OPTIMAL BANDWIDTH SELECTION FOR
LOCAL LINEAR REGRESSION

BY

LI - SHAN HUANG* AND JIANQING FAN+

FSU STATISTICS REPORT M912
Optimal Bandwidth Selection for Local Linear Regression

Li-Shan Huang* and Jianqing Fan*

September 23, 1996

FSU Technical Report No. M-912

*Department of Statistics
Florida State University
Tallahassee, FL 32306-3033

+Department of Statistics
University of North Carolina
Chapel Hill, N.C. 27599-3260
Optimal Bandwidth Selection for Local Linear Regression

Jianqing Fan
Department of Statistics
University of North Carolina
Chapel Hill, N.C. 27599-3260

Li-Shan Huang
Department of Statistics
Florida State University
Tallahassee, F.L. 32306-3033

Abstract: A bandwidth selection method is proposed for local linear regression. Our approach is to combine the ideas of optimal bandwidth selection of Hall et al. (1991) in kernel density estimation, and use of direct bias and variance in Fan and Gijbels (1995) for local linear regression. We show that the bandwidth selector has an optimal relative rate of convergence of $n^{-1/2}$ with $n$ the sample size.

Key words and phrases. Bandwidth selection, local linear regression, kernel density estimation, convergence rate.
Abbreviated title: Bandwidth selection.

1 Introduction

Local linear regression is a method for curve estimation by fitting locally weighted least squares linear regression. Recent work on local linear regression includes Fan (1992, 1993), and Hastie and Loader (1993), etc. An extension of local linear regression, local polynomial regression, is discussed in Ruppert and Wand (1994) and the monograph by Fan and Gijbels (1996) and the references therein.

One critical and inevitable step for fitting local regression methods is the choice of the smoothing parameter $h$, or called "bandwidth", which controls the degree of smoothing. Different values of the bandwidth results in different estimated curves. A natural method for selecting the bandwidth is to plot out several curves and choose one estimate subjectively, according to one's preference. From the model selection point of view, choosing a bandwidth is equivalent to selecting a model for the data, in the sense that when $h$ is chosen to be a fairly large number (or theoretically $h = \infty$), ordinary global polynomial regression is fitted since all data points receiving same weights; when $h$ is close to zero, modeling degenerates to data interpolation. Hence one has infinity number of models ($h \in (0, \infty)$) to choose from. It is desirable to provide a data-driven and automatic procedure that suits the practical needs and has a fast convergence rate to the true bandwidth that is unknown in practice.
The problem of bandwidth selection has stimulated much research especially in kernel density estimation. The approaches include cross-validation and "plug-in" methods; see the review paper by Jones, Sheather, and Marron (1996). Particularly, Hall et al. (1991) consider this problem in an asymptotically framework and the result is a fast $n^{-1/2}$-consistent bandwidth selector for kernel density estimates with $n$ the sample size. Based on the same idea, Chiu (1991a, 1991b) constructs $n^{-1/2}$-rate bandwidth selectors for the kernel-type estimators of density and regression functions. Ruppert, Sheather, and Wand (1996) propose a bandwidth selection method for local linear regression. Their approach is based on the "plug-in" idea for the unknown quantities of the asymptotically optimal bandwidth, and has a relative rate of convergence is $n^{-5/14}$. Hart and Yi (1996) focus on a non-random design model and study a cross-validation type bandwidth selector with rate $n^{-3/10}$, while Schucany (1995) study local bandwidth selection for Priestly-Chao kernel estimators (Priestly and Chao (1972)) with possible extension to local linear regression. Opsomer (1995) give a "plug-in" bandwidth selection method for additive models with local linear fitting. Fan and Gijbels (1995) study the problem of variable bandwidth for local polynomial fitting with the appealing idea of assessing the "exact" bias and variance expressions. This approach saves the effort of estimating unknown terms related to the design density. Note that the estimation would be necessary if one uses the asymptotic expression of the optimal bandwidth such as the "plug-in" methods.

In this paper, we combine the ideas of Hall et al. (1991) and Fan and Gijbels (1995) to propose a fast $n^{-1/2}$-rate bandwidth selector for local linear regression. Define the optimal bandwidth being the one that minimizes the weighted mean integrated square error (see (3.1)) of local linear regression. A second-order asymptotic expansion of the bias and variance is shown in Theorem 1. Note that the expansion is necessary for the $n^{-1/2}$ rate of convergence, as in Hall et al. (1991) for kernel density estimates. The resulting asymptotically optimal bandwidth involves many unknown quantities (see (3.3)), and thus, unlike the density estimation problem, the "plug-in" approach is not feasible in this case. Our attempt is to construct a bandwidth selector that does not depend on the asymptotic expressions, such as using the "exact" bias and variance by Fan and Gijbels (1995). We show that the asymptotic rate of convergence of the estimated bandwidth is to its best possible theoretical value $n^{-1/2}$ by applying the results of Huang and Fan (1995).
The paper is organized as follows. Section 2 gives some background theory of local polynomial fitting. Section 3 is on the asymptotic expansion of the optimal bandwidth for local linear regression. In Section 4 we propose a bandwidth selection method and give its theoretical justification. Section 5 contains some simulation results with discussions.

2 Local polynomial fitting

Let \((X_i, Y_i), i = 1, \ldots, n\), denote independent data generated by a random design model:

\[ Y = m(X) + \sigma(X)\varepsilon, \quad E(\varepsilon) = 0, \text{var}(\varepsilon) = 1, \quad (2.1) \]

where \(m(\cdot)\) is an unknown regression function, \(\varepsilon\) is an error variable, and \(X\) is independent of \(\varepsilon\) with density \(f(\cdot)\). The nonparametric regression problems are to estimate the conditional regression function \(m(x) = E(Y|X = x)\) and to find interesting features in the data. An effective method for estimating \(m(\cdot)\) and its derivatives is local polynomial fitting. Assume that the \((p + 1)\)-th derivative of \(m(\cdot)\) at point \(x\) exists. For \(X_i\)'s in a neighborhood of \(x\), a Taylor's expansion gives

\[ m(X_i) \approx m(x) + m'(x)(X_i - x) + \ldots + m^{(p)}(x)(X_i - x)^p / p!. \]

Let \(\beta_j(x) = m^{(j)}(x)/j!\) and \(\beta(x) = (\beta_0(x), \ldots, \beta_p(x))^T\). Fitting local polynomial of order \(p\) is to find the solution of \(\beta(x)\) to the following weighted least squares problem:

\[ \min_{\beta} \sum_{i=1}^{n} \left( Y_i - \sum_{j=0}^{p} \beta_j(X_i - x)^j \right)^2 K \left( \frac{X_i - x}{h} \right), \quad (2.2) \]

where \(K(\cdot)\) is a kernel function, usually a symmetric probability density, and \(h\) is the smoothing parameter which controls the size of neighborhood for local modeling. Note that the dependence of \(\beta\) on \(x\) is suppressed. It is easy to see that (2.2) can be written as

\[ \min_{\beta} (Y - X\beta)^T W(Y - X\beta), \]

where

\[ X = \begin{pmatrix} 1 & (X_1 - x) & \ldots & (X_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (X_n - x) & \ldots & (X_n - x)^p \end{pmatrix}, \quad (2.3) \]
\[ W = \text{diag}(K(\frac{X_i - x}{h}))_{n \times n} = \begin{pmatrix}
K(\frac{X_1 - x}{h}) \\
\vdots \\
K(\frac{X_n - x}{h})
\end{pmatrix}, \]

and \( Y = (Y_1, \ldots, Y_n)^T \). Hence the solution \( \hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^T \) is the usual weighted least squares estimates

\[ \hat{\beta} = (X^TWX)^{-1}X^TWY. \quad (2.4) \]

See for example, Ruppert and Wand (1994), and Fan and Gijbels (1996) for more details on local polynomial modeling and its applications. When \( p = 1 \) in (2.2), i.e. the case of fitting local linear regression, the solution \( \hat{\beta}_0(x) = \hat{m}_\ell(x) \) is the local linear regression estimator for \( m(x) \), and

\[ \hat{m}_\ell(x) = \frac{\sum_{i=1}^{n} (S_{n,2} - (X_i - x)S_{n,1}) K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^{n} (S_{n,2} - (X_i - x)S_{n,1}) K\left(\frac{X_i - x}{h}\right)}, \]

where \( S_{n,j} = \sum_{i=1}^{n} (X_i - x)^j K\left(\frac{X_i - x}{h}\right) \), \( j = 0, 1, \ldots \). Again, the dependence of \( S_{n,j} \) on \( x \) is suppressed.

### 3 Asymptotic expansion

Consider fitting local linear regression on an interval \([a, b]\). A criterion to measure the difference between the estimated curve \( \hat{m}_\ell(x) \) and the true regression function is the conditional weighted mean integrated square error (MISE) given by

\[ M(h) = E \left\{ \left( \int (\hat{m}_\ell(x) - m(x))^2 w(x) dx \right) | X_1, \ldots, X_n \right\}, \quad (3.1) \]

where \( w(\cdot) \) is a nonnegative weight function with support \([a, b]\). Define the optimal bandwidth \( h_{OPT} \) to be the minimizer of the \( M(h) \). The first-order expansion of \( h_{OPT} \) is well known; see for example Fan and Gijbels (1995), and Ruppert, Sheather, and Wand (1995) with \( w(x) = f(x) \). The goal of this section is to derive a second-order asymptotic expansion of \( h_{OPT} \).

We give in the following theorem the first and second order terms of the asymptotic bias and variance of \( \hat{m}_\ell(x) \), \( x \in [a, b] \). The technical assumptions are:
Conditions (A):

(A1) The regression function \( m(\cdot) \) is “smooth” with continuous derivatives to the 5-th order on \([a, b]\).

(A2) The design density \( f(\cdot) \) has a continuous second derivative and is bounded away from 0 on \([a, b]\).

(A3) The bandwidth \( h \) lies in an interval \((\alpha_1 n^{-t_1}, \alpha_2 n^{-t_2})\) such that \((\gamma_1 n^{-1/5}, \gamma_2 n^{-1/5}) \subset (\alpha_1 n^{-t_1}, \alpha_2 n^{-t_2})\), for some positive constants \( \alpha_1, \alpha_2, \gamma_1, \gamma_2, t_1 \) and \( t_2 \), and \( \alpha_1 < \alpha_2, \gamma_1 < \gamma_2, \) and \( 0 < t_1 \leq t_2 < 1 \).

(A4) The kernel function \( K(\cdot) \) is a bounded symmetric probability density of order 2 defined on a compact interval.

(A5) The variance function \( \sigma^2(\cdot) \) is continuous on \([a, b]\).

**Theorem 1** Under Conditions (A), the asymptotic bias and variance of \( \hat{m}_n(x) \) are:

\[
E(\hat{m}_n(x)|X_1, \ldots, X_n) - m(x) = h^2 \mu_2 \beta_2(x) + h^4 \{(\mu_2 - \mu_4) \beta_2(x) (f''(x)/f(x)) \}
- f'''(x)/(2f(x))) + \beta_4(x) \mu_4 \}
+ n^{-1/2} h^{3/2} \nu_{n,1}(\mu_2 \beta_2(x) f(x) + 1/f(x))
+ o_P(h^4 + n^{-1/2} h^{3/2}),
\]

\[
\text{var}(\hat{m}_n(x)|X_1, \ldots, X_n) = \sigma^2(x) \{n^{-1} h^{-1} \nu_0/f(x) + n^{-1} h(\nu_0 \mu_2 - \nu_2/2)
\times (2f''(x)/f^3(x) - f'''(x)/f^2(x))
+ n^{-3/2} h^{-3/2} \nu_{n,2}(2\nu_0 + f(x)^{-2}) \}
+ o_P(n^{-1} h + n^{-3/2} h^{-3/2}),
\]

(3.2)

where \( \mu_j = \int u^j K(u) du, \nu_j = \int u^j K^2(u) du, j = 0, 1, \ldots, \) and \( \nu_{n,1} = O_P(1) \) and \( \nu_{n,2} = O_P(1) \) are random variables depending on the design points \( X_i \)'s.

The proof of this theorem is given in the Appendix. Note that the expansion of the \( \text{var}(\hat{m}_n(x)|X_1, \ldots, X_n) \) (see (3.2)) shows that the \( n^{-1} \) term is 0 and hence the second leading term is of order \( n^{-1} h \). Remark that Hall et al. (1991) do not expand the variance for kernel density estimates since the \( n^{-1} \) term does not involve \( h \) and the term of order \( n^{-1} h \) is 0.
Writing $\theta_v = \int \beta_\nu^2(x)w(x)dx$, we show in Theorem 2 that the optimal bandwidth $h_{OPT}$ is asymptotically equivalent to

$$h_O = a_1(\beta_2, \sigma)n^{-1/5} + a_2(\beta_2, \beta_4, \sigma, \nu_{n,1}, \nu_{n,2})n^{-3/5} \tag{3.3}$$

with

$$a_1(\beta_2, \sigma) = \left(\nu_0 \int \sigma^2(x)f^{-1}(x)w(x)dx/4\mu_2^2\theta_2\right)^{1/5},$$

$$a_2(\beta_2, \beta_4, \sigma, \nu_{n,1}, \nu_{n,2}) = (-3a_1^3/5\mu_2^2\theta_2) \left(\mu_2^2 - \mu_4 \int \beta_2^2(x)dx\right) \times (f'^2(x)/f^2(x) - f''(x)/(2f(x)))w(x)dx \left[ \mu_2^2 + \mu_4 \int \beta_2(x)\beta_4(x)w(x)dx \right] + (a_1^2/20\mu_2^2\theta_2)(2\nu_0\mu_2 - \nu_2)

\left[ \int \sigma^2(x)(f'^2(x)/f^3(x) - f''(x)/(2f^2(x)))w(x)dx \right]

+ 7a_1^{1/2}\nu_{n,1} \left( \int \beta_2(x)(\beta_2(x)f(x) + 1/f(x))w(x)dx \right) / (2\theta_2)

- 1.5a_1^{-7/2}\nu_{n,2} \left( \int \sigma^2(x)(2\nu_0 + f(x)^{-2})w(x)dx \right) / (2\mu_2^2\theta_2). \tag{3.4}$$

**Theorem 2 Under Conditions (A),**

$$\frac{(h_{OPT} - h_O)}{h_{OPT}} = o_P(n^{-1/2}).$$

Hall et al. (1991) give a similar second-order approximation of the optimal bandwidth for the kernel density estimator; see equation (1) in the paper. Their asymptotically optimal bandwidth involves two unknown roughness functionals of the density, and hence a “plug-in” bandwidth selector is proposed via estimating the unknown roughness quantities. However, $h_O$ for local linear regression is much more complicated than that of kernel density estimates. It involves many unknown quantities such as the design density and the derivatives of the regression function. Thus we can not simply follow the path of Hall et al. (1991) by taking the “plug-in” approach. Another reason is that it may require a large sample size to show the strength of the second-order “plug-in” rule, as demonstrated by Jones, Marron, and Sheather (1996).
A bandwidth selector

In this section we propose a consistent bandwidth selector which mimics the \( n^{-1/2} \) asymptotic structure in Section 3 via estimating the unknown quantities implicitly. The homoscedastic model \(((2.1)\ \text{with} \ \sigma^2(x) = \sigma)\) is adopted here for simplicity.

Fan and Gijbels (1995) study adaptive bandwidth selection in local polynomial fitting. Instead of using the asymptotic bias and variance, their procedure involves using equation (2.4) to assess the bias and variance:

\[
\text{bias}(\hat{m}_\ell(x)|X_1, \ldots, X_n) = e_0^T (X_\ell^T W X_\ell)^{-1} X_\ell^T W (m - X_\ell \beta_0, \beta_1)^T,
\]

\[
\text{var}(\hat{m}_\ell(x)|X_1, \ldots, X_n) = \sigma^2 e_0^T (X_\ell^T W X_\ell)^{-1} (X_\ell^T W^2 X_\ell) (X_\ell^T W X_\ell)^{-1} e_0,
\]

(4.1)

where \( X_\ell \) is the design matrix for fitting local linear regression at point \( x \) (i.e. a special case of (2.3) when \( p = 1 \)) and \( e_0 = (1, 0)^T \). Note that the unknown terms in (4.1) are \((m - X_\ell \beta_0, \beta_1)^T\) and the variance parameter \( \sigma^2 \). Estimated MSE for \( \hat{m}_\ell(x) \) can then be obtained by approximating only \((m - X_\ell \beta_0, \beta_1)^T\) and \( \sigma^2 \). Estimation of the conditional variance \( \sigma^2 \) is studied by Rice (1984) and Hall, Kay and Titterington (1990), for example. Thus, variance estimation is not the goal of our study. To approximate \((m - X_\ell \beta_0, \beta_1)^T\), a Taylor expansion may be helpful:

\[
m - X_\ell \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \approx \begin{pmatrix} \beta_2 (X_1 - x)^2 + \beta_3 (X_1 - x)^3 + \beta_4 (X_1 - x)^4 \\ \vdots \\ \beta_2 (X_n - x)^2 + \beta_3 (X_n - x)^3 + \beta_4 (X_n - x)^4 \end{pmatrix}.
\]

(4.2)

Fan and Gijbels (1995) give a slightly different expression of (4.2) by expanding only to \( \beta_3(x)(X_i - x)^3 \) and claim that it has the advantage of computational simplicity while not far from \( n^{-1/2} \)-consistency. Here we expand to the term \( \beta_4(x)(X_i - x)^4 \) to ensure the fast \( n^{-1/2} \)-convergence.

From (4.1) and (4.2), we need only estimate \( \sigma^2, \beta_2(x), \beta_3(x), \) and \( \beta_4(x) \). Following the development of local polynomial fitting, we use local 5-th degree polynomial approximation with a pilot bandwidth \( g \) to estimate \( \sigma^2, \beta_2(x), \beta_3(x) \) and \( \beta_4(x) \). Note that the difference of our approach from the “plug-in” method is that we do not need to estimate the unknown terms involving the design density or its derivatives although they are present in the asymptotic optimal bandwidth \( h_O \), and the procedure still has a fast rate of convergence.
We now give the details of the proposed bandwidth selector.

Suppose the goal is to estimate \( m(\cdot) \) at grid points \( x_1, \ldots, x_N \) by local linear regression. **LOOP:** For a set of \( h \) values:

**LOOP:** For each grid point \( x_j \):

**STEP 1:** Choose a pilot bandwidth \( g = O(n^{-1/10}) \) and then fit a local polynomial of order 5 to compute the estimated regression coefficients \( \hat{\beta}_2(x_j), \hat{\beta}_3(x_j), \hat{\beta}_4(x_j), \) and \( \hat{\sigma}^2 \), where \( \hat{\sigma}^2 \) may be obtained by the difference-based estimators in Rice (1984) or Hall, Kay and Titterington (1990).

**STEP 2:** An estimated MSE can be obtained via \( \text{MSE}(\hat{m}_h(x_j)) = \text{bias}^2(\hat{m}_h(x_j)) + \text{var}(\hat{m}_h(x_j)) \), where

\[
\text{bias}(\hat{m}_h(x_j)|X_1, \ldots, X_n) = e_0^T (X_\ell^T W X_\ell)^{-1} X_\ell^T W \tau,
\]

\[
\text{var}(\hat{m}_h(x_j)|X_1, \ldots, X_n) = \hat{\sigma}^2 e_0^T (X_\ell^T W X_\ell)^{-1} (X_\ell^T W^2 X_\ell) (X_\ell^T W X_\ell)^{-1} e_0,
\]

with

\[
\tau = \begin{pmatrix}
\hat{\beta}_2(X_1 - x_j)^2 + \hat{\beta}_3(X_1 - x_j)^3 + \hat{\beta}_4(X_1 - x_j)^4 \\
\vdots \\
\hat{\beta}_2(X_n - x_j)^2 + \hat{\beta}_3(X_n - x_j)^3 + \hat{\beta}_4(X_n - x_j)^4
\end{pmatrix}.
\]

**END LOOP** of all grid points.

Compute the average MSE (MASE) of all grid points,

\[
\tilde{\text{MASE}}(h) = \frac{1}{N} \sum_{j=1}^{N} \text{MSE}(x_j; h).
\] (4.3)

**END LOOP** of all \( h \) values.

**STEP 3:** Choose the bandwidth that minimizes \( \tilde{\text{MASE}} \) as the empirically optimal bandwidth \( \hat{h}_{OPT} \) to fit local linear regression.

The MASE criterion in (4.3) is used for simplicity in finite-sample cases. One may want to adjust it with a weight function if boundary effects are concerned. The theory of the pilot bandwidth \( g \) of order \( O(n^{-1/10}) \) is based on Theorem 4.2 part (a) in Huang and Fan (1995). Practically, a pilot smoothing parameter may be selected by using the residual squares criterion (RSC) in Fan and Gijbels (1995) or other methods. The rate of convergence of the selected bandwidth \( \hat{h}_{OPT} \) is given in the following theorem.
Theorem 3 Assume that \( h = h(n) \to 0, nh + \log(h) \to \infty \) as \( n \to \infty \), and \( \theta_2 \) and \( \sigma^2 \) are independent. Under Conditions (A1)-A(4), if \( \theta_2 \) and \( \sigma^2 \) can be estimated at a \( n^{-1/2} \) rate and \( \int \beta_2(x)\beta_4(x)w(x)dx \) can be estimated a rate of \( O_P(n^{-1/10}) \), then

\[
\frac{\hat{h}_{OPT} - h_{OPT}}{h_{OPT}} = O_P(n^{-1/2}).
\]

The proof is given in Appendix.

From Theorem 4.2 (a) of Huang and Fan (1995), the requirements for the \( n^{-1/2} \)-convergence of estimating \( \int \beta_2^2(x)w(x)dx \) are that \( m(\cdot) \) has a 5 degree of smoothness, \( g = O(n^{-1/10}) \), and local polynomial fitting of order \( p = 7 \). Note that the smoothness condition here is slightly stronger, comparing to 4.25 of Hall et al. (1991). Although the condition may be reduced to 4.5 (Theorem 4.1 (a) in Huang and Fan (1995)) or possibly 4.25, our objective is not to find the minimal smoothness condition but to give a reasonable condition for a “natural” estimator of \( \theta_2 \). The required order \( p = 7 \) of local polynomial fitting is the same as in density estimation with a kernel of order 6. Empirically, fitting local polynomial of the 7-th order may introduce large variance. Therefore, we suggest fitting local polynomial regression of order 5 to estimate \( \beta_2, \beta_3, \) and \( \beta_4 \) in the procedure. In the density estimation setting, Hall et al. (1991) suggest using a kernel function of order 4 instead of 6 for the same reason.

5 Simulation results

The simulation study reported here is to illustrate the consistency of the proposed bandwidth selection algorithm. We generated 400 realizations of data sets of sizes \( n = 50, 200, 800 \) with random design points from each of three test examples. The grid points considered are \( x_i = -1.8 + i \times 0.036, i = 0, \ldots, 100 \). For each sample, the bandwidth selection algorithm is applied and the pilot bandwidth \( g \) is taken as \( g = 3.155h \) for simplicity of implementation. The constant 3.155 is the ratio of the asymptotic variance of the estimator for \( m(x) \) in STEP 1 (fitting a local polynomial of order 5) to that of \( m_\ell(x) \). To solve the minimization of \( \text{MASE} \), we choose the bandwidth with the smallest \( \text{MASE} \) among a set of 100 bandwidths (in log scale) ranging from 0 to 1, as the empirically optimal bandwidth \( \hat{h}_{OPT} \). A kernel density estimate of \( \log_{10}(h) \)'s is shown to summarize these 400 selected bandwidths with the
smoothing parameter $1.06 \times (\text{sample standard deviation}) (400)^{-1/5}$, as suggested in Silverman (1986).

**Example 1** Consider the regression model:

$$Y = X + 2\exp(-16X^2) + \epsilon, \, X \sim \text{Uniform}(-2, 2), \epsilon \sim N(0, 0.5^2).$$

Let $h_f = a_1(\beta_2, \sigma)n^{-1/5}$ be the first order optimal bandwidth, and $h_s$ be the optimal bandwidth to its deterministic second order $(n^{-3/5})$ terms, i.e.

$$h_s = h_o + n^{-3/5} \left( -7a_1^{1/2} \mathcal{V}_{n,1} \left( \int \beta_2(x)(\beta_2(x)f(x) + 1/f(x))w(x)dx \right) / (20\theta_2) \\
+ 1.5a_1^{-7/2} \mathcal{V}_{n,2} \left( \int \sigma^2(x)(2\nu_0 + f(x)^{-1})w(x)dx \right) / (20\mu^2\theta_2) \right).$$

The bandwidth $\bar{h}_{OPT}$ denotes the average of the optimal bandwidths $h_{OPT}$ over 400 simulations. Note that since the designs are random, the conditionally optimal bandwidth $h_{OPT}$ varies from simulation to simulation. The following table shows the numerical values of $h_f$, $h_s$, and $\bar{h}_{OPT}$ with the sample standard deviation $s$.

<table>
<thead>
<tr>
<th>sample size</th>
<th>$h_f$</th>
<th>$h_s$</th>
<th>$\bar{h}_{OPT} \pm s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.2625</td>
<td>0.2935</td>
<td>0.3771 ± 0.0725</td>
</tr>
<tr>
<td>200</td>
<td>0.1989</td>
<td>0.2124</td>
<td>0.2203 ± 0.0122</td>
</tr>
<tr>
<td>800</td>
<td>0.1143</td>
<td>0.1168</td>
<td>0.1558 ± 0.0032</td>
</tr>
</tbody>
</table>

It is clear that the second order approximation $h_s$ is closer to $h_{OPT}$ than $h_f$. When the sample size is large enough ($n = 200$ and $n = 800$ in this example), the difference between $h_f$ and $h_s$ becomes smaller. Figure 1 shows a set of simulated data with the true regression curve and the estimated density curves of $\log_{10}(\bar{h}_{OPT})$ for sample sizes 50, 200, 800. We also indicate the locations of $h_f$ (dotted line), $h_s$ (dash line), and $\bar{h}_{OPT}$ (solid line) in Figure 1. Note that the center of the distribution of $\bar{h}_{OPT}$ is a random variable $h_{OPT}$, not the average $\bar{h}_{OPT}$. The normal distribution provides an excellent approximation for the distribution of the selected bandwidth for sample sizes 200 and 800. The density curve of $n = 50$ is not quite satisfactory, since asymptotic theory may not take effect for data with small sample sizes.

**Example 2**

$$Y = \sin(2X) + 2\exp(-16X^2) + \epsilon, \, X \sim \text{Uniform}(-2, 2), \epsilon \sim N(0, 0.4^2).$$
<table>
<thead>
<tr>
<th>sample size</th>
<th>$h_f$</th>
<th>$h_s$</th>
<th>$\hat{h}_{OPT} \pm s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.2387</td>
<td>0.2614</td>
<td>0.3473 $\pm$ 0.0678</td>
</tr>
<tr>
<td>200</td>
<td>0.1809</td>
<td>0.1907</td>
<td>0.1987 $\pm$ 0.0103</td>
</tr>
<tr>
<td>800</td>
<td>0.1371</td>
<td>0.1414</td>
<td>0.1404 $\pm$ 0.0027</td>
</tr>
</tbody>
</table>

In this example, the regression curve is more complicated than that of Example 1, but a similar phenomenon is observed from Figure 2. Hence the same interpretation as Example 1 follows.

We consider a simple linear model in Example 3. In terms of bandwidth selection, a linear model is one of the most difficult cases because the theoretical MASE($h$) decays very slowly as $h$ increases. But, the regression function itself is not difficult to estimate, i.e. the precision of the smoothing parameter is less crucial in simple linear models.

**Example 3**

\[ Y = 0.4X + 1 + \varepsilon, \, X \sim Uniform(-2, 2), \varepsilon \sim N(0, 0.3^2). \]

The theoretically optimal bandwidth for a linear model is infinity, in the simulation, the largest being $h_{OPT} = 1$. We indicate in Figure 3 the location of $h = 1$ since $\log_{10}(1) = 0$. For the sample size $n = 200$, the dashed curve is the estimated density based on the pilot bandwidth $g = 3.155h$, and the solid curve, for comparison, is obtained by modifying the pilot bandwidth so that it will not exceed the range of the data. The selected bandwidth $\hat{h}_{OPT}$ of size $n = 200$ converges although the smoothing parameter is difficult to estimate for such a linear model. Note that in all three examples, a moderate sample size of $n = 200$ is adequate for the convergence of the proposed procedure and this is not the smallest sample size in order to have such a convergence.

The simulations by Fan and Gijbels (1995) are based on the RSC pilot bandwidth selector and a simpler expansion of the optimal bandwidth. Here our important contribution is to show that the $n^{-1/2}$ relative rate of convergence is possible for the bandwidth selection problem of local linear regression, although the sample size required to show the improvement may be large, as evidenced by Jones, Sheather, and Marron (1996) for kernel density estimation.
6 Appendix

Proof of Theorem 1: We begin by estimating the conditional bias. From (4.1) and (4.2), the bias term can be written as
\[
\text{bias}(\hat{m}_t(x)|X_1, \ldots, X_n) = e_0^T \left( \begin{array}{cc}
S_{n,0} & S_{n,1} \\
S_{n,1} & S_{n,2}
\end{array} \right)^{-1} \left( \begin{array}{c}
\beta_2 S_{n,2} + \beta_3 S_{n,3} + \beta_4 S_{n,4} \\
\beta_2 S_{n,3} + \beta_3 S_{n,4} + \beta_4 S_{n,5}
\end{array} \right) + r(x, X_1, \ldots, X_n),
\]
where \( r(\cdot) \) denotes the remainder terms. Using similar arguments as in the proof of Theorem 1 of Fan et al. (1996), we obtain
\[
\text{bias}(\hat{m}_t(x)|X_1, \ldots, X_n) = e_0^T \left( \begin{array}{cc}
s_0^* & s_1^* \\
s_1^* & s_2^*
\end{array} \right) + O_P(a_n)
\times \left( h^2 \left( \begin{array}{c}
\beta_2 s_2^* + h \beta_3 s_3^* + h^2 \beta_4 s_4^* \\
\beta_2 s_3^* + h \beta_3 s_4^*
\end{array} \right) + O_P(h^2 a_n) \right),
\]
where \( a_n \) denotes a quantity of order \( (h^3 + n^{-1/2} h^{-1/2}) \), \( s_0^* = f(x) + \frac{h^2}{2} \mu_2 f''(x) \), \( s_1^* = h \mu_2 f'(x) \), \( s_2^* = f(x) \mu_2 + h^2 \mu_4 f''(x) / 2 \), \( s_3^* = h \mu_4 f'(x) \), \( s_4^* = \mu_4 f(x) \), and \( s_5^* = h \mu_6 f'(x) \). It is easy to see that
\[
\left( \begin{array}{cc}
s_0^* & s_1^* \\
s_1^* & s_2^*
\end{array} \right) = f(x) \left( \begin{array}{cc}
1 & 0 \\
0 & \mu_2
\end{array} \right) + h f'(x) \left( \begin{array}{cc}
0 & \mu_2 \\
\mu_2 & 0
\end{array} \right) + h^2 f''(x) \left( \begin{array}{cc}
\mu_2 / 2 & 0 \\
0 & \mu_4 / 2
\end{array} \right).
\]
Note that for square invertible matrices \( A, B, \) and \( C \),
\[
(A + hB + h^2C)^{-1} = A^{-1} - h A^{-1} B A^{-1} - h^2 A^{-1} C A^{-1} + h^2 A^{-1} B A^{-1} B A^{-1} + O(h^3).
\]
Applying (6.2) to (6.1) with some matrix algebra, the bias expression in Theorem 1 is obtained.

Similarly,
\[
\text{var}\{\hat{m}_t(x)|X_1, \ldots, X_n\} = \frac{1}{nh} e_0^T \left( \begin{array}{cc}
s_0^* & s_1^* \\
s_1^* & s_2^*
\end{array} \right) + O_P(a_n)
\times \left( \begin{array}{cc}
r_0^* & r_1^* \\
r_1^* & r_2^*
\end{array} \right) + O_P(a_n)
\times \left( \begin{array}{cc}
s_0^* & s_1^* \\
s_1^* & s_2^*
\end{array} \right) + O_P(a_n)
\times e_0 \sigma^2(x)
\]
\]
(6.3)
with \( r^*_0 = f(x)\nu_0 + \frac{h^2}{2} \nu_2 f''(x) \), \( r^*_1 = h\nu_2 f'(x) \), and \( r^*_2 = f(x)\nu_2 + \frac{h^2}{2} \nu_4 f''(x) \). The asymptotic variance follows by some matrix computations.

**Proof of Theorem 2:** The proof of Theorem 2 is similar to the proof of Theorem 1 in Hall *et al.* (1991) so we omit it. The proof can be found in Huang (1995).

**Proof of Theorem 3:** Let

\[
\hat{\mathcal{M}}(h) = \int \tilde{\text{MSE}}(x)w(x)dx,
\]

be the integrated version of \( \mathcal{MSE}(h) \) in (4.3), and \( \hat{h}_O = \hat{h}_O(\hat{\beta}_2, \hat{\beta}_4, \hat{\sigma}, \nu_{n,1}, \nu_{n,2}, \nu_{n,1}) \) be the asymptotic minimizer \( h_O \) with unknown terms \( \beta_2(x), \beta_4(x) \) and \( \sigma^2 \) being substituted by the estimates. Note that \( \hat{h}_O \) is not a statistic since we do not estimate the design density, but \( \hat{h}_{OPT} \) is a statistic. To prove Theorem 3, we approach through the following two Lemmas which reduce the convergence-rate problem of \( \hat{h}_{OPT} \), the bandwidth that minimizes \( \hat{\mathcal{M}}(h) \), to that of \( \hat{h}_O \). The advantage of this reduction is that \( \hat{h}_O \) and \( h_O \) have explicit forms which are easier for further study.

**Lemma 1** Under the Conditions of Theorem 3, assume that the estimated \( \hat{\beta}_2(x), \hat{\beta}_4(x), \) and \( \hat{\beta}_4(x) \) are stochastically bounded. Then

\[
\frac{(\hat{h}_{OPT} - \hat{h}_O)}{\hat{h}_{OPT}} = o_p(n^{-1/2}),
\]

and

\[
\frac{(\hat{h}_{OPT} - h_{OPT})}{h_{OPT}} = \frac{(\hat{h}_O - h_O)}{h_{OPT}} + o_p(n^{-1/2}). \tag{6.4}
\]

This Lemma follows directly from Theorem 2.

**Lemma 2** Under the Conditions in Theorem 3, if \( \sqrt{n}(\hat{h}_O - h_O) \) is asymptotically normally distributed with mean 0 and finite variance, then \( \sqrt{n}(\hat{h}_{OPT} - h_{OPT}) \) has an asymptotically normal distribution with mean 0 and finite variance, and \( n\frac{\text{MISE}(\hat{h}_{OPT})}{\text{MISE}(h_{OPT})} \sim c\chi^2(1) \), where \( c \) is a constant.

**Proof:** Since

\[
\sqrt{n}(\hat{h}_{OPT} - h_{OPT}) = \frac{\hat{h}_{OPT}}{h_{OPT}} \sqrt{n}(\hat{h}_O - h_O) + o_p(1),
\]

13
The first conclusion follows. To show the second result, note that by a Taylor’s expansion,

\[ M(\hat{h}_{OPT}) = M(h_{OPT}) + M'(h_{OPT})(h_{OPT} - \hat{h}_{OPT}) + M''(\bar{h})(h_{OPT} - \hat{h}_{OPT})^2/2, \]

where \( \bar{h} = O(n^{-1/5}) \) lies between \( \hat{h}_{OPT} \) and \( h_{OPT} \). Since \( h_{OPT} \) minimizes \( M(h) \), \( M'(h_{OPT}) \) must be zero. Thus,

\[
\frac{n(M(\hat{h}_{OPT}) - M(h_{OPT}))/M(h_{OPT})}{n((h_{OPT} - \hat{h}_{OPT})^2M''(\bar{h})/(2M(h_{OPT})))} = \frac{(\sqrt{n}(h_{OPT} - \hat{h}_{OPT})/h_{OPT})^2(h_{OPT}^2M''(\bar{h})/(2M(h_{OPT})))}{(\sqrt{n}(h_{OPT} - \hat{h}_{OPT})/h_{OPT})^2(h_{OPT}^2M''(\bar{h})/(2M(h_{OPT})))}.
\]

The second result follows by noting that \( (h_{OPT}^2M''(\bar{h})/(2M(h_{OPT}))) \) converges to a constant.

Now the terms of \( \hat{h}_O \) which do not involve the design density or its derivatives are \( \theta_2 \), \( \sigma^2 \) and \( \int \beta_2(x)\beta_4(x)w(x)dx \). Hence the theorem follows if those quantities can be estimated at the required rates. \( \Box \)

References


Figure 1: Example 1.
Figure 2: Example 2.

A typical simulated data

Bandwidth Selection, n = 50

Bandwidth Selection, n = 200

Bandwidth Selection, n = 800
Figure 3: Example 3.

A typical simulated data

Bandwidth Selection, $n = 50$

Bandwidth Selection, $n = 200$

Bandwidth Selection, $n = 800$