# An Asymptotic Study of Variable Bandwidth Selection for Local Polynomial Regression with Application to Density Estimation


Jianqing Fan†
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

Tien-Chung Hu
Institute of Mathematics
National Tsing Hua University
Hsinchu, Taiwan 30043, R.O.C.

Li-Shan Huang
Department of Statistics
University of North Carolina
Chapel Hill, N.C. 27599-3260


June 2, 1994


**Abstract**

A decisive question in nonparametric smoothing techniques is the choice of the bandwidth or smoothing parameter. The present paper addresses this question when using local polynomial approximations for estimating the regression function and its derivatives. A fully-automatic bandwidth selection procedure has been proposed by Fan and Gijbels (1995), and the empirical performance of it was tested in detail via a variety of examples. Those experiences supported the methodology towards a great extend. In this paper we establish asymptotic results for the proposed variable bandwidth selector. We provide the rate of convergence of the bandwidth estimate, and obtain the asymptotic distribution of its error relative to the theoretical optimal variable bandwidth. Those asymptotic properties give extra support to the developed bandwidth selection procedure. It is also demonstrated how the proposed selection method can be applied in the density estimation setup. Some examples illustrate this application.


---


† Supported by NSF Grant DMS-9203135, an NSF postdoctoral fellowship, and a grant of the research council of UNC Chapel Hill.

‡ Supported by the National Science Foundation, Belgium, and the NSC, R.O.C.

*Abbreviated title*: Asymptotics for bandwidth selection.

*AMS 1991 subject classification.* Primary 62G07. Secondary 62J02, 60F05.

*Key words and phrases*: Assess of bias and variance, asymptotic normality, binning, density estimation, local polynomial fitting, variable bandwidth selector.

# 1 Introduction

In this paper primary interest focuses on studying the regression relationship between two variables $X$ and $Y$. In nonparametric estimation no prior assumption is made about the form of the regression function: the data itself will determine this form. Various smoothing techniques can be used to detect the underlying structure, but each of them involves the decisive choice of a smoothing parameter or bandwidth. In this paper we discuss how to choose a bandwidth when local polynomial fitting is used. It was shown extensively in the literature that the local polynomial approximation method has various nice features, among which, nice minimax properties, satisfactory boundary behavior, applicability for a variety of design-situations and easy interpretability. See, for example, Stone (1977), Fan (1992, 1993), Fan and Gijbels (1992), Fan, Gasser, Gijbels, Brockmann and Engel (1993) and Ruppert and Wand (1995). The papers by Cleveland (1979) and Cleveland and Devlin (1988) contain ample of nice examples, in various fields of applications, illustrating the performance of locally-weighted regression. Hastie and Tibshirani (1986) discussed the particular class of locally-weighted running-line smoothers.

Assume that the bivariate data $(X_1, Y_1), \ldots, (X_n, Y_n)$ form an i.i.d. sample from the model

$$Y = m(X) + \sigma(X)\varepsilon, \qquad E\varepsilon = 0 \quad \text{and} \quad \mathrm{Var}(\varepsilon) = 1,$$

where $X$ and $\varepsilon$ are independent random variables. The objective is to estimate the regression function $m(x) = E(Y|X = x)$ and its derivatives, based on the observations $(X_1, Y_1), \cdots, (X_n, Y_n)$. The marginal density of $X$ will be denoted by $f_X(\cdot)$. The above location-scale model is not a necessity, but a convenient way to introduce the notations. The i.i.d. assumption on the bivariate data suffices.

Suppose that the $(p+1)^{th}$-derivative of $m(x)$ at the point $x_0$ exists. We then approximate $m(x)$ locally by a polynomial of order $p$:

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) + \cdots + m^{(p)}(x_0)(x - x_0)^p/p!, \tag{1.1}$$

for $x$ in a neighborhood of $x_0$, and do a local polynomial regression fit

$$\min_{\beta} \sum_{i=1}^{n} \left( Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \right)^2 K\left( \frac{X_i - x_0}{h} \right), \tag{1.2}$$

where $\beta = (\beta_0, \cdots, \beta_p)^T$. Here $K(\cdot)$ denotes a nonnegative weight function and $h$ is a smoothing parameter — determining the size of the neighborhood of $x_0$. Let $\{\hat{\beta}_\nu\}$ denote the solution to the weighted least squares problem (1.2). Then it is obvious from the Taylor expansion in (1.1) that $\nu!\hat{\beta}_\nu$ estimates $m^{(\nu)}(x_0), \nu = 0, \cdots, p$.

For convenience we introduce some matrix notations. Let $\mathbf{W}$ be the diagonal matrix of weights, with entries $W_i \equiv K\left(\frac{X_i - x_0}{h}\right)$. Denote by $\mathbf{X}$ the design matrix whose $(l, j)^{th}$ element is $(X_l - x_0)^{j-1}$ and put $\mathbf{y} = (Y_1, \cdots, Y_n)^T$. Then, the weighted least squares problem (1.2) reads as follows:

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

The solution vector is provided via ordinary least squares and is given by

$$\hat{\beta} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}. \tag{1.3}$$

An important issue for the performance of the estimator $\hat{\beta}_\nu$ is the choice of the smoothing parameter $h$. A constant bandwidth can be sufficient if the unknown regression function behaves quite smoothly. In other situations a local variable bandwidth (which changes with the location point $x_0$) is a necessity. The problem of choosing a smoothing parameter received a lot of attention in the literature. In the nonparametric regression setting Härdle and Marron (1985) considered a constant cross-validated bandwidth and Rice (1986) used a criterion based on the residual sum of squares and the estimated variance. Müller and Stadtmüller (1987) studied the choice of a local variable bandwidth for convolution type regression estimators. A method for automatic selection of a constant bandwidth for convolution type kernel estimators of derivatives of a regression function was proposed in Müller, Stadtmüller and Schmitt (1987). Data-driven bandwidth selection rules based on "plug-in" techniques were considered in Gasser, Kneip and Köhler (1991), and Brockmann, Gasser and Hermann (1993). Vieu (1991) discussed a local cross-validated bandwidth selection rule. Ruppert, Sheather and Wand (1993) relied on plug-in techniques to derive a constant data-driven bandwidth for local polynomial regression. In the density estimation setting Sheather and Jones (1991) and Hall, Sheather, Jones and Marron (1991) discussed data-driven bandwidth selection using plug-in techniques. An interesting survey on recent developments in bandwidth selection for density estimation is given in Jones, Marron and Sheather (1992).

Automatic procedures for selecting constant and local variable bandwidths in local polynomial regression have been developped recently by Fan and Gijbels (1995). The basic idea for their procedure is to assess the conditional Mean Squared Error (MSE) of the estimators via deriving good finite sample estimates of this theoretical conditional MSE. See also Section 2. Working with the conditional Mean Squared Error means that one opts for finding an optimal bandwidth for the given covariate values. If the unconditional Mean Squared Error is used, then one focuses on a bandwidth that is optimal for the average of many repeated experiments. This is less attractive from phylosophical viewpoint and creates unnecessary technicalities. Another possible measure of performance is the weighted Average Squared Error $n^{-1}\sum_{i=1}^{n}\left(\hat{m}(X_i)-m(X_i)\right)^2 w_{n,x}(X_i)$, with $w_{n,x}(\cdot)$ a nonnegative weight function, which is a discretized version of the Integrated Squared Error. Asymptotically it would make no difference whether MSE or ASE is used as a measure of performance. Phylosophical arguments on the favorability of the conditional MSE versus the ASE is beyond our scope, but the former is more attractive in our setup since it is conceptually simpler to estimate.

The performance of the selection methodology proposed by Fan and Gijbels (1995) was investigated in detail via simulation studies. In the present paper we provide further theoretical foundations for the proposed bandwidth selection rule. More precisely, we establish the rate of convergence of the bandwidth selector as well as the asymptotic distribution of its error relative to the theoretical optimal variable bandwidth. Furthermore, we show how to convert a density estimation problem into a regression problem. This inner connection makes it possible to apply the regression techniques for density estimation.

The organization of the paper is as follows. In the next section, we explain briefly the bandwidth selection rule and present the asymptotic results for the estimated optimal bandwidth. In Section 3 we discuss how the bandwidth selection procedure can be applied in density estimation. Some examples illustrate the application in this setup. The last section contains the proofs of the theoretical results established in Section 2.

# 2 Bandwidth selection method and main results

Let $\hat{\beta}$ be the vector of the local polynomial regression estimates, defined in (1.3). Clearly, its bias and variance, conditionally upon $\underset{\sim}{X} = \{X_1, \cdots, X_n\}$, are given by

$$\begin{aligned} \text{Bias}(\hat{\beta}|\underset{\sim}{X}) &= \text{E}(\hat{\beta}|\underset{\sim}{X}) - \beta = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\mathbf{m} - \mathbf{X}\beta) \\ \text{Var}(\hat{\beta}|\underset{\sim}{X}) &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}^T\Sigma\mathbf{X})(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}, \end{aligned} \tag{2.1}$$

where $\mathbf{m} = (m(X_1), \ldots, m(X_n))^T$, and $\Sigma = \text{diag}\left(K^2(\frac{X_i - x_0}{h})\sigma^2(X_i)\right)$.

Note that the expressions in (2.1) give the exact conditional bias and variance of the local polynomial fit. We next study asymptotic expansions of this bias vector and variance matrix. These expansions serve as a gateway to prove the asymptotic normality of the selected bandwidth. The following assumptions will be needed.

**Assumptions:**

(i) The kernel $K$ is a bounded and continuous density function with finite $(2p+5)^{th}$ moment;

(ii) $f_X(x_0) > 0$ and $f''_X(x)$ is bounded in a neighborhood of $x_0$;

(iii) $m^{(p+3)}(\cdot)$ exists and is continuous in a neighborhood of $x_0$;

(iv) $\sigma^2(\cdot)$ has a bounded second derivative in a neighborhood of $x_0$;

(v) $m^{(p+1)}(x_0) \neq 0$ and the kernel $K$ is symmetric.

In the sequel we will use the following notations. Let $\mu_j = \int u^j K(u)du$ and $\nu_j = \int u^j K^2(u)du$ and denote

$$S_p = (\mu_{i+j-2})_{1 \le i,j \le (p+1)}\ , \qquad \tilde{S}_p = (\mu_{i+j-1})_{1 \le i,j \le (p+1)}.$$
$$S_p^* = (\nu_{i+j-2})_{1 \le i,j \le (p+1)}\ , \qquad \tilde{S}_p^* = (\nu_{i+j-1})_{1 \le i,j \le (p+1)}.$$

Further, set $c_p = (\mu_{p+1}, \ldots, \mu_{2p+1})^T$ and $\tilde{c}_p = (\mu_{p+2}, \ldots, \mu_{2p+2})^T$.

The asymptotic expansions for the conditional bias and variance are provided in the next theorem.

**Theorem 1.** *Under Assumptions (i) — (iv), the bias vector and variance matrix of $\hat{\beta}$ have the following expansions:*

$$Bias(\hat{\beta}|\underset{\sim}{X}) = h^{p+1} H^{-1} \left[ \beta_{p+1} S_p^{-1} c_p + h\ b^*(x_0) + O_P\left( h^2 + \frac{1}{\sqrt{nh}} \right) \right] \qquad (2.2)$$

*and*

$$Var(\hat{\beta}|\underset{\sim}{X}) = \frac{\sigma^2(x_0)}{f_X(x_0)nh} H^{-1} \left[ S_p^{-1} S_p^* S_p^{-1} + h\, V^*(x_0) + O_P\left(h^2 + \frac{1}{\sqrt{nh}}\right)\right] H^{-1}, \quad (2.3)$$

*where* $H = diag(1, h, \ldots, h^p)$,

$$b^*(x_0) = \frac{f_X'(x_0)\beta_{p+1} + \beta_{p+2} f_X(x_0)}{f_X(x_0)} S_p^{-1}\tilde{c}_p - \frac{f_X'(x_0)}{f_X(x_0)}\beta_{p+1} S_p^{-1}\tilde{S}_p S_p^{-1} c_p,$$

*and*

$$\begin{aligned} V^*(x_0) &= \frac{2\sigma'(x_0) f_X(x_0) + \sigma(x_0) f_X'(x_0)}{\sigma(x_0) f_X(x_0)} S_p^{-1}\tilde{S}_p^* S_p^{-1} \\ &\quad - \frac{f_X'(x_0)}{f_X(x_0)} \left( S_p^{-1}\tilde{S}_p S_p^{-1} S_p^* S_p^{-1} + S_p^{-1} S_p^* S_p^{-1}\tilde{S}_p S_p^{-1}\right). \end{aligned}$$

**Remark 1.** For a symmetric kernel $K$, we will show in the proof of Theorem 2 that the $(\nu+1)^{th}$-diagonal element of $V^*(x_0)$ and the $(\nu+1)^{th}$-element of $b^*(x_0)$ are zero when $p-\nu$ is odd. In other words, if $p-\nu$ is odd, the second leading terms in the expansions of the bias and variance of $\hat{\beta}_\nu$ are zero. The requirement for odd $p-\nu$ is natural, since the odd order approximations outperform the even order ones, as demonstrated in Fan and Gijbels (1994) and Ruppert and Wand (1995).

**Remark 2.** Consider the special case that the regression function has two continuous derivatives in the neighborhood of $x_0$ and a local linear fit ($p=1$) is carried through. In this case the asymptotic expressions for the conditional bias and variance of the resulting estimator $\hat{m}(x_0) = \hat{\beta}_0$ are simply given by

$$\text{Bias}(\hat{\beta}_0|\underset{\sim}{X}) = \frac{1}{2}\mu_2 m''(x_0)h^2[1+o_P(1)] \qquad \text{and} \qquad \text{Var}(\hat{\beta}_0|\underset{\sim}{X}) = \frac{\nu_0\sigma^2(x_0)}{f_X(x_0)nh}[1+o_P(1)],$$

as was first established by Fan (1993). Here we specify further the rate of the second order term, namely $o_P(1) = O_P\left(h^2 + 1/\sqrt{nh}\right)$, under the assumption that $m$ has a fourth continuous derivative in a neighborhood of $x_0$.

We would like to mention also that the $O_P\left(h^2+1/\sqrt{nh}\right)$ term in Theorem 1 holds uniformly in $h$. Indeed, this can be shown using the same arguments as in the proof of Theorem 1 and the fact that $(nh^{j+1})^{-1}\sum_{i=1}^n (X_i-x_0)^j K((X_i-x_0)/h) = f_X(x_0)\mu_j + hf'_X(x_0)\mu_{j+1} + O_P\left(h^2+1/\sqrt{nh}\right)$, uniformly in $h$.

Based on Theorem 1, we can derive the rate of convergence for bandwidth selection. We need the following simple lemma.

**Lemma 1.** *Suppose that a function $M(h)$ has the following asymptotic expansion:*

$$M(h) = ah^{2(p+1-\nu)}\left(1+O_P(h^2+\frac{1}{\sqrt{nh}})\right)+\frac{b}{nh^{2\nu+1}}\left(1+O_P(h^2+\frac{1}{\sqrt{nh}})\right) \qquad \text{as } h\to 0.$$

*Let $h_{opt}$ be the minimizer of $M(h)$. Then*

$$h_{opt} = \left(\frac{(2\nu+1)b}{2(p+1-\nu)a\ n}\right)^{\frac{1}{2p+3}}\left(1+O_P\left(n^{-\frac{2}{2p+3}}\right)\right),$$

*and*

$$M(h_{opt}) = a^{1-s}b^s(2p+3)(2\nu+1)^{-(1-s)}[2(p+1-\nu)]^{-s}n^{-s}\left(1+O_P\left(n^{-\frac{2}{2p+3}}\right)\right),$$

*with $s=\frac{2(p+1-\nu)}{2p+3}$, provided that $a,\ b>0$.*

We now study the theoretical optimal variable bandwidth for estimating the $\nu^{th}$-derivative $m^{(\nu)}(x_0)$. Let

$$\mathrm{MSE}_\nu(x_0) = b_{p,\nu}^2(x_0) + V_{p,\nu}(x_0),$$

where $b_{p,\nu}(x_0)$ and $V_{p,\nu}(x_0)$ are respectively the $(\nu+1)^{th}$-element of $\mathrm{Bias}(\hat\beta|\underset{\sim}{X})$ and the $(\nu+1)^{th}$-diagonal element of $\mathrm{Var}(\hat\beta|\underset{\sim}{X})$. Define

$$h_{\nu,opt} = \arg\min_h \mathrm{MSE}_\nu(x_0).$$

The quantity $\mathrm{MSE}_\nu(x_0)$ is, in a sense, an ideal assess of the conditional MSE of $\hat\beta_\nu$, and $h_{\nu,opt}$ is the ideal bandwidth selector.

We have the following expression for this ideal optimal bandwidth.

**Theorem 2.** *Under Assumptions (i) — (v) it holds that*

$$\frac{h_{\nu,opt} - h_{\nu,o}}{h_{\nu,o}} = O_P(n^{-\frac{2}{2p+3}})$$

*provided that* $p - \nu$ *is odd, where, with* $e_{\nu+1}$ *the* $(p+1) \times 1$ *unit vector containing 1 on the* $(\nu+1)^{th}$*-position,*

$$h_{\nu,o} = \left( \frac{(2\nu+1)\sigma^2(x_0) e_{\nu+1}^T S_p^{-1} S_p^* S_p^{-1} e_{\nu+1}}{2(p+1-\nu) f_X(x_0) (\beta_{p+1} e_{\nu+1}^T S_p^{-1} c_p)^2 \, n} \right)^{\frac{1}{2p+3}} . \tag{2.4}$$

Hence the ideal optimal bandwidth $h_{\nu,opt}$ behaves in first order like $h_{\nu,o}$, the asymptotical optimal bandwidth. From (2.4) it is also clear that there is a relationship between the asymptotical optimal bandwidths $h_{\nu,o}$ and $h_{0,o}$, namely

$$h_{\nu,o} = \left( \frac{(2\nu+1)(p+1)}{p+1-\nu} \right)^{1/(2p+3)} \left( \frac{e_{\nu+1}^T S_p^{-1} S_p^* S_p^{-1} e_{\nu+1}}{e_1^T S_p^{-1} S_p^* S_p^{-1} e_1} \frac{(e_1^T S_p^{-1} c_p)^2}{(e_{\nu+1}^T S_p^{-1} c_p)^2} \right)^{1/(2p+3)} h_{0,o}.$$

So, $h_{\nu,o}$ can be obtained by multiplying $h_{0,o}$ by a known constant which only depends on the kernel function $K$ and $p$ and $\nu$. Hence an estimator for $h_{\nu,o}$ can be derived from an estimator for $h_{0,o}$ and the above relationship. This idea was exploited by Müller, Stadtmüller and Schmitt (1987) and led to the so-called factor method.

We next briefly motivate a natural estimator for assessing the bias and the variance in (2.1). Using a Taylor expansion around the point $x_0$, we have

$$\mathbf{m} - \mathbf{X}\beta \approx \left( \beta_{p+1}(X_i - x_0)^{p+1} \right)_{1 \le i \le n}, \quad \text{and} \quad \sigma(X_i) \approx \sigma(x_0).$$

Substituting this into (2.1) leads to

$$\begin{aligned} \text{Bias}(\hat{\beta} | \underset{\sim}{X}) &\approx \beta_{p+1} S_n^{-1} \left( s_{n,p+1}, \ldots, s_{n,2p+1} \right)^T \\ \text{Var}(\hat{\beta} | \underset{\sim}{X}) &\approx \sigma(x_0) S_n^{-1} (\mathbf{X}\mathbf{W}^2\mathbf{X}) S_n^{-1}, \end{aligned} \tag{2.5}$$

where

$$s_{n,j} = \sum_{i=1}^{n} (X_i - x_0)^j K\left(\frac{X_i - x_0}{h}\right) \text{ and } S_n = \mathbf{X}^T \mathbf{W} \mathbf{X} = (s_{n,i+j-2})_{1 \le i,j \le (p+1)}.$$

A natural estimate of the bias and the variance is obtained by estimating $\beta_{p+1}$ and $\sigma^2(x_0)$ in (2.5) from another least squares problem.

Let $\beta_{p+1}$ be estimated by using a local polynomial regression of order $r$ ($r > p$) with a bandwidth $h_*$. Further, estimate $\sigma^2(x_0)$ by the standardized residual sum of squares:

$$\hat{\sigma}^2(x_0) = \frac{1}{\operatorname{tr}(\mathbf{W}^*) - \operatorname{tr}\left((\mathbf{X}^{*^T} \mathbf{W}^* \mathbf{X}^*)^{-1} \mathbf{X}^* \mathbf{W}^{*^2} \mathbf{X}^*\right)} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 K\left(\frac{X_i - x_0}{h_*}\right),$$

where $\mathbf{X}^* = ((X_i - x_0)^j)_{n \times (r+1)}$ is the design matrix, $\mathbf{W}^* = \operatorname{diag}\left(K(\frac{X_i - x_0}{h_*})\right)$ is the matrix of weights and $\hat{\mathbf{y}} = \mathbf{X}^*(\mathbf{X}^{*^T} \mathbf{W} \mathbf{X}^*)^{-1} \mathbf{X}^{*^T} \mathbf{W}^* \mathbf{y}$ is the vector of "predicted" values after the local $r^{th}$-order polynomial fit. The estimated conditional MSE of $\hat{\beta}_\nu$ is naturally defined by

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

where $\hat{b}_{p,\nu}(x_0)$ and $\hat{V}_{p,\nu}(x_0)$ are respectively the $(\nu+1)^{th}$-element of the bias vector and the $(\nu+1)^{th}$-diagonal element of the variance matrix in (2.5), with $\beta_{p+1}$ and $\sigma(x_0)$ being estimated. More precisely, $\hat{b}_{p,\nu}(x_0)$ refers to the $(\nu+1)^{th}$-element of $\hat{\beta}_{p+1} S_n^{-1} \left(s_{n,p+1}, \ldots, s_{n,2p+1}\right)^T$ and $\hat{V}_{p,\nu}(x_0)$ is the $(\nu+1)^{th}$-diagonal element of $\hat{\sigma}^2(x_0)$ $S_n^{-1} \left(\mathbf{X}^T \mathbf{W}^2 \mathbf{X}\right) S_n^{-1}$.

Define the estimated optimal bandwidth as

$$\hat{h}_\nu = \arg\min_h \widehat{\text{MSE}}_\nu(x_0).$$

We now study the sampling properties of $\hat{h}_\nu$. Using the arguments provided in Theorems 1 and 2 on the set $\{|\hat{\beta}_{p+1}| \le c, |\hat{\sigma}(x_0)| \le c\}$, where $c$ is a large constant, we find that

$$\frac{\hat{h}_\nu - \hat{h}_{\nu,o}}{\hat{h}_{\nu,o}} = O_P(n^{-\frac{2}{2p+3}}), \tag{2.6}$$

where $\hat{h}_{\nu,o}$ is defined similarly as $h_{\nu,o}$ in equation (2.4), but now with $\beta_{p+1}$, respectively $\sigma^2(x_0)$, replaced by $\hat{\beta}_{p+1}$, respectively $\hat{\sigma}^2(x_0)$.

Using equation (2.6) and Theorem 2, we obtain

$$\frac{\hat{h}_\nu - h_{\nu,opt}}{h_{\nu,opt}} = \frac{\hat{h}_{\nu,o} - h_{\nu,o}}{h_{\nu,o}} + O_P(n^{-\frac{2}{2p+3}}),$$

and the following lemma.

**Lemma 2.** *If the estimators $\hat{\beta}_{p+1}$ and $\hat{\sigma}^2(x_0)$ are stochastically bounded away from zero and infinity, then we have for the relative error*

$$\frac{\hat{h}_\nu - h_{\nu,opt}}{h_{\nu,opt}} = \frac{(\hat{\sigma}^2(x_0)/\sigma^2(x_0))^{\frac{1}{2p+3}} - (\hat{\beta}^2_{p+1}/\beta^2_{p+1})^{\frac{1}{2p+3}}}{(\hat{\beta}^2_{p+1}/\beta^2_{p+1})^{\frac{1}{2p+3}}} + O_P(n^{-\frac{2}{2p+3}}).$$

**Remark 3.** Lemma 2 holds for any estimators $\hat{\beta}_{p+1}$ and $\hat{\sigma}^2(x_0)$ which are stochastically bounded away from zero and infinity. So, these estimators should not necessarily be obtained via local polynomial fitting.

In order to discuss the asymptotic distribution of $\dfrac{\hat{h}_\nu - h_{\nu,opt}}{h_{\nu,opt}}$, we make some further assumptions.

**Assumptions:**

(vi) $E(\varepsilon^4) < \infty$;

(vii) The bandwidth $h_*$ satisfies $n^{-\frac{1}{2p+3}} << h_* << n^{-\frac{2p-1}{(2p+3)^2}}$.

Remark that $\hat{\sigma}^2(x)$ converges to $\sigma^2(x)$ much faster than $\hat{\beta}_{p+1}$ to $\beta_{p+1}$. Thus, it follows from Lemma 2 that the behavior of $\hat{\beta}_{p+1}$ dictates that of the relative error. For this reason, we introduce the following notations. Let $\alpha$ be the $(p+2)^{th}$-element of the vector $S_r^{-1}c_r$ and $\gamma$ be the $(p+2)^{th}$-diagonal element of the matrix $S_r^{-1}S_r^*S_r^{-1}$.

Remark that $\alpha$ and $\gamma$ are the constant factors in the asymptotic bias and variance expressions of $\hat{\beta}_{p+1}$ using the local $r^{th}$-order polynomial regression.

**Theorem 3.** *Suppose that $p-\nu$ is odd. Then, under Assumptions (i)- (vii),*

$$\sqrt{nh_*^{2p+3}}\left(\frac{\hat{h}_\nu - h_{\nu,opt}}{h_{\nu,opt}} + \frac{2\alpha\beta_{r+1}}{(2p+3)\beta_{p+1}}h_*^{r-p}\right) \xrightarrow{D} N\left(0, \frac{4\gamma\sigma^2(x_0)}{(2p+3)^2\beta_{p+1}^2 f_X(x_0)}\right). \quad (2.7)$$

**Remark 4.** The asymptotic distribution of the relative error is independent of $\nu$. When $r = p+2$ and $h_* \sim n^{-\frac{1}{2r+3}}$, the relative error is of order $O_P(n^{-\frac{2}{2r+3}})$. Specifically the relative error is of order $O_P(n^{-2/9})$ when $p = 1$.

The above result informs us about the asymptotic behavior of $\hat{h}_\nu$ in the case that $m^{(p+1)}(x_0) \neq 0$ (see Assumption (v)). The asymptotic behavior of $\hat{h}_\nu$ when $m^{(p+1)}(x_0) = 0$ has not been established yet. In order to derive such a result a higher order expansion of the conditional bias of the estimator is needed. Indeed in that case both leading terms in the asymptotic expression for the conditional bias are zero for $p-\nu$ odd. We do not investigate the selection of $\hat{h}_\nu$ when $p-\nu$ is even since, as mentioned before, these polynomial approximations are unfavorable.

To illustrate Theorem 3 let us consider the special case of estimating the regression function ($\nu = 0$) using a local linear fit ($p = 1$). Taking for example $r = 3$ we find via straightforward calculations that $\alpha = (\mu_6 - \mu_2\mu_4)/(\mu_4 - \mu_2^2)$ and $\gamma = (\mu_2^2\nu_0 - 2\mu_2\nu_2 + \nu_4)/(\mu_4 - \mu_2^2)^2$. Then, if the bandwidth $h_*$ satisfies $n^{-1/5} << h_* << n^{-1/25}$, Theorem 3 provides the following asymptotic normality result for $\hat{h}_0$:

$$\sqrt{nh_*^5}\left(\frac{\hat{h}_0 - h_{0,opt}}{h_{0,opt}} + \frac{\alpha m^{(4)}(x_0)}{30m''(x_0)}h_*^2\right) \xrightarrow{D} N\left(0, \frac{16\gamma\sigma^2(x_0)}{25(m''(x_0))^2 f_X(x_0)}\right).$$

In Theorem 3 we do not specify a method for selecting $h_*$. A data-driven procedure for choosing $h_*$ is provided in Fan and Gijbels (1995) via a RSC (Residual Sum of Squares)-criterion. See that paper for details of the particular criterion, as well as for details on the implementation of selecting a variable bandwidth, including the detailed description of an algorithm. There the reader can also find simulation studies illustrating the behavior of the relative error $(\hat{h}_\nu - h_{\nu,opt})/h_{\nu,opt}$, which converges very fast.

# 3 Application to Density Estimation

The bandwidth selection procedure for the regression problem can also be applied to the density estimation setup. Let $X_1, \ldots, X_n$ be independent identically distributed random variables. The interest is to estimate the common density $f(x)$ and its derivatives on an interval $[a, b]$. Partition the interval $[a, b]$ into $N$ subintervals $\{I_k, k = 1, \ldots, N\}$. Let $x_k$ be the center of $I_k$ and $y_k$ be the proportion of data $\{X_i, i = 1, \ldots, n\}$, falling in the partition $I_k$, divided by the bin length. Then, we use a local polynomial of order $p$ to fit the data $\{(x_k, y_k), k = 1, \ldots, N\}$. Let $\hat{\beta}_0, \ldots, \hat{\beta}_p$ be the solution of the local polynomial regression problem:

$$\sum_{k=1}^{N}(y_k - \sum_{j=0}^{p} \beta_j (x_k - x)^j)^2 K\left(\frac{x_k - x}{h}\right).$$

Then, the estimator for the $\nu^{th}$ derivative is

$$\hat{f}^{(\nu)}(x) = \nu! \hat{\beta}_\nu.$$

In order to have the same asymptotic properties for $\hat{f}^{(\nu)}(x)$ as those available for the kernel density estimator, we require that $Nh \longrightarrow \infty$ (see Cheng, Fan, and Marron (1993)). Now, the bandwidth selection procedure discussed in the previous section can also be applied to this setting. We would expect the resulting bandwidth selector to have the following property (compare with (2.7)):

$$\sqrt{nh_*^{2p+3}}\left(\frac{\hat{h}_\nu - h_{\nu,opt}}{h_{\nu,opt}} + \frac{2\alpha(p+1)! f^{(r+1)}(x)}{(2p+3)(r+1)! f^{(p+1)}(x)} h_*^{r-p}\right)$$

$$\xrightarrow{D} N\left(0, \frac{4\gamma}{(2p+3)^2(f^{(p+1)}(x)/(p+1)!)^2}\right).$$

We now illustrate the data-driven density estimation procedure via four simulated examples. These four densities were chosen from the 15 normal mixture densities discussed in Marron and Wand (1992). They were recommended by J.S. Marron in a private conversation, on the ground that they range from densities easy to estimate to densities very hard to estimate. In the implementation, we took $N = n/4$ and used Anscombe's (1948) variance-stabilizing transformation to the bin count $c_k = n\ y_k \times$ length$(I_k)$:

$$y_k^* = 2\sqrt{c_k + \frac{3}{8}}.$$

Let $\hat{\beta}_0^*$ be the local linear regression smoother for the transformed data. Then the density estimator is obtained by considering the inverse of Anscombe's transformation:

$$\hat{f}(x) = C\left(\frac{(\hat{\beta}_0^*)^2}{4} - \frac{3}{8}\right)_+,$$

where the constant $C$ is a normalization constant such that $\hat{f}(x)$ has area 1. This transformation step stabilized our simulation results somewhat, but did not result into a significant improvement. The variable bandwidth is, as described above, selected by using the techniques developed in Fan and Gijbels (1995) for the regression setup. Another possible, but more complicated, way of converting the density estimation problem into a regression problem is given in Wei and Chu (1993). We have no intention to recommend the scientific community to use regression techniques for density estimation. Here, we only point out the inner connection between nonparametric regression and density estimation. This connection makes that nonparametric regression techniques such as data-driven bandwidth selection are also available for density estimation. These include procedures proposed by Gasser, Kneip and Köhler

(1991) and those in Fan and Gijbels (1995) — in particular the constant bandwidth selection procedures such as RSC- and Refined bandwidth- selectors.

The four simulated examples concern a Gaussian density, a Strongly Skewed density, a Separated Bimodal density and a Smooth Comb density. Those correspond with respectively Densities #1, #3, #7 and #14 in Marron and Wand (1992). For each of the simulated examples we used the Epanechnikov kernel and estimated the density using a local linear approximation, based on samples of size 200 and 800. The number of simulations was 400. For each estimated curve we calculated the conditional Mean Squared Error over all grid points in which the curve was evaluated. We then ranked the 400 estimated curves according to this global measure. As representatives of all estimated curves we took the curves corresponding to the $10^{th}\%$, the $50^{th}\%$ and the $90^{th}\%$ among those rank-observations. Each of the presented pictures contains the true density function (the solid line) and three typical estimated curves (other line types).

*Put Figures 1.1 — 4.2 about here*

Figures 1.1 and 1.2: *A Gaussian Density (solid curve) and three estimated curves (dashed lines), based on* $n = 200$*, respectively* $n = 800$.
Figures 2.1 and 2.2: *A Strongly Skewed Density (solid curve) and three estimated curves (dashed lines), based on* $n = 200$*, respectively* $n = 800$.
Figures 3.1 and 3.2: *A Separated Bimodal Density (solid curve) and three estimated curves (dashed lines), based on* $n = 200$*, respectively* $n = 800$.
Figures 4.1 and 4.2: *A Smooth Comb Density (solid curve) and three estimated curves (dashed lines), based on* $n = 200$*, respectively* $n = 800$.

The Gaussian density in Example 1 and the Separated Bimodal density in Example 3 are easiest to estimate, and hence estimates based on samples of size 200 show a good performance. The densities in Examples 2 and 4 (i.e. the Strongly Skewed density and the Smooth Comb density) form a more difficult task. Estimates based on samples of size 200 do not succeed in capturing the sharp peaks appearing in the densities. Samples of size 800 result into better estimated curves.

# 4 Proofs

**Proof of Theorem 1**. We first remark that

$$\begin{aligned}
s_{n,j} &= \sum_{i=1}^{n}(X_i - x_0)^j K\left(\frac{X_i - x_0}{h}\right) \\
&= nh^{j+1}\int u^j K(u) f_X(x_0 + hu)du + O_P\left(\sqrt{nE\left[(X_1 - x_0)^{2j}K^2(\frac{X_1 - x_0}{h})\right]}\right) \\
&= nh^{j+1}\left(f_X(x_0)\mu_j + hf'_X(x_0)\mu_{j+1} + O_P(a_n)\right), \qquad (4.1)
\end{aligned}$$

where $a_n = h^2 + \frac{1}{\sqrt{nh}}$. Thus

$$S_n = nhH\left(f_X(x_0)S_p + hf'_X(x_0)\tilde{S}_p + O_P(a_n)\right)H. \qquad (4.2)$$

It follows from equation (2.1) and Taylor's expansion that

$$\begin{aligned}
&\text{Bias}(\hat{\beta}|\underset{\sim}{X}) \\
&= S_n^{-1}\mathbf{X}^T W\left(\beta_{p+1}(X_i - x_0)^{p+1} + \beta_{p+2}(X_i - x_0)^{p+2} + O_P\left((X_i - x_0)^{p+3}\right)\right)_{1\le i\le n} \\
&= S_n^{-1}\left(\beta_{p+1}c_n + \beta_{p+2}\tilde{c}_n + o_P(h^{p+3})\right), \qquad (4.3)
\end{aligned}$$

where $c_n = (s_{n,p+1}, \ldots, s_{n,2p+1})^T$ and $\tilde{c}_n = (s_{n,p+2}, \ldots, s_{n,2p+2})^T$. Combination of equations (4.2) and (4.3) leads to

$$\begin{aligned}
\text{Bias}(\hat{\beta}|\underset{\sim}{X}) &= H^{-1}\left(f_X(x_0)S_p + hf'_X(x_0)\tilde{S}_p + O_P(a_n)\right)^{-1} \\
&\quad \times h^{p+1}\left(\beta_{p+1}f_X(x_0)c_p + h\left(f'_X(x_0)\beta_{p+1} + \beta_{p+2}f_X(x_0)\right)\tilde{c}_p + O_P(a_n)\right).
\end{aligned} \qquad (4.4)$$

Remark that

$$(A + hB)^{-1} = A^{-1} - hA^{-1}BA^{-1} + O(h^2). \qquad (4.5)$$

Applying this to equation (4.4) and using some simple algebra, we obtain the bias expression (2.2) in Theorem 1.

For the variance matrix remark from (2.1) that

$$\mathrm{Var}(\hat{\beta}|\underset{\sim}{X}) = S_n^{-1} S_n^* S_n^{-1}, \tag{4.6}$$

where $S_n^* = (s_{n,i+j-2}^*)_{1 \le i,j \le (p+1)}$, with $s_{n,j}^* = \sum_{i=1}^n (X_i - x_0)^j K^2\left(\frac{X_i - x_0}{h}\right)\sigma^2(X_i)$. Following similar arguments as in (4.1) we find that,

$$s_{n,j}^* = nh^{j+1}\left(g(x_0)\nu_j + hg'(x_0)\nu_{j+1} + O_P(a_n)\right), \tag{4.7}$$

where $g(x_0) = f_X(x_0)\sigma^2(x_0)$. From (4.6), (4.7) and (4.2) we obtain that

$$\begin{aligned} \mathrm{Var}(\hat{\beta}|\underset{\sim}{X}) &= \frac{nhg(x_0)}{(f_X(x_0)nh)^2} H^{-1}\left(S_p + \frac{hf_X'(x_0)}{f_X(x_0)}\tilde{S}_p + O_P(a_n)\right)^{-1} \\ &\quad \times \left(S_p^* + h\frac{g'(x_0)}{g(x_0)}\tilde{S}_p^* + O_P(a_n)\right)\left(S_p + \frac{hf_X'(x_0)}{f_X(x_0)}\tilde{S}_p + O_P(a_n)\right)^{-1} H^{-1}. \end{aligned}$$

The conclusion in (2.3) now follows from equation (4.5) and some simple algebra. □

**Proof of Lemma 1**. Let

$$h_o = \left(\frac{(2\nu+1)b}{2(p+1-\nu)a\, n}\right)^{\frac{1}{2p+3}}.$$

Then

$$M(h_o) = a^{1-s}b^s(2p+3)(2\nu+1)^{-(1-s)}[2(p+1-\nu)]^{-s}n^{-s}(1 + O_P(n^{-\frac{2}{2p+3}})).$$

It is easy to see that for any $h$ such that $|h/h_o - 1| > Cn^{-\frac{2}{2p+3}}$, $M(h) > M(h_o)$ when $C$ is large. In other words, such an $h$ can not be a minimizer of $M(h)$. Hence $|h/h_o - 1| \le Cn^{-\frac{2}{2p+3}}$. This proves the first conclusion. The second conclusion follows directly from the first. □

**Proof of Theorem 2**. Since the kernel $K$ is symmetric, $\mu_{2j+1} = 0$, for $j = 0, 1, \ldots, (p+1)$. Thus , the matrix $S_p = (\mu_{i+j-2})_{1\le i,j\le (p+1)}$ has the following structure:

$$S_p = \begin{pmatrix} \times & 0 & \times & 0 & \ldots \\ 0 & \times & 0 & \times & \ldots \\ \times & 0 & \times & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where "$\times$" denotes any number. It can be shown that $S_p^{-1}$ will have exactly the same structure as $S_p$, by examining the determinant of the accompanying submatrix: this submatrix contains $[(p+1)/2]$ rows which each have at most $[(p+1)/2] - 1$ nonzero elements. Thus, these rows are linearly dependent and the determinant of the accompanying submatrix must be zero. Therefore, each odd-position (e.g. (3,4)) -element is zero. Similar arguments hold for the matrix

$$\tilde{S}_p = (\mu_{i+j-1})_{1\le i,j\le (p+1)} = \begin{pmatrix} 0 & \times & 0 & \times & \ldots \\ \times & 0 & \times & 0 & \ldots \\ 0 & \times & 0 & \times & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Using the sparsity structure of the matrices $S_p$ and $\tilde{S}_p$, and of $c_p$ and $\tilde{c}_p$, we can see that the $(\nu+1)^{th}$-element of $S_p^{-1}\tilde{c}_p$ and of $S_p^{-1}\tilde{S}_p S_p^{-1} c_p$ are zero. Thus by Theorem 1,

$$b_{p,\nu}(x_0) = \beta_{p+1} h^{p+1-\nu} e_{\nu+1}^T S_p^{-1} c_p \left(1 + O_P(h^2 + \frac{1}{\sqrt{nh}})\right). \tag{4.8}$$

A similar argument shows that the $(\nu+1)^{th}$-diagonal element of $V^*(x_0)$ in Theorem 1 is zero. Therefore, by Theorem 1,

$$V_{p,\nu}(x_0) = \frac{\sigma^2(x_0)}{f_X(x_0)nh} h^{-2\nu} e_{\nu+1}^T S_p^{-1} S_p^* S_p^{-1} e_{\nu+1} \left(1 + O_P(h^2 + \frac{1}{\sqrt{nh}})\right). \tag{4.9}$$

Combining equations (4.8) and (4.9) gives

$$\begin{aligned}\mathrm{MSE}_\nu(x_0) &= \left(\beta_{p+1}e_{\nu+1}^T S_p^{-1}c_p\right)^2 h^{2(p+1-\nu)}\left(1+O_P(h^2+\frac{1}{\sqrt{nh}})\right)\\ &\quad+\frac{\sigma^2(x_0)}{f_X(x_0)nh^{2\nu+1}}e_{\nu+1}^T S_p^{-1}S_p^* S_p^{-1}e_{\nu+1}\left(1+O_P(h^2+\frac{1}{\sqrt{nh}})\right).\end{aligned}$$

The result now follows directly from Lemma 1. □

**Proof of Theorem 3**. By Theorem 1 of Fan and Gijbels (1995),

$$E(\hat{\sigma}^2(x_0)|\underset{\sim}{X}) = \sigma^2(x_0) + O_P(h_*^{2(r+1)}),$$

and it can be shown that

$$\mathrm{Var}(\hat{\sigma}^2(x_0)|\underset{\sim}{X}) = O_P\left(\frac{1}{nh_*}\right).$$

Thus, using Assumption (vii), we obtain

$$\hat{\sigma}^2(x_0) = \sigma^2(x_0) + O_P\left(h_*^{2(r+1)} + \frac{1}{\sqrt{nh_*}}\right) = \sigma^2(x_0) + o_P((nh_*^{2p+3})^{-\frac{1}{2}}),$$

and hence

$$\left(\frac{\hat{\sigma}^2(x_0)}{\sigma^2(x_0)}\right)^{\frac{1}{2p+3}} = 1 + o_P\left((nh_*^{2p+3})^{-\frac{1}{2}}\right).$$

Now, by Theorem 1 of Fan, Heckman and Wand (1995),

$$\sqrt{\frac{nh_*^{2p+3}}{\gamma\sigma^2(x_0)/f_X(x_0)}}\left(\hat{\beta}_{p+1} - \beta_{p+1} - \alpha\beta_{r+1}h_*^{r-p}\right) \xrightarrow{D} N(0,1). \tag{4.10}$$

Note that equation (4.10) can also be proved directly by checking Lindeberg's condition. Evidently,

$$\left(\frac{\hat{\beta}_{p+1}}{\beta_{p+1}}\right)^{\frac{2}{2p+3}} = 1 + \frac{2(\hat{\beta}_{p+1}-\beta_{p+1})}{(2p+3)\beta_{p+1}} + O_P\left((\hat{\beta}_{p+1}-\beta_{p+1})^2\right).$$

Therefore,

$$\begin{aligned}&\sqrt{\frac{nh_*^{2p+3}}{\gamma\sigma^2(x_0)/f_X(x_0)}}\left(\left(\frac{\hat{\beta}_{p+1}}{\beta_{p+1}}\right)^{\frac{2}{2p+3}} - 1 - \frac{2\alpha\beta_{r+1}}{(2p+3)\beta_{p+1}}h_*^{r-p}\right)\\ &\qquad\xrightarrow{D} N\left(0, \left(\frac{2}{(2p+3)\beta_{p+1}}\right)^2\right).\end{aligned}$$

Hence, it follows from Lemma 2 that

$$\sqrt{\frac{nh_*^{2p+3}}{\gamma\sigma^2(x_0)/f_X(x_0)}}\left(\frac{\hat{h}_\nu - h_{\nu,opt}}{h_{\nu,opt}} + \frac{2\alpha\beta_{r+1}}{(2p+3)\beta_{p+1}}h_*^{r-p}\right)$$
$$\xrightarrow{D} N\left(0, \left(\frac{2}{(2p+3)\beta_{p+1}}\right)^2\right),$$

which completes the proof. □

**Acknowledgement**

The authors thank a referee for giving very valuable remarks, which led to an improvement of the paper.

## REFERENCES

Anscombe, F.J. (1948). The transformation of poisson, binomial and negative-binomial data. *Biometrika*, **35**, 246 – 254.

Brockmann, M., Gasser, T. and Hermann, E. (1993). Locally adaptive bandwidth choice for kernel regression estimators. *Jour. Amer. Statist. Assoc.*, to appear.

Cheng, M.Y., Fan, J. and Marron, J.S. (1993). Minimax efficiency of local polynomial fit estimators at boundaries. *Inst. of Statist. Mimeo Series* # 2098. University of North Carolina, Chapel Hill.

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

Cleveland, W.S. and Devlin, S.J. (1988). Locally-weighted regression: an approach to regression analysis by local fitting. *Jour. Amer. Statist. Assoc.* **83** 597 – 610.

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

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

Fan, J. , Gasser, T., Gijbels, I. , Brockmann, M. and Engel, J. (1993). Local polynomial fitting: a standard for nonparametric regression. *Institut de Statistique, Discussion Paper Series* # 9315, Université Catholique de Louvain, Louvain-la-Neuve.

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

Fan, J. and Gijbels, I. (1994). Adaptive order polynomial fitting: bandwidth robustification and bias reduction. *Institut de Statistique, Discussion Paper Series* # 9313 (revised version). Université Catholique de Louvain, Louvain-la-Neuve.

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

Fan, J. , Heckman, N. and Wand, M.P. (1995). Local Polynomial Kernel regression for generalized linear models and quasi-likelihood functions. *Jour. Amer. Statist. Assoc.*, to appear.

Gasser, T., Kneip, A. and Köhler, W. (1991). A flexible and fast method for automatic smoothing. *Jour. Amer. Statist. Assoc.*, **86**, 643 – 652.

Hall, P., Sheather, S.J., Jones, M.C. and Marron, J.S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*, **78**, 263 – 271.

Härdle, W. and Marron, J.S. (1985). Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression. *Biometrika*, **72**, 481 – 484.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models (with discussion). *Statist. Sci.*, **1**, 297 – 318.

Jones, M.C., Marron, J.S. and Sheather, S.J. (1992). Progress in data based bandwidth selection for kernel density estimation. Department of Statistics, University of North Carolina. *Mimeo Series # 2088.*

Marron, J.S. and Wand, M.P. (1992). Exact mean integrated squared error. *Ann. Statist.*, **20**, 712 – 736.

Müller, H.–G. and Stadtmüller, U. (1987). Variable bandwidth kernel estimators of regression curves. *Ann. Statist.*, **15**, 182 – 201.

Müller, H.–G., Stadtmüller, U. and Schmitt, T. (1987). Bandwidth choice and confidence intervals for derivatives of noisy data. *Biometrika*, **74**, 743 – 749.

Rice, J.A. (1986). Bandwidth choice for differentiation. *Jour. Mult. Anal.* , **19**, 251 – 264.

Ruppert, D. and Wand, M.P. (1995). Multivariate weighted least squares regression. *Ann. Statist.*, to appear.

Ruppert, D., Sheather, S.J. and Wand, M.P. (1993). An effective bandwidth selector for local least squares regression. *Manuscript.*

Sheather, S.J. and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Jour. Roy. Statist. Soc., Series B*, **53**, 683 – 690.

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

Vieu, P. (1991). Nonparametric regression: optimal local bandwidth choice. *Jour. Roy. Statist. Soc., Series B*, **53**, 453 – 464.

Wei, C.Z. and Chu, J.K. (1993). A regression point of view toward density estimation. *Manuscript.*