Statistica Sinica **12**(2002), 429-447

ERROR-DEPENDENT SMOOTHING RULES IN LOCAL LINEAR REGRESSION

Ming-Yen Cheng[1,2] and Peter Hall[1]

[1]*Australian National University and* [2]*National Taiwan University*

Abstract: We suggest an adaptive, error-dependent smoothing method for reducing the variance of local-linear curve estimators. It involves weighting the bandwidth used at the ith datum in proportion to a power of the absolute value of the ith residual. We show that the optimal power is 2/3. Arguing in this way, we prove that asymptotic variance can be reduced by 24% in the case of Normal errors, and by 35% for double-exponential errors. These results might appear to violate Jianqing Fan's bounds on performance of local-linear methods, but note that our approach to smoothing produces nonlinear estimators. In the case of Normal errors, our estimator has slightly better mean squared error performance than that suggested by Fan's minimax bound, calculated by him over all estimators, not just linear ones. However, these improvements are available only for single functions, not uniformly over Fan's function class. Even greater improvements in performance are achievable for error distributions with heavier tails. For symmetric error distributions the method has no first-order effect on bias, and existing bias-reduction techniques may be used in conjunction with error-dependent smoothing. In the case of asymmetric error distributions an overall reduction in mean squared error is achievable, involving a trade-off between bias and variance contributions. However, in this setting, the technique is relatively complex and probably not practically feasible.

Key words and phrases: Bandwidth, kernel method, nonparametric regression, tail weight, variance reduction.

1. Introduction

There is a great variety of bias reduction methods for nonparametric curve estimation, ranging from high-order kernel techniques (e.g., Wand and Jones (1995), Chapters 2 and 5) to local bandwidth adjustments (e.g., Abramson (1982, 1984), Jones (1990)), methods based on varying location and scale (e.g., Samiuddin and el-Sayyad (1990), Jones, McKay and Hu (1994)), empirical transformations (e.g., Ruppert and Cline (1994)), weights (e.g., Jones, Linton and Nielsen (1995)), and skew computation (e.g., Choi and Hall (1998)). All have variants for both nonparametric density estimation and nonparametric regression, although often they are introduced first in the former setting. However, very few methods have been suggested for reducing the impact of variance. Those that do exist involve principally deterministic adjustments to bandwidth, altering the local

trade-off between variance and squared bias in the context of mean integrated squared error.

In the present paper, and in the context of nonparametric regression, we suggest a new and entirely different approach to variance reduction. It involves adjusting bandwidth in a stochastic, rather than deterministic, way, with the aim of providing improved performance by giving greater weight to data pairs that correspond to smaller absolute errors. In an extreme case, if an error were exactly zero then we would wish to use the corresponding data pair for interpolation, rather than simply for smoothing. Interpolation would correspond to using a bandwidth of zero, and so our "error dependent" method involves taking the bandwidth to be a function of the error — or more practically, of the residual. The function is non-degenerate, even in the asymptotic limit. For local-linear estimators we suggest that it be taken proportional to the two-thirds power of the absolute value of the error (or residual). Different powers are appropriate for higher-order methods, with the power increasing to 1 as order increases.

The idea of allowing the bandwidth for the ith datum to depend nontrivially on the ith error is reminiscent of Abramson's (1982) approach to bias reduction in density estimation. There, the bandwidth for smoothing X_i when estimating the density f is ideally taken proportional to a negative power of $f(X_i)$. While this approach has an analogue for nonparametric regression (Hall (1990)), it nevertheless reduces bias rather than variance, and is not closely related to the technique suggested here. In particular, error-dependent smoothing reduces variance by a constant factor, and has no first-order impact on bias (in the case of symmetric error distributions); most bias reduction methods reduce bias by an order of magnitude and inflate variance by a constant factor.

In large samples, our method reduces the variance contribution to mean integrated squared error by 24% in the case of Normal errors, by 35% for double-exponentially distributed errors, and by even greater amounts for very heavy-tailed error distributions. These figures might appear to violate the bounds asserted by Fan (1993) for performance of local-linear estimators. There is in fact no contradiction, however, principally for the following reason. Fan's minimax theory applies to classes of regression functions that have, in effect, two bounded derivatives. By way of contrast, our results require the functions to have two continuous derivatives. Our claim about improved performance does need continuity of the second derivative, and in particular does not hold uniformly across the function class addressed by Fan. Other dissimilarities too should be born in mind. For example, Fan's (1993) result about local-linear estimators applies only within the class of linear techniques, and (after error-dependent smoothing) our estimators are nonlinear. Furthermore, Fan's other minimax bounds, measuring performance against nonlinear techniques, do not specifically address the range of heavy-tailed error distributions that are considered in the present paper, since

his class $\mathcal{C}_2$ of models for the "max" part of "minimax" contains models with Normal errors.

Our methods are relatively simple when the error distribution is symmetric or nearly symmetric, and that is the context in which they have greatest practical significance. In the case of asymmetry, however, error-dependent smoothing can introduce an additional bias term to the estimator. This makes it relatively complex to minimise mean squared error. For the sake of completeness we briefly explore the general case from a theoretical viewpoint but, since the practical attraction of the asymmetric setting is not high, we do not describe its numerical properties.

From a practical viewpoint, empirical bandwidth choice methods that produce overly variable bandwidths can require relatively large samples in order to achieve theoretically optimal levels of performance. The rule proposed in this paper, where the bandwidth is proportional to a power of a residual, can be subject to excessive fluctuation when residuals are either close to 0 or large in absolute value. This suggests that a truncation argument be used to dampen variability and improve performance. The need to choose truncation points adds to the complexity of the method, however, as does the necessity of selecting a pilot bandwidth in order to calculate residuals. The result is that, in small to moderate sized samples, mean squared error reductions offered by a practical version of our method may be some distance from theoretically optimal levels. Nevertheless, our theory and simulation analysis show clearly the potential of the methods.

Provided errors are stochastically independent, conditional on design points, there is no difficulty in combining error-dependent smoothing with a bias-reduction method. The effect is to reduce bias by an order of magnitude, and reduce variance by a constant factor relative to the value it would assume if only bias reduction were employed. Likewise, error-dependent smoothing may be applied in conjunction with a spatially-local bandwidth choice procedure, such as that suggested by Fan and Gijbels (1995): one simply takes the scale factor h, in formulae such as $h_i = h\,H(\hat{\epsilon}_i)$ discussed in Section 2, to depend on spatial location x. The method is also applicable to the case of heteroscedasticity, no matter whether the variance function is modelled parametrically or estimated nonparametrically. See Section 2.7 for discussion.

2. Methodology and Main Results

2.1. Model and estimator

Assume that independent and identically distributed data pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are generated by the model

$$Y_i = g(X_i) + \epsilon_i\,, \tag{2.1}$$

where X_i is independent of ϵ_i, and $E(\epsilon_i) = 0$. In the "ideal" case, where the errors ϵ_i are known, we define a bandwidth h_i by $h_i = h\,H(\epsilon_i)$. Here, $h = h(n)$ denotes a sequence of positive constants, and H is a fixed positive function. More realistically, we might approximate ϵ_i by a residual $\hat{\epsilon}_i$, and put $h_i = h\,H(\hat{\epsilon}_i)$. In either case, and for fixed x, let $(\hat{a}, \hat{b})$ denote the pair that minimises

$$\sum_{i=1}^{n} \{Y_i - a - b\,(X_i - x)\}^2\, h_i^{-1}\, K\{(X_i - x)/h_i\}\,,$$

where K is a kernel function, and put $\hat{g}(x) = \hat{a}$.

We expect $H(x)$ to be an increasing function of $|x|$, in which case smaller absolute errors (or residuals) produce less smoothing. In the "realistic" case, where residuals are used rather than the errors themselves, particularly large absolute values of residuals may not necessarily reflect similar values of errors. Rather, they might result from inaccuracies in the pilot estimator of g that is used to compute residuals. This means that we should usually threshold the smoothing parameter, to guard against outlying values. Thresholding is also useful if we are to ward off problems with sparse design, which do not make themselves felt through the asymptotic distribution of $\hat{g}$. These difficulties do not arise in the "ideal" case, which therefore offers more insight into the operation of level-dependent thresholding. We address the "ideal" and "realistic" cases in Sections 2.4 and 2.5, respectively.

2.2. Overview of theory

First we deal with the "ideal" setting. In the case of classical local-linear smoothing, where $H \equiv 1$, it is well-known that the local-linear estimator has bias of size h^2 and error about the mean of size $(nh)^{-1/2}$. Specifically, under mild regularity conditions (see Fan (1993)),

$$\hat{g} = g + \tfrac{1}{2} h^2 g'' \kappa_2 + (nh)^{-1/2} f^{-1/2} \kappa^{1/2} \sigma N_n + o_p(h^2)\,, \tag{2.2}$$

where $\kappa_2 = \int u^2\, K(u)\, du$, $\kappa = \int K^2$, $\sigma^2 = \mathrm{var}(\epsilon)$, and N_n denotes a random variable whose distribution is asymptotically Normal $\mathrm{N}(0, 1)$. The second term on the right-hand side of (2.2) represents systematic error, and the third is stochastic error.

More generally, suppose H is a non-degenerate function. If $E\{H(\epsilon)^2\} < \infty$ then we may standardise H so that $E\{H(\epsilon)^2\} = 1$. We claim that in these circumstances, provided the error distribution is symmetric and H is an even function, the expansion at (2.2) continues to hold, except that the third term is multiplied by the factor ρ, where $\rho^2 = E\{\epsilon^2\, H(\epsilon)^{-1}\}/\sigma^2$:

$$\hat{g} = g + \tfrac{1}{2} h^2 g'' \kappa_2 + (nh)^{-1/2} \rho f^{-1/2} \kappa^{1/2} \sigma N_n + o_p(h^2)\,. \tag{2.3}$$

As in (2.2), N_n denotes a random variable that is asymptotically Normal $\mathrm{N}(0,1)$, although it will assume different numerical values in appearances at (2.2) and (2.3).

Minimising ρ subject to $E\{H(\epsilon)^2\} = 1$ is an elementary variational problem. The minimum is achieved when $H(\epsilon)$ equals a constant multiple of $|\epsilon|^{2/3}$, in which case $\rho^2 = \rho_0^2$ where

$$\rho_0^2 \equiv \{E(|\epsilon|^{4/3})\}^{3/2}/\sigma^2 < 1\,. \tag{2.4}$$

Of course, taking $H(\epsilon)$ proportional to $|\epsilon|^{2/3}$ is all that is needed for optimal reduction of asymptotic mean squared error. The particular constant is absorbed into the non-random multiplier, h, and so is immaterial.

For Normal errors, and for our estimator rather than a conventional local-linear estimator, our results are suggestive of a version of Fan's (1993) Theorem 4 in which his constant 0.896^2 is replaced by 1.00, this being the value (to two decimal places) of

$$(1.243\,\rho_0^{8/5})^{-1} = (1.243 \times [2\,\{\Gamma(7/6)/\pi^{1/2}\}^{3/2}]^{4/5})^{-1} \approx 1.0024^2\,. \tag{2.5}$$

The formula here represents the efficiency, defined as the ratio of mean squared errors, of our estimator relative to the "optimal" nonlinear regression function estimator, when estimating a fixed, twice continuously differentiable target g. The "optimal" estimator here is the one that gives minimum mean squared error in a minimax sense, uniformly over functions that have two bounded derivatives; it does not make use of continuity of the second derivative. Details of the origin of the formula are given by Donoho and Liu (1991), and in fact the figure 1.243 is taken from Donoho and Liu's Table 1. It was apparently derived by combining, in a conservative way, numerical results obtained by Donoho, Liu and MacGibbon (1990). The exact value of ρ_0^2 in the Normal case may be shown to equal $2\,\{\Gamma(7/6)/\pi^{1/2}\}^{3/2} \approx 0.757$.

Note that our results rely on *continuity* of the second derivative of g, whereas Fan's results apply uniformly over a class of g's that have only a uniformly *bounded* derivative. The requirement for continuity is the key to reconciling Fan's results with our own. The fact that the "ideal" estimator is not a true estimator in the usual sense is not an issue in comparing Fan's results with our own. Indeed, we show in Section 2.5 that, for any particular functions f and g with two continuous derivatives, the level of asymptotic performance evinced by the ideal estimator is achievable by a realistic version. In this case a pilot estimator $\tilde{g}$ is constructed, and used to calculate $\hat{\epsilon}_i = Y_i - \tilde{g}(X_i)$. Then, possibly after centering or thresholding these residuals, we compute the empirical bandwidth $\hat{h}_i = h\,|\hat{\epsilon}_i|^{2/3}$. We use $\hat{h}_i$ in place of h_i to construct $\hat{g}$. Modulo regularity conditions, and truncation to alleviate difficulties caused by too large or too small values of $|\hat{\epsilon}_i|$, formula (2.3) continues to hold.

Since Fan's minimax function class $\mathcal{C}_2$ contains models where the error distribution is Normal, results such as those discussed above do not relate to improvements in performance that error-dependent smoothing can achieve in the case of heavy-tailed error distributions. Indeed, the value of ρ_0^2 tends to be smaller for distributions with heavier tails. Values in the cases of double Exponential, Normal and Uniform errors are respectively 0.650, 0.757 and 0.842. For errors with Student's t distribution on 5, 10 or 20 degrees of freedom, the values are 0.676, 0.726 and 0.743, respectively. Figure 2.1 plots ρ^2 as a function of the number of degrees of freedom (interpreted in the continuum) for Student's t errors. For non-Normal errors one can also construct estimators that are more efficient than $\hat{g}$ by replacing local least-squares by a robust method such as local M-estimation.

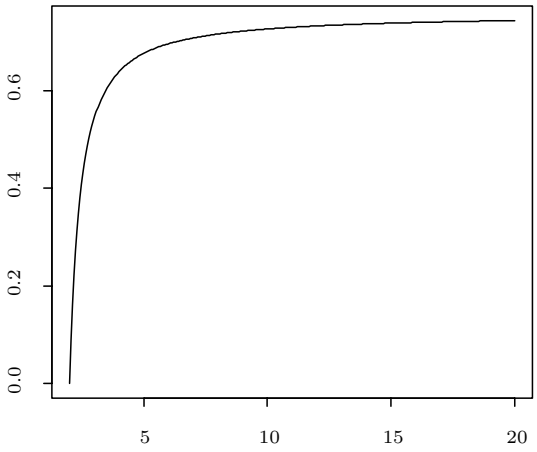

Figure 2.1. Values of ρ^2 versus number of degrees of freedom for Student's t error distributions.

In the case of asymmetric errors, or when H is not an even function, a formula similar to (2.3) is valid, except that an extra bias term of size h^2 is introduced. This quantity is proportional to $E\{\epsilon\, H(\epsilon)^2\}$, and so is typically small if the error distribution is close to being symmetric. More generally, it is possible to either choose both h and H (in the formula $h_i = h\, H(\epsilon_i)$) so that asymptotic mean squared error is minimised; or to choose H to minimise the asymptotic variance term subject to both $E\{H(\epsilon)^2\} = 1$ and $E\{\epsilon\, H(\epsilon)^2\} = 0$. The first rule produces a particularly complex function H, and is quite unattractive for practical implementation. The second rule gives

$$H(u) = \begin{cases} c_1\, u^{2/3} & \text{if } u > 0, \\ c_2\, |u|^{2/3} & \text{if } u < 0, \end{cases} \tag{2.6}$$

where

$$c_1 = \Big[E\{\epsilon^{4/3}\, I(\epsilon > 0)\} + \frac{E\{|\epsilon|^{4/3}\, I(\epsilon < 0)\}\, E\{\epsilon^{7/3}\, I(\epsilon > 0)\}}{E\{|\epsilon|^{7/3}\, I(\epsilon < 0)\}} \Big]^{-1/2},$$

$$c_2/c_1 = \Big[E\{\epsilon^{7/3}\, I(\epsilon > 0)\} / E\{|\epsilon|^{7/3}\, I(\epsilon < 0)\} \Big]^{1/2}.$$

Thus, using error-dependent smoothing in the case of asymmetric errors requires relatively detailed inference about the error distribution, and is not as attractive as in the case of symmetry.

Note, however, that there is always a strict reduction in the variance contribution to asymptotic mean squared error. For example, if H is given by (2.6), with the constants c_1, c_2 defined above, then (without requiring the assumption of symmetry) (2.3) holds with $\rho < 1$.

2.3. Generalisations and extensions

The main idea behind these results — that the variance of a nonparametric regression estimator may be reduced by using an error-dependent smoothing parameter — is applicable to a wide range of settings. For example, it is valid for linear, second-order methods such as the Nadaraya-Watson estimator (see e.g., Härdle (1990), p.25; Wand and Jones (1995), p.119), the Gasser-Müller estimator (Wand and Jones (1995), p.131) and the Priestley-Chao estimator (Wand and Jones (1995), p.130). In all these cases the asymptotically optimal form of H is the same as that described above in the local-linear context. Moreover, if the error distribution is symmetric then bias is unaffected, to first order, by error-dependent smoothing.

Error-dependent smoothing is also applicable to general local-polynomial estimators of regression means and their derivatives (see e.g., Ruppert and Wand (1994)). There, if the method is of order r (meaning that, in the case of a non-random bandwidth h, bias is of size h^r), the natural analogue of the condition $E\{H(\epsilon)^2\} = 1$ is $E\{H(\epsilon)^r\} = 1$. This ensures that, if the error-dependent bandwidth is $h_i = h\, H(\epsilon_i)$, if H is an even function, and if the error distribution is symmetric, the bias term is the same (to first order) as it would be if the bandwidth were simply h. If we are estimating the sth derivative of g, for $s \geq 0$, then the variance contribution to asymptotic mean squared error is also the same to first order, except that it is reduced by the multiplicative factor $\rho = E\{\epsilon^2\, H(\epsilon)^{-(2s+1)}\}/\sigma^2$.

The minimum of this quantity, subject to the constraint $E\{H(\epsilon)^r\} = 1$, is achieved when $H(\epsilon)$ is proportional to $|\epsilon|^{2/(r+2s+1)}$. Thus, the general form of ρ_0^2 (given at (2.4) in the case $(r, s) = (2, 0)$) is

$$\rho_0^2 = \{E(|\epsilon|^{2r/(r+2s+1)})\}^{(r+2s+1)/r}/\sigma^2 < 1\,.$$

In the case of symmetric errors, this represents the greatest amount by which variance may be reduced using error-dependent smoothing, when estimating an sth derivative by an rth order method.

2.4. Theory in "ideal" case

Assume that $h = h(n) \to 0$ and $nh \to \infty$ as $n \to \infty$, and that H is a continuous, positive function satisfying

$$E\{H(\epsilon)^3\} < \infty\,, \quad E\{H(\epsilon)^{-1}\} < \infty\,, \quad E[\epsilon^2\{H(\epsilon) + H(\epsilon)^{-1}\}] < \infty\,. \tag{2.7}$$

Let K be a bounded, compactly supported, symmetric probability density, and let f denote the marginal density of X_i, assumed to exist in a neighbourhood of x. Assume that f and g have two continuous derivatives in a neighbourhood of x, and that $f(x) > 0$.

Theorem 2.1. *Under the above conditions,*

$$\begin{aligned}\hat g(x) = g(x) + \tfrac{1}{2}\,h^2\,\kappa_2\,[g''(x)\,E\{H(\epsilon)^2\} + f''(x)\,f(x)^{-1}\,E\{\epsilon\,H(\epsilon)^2\}] \\ +[(nh)^{-1}\,\kappa\,f(x)^{-1}\,E\{\epsilon^2\,H(\epsilon)^{-1}\}]^{1/2}\,N_n(x) + o_p(h^2)\,,\end{aligned} \tag{2.8}$$

where $N_n(x)$ denotes a random variable whose distribution is asymptotically Normal $N(0,1)$.

Corollary. *Assume the conditions of Theorem* 2.1. *If $E\{\epsilon\,H(\epsilon)^2\} = 0$ — in particular, if the distribution of ϵ is symmetric and H is an even function — and if $E\{H(\epsilon)^2\} = 1$, then* (2.3) *holds at x, with $\rho^2 = E\{\epsilon^2\,H(\epsilon)^{-1}\}/\sigma^2$.*

Conditions (2.7) hold if, for example, (a) $C_1|\epsilon|^{1-\delta} \le H(\epsilon) \le C_2(1 + |\epsilon|)$ for constants $C_1, C_2, \delta > 0$, (b) $E(|\epsilon|^3) < \infty$, and (c) the probability density of ϵ exists in a neighbourhood of the origin and is bounded away from 0 there. If $H(\epsilon)$ has the optimal form $C\,|\epsilon|^{2/3}$ then (2.7) is valid if (c) holds and $E(|\epsilon|^{8/3}) < \infty$. The theorem may be extended to a variety of other settings, for example to the case where x is a boundary point.

2.5. Theory in "realistic" case

Suppose f and g have two continuous derivatives in a neighbourhood of x, and that $f(x) > 0$. Assume too that K is a bounded, compactly supported, symmetric probability density, and K' exists and satisfies a Hölder condition; that H' exists and satisfies a Hölder condition, and $C_1 \le H \le C_2$ for constants $0 < C_1 \le C_2 < \infty$; that $h = h(n) \to 0$ and $nh \to \infty$ as $n \to \infty$; that $E(\epsilon^2) < \infty$. Call these assumptions (A_1). (We say that a function ψ satisfies a Hölder condition, or is Hölder continuous with exponent η, if there exists a constant $C > 0$ such that $|\psi(x) - \psi(y)| \le C\,|x - y|^\eta$ for all x and y.)

In view of the continuity of f'' and g'', the quantity

$$\xi(\delta) \equiv \sup_{u\,:\,|x-u|\le\delta} \{|f''(x) - f''(u)| + |g''(x) - g''(u)|\} \tag{2.9}$$

converges to 0 as $\delta \to 0$. Let $h_0 = h_0(n)$ denote a bandwidth, chosen to converge to 0 sufficiently slowly for $n^{1/5}\, h_0 \to \infty$, and sufficiently quickly for $n^{1/5}\, h_0 = O(n^\eta)$ for all $\eta > 0$ and $n^{1/5}\, h_0\, \xi(h_0)^{1/2} \to 0$. Call this assumption (A_2). For example, if both f'' and g'' are Hölder continuous then $\xi(\delta) = O(\delta^\eta)$ for some $\eta > 0$, and so $h_0(n) \asymp n^{-1/5} \log n$ is an adequate choice.

Let $\tilde g$ denote a standard local-linear estimator of g, computed using bandwidth h_0. Put $\hat\epsilon_i = Y_i - \tilde g(X_i)$ and $\hat h_i = h\, H(\hat\epsilon_i)$, and redefine $\hat g(x)$ as the value $\hat a$ in the pair $(a, b) = (\hat a, \hat b)$ that minimises

$$\sum_{i=1}^n \{Y_i - a - b\,(X_i - x)\}^2\, \hat h_i^{-1}\, K\{(X_i - x)/\hat h_i\}\,.$$

Theorem 2.2. *Under assumptions* (A_1) *and* (A_2), (2.8) *holds for the new estimator* $\hat g$, *with* N_n *again denoting a random variable that is asymptotically Normal* $N(0, 1)$.

Neither Theorem 2.1 nor Theorem 2.2 is available uniformly in a function class such as Fan's $\mathcal{C}_2$. One reason, complementary to that given in Section 2.2, is that the rate at which the "$o_p(h^2)$" terms converge to 0 depends explicitly on the modulus of continuity of both f'' and g'', and can be arbitrarily slow. In the case of Theorem 2.2 this in turn influences choice of the pilot-estimator bandwidth h_0. Taking that quantity to be equal to a fixed constant multiple of $n^{-1/5}$, rather than of larger order than $n^{-1/5}$ (as required by assumption (A_2)), results in inflation of variance relative to that achieved by the "ideal" estimator. For sufficiently heavy-tailed error distributions, such an inflation still produces a reduction in variance, relative to that for a classical local-linear estimator. This is readily apparent in both theoretical analysis and numerical simulations. For Normal data, however, slight oversmoothing of the pilot estimator and relatively large sample sizes are necessary in order to achieve obvious improvements.

The conditions imposed on H in Theorem 2.2 preclude the optimal form $H(u) = \gamma\,|u|^{2/3}$. However, the level of performance in that case may be achieved, up to a constant factor that converges to 1 as $n \to \infty$, by considering successive approximations to the optimal H by functions satisfying the conditions of the theorem. Likewise, conditions on the kernel K prevent it from being the Epanechnikov function that attains asymptotically minimal mean squared error, but we may circumvent this problem by considering a sequence of approximations.

For example, the following is an empirical, thresholded version of the asymptotically optimal "ideal" procedure suggested in Section 2.2. Let $\tilde g$ be defined

as before, and assume the conditions imposed in Theorem 2.2 on K and on the bandwidths h and h_0. Additionally, suppose K is a compactly supported probability density with two bounded derivatives, that f and g have three bounded derivatives in a neighbourhood of x, that $f(x) > 0$, that $E(\epsilon^4) < \infty$, and that H is defined by

$$H(u) = \begin{cases} \gamma\,|u|^{2/3} & \text{if } n^{-\alpha} \le |u| \le n^{\beta}, \\ \gamma\,n^{-2\alpha/3} & \text{if } |u| < n^{-\alpha}, \\ \gamma\,n^{2\beta/3} & \text{if } |u| > n^{\beta}\,, \end{cases} \tag{2.10}$$

where α, β, γ denote positive constants. Then it is possible to choose $\alpha, \beta > 0$ such that, for all $\gamma > 0$, (2.7) holds for the new version of $\hat{g}$ (constructed using the bandwidths $\hat{h}_i = H(\hat{\epsilon}_i)$). The proof is particularly complex, and so will not be given here. The function H at (2.10) achieves the asymptotic performance represented by (2.8) with $H(u) \equiv \gamma\,|u|^{2/3}$.

2.6. Bandwidth choice

Note that in view of property (2.3) the optimal bandwidth, $h^{\rm ed}_{\rm opt}$ say, for an error-dependent smoothing rule is $h^{\rm ed}_{\rm opt} = h^{\rm llin}_{\rm opt}\,\rho^{2/5}$, where $h^{\rm llin}_{\rm opt}$ is the optimal bandwidth for the conventional local-linear estimator, and $\rho^2 = E\{\epsilon^2\,H(\epsilon)^{-1}\}/\sigma^2$. (It is assumed here that H has been standardised so that $E\{H(\epsilon)^2\} = 1$.) Therefore, to construct an empirical approach to bandwidth choice for an error-dependent smoothing rule we can use our "favourite" technique (such as a plug-in method or cross-validation) to compute an empirical approximation $\hat{h}^{\rm llin}$ to $h^{\rm llin}_{\rm opt}$, and take $\hat{h}^{\rm ed} = \hat{h}^{\rm llin}\,\hat{\rho}^{2/5}$ to be our empirical approximation to $h^{\rm ed}_{\rm opt}$, where

$$\hat{\rho}^2 = \biggl\{n^{-1}\sum_{i=1}^n \hat{\epsilon}_i^2\,H(\hat{\epsilon}_i)^{-1}\biggr\}\biggl\{n^{-1}\sum_{i=1}^n H(\hat{\epsilon}_i)^2\biggr\}^{1/2}\biggl(n^{-1}\sum_{i=1}^n \hat{\epsilon}_i^2\biggr)^{-1}.$$

(The second factor here gives an empirical version of the standardisation $E\{H(\epsilon)^2\} = 1$.) It may be proved that if the error distribution is symmetric, and if (in the construction of $\hat{g}(x)$) we take $\hat{h}_i = \hat{h}^{\rm llin}\,\hat{\rho}^{2/5}\,H(\hat{\epsilon}_i)$, and replace h on the right-hand side of (2.8) by $h^{\rm ed}_{\rm opt}$, then (2.3) continues to hold.

There is also a more complex cross-validation approach which adjusts implicitly for error-dependent smoothing, and does not require the multiplicative adjustment by $\hat{\rho}^{2/5}$. It is very computer intensive, however.

2.7. Heteroscedastic errors

For simplicity and brevity we have presented results only for models with independent and identically distributed errors. However, our methods apply virtually without change in heteroscedastic cases, where the error variance changes

with x. The simplicity with which this setting can be treated derives from the fact that, provided error variance is a smooth function of x, the model is "locally homoscedastic".

Indeed, suppose $Y_i = g(X_i) + \epsilon_i$ where $\epsilon_i = \zeta_i \, \sigma(X_i)$, the variables $X_1, \zeta_1, \ldots, X_n, \zeta_n$ are independent, the X_i's are identically distributed, the ζ_i's are identically distributed with zero mean and unit variance, and the function σ is continuous. Denote these conditions by assumption (A_3) (it replaces the overarching assumption made in Section 2.1 that the pairs (X_i, Y_i) are independent and identically distributed, which implies that the ϵ_i's are independent and identically distributed). If we continue to define $\hat{\epsilon}_i = Y_i - \tilde{g}(X_i)$, if we continue to take $\hat{h}_i = h \, H(\hat{\epsilon}_i)$, if on the right-hand side of (2.8) we replace ϵ in all the expectations by $\zeta \, \sigma(x)$, and if we assume (A_1)–(A_3), then Theorem 2.2 continues to hold. The method of proof is virtually identical.

The only essential difference between homoscedastic and heteroscedastic cases lies in the way bandwidth is computed. We discuss this issue in the setting of local bandwidth choice, which seems more appropriate when error variance changes with location. Assume the distribution of ζ is symmetric, and for simplicity consider the case where $H(u) = |u|^{2/3}$. (Truncated versions of H, such as those considered at the end of Section 2.5, are approximations to this function. There is no need to include a constant multiplier, since it may be incorporated into the bandwidth.) Put $\mu = E(|\zeta|^{4/3})$. Then, noting the result reported in the previous paragraph, we show that

$$\begin{aligned}\hat{g}(x) = g(x) &+ \tfrac{1}{2} \, h^2 \, \kappa_2 \, g''(x) \, \mu \, \sigma(x)^{4/3} \\ &+ [(nh)^{-1} \, \kappa \, f(x)^{-1} \, \mu \, \sigma(x)^{4/3}]^{1/2} \, N_n(x) + o_p(h^2) \, .\end{aligned}$$

It may be proved from this formula that the bandwith $h^{\rm ed}_{\rm opt}(x)$ that minimises asymptotic mean squared error at x may be expressed as $h^{\rm ed}_{\rm opt}(x) = h^{\rm llin}_{\rm opt}(x) \{\mu \sigma(x)^{4/3}\}^{-1/5}$, where $h^{\rm llin}_{\rm opt}(x) = \{\kappa \sigma(x)^2 / n \kappa_2 f(x) g''(x)^2\}^{1/5}$ is the bandwidth that minimises mean squared error of the standard local linear estimator.

Therefore, given an empirical version $\hat{h}^{\rm llin}_{\rm opt}(x)$ of $h^{\rm llin}_{\rm opt}(x)$ (see e.g., Fan and Gijbels (1995)), and estimators $\widehat{\mu}$ and $\hat{\sigma}(x)$ of μ and $\sigma(x)$ respectively, we may calculate an empirical version $\hat{h}^{\rm ed}_{\rm opt}(x) = \hat{h}^{\rm llin}_{\rm opt}(x) \, \{\widehat{\mu} \, \hat{\sigma}(x)^{4/3}\}^{-1/5}$ of $h^{\rm ed}_{\rm opt}(x)$. There are several ways of computing $\widehat{\mu}$ and $\hat{\sigma}(x)$; we give only two here. In the first, $\sigma(\cdot)$ is modelled parametrically by a smooth function, for example a linear or a quadratic function. Parameters of the model, and hence $\sigma(\cdot)$, can be consistently estimated by treating centered residuals as though they were true values of the ϵ_i's. In this way, an estimator $\hat{\sigma}(\cdot)$ of $\sigma(\cdot)$ can be calculated. Residual values of the ζ_i's can now be computed as $\hat{\zeta}_i = \hat{\epsilon}_i / \hat{\sigma}(X_i)$, and μ estimated by $\widehat{\mu} = n^{-1} \sum_i |\hat{\zeta}_i|^{4/3}$. In the second approach, $\sigma(\cdot)$ is estimated nonparametrically

by considering the nonparametric regression problem in which squares of centered residuals are regressed on their expected values. Once $\hat{\sigma}(\cdot)$ has been calculated in this way, $\widehat{\mu}$ can be computed as before.

3. Numerical Properties

In this section we report a simulation study conducted to examine numerical properties of error-dependent smoothing rules. In this work we took $g(x) = 4\sin(2\pi x)$ and used equally-distributed design points on $(0, 1)$. The error distribution was either Normal $\mathrm{N}(0, 2.25)$ or Student's t with 5 degrees of freedom. Sample size n was 50, 100, 200 or 500. The biweight kernel $K(u) = (1 - u^2)^2\, I_{(-1<u<1)}$ was employed.

When plotting integrated squared biases, variances and mean squared errors against bandwidth, h ranged over 51 logarithmically equispaced values. We do not explore empirical bandwidth choice, since the additional variation that it introduces may confound differences between conventional local-linear techniques and our method. In theory the second bandwidth, h_0, used for the for the pilot estimator has only a second-order effect on the results, although it should be taken larger than the theoretically optimal bandwidth. In all the work reported here we chose h_0 to be 25% larger than the standard optimal value.

Three estimators were considered: (i) $\hat{g}_L$, the standard local-linear estimator; (ii) $\hat{g}_I$, the "ideal" error-dependent local-linear estimator with $h_i = h|\epsilon_i|^{2/3}$; and (iii) $\hat{g}_R$, the "realistic" error-dependent local-linear estimator using the function H at (2.10) with $(n^{-\alpha}, n^{\beta}) = (0.2^{2/3}, 8^{2/3})$ and residuals obtained from a pilot estimation. For every setting, 1000 random samples were generated. Each estimator was evaluated over a equispaced grid of 400 points with the interpolation method of Hall and Turlach (1997) used to guard against sparse design problems caused by too-small choices of bandwidth. The mean integrated squared biases and variances were approximated by averaging over the 1000 realizations.

Figure 3.1 summarises results when $n = 100$ and the error distribution was Student's t with 5 degrees of freedom. First, we note that error-dependent smoothing rules lead to significant decrease in the mean integrated variance (see panel (c)) while maintaining almost the same level of mean integrated squared bias (panel (b)). This is as predicted by our asymptotic theory. Furthermore, the minimum mean integrated squared error (MISE) values for $\hat{g}_L$, $\hat{g}_I$ and $\hat{g}_R$ are 0.137, 0.096 and 0.119. Equivalently, the "ideal" and "realistic" error-dependent smoothing rules reduced MISE by 30% and 13% respectively. Greater reductions occurred for larger values of n until they asymptoted to the large-sample limit predicted in Section 2.

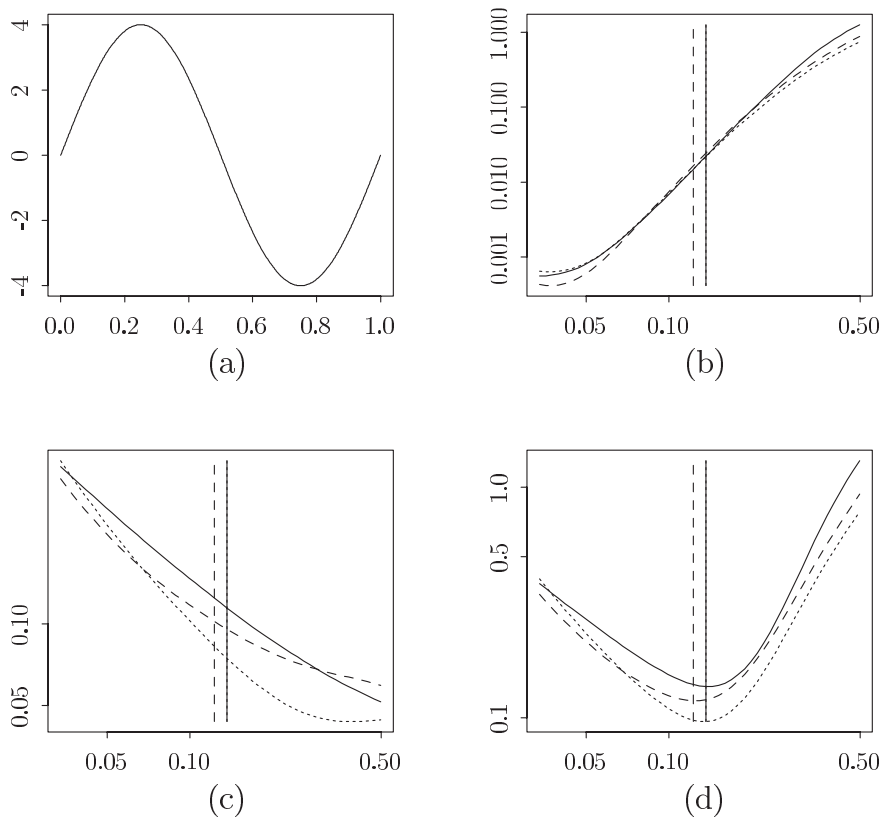

Figure 3.1. Simulation results for the sine regression function, depicted in panel (a), and for sample size $n = 100$ and t_5 errors. In panels (b), (c) and (d), respectively, the integrated squared biases, variances and mean squared errors of $\hat{g}_L$ (solid lines), $\hat{g}_I$ (dotted lines) and $\hat{g}_R$ (dashed lines) are plotted against bandwidth on a log-log scale. The vertical lines, with consistent line types, locate the optimal bandwidths that produced minimum mean integrated squared errors.

Figure 3.2 is the analogue of Figure 3.1 except that the error distribution is now Normal $N(0, 2.25)$. The "ideal" error-dependent smoothing rule again produces significant MISE reduction. However, MISE reductions in the "realistic" case only become significant for $n = 500$. In the case of Normal errors, error-dependent smoothing rules tend to slightly inflate the bias while reducing the variance. The increase in bias is of course a second-order effect, since it is not evident in the first-order theoretical analysis in Section 2. More generally, simulations with Student's t errors with a range of degrees of freedom show that the extent of MISE improvement offered by $\hat{g}_R$ declines, and the extent of bias inflation increases, as the error tail-weight decreases.

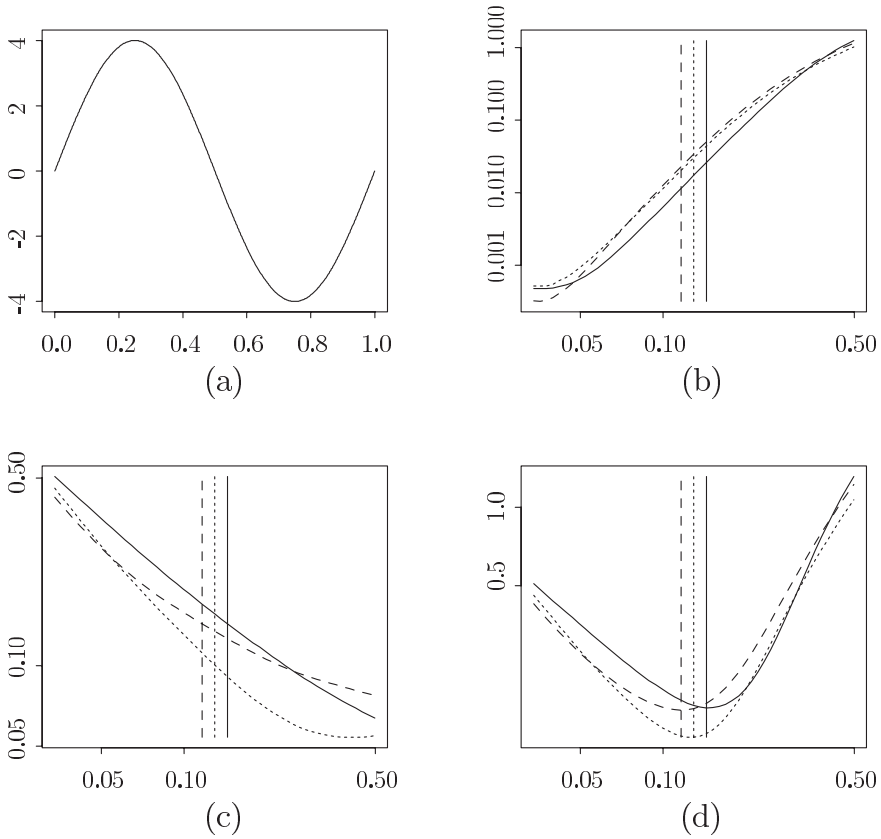

Figure 3.2. Simulation results for the sine regression function, $n = 100$ and Normal $(0, 2.25)$ errors. Functions and line types are the same as in Figure 3.1.

In summary, tail weight of the error distribution, and second-order contributions from bias, can have significant affect on the performance of error-dependent smoothing rules. In relation to the bias issue we found that, for a given error distribution, performance of $\hat{g}_R$ relative to $\hat{g}_L$ could be made better or worse by using target functions that produced lesser or greater amounts of bias, respectively. For example, by taking the sinusoid g to have only half a wavelength over the interval $(0, 1)$ we could enhance performance of $\hat{g}_R$ relative to $\hat{g}_L$; and by giving it more than one wavelength we could reduce relative performance.

A reviewer expressed concern that our use of grids for computation implied the method might be excessively computationally expensive. We use grids only to numerically determine mean squared integrated errors. In particular, no grid search methods are required. All our techniques involve only explicit calculation; nothing is defined implicitly, and no equations have to be solved in order to con-

struct our estimators. In practice one could use binning procedures to accelerate calculation. See for example page 238 of Scott (1992).

Acknowledgements

We are grateful to Iain Johnstone for very helpful discussion, and to three reviewers for helpful comments on the first version of the paper.

Appendix

A.1 Proof of Theorem 2.1.

We may write $\hat{g} = (S_2T_0 - S_1T_1)/(S_2S_0 - S_1^2)$, where, defining $K_i(x) = K\{(X_i - x)/h_i\}$ and, for a vector $a = a(x) = (a_1, \ldots, a_n)$, setting $U(a) = n^{-1}\sum_i h_i^{-1} a_i K_i$, we put $S_j = U(a)$ for $a_i(x) = (X_i - x)^j$, and $T_j = U(a)$ for $a_i(x) = Y_i (X_i - x)^j$. Let $T_{j1} = U(a)$ with $a_i(x) = g(X_i)(X_i - x)^j$, $T_{j2} = U(a)$ with $a_i(x) = \epsilon_i (X_i - x)^j$, and $R_j = U(a)$ with $a_i = h_i^j$. Then, $T_j = T_{j1} + T_{j2}$. (For the sake of simplicity we shall often suppress the argument x.)

It may be proved by Taylor expansion that

$$S_2T_{01} - S_1T_{11} = g\,(S_2S_0 - S_1^2) + \tfrac{1}{2}\,g''\,[\kappa_2\,h^2\,f\,E\{H(\epsilon)^2\}]^2 + o_p(h^4) \qquad \text{(A.1)}$$

and $S_2S_0 - S_1^2 = \kappa_2\,h^2\,f^2\,E\{H(\epsilon)^2\} + o_p(h^2)$. Therefore,

$$\hat{g}_1 \equiv (S_2T_{01} - S_1T_{11})/(S_2S_0 - S_1^2) = g + \tfrac{1}{2}\,g''\,\kappa_2\,h^2\,E\{H(\epsilon)^2\} + o_p(h^2)\,. \qquad \text{(A.2)}$$

Noting that (2.7) implies $E\{|\epsilon|\,H(\epsilon)^2\} < \infty$, it may be proved that for $j = 0, 1$,

$$\begin{aligned} E(T_{j2}) = {} & h^j\,\kappa_j\,f(x)\,E\{\epsilon\,H(\epsilon)^j\} + h^{j+1}\,\kappa_{j+1}\,f'(x)\,E\{\epsilon\,H(\epsilon)^{j+1}\} \\ & + \tfrac{1}{2}h^{j+2}\,\kappa_{j+2}\,f''(x)\,E\{\epsilon\,H(\epsilon)^{j+2}\} + o(h^{j+2})\,, \qquad \text{(A.3)} \\ \operatorname{var}(T_{j2}) \sim {} & n^{-1}h^{2j-1}\,E\{\epsilon^2 H(\epsilon)^{2j-1}\}\,f\int y^{2j}K(y)^2\,dy\,, \qquad \text{(A.4)} \end{aligned}$$

where in (A.3), when $j = 1$, it is understood that we take the expansion only up to a remainder of $o(h^2)$. Lindeberg's condition for the series T_{02}, normalised by its standard deviation, holds provided that, for any $C_1, C_2 > 0$,

$$(nh)^{-1}\sum_{i=1}^{n} E[\epsilon_i^2\,H(\epsilon_i)^{-2}\,I(|X_i - x| \le C_1h_i)\,I\{|\epsilon_i|\,H(\epsilon_i)^{-1} > C_2\,(nh)^{1/2}\}] \to 0\,. \qquad \text{(A.5)}$$

(See e.g., Chung (1974), p.205, for Lindeberg's condition.) Result (A.5) follows from the fact that $E\{\epsilon^2 H(\epsilon)^{-1}\} < \infty$, and so by Lindeberg's central limit theorem, $(T_{02} - ET_{02})/(\operatorname{var} T_{02})^{1/2}$ is asymptotically Normal $\mathrm{N}(0, 1)$. Combining this

result with (A.3) and (A.4) we deduce that, for fixed values of the argument x,

$$\begin{aligned}\hat{g}_2 &\equiv (S_2T_{02} - S_1T_{12})/(S_2S_0 - S_1^2)\\ &= \tfrac{1}{2}\, f'' f^{-1}\, \kappa_2\, h^2\, E\{\epsilon\, H(\epsilon)^2\} + [(nh)^{-1}\, f^{-1}\, \kappa\, E\{\epsilon^2\, H(\epsilon)^{-1}\}]^{1/2}\, N_n + o_p(h^2)\,,\end{aligned}$$

where N_n is asymptotically Normal $N(0,1)$. The theorem follows from this expansion and (A.2), on noting that $\hat{g} = \hat{g}_1 + \hat{g}_2$.

A.2. Proof of Theorem 2.2.

Put $\widehat{K_i} = K\{(X_i - x)/\hat{h}_i\}$ and $\widehat{U}(a) = n^{-1}\, \sum_i \hat{h}_i^{-1}\, a_i\, \widehat{K_i}$, and let $\widehat{S}_j$, $\widehat{T}_{j1}$, $\widehat{T}_{j2}$, $\widehat{T}_j$ and $\widehat{R}_j$ denote the versions of $\widehat{U}(a)$ that arise with $a_i(x) = (X_i - x)^j$, $g(X_i)\,(X_i - x)^j$, $\epsilon_i\,(X_i - x)^j$, $Y_i\,(X_i - x)^j$ and $\hat{h}_i^j$, respectively. Then, $\widehat{T}_j = \widehat{T}_{j1} + \widehat{T}_{j2}$ and $\hat{g} = (\widehat{S}_2\widehat{T}_0 - \widehat{S}_1\widehat{T}_1)/(\widehat{S}_2\widehat{S}_0 - \widehat{S}_1^2)$. Since H is assumed bounded away from zero and infinity then $\widehat{R}_j = O_p(h^j)$, for each j. Therefore, the following analogue of (A.1) holds, derived in a similar manner:

$$\widehat{S}_2\widehat{T}_{01} - \widehat{S}_1\widehat{T}_{11} = g\,(\widehat{S}_2\widehat{S}_0 - \widehat{S}_1^2) + \tfrac{1}{2}\, g''\,(\widehat{S}_2^2 - \widehat{S}_1\widehat{S}_3) + o_p(h^4)\,. \tag{A.6}$$

Let $C_1, C_2, \ldots$ denote generic finite, strictly positive constants, let $\eta > 0$ be a Hölder exponent appropriate for both H' and K', and put $\Delta = \tilde{g} - g$, $H_1 = H'/H$ and $h_i = h\, H(\epsilon_i)$. Since $C_1 \le H \le C_2$ then we may choose $C_3, C_4, C_5 > 0$ such that, by Taylor expansion and uniformly in i,

$$\left|\frac{H(\hat{\epsilon}_i) - H(\epsilon_i)}{H(\epsilon_i)} + \Delta(X_i) H_1(\epsilon_i)\right| \le C_3\, |\Delta(X_i)|^{1+\eta}\,, \tag{A.7}$$

$$\begin{aligned}&\left|K\Big(\frac{X_i - x}{\hat{h}_i}\Big) - K\Big(\frac{X_i - x}{h_i}\Big) + \Big(\frac{\hat{h}_i - h_i}{h_i}\Big)\Big(\frac{X_i - x}{h_i}\Big) K'\Big(\frac{X_i - x}{h_i}\Big)\right|\\ &\le C_4 \left|\frac{\hat{h}_i - h_i}{h_i}\right|^{1+\eta} I(|X_i - x| \le C_5 h)\,.\end{aligned} \tag{A.8}$$

Combining (A.7) and (A.8) we see that, with $L(u) = u\, K'(u)$,

$$\begin{aligned}&\left|K\Big(\frac{X_i - x}{\hat{h}_i}\Big) - K\Big(\frac{X_i - x}{h_i}\Big) - \Delta(X_i)\, H_1(\epsilon_i)\, L\Big(\frac{X_i - x}{h_i}\Big)\right|\\ &\le C_6\, |\Delta(X_i)|^{1+\eta}\, I(|X_i - x| \le C_5 h)\,.\end{aligned} \tag{A.9}$$

Similarly but more simply, with $H_2 = H'/H^2$ we have $|h(\hat{h}_i^{-1} - h_i^{-1}) - \Delta(X_i)\, H_2(\epsilon_i)| \le C_7\, |\Delta(X_i)|^{1+\eta}$. Combining this result with (A.9), and defining $M = K + L$, we deduce that for any sequence $A_1, \ldots, A_n$,

$$\left|\sum_{i=1}^n \hat{h}_i^{-1}\, (X_i - x)^j\, A_i\, K\{(X_i - x)/\hat{h}_i\} - \sum_{i=1}^n h_i^{-1}\, (X_i - x)^j\, A_i\, K\{(X_i - x)/h_i\}\right.$$

$$-h^{-1}\sum_{i=1}^{n}\Delta(X_i)\,(X_i-x)^j\,A_i\,H_2(\epsilon_i)\,M\{(X_i-x)/h_i\}\Big|$$
$$\le C_8\,h^{-1}\sum_{i=1}^{n}|\Delta(X_i)|^{1+\eta}\,|X_i-x|^j\,|A_i|\,I(|X_i-x|\le C_5h)\,. \tag{A.10}$$

Put $\lambda_n = n^{1/5}h_0$. It is straightforward to prove that $\Delta(X_i) = O_p(n^{-2/5}\lambda_n^2)$ uniformly in values of i such that $|X_i - x| \le C_5h$. Hence, when $A_i \equiv 1$ we obtain from (A.10) the result

$$\Big|\sum_{i=1}^{n}\hat h_i^{-1}\,(X_i-x)^j\,K\{(X_i-x)/\hat h_i\} - \sum_{i=1}^{n}h_i^{-1}\,(X_i-x)^j\,K\{(X_i-x)/h_i\}\Big|$$
$$= O_p(n^{3/5}\,h^j\,\lambda_n^2)\,.$$

Equivalently, $\widehat S_j - S_j = O_p(n^{-2/5}h^j\lambda_n^2)$, where S_j is as in the proof of Theorem 2.1. From this result and the property $n^{-2/5}\lambda_n^2 = o(1)$ it may be shown that $\widehat S_j = \kappa_j\,h^j\,f\,E\{H(\epsilon)^j\} + o_p(h^j)$. The latter relation, and (A.6), imply that

$$\widehat S_2\widehat T_{01} - \widehat S_1\widehat T_{11} = g(\widehat S_2\widehat S_0 - \widehat S_1^2) + \tfrac12\,g''\,[\kappa_2\,h^2\,f\,E\{H(\epsilon)^2\}]^2 + o_p(h^4)\,, \tag{A.11}$$
$$\widehat S_2\widehat S_0 - \widehat S_1^2 = \kappa_2\,h^2\,f^2\,E\{H(\epsilon)^2\} + o_p(h^2)\,. \tag{A.12}$$

The only other sequence $A_1,\dots,A_n$ in which we are interested is $A_i \equiv \epsilon_i$, and there we may deduce from (A.10) the bound

$$\Big|\sum_{i=1}^{n}\hat h_i^{-1}\,(X_i-x)^j\epsilon_iK\{(X_i-x)/\hat h_i\} - \sum_{i=1}^{n}h_i^{-1}(X_i-x)^j\epsilon_iK\{(X_i-x)/h_i\} - V_j\Big|$$
$$= O_p(n^{(3-2\eta)/5}\,h^j\,\lambda_n^{2(1+\eta)})\,, \tag{A.13}$$

where $V_j \equiv h^{-1}\sum_i\,\Delta(X_i)(X_i-x)^j\epsilon_iH_2(\epsilon_i)M\{(X_i-x)/h_i\}$. By Taylor-expanding the ratio formula for a local-linear estimator (see e.g., the expression for $\hat a$ given by Fan (1993, p.197) it may be proved that $\Delta(u) = h_0^2\psi_n(u) + (nh_0)^{-1}f(u)^{-1}\sum_i\,\epsilon_i$ $\times K\{(X_i-u)/h_0\} + O_p(h_0^4)$, uniformly in values u in a neighbourhood of x, where ψ_n denotes a deterministic function that satisfies $\psi_n(u) = \frac12\,g''(u)\,\kappa_2 + o(\xi_n)$ uniformly in a neighbourhood of x, and $\xi_n = \xi(h_0)$, $\xi(\cdot)$ being as defined at (2.9). Therefore, $V_j = V_{j1} + V_{j2} + o_p(n^{3/5}h^j)$, where

$$V_{j1} = h^{-1}\,h_0^2\sum_{i=1}^{n}(X_i-x)^j\,\psi_n(X_i)\,\epsilon_i\,H_2(\epsilon_i)\,M\{(X_i-x)/h_i\}\,,$$
$$V_{j2} = (nhh_0)^{-1}f(x)^{-1}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\epsilon_{i_1}\epsilon_{i_2}H_2(\epsilon_{i_2})(X_{i_2}-x)^jK\{\frac{X_{i_1}-X_{i_2}}{h_0}\}M\{\frac{X_{i_2}-x}{h_{i_2}}\}\,.$$

Noting that $\int u^j M(u)\,du = 0$ for $j = 0, 1$ we may prove that for $j = 0, 1$ and $k = 1, 2$, $E(V_{j1}) = o(n^{3/5}h^j)$, $E(V_{j2}) = o(n^{3/5}h^j)$, $V_{jk} - E(V_{jk}) = o_p(n^{3/5}h^j)$. When treating V_{j2} we consider the diagonal and off-diagonal terms separately. The absolute value of the sum of the diagonal terms is adequately bounded by the sum of the absolute values of the summands. Since, conditional on the X_i's, the ϵ_i's are independent, then the variance of the sum of the off-diagonal terms in V_{j2} is readily computed under the assumption that $E(\epsilon^2) < \infty$; in particular, we do not need finite fourth moments. However, to prove that $\text{var}(V_{j2}) = o(n^{6/5}h^{2j})$ we do require the assumption that $n^{1/5}h_0 \to \infty$.

Therefore, $V_j = o_p(n^{3/5}h^j)$ for $j = 0, 1$, and so by (A.13),

$$\widehat{T}_{j2} = T_{j2} + n^{-1}(V_{j1} + V_{j2}) + O_p(n^{-2(1+\eta)/5}\, h^j\, \lambda_n^{2(1+\eta)}) = T_{j2} + o_p(n^{-2/5}\, h^j) \quad \text{(A.14)}$$

for $j = 0, 1$, where T_{j2} is as in the proof of Theorem 2.1. We know from that proof that $|T_{02}| + |h^{-1}T_{12}| = O_p\{h^2 + (nh)^{-1/2}\}$. Hence, by (A.12) and (A.14),

$$\widehat{S}_2\widehat{T}_{02} - \widehat{S}_1\widehat{T}_{12} = f^{-1}\,(\widehat{S}_2\widehat{S}_0 - \widehat{S}_1^2)\,T_{02} + o_p[h^2\,\{h^2 + (nh)^{-1/2}\} + n^{-2/5}h^2]\,. \quad \text{(A.15)}$$

Combining (A.11), (A.12) and (A.15) we deduce that

$$\begin{aligned}\hat{g} &= \{(\widehat{S}_2\widehat{T}_{01} - \widehat{S}_1\widehat{T}_{11}) + (\widehat{S}_2\widehat{T}_{02} - \widehat{S}_1\widehat{T}_{12})\}\,(\widehat{S}_2\widehat{S}_0 - \widehat{S}_1^2)^{-1} \\ &= g + \tfrac{1}{2}\,g''\,\kappa_2\,h^2\,E\{H(\epsilon)^2\} + f^{-1}\,T_{02} + o_p\{h^2 + (nh)^{-1/2}\}\,. \end{aligned} \quad \text{(A.16)}$$

We showed during the proof of Theorem 2.1 that $T_{02} = \frac{1}{2}\,f''\,\kappa_2\,h^2\,E\{\epsilon\,H(\epsilon)^2\} + [(nh)^{-1}\,f\,\kappa\,E\{\epsilon^2\,H(\epsilon)^{-1}\}]^{1/2}\,N_n' + o_p(h^2)$, where N_n' converges to Normal $\text{N}(0, 1)$. Theorem 2.2 follows from this result and (A.16).

References

Abramson, I. S. (1982). On bandwidth variation in kernel estimates — a square root law. *Ann. Statist.* **9**, 168-176.

Abramson, I. S. (1984). Adaptive density flattening — a metric distortion principle for combating bias in nearest neighbour methods. *Ann. Statist.* **12**, 880-886.

Choi, E. and Hall, P. (1998). On bias reduction in local linear smoothing. *Biometrika* **85**, 333-346.

Chung, K. L. (1974). *A Course in Probability Theory.* Academic Press, New York.

Donoho, D. L. and Liu, R. C. (1991). Geometrizing rates of convergence, III. *Ann. Statist.* **19**, 668–701.

Donoho, D. L., Liu, R. C. and MacGibbon, B. (1990). Minimax risk over hyperrectangles, and implications. *Ann. Statist.* **18**, 1416-1437.

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

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

Hall, P. (1990). On the bias of variable bandwidth curve estimators. *Biometrika* **77**, 529-536.

Hall, P. and Turlach, B. A. (1997). Interpolation methods for adapting to sparse design in nonparametric regression. (With discussion and rejoinder.) *J. Amer. Statist. Assoc.* **92**, 466-472.

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

Jones, M. C. (1990). Variable kernel density estimates and variable kernel density estimates. *Austral. J. Statist.* **32**, 361-371.

Jones, M. C., Linton, O. and Nielsen, J. P. (1995). A simple bias reduction method for density estimation. *Biometrika* **82**, 327-338.

Jones, M. C., McKay, I. J. and Hu, T.-C. (1994). Variable location and scale kernel density estimation. *Ann. Statist. Math.* **46**, 521-535.

Ruppert, D. and Cline, B. H. (1994). Bias reduction in kernel density estimation by smoothed empirical transformations. *Ann. Statist.* **22**, 185-210.

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

Samiuddin, M. and el-Sayyad, G. M. (1990). On nonparametric kernel density estimates. *Biometrika* **77**, 865-874.

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

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

Department of Mathematics, National Taiwan University, Taipei 106, Taiwan.

E-mail: cheng@math.ntu.edu.tw

Centre for Mathematics and Its Applications, Australian National University, Canberra, ACT 0200, Australia.

E-mail: halpstat@pretty.anu.edu.au

(Received October 2000; accepted October 2001)