A Blueprint of Local Maximum Likelihood Estimation

Jianqing Fan[†] and Mark Farmen
Department of Statistics
University of North Carolina
Chapel Hill, N.C. 27599-3260

Irène Gijbels[‡]
Institut de Statistique
Université Catholique de Louvain
B-1348 Louvain-la-Neuve, Belgium

February 27, 1996

Abstract

Local maximum likelihood estimation is a nonparametric counterpart of the widely-used parametric maximum likelihood techniques. It extends the scope of the parametric maximum likelihood method to a much wider class of parametric spaces. Associated with this nonparametric estimation scheme is the issue of bandwidth selection and bias and variance assessment. This article provides a unified approach to selecting bandwidth and constructing confidence intervals for local maximum likelihood estimation. The approach is then applied to least-squares nonparametric regression and to nonparametric logistic regression. Our experiences in these two settings show that the general idea outlined here is powerful and encouraging.

1 Introduction

Maximum likelihood estimation provides a useful blueprint for various statistical estimation problems. It also provides a unified method for constructing confidence intervals of estimated parameters. A drawback of this method is that one has to assume a particular parametric form for the unknown targeting function. This restrictive assumption can be removed by using a local maximum likelihood estimator. An important issue is then to choose the size of the neighborhood. Further, the question arises of how to construct confidence intervals. These problems are challenging and

[†] Supported by NSF Grant DMS9504414 and an NSA grant.

[‡] Supported by 'Projet d'Actions de Recherche Concertées' (No. 93/98 - 164), by an FNRS-grant (No. 1.5.001.95F) and by the European Human Capital and Mobility Programme (CHRX-CT94-0693).

Abbreviated title: Local likelihood estimation.

AMS 1991 subject classification. Primary 62G07. Secondary 62A10.

Key words and phrases: Bandwidth selection; confidence intervals; generalized linear models; logit regression; maximum likelihood; nonparametric regression.

so far they have no satisfactory answer in the literature. This article attempts to provide a unified approach to these problems. It can serve as a kind of blueprint for local maximum likelihood estimation.

Local maximum likelihood estimation is based on the idea of local fitting. Scatterplot smoothing by local fitting has been around for many years. Local fitting is indeed a particular useful technique in nonparametric estimation. Among the earlier papers in the context of nonparametric regression are Stone (1977, 1980), Cleveland (1979) and Friedman and Stuetzle (1981). Further, there is a vast literature on likelihood-based models in various domains of applications, and these models mainly appear in parametric estimation problems.

The idea of using local fitting to likelihood-based regression models was applied by Tibshirani and Hastie (1987) to the class of generalized linear models (see Nelder and Wedderburn (1972)) and to the proportional hazards model of Cox (1972). Fan, Heckman and Wand (1995) show that the approach has good sampling properties when used with local polynomial fitting in the context of generalized linear models. Staniswalis (1989) approached the same problem using a kernel method. These nice sampling properties carry further to the hazards regression setup. See Fan, Gijbels and King (1995). Recently, there is a vast interest in applying the local likelihood method to the problem of density estimation or hazard rate estimation. See for example Hjort (1991, 1995), Jones (1994), Copas (1995), Hjort and Glad (1995), Loader (1995) and Hjort and Jones (1996).

An important issue when using local techniques is the determination of the 'local neighborhood', which is commonly described by a kernel function K and a bandwidth parameter h. It is well-known that among these two quantities, the choice of the bandwidth is the most crucial one. This bandwidth controls the size of the local neighborhood, and can be chosen to be constant or to depend on the locations.

The building blocks for bandwidth selection are to estimate bias and variance of a nonparametric estimator. A general idea for this estimation task is proposed in the paper, and is applicable to most of the likelihood-based models. The idea is a non-trivial extension of the method developed by Fan and Gijbels (1995) in the least-squares context. Fan and Gijbels (1995) provides extensive evidence showing that the resulting procedure performs very well and we show further in this paper that the generalized idea is appealing for logistic regression. Therefore, it is expected that the extension will work well in a more general likelihood context.

The assessed bias and variance have also important applications in constructing confidence intervals and even confidence bands. Indeed, one can use the estimated bias and variance and rely on the asymptotic normality of the estimator to construct confidence intervals and confidence bands. See for example Eubank and Speckman (1993).

The organization of the paper is as follows. In the next section we discuss briefly the idea of local likelihood techniques using local polynomial fitting. Sections 3 and 4 show how to access the bias and variance of the local maximum likelihood estimator. Section 5 discusses a simple approach to select a pilot bandwidth. The applications to bandwidth selection are discussed in Section 6. Section 7 discusses how to construct confidence intervals based on the assessed bias and variance. The proposed methodology is then illustrated for local least-squares regression and for local logistic regression in Section 8.

2 Local log-likelihood estimation

In order to introduce the local likelihood idea, we recall the maximum likelihood estimation method for a parametric model. Suppose that the i^{th} observation (X_i, Y_i) in a sample $(X_1, Y_1), \cdots, (X_n, Y_n)$ has a contribution $\ell\{g(X_i), Y_i\}$ to the conditional log-likelihood, where $g(\cdot)$ is an unknown parametrized function of interest, i.e. $g(x) = g_\theta(x)$ where θ is an unknown parameter. The conditional log-likelihood of the n observations is then given by $\sum_{i=1}^n \ell\{g_\theta(X_i), Y_i\}$. We require that θ is the solution to the likelihood equation

$$E\{\ell'\{g_\theta(x), Y\}|X = x\} = 0, \tag{2.1}$$

where $\ell'(t, u) = \frac{\partial}{\partial t}\ell(t, u)$. Note that one can regard (2.1) as the definition of the parameter θ.

We now turn to nonparametric estimation of $g(\cdot)$ in which the form of $g(\cdot)$ is completely unknown. Suppose that we want to estimate $g(x_0)$. For data points X_i in a neighborhood of x_0 we approximate $g(X_i)$ via a Taylor expansion by a polynomial of degree p:

$$g(X_i) \approx g(x_0) + g'(x_0)(X_i - x_0) + \cdots + \frac{g^{(p)}(x_0)}{p!}(X_i - x_0)^p \equiv \mathbf{X}_i^T \beta,$$

where $\mathbf{X}_i = (1, X_i - x_0, \cdots, (X_i - x_0)^p)^T$ and $\beta = (\beta_0, \cdots, \beta_p)^T$, with $\beta_\nu = g^{(\nu)}(x_0)/\nu!$, $\nu = 0, 1, \cdots, p$. For data points (X_i, Y_i) in a neighborhood of x_0, the contribution to the log-likelihood is $\ell(\mathbf{X}_i^T\beta, Y_i)$, weighted by $K_h(X_i - x_0)$, with $K_h(\cdot) = K(\cdot/h)/h$. These considerations yield the conditional *local log-likelihood*

$$L_p(\beta; h, x_0) = \sum_{i=1}^{n} \ell(\mathbf{X}_i^T\beta, Y_i) K_h(X_i - x_0). \tag{2.2}$$

The subscript p indicates the degree of the polynomial used for the local fitting. Maximizing the local log-likelihood (2.2) with respect to β, gives the vector of estimators $\widehat{\beta} = (\widehat{\beta}_0, \cdots, \widehat{\beta}_p)^T$. Estimators $\widehat{g}_\nu(x_0)$ for $g^{(\nu)}(x_0)$, $\nu = 0, 1, \cdots, p$ are then given by

$$\widehat{g}_\nu(x_0) = \nu! \widehat{\beta}_\nu. \tag{2.3}$$

To illustrate the above concept we consider the normal regression model $Y = g(X) + \varepsilon$ with $\varepsilon \sim \mathrm{N}(0; \sigma^2)$, and X and ε independent. The conditional local log-likelihood is

$$-\log(\sqrt{2\pi}\sigma) \sum_{i=1}^{n} K_h(X_i - x_0) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \Big\{ Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \Big\}^2 K_h(X_i - x_0), \tag{2.4}$$

which has to be maximized with respect to β. This is equivalent to minimizing

$$\sum_{i=1}^{n} \Big\{ Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \Big\}^2 K_h(X_i - x_0), \tag{2.5}$$

leading to local polynomial regression, also referred to as locally weighted least-squares regression.

In the above example, the unknown function was a mean regression function. In other contexts, $g(\cdot)$ will be another unknown function of interest. For example $g(\cdot)$ can be a transformed conditional mean function in generalized linear models, or the risk contribution function of covariates to the conditional hazard function in a proportional hazards model encountered in survival analysis, or the logarithm of the density function in a density estimation problem.

3 Assessing the bias of the estimator

The bias of the estimator $\widehat{\beta}$ comes from the approximation error in the Taylor expansion. Let $r(X_i) = g(X_i) - \sum_{j=0}^{p} g^{(j)}(x_0)(X_i - x_0)^j/j!$ denote this approximation error. Then, a further expansion of $g(X_i)$ gives approximation to the approximation error:

$$r(X_i) \approx \beta_{p+1}(X_i - x_0)^{p+1} + \cdots + \beta_{p+a}(X_i - x_0)^{p+a} \equiv r_i \tag{3.1}$$

where a denotes the order of the approximation. The choice of a will have some effect on the performance of the estimated bias. A discussion on the choice of a can be found in Fan and Gijbels (1995). A good practical performance is obtained with $a = 2$.

Suppose for a moment that the quantities r_i are known. Then, a more precise local log-likelihood is

$$L_p^*(\beta; h, x_0) = \sum_{i=1}^{n} \ell(\mathbf{X}_i^T \beta + r_i, Y_i) K_h(X_i - x_0). \tag{3.2}$$

Denote by $\widehat{\beta}^* = \widehat{\beta}^*(x_0)$ the maximizer of the local log-likelihood $L_p^*(\beta; h, x_0)$. The bias of $\widehat{\beta}(x_0)$ can then be estimated by $\widehat{\beta}(x_0) - \widehat{\beta}^*(x_0)$. However, the computation of $\widehat{\beta}^* = \widehat{\beta}^*(x_0)$ can be avoided, as follows. Denote by

$$L_p^{*\prime}(\beta; h, x_0) = \frac{\partial}{\partial \beta} L_p^*(\beta; h, x_0) \quad \text{and} \quad L_p^{*\prime\prime}(\beta; h, x_0) = \frac{\partial^2}{\partial \beta^2} L_p^*(\beta; h, x_0)$$

respectively the gradient vector and the Hessian matrix of the log-likelihood L_p^*. Since $\widehat{\beta}^*(x_0)$ is the maximizer of $L_p^*(\beta; h, x_0)$, a Taylor expansion gives

$$0 = L_p^{*\prime}(\widehat{\beta}^*; h, x_0) \approx L_p^{*\prime}(\widehat{\beta}; h; x_0) + L_p^{*\prime\prime}(\widehat{\beta}; h, x_0)\{\widehat{\beta}^*(x_0) - \widehat{\beta}(x_0)\},$$

and this leads us to define the estimated bias vector

$$\widehat{b}_p(x_0) = \left\{ L_p^{*\prime\prime}(\widehat{\beta}; h, x_0) \right\}^{-1} L_p^{*\prime}(\widehat{\beta}; h, x_0). \tag{3.3}$$

To get a better insight into the bias approximation (3.3) let us look at the normal likelihood (2.5). Denote by $\mathbf{X}$ the design matrix of the regression problem, i.e. the $n \times (p+1)$ matrix whose $(i,j)^{th}$-element is $(X_i - x_0)^{j-1}$, and let $\mathbf{W} = \text{diag}\,\{K_h(X_i - x_0)\}$ be the diagonal matrix containing the weights. Then, except for a constant factor,

$$L_p^*(\beta; h, x_0) = (\mathbf{y} - \mathbf{X}\beta - \mathbf{r})^T \mathbf{W} (\mathbf{y} - \mathbf{X}\beta - \mathbf{r}),$$

where $\mathbf{y} = (Y_1, \cdots, Y_n)^T$ and $\mathbf{r} = (r_1, \cdots, r_n)^T$. Further, $\widehat{\beta} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}\, \mathbf{X}^T \mathbf{W} \mathbf{y}$, and hence

$$L_p^{*\prime}(\widehat{\beta}; h, x_0) = 2\mathbf{X}^T \mathbf{W} \mathbf{r} \quad \text{and} \quad L_p^{*\prime\prime}(\widehat{\beta}; h, x_0) = 2\mathbf{X}^T \mathbf{W} \mathbf{X}. \tag{3.4}$$

Therefore

$$\widehat{b}_p(x_0) = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{r},$$

which is equal to the approximation of the bias $E(\widehat{\beta}|\mathbb{X})-\beta$, where $\mathbb{X}$ denotes $(X_1,\cdots,X_n)$, obtained in Fan and Gijbels (1995).

Recall that the approximated bias (3.3) depends on the quantities $r_1,\cdots,r_n$, which are unknown. These quantities will be estimated by fitting a polynomial of degree $p+a$ locally via (2.2), using a pilot bandwidth h^*. This gives estimates $\widehat{\beta}^{(p+a)}=(\widehat{\beta}_0,\cdots,\widehat{\beta}_{p+a})^T$, which are then substituted into expression (3.1), yielding estimates $\widehat{r}_1,\cdots,\widehat{r}_n$ of $r_1,\cdots,r_n$. These estimates are then substituted into (3.2), leading to the estimated bias as in (3.3). Denote the estimated bias of $\widehat{\beta}_\nu$ by $\widehat{B}_{p,\nu}(x_0;h)$. The choice of the pilot bandwidth h^* will be discussed in Section 5.

4 Assessing the variance of the estimator

To get a grip on the variance, first note that, with $\beta=(\beta_0,\cdots,\beta_p)^T$ being the true local parameters,

$$0=L_p'(\widehat{\beta};h,x_0)\approx L_p'(\beta;h,x_0)+L_p''(\beta;h,x_0)(\widehat{\beta}-\beta).$$

This leads to

$$\widehat{\beta}-\beta\approx-\left\{L_p''(\beta;h,x_0)\right\}^{-1}L_p'(\beta;h,x_0),$$

and an approximation for the conditional variance is

$$\mathrm{Var}(\widehat{\beta}|\mathbb{X})\approx\left\{L_p''(\beta;h,x_0)\right\}^{-1}\mathrm{Var}\left\{L_p'(\beta;h,x_0)|\mathbb{X}\right\}\left\{L_p''(\beta;h,x_0)\right\}^{-1}. \tag{4.1}$$

The Hessian matrix $L_p''(\beta;h,x_0)$ can be estimated by $L_p''(\widehat{\beta};h,x_0)$, and the conditional variance on the right-hand side of (4.1) can be approximated as follows. From (2.2) we obtain

$$\begin{aligned}\mathrm{Var}\left\{L_p'(\beta;h,x_0)|\mathbb{X}\right\} &= \sum_{i=1}^n\mathrm{Var}\left\{\frac{\partial}{\partial\beta}\ell(\mathbf{X}_i^T\beta,Y_i)|\mathbb{X}\right\}K_h^2(X_i-x_0)\\ &= \sum_{i=1}^n\mathrm{Var}\left\{\ell'(\mathbf{X}_i^T\beta,Y_i)|X_i\right\}\mathbf{X}_i\mathbf{X}_i^TK_h^2(X_i-x_0).\end{aligned}$$

Since we deal with X_i in a neighborhood around x_0,

$$\mathrm{Var}\{\ell'(\mathbf{X}_i^T\beta,Y_i)|X_i\}\approx\mathrm{Var}[\ell'\{g(x_0),Y\}|X=x_0].$$

Thus, we have

$$\mathrm{Var}\left\{L_p'(\beta;h,x_0)|\mathbb{X}\right\}\approx\mathrm{Var}\left\{\ell'(g(x_0),Y)|X=x_0\right\}\overline{S}_n, \tag{4.2}$$

where $\overline{S}_n = \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T K_h^2(X_i - x_0)$. Combining (4.1) and (4.2) we obtain the following approximation of the conditional variance of $\widehat{\beta}$:

$$\operatorname{Var}(\widehat{\beta}|\mathbb{X}) \approx \Sigma(x_0) = \operatorname{Var}\{\ell'(g(x_0), Y)|X = x_0\} \left\{L_p''(\beta; h, x_0)\right\}^{-1} \overline{S}_n \left\{L_p''(\beta; h, x_0)\right\}^{-1}. \qquad (4.3)$$

The unknown local parameter β in (4.3) can be estimated by $\widehat{\beta}$. The first factor in (4.3) is also unknown and has to be estimated. We separate this into two cases. The first case is that $\operatorname{Var}\{\ell'(g(x_0), Y)|X = x_0\} = V\{g(x_0)\}$ for some known function $V(\cdot)$ such as for the Bernoulli, Poisson and Exponential distributions in the context of generalized linear models. In this case, we estimate the conditional variance by $V\{\widehat{g}(x_0)\}$. The second case is that we do not have such a form. This is for example the case in the normal likelihood model. In this second case we have that, by (2.1),

$$\operatorname{Var}\{\ell'(g(x_0), Y)|X = x_0\} = E[\{\ell'(g(x_0), Y)\}^2|X = x_0].$$

This quantity can be estimated by

$$\frac{\sum_{i=1}^n \left\{\ell'(\mathbf{X}_i^{*T}\widehat{\beta}^{(p+a)}, Y_i)\right\}^2 K_{h^*}(X_i - x_0)}{\sum_{i=1}^n K_{h^*}(X_i - x_0)}, \qquad (4.4)$$

where $\widehat{\beta}^{(p+a)} = (\widehat{\beta}_0, \cdots, \widehat{\beta}_{p+a})^T$ is the result of a $(p+a)^{th}$-order local polynomial fit (2.2) using the pilot bandwidth h^* and $\mathbf{X}_i^* = (1, X_i - x_0, \cdots, (X_i - x_0)^{p+a})^T$.

In many practical situations, it is possible to encounter over-dispersion in the first case above, i.e.

$$\operatorname{Var}\{\ell'(g(x_0), Y)|X = x_0\} = \phi V\{g(x_0)\},$$

where ϕ is an unknown parameter. See McCullagh and Nelder (1989) for a more detailed description. In this case, we will use (4.4), instead of $V\{\widehat{g}(x_0)\}$ to estimate the conditional variance in (4.3).

To illustrate the idea, let use consider again the normal likelihood, for which $\ell(x, y) = -(y - x)^2/2$. This implies

$$\operatorname{Var}\{\ell'(g(x_0), Y)|X = x_0\} = \operatorname{Var}\{Y - g(x_0)|X = x_0\} = \sigma^2(x_0),$$

and (4.3) can be expressed as

$$\Sigma(x_0) = \sigma^2(x_0) S_n^{-1} \overline{S}_n S_n^{-1}, \qquad (4.5)$$

where $S_n = \mathbf{X}^T \mathbf{W} \mathbf{X}$.

The right-hand side of (4.5) is exactly the approximation for the conditional variance derived by Fan and Gijbels (1995) for the locally weighted least-squares regression problem.

What does the estimator (4.4) reduce to in this special case? Here $\ell'(u, y) = y - u$ and $\mathbf{X}_i^{*T} \widehat{\beta}^{(p+a)} = \widehat{Y}_i$, the fitted value from a local $(p+a)^{th}$-order fit. Hence (4.4) reduces to

$$\widehat{\sigma}^2(x_0) = \frac{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2 K_{h^*}(X_i - x_0)}{\sum_{i=1}^n K_{h^*}(X_i - x_0)}, \tag{4.6}$$

which is asymptotically the same as the estimator for $\sigma^2(x_0)$ provided in Fan and Gijbels (1995) (see expression (2.3) in that paper or (5.2) below for a similar expression).

5 Pilot bandwidth selector

The estimated bias discussed in Section 3 depends on the pilot estimation of $\beta_{p+1}, \cdots, \beta_{p+a}$. This in turn requires a selection of bandwidth. Also the estimation of the variance, described in the previous section, requires selection of a pilot bandwidth.

To motivate our selection procedure, let us consider the least-squares case studied in Fan and Gijbels (1995). Suppose the goal is to estimate $g^{(\nu)}(\cdot)$ using a local polynomial fit of order p. Let $h_{\text{opt}}(x_0)$ be the asymptotic optimal bandwidth that minimizes the asymptotic optimal MSE of $\widehat{\beta}_\nu(x_0)$. In the least-squares case, Fan and Gijbels (1995) define the following Residual Squares Criterion (RSC):

$$\text{RSC}(x_0; h) = \widehat{\sigma}^2(x_0)\{1 + (p+1)/N\}, \tag{5.1}$$

where $\widehat{\sigma}^2(\cdot)$ is the normalized weighted residual sum of squares after fitting locally a p^{th}-order polynomial given by

$$\widehat{\sigma}^2(x_0) = \frac{\sum_{i=1}^n \left(Y_i - \widehat{Y}_i\right)^2 K_h(X_i - x_0)}{\text{tr}\,(\mathbf{W}) - \text{tr}\left\{(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^2\mathbf{X}\right\}}. \tag{5.2}$$

and N^{-1} is the first diagonal element of the matrix $(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{W}^2\mathbf{X})(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$. Note that N in fact reflects the effective number of local data points, since $\text{Var}\{\widehat{\beta}_0|\mathbb{X}\} \approx \sigma^2(x_0)/N$ by (4.5). The intuition behind (5.1) is as follows. When the bandwidth h is too large, the polynomial does not fit well. The bias is large and hence also the Residual Sum of Squares $\widehat{\sigma}^2(x_0)$. When the bandwidth h is too small, the variance of the fit will be large and hence also N^{-1}. Since both factors $\widehat{\sigma}^2(x_0)$ and N are incorporated in the RSC this quantity will prevent both extreme choices of the bandwidth. It is shown in Fan and Gijbels (1995) that the minimizer of (5.1) is only a constant factor away from the targeted optimal bandwidth:

$$h_{\text{opt}}(x_0) = \text{adj}_{p,\nu}(K) h_o(x_0), \tag{5.3}$$

where $h_o(x_0)$ is the asymptotic optimal bandwidth that minimizes the main terms of the expected value of (5.1), and $\text{adj}_{p,\nu}(K)$ is a known constant that depends only on K, p and ν (see definition below), and is tabulated in Fan and Gijbels (1995). The exact expression of this constant is as follows. Let $\mu_j = \int t^j K(t)dt$. Define the $(p+1) \times (p+1)$ matrix S with the $(i+j-2)^{th}$-moment μ_{i+j-2} of K as its $(i,j)^{th}$-element. Let $K_\nu^*(t) = \{\sum_{j=0}^p s^{\nu+1,j} t^j\} K(t)$ be the equivalent kernel, where $s^{\nu+1,j}$ is the $(\nu+1,j)^{th}$-element of S^{-1}. The constants $\text{adj}_{p,\nu}(K)$ are defined as

$$\text{adj}_{p,\nu}(K) = \left[\frac{(2\nu+1) C_p \int K_\nu^{*2}(t)dt}{(p+1-\nu)\{\int t^{p+1} K_\nu^*(t)dt\}^2 \int K_0^{*2}(t)dt} \right]^{1/(2p+3)}, \tag{5.4}$$

where $C_p = \mu_{2p+2} - (\mu_{p+1}, \cdots, \mu_{2p+1}) S^{-1} (\mu_{p+1}, \cdots, \mu_{2p+1})^T$.

The above criterion can also be used when the function $g(\cdot)$ is a transform of the mean regression function: $g(\cdot) = L\{\mu(\cdot)\}$ where L is a link function and μ is the mean regression function. In this case, $\widehat{Y}_i = L^{-1}(\mathbf{X}_i^T \widehat{\beta})$ in equation (5.2). This RSC-criterion corresponds to the approximately weighted squared errors in the domain of g using weight $[L'\{\mu(x)\}]^{-2}$.

An extension of the above idea is to regard the local likelihood problem as iterative local least-squares problems. Given the current value β_c of β, update β_c via the local p^{th}-order polynomial regression of the working variable

$$Z_i = \mathbf{X}_i^T \beta_c - \frac{\ell'(\mathbf{X}_i^T \beta_c, Y_i)}{E\{\ell''(g(x_0), Y) | X = x_0\}} \tag{5.5}$$

on X_i, where the conditional expectation is computed using the parameter β_c. The justification of this is given in the appendix. Thus, at the last step of the iteration, we can regard the local likelihood problem as a local polynomial regression problem, and use the residual squares criterion

$$\text{ERSC}(x_0; h) = \widehat{\sigma_*}^2(x_0)\{1 + (p+1)/N\}, \tag{5.6}$$

where $\widehat{\sigma_*}^2(x_0)$ is the normalized residual sum of squares using the working variable Z_i. Compare with (5.1). We will refer to the criterion (5.6) as the Extended Residuals Squares Criterion (ERSC). The heuristic justification of this is simple. First of all, the bias of $\widehat{\beta}$ comes from local polynomial approximation of g. Hence, it is the same for the local likelihood method as for the local least-squares problem. Using (4.3) together with approximation (A.2),

$$\mathrm{Var}(\widehat{\beta}|\mathbb{X}) \approx \sigma_*^2(x_0) S_n^{-1} \bar{S}_n S_n^{-1},$$

where

$$\sigma_*^2(x_0) = \mathrm{Var}\{\ell'(g(x_0), Y)|X = x_0\}[E\{\ell''(g(x_0), Y)|X = x_0\}]^{-2}.$$

Comparing (4.3) with (4.5), the asymptotic variance of the local likelihood problem corresponds to that of the least-squares problem with $\sigma = \sigma_*$. Treating β_c in (5.5) as fixed, the working variable Z_i has the same variance structure, namely

$$\mathrm{Var}(Z_i|X_i) \approx \sigma_*^2(x_0).$$

We now can select the pilot bandwidth as follows. Let

$$\widehat{h}_{p,\nu}^* = \arg\min_h \int \mathrm{ERSC}(x; h) w(x) dx,$$

for some given weight function w. Then, define the ERSC-selector as follows

$$\widehat{h}_{p,\nu}^{\mathrm{ERSC}} = \mathrm{adj}_{p,\nu}(K) \widehat{h}_{p,\nu}^*. \tag{5.7}$$

As noted above, in the case that $g(\cdot) = L\{\mu(\cdot)\}$, when the weight $w(x) = [L'\{\mu(x)\}]^{-2}$ is used, this criterion is approximately the same as the RSC-criterion (5.1). In the least-squares problem, this bandwidth selector was investigated in Fan and Gijbels (1995). It performs reasonably well, but the rate of convergence can be improved. For this reason, we only use it in the pilot stage.

6 Bandwidth selection

Recall that the estimation procedure consists of maximizing the local log-likelihood (2.2), leading to the estimated vector $\widehat{\beta}$. The complexity of the model is determined by the bandwidth h. If $h \to \infty$ then (2.2) results in a global fit of a polynomial of degree p. If on the other hand $h \to 0$ then we end up with interpolation of the data. Many interesting models lie between these two extreme choices. In this section we discuss data-driven choices of a constant and local variable bandwidth.

The basic idea for bandwidth selection is very simple. First a pilot bandwidth $\widehat{h}^*_{p+a,p+1}$ should be selected, and this can be done by either the RSC-criterion (5.1) or the ERSC-criterion (5.6). As noted at the end of Section 5, the difference is only a matter of weighting scheme. Given a pilot bandwidth $\widehat{h}^*_{p+a,p+1}$, we then first fit a polynomial of degree $p+a$ locally via maximizing (2.2), resulting in the estimator $\widehat{\beta}^{(p+a)} = (\widehat{\beta}_0, \cdots, \widehat{\beta}_{p+a})^T$. With these estimated parameters we obtain the estimated bias $\widehat{B}_{p,\nu}(x_0; h)$ and variance $\widehat{V}_{p,\nu}(x_0; h)$ of $\widehat{\beta}_\nu$, which are respectively the $(\nu+1)^{th}$-element of (3.3) and the $(\nu+1)^{th}$-diagonal element of the estimated expression (4.3). An estimator for the Mean Squared Error (MSE) of $\widehat{\beta}_\nu$ is then given by

$$\widehat{\mathrm{MSE}}_{p,\nu}(x_0; h) = \widehat{B}^2_{p,\nu}(x_0; h) + \widehat{V}_{p,\nu}(x_0; h). \tag{6.1}$$

This leads to the following bandwidth selector:

$$\widehat{h}_{p,\nu} = \arg\min_h \int \widehat{\mathrm{MSE}}_{p,\nu}(x; h) w(x) dx, \tag{6.2}$$

where $w(\cdot)$ is a given weight function. A common choice of $w(\cdot)$ is the indicator function of the interval where the curve $g^{(\nu)}(\cdot)$ is to be estimated.

The constant bandwidth $\widehat{h}_{p,\nu}$, which is independent of the location x_0, suffices in many applications. However, when the curve $g^{(\nu)}(\cdot)$ admits various degrees of smoothness at different locations, a variable bandwidth selector is needed in order to enhance the spatial adaptation. The basic ideas for selecting such a variable bandwidth are simple: select a bandwidth $\widehat{h}_{p,\nu}(x_0)$ that minimizes the locally weighted $\widehat{\mathrm{MSE}}_{p,\nu}(x; h)$ around the point x_0. This average stabilizes the estimated MSE. The implementation is analogous to that discussed in Fan and Gijbels (1995) in the least-squares regression context. We omit details here.

7 Confidence Intervals

A confidence interval is a very important tool for understanding the sampling variability of an estimator. In the context of nonparametric function estimation, the task of constructing such an interval is difficult, due to non-negligible bias. However, with our estimated bias and variance, one can easily construct a confidence interval.

Denote by

$$\widehat{B}^A_{p,\nu}(x_0;h) = \int \widehat{B}_{p,\nu}(x;h)K_h(x-x_0)dx \tag{7.1}$$
$$\widehat{V}^A_{p,\nu}(x_0;h) = \int \widehat{V}_{p,\nu}(x;h)K_h(x-x_0)dx. \tag{7.2}$$

These two quantities are just local weighted averages of respectively the estimated bias and variance. Note that the estimated bias involves the estimation of higher order derivative curves, whose estimation can be unstable. The purpose of the average is to stabilize the estimated bias and variance function, and to prevent them from abrupt change.

The local maximum likelihood estimator is usually asymptotically normal. In the context of generalized linear models, this has been shown by Fan, Heckman and Wand (1995) and in the context of hazards regression by Fan, Gijbels and King (1995). By invoking the asymptotic normality, we construct the pointwise confidence interval as follows. With approximately $1-\alpha$ coverage probability, the unknown function $g^{(\nu)}(x_0)$ falls in the random intervals

$$\widehat{g}^{(\nu)}(x_0) - \widehat{B}^A_{p,\nu}(x_0;h) \pm \Phi^{-1}(1-\alpha)\ \widehat{V}^A_{p,\nu}(x_0;h). \tag{7.3}$$

For constructing simultaneous confidence bands, one can use the raw materials given in (7.1) and (7.2) and the ideas given in Eubank and Speckman (1993).

8 Applications to logistic regression

We have illustrated the key idea of this paper in the context of the least-squares regression problem. Extensive simulations in Fan and Gijbels (1995) indicate a good performance of the resulting procedure. We now consider nonparametric logistic regression to further clarify the idea.

8.1 Illustration of the blueprint

We assume that the data (X_i, Y_i) are i.i.d. and that the conditional distribution of Y_i given X_i is a Bernoulli distribution:

$$P(Y_i = 1|X_i = x) = p(x); \qquad P(Y_i = 0|X_i = x) = 1 - p(x) \equiv q(x).$$

In this case, the mean regression function is $p(x) = E(Y|X = x)$. The parameter of interest is

$$g(x) = \text{logit}(p(x)) = \log \frac{p(x)}{1 - p(x)}.$$

In this nonparametric regression context, estimating $p(\cdot)$ is equivalent to estimating $g(\cdot)$. However, we prefer to working on the logit domain, since the log-likelihood is concave, and the logistic linear regression model corresponds to our case with $h = \infty$. See Fan, Heckman and Wand (1995) for more detailed arguments.

The log-likelihood is determined via

$$\ell(g(X), Y) = \log\{p(X)^Y q(X)^{1-Y}\} = Y g(X) - \log[1 + \exp\{g(X)\}].$$

Thus, $\ell'(u, y) = y - e^u/(1 + e^u)$ and

$$E\{\ell'(g(X), Y)|X\} = 0, \quad \text{Var}\{\ell'(g(X), Y)|X\} = p(X)q(X). \tag{8.1}$$

For this case, we have

$$L_p^*(\beta; h, x_0) = \sum_{i=1}^{n} \Big[Y_i(X_i^T\beta + r_i) - \log\{1 + \exp(\mathbf{X}_i^T\beta + r_i)\}\Big] K_h(X_i - x_0).$$

The local likelihood L_p corresponds to L_p^* with $r_i \equiv 0$. Denote by

$$p_i^* = \frac{\exp(\mathbf{X}_i^T\beta + r_i)}{1 + \exp(\mathbf{X}_i^T\beta + r_i)}, \quad \text{and} \quad p_i = \frac{\exp(\mathbf{X}_i^T\beta)}{1 + \exp(\mathbf{X}_i^T\beta)}. \tag{8.2}$$

Then, simple algebra shows that

$$\begin{aligned} L_p^{*\prime}(\beta; h, x_0) &= \sum_{i=1}^{n} (Y_i - p_i^*)\mathbf{X}_i K_h(X_i - x_0), \\ L_p^{*\prime\prime}(\beta; h, x_0) &= -\sum_{i=1}^{n} p_i^*(1 - p_i^*)\mathbf{X}_i\mathbf{X}_i^T K_h(X_i - x_0). \end{aligned}$$

The approximated bias vector and variance matrix are now easily obtained, and are given by

$$\begin{aligned} \widehat{b}_p(x_0) &= \left\{L_p^{*\prime\prime}(\widehat{\beta}; h, x_0)\right\}^{-1} L_p^{*\prime}(\widehat{\beta}; h, x_0) \\ \Sigma(x_0) &= p(x_0)q(x_0)\{\sum_{i=1}^{n} p_i q_i \mathbf{X}_i\mathbf{X}_i^T K_h(X_i - x_0)\}^{-1}(\mathbf{X}^T\mathbf{W}^2\mathbf{X})\{\sum_{i=1}^{n} p_i q_i \mathbf{X}_i\mathbf{X}_i^T K_h(X_i - x_0)\}^{-1}, \end{aligned}$$

where $q_i = 1 - p_i$.

It can also easily be computed that

$$\text{ERSC}(x_0; h_0) = 2\{1 + (p+1)/N\} \sum_{i=1}^{n} [Y_i \log(Y_i/\widehat{p}_i) + (1 - Y_i)\log\{(1 - Y_i)/(1 - \widehat{p}_i)\}] K_h(X_i - x_0),$$

where $\widehat{p}_i$ is given by (8.2) with β being estimated. Minimizing the average of ERSC gives a pilot bandwidth selector as discussed in Section 5.

8.2 Implementation

In the context of logistic regression, the two stage bandwidth selector was tested on simulated data from a variety of target curves. The Epanechnikov kernel

$$K(u) = \frac{3}{4}\left(1 - u^2\right) I_{[-1,1]}(u)$$

and order $p = 1$ were used for estimating g at a set of equally spaced grid points $x_1, ..., x_{\text{nbin}}$. The RSC-criterion was used to choose the pilot bandwidth with a $p = 3$ order fit for estimating the curvature g''. The sum

$$\text{ARSC}(h_j) = \sum_{i=\tau}^{\text{nbin}-\tau} RSC(x_i; h_j)$$

where τ is the largest integer less than $0.05 \cdot \text{nbin}$, was computed on a multiplicative grid of bandwidths $h_j = C^j h_{\text{min}}$. Restricting the sum helps to reduce boundary effects. At the boundaries, RSC and estimated derivatives can be too large due to numerical instabilities and scarcity of data.

The implementation of our idea is somewhat tricky since it involves iterative solution. For small values of h, the solution may not even exist since locally one may get a set of all zeros or ones. To avoid such a difficulty, we only consider the bandwidths that are large enough such that $\boldsymbol{\beta}'s$ are defined. To this goal, let l_{runs} be the maximum span among lengths of runs of ones or zeros. For instance, if $X_{(1)}, ..., X_{(n)}$ are the order statistics of the X sample, then $X_{(k)} - X_{(j)}$ is the length of a run of ones provided

$$Y_{j-1} = 0, Y_j = 1, Y_{j+1} = 1, ..., Y_k = 1, Y_{k+1} = 0.$$

The choice of the minimum value of h was given by

$$h_{\text{min}} = (1.25 l_{\text{runs}} + 6\Delta_g)/2$$

where Δ_g is the grid spacing. The maximum value of the bandwidth was set at

$$h_{\text{max}} = \max_j \left\{ C^j h_{\text{min}} : C^j h_{\text{min}} < (x_{\text{nbin}} - x_1)/2 \right\}$$

with $C = 1.2$.

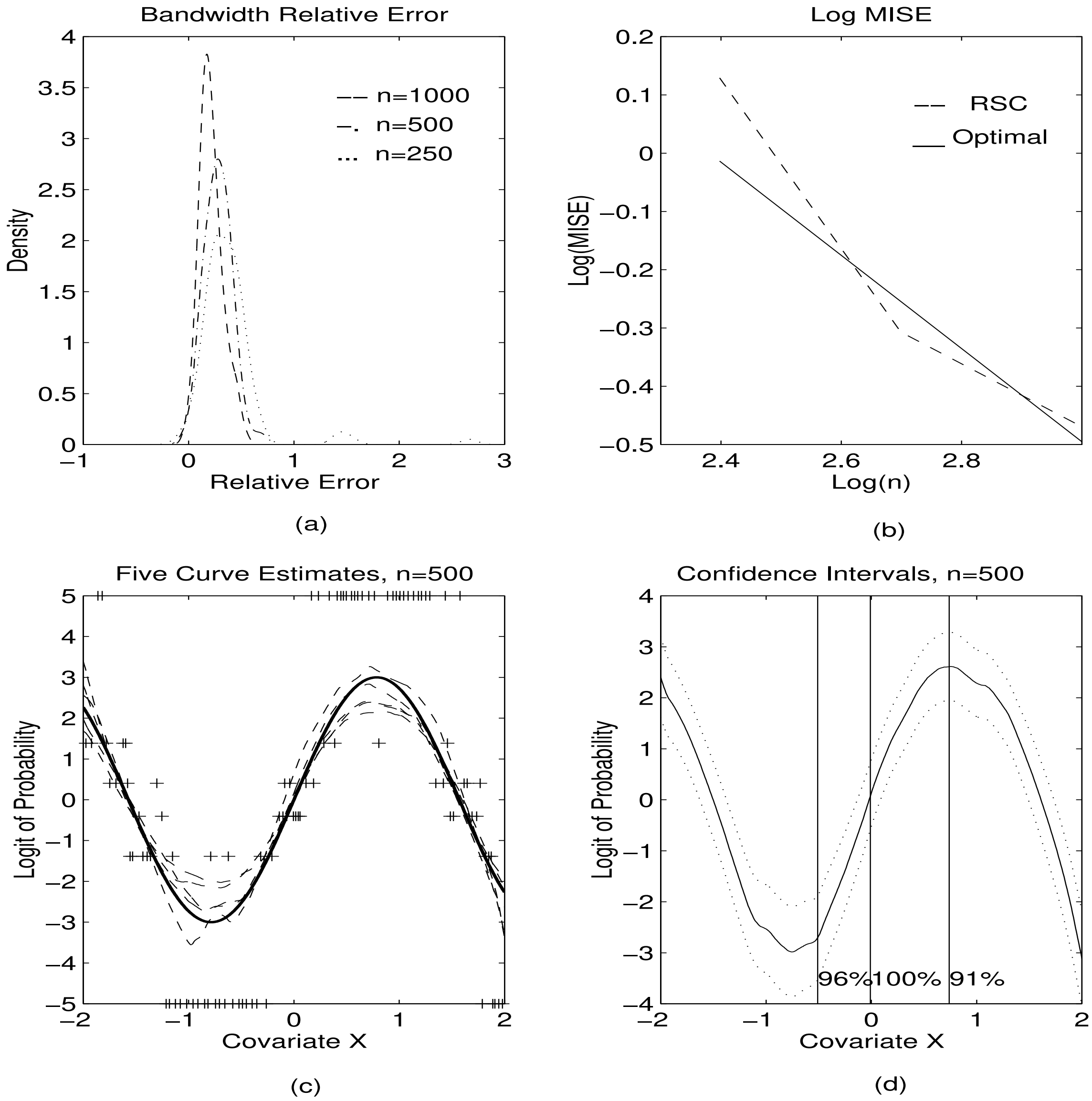

Figure 8.1: *Example 1. (a) Kernel Density Estimate of bandwidth relative errors for three sample sizes. (b) Base 10 log of MISE versus sample size. The asymptotically optimal MISE is indicated by the solid curve. (c) Five sample curve estimates with true curve (solid line) and sample logits (+) using five contiguous observations. (d) Sample curve estimate (solid line) with confidence interval (dashed line) and the percentage of confidence intervals containing the true curve at three separate points indicated with vertical lines.*

8.3 Simulation Results

Results for three underlying target curves on the logit scale are presented. Sample sizes of $n = 250, 500, 1000$ were used. The design is from a uniform distribution on $[-2, 2]$. In other words, the marginal distribution for X is uniform on $[-2, 2]$. The logit transform of the conditional probability

$Y = 1$ is given by

Example 1. $g(x) = 3\sin(2x)$

Example 2. $g(x) = 7[\exp\{-(x+1)^2\} + \exp\{-(x-1)^2)\}] - 5.5$

Example 3. $g(x) = 2 - x^2$.

These curves appear as the thick line in part (c) of the following figures. Only 100 simulations were performed for a given target curve at a given sample size. This was a practical constraint due to the fact that it takes approximately 20 minutes to compute the two stage bandwidth using a primitive algorithm implemented in Matlab on a Sparc 10 station. This can be improved upon by at least a factor of 10, if a low level language such as C is used. The asymptotically optimal bandwidth from Fan, Heckman and Wand (1995)

$$h_{\text{opt}} = \left\{ \frac{\int K(u)^2 du \int \text{Var}(Y|X=x) L'(\mu(x))^2 w(x)/f_X(x) dx}{n \left(\int u^2 K(u) du\right)^2 \int g''(x)^2 w(x) dx} \right\}^{1/5} \tag{8.3}$$

with w taken to be the indicator function on $[-2, 2]$ was used to judge performance.

The relative error of the two stage bandwidth $\hat{h}$ is computed as $(\hat{h} - h_{\text{opt}})/h_{\text{opt}}$. In addition to relative error, the MISE of the curves was calculated based on a sample of 100 and the percentage of the 100 confidence intervals that contained the true curve at a given point was also computed. The results are summarized in Figures 8.1, 8.3 and 8.3. Note that the standard error for the sample coverage is $(0.05 * 0.95/100)^{1/2} = 2.18\%$.

Note that the relative errors become more and more concentrated near zero as n increases. There are a few bandwidths that appear to have quite large relative errors but the number of large errors decreases with n. From part (b), the MISE becomes closer to the asymptotically optimal MISE as n increases. Part (c) indicates the difficulty in estimating these curves by the spread of the sample logits based on five observations. The values of these logits were truncated at ± 5 and some were actually infinite. These large sample logits are indicated by the plus marks along the axes. Details of computing sample logits are as follows: Group the data points according to their covariate values so that each group consists of 5 data points. For each group, compute the sample proportion of ones and do a logit transform of it. This sample logit is then plotted against mean covariate values of its corresponding group. The sample curve estimates indicate that the estimator performs quite well. The pointwise confidence intervals also have good performance. Figures 8.1

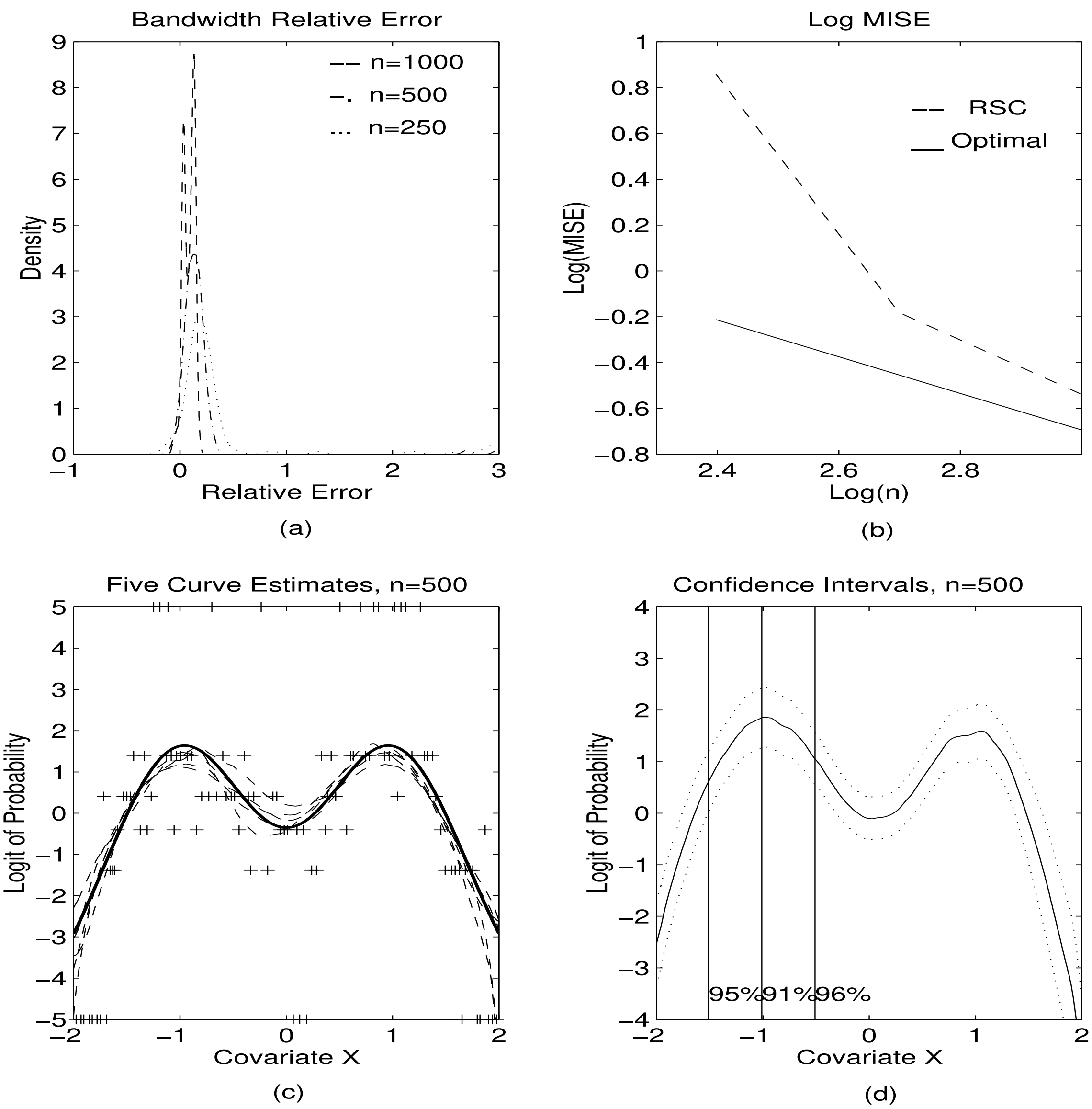

Figure 8.2: *Example 2. (a) Kernel Density Estimate of bandwidth relative errors for three sample sizes. (b) Base 10 log of MISE versus sample size. The asymptotically optimal MISE is indicated by the solid curve. (c) Five sample curve estimates with true curve (solid line) and sample logits (+) using five contiguous observations. (d) Sample curve estimate (solid line) with confidence interval (dashed line) and the percentage of confidence intervals containing the true curve at three separate points indicated with vertical lines.*

and 8.3 indicate some difficulty near the peaks but this difficulty is expected because the bias can be quite large near sharp peaks. The curve estimate in the confidence interval is the bias corrected curve estimate $\widehat{g}^{(0)}(x) - \widehat{B}^{A}_{1,0}(x;h)$ (see (7.3)).

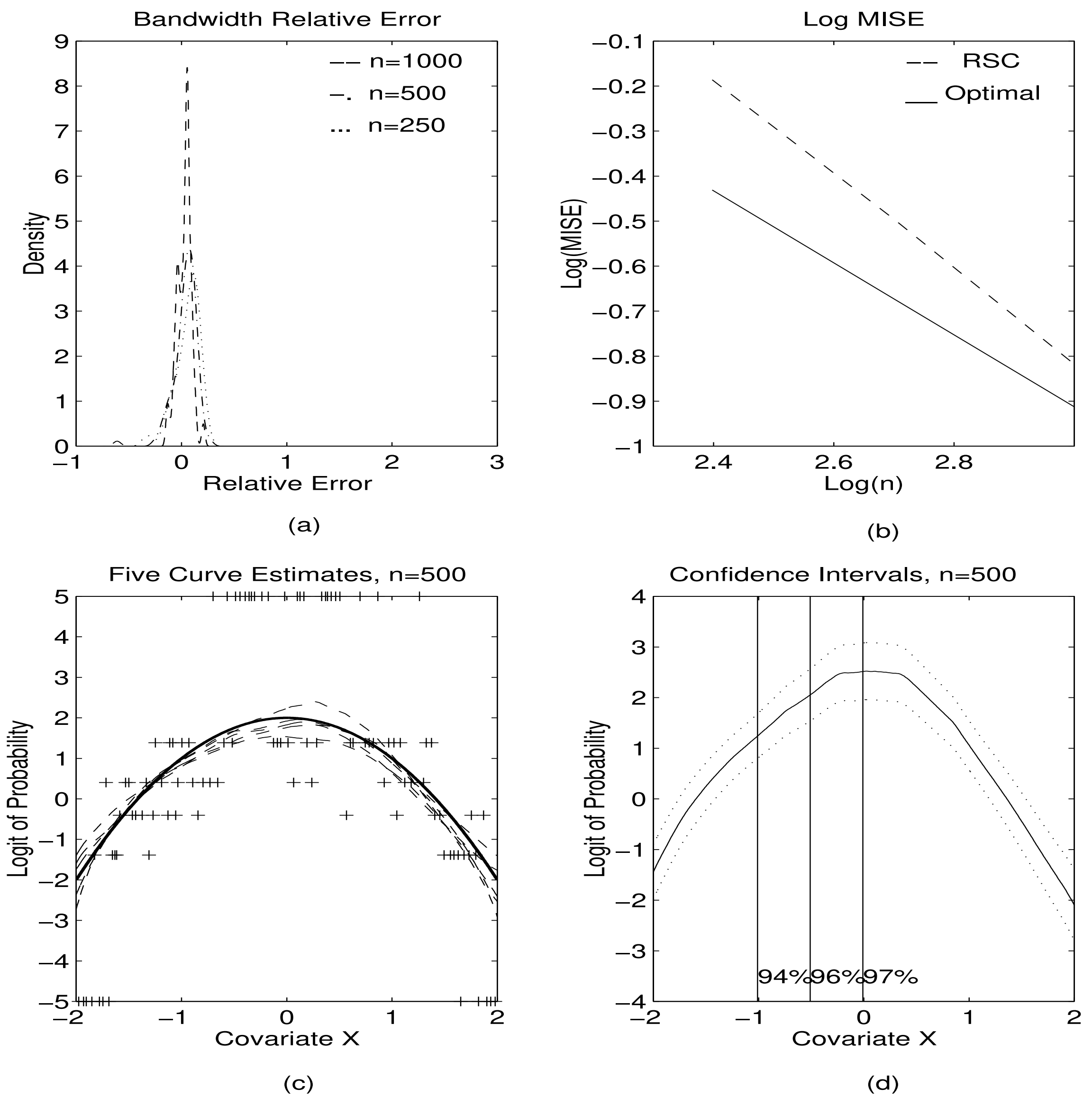

Figure 8.3: *Example 3. (a) Kernel Density Estimate of bandwidth relative errors for three sample sizes. (b) Base 10 log of MISE versus sample size. The asymptotically optimal MISE is indicated by the solid curve. (c) Five sample curve estimates with true curve (solid line) and sample logits (+) using five contiguous observations. (d) Sample curve estimate (solid line) with confidence interval (dashed line) and the percentage of confidence intervals containing the true curve at three separate points indicated with vertical lines.*

9 Concluding Remarks

We have laid out a blueprint for nonparametric smoothing, bandwidth selection and confidence interval construction. The purpose of this article is to indicate that there is a unified approach to nonparametric smoothing. We have only extensively tested the idea in the context of the least-squares case and in a few other cases. Further studies are needed to test the approach in other contexts. We wish this article will stimulate future research in this topic.

REFERENCES

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

Copas, J.B. (1995). Local likelihood based on kernel censoring. *Jour. Royal Statist. Society* Series B, **57**, 221–235.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *Jour. Royal Statist. Society* Series B, **4**, 187–220.

Eubank, R. and Speckman, P. (1993). Confidence Bands in Nonparametric Regression. *Jour. Amer. Statist. Assoc.*, **88**, 1287–1301.

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

Fan, J., Gijbels, I. and King, M. (1995). Local likelihood and local partial likelihood in hazard regression. *Technical report 95.3*, Department of Statistics, Chinese University of Hong Kong.

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

Friedman, J.H. and Stuetzle, W. (1981). Projection Pursuit Regression. *Jour. Amer. Statist. Assoc.*, **76**, 817–823.

Hjort, N.L. (1991). Semiparametric estimation of parametric hazard rates. In *Survival Analysis: State of the Art*, Kluwer, Dordrecht, 211–236. Proceedings of the *NATO Advanced Study Workshop on Survival Analysis and Related Topics*, eds. P.S. Goel and J.P. Klein.

Hjort, N.L. (1995). Dynamic likelihood hazard rate estimation. *Biometrika*, to appear.

Hjort, N.L. and Glad, I.K. (1995). Nonparametric density estimation with a parametric start. *Ann. Statist.*, **23**, 882–904.

Hjort, N.L. and Jones, M.C. (1996). Locally parametric nonparametric density estimation. *Ann. Statist.*, to appear.

Jones, M.C. (1994). On close relations of local likelihood density estimation. *Manuscript.*

Loader, C.R. (1995). Local likelihood density estimation. *Ann. Statist.*, to appear.

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

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Jour. Royal Statist. Society* Series A, **135**, 370–384.

Staniswalis, J.G. (1989). The kernel estimate of a regression function in likelihood-based models. *Jour. Amer. Statist. Assoc.*, **84**, 276–283.

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.

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

Appendix

Proof of (5.5). The Fisher scoring method for computing the local maximum likelihood estimator is to update $\widehat{\beta}$ via

$$\widehat{\beta} = \beta_0 - [E\{L_p''(\beta_0; h, x_0)|\mathbb{X}\}]^{-1} L_p'(\beta_0; h, x_0). \tag{A.1}$$

Using the continuity assumption of the conditional expectation, it follows that

$$\begin{aligned} E\{L_p''(\beta_0; h, x_0)|\mathbb{X}\} &= \sum_{i=1}^n E\{\ell''(\mathbf{X}_i^T \beta_0, Y_i)|X_i\} \mathbf{X}_i \mathbf{X}_i^T K_h(X_i - x_0) \\ &\approx E\{\ell''(g(x_0), Y)|X = x_0\} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T K_h(X_i - x_0). \end{aligned} \tag{A.2}$$

Note that even though the computational expectation is computed approximately, the algorithm is exact, namely it converges to the local MLE even when we use the approximated Hessian matrix (A.2). Therefore, substituting (A.2) into (A.1) and using (2.2), we have

$$\widehat{\beta} = [\sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T K_h(X_i - x_0)]^{-1} \sum_{i=1}^n Z_i \mathbf{X}_i K_h(X_i - x_0),$$

namely, $\widehat{\beta}$ is obtained via regressing Z_i on X_i using locally the polynomial of order p.