conditional quantile function. In parametric modeling, the function $\theta(x)$ is modeled by a parametric family $\theta(x) = g(x; \beta)$ and one tries to minimize (or maximize) an objective function:

$$L(\beta) = \sum_{i=1}^{n} \ell\{X_i, Y_i, g(X_i; \beta)\}. \tag{5.1}$$

Usually ℓ is a discrepancy loss function or the log-likelihood function of an individual observation. The local modeling method aims simply at relaxing the global parametric model assumption to the local parametric model assumption. The benifit of this is that the local modeling method is applicable to any smooth function $\theta(\cdot)$, not just to the family $g(x; \beta)$. Hence, the modeling bias can be reduced significantly.

For a given point x_0, the *local modeling* method consists simply of replacing the objective function in (5.1) by its local version

$$L(\beta(x_0)) = \sum_{i=1}^{n} \ell\{X_i, Y_i, g(X_i; \beta(x_0))\} K\left(\frac{X_i - x_0}{h}\right). \tag{5.2}$$

Optimizing (5.2) yields an estimate $\widehat{\beta}(x_0)$ of the local parameter $\beta(x_0)$ and an estimate of the function $\theta(\cdot)$ at the point x_0 is simply $\widehat{\theta}(x_0) = g\{x_0; \widehat{\beta}(x_0)\}$.

Which choice of the objective function should one use? What are the relations between the population (X, Y), and the functions ℓ and θ? Since the local estimate $\widehat{\beta}(x_0)$ optimizes (5.2), the estimate $g\{x_0; \widehat{\beta}(x_0)\}$ should converge to its population version, namely

$$\theta(x_0) = \operatorname{argmin}_a E\{\ell(X, Y, a) | X = x_0\}. \tag{5.3}$$

Here argmin should be replaced by argmax if ℓ is a log-likelihood function. In other words, the estimator $\widehat{\theta}(x_0)$ is a consistent estimator of the function $\theta(x_0)$ if $h \to 0$ in such a way that $nh \to \infty$.

One can choose a prefered parametric model $g(x; \beta)$ in the local model. An advantage of this is that the local model is parsimony to the global parametric model in the sense that when $h = \infty$ the local model reduces to the global model. In other words, the global parametric model $g(x; \beta)$ is embeded in the family of local models indexed by h. This can provide useful diagnostic tools for checking the goodness-of-fit of the parametric model $g(x; \beta)$.

In practice, the function $g(x;\beta)$ is often chosen to be a polynomial function:

$$g(x;\beta) = a_0 + a_1 x + \cdots + a_p x^p.$$

For each given x_0, one can reparametrize the function as

$$g(x;\beta) = \beta_0(x_0) + \beta_1(x_0)(x - x_0) + \cdots + \beta_p(x_0)(x - x_0)^p. \qquad (5.4)$$

The theoretical justification of this comes from Taylor's expansion. With such a reparametrization and suppressing the dependence of $\beta's$ on x_0, (5.2) becomes

$$\sum_{i=1}^{n} \ell\{X_i, Y_i, \beta_0 + \beta_1(X_i - x_0) + \cdots + \beta_p(X_i - x_0)^p\} K\left(\frac{X_i - x_0}{h}\right). \qquad (5.5)$$

Let $\widehat{\beta}_j, (j = 0, \cdots, p)$ optimize (5.5). Then, as in (1.6),

$$\widehat{\theta}_\nu(x_0) = \nu! \widehat{\beta}_\nu, \quad \nu = 0, \cdots, p, \qquad (5.6)$$

estimates the function $\theta^{(\nu)}(x_0)$.

Existing software for fitting parametric models can easily be used to compute $\widehat{\beta}_\nu$, $\nu = 0, \cdots, p$. For each given x_0, the objective function (5.2) or (5.5) is just a weighted version of (5.1) for the parametric model. Thus, the local parameters β can easily be computed by using the software for optimizing (5.1). This is another advantage of the local modeling approach.

It is clear that local polynomial regression estimation is a specific case of the above local modeling framework with a quadratic loss function $\ell(x, y, \theta) = (y - \theta)^2$. An extension of the ideas for estimating bias and variance in Section 1.3 to the above local modeling framework can be found in Fan, Farmen and Gijbels (1996), where methods for selecting bandwidths and constructing confidence intervals are also proposed. A closely related framework is the local estimating equation method introduced by Carroll, Ruppert and Welsh (1996) who propose to use the empirical-bias method to select bandwidths.

1.5.2 Quantile regression and robust regression

Conditional quantiles are very useful and informative summary statistics and have important applications for constructing predictive intervals and detecting heteroscedasticity. In particular, the conditional median regression function is more informative than the conditional mean regression function when the error distribution is asymmetric.

Let $G(y|x)$ be the conditional distribution function of Y given $X = x$. Then the *conditional α-quantile function* is $\xi_\alpha(x) = G^{-1}(\alpha|x)$. This function minimizes (5.3) with the loss function $\ell(x, y, \theta) = \ell_\alpha(y - \theta)$ where

$$\ell_\alpha(t) = |t| + (2\alpha - 1)t. \qquad (5.7)$$

Figure 1.9 (a) presents a few examples of functions $\ell_\alpha(t)$. Thus, to estimate the conditional quantile function, one needs to minimize

$$\sum_{i=1}^{n} \ell_\alpha\{Y_i - \beta_0 - \beta_1(X_i - x_0) - \cdots - \beta_p(X_i - x_0)^p\} K\left(\frac{X_i - x_0}{h}\right) \qquad (5.8)$$

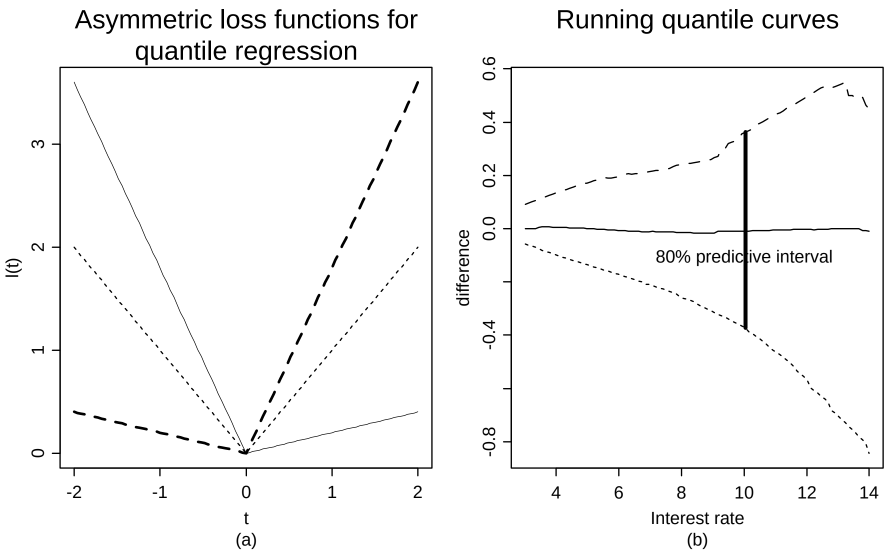

Figure 1.9: Quantile regression. (a) Loss function for quantile regression with $\alpha = 0.1$ (solid lines), 0.5 (short-dashed lines) and 0.9 (long-dashed lines). (b) Conditional quantiles for the data presented in Figure 1.2 (d): Short-dashed curve — $\alpha = 0.1$, solid curve — $\alpha = 0.5$, long-dashed curve — $\alpha = 0.9$. The vertical bar indicates the 80%-predictive interval at the point $x = 10$.

and the resulting estimator for $\xi_\alpha^{(\nu)}(x_0)$ is simply $\nu!\widehat{\beta}_\nu$.

As an illustration, consider the twelve-month Treasury bill data presented in Figure 1.2 (d). Figure 1.9 (b) depicts the conditional median, conditional 10th percentile and the conditional 90th percentile. The fan shape of the conditional quantiles indicates that the variability gets larger as the interest rate gets higher. The intervals sandwiched by the conditional 10th and 90th percentiles are 80%-predictive intervals. For example, given the current interest rate being 10%, with probability 80% the difference of the next week's rate and this week's rate falls in the interval [-0.373%, 0.363%].

For robust estimation of the regression function, one can simply replace the loss function in (5.8) by an outlier-resistent loss function such as

$$\ell(t) = \begin{cases} t^2/2 & \text{when } |t| \le c \\ c|t| - c^2/2 & \text{when } |t| > c, \end{cases}$$

namely, taking the derivative of $\ell(t)$ to be Huber's ψ-function: $\psi_c(t) = \max\{-c, \min(c,t)\}$. When the conditional distribution of Y given $X = x$ is symmetric about the regression function $m(x)$, the resulting estimates are consistent for all $c \ge 0$. Another useful robust procedure is LOWESS, introduced by Cleveland (1979), which reduces the influence of outliers by an iterative reweighted least-squares scheme with weights proportional to the residuals from the previous iteration.

There is a large literature in nonparametric quantile regression and robust regression. Härdle and Gasser (1984) and Tsybakov (1986) considered respectively local constant and local polynomial fitting. Other contributions in this area include Truong (1989), Hall and Jones (1990), Chaudhuri (1991) and Koenker, Portnoy and Ng (1992). For further references see Sections 5.5 and 5.7 of Fan and Gijbels (1996).

1.5.3 Generalized linear models

Generalized linear models are an extension of the normal linear regression model. They contain a family of useful stochastic distributions that can be used to deal with both discrete and continuous responses. In particular, it is assumed that the conditional distribution of Y given X belongs to the canonical exponential family:

$$f(y|x) = \exp([\theta(x)y - b\{\theta(x)\}]/a(\phi) + c(y,\phi)), \tag{5.9}$$

for some known functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot,\cdot)$. The parameter $\theta(\cdot)$ is called the canonical parameter and ϕ is called the dispersion parameter. These two parameters are related to the conditional mean and conditional variance as follows:

$$\begin{aligned} m(x) &= E(Y|X=x) = b'\{\theta(x)\}, \\ \mathrm{Var}(Y|X=x) &= a(\phi)b''\{\theta(x)\}. \end{aligned} \tag{5.10}$$

The function $(b')^{-1}$, which links the mean regression function to the canonical parameter function $\theta = (b')^{-1}(m)$ is called a canonical link. More details on generalized linear models can be found in McCullagh and Nelder (1989) and Section 5.4 of Fan and Gijbels (1996). In the parameter model, the canonical function $\theta(x)$ is typically modeled as a linear function or more generally a polynomial function of x. The aim of the local modeling approach is to release this assumption and to let data search for a function that describes well the data.

As an example, consider the bivariate data in Figure 1.2 (a). It is reasonable to assume that the conditional distribution of Y given $X = x$ is a Bernoulli distribution with probability of success $p(x)$. Then, its conditional density (with respect to the counting measure) can be written as

$$p(x)^y\{1-p(x)\}^{1-y} = \exp\Big[y\theta(x) - b\{\theta(x)\}\Big], \tag{5.11}$$

where $\theta(x) = \log\frac{p(x)}{1-p(x)}$ is the logit transformation of $p(x)$ and $b(\theta) = \log\{1+\exp(\theta)\}$. The canonical link is the logit transformation $\log\{t/(1-t)\}$.

For model (5.9), the function $\ell(x, y, \theta)$ is the log-likelihood function, which is proportional to $\theta y - b(\theta)$. Thus, the local likelihood (5.5) reduces to

$$\sum_{i=1}^n [\{\beta_0 + \beta_1(X_i - x_0) + \cdots + \beta_p(X_i-x_0)^p\}Y_i - b\{\beta_0 + \beta_1(X_i - x_0) + \cdots + \beta_p(X_i-x_0)^p\}]K\left(\frac{X_i - x_0}{h}\right). \tag{5.12}$$

Let $\widehat{\beta}_\nu$, $\nu = 0, \cdots, p$, maximize (5.12). Then, an estimator of the function $\theta^{(\nu)}(x)$ is simply $\widehat{\theta}_\nu(x) = \nu!\widehat{\beta}_\nu$.

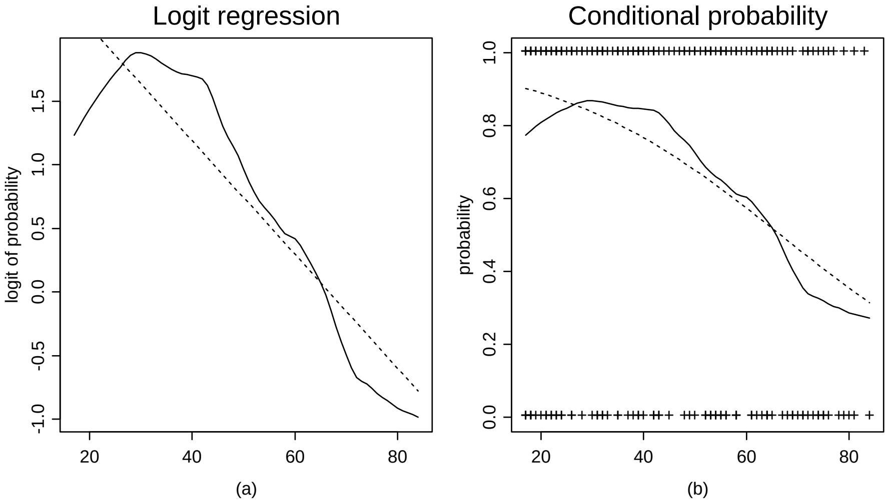

Figure 1.10: Illustration of local likelihood approach for the burn data. (a) Estimated logit transform of the conditional probability. (b) Estimated conditional probability. Solid curve — local modeling with about 40% of the data points in each local neighborhood; dashed curve — global parametric logit linear model.

We now use the burn data in Figure 1.2 (a) to illustrate the application of the local modeling approach to the generalized linear models setting. Of interest is the estimation of the probability of surviving given the age of victims. We employed the local linear fit with the Epanechnikov kernel function and chose a nearest neighbor type of bandwidth with a span of 0.4 of the data range. The results are summarized in Figure 1.10. Note that the conditional probability function must be monotonic for the parametric linear model, whereas for the local linear model, the conditional probability function can be any curve. The former model can overstate the probability of survival for the younger group and for the senior group. The solid curves in Figure 1.10 suggest that the conditional probability function is unimodal, which is reasonable in the current context.

For some literature on nonparametric generalized linear models, see Green and Yandell (1985), Tibshirani and Hastie (1987), Gu (1990), Severini and Staniswalis (1994), Stone (1994) and Fan, Heckman and Wand (1995), among others.

1.5.4 Proportional hazards model

A useful model for studying the effect of a covariate X on a survival time T is the proportional hazards model. Let $\lambda(t|X)$ be the *conditional hazard rate* function defined by

$$\lambda(t|X) = \lim_{\Delta \downarrow 0} \frac{P(t \le T < t + \Delta | T \ge t, X)}{\Delta},$$

and let $\lambda_0(t)$ be the *baseline hazard function*, representing the hazard risk at a certain given value of the covariate, say 0. The *proportional hazards model* assumes that the covariate x inflates the hazard risk by a factor of $\exp\{\theta(x)\}$, namely,

$$\lambda(t|x) = \lambda_0(t) \exp\{\theta(x)\}. \qquad (5.13)$$

The function $\theta(x)$ is called *hazard regression function*, and characterizes the risk contribution of the covariate at value x. Model (5.13) with time-dependent covariates was discussed by Cox (1972).

A salient feature of survival analysis is that the survival time is often not fully observable due, for example, to the termination of a study or early withdrawel. For the i^{th} individual, the observed data are the minimum $Z_i = \min(T_i, C_i)$ of the survival time T_i and the censoring time C_i, and the censoring indicator $\delta_i = I(T_i \le C_i)$, indicating if the observed time is the survival time, along with the covariate X_i. In other words, the available data are

$$\{(X_i, Z_i, \delta_i) : i = 1, \cdots, n\},$$

which are assumed a random sample from a population (X, Z, δ). Based on these censored data, one is interested in estimating the hazard regression function $\theta(x)$ and the baseline hazard function $\lambda_0(t)$. For a thorough introduction to survival analysis, see recent books by Fleming and Harrington (1991) and Andersen, Borgan, Gill and Keiding (1993).

In the parametric analysis, a linear form $\theta(x) = \beta x$ is imposed on the hazard regression function. The local modeling approach aims at removing this restriction and is applicable to any smooth hazard regression function. For simplicity of discussions, we restrict the covariate to be univariate. For multivariate settings, a dimensionality reduction technique such as additive models should be incorporated. See Hastie and Tibshirani (1990).

A commonly-used technique for estimating the hazard regression function is the *partial likelihood* technique introduced by Cox (1975). Let $t_1^o < \cdots < t_N^o$ denote the ordered observed failure times. Let (j) provide the label for the item failing at t_j^o so that the covariates associated with the N failures are $X_{(1)}, \cdots, X_{(N)}$. Denote by $R_j = \{i : Z_i \ge t_j^o\}$, the risk set at time instantaneously before t_j^o. Then, the log-partial likelihood in our context is given by

$$\sum_{j=1}^{N} \Big[\theta(X_{(j)}) - \log\Big(\sum_{k \in R_j} \exp\{\theta(X_k)\}\Big)\Big]. \qquad (5.14)$$

Substituting the parametric form of $\theta(\cdot)$ into (5.14) yields a maximum partial likelihood estimate of the hazard regression function.

We now apply the local modeling technique to estimate the hazard regression function $\theta(\cdot)$. For a given x_0, approximate $\theta(x)$ by

$$\theta(x) \approx \beta_0 + \cdots + \beta_p (x - x_0)^p, \qquad (5.15)$$

for x in a neighborhood of x_0. Let

$$\beta = (\beta_1, \cdots, \beta_p)^T \text{ and } \mathbf{X}_j = \{(X_j - x_0), \cdots, (X_j - x_0)^p\}^T.$$

Then the *local partial likelihood* is

$$\sum_{j=1}^{N} K_h(X_{(j)} - x_0)\Big[\mathbf{X}_{(j)}^T\beta - \log\Big\{\sum_{k\in R_j} \exp(\mathbf{X}_k^T\beta)K_h(X_k - x_0)\Big\}\Big], \tag{5.16}$$

where $K_h(\cdot) = K(\cdot/h)/h$. The kernel function is imposed to ensure that model (5.15) is only used locally. See Fan, Gijbels and King (1997) for a derivation of the local partial likelihood (5.16). When the kernel function is uniform and the bandwidth is of the nearest neighbor type, the local likelihood (5.16) was introduced by Tibshirani and Hastie (1987). For a related approach based on the local likelihood, see Gentleman and Crowley (1991).

The function value $\theta(x_0)$ is not directly estimable since (5.16) does not depend on the intercept β_0. However, the derivative functions are directly estimable. Let $\widehat{\beta}(x_0)$ be the maximum local log-partial likelihood estimate that maximizes (5.16). An estimate $\widehat{\theta}_\nu(x_0)$ of $\theta^{(\nu)}(x_0)$ is given by $\nu!\widehat{\beta}_\nu(x_0)$.

We impose the condition $\theta(0) = 0$ for identifiability. With this extra constraint, the function $\theta(x)$ can be estimated by

$$\widehat{\theta}(x) = \int_0^x \widehat{\theta}'(t)dt, \tag{5.17}$$

where $\widehat{\theta}'(t) = \widehat{\theta}_1(t)$ is the derivative estimator. In practice, the function $\widehat{\theta}_1(x)$ is often evaluated at either grid points or the design points. Assume that $\widehat{\theta}_1(x_j) = \widehat{\beta}_1(x_j)$ are computed at points $\{x_0, \cdots, x_m\}$. Then, $\widehat{\theta}(x_i)$ can be approximated by the trapezoidal rule

$$\widehat{\theta}(x_i) = \sum_{j=1}^{i}(x_j - x_{j-1})(\widehat{\beta}_{1,j} + \widehat{\beta}_{1,j-1})/2,$$

where $\widehat{\beta}_{1,j} = \widehat{\beta}_1(x_j)$. Once $\widehat{\theta}(\cdot)$ is estimated, one can use Breslow's (1974) type of estimator for the *cumulative hazard function* $\Lambda_0(t) = \int_0^t \lambda_0(u)du$:

$$\widehat{\Lambda}_0(t) = \sum_{j=1}^{N}\left[\sum_{i\in\mathcal{R}_j} \exp\{\widehat{\psi}(X_i)\}\right]^{-1} I\{t_j \le t\}. \tag{5.18}$$

A kernel smoothing technique can then be employed to obtain an estimate of $\lambda_0(t)$ via

$$\widehat{\lambda}_0(t) = \int W_g(t - x)d\widehat{\Lambda}_0(x), \tag{5.19}$$

where $W_g(\cdot) = W(\cdot/g)/g$ with W a kernel function and g a bandwidth. With the estimated hazard regression function and the estimated baseline hazard function, one can construct the estimated hazard risk function of individuals with covariate x via (5.13), namely

$$\widehat{\lambda}(t|x) = \widehat{\lambda}_0(t)\exp\{\widehat{\theta}(x)\}.$$

Using the relationship between the *survivor function* and the hazard risk function, individuals with the covariate value x have estimated survivor function

$$\widehat{P}(T > t|X = x) = \exp\{-\int_0^t \widehat{\lambda}(u|x)du\}.$$

From this function, one can easily get an estimate of their mean survival time:

$$\widehat{E}(T|X = x) = \int_0^\infty \widehat{P}(T > t|X = x)dt,$$

namely the area of the survivor curve.

We conclude this section with an analysis of the Primary Biliary Cirrhosis (PBC) data set that can be found in Fleming and Harrington (1991). PBC is a rare but fatal chronic liver disease of unknown cause. The analysis is here based on the data collected at Mayo Clinic between January 1974 and May 1984. Of 312 patients who participated in the randomized trial, 187 cases were censored. In our analysis, we take the time (in days) between registration and death, or liver transplantation or the time of the study analysis (July 1986) as response and the ages of the patients as a covariate. The observed data are presented in Figure 1.11 (a). The local partial likelihood method (5.16) with $p = 2$ was employed for three different bandwidths $h = 10$, 20 and 30. The estimated hazard regression function and its derivative function are respectively given in Figure 1.11 (c) and (b). Note that since the hazard regression function is only identifiable within a constant, the curves in Figure 1.11 (c) are normalized to have the same average height so that they can be better compared. Figure 1.11 (c) reveals the fact that it is reasonable to model linearly the hazard regression function of covariate age.

1.6 Applications

This section discusses briefly a few more applications of local polynomial fitting. We first discuss how to estimate a conditional variance and how to detect a discontinuity of a regression curve, and then show how to estimate conditional densities.

1.6.1 Estimation of variance

Conditional variance functions have a variety of statistical applications such as for example measuring the volatility or risk in finance, monitoring the reliability in nonlinear prediction, identifying homoscedastic transforms in regression, and monitoring the signal-to-noise ratios in quality control of experimental design. Because of their important applications in finance where data are often

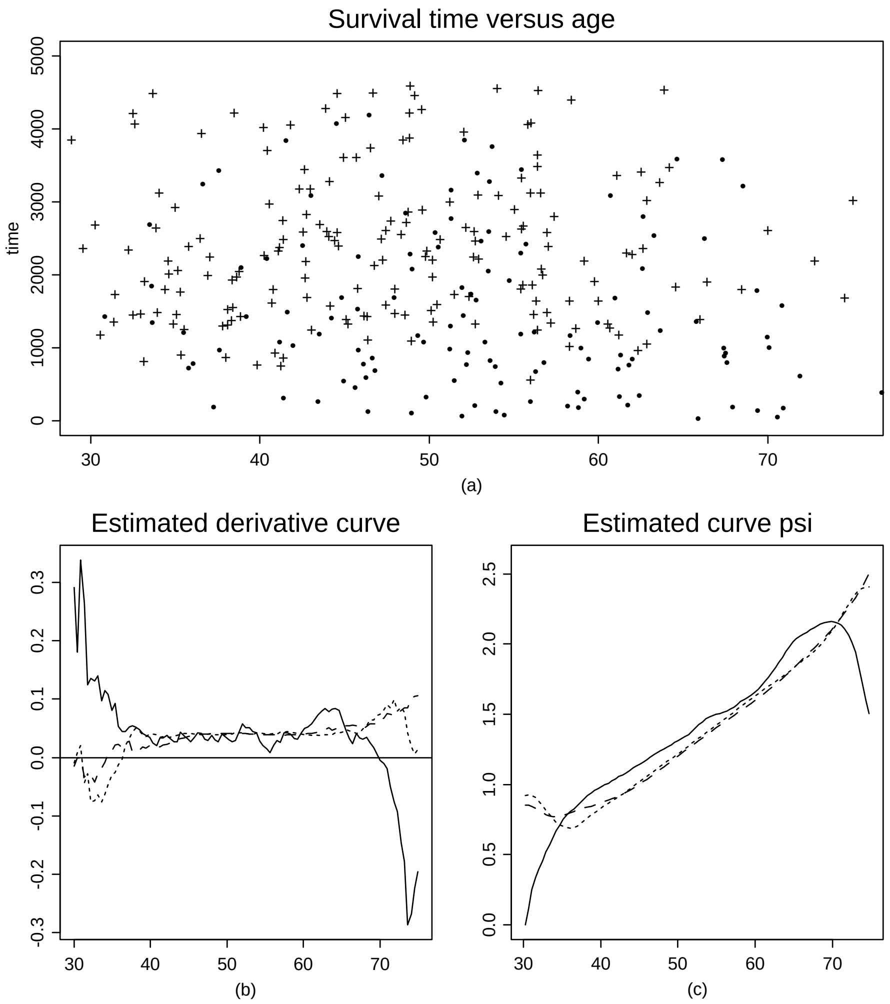

Figure 1.11: Local partial likelihood estimation of the hazard regression function. (a) Observed time versus age with '+' indicating censored observations. (b) Estimated derivative function $\theta'(\cdot)$. (c) Estimated hazard regression function $\theta(\cdot)$; solid curve — bandwidth = 10; short-dashed curve — bandwidth = 20; long-dashed curve — bandwidth = 30.

dependent, we formulate our problems in the stochastic setup. For simplicity of presentation, we focus only on local linear fitting.

Let $\{(X_i, Y_i)\}$ be a two-dimensional strictly stationary process having the same joint distribution as (X, Y). Let $m(x) = E(Y|X = x)$ and $\sigma^2(x) = \mathrm{Var}(Y|X = x)$ be respectively the regression function and the conditional variance function. Our approach is based on the residuals of the local linear fit. Let $\widehat{m}_{h_1,K}(\cdot)$ be the local linear fit of $m(\cdot)$ using a kernel K and a bandwidth

h_1. Consider the squared residuals

$$\widehat{r}_i = \{Y_i - \widehat{m}_{h_1,K}(X_i)\}^2. \tag{6.1}$$

Note that the conditional variance function can be expressed as

$$\sigma^2(x) = E[\{Y - m(X)\}^2 | X = x],$$

which is the regression function of the squared residuals. Therefore, a natural procedure is to run a local linear fit on the squared residuals. Let $\widehat{\sigma}^2_{h_2,W}(x)$ be the local linear fit based on the data $\{(X_i, \widehat{r}_i), i = 1, \cdots, n\}$, using a bandwidth h_2 and a kernel W. Then, it was shown by Ruppert, Wand, Holst and Hössjer (1995), and Fan and Yao (1996) that the estimator $\widehat{\sigma}^2_{h_2,W}(x)$ performs as well as the ideal estimator, which is a local linear fit to the true squared residuals

$$\{(X_i, \{Y_i - m(X_i)\}^2), i = 1, \cdots, n\}$$

using the same bandwidth h_2 and the same kernel W. In particular, the bias of the estimator $\widehat{\sigma}^2_{h_2,W}(x)$ is of order $O(h_2^2 + h_1^4 + (nh_1)^{-1})$ though the bias of $\widehat{m}_{h_1,K}$ is of order $O(h_1^2)$, and the variance of $\widehat{\sigma}^2_{h_2,W}(x)$ is of order $O\{(nh_2)^{-1}\}$.

The above discussions suggest that if the bandwidth h_1 is of order $n^{-1/5}$, then the residual-based *conditional variance estimator* performs asymptotically as well as the ideal one. In particular, the optimal bandwidth for estimating the mean regression function is permitted to be used for computing the residuals. This leads to the following data-driven conditional variance estimator.

Let $\widehat{h}(X_1, \cdots, X_n; Y_1, \cdots, Y_n)$ be a data-driven bandwidth selection rule for local linear regression based on the data $(X_1, Y_1), \cdots, (X_n, Y_n)$. This can be in one case a cross-validation bandwidth rule, and in another case one of the bandwidth selection methods discussed in Section 1.3.3. Our data-driven version of conditional variance estimation reads as follows:

1. Use bandwidth $h_1 = \widehat{h}(X_1, \cdots, X_n; Y_1, \cdots, Y_n)$ in local linear regression to obtain the estimate $\widehat{m}_{h_1,K}(X_i)$ for $i = 1, \cdots, n$;

2. Compute squared residuals $\widehat{r}_i = \{Y_i - \widehat{m}_{h_1,K}(X_i)\}^2$, $i = 1, \cdots, n$;

3. Apply bandwidth $h_2 = \widehat{h}(X_1, \cdots, X_n; \widehat{r}_1, \cdots, \widehat{r}_n)$ in local linear regression to obtain $\widehat{\sigma}^2_{h_2,W}(\cdot)$.

To illustrate the usefulness of the above automatic method, consider the yields of twelve-month Treasury bill. We use the refined global bandwidth selector of Fan and Gijbels (1995a) (see also Section 1.3.3) and the Epanechnikov kernel. Figure 1.12 (a) gives the estimated mean regression function for the data in Figure 1.2 (d). The bandwidth $\widehat{h}_1 = 3.99$ was chosen by the software. The two dashed curves around a solid one indicate one standard error above and below the estimated mean regression function. Figure 1.12 (b) depicts the estimated conditional standard deviation (the volatility function) and the conditional variance function. The bandwidth $\widehat{h}_2 = 3.63$ was selected by the software. Visual inspection suggests that the volatility function should

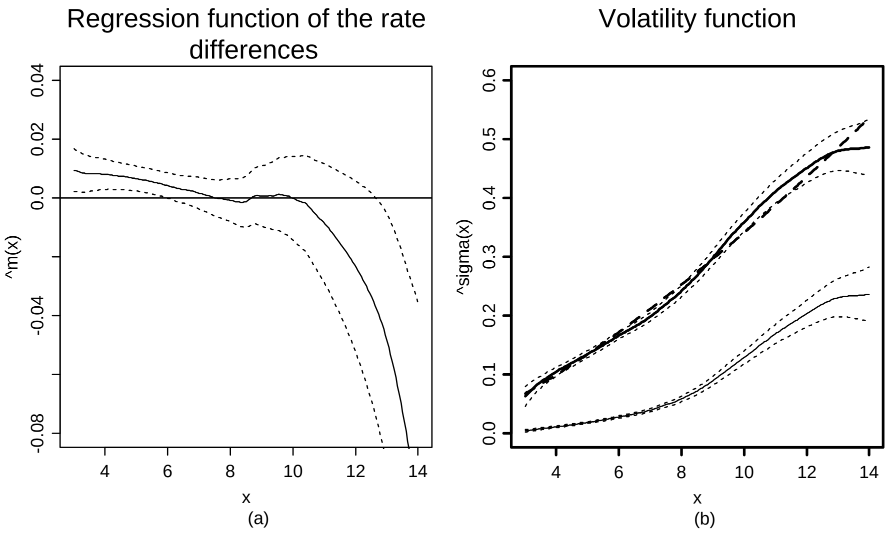

Figure 1.12: The regression function and the volatility function for the twelve-month Treasury bill data. (a) The estimated mean regression function and, (b) Estimated volatility function (thick curve) and the estimated conditional variance function (thin curve) for the data in Figure 1.2 (d).

be a power function. Indeed, the correlation coefficient between $\{\log(x_j)\}$ and $\{\log(\widehat{\sigma}(x_j))\}$ is 0.997!, where $x_j, (j = 1, \cdots, 201)$ are grid points in the interval [3, 14]. Fitting a line through the data $\{(\log(x_j), \log\{\widehat{\sigma}(x_j)\}), j = 1, \cdots, 201\}$, we obtain the estimate

$$\widehat{\sigma}(x) = 0.0154x^{1.3347}.$$

This estimate is presented as a thick-dashed curve in Figure 1.12 (b). This is an example where the nonparametric analyses yield a good parametric model $\sigma(x) = \alpha x^{\beta}$. Based on the linear regression on the data $(\log(X_i), \log(\widehat{r}_i))$, one can also obtain directly an estimate of α and β.

1.6.2 Change point detection

Jump points arise for example when the treatment effects change suddenly without warning or planning. Examples of this kind occur in the context of medical monitoring and quality control. The statistical problem can be abstracted as follows.

Assume that the bivariate data $(X_1, Y_1), \cdots, (X_n, Y_n)$ are a random sample from a population (X, Y) with conditional mean function m, which is smooth except for a few number of jump discontinuities. For simplicity, we assume that there is only one single discontinuity point, also called *change point*.

One approach to the problem is to regard the change point as the location where the derivative function $|m'(\cdot)|$ is maximized. Thus, a natural approach is to first estimate the derivative curve and then find the maximizer of the absolute value of the estimated derivative function. Let $D(x,h)$ be a derivative estimator resulting from a local polynomial fit of order p with bandwidth h and kernel K. For simplicity, assume that the support of K is $[-1,1]$. The above idea translates into the following estimating scheme: plot the function $|D(\cdot,h)|$ for a range of values of h and identify the jump as the point x in the vicinity of which $|D(x,h)|$ is consistently large for a *range of values of* h. More precisely, let $\widetilde{x}(h)$ be the global maximum of the function $|D(\cdot,h)|$. Put

$$\widetilde{x}_-(h) = \sup_{h_1\in[h,\eta_n]}\{\widetilde{x}(h_1) - 2h_1\}, \quad \widetilde{x}_+(h) = \inf_{h_1\in[h,\eta_n]}\{\widetilde{x}(h_1) + 2h_1\}, \tag{6.2}$$

for $h \le \eta_n$, where $\eta_n > 0$ is a prescribed number, tending to zero more slowly than $n^{-1}\log n$. Let $\widetilde{h}$ denote the infimum of values h such that $\widetilde{x}_-(h) \le \widetilde{x}_+(h)$. The proposed jump point estimator is $\widehat{x}_0 = \widetilde{x}(\widetilde{h})$. Such a kind of idea is due to Gijbels, Hall and Kneip (1995).

Gijbels, Hall and Kneip (1995) also propose a further refinement of the above idea. For a given bandwidth h, pretend the change point lies in the interval $\widetilde{x}(h) \pm 2h$ and the regression function is a step function on this interval. Then, find the unknown location of the jump such that it minimizes the residual sum of squares, using only the data in the strip $\widetilde{x}(h) \pm 2h$. The resulting estimator is a refinement of the estimator $\widetilde{x}(h)$. In particular, we can take $h = \widetilde{h}$ to yield a refinement of $\widehat{x}_0$.

Another intuitive method is a one-sided kernel approach due to Müller (1992). The idea extended to the local polynomial setting works as follows. Let K_- be a kernel function supported on $[-1,0]$. Let $\widehat{m}_-(x,h)$ be a local polynomial fit using the bandwidth h and the kernel K_-. Note that the estimator $\widehat{m}_-(x,h)$ uses only the local data on the left-side of the point x. Analogously, let K_+ be a kernel function supported on $[0,1]$ and $\widehat{m}_+(x,h)$ be a local polynomial fit using the bandwidth h and the kernel K_+. Then, $\widehat{m}_+(x,h)$ uses only the data on the right-hand side of the point x. At the smooth locations, the estimates $\widehat{m}_-(x,h)$ and $\widehat{m}_+(x,h)$ are about the same, since both are consistent estimates of $m(x)$. At the discontinuity point, however, they estimate respectively the left-limit and the right-limit of the function m at the point x. Thus, a natural estimator is the location such that the difference function $|\widehat{m}_+(x,h) - \widehat{m}_-(x,h)|$ is maximized. The bandwidth for detecting the change point is typically much smaller than the optimal bandwidth for curve estimation. For a wavelet-based method, see Wang (1995).

To illustrate the above two approaches, consider the simulated data in Figure 1.2 (b). We apply a local linear fit with the Epanechnikov kernel. Figure 1.13 (a) presents the absolute value of the estimated derivative curve for a wide range of bandwidths, ranging from 0.05 to 0.3. Though the bandwidths have been increased by a factor of 6, the location of the maximum remains about the same — all of them are close to the true value. Using the bandwidth $h = 0.05$, we obtain the jump point $\widehat{x} = 0.49$. Figure 1.13 (b) depicts the difference curves $|\widehat{m}_+(x,h) - \widehat{m}_-(x,h)|$ for two bandwidths $h = 0.1$ and $h = 0.3$. Since the kernel is one-sided, we do not have enough data at the boundary points to fit a line. Thus, we only compute the difference curves for $x \in [0.1, 0.9]$. The choice $h = 0.3$ gives a correct estimate $\widehat{x} = 0.51$ of the change point, while the

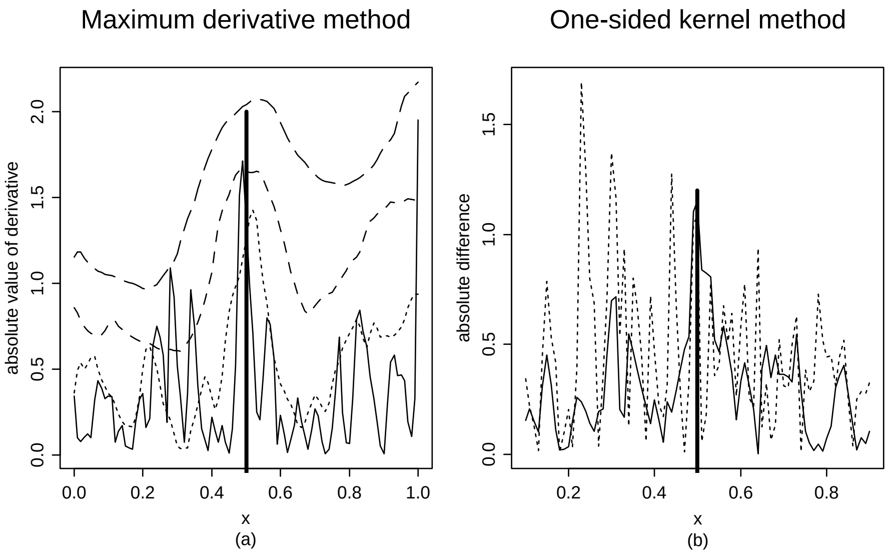

Figure 1.13: Change-point detection. (a) Maximum derivative method. Presented is the absolute value of the estimated derivative curve with bandwidths $h = 0.3$ (long-dashed curve), 0.2 (dashed curve), 0.1 (short-dashed) and 0.05 (solid curve). (b) One-sided kernel method. Presented are $|\widehat{m}_+(x,h) - \widehat{m}_-(x,h)|$ with bandwidth $h = 0.3$ (solid curve) and $h = 0.1$ (dashed curve).

choice $h = 0.1$ results in a wrong estimated value 0.23 of the location of the change point. This example shows two advantages of the derivative method: the estimator of the jump point is less sensitive to the choice of the bandwidth and the sparsity of local data points occurs less often.

1.6.3 Estimation of conditional densities

In the univariate setting, the population mean, while easier to estimate, is far less informative than the probability density function. Likewise, in the regression setting, the regression function is far less informative than the conditional probability density function. The conditional density function provides information about not only its center (the regression function) but also the spreadness from the center. If the shape of the *conditional density estimator* appears multimodal or highly skewed, it is not appropriate to summarize the density via the conditional mean regression function.

Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be a random sample from the population (X, Y) with the conditional density $g(y|x)$. Note that

$$E\left\{K_{h_2}(Y-y)|X=x\right\} \approx g(y|x), \quad \text{as } h_2 \to 0, \tag{6.3}$$

where $K_h = K(\cdot/h)/h$ is a rescaled kernel function. Thus, $g(y|x)$ can be regarded approximately as the regression function of the variable $K_{h_2}(Y-y)$

on X. Considerations of this nature lead to the following local polynomial regression problem:

$$\sum_{i=1}^{n}\left\{K_{h_2}(Y_i-y)-\sum_{j=0}^{p}\beta_j\,(X_i-x)^j\right\}^2 W_{h_1}(X_i-x), \tag{6.4}$$

for a given bandwidth h_1 and a kernel function W. Let $\{\widehat{\beta}_j(x,y), j=0,\cdots,p\}$ be the solution of the least-squares problem. Then, from Section 1.1.3, an estimator of $g^{(\nu)}(y|x)=\frac{\partial^\nu g(y|x)}{\partial y^\nu}$ is $\nu!\widehat{\beta}_\nu(x,y)$. We write $\widehat{g}(y|x)=\widehat{\beta}_0(x,y)$ as the estimator of the conditional density.

The bandwidth selection problem appears somewhat challenging for this problem. For a given bandwidth h_2, the estimator $\widehat{g}(y|x)$ is derived from a local polynomial fit. Hence, any bandwidth selection method in Section 1.3.3 can be used to select h_1. For selection of h_2, a simple method is to apply a rule of thumb bandwidth selector such as the normal reference rule (p.45 of Silverman, 1986)

$$\widehat{h}_2=\left[\frac{8\pi^{1/2}\int K^2(x)dx}{3\{\int x^2K(x)dx\}^2}\right]^{1/5} S_y n^{-1/5}, \tag{6.5}$$

where S_y is the sample standard deviation of $Y_1,\cdots,Y_n$. When K is the Gaussian kernel, $\widehat{h}_2=1.06S_yn^{-1/5}$; for the Epanechnikov kernel, $\widehat{h}_2=2.34S_yn^{-1/5}$. The method outlined above is proposed in Fan, Yao and Tong (1996), where one can also find many interesting numerical examples.

With the estimated conditional density function, one can derive many statistical estimators. For example, the mean regression function can simply be estimated by

$$\widehat{m}(x)=\int y\widehat{g}(y|x)dy.$$

It can be shown that this estimator is the same as the local polynomial regression estimator when the kernel function K has mean zero. Similarly, one can derive estimates for the conditional variance and conditional quantile function.

Acknowledgments

This project was supported by a direct grant from the Faculty of Science of The Chinese University of Hong Kong. The first author was supported partially by NSF grant DMS-9504414 and NSA Grant 96-1-0015. This chapter was written while the second author was visiting the Department of Statistics of The Chinese University of Hong Kong. The author gratefully acknowledges the Department for its hospitality and support. The second author was partially supported by 'Projet d'Actions de Recherche Concertées' (No. 93/98-164) from the Belgian government and by research grant (No. 15.001.95F) of the National Science Foundation (FNRS), Belgium.

1.7 References

Andersen, P.K., Borgan, Ø., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag, New York.

Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics*, 30:89–99.

Carroll, R.J., Ruppert, D. and Welsh, A.H. (1996). Nonparametric estimation via local estimating equations, *Unpublished manuscript.*

Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local Bahadur representation. *Ann. Statist.*, 19:760–777.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, 74:829–836.

Cleveland, W.S., Grosse, E. and Shyu, W.M. (1992). Local regression models. In Chambers, J.M. and Hastie, T.J., editors, *Statistical Models in S*, pages 309–376. Wadsworth & Brooks, California.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *J. Royal Statist. Soc. B*, 34:187–220.

Cox, D.R. (1975). Partial likelihood. *Biometrika*, 62:269–276.

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.*, 87:998–1004.

Fan, J., Farmen, M. and Gijbels, I. (1996). A blueprint of local maximum likelihood estimation. Discussion Paper #9604, Institute of Statistics, Catholic University of Louvain, Louvain-la-Neuve, Belgium.

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.*, 20:2008–2036.

Fan, J. and Gijbels, I. (1995a). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Royal Statist. Soc. B*, 57:371–394.

Fan, J. and Gijbels, I. (1995b). Adaptive order polynomial fitting: bandwidth robustification and bias reduction. *J. Comp. Graph. Statist.*, 4:213–227.

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

Fan, J., Gijbels, I. and King, M. (1997). Local likelihood and local partial likelihood in hazard regression. *Ann. Statist.*, to appear.

Fan, J., Heckman, N.E. and Wand, M.P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.*, 90:141–150.

Fan, J. and Yao, Q. (1996). Efficient estimation of conditional variance functions in stochastic regression. *Unpublished manuscript.*

Fan, J., Yao, Q. and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83:189–206.

Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis.* Wiley, New York.

Gentleman, R. and Crowley, J. (1991). Local full likelihood estimation for the proportional hazards model. *Biometrics*, 47:1283–1296.

Gijbels, I., Hall, P. and Kneip, A. (1995). On the estimation of jump points in smooth curves. Discussion Paper #9515, Institute of Statistics, Catholic University of Louvain, Louvain-la-Neuve, Belgium.

Green, P.J. and Yandell, B. (1985). Semiparametric generalized linear models. In *Proceedings of the 2nd International GLIM Conference.* Lecture Notes in Statistics, 32. Springer-Verlag, Berlin.

Gu, C. (1990). Adaptive spline smoothing in non-Gaussian regression models. *J. Amer. Statist. Assoc.*, 85:801–807.

Hall, P. and Jones, M.C. (1990). Adaptive M-estimation in nonparametric regression. *Ann. Statist.*, 18:1712–1728.

Härdle, W. and Gasser, T. (1984). Robust non-parametric function fitting. *J. Royal Statist. Soc. B*, 46:42–51.

Hastie, T.J. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Huang, L.-S. (1995). *On nonparametric estimation and goodness-of-fit.* PhD thesis, Department of Statistics, The University of North Carolina, Chapel Hill.

Koenker, R., Portnoy, S. and Ng, P. (1992). Nonparametric estimation of conditional quantile function. In Dodge, Y., editor, *Proceedings of the conference on L_1 - Statistical Analysis and Related Methods*, pages 217–229. Elsevier.

Marron, J.S. and Nolan, D. (1988). Canonical kernels for density estimation. *Statist. Prob. Lett.*, 7:195–199.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* Chapman and Hall, London.

Müller, H.-G. (1992). Change-points in nonparametric regression analysis. *Ann. Statist.*, 20:737–761.

Ruppert, D. (1995). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Technical Report* #1137, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York.

Ruppert, D., Sheather, S.J. and Wand, M.P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.*, 90:1257–1270.

Ruppert, D. and Wand, M.P. (1994). Multivariate weighted least squares regression. *Ann. Statist.*, 22:1346–1370.

Ruppert, D., Wand, M.P., Holst, U. and Hössjer, O. (1995). Local polynomial variance function estimation. *Unpublished manuscript.*

Seifert, B. and Gasser, T. (1996). Finite-sample variance of local polynomials: analysis and solutions. *J. Amer. Statist. Assoc.*, 91:267–275.

Severini, T.A. and Staniswalis, J. (1994). Quasi-likelihood estimation in semiparametric models. *J. Amer. Statist. Assoc.*, 89:501–511.

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

Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10:1040–1053.

Stone, C.J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.*, 22:118–184.

Tibshirani, R. and Hastie, T.J. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.*, 82:559–567.

Truong, Y.K. (1989). Asymptotic properties of kernel estimators based on local medians. *Ann. Statist.*, 17:606–617.

Tsybakov, A.B. (1986). Robust reconstruction of functions by the local-approximation method. *Problems of Information Transmission*, 22:133–146.

Wang, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika*, 82:385–397.