1 Local Modeling: Density Estimation and Nonparametric Regression

Jianqing FAN and Runze LI

Chinese University of Hong Kong and Pennsylvania State University

1.1 INTRODUCTION

Local modeling approaches are useful tools for exploring features of data without imposing a parametric model. These approaches have been received increasing attention in last two decades and successfully applied to various scientific disciplines, such as, economics, engineering, medicine, environmental science, health science and social science. There are a vast amount of literature on this topic. A comprehensive account of local modeling can be found in the books by Silverman (1986), Härdle (1990), Scott (1992), Wand and Jones (1995), Fan and Gijbels (1996), Siminoff (1996) and Bowman and Azzalini (1997). See also Fan and Gijbels (2000) and Fan and Müller (1995) for a brief overview on this topic. In this chapter, we will introduce fundamental ideas of local modeling and illustrate the ideas by real data examples. For ease of presentation, we will omit all technical parts. A list of references will be given at the end of chapter. Readers may further pursue theoretical results from the references therein.

This chapter basically consists of two parts: kernel density estimation and local polynomial fitting. In Section 1.2, the kernel density estimation method will be introduced. Important issues, including bandwidth selection, will be addressed. Real data examples will be used to illustrate the ideas how to implement this type of method. Local polynomial regression will be introduced in Section 1.3. In this section, we also discuss how to decide the amount of smoothing, and extend the ideas of local polynomial regression to other contexts. The idea is further extended to the local likelihood and local partially likelihood in Section 1.4. Section 1.5 introduces the ideas of nonparametric smoothing tests. Section 1.6 summarizes some applications of local modeling, including estimation of conditional quantile functions, conditional variance functions and conditional densities, and change point detection.

1.2 DENSITY ESTIMATION

Suppose that $X_1, \cdots, X_n$ are an independent and identically distributed sample from a population with an unknown probability density $f(x)$. Of interest is to estimate the density f. In explanatory data analysis, we

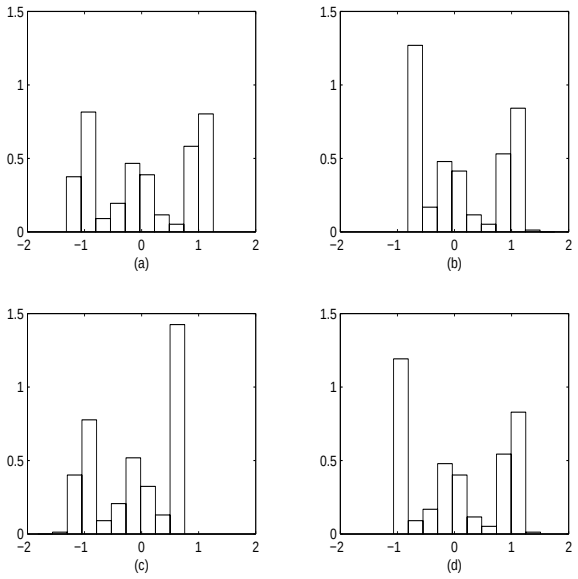

Fig. 1.1 Histograms of a sample of size 300 from a mixture of normal distribution $1/3N(-1, 0.1^2)+1/3N(0, 0.25^2)+1/3N(1, 0.1^2)$.

may construct a histogram for the data. If the resulting histogram has a bell shape, then we may assume that the samples were taken from a normal distribution. In this situation, one may just estimate the population mean and variance using the sample mean and sample variance because a normal distribution is completely determined by its mean and variance. In general, parametric approaches to estimation of a density function assume that the density belongs to a parametric family of distributions, such as normal, gamma or beta family. In order to fully specify the density function, one has to estimate the unknown parameters using, for example, maximum likelihood estimation. One may use prior knowledge or scientific reasons to determine a parametric distribution family. In explanatory data analysis, data analysts frequently construct a histogram based on the sample, and then draw reasonable conclusions on the population density.

1.2.1 Histogram

A histogram is usually formed by partitioning the range of data into equally length intervals, called bins, and then drawing a block over each interval with height being the proportion of the data falling in the bin divided by the width of the bin. Specifically, the histogram estimate at a point x is given by

$$\widehat{f}(x, h) = \frac{\text{number of observations in the bin containing } x}{nh},$$

where h is the width of the bins, namely binwidth. For a fixed choice of bins, it can be shown that under some mild conditions, $\widehat{f}(\cdot, h)$ is a maximum likelihood estimate of the unknown density f. It is worthwhile to note that the nonparametric maximum likelihood estimate of the unknown density f without any further

restriction does not exist, since

$$\max_{\{f: f\geq 0, \int f=1\}} \prod_{i=1}^{n} f(X_i) = \infty.$$

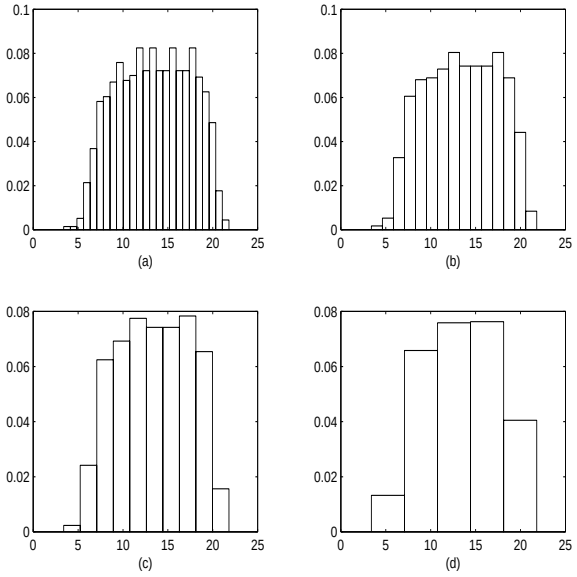

Fig. 1.2 Histograms for crab sizes. The data is the length of crab (cm).

When one constructs a histogram, one has to choose the binwidth and the centers of bins. Figure 1.1 depicts four histograms based on the same data set and the same binwidth, but using different locations of bin centers. It can be seen from Figure 1.1 that the shapes of the resulting histograms are quite different. This implies that the histogram suffers the "edge" effect. Figure 1.2 shows four histograms of the lengths of crabs, collected from 1973 to 1986, but with different binwidths. The crab data set is available from the website of statlib at Carnegie Mellon University at http://lib.stat.cmu.edu. From Figure 1.2, if the binwidth h is too small, then the resulting histogram is rough, on the other hand, if the binwidth is too large, then the resulting histogram is too smooth. Thus constructing a histogram actually is not so simple! Usually one may start from an undersmoothed histogram, and then increase gradually the binwidth until getting a satisfactory result.

The histogram is the oldest and most widely used nonparametric estimate of density. The choice of binwidth is a smoothing problem. The edge effect of histograms can be repaired by the kernel density estimation introduced in next section. Furthermore, the kernel estimate will result in a smooth density curve rather than a step function as in histograms. It is an improved technique over the kernel density estimation.

1.2.2 Kernel Density Estimation

A kernel density estimate is defined as

$$\widehat{f}_h(x) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\},$$

where $K(\cdot)$ is a function satisfying $\int K(x)\,dx = 1$, called a *kernel function* and h is a positive number, called a *bandwidth* or a *smoothing parameter*. A density function such as the plot (thick curve) in Figure 1.3 is usually obtained by evaluating the function $\widehat{f}_h(x)$ over a few hundred of grid points. From the definition, indeed, the kernel estimate is the average of density functions $h^{-1}K\{(x - X_i)/h\}$, which smoothly redistribute the point mass at the point X_i. Figure 1.3 depicts the redistribution of point masses. To facilitate notation, let $K_h(t) = \frac{1}{h}K(t/h)$ be a rescaling function of K. This allows us to write

$$\widehat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i). \tag{2.1}$$

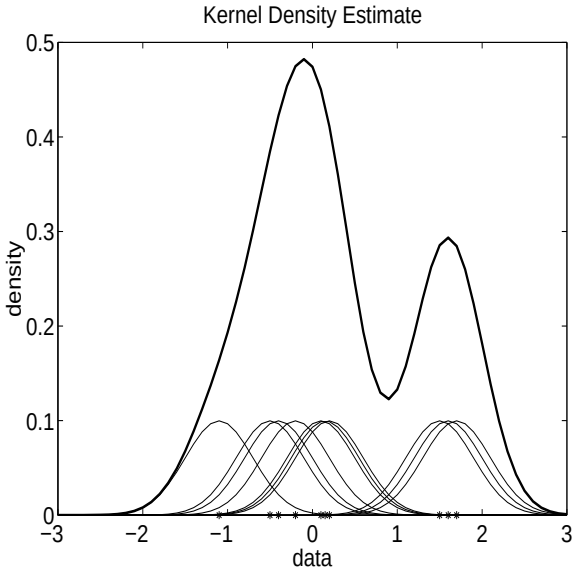

Fig. 1.3 Kernel density estimate for an hypothetical data set (thick curve). It smoothly redistributes the point mass at X_i by the function $(nh)^{-1}K\{(x - X_i)/h\}$. The small bumps show how point masses are redistributed.

It is well known that the choice of K is not very sensitive, scaled in a canonical form as discussed by Marron and Nolan (1988), to the estimate $\widehat{f}_h(x)$. Thus it is assumed throughout this chapter that the kernel function is a symmetric probability density function. The most commonly used kernel function is the Gaussian density function given by

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2). \tag{2.2}$$

Other popular kernel functions include the symmetric beta family

$$K(t) = \frac{1}{Beta(1/2, \gamma+1)}(1-t^2)_+^\gamma, \qquad \gamma = 0, 1, \cdots, \tag{2.3}$$

where $+$ denotes the positive part, which is assumed to be taken before exponentiation, so that the support of K is $[-1, 1]$, and $Beta(\cdot, \cdot)$ is a beta function. The corresponding kernel functions when $\gamma = 0, 1, 2$ and 3 are the uniform, the Epanechnikov, the biweight and the triweight kernel functions. Figure 1.4 shows these kernel functions.

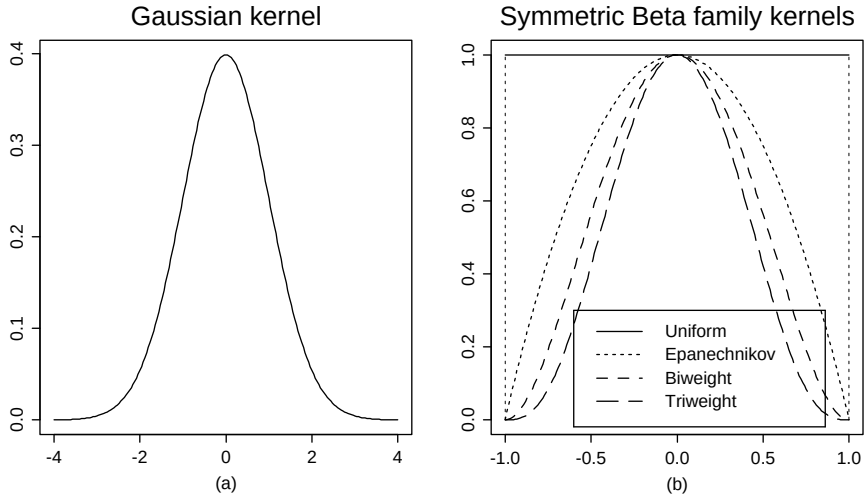

Fig. 1.4 Commonly-used kernels. (a) Gaussian kernel; (b) Symmetric Beta family of kernels that are renormalized to have maximum height 1.

The smoothing parameter h controls the smoothness of density estimates, acting as the binwidth in histograms. The choice of the bandwidth is of crucial importance. If h is chosen too large, then the resulting estimate misses fine features of the data, while if h is selected too small then spurious sharp structure become visible. See Figure 1.6 for example. In fact, it can be shown that under some mild conditions, when $n \to \infty$, $h \to 0$ and $nh \to \infty$,

$$E\widehat{f}_h(x) - f(x) = \frac{f''(x)}{2}\mu(K)h^2 + o(h^2) \tag{2.4}$$

and

$$\text{var}\{\widehat{f}_h(x)\} = \frac{R(K)f(x)}{nh}(1 + o(1)), \tag{2.5}$$

where $\mu(K) = \int t^2 K(t)dt$ and $R(K) = \int K^2(t)dt$. Thus, from (2.4) and (2.5), a large bandwidth h results in a large bias while a small bandwidth produces an estimate with a large variance. A good choice of bandwidth would balance the bias and variance trade-off. This is conveniently assessed by the *Asymptotic Mean Integrated Square Error* (AMISE) which is defined as

$$\text{AMISE}(h) = \frac{\mu^2(K)h^4}{4}\int \{f''(x)\}^2 dx + \frac{R(K)}{nh}. \tag{2.6}$$

Minimizing (2.6) with respect to h gives the ideal bandwidth

$$h_I = \left(\frac{R(K)}{\mu^2(K) \int \{f''(x)\}^2 dx} \right)^{1/5} n^{-1/5}, \tag{2.7}$$

which involves the unknown density function, and cannot be directly used in kernel smoothing. Since the choice of bandwidth is critical to kernel density estimation, there has a large literature on this topic. See Jones, Marron and Sheather (1996a, b) for a survey. In practice, we may take the Gaussian density with variance σ^2 as a reference density. In this situation, equation (2.7) becomes

$$h_I = \left(\frac{8\sqrt{\pi} R(K)}{3\mu^2(K)} \right)^{1/5} \sigma n^{-1/5}. \tag{2.8}$$

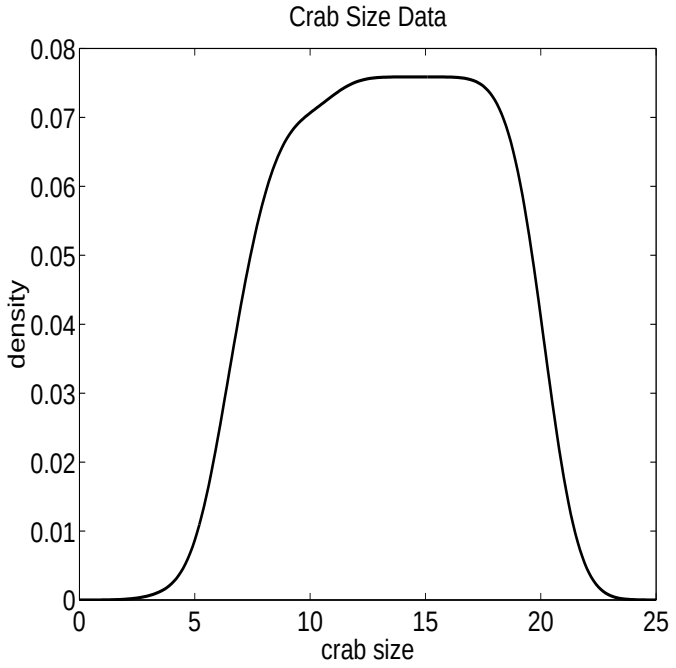

Fig. 1.5 Automatic kernel density estimates using the bandwidth according the rule of thumb. The data set is the crab size data collected from 1973 to 1986.

Here we focus on a rule of thumb. See Silverman (1986). The rule of thumb of bandwidth selection is to replace σ by the sample standard deviation s_n. Thus for the Gaussian kernel,

$$\widehat{h}_I = 1.06 s_n n^{-1/5},$$

and for the symmetric beta family

$$\widehat{h}_I = \left[\frac{8\sqrt{\pi} Beta(1/2, 2\gamma + 1)}{\{Beta(3/2, \gamma + 1)\}^2} \right]^{1/5} s_n n^{-1/5}.$$

Figure 1.5 depicts a kernel density estimate of the length of crab using the bandwidth $\widehat{h}_I$ with the Gaussian

kernel. From the shape of the estimated density curve, it seems that a normal distribution is not appropriate for modeling the crab size.

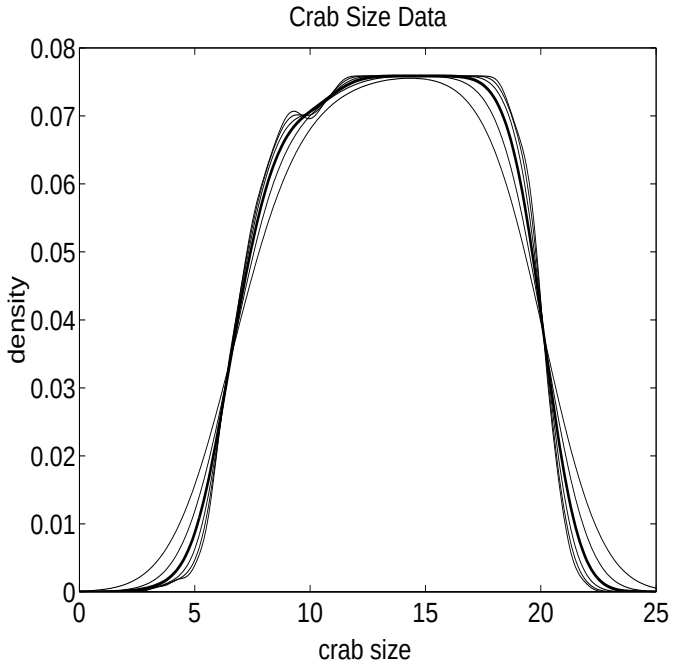

Fig. 1.6 A family of kernel estimates. The data set is the crab size data. The thick curve corresponds to $\widehat{h}_I$.

While the rule of thumb works well for many data sets, it tends to produce oversmooth estimates as the referenced density is a Gaussian density. Another method to avoid choosing a single optimal bandwidth is the family smoothing approach. This can be done by using a family of estimates

$$\{\widehat{f}_h(x), h = 1.4^j \widehat{h}_I, j = -3, -2, -1, 0, 1, 2\} \tag{2.9}$$

and then overlaying them in the same plot. The family smoothing approach allows us to explore possible patterns contained in data using different scale of bandwidths. This is closely related to scale space ideas in computer science. Choosing a smaller bandwidth acts as "zoom in", while selecting a larger bandwidth corresponds to "zoom out" in the scale space. These ideas have been further developed in a SiZer map, proposed by Chaudhuri and Marron (1999). The SiZer map can detect significant features in estimated curve with different scales. Figure 1.6 depicts a family smoothing plot for the crab size data.

The density estimation method is also a powerful graphic tool for comparing the results of two experiments. This is related to the classical two-sample mean problem. The advantage of the kernel smoothing approach over the traditional two sample tests is that the smoothing approach can show an overall pattern of the experiments, including the locations of centers and the dispersions of the data. Further, it gives us some ideas of two population distributions. To illustrate the idea, we applied the smoothing techniques for two subsets of the crab size data. One contains the data set of Year 1976, and the other consists of the data set of Year 1986. The two estimated density curves are depicted in Figure 1.7. They have different centers and dispersions.

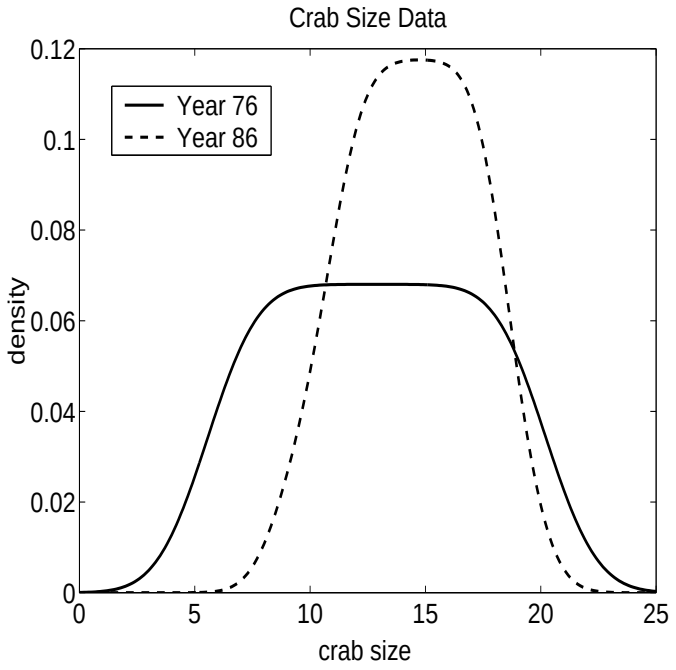

Fig. 1.7 Comparison of the length of crabs between Year 1976 and Year 1986. The sample mean and standard deviation for Year 1976 are 12.9020 and 4.2905, while the sample mean and standard deviation for Year 1986 are 14.4494 and 2.6257, respectively.

In this section, the bandwidth remains constant, that is, it depends on neither the location x nor the datum point X_i. This kind of bandwidth is referred as a global bandwidth. From (2.7), it is desirable to use a larger bandwidth when changes of curvature is small and use a smaller bandwidth when curvature of underlying density dramatically changes. This leads to studying variable bandwidth selection, which suggests the use of different bandwidth at different location of x. Usually, a global bandwidth is easier to choose than the variable bandwidth. In order to use a constant bandwidth, one may first transform the data by

$$Y_i = g(X_i), \qquad i = 1, \cdots, n,$$

where g is a given monotone increasing function. The transformation g should be chosen so that the transformed data have a density with more homogeneous degree of smoothness so that a global bandwidth for the transformed data is more appropriate. Then apply the kernel density estimate to the transformed data set and obtain the estimate $\widehat{f}_Y(y)$. Finally apply the inverse transform to obtain the density of X:

$$\widehat{f}_X(x) = g'(x)\widehat{f}_Y(g(x)) = g'(x)n^{-1}\sum_{i=1}^{n} K_h(g(x) - g(X_i)).$$

The performance of this type estimate has been illustrated in Wand, Marron and Ruppert (1991). Marron and Yang (1999) proposed an approach to selecting a good transformation g.

1.3 LOCAL POLYNOMIAL FITTING

Regression is one of the most useful techniques in statistics. Consider the $(d+1)$-dimensional data $(\mathbf{X}_1, Y_1)$, $\cdots$, $(\mathbf{X}_n, Y_n)$, which form an independent and identically distributed sample from a population $(\mathbf{X}, Y)$, where $\mathbf{X}$ is a d-dimensional random vector and Y is a random variable. Of interest is to estimate the regression function $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$. In other words, the data are regarded as realizations from the model:

$$Y = m(\mathbf{X}) + \varepsilon,$$

where ε is a random error with zero mean. For a given data set, one may try to fit the data by using a linear regression model. If a nonlinear pattern appears in the scatter plot of Y against $\mathbf{X}$, one may employ polynomial regression to reduce the modeling bias of linear regression. Consider for example the data plotted in Figure 1.8, 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 Figure 1.8, it can be seen that the relationship between the concentration of nitric oxides and the equivalence ratio is highly nonlinear. Polynomial regression is used to fit the data. Figure 1.8 presents the resulting fits by using four different degrees of polynomials. One can easily see that all resulting fits have subtantial biases. Because polynomial functions have all orders of derivatives everywhere, and polynomial degree cannot be controlled continuously, polynomial functions are not very flexible in modeling features encountered in practice. Further individual observations can have a large influence on remote parts of the curve in polynomial regression models. Nonparametric regression techniques can be used to repair the drawbacks of polynomial fitting. Fan and Gijbels (1996) give detailed background and excellent overview on various nonparametric regression techniques, which can be classified into two categories. One is to approximate the regression function globally and the other one is to parameterize the regression function $m(\mathbf{x})$ locally. Two common methods of global approximation are the *spline approach* and the *orthogonal series method*. In this section, we focus on the techniques of local modeling.

1.3.1 Kernel Regression

Consider the bivariate data (X_1, Y_1), $\cdots$, (X_n, Y_n), an i.i.d. sample from the model:

$$Y = m(X) + \varepsilon,$$

where ε is random error with $E(\varepsilon|X) = 0$ and $\text{var}(\varepsilon|X = x) = \sigma^2(x)$. The nonparametric regression problem is to estimate the regression function $m(\cdot)$ with imposing a form. Usually, a datum point closer to x carries more information about the value of $m(x)$. Therefore an intuitive estimator for the regression function $m(x)$ is the running local average. An improved version of this is the locally weighted average. That is

$$\widehat{m}(x) = \sum_{i=1}^{n} w_i(x) Y_i / \sum_{i=1}^{n} w_i(x).$$

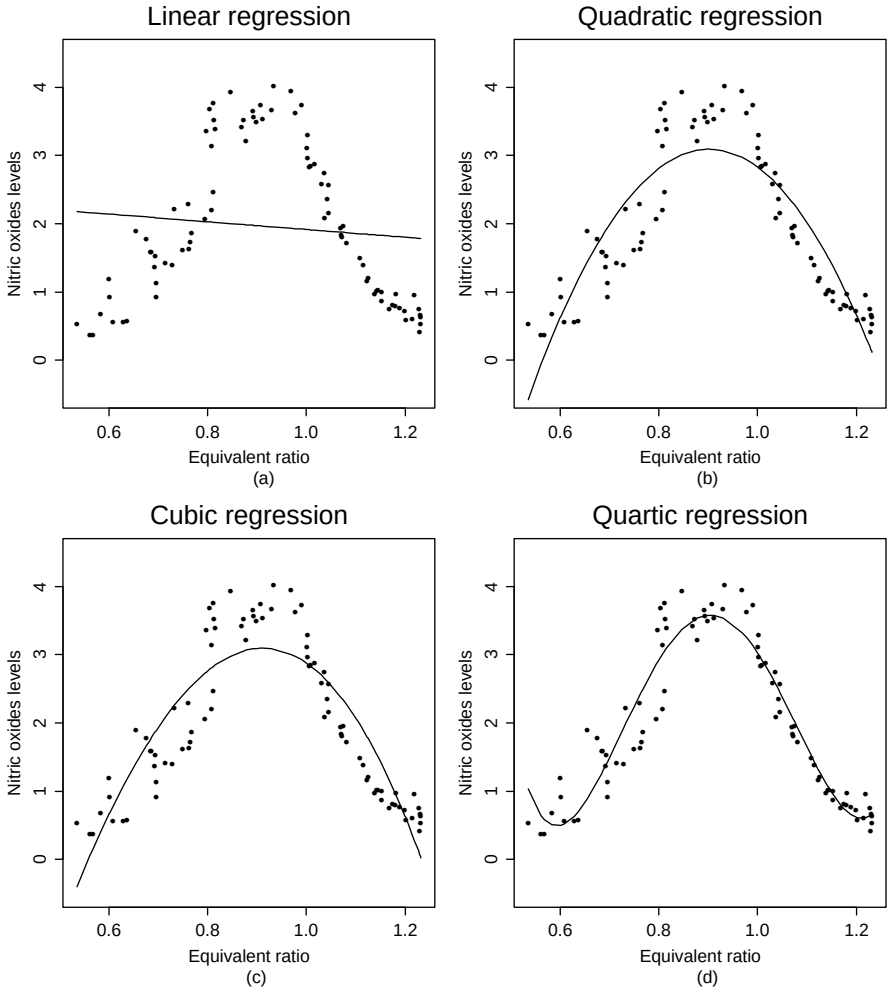

Fig. 1.8 Polynomial fits to the ethanol data. Presented are the scatter plots of the concentration of nitric oxides against the equivalence ratio along with the fitted polynomial regression functions. Adapted from Fan and Gijbels (2000).

An alternative interpretation of locally weighted average estimators is that the resulting estimator is the solution to the following weighted least-squares problem:

$$\min_{\theta} \sum_{i=1}^{n} (Y_i - \theta)^2 w_i(x).$$

In other words, the kernel regression estimators are a weighted least squares estimate at the point x using a local constant approximation.

Setting the weights $w_i(x) = K_h(X_i - x)$ results in the NW kernel regression estimator, which is given by

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h(X_i - x) Y_i}{\sum_{i=1}^{n} K_h(X_i - x)}. \tag{3.1}$$

See Nadaraya (1964) and Watson (1964).

Since the denominator in (3.1) is a random variable, it is inconvenient to take derivatives with respect to x and to derive the asymptotic properties of the estimator. Assume that the data have already been sorted according to the X-variable. Taking the local weights $w_i(x) = \int_{s_{i-1}}^{s_i} K_h(u-x)du$ with $s_i = (X_i + X_{i+1})/2$, $X_0 = -\infty$ and $X_{n+1} = +\infty$, we obtain the GM regression estimator given by

$$\widehat{m}_h(x) = \sum_{i=1}^{n} \int_{s_i-1}^{s_i} K_h(u-x)\, du Y_i.$$

See Gasser and Müller (1984).

Just like the kernel density estimate, the choice of bandwidth is critical to the quality of the estimate. A too large bandwidth yields an oversmooth estimate, while a too small bandwidth gives a rough estimate. The basic asymptotic properties of the NW and GM regression estimators have been well established. The asymptotic biases and variances of these two estimators are depicted in Table 1.1, taken from Fan (1992). The properties on the GM estimator were established in Mack and Müller (1989) and Chu and Marron (1991).

Table 1.1 Leading Terms in the Asymptotic Biases and Variances

Method	Bias	Variance
NW estimator	$\{m''(x) + \frac{2m'(x)f'(x)}{f(x)}\}b_n$	V_n
GM estimator	$m''(x)b_n$	$1.5V_n$
Local linear	$m''(x)b_n$	V_n

Here $b_n = \frac{1}{2}\int_{-\infty}^{+\infty} u^2 K(u)\, du\, h^2$ and $V_n = \frac{\sigma^2(x)}{f(x)nh}\int_{-\infty}^{+\infty} K^2(u)\, du$.

1.3.2 Local Polynomial Regression

As indicated in the last section, both the NW estimator and the GM estimator are a local constant fit. It is natural to extend this to a local polynomial fit. The idea of local polynomial regression has been around for a long time. Since both a local constant and local polynomial fits use effectively datum points in a local neighborhood, this idea is referred as "local modeling". It appeared in the statistical literature in Stone (1977) and Cleveland (1979). Stone (1980, 1982) shows that local regression achieves optimal rates in a minimax sense. Müller (1987) establishes the equivalence between a local polynomial fit and a local constant fit under an equally-spaced design model. Fan (1992, 1993) focus on local linear regression in the random design case and show that it has many advantages, such as simple expression for local bias and variance, spatial adaptation and high minimax efficiency. Fan and Gijbels (1992) proved that theoretically the local linear regression estimator adapts automatically to the boundary. This was also empirically observed by Tibshirani and Hastie (1987). Ruppert and Wand (1994) extended the results of Fan and Gijbels (1992) to the case of local polynomial estimation. A thorough study of this topic can also be found in Chapters 3 and 4 of Fan and Gijbels (1996).

Suppose that the regression function m is smooth. For z in a neighborhood of x, it follows from using Taylor's expansion that

$$m(z) \approx \sum_{j=1}^{p} \frac{m^{(j)}(x)}{j!}(z-x)^j \equiv \sum_{j=1}^{p} \beta_j (z-x)^j. \tag{3.2}$$

Thus for X_i close enough to x,

$$m(X_i) \approx \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \equiv \mathbf{X}_i^T \beta,$$

where $\mathbf{X}_i = (1, (X_i - x_0), \cdots, (X_i - x_0)^p)^T$ and $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^T$. Intuitively datum points further from x have less information about $m(x)$. This suggests using a locally weighted polynomial regression

$$\sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \beta)^2 K_h(X_i - x). \tag{3.3}$$

Denote by $\widehat{\beta}_j$ $(j = 0, \cdots, p)$ the minimizer of (3.3). The above exposition suggests that an estimator for the regression function $m(x_0)$ is

$$\widehat{m}(x_0) = \widehat{\beta}_0(x_0). \tag{3.4}$$

Furthermore, an estimator for the ν-th order derivative of $m(x_0)$ at x_0 is

$$\widehat{m}_\nu(x_0) = \nu! \widehat{\beta}_\nu(x_0).$$

In general, local polynomial fitting has certain advantages over the NW and the GM estimators not only for estimating regression curves, but also for derivative estimation.

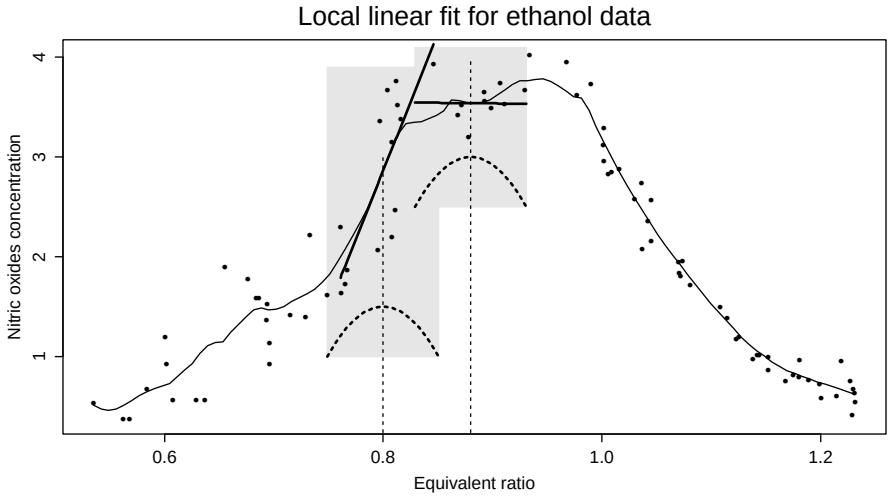

Fig. 1.9 Illustration of the local linear fit. For each given x_0, a linear regression is fitted through the data contained in the strip $x_0 \pm h$, using the weight function indicated at the bottom of the strip. The interactions of the fitted lines and the short dashed lines are the local linear fits. Adapted from Fan and Gijbels (2000).

To better appreciate the above local polynomial regression, consider the Ethanol data presented in Figure 1.8. The window size h is taken to be 0.051 and the kernel is the Epanechnikov kernel. To estimate the

regression function at the point $x_0 = 0.8$, we use the local data in the strip $x_0 \pm h$ to fit a regression line (c.f. Figure 1.9). The local linear estimate at x_0 is simply the intersection of the fitted line and the line $x = x_0$. Suppose that we wish to estimate the regression function at another point $x_0 = 0.88$, another line is fitted using the data in the window 0.88 ± 0.051. The whole curve is obtained by estimating the regression function in a grid of points. Indeed, the curve in Figure 1.9 was obtained by 101 local linear regressions, taking the 101 grid points from 0.0535 to 1.232.

The *local linear regression smoother* is particularly simple to implement. Indeed, the estimator has the simple expression

$$\widehat{m}_L(x) = \sum_{i=1}^{n} w_i(x) Y_i, \tag{3.5}$$

where with $S_{n,j}(x) = \sum_{i=1}^{n} K_h(X_i - x)(X_i - x)^j$,

$$w_i(x) = K_h(X_i - x)\{S_{n,2}(x) - (X_i - x)S_{n,1}(x)\}/(S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)). \tag{3.6}$$

We can either use the explicit formula (3.6) or a regression package to compute it. It has several nice properties such as high statistical efficiency (in an asymptotic minimax sense), design adaption (Fan, 1993) and good boundary behavior (Fan and Gijbels, 1992; Ruppert and Wand, 1994). The asymptotic bias and variance for this estimator is

$$E\{\widehat{m}_L(x)|X_1, \ldots, X_n\} - m(x) = \mu(K)\frac{m''(x)}{2}h^2 + o(h^2) \tag{3.7}$$

and

$$\text{var}\{\widehat{m}_L(x)|X_1, \ldots, X_n\} = R(K)\frac{\sigma^2(x)}{f(x)nh} + o\left(\frac{1}{nh}\right), \tag{3.8}$$

provided that the bandwidth h tends to zero in such a manner that $nh \to \infty$, where f is the marginal density of X, namely, *the design density*. See Fan (1993). Table 1.1 lists the leading term in the asymptotic bias and variance. By comparing the leading terms in the asymptotic variance, clearly the local linear fit uses locally one extra parameter without increasing its variability. But this extra parameter creates opportunities for significant bias reduction, particularly at the boundary regions and slope regions. This is evidenced by comparing their asymptotic biases.

Local linear fitting requires a choice for the smoothing parameter h 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 truely for local polynomial regression. It has been shown that the Epanechnikov kernel is optimal in some sense. See Gasser, Müller, and Mamitzsch (1985), Granovsky and Müller (1991) and Chapter 3 of Fan and Gijbels (1996).

The bandwidth selection is critical to all nonparametric estimators. A too large bandwidth creates excessive biases in nonparametric estimates and a too small bandwidth results in a large variance in nonparametric estimate. There are two basic choices of bandwidth: subjective and data-driven. In subjective choices, data analysts use different bandwidths to estimate the regression function and choose the one that visually balances the bias and variance trade-off. Trials-and-errors are needed in this endeavor. Alternatively, one can present the nonparametric estimates using a few different bandwidths (c.f. Figure 1.6 for a similar idea). The data-driven bandwidth is to let data themselves choose a bandwidth that balances the bias and variance, via minimizing certain estimated *Mean Integrated Square Errors* (MISE).

We now briefly discuss some data-driven choices of the bandwidth. By (3.7) and (3.8), the weighted MISE of the local linear estimator is

$$\frac{\mu(K)^2 h^4}{4} \int \{m''(x)\}^2 w(x) dx + \frac{R(K)}{nh} \int \frac{\sigma^2(x)}{f(x)} w(x) dx.$$

The asymptotic optimal bandwidth, that minimizes the asymptotic weighted MISE of $\widehat{m}_L(x)$, is given by

$$h_{opt} = \left(\frac{R(K) \int \sigma^2(x) f^{-1}(x) w(x)\, dx}{\mu^2(K) \int \{m''(x)\}^2 w(x)\, dx} \right)^{1/5} n^{-1/5}, \tag{3.9}$$

where $w(x)$ is a weight function.

The optimal bandwidth involves the unknown regression function and the unknown density function of X. Hence it cannot be applied directly. There are many references on the topic of bandwidth selection. See Chapter 4 of Fan and Gijbels (1996) and references therein. Here we focus on the cross-validation method, which is conceptually simple, but needs intensive computation. Let $\widehat{m}_{h,(-i)}(x)$ be the local linear regression estimator (3.3) without using the i^{th}-observation (X_i, Y_i). We now analogously validate the "goodness-of-fit" by measuring the "prediction error" $Y_i - \widehat{m}_{h,(-i)}(X_i)$. The cross-validation criterion measures the overall "prediction errors", which is defined by

$$CV(h) = n^{-1} \sum_{i=1}^{n} \{Y_i - \widehat{m}_{h,(-i)}(X_i)\}^2. \tag{3.10}$$

The cross-validation bandwidth selector $\widehat{h}_{CV}$ chooses the one that minimizes $CV(h)$.

In what follows, we illustrate the methodology of local linear regression in details by an environmental data set. This data set consists of 612 observations of 15 variables and has been analyzed by Rawlings and Spruill (1994). See Section 2 of Rawlings and Spruill (1994) for a detailed description. Here we are interested in how *depth to mottling* (DMOT) of soil affects the increment of diameter growth of some kinds of pine. Thus we take the increment of diameter as a response variable Y and the DMOT of soil as an independent variable X. After excluding the data points with missing values, we have 216 observations. The scatter plot of the data is depicted in Figure 1.10.

The cross-validation method was used to search a bandwidth over 20 grid points $0.15 * 1.1^j$ multiplying the range of X variable, $j = 0, \cdots, 19$. With the smallest bandwidth 0.15 multiplying the range of X, we used 15 percent of data around x_0 to estimate $m(x_0)$, while with the largest bandwidth 0.15×1.1^{19} multiple the range of X, we used about 92 percent of data around x_0 to estimate $m(x_0)$. Here the Epanechnikov kernel was used. The plot of cross-validation scores against candidate bandwidths is depicted in Figure 1.10(a). The corresponding $\widehat{h}_{CV}$ is 12.776.

With the selected bandwidth, we are able to estimate the regression function. In nonparametric regression, one usually plots the curve of the estimated regression function. Thus one has to evaluate the regression function over a grid of points. Usually we take the grid of points evenly distributing over the range of X. A natural question arises here is how many grid points at which the estimate needed to be evaluated. Figure 1.10 (c)-(e) depicts the resulting estimated curve with the number of grid points (Ngrid) being 100, 200 and 400, respectively. The plots shows nonlinear between the increment and the DMOT. From these plots, the estimated curves are almost the same, since the underline estimate is relatively smooth. In practice, we

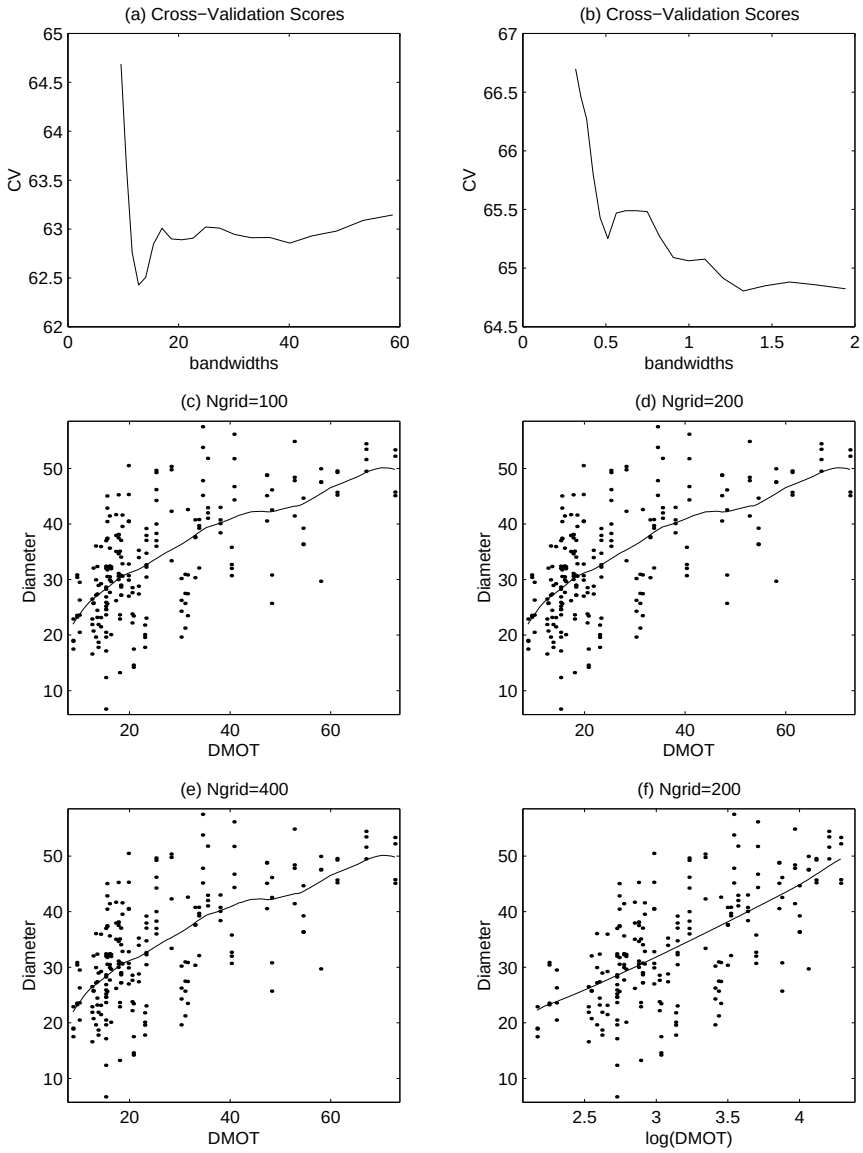

Fig. 1.10 Estimated regression functions. (a) and (b) are plots of cross-validation scores for increment of diameter versus depth of mottling (DMOT) and for versus log(DMOT), respectively. (c)—(e) are estimated regression function curves $E(\text{increment}|\text{DMOT})$ with scatter plot of data, corresponding to the number of grid points 100, 200 and 400, respectively. (h) is the estimated regression curve $E(\text{increment}|\log(\text{DMOT}))$.

recommend using 100 or 200 grid points to evaluate estimated regression functions.

Now we take the natural logarithm of DMOT as the X-variable, and then use the cross-validation method to choose a bandwidth. The CV scores are depicted in Figure 1.10 (b). This yields $\widehat{h}_{CV} = 1.3271$. The estimated curve is depicted in Figure 1.10 (h). Figure 1.10 (h) shows that increment of diameter growth versus log(DMOT) is nearly linear. For such an implementation, it spent about 2 seconds (using MATLAB on PC Pentium II 450MHz) to compute the estimated function over 200 grid points, including bandwidth

selection using the cross-validation method.

Direct implementation of local polynomial regression for a large data set needs a considerable amount of computation. Fast computation algorithms have been proposed in Fan and Marron (1994). Many computer codes are available through internet. For example, S-plus codes can be downloaded from Matt Wand's homepage at

http://www.biostat.harvard.edu/ mwand/software.html

while Matlab codes can be downloaded from James S. Marron's homepage at

http://www.stat.unc.edu/faculty/marron/marron_software.html

or through the authors. These codes can be easily implemented by directly plugging-in data. There is also a procedure of kernel smoothing in the latest version of SAS.

1.4 LOCAL LIKELIHOOD AND LOCAL PARTIAL LIKELIHOOD

The local likelihood approach was first proposed by Tibshirani and Hastie (1987) based on the running line smoother. As an extension of the local likelihood approach, local quasi-likelihood estimation using local constant fits, was considered by Severini and Staniwalis (1994). Fan, Heckman and Wand (1995) investigated the asymptotic properties of the local quasi-likelihood method using local polynomial modeling. Fan, Farmen and Gijbels (1998) addressed the issue of bandwidth selection, bias and variance assessment and constructed confidence intervals in local maximum likelihood estimation. Fan and Chen (1999) proposed one-step local quasi-likelihood estimation, and demonstrated that the one-step local quasi-likelihood estimator performs as well as the maximum local quasi-likelihood estimator using the ideal optimal bandwidth. Fan, Gijbels and King (1998) extended the idea of the local likelihood approach to local partial likelihood in the context of censored survival data analysis, such as Cox's regression model. The ideas in this section are motivated from Fan, Heckman and Wand (1995) and Fan, Gijbels and King (1998). Carroll, Ruppert and Wand (1998) extend the idea further to the likelihood equations.

1.4.1 A. Generalized Linear Models

Generalized linear models introduced by Nelder and Wedderburn (1972) extend the scope of the traditional least squares fitting of linear models. The relationship between a response variable and a set of covariates is modeled as a linear fit to the transformed conditional mean. A comprehensive account of generalized linear models can be found in McMullagh and Nelder (1989). Suppose that we have n independent observations $(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)$ of random vector $(\mathbf{X}, Y)$, where $\mathbf{X}$ is a d-dimensional real vector of covariates, and Y is a scalar response variable. The conditional density of Y given covariate $\mathbf{X} = \mathbf{x}$ belongs to the canonical exponential family:

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = \exp\{[\theta(\mathbf{x})y - b\{\theta(\mathbf{x})\}]/a(\phi) + c(y, \phi)\} \tag{4.11}$$

for known functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot,\cdot)$. In parametric generalized linear models it is usual to model a transformation of the regression function $m(\mathbf{x}) = E(Y|\mathbf{X}=\mathbf{x})$ as linear, that is

$$\eta(\mathbf{x}) = g\{m(\mathbf{x})\} = \mathbf{x}^T\beta,$$

and g is a known *link* function. If $g = (b')^{-1}$, then g is called the canonical link because it transform the regression function into the canonical parameter: $(b')^{-1}\{m(\mathbf{x})\} = \theta(\mathbf{x})$.

Here are a few examples that illustrate the model (4.11). The first example is that the conditional distribution of Y given $\mathbf{X}=\mathbf{x}$ has a normal distribution with mean $m(\mathbf{x})$ and variance σ^2. The normal density can be rewritten as

$$f_{Y|\mathbf{X}} = \exp\left\{\frac{m(\mathbf{x})y - m^2(\mathbf{x})/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \log(\sqrt{2\pi\sigma^2})\right\}.$$

It can be easily seen that

$$\phi = \sigma^2,\ a(\phi) = \phi,\ b(m) = m^2/2$$

and

$$c(y,\phi) = -y^2/(2\phi) - \log(\sqrt{2\pi\phi}).$$

The canonical link function is the identity link $g(t) = t$. This model is useful for a continuous response with homoscedastic errors.

Suppose that the conditional distributions of Y given $\mathbf{X}=\mathbf{x}$ is a Bernoulli distribution with the probability of success $p(\mathbf{x})$, in which case it can be seen that

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = \exp\left(y\log[p(\mathbf{x})/\{1-p(\mathbf{x})\}] + \log\{1-p(\mathbf{x})\}\right).$$

The canonical parameter in this example is $\theta(\mathbf{x}) = \text{logit}\{p(\mathbf{x})\}$, and the logit function is the canonical link.

Under model (4.11), it can be easily shown that the conditional mean and conditional variance are given respectively by $m(\mathbf{x}) = E(Y|\mathbf{X}=\mathbf{x}) = b'\{\theta(\mathbf{x})\}$, and $\text{var}(Y|\mathbf{X}=\mathbf{x}) = a(\phi)b''\{\theta(\mathbf{x})\}$. Hence,

$$\theta(\mathbf{x}) = (b')^{-1}\{m(\mathbf{x})\}.$$

Using the definition of $\eta(\cdot)$, we have

$$\theta(\mathbf{x}) = (b')^{-1}\{g^{-1}[\eta(\mathbf{x})]\}. \tag{4.12}$$

Since our primary interest is to estimate the mean function, without loss of generality, the factors related to the dispersion parameter ϕ are omitted. This leads to the following conditional log-likelihood function

$$\ell\{\theta, y\} = \theta(\mathbf{x})y - b\{\theta(\mathbf{x})\}.$$

By (4.12), the above log-likelihood can be expressed as

$$\ell\{\theta, y\} = \left[y(b')^{-1} \circ g^{-1}(\eta(\mathbf{x})) - b\{(b')^{-1} \circ g^{-1}(\eta(\mathbf{x}))\}\right], \tag{4.13}$$

where $\circ$ denotes composition. In particular, when g is the canonical link,

$$\ell\{\theta, y\} = \eta(\mathbf{x})y - b\{\eta(\mathbf{x})\}.$$

B. Local likelihood estimate

It has been of interest to adapt these models to situations where the functional form for the dependence of $g(m(\mathbf{x}))$ on $\mathbf{x}$ is unknown. In what follows, the covariate $\mathbf{X}$ is assumed to be a scalar random variable. If $\eta(x)$ is a smooth function of x, then for X_i close enough to a given point x_0,

$$\eta(X_i) \approx \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \equiv \mathbf{X}_i^T \beta, \tag{4.14}$$

where $\mathbf{X}_i = (1, (X_i - x_0), \cdots, (X_i - x_0)^p)^T$ and $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^T$. Intuitively data points close to x_0 have more information about $\eta(x_0)$ than those away from x_0. Therefore, by (4.13), the local log-likelihood function based on the random sample $\{(X_i, Y_i)\}_{i=1}^n$ is

$$\ell(\beta) = \sum_{i=1}^{n} \left[Y_i (b')^{-1} \circ g^{-1}(\mathbf{X}_i^T \beta) - b\{(b')^{-1} \circ g^{-1}(\mathbf{X}_i \beta)\} \right] K_h(X_i - x_0). \tag{4.15}$$

Define the local maximum likelihood estimator of β to be

$$\widehat{\beta} = \text{argmax}_{\beta \in R^{p+1}} \ell(\beta).$$

Thus $\eta(x_0)$ and the ν-th derivative of $\eta(x_0)$ can be estimated by

$$\widehat{\eta}(x_0) = \widehat{\beta} \quad \text{and} \quad \widehat{\eta}^{(\nu)} = \nu! \widehat{\beta}_\nu$$

respectively, assuming that η has p derivatives. When the canonical like $g = (b')^{-1}$ is used, (4.15) becomes

$$\ell(\beta) = \sum_{i=1}^{n} \left[Y_i (\mathbf{X}_i^T \beta) - b(\mathbf{X}_i \beta) \right] K_h(X_i - x_0).$$

The log-likelihood function (4.15) is really a weighted log-likelihood and hence can be computed by using the existing software. In fact, suppose that we want to estimate $\widehat{\eta}(\cdot)$ in a given interval. Take a grid of points (200, say) in that interval. For each given grid point x_0, (4.15) can be maximized by using existing software packages such as SAS and Splus that contains the parametric Glim function. The whole estimated function is obtained by plotting the estimates obtained at grid points.

The choice of the link function g is not as crucial as for parametric generalized linear models, because the fitting is localized. Indeed it is conceivable to dispense with the link function and just estimate $m(x)$ directly. But there are several drawbacks to having the link equal to the identity. An identity link may yield a local likelihood that is not convex, allowing for the possibility of multiple maxima, inconsistency and computational problems. Use of the canonical link guarantees convexity. Furthermore the canonical link ensures that the final estimate has the correct range. For example, in the logistic regression context using the logit link leads to an estimate that is always a probability whereas using the identity link does not have. A final reason for preferring the canonical link is that the estimate of $m(x)$ approaches the usual parametric estimate as the bandwidth becomes large. This can be useful as a diagnostic tool. See Fan, Heckman and Wand (1995) for details.

We now illustrate the local likelihood approach via analyzing the data set: *Burns data*, collected by General

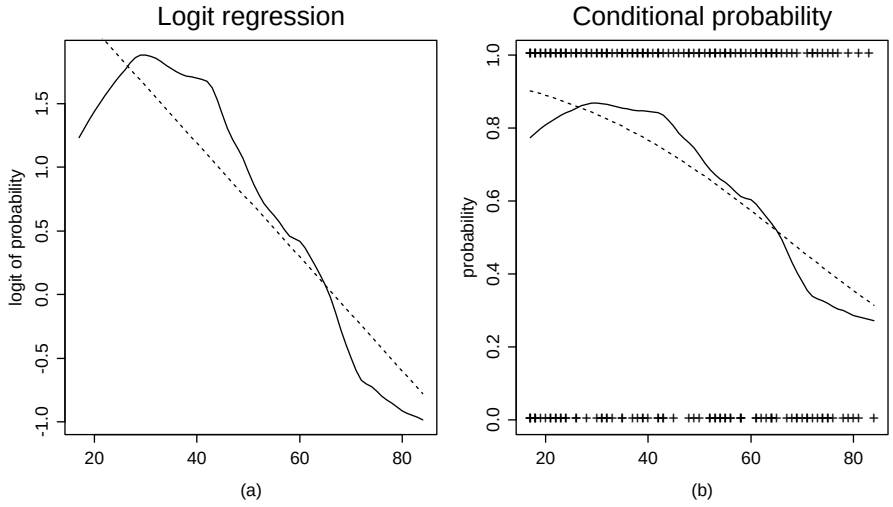

Fig. 1.11 Illustration of local likelihood approach for the burn data. (a) Estimated logit transform of the conditional probability. (b) Estimated conditional probability. Solid curve — local modeling with about 40% of the data points in each local neighborhood; dashed curve — global parametric logit linear model. Taken from Fan and Gijbels (2000).

Hospital Burn Center at the University of South California. It is of interest to estimate the probability of surviving given the age of victims. Local likelihood estimate was computed over a grid of points with bandwidth 0.4 multiplying the range of X, and the estimated curves are depicted in Figure 1.11. 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 1.11 suggest that the conditional probability function is unimodal, which is reasonable in the current context.

1.4.2 Local partial likelihood estimate

In this section, we apply the local likelihood techniques to survival data analysis. Let T, C and X be respectively the survival time, the censoring time and their associated covariates. Correspondingly, let $Z = \min\{T, C\}$ be the observed time and $\delta = I(T \leq C)$ be the censoring indicator. It is assumed that T and C are conditionally independent given X and that the censoring mechanism is noninformative. Suppose that $\{(X_i, Z_i, \delta_i) : i = 1, \cdots, n\}$ are an iid sample from the population (X, Z, δ). For a thorough introduction to survival analysis, see books by Fleming and Harrington (1991) and Andersen, Borgan, Gill and Keiding (1993).

Let $h(t|x)$ be the conditional hazard rate function. The proportional hazards model assumes that

$$h(t|x) = h_0(t) \exp\{\theta(x)\}. \tag{4.16}$$

This model indicates that the covariate x inflates or deflates the hazard risk by a factor of $\exp\{\theta(x)\}$. The function $\theta(x)$ is called a hazard regression function, and characterizes the risk contribution of the covariate at value x. See Cox (1972) for proportional hazard models with time-dependent covariates.

In the parametric model, a linear form $\theta(x) = \beta x$ is imposed on the hazard regression function. The local modeling methodology aims at removing this restriction and exploring possible nonlinearity, and is applicable to any smooth hazard regression function. For simplicity of discussion, we focus on the univariate cases. For multivariate settings, a dimensionality reduction technique such as additive models should be used. 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 $R_j = \{i : Z_i \geq t_j^o\}$, the risk set at time instantaneously before t_j^o. Then, the log-partial likelihood in our context is given by

$$\sum_{j=1}^{N} \Big[\theta(X_{(j)}) - \log\Big(\sum_{k \in R_j} \exp\{\theta(X_k)\} \Big) \Big]. \tag{4.17}$$

See Cox (1975), Fleming and Harriton (1991) and Fan and Gijbels (1996). Substituting the parametric form of $\theta(\cdot)$ into (4.17) 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

$$\theta(x) \approx \beta_0 + \cdots + \beta_p (x - x_0)^p, \tag{4.18}$$

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

$$\sum_{j=1}^{N} K_h(X_{(j)} - x_0) \Big[\mathbf{X}_{(j)}^T \beta - \log\Big\{ \sum_{k \in R_j} \exp(\mathbf{X}_k^T \beta) K_h(X_k - x_0) \Big\} \Big]. \tag{4.19}$$

See Fan, Gijbels and King (1997) for a derivation of the local partial likelihood (4.19). When the kernel function is uniform and the bandwidth is of the nearest neighbor type, the local likelihood (4.19) 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 (4.19) 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 (4.19). 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

$$\widehat{\theta}(x) = \int_0^x \widehat{\theta}'(t) dt, \tag{4.20}$$

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,

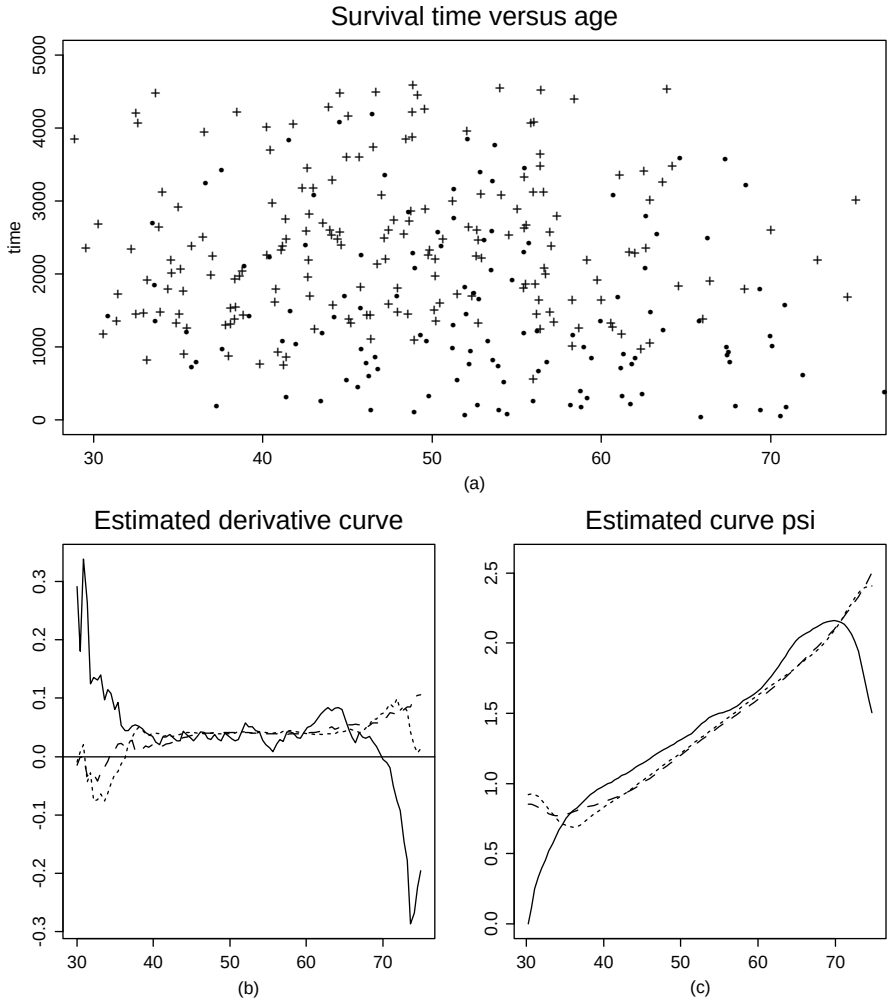

Fig. 1.12 Local partial likelihood estimation of the hazard regression function. (a) Observed time versus age with '+' indicating censored observations. (b) Estimated derivative function $\theta'(\cdot)$. (c) Estimated hazard regression function $\theta(\cdot)$; solid curve — bandwidth = 10; short-dashed curve — bandwidth = 20; long-dashed curve — bandwidth = 30. From Fan and Gijbels (2000).

$\widehat{\theta}(x_i)$ 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 $\widehat{\beta}_{1,j} = \widehat{\beta}_1(x_j)$. The coefficients can simply be computed by using existing software packages for parametric Cox's proportional hazards model.

We conclude this section with an analysis of the Primary Biliary Cirrhosis (PBC) data set, which 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 1.12 (a). The local partial likelihood method (4.19) 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 1.12 (c) and (b). Note that since the hazard regression function is only identifiable within a constant, the curves in Figure 1.12 (c) are normalized to have the same average height so that they can be better compared. Figure 1.12 (c) reveals the fact that it is reasonable to model linearly the hazard regression function of covariate age.

1.5 NONPARAMETRIC GOODNESS OF FIT TESTS

Nonparametric goodness of fit test has received increasing attention recently. A motivating and simple example is to consider a simple nonparametric regression model. Suppose that $(X_1, Y_1), \cdots, (X_n, Y_n)$ are a random sample from the nonparametric regression model

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

with $E(\varepsilon_i|X_i) = 0$ and $\text{var}(\varepsilon_i|X_i) = \sigma^2$. Of interest is to test the hypothesis

$$H_0 : m(x) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x^p \quad \text{versus} \quad H_1 : m(x) \neq \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x^p.$$

This testing problem is well known as test of linearity in the context of model diagnostic where the question arises whether a family of parametric models fit adequately the data. It is natural to use the nonparametric model as an alternative hypothesis. On the other hand, it is known that nonparametric regression may yield a complicated model. Thus after fitting a data set by a nonparametric model, we may check whether the data can be fitted by a less complicated parametric model. This leads to a nonparametric goodness of fit test. Hart (1997) gives a comprehensive study and presents many examples on this topic. Fan (1996) and Fan and Huang (2001) proposed some goodness of fit tests for various parametric models and nonparametric models. Fan, Zhang and Zhang (2001) proposed generalized likelihood ratio tests and established a general framework for nonparametric smoothing tests. For related literature, see Azzalini, Bowman and Härdle (1989), Eubank and Hart (1992), Eubank and LaRiccia (1992), Azzalini and Bowman (1993), Härdle and Mammen (1993) Inglot, Kallenberg and Ledwina (1994), Spokoiny (1996), Kallenberg and Ledwina (1997), Aerts, Claeskens and Hart (1999), among others. In this section, we illustrate the idea of nonparametric likelihood ratio test by generalized varying coefficient models. Some material of this section was extracted from Cai, Fan and Li (2000), referred as CFL. See the paper for details.

1.5.1 Generalized varying coefficient models

A generalized varying-coefficient model has the form

$$\eta(\mathbf{u}, \mathbf{x}) = g\{m(\mathbf{u}, \mathbf{x})\} = \sum_{j=1}^{p} a_j(\mathbf{u})\, x_j \tag{5.1}$$

for some given link function $g(\cdot)$, where $\mathbf{x} = (x_1, \ldots, x_p)^T$, and $m(\mathbf{u}, \mathbf{x})$ is the mean regression function of the response variable Y given the covariates $\mathbf{U} = \mathbf{u}$ and $\mathbf{X} = \mathbf{x}$ with $\mathbf{X} = (X_1, \ldots, X_p)^T$. Clearly, model (5.1) includes both the parametric generalized linear model (McCullagh and Nelder 1989) and the generalized partially linear model (Chen 1988; Speckman 1988; Green and Silverman 1994; Carroll, Fan, Gijbels and Wand 1997). An advantage of model (5.1) is that by allowing the coefficients $\{a_j(\cdot)\}$ to depend on $\mathbf{U}$, the modeling bias can be reduced significantly and the "curse of dimensionality" is avoided.

In this section, we focus on the cases in which the response is discrete. For continuous responses, many works have been done. In the least-squares setting, model (5.1) with the identity link was introduced by Cleveland, Grosse and Shyu (1992) and extended by Hastie and Tibshirani (1993) to various aspects. Varying-coefficient models are a simple and useful extension of classical generalized linear models. They are particularly appealing in longitudinal studies where they allow one to explore the extent to which covariates affect responses changing over time. See Hoover *et al.* (1998), Brumback and Rice (1998) and Fan and Zhang (2000) for details on novel applications of the varying-coefficient models to longitudinal data. Also see the chapter written by Colin O. Wu in this book and references therein. For nonlinear time series applications, see Chen and Tsay (1993) and Cai, Fan and Yao (2000) for statistical inferences based on the functional-coefficient autoregressive models. Kauermann and Tutz (1999) used varying coefficient models for model disgnostics.

1.5.2 Estimation Procedures

For simplicity, we consider the important case that $\mathbf{u}$ is one-dimensional. Extension to multivariate $\mathbf{u}$ involves no fundamentally new ideas. However, implementations with $\mathbf{u}$ having more than two dimensions may have some difficulties due to the "curse of dimensionality."

In this section, it is assumed that the conditional log-likelihood function $\ell(v, y)$ is known and linear in y for fixed v. This assumption is satisfied for the canonical exponential family, which is the focus of this section. The methods, introduced in this section, are directly applicable to situations in which one can not specify fully the conditional log-likelihood function $\ell(v, y)$, but can model the relationship between the mean and variance by $\mathrm{Var}(Y \mid \mathbf{U} = \mathbf{u}, \mathbf{X} = \mathbf{x}) = \sigma^2 \mathbf{V}\{\mathbf{m}(\mathbf{u}, \mathbf{x})\}$ for a known variance function $V(\cdot)$ and unknown σ. In this case, one needs only to replace the log-likelihood function $\ell(v, y)$ by the quasi-likelihood function $Q(\cdot, \cdot)$, defined by $\frac{\partial}{\partial \mu} Q(\mu, y) = \frac{y-\mu}{V(\mu)}$.

A. Local MLE

Local linear modeling will be used here, though general local polynomial methods are also applicable. Suppose that $a_j(\cdot)$ has a continuous second derivative. For each given point u_0, $a_j(u)$ can be approximated locally by a linear function $a_j(u) \approx a_j + b_j\,(u - u_0)$ for u in a neighborhood of u_0. Based on a random sample

$\{(U_i, \mathbf{X}_i, Y_i)\}_{i=1}^n$, one may use the following local likelihood method to estimate the coefficient functions

$$\ell_n(\mathbf{a}, \mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} \ell \left[g^{-1} \left\{ \sum_{j=1}^{p} (a_j + b_j(U_i - u_0)) X_{ij} \right\}, Y_i \right] K_h(U_i - u_0), \tag{5.2}$$

where $\mathbf{a} = (a_1, \ldots, a_p)^T$ and $\mathbf{b} = (b_1, \ldots, b_p)^T$. Note that a_j and b_j depend on u_0, and so does $\ell_n(\cdot, \cdot)$. Maximizing the local likelihood function $\ell_n(\mathbf{a}, \mathbf{b})$ results in estimates $\widehat{\mathbf{a}}(u_0)$ and $\widehat{\mathbf{b}}(u_0)$. The components in $\widehat{\mathbf{a}}(u_0)$ provide an estimate of $a_1(u_0), \ldots, a_p(u_0)$. For simplicity of notation, let $\beta = \beta(u_0) = (a_1, \ldots, a_p, b_1, \ldots, b_p)^T$, and write the local likelihood function (5.2) as $\ell_n(\beta)$. Likewise, the local MLE is denoted by $\widehat{\beta}_{\mathrm{MLE}} = \widehat{\beta}_{\mathrm{MLE}}(u_0)$. The sampling properties have been established in CFL.

B. One-step local MLE

Computation for the above local MLE is expensive. We have to maximize the local likelihood (5.2) for usually hundreds of distinct values of u_0, with each maximization requiring an iterative algorithm, in order to obtain the estimated functions $\{\widehat{a}_j(\cdot)\}$. To alleviate this expense, we replace an iterative local MLE by the one-step estimator, which has been frequently used in parametric models (Bickel 1975; Lehmann 1983). The one-step local MLE does not lose any statistical efficiency provided that the initial estimator is good enough. See CFL for theoretic insights.

Let $\ell_n'(\beta)$ and $\ell_n''(\beta)$ be the gradient and Hessian matrix of the local log-likelihood $\ell_n(\beta)$. Given an initial estimator $\widehat{\beta}_0 = \widehat{\beta}_0(u_0) = \left(\widehat{\mathbf{a}}(u_0)^T, \widehat{\mathbf{b}}(u_0)^T\right)^T$, one-step of the Newton-Raphson algorithm updates its solution by

$$\widehat{\beta}_{\mathrm{OS}} = \widehat{\beta}_0 - \left\{ \ell_n'' \left(\widehat{\beta}_0 \right) \right\}^{-1} \ell_n' \left(\widehat{\beta}_0 \right), \tag{5.3}$$

thus featuring the computational expediency of least-squares local polynomial fitting. Furthermore, the sandwich formula can be used as an estimate for standard errors of the resulting estimate

$$\widehat{\mathrm{cov}}(\widehat{\beta}_{\mathrm{OS}}) = \left\{ \ell_n'' \left(\widehat{\beta}_0 \right) \right\}^{-1} \widehat{\mathrm{cov}} \{ \ell_n' \left(\widehat{\beta}_0 \right) \} \left\{ \ell_n'' \left(\widehat{\beta}_0 \right) \right\}^{-1}.$$

This formula has been tested in CFL to be accuracy enough for most of practical purpose.

In univariate generalized linear models, Fan and Chen (1999) carefully studied properties of the local one-step estimator. In that setting, the least-squares estimate serves a natural candidate as an initial estimator. However, in the multivariate setting, it is not clear how an initial estimator can be constructed. The following is proposed in CFL. Suppose that we wish to evaluate the functions $\widehat{\mathbf{a}}(\cdot)$ at grid points u_j, $j = 1, \ldots, n_{\mathrm{grid}}$. Our idea of finding initial estimators is as follows. Take a point u_{i_0}, usually the center of the grid points. Compute the local MLE $\widehat{\beta}_{\mathrm{MLE}}(u_{i_0})$. Use this estimate as the initial estimate for the point u_{i_0+1} and apply (5.3) to obtain $\widehat{\beta}_{\mathrm{OS}}(u_{i_0+1})$. Now, use $\widehat{\beta}_{\mathrm{OS}}(u_{i_0+1})$ as the initial estimate at the point u_{i_0+2} and apply (5.3) to obtain $\widehat{\beta}_{\mathrm{OS}}(u_{i_0+2})$ and so on. Likewise, we can compute $\widehat{\beta}_{\mathrm{OS}}(u_{i_0-1})$, $\widehat{\beta}_{\mathrm{OS}}(u_{i_0-2})$, etc. In this way, we obtain our estimates at all grid points.

A refine alternative of the above proposal is to calculate a fresh local MLE as a new initial value after iterating along the grid points for a while. For example, if we wish to evaluate the functions at 200 grid points and are willing to compute the local maximum likelihood at five distinct points. A sensible placement of these points is u_{20}, u_{60}, u_{100}, u_{140} and u_{180}. Use for example $\widehat{\beta}_{\mathrm{MLE}}(u_{60})$ along with the idea in

the last paragraph to compute $\widehat{\beta}_{\rm OS}(u_i)$ for $i = 40, \ldots, 79$, and use $\widehat{\beta}_{\rm MLE}\ (u_{100})$ to compute $\widehat{\beta}_{\rm OS}(u_i)$ for $i = 80, \ldots, 119$, and so on.

Note that $\ell_n''(\widehat{\beta}_0)$ can be nearly singular for certain u_0, due to possible data sparsity in certain local regions. Seifert and Gasser (1996) and Fan and Chen (1999) explored the use of the ridge regression as an approach to handling such problems in the univariate setting. See CFL for details.

1.5.3 Hypothesis testing

When fitting a varying-coefficient model, it is natural to ask whether the coefficient functions are actually varying or whether any particular covariate is significant in the model. For simplicity of description, we only consider the first hypothesis testing problem

$$H_0 : a_1(u) \equiv a_1, \cdots, a_p(u) \equiv a_p, \tag{5.4}$$

though the technique also applies to other testing problems. A useful procedure is based on the nonparametric likelihood ratio test statistic

$$T = 2\{\ell(H_1) - \ell(H_0)\}, \tag{5.5}$$

where $\ell(H_0)$ and $\ell(H_1)$ are respectively the log-likelihood functions computed under the null and alternative hypotheses. Note that the normalization constant in (5.5) does not change the testing procedure. However, in order for it to possess a χ^2 distribution, it needs to be normalized as (see Fan, Zhang and Zhang, 2001)

$$T_K = r_K\{\ell(H_1) - \ell(H_0)\}, \tag{5.6}$$

where

$$r_K = \frac{K(0) - \frac{1}{2}\int K^2(t)dt}{\int (K(t) - \frac{1}{2}K * K(t))^2 dt}.$$

The following table gives the value of r_K for a few commonly used kernels.

Table 1.2 Normalization constant r_K

Kernel	Uniform	Epanechnikov	Biweight	Triweight	Gaussian
r_K	1.2000	2.1153	2.3061	2.3797	2.5375

For parametric models, it is well known that the likelihood ratio statistic follows asymptotically a χ^2-distribution. The asymptotic null distribution is independent of nuisance parameters under the null hypothesis. This is the Wilks type of phenomenon. Fan, Zhang and Zhang (2001) has shown the Wilks phenomenon still holds for the nonparametric likelihood ratio tests. Furthermore, they showed that the null distribution of the nonparametric likelihood ratio test is a χ^2-distribution in some sense and does not depend on the values of $a_1, \cdots, a_p$. Thus one may use the following *conditional bootstrap* to construct the null distribution of T_K and hence the P-value. Let $\{\widehat{a}_j\}$ be the MLE under the null hypothesis. Given the covariates $(U_i, \mathbf{X}_i)$, generate a bootstrap sample Y_i^* from the given distribution of Y with the estimated linear predic-

tor $\widehat{\eta}(U_i, \mathbf{X}_i) = \sum_{j=1}^{p} \widehat{a}_j X_{ij}$ and compute the test statistic T_K^* in (5.5). Use the distribution of T_K^* as an approximation to the distribution of T_K.

Note that the above conditional bootstrap method applies readily to setting without presence of dispersion parameter, such as the Poisson and Bernoulli distributions. It is really a simulation approximation to the conditional distribution of T_K given observed covariates under the particular null hypothesis: $H_0 : a_j(u) = \widehat{a}_j$ $(j = 1, \cdots, p)$. As pointed out above, this approximation is valid under both H_0 and H_1 as the null distribution does not asymptotically depend on the values of $\{a_j\}$. In the case where model (5.1) involves a dispersion parameter (e.g., the Gaussian model), the dispersion parameter should be estimated based on the residuals from the *alternative* hypothesis.

It is also of interest to investigate whether some covariates are significant. For example, we want to check whether the covariate X_p can be excluded from the model. This is equivalent to testing the hypothesis $H_0 : a_p(\cdot) = 0$, the above conditional bootstrap idea can be employed to obtain the null distribution of T_K under the model (5.1) and the generalized likelihood ratio statistics continue to apply. In this case, the data should be generated from the mean function $g\{m(\mathbf{u}, \mathbf{x})\} = \sum_{j=1}^{p-1} \widehat{a}_j(\mathbf{u})\, x_j$, where $\widehat{a}_j(\cdot)$ is an estimate under the alternative hypothesis.

1.5.4 An Application

We conclude this section via illustrating the proposed methodology to analyze the *Burn Data* set. The binary response variable Y is 1 for those victims who survived their burns and 0 otherwise, and covariates X_1=*age*, X_2=*sex*, $X_3 = \log(\text{burn area}+1)$ and binary variable X_4 =*Oxygen* (0 if oxygen supply is normal, 1 otherwise) are considered. Of interest is to study how burn areas and the other variables affect the survival probabilities for victims at different age groups. This naturally leads to the following varying-coefficient model

$$\text{logit}\{p(x_1, x_2, x_3, x_4)\} = a_1(x_1) + a_2(x_1)\, x_2 + a_3(x_1)\, x_3 + a_4(x_1)\, x_4. \tag{5.7}$$

Figure 1.13 presents the estimated coefficients for model (5.7) via the one-step approach with bandwidth $h = 65.7882$, selected by a cross-validation method. See CFL for details.

A natural question arises whether the coefficients in (5.7) are actually varying. To see this, we consider the parametric logistic regression model

$$\text{logit}\{p(x_1, x_2, x_3, x_4)\} = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \beta_3\, x_3 + \beta_4\, x_4 \tag{5.8}$$

as the null model. As a result, the MLE of $(\beta_0, \cdots, \beta_4)$ in model (5.8) and its standard deviation are (23.2213, −6.1485, −0.4661, −2.4496, −0.9683) and (1.9180, 0.6647, 0.2825, 0.2206, 0.2900), respectively. The likelihood ratio test T_K is 58.1284 with p-value 0.000, based on 1000 bootstrap samples (the sample mean and variance of T_K^* are 6.3201 and 11.98023, respectively). This implies that the varying-coefficient logistic regression model fits the data much better than the parametric fit. It also allows us to examine the extent to which the regression coefficients vary over different ages. The estimated density of T_K^* is depicted in Figure 1.14, from which we can seen that the null distribution is well approximated by a χ^2 distribution with 6.5 degrees of freedom (a gamma distribution).

To examine whether there is any gender gap for different age groups or if the variable X_4 affects the

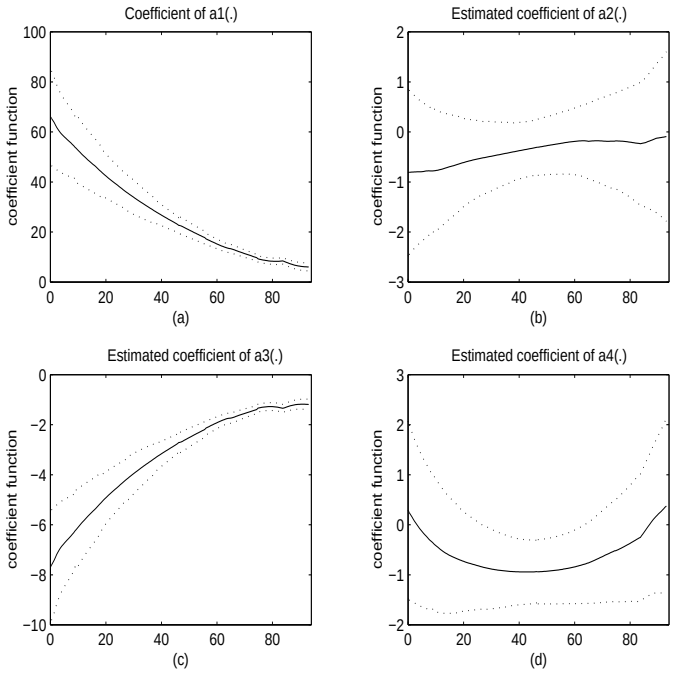

Fig. 1.13 *The estimated coefficient functions (the solid curves) via the one-step approach with bandwidth chosen by the CV. The dot curves are the estimated functions plus/minus twice estimated standard errors. Adapted from Cai, Fan and Li (2000).*

survival probabilities for different age of burn victims, we consider testing the null hypothesis H_0 : both $a_2(\cdot)$ and $a_4(\cdot)$ are constant under model (5.7). The corresponding test statistic T_K is 3.4567 with p-value 0.7050, based on 1000 bootstrap samples. This in turn suggests that the coefficient functions $a_2(\cdot)$ and $a_4(\cdot)$ are independent of age and indicates that there are no gender differences for different age groups.

Finally, we examine whether both covariates *sex* and *Oxygen* are statistically significant in model (5.7). The likelihood ratio test for this problem is $T_K = 11.9256$ with p-value 0.0860, based on 1000 bootstrap samples (the sample mean and variance of T_K^* are 5.5915 and 10.9211, respectively). Both covariates *sex* and *Oxygen* are not significant at level 0.05. This suggests that gender and oxygen do not play a significant role in determining the survival probability of a victim.

1.6 OTHER APPLICATIONS

There are many other applications of local modeling methods. This section briefly introduces some of them and gives some relevant references for those who wish for more details. Suppose that $(X_1, Y_1), \cdots, (X_n, Y_n)$ are a random sample from a population (X, Y). We are interested in estimating a population parameter function θ. The function $\theta(\cdot)$ can be, for example, the conditional mean function $E(Y|X)$ and the conditional quantile function. In parametric settings, we model $\theta(x)$ using a parametric family $\theta(x) = g(x; \beta)$. To get

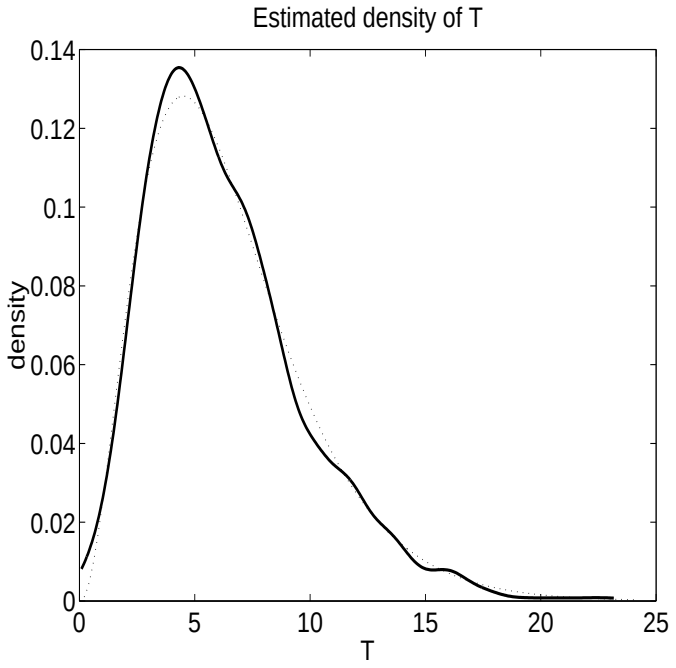

Fig. 1.14 *The estimated density of T_K by Monte Carlo simulation. The solid curve is the estimated density, and the dashed curve stands for the density of chi-squared distribution (gamma distribution) with 6.5 degrees of freedom.*

an estimator of β, we optimize (either minimize or maximize) an objective function

$$L(\beta) = \sum_{i=1}^{n} \ell\{X_i, Y_i, g(X_i, \beta)\}. \tag{6.1}$$

Here ℓ is a discrepancy loss function or the log-likelihood function of an individual observation. For example, the L_2-loss function leads to a least squares estimate, while the L_1-loss function corresponds to a robust linear regression.

The local modeling method can be used to relax the global parametric model assumption and to significantly reduce the modeling bias. For a given point x_0, we replace the objective function by its local version

$$L\{\beta(x_0)\} = \sum_{i=1}^{n} \ell\{X_i, Y_i, g(X_i, \beta(x_0))\} K_h(X_i - x_0). \tag{6.2}$$

Optimizing (6.2) yields an estimate $\widehat{\beta}(x_0)$, just like the local likelihood estimate discussed in the last section. Thus an estimate of the function $\theta(\cdot)$ by $\widehat{\theta}(x_0) = g\{x_0; \widehat{\beta}(x_0)\}$. Since the local estimate $\widehat{\beta}(x_0)$ optimizes (6.2), the estimate $g\{x_0, \widehat{\beta}(x_0)\}$ should converge to its population version. Therefore the estimate $\widehat{\theta}(x_0)$ is a consistent estimator of the function $\theta(x_0)$ if $h \to 0$ in such a way that $nh \to \infty$.

For a given x_0, by Taylor's expansion, we can parametrize the function in a local neighborhood of x_0 as

$$g(x; \beta) = \beta_0(x_0) + \beta_1(x_0)(x - x_0) + \cdots + \beta_p(x_0)(x - x_0)^p. \tag{6.3}$$

With suppressing the dependence of β's on x_0, (6.2) can be rewritten as

$$L\{\beta(x_0)\} = \sum_{i=1}^{n} \ell\{X_i, Y_i, \beta_0 + \beta_1(X_i - x_0) + \cdots + \beta_p(X_i - x_0)^p\} K_h(X_i - x_0). \tag{6.4}$$

Let $\widehat{\beta}_j$ $(j = 0, 1, \cdots, p)$ optimize (6.4). Then as in last section,

$$\widehat{\theta}(x_0) = \widehat{\beta}_0$$

and

$$\widehat{\theta}_\nu(x_0) = \nu!\widehat{\beta}_\nu, \quad \nu = 1, \cdots, p$$

estimates the ν-th derivative of the function $\theta(x)$ at $x = x_0$.

It is clear that local polynomial regression and local likelihood approach are special cases hereof. An extension of the ideas for estimating bias and variance can be found in Fan, Farmen and Gijbels (1998), in which 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 (1998) and the kernel generalized estimating equation (GEE) proposed by Lin and Carroll (2000).

1.6.1 Estimation of conditonal quantiles and median

In explanatory data analysis, quantiles provide us informative summary of a population. In regression analysis, conditional quantiles have important applications for constructing predictive intervals and detecting heteroscedasticity. When the error distribution is asymmetric, the conditional median regression function is more informative than the conditional mean regression.

Take the loss function in (6.1) to be $\ell(x, y, \theta) = \ell_\alpha(y - \theta)$ with

$$\ell_\alpha(t) = |t| + (2\alpha - 1)t. \tag{6.5}$$

The minimizer of $E\ell(X, Y, \theta)$ in this situation is the conditional α-quantile function $\xi_\alpha(x) = G^{-1}(\alpha|x)$, where $G^{-1}(y|x)$ is the conditional distribution of Y given $X = x$.

Now we apply the local modeling approach to estimate the conditional quantile function. Minimize

$$\sum_{i=1}^{n} \ell_\alpha\{Y_i - \beta_0 - \beta_1(X_i - x_0) - \cdots - \beta_p(X_i - x_0)^p\} K_h(X_i - x_0) \tag{6.6}$$

and the resulting estimator for $\xi_\alpha(x_0)$ is simply $\widehat{\beta}_0$.

Now we apply the proposed approach to the twelve-month Treasury bill data presented in Figure 1.15 (a). Figure 1.15 (b) depicts the estimated conditional median, the conditional 10th percentile and the conditional 90th percentile. The fan shape of the conditional quantiles shows that the variability gets larger as the interest rate gets higher. The intervals sandwiched by 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 $[-0.373\%, 0.363\%]$.

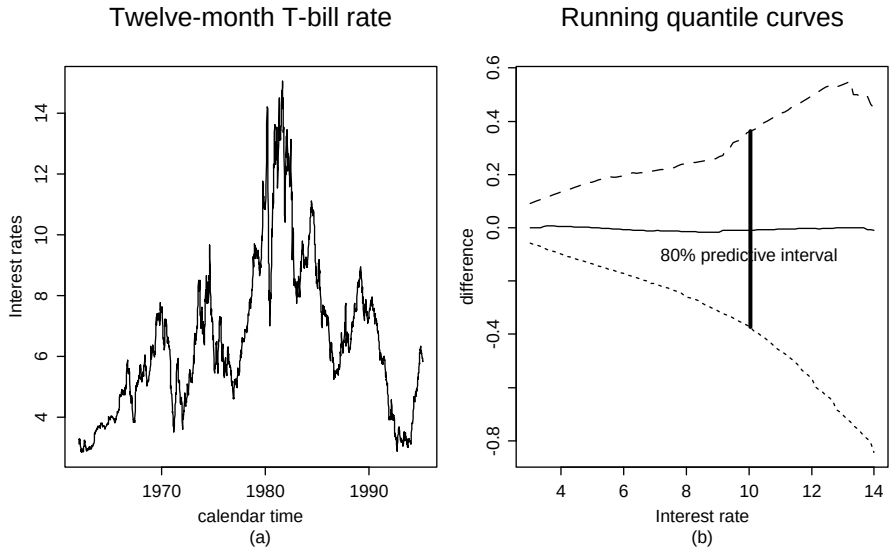

Fig. 1.15 Quantile regression. (a) The yields of twelve-month Treasury bill. (b) Conditional quantiles for the data presented in Figure 1.15 (a): Short-dashed curve — $\alpha = 0.1$, solid curve — $\alpha = 0.5$, long-dashed curve — $\alpha = 0.9$. The vertical bar indicates the 80%-predictive interval at the point $x = 10$. Taken from Fan and Gijbels (2000).

For robust estimation of the regression function, one can simply replace the loss function in (6.5) by an outlier-resistant loss function such as

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

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 $c \geq 0$. 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 on 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).

1.6.2 Estimation of conditional variance

Conditional variance functions have many statistical applications, particularly in finance. Because of their important applications in finance in which data are often dependent, we formulate the problems in stochastic setup.

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 $\sigma^2(x) = \text{var}(Y|X = x)$ be respectively the regression function and the conditional variance function. Our approach is based on the residuals of the local fit. Let $\widehat{m}_{h_1,K}(\cdot)$ be the local fit of $m(\cdot)$ using a kernel K and a bandwidth h_1. Consider the squared residuals

$$\widehat{r}_i = \{Y_i - \widehat{m}_{h_1,K}(X_i)\}^2. \tag{6.7}$$

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 fit on the squared residuals. Let $\widehat{\sigma}^2_{h_2,W}(x)$ be the local 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 (1997), and Fan and Yao (1998) 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. They also obtained the order of bias and variance of the resulting estimators. Their results 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. Thus a data-driven procedure can be established. See Fan and Yao (1998) for details.

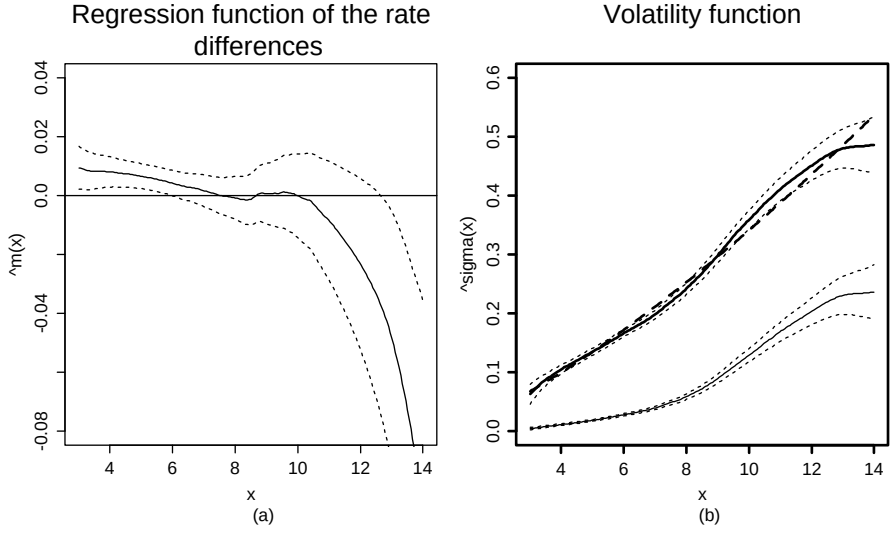

Fig. 1.16 The regression function and the volatility function for the twelve-month Treasury bill data. (a) The estimated mean regression function and, (b) Estimated volatility function (thick curve) and the estimated conditional variance function (thin curve). The two dashed curves around a solid one indicate one standard error above and below the estimated mean regression function. From Fan and Gijbels (2000).

To illustrate the usefulness of the above automatic method, consider the yields of twelve-month Treasury bill. The refined global bandwidth selector of Fan and Gijbels (1995a) and the Epanechnikov kernel. Figure

1.16 (a) gives the estimated mean regression function. The bandwidth $\widehat{h}_1 = 3.99$ was chosen by the software. Figure 1.16 (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 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 1.16 (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.

1.6.3 Estimation of conditional density

It is well known that probability density function is much more informative than the mean and the variance. Similarly, in regression settings, the conditional probability density function provided more information about the population than the conditional regression function. The probability density function plots can show us about the center as well as the spreadness of the population. The shape of conditional probability density function tells us whether it is symmetric. This provides a guidance for us to summarize the population via the conditional mean regression function or the conditional median regression function.

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

$$E\{K_{h_2}(Y - y)|X = x\} \approx g(y|x), \quad \text{as } h_2 \to 0. \tag{6.8}$$

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:

$$\sum_{i=1}^{n}\Big\{K_{h_2}(Y_i - y) - \sum_{j=0}^{p}\beta_j\,(X_i - x)^j\Big\}^2 W_{h_1}(X_i - x), \tag{6.9}$$

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 an estimator of $g^{(\nu)}(y|x) = \frac{\partial^\nu g(y|x)}{\partial x^\nu}$ 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.

To apply the proposed approach of conditional density estimation, we have to choose two bandwidths h_1 and h_2. The method for constructing a data-driven bandwidth for local polynomial regression can be used to compute a bandwidth for h_1, and the method for choosing a bandwidth for kernel density estimation can be employed to find a bandwidth h_2. See Fan, Yao and Tong (1996) for details.

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 functions.

1.6.4 Change point detection

Change point detection is useful in medical monitoring and quality control. For example, when the treatment effects change suddenly without warning or planning, *jump points* arise. The statistical problem can be formulated as follows.

Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be 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 a *change point*.

One may regard the change point as the location where the derivative function $|m'(\cdot)|$ is maximized. Thus, a naive method 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

$$\widetilde{x}_-(h) = \sup_{h_1 \in [h, \eta_n]} \{\widetilde{x}(h_1) - 2h_1\}, \quad \widetilde{x}_+(h) = \inf_{h_1 \in [h, \eta_n]} \{\widetilde{x}(h_1) + 2h_1\}, \tag{6.10}$$

for $h \leq \eta_n$, 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 $\widetilde{x}_-(h) \leq \widetilde{x}_+(h)$. 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 $\widehat{x}_0$.

Müller (1992) proposed an alternative method based on a one-sided kernel approach. The idea can be extended to the local polynomial setting as follows. Denote by K_- a kernel function supported on $[-1, 0]$ and $\widehat{m}_-(x, h)$ 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-hand 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. Müller (1992) and Gijbels, Hall and Kneip (1995) also gave some interesting examples. See the two papers for applications.

REFERENCES

Aerts, M., Claeskens, G. & Hart, J.D. (1999). Testing the fit of a parametric function. *J. Amer. Statist. Assoc.*, **94**, 869–879.

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

Azzalini, A. & Bowman, A.N. (1993). On the use of nonparametric regression for checking linear relationships. *J. Roy. Statist. Soc. Ser.B* **55**, 549-557.

Azzalini, A., Bowman, A.N. & Härdle, W. (1989). On the use of nonparametric regression for model checking. *Biometrika* **76**, 1-11.

Bickel, P.J. (1975), "One-step Huber estimates in linear models," *Journal of the American Statistical Association*, **70**, 428-433.

Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis, The Kernel Approach with S-Plus Illustrations.* Oxford Science Publications, Oxford.

Brumback, B. and Rice, J. (1998), "Smoothing spline models for the analysis of nested and crossed samples of curves," *Journal of the American Statistical Association*, **93**, 961–976.

Cai, Z., Fan, J and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models, *J. Amer. Statist. Assoc.*, **95**, 888-902.

Cai, Z., Fan, J. and Yao, Q. (2000), Functional-coefficient regression models for nonlinear time series, *Journal of the American Statistical Association*, **95**, 941-956.

Carroll, R.J., Fan, J., Gijbels, I. and Wand, M.P. (1997), "Generalized partially linear single-index models," *Journal of the American Statistical Association*, **92**, 477-489.

Carroll, R.J., Ruppert, D. and Welsh, A.H. (1998), "Local estimating equations", *Journal of the American Statistical Association*, **93**, 214-227.

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

Chaudhuri, P. and Marron, J. S. (1999). SiZer for exploration of structures in curves, *J. Amer. Statist. Assoc.*, **94**, 807-822.

Chen, H. (1988), Convergence rates for parametric components in a partly linear model,*The Annals of Statistics*, **16**, 136-146.

Chen, R. and Tsay, R.S. (1993), Functional-coefficient autoregressive models, *Journal of the American Statistical Association*, **88**, 298-308.

Chu, C. K. and Marron, J. S. (1991). Choosing a kernel regression estimator (with discussions). *Statist. Sci.*, **6**, 404-436.

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.

Eubank, R.L. and Hart, J.D. (1992). Testing goodness-of-fit in regression via order selection criteria. *Ann. Statist.*, **20**, 1412-1425.

Eubank, R.L. and LaRiccia, V.M. (1992). Asymptotic comparison of Cramér-von Mises and nonparametric function estimation techniques for testing goodness-of-fit. *Ann. Statist.*, **20**, 2071-86.

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

Fan, J. (1993). Local linear regression smoothers and their minimax. *Ann. Statist.*, **21**, 196–216.

Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *J. Amer. Statist. Assoc.*, **91**, 674-688.

Fan, J. and Chen, J. (1999). One-step local quasi-likelihood estimation, *J. Royal Statist. Soc. B*, **61**, 927-943.

Fan, J., Farmen, M. and Gijbels, I. (1998). A blueprint of local maximum likelihood estimation. *J. Royal Statist. Soc. B*, **60**, 591-608.

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

Fan, J. and Gijbels, I. (2000). Local polynomial fitting, *Smoothing and Regression. Approaches, Computation and Application* (M.G. Schimek ed.), 228-275, John Wiley and Sons

Fan, J., Gijbels, I. and King, M. (1997). Local likelihood and local partial likelihood in hazard regression. *Ann. Statist.*, **25**, 1661-1690.

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 Huang, L. (2001). Goodness-of-fit test for parametric regression models, *J. Amer. Statist. Assoc.*, to appear.

Fan, J. and Marron, J. S. (1994). Fast implementations of nonparametric curve estimators, *J. Comp. Graph. Statist.*, **3**, 35-56.

Fan, J. and Müller, M. (1995). Density and regression smoothing, in *XploRe: an interactive statistical computing environment* (W. Härdle, S.Klinke, B.A. Turlach, eds.), 77-99, Springer, Berlin.

Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, **85**, 645-660.

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

Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks Phenomenon, *Ann. Statist.*, **29**, to appear.

Fan, J. and Zhang, J. (2000), Functional linear models for longitudinal data, *Journal of the Royal Statistical Society, Series B*, **62**, 303–332.

Fan, J. and Zhang, W. (1999), Statistical estimation in varying-coefficient models, *The Annals of Statistics*, **27**, 1491–1518.

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

Gasser, T. and Müller, H.-G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. of Statist.*, **11**, 171–185.

Gasser, T. , Müller, H.-G. and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *J. Royal Statist. Soc. B*, **47**, 238–252.

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.

Granovsky, B.L. and Müller, H.-G. (1991), Optimizing kernel methods: a unifying variational principle. *Inter. Statist. Rev.*, **59**, 373–388.

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

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

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

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

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, **21**, 1926–47.

Hart, J. D. (1997). *Nonparametric Smoothing and Lack-of-fit Tests*, Springer, New York.

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

Hastie, T.J. and Tibshirani, R.J. (1993), Varying-coefficient models (with discussion), *Journal of the Royal Statistical Society, Series B*, **55**, 757-796.

Hoover, D.R., Rice, J.A., Wu, C.O. and Yang, L.P. (1998), Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data, *Biometrika*, **85**, 809–822.

Inglot, T., Kallenberg, W.C.M. & Ledwina, T. (1994). Power approximations to and power comparison of smooth goodness-of-fit tests. *Scand. J. Statist.* **21**, 131-45.

Jones, M. C., Marron, J. S. and Sheater, S. J. (1996a). A brief survey of bandwidth selection for density estimation, *J. Amer. Statist. Assoc.*, **91**, 401-407.

Jones, M. C., Marron, J. S. and Sheater, S. J. (1996b). Progress in data-based bandwidth selection for kernel density estimation, *Computational Statistics*, **11**, 337-381.

Kallenberg, W.C.M. and Ledwina, T. (1997). Data-Driven smooth tests when the hypothesis is composite. *Jour. Ameri. Statist. Assoc.*, **92**, 1094 –1104.

Kauermann, G. and Tutz, G. (1999), On model diagnostics using varying coefficient models, *Biometrika*, **86**, 119-128.

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.

Lehmann, E.L. (1983), *Theory of Point Estimation*, Pacific Grove, California: Wadsworth & Brooks/Cole.

Lin, X. and Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J. Amer. Statist. Assoc.*, **95**, 520-534.

Mack, Y. P. and Müller, H. G. (1989). Convolution type estimators for nonparametric regression estimation. *Statist. Prob. Lett.*, **7**, 229-239.

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. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *J. Amer. Statist. Assoc.*, **82**, 231–238.

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

Nadaraya, E.A. (1964). On estimating regression. *Theory Prob. Appl.*, **9**, 141–142.

Rawlings, J. O. and Spruill, S. E. (1994). Estimating pine seedling response to ozone and acidic rain. In *Case Studies in Biometry*, (Lange, N, Ryan, L., Billard, L., Brillinger, D., Conquest, L. and Greenhouse, J. eds.), 81-106, Wiley, New York.

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. (1997). Local polynomial variance function estimation. *Technometrics*, **39**, 262-73.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley & sons, New York.

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.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics.* Springer, New York.

Speckman, P. (1988), Kernel smoothing in partial linear models, *Journal of the Royal Statistical Society, Series B*, **50**, 413-436.

Spokoiny, V.G. (1996). Adaptive hypothesis testing using wavelets. *Ann. Statist.*, **24**, 2477-2498.

Stone, C.J. (1977). Consistent nonparametric regression. *Ann. Statist.*, **5**, 595–645.

Stone, C.J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, **8**, 1348–1360.

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

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.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing.* Chapman and Hall, London.

Wand, M. P, Marron, J. S. and Ruppert, D. (1991). Transformations in density estimation, *J. Amer. Statist. Assoc.*, **86**, 343-361.

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

Yang, L. and Marron, J. S. (1999). Iterated transformation – kernel density estimation, *J. Amer. Statist. Assoc.*, **94**, 580-589.