Bernoulli **12**(3), 2006, 469–490

A simple nonparametric estimator of a strictly monotone regression function

HOLGER DETTE*, NATALIE NEUMEYER** and KAY F. PILZ†

Fakultät für Mathematik, Ruhr-Universität Bochum, 44780 Bochum, Germany.
*E-mail: *holger.dette@ruhr-uni-bochum.de; **natalie.neumeyer@ruhr-uni-bochum.de;*
†kay.pilz@ruhr-uni-bochum.de

A new method for monotone estimation of a regression function is proposed, which is potentially attractive to users of conventional smoothing methods. The main idea of the new approach is to construct a density estimate from the estimated values $\hat{m}(i/N)$ $(i = 1, \ldots, N)$ of the regression function and to use these 'data' for the calculation of an estimate of the inverse of the regression function. The final estimate is then obtained by a numerical inversion. Compared to the currently available techniques for monotone estimation the new method does not require constrained optimization. We prove asymptotic normality of the new estimate and compare the asymptotic properties with the unconstrained estimate. In particular, it is shown that for kernel estimates or local polynomials the bandwidths in the procedure can be chosen such that the monotone estimate is first-order asymptotically equivalent to the unconstrained estimate. We also illustrate the performance of the new procedure by means of a simulation study.

Keywords: isotone regression; local linear regression; Nadaraya–Watson estimator; order-restricted inference

1. Introduction

Smoothing as a means of modelling nonlinear structure in data has become increasingly popular in numerous applications. However, in many cases monotone estimates of the regression function are required, because physical considerations suggest that the response is a monotone function of the explanatory variable. There exists a vast literature on the problem of estimating a regression function m which is believed to be monotone; see the recent reviews by Delecroix and Thomas-Agnan (2000) or Gijbels (2005). Brunk (1955) proposed a modified maximum likelihood. Because this estimate is not smooth in general, Mukerjee (1988) modified it and obtained a monotone estimate with properties similar to those of nonparametric regression estimators; see also Cheng and Lin (1981), Wright (1981), Friedman and Tibshirani (1984) and Mammen (1991) for similar procedures. Monotone nonparametric regression estimators based on constrained spline smoothing have been proposed by Ramsay (1988, 1998), Kelly and Rice (1990), Mammen and Thomas-Agnan (1999), while Mammen *et al.* (2001) suggested projection-based techniques for constrained smoothing. Recently, Hall and Huang (2001) proposed a new method for

1350–7265

monotonizing a general kernel type estimator, which modifies the weights in a kernel estimator such that the modified function is monotone.

In the present paper we propose an alternative construction of monotone regression functions. The method can easily be motivated by considering an independent and identically distributed (i.i.d.) sample of uniform random variables, say $U_1, \ldots, U_N \sim \mathcal{U}([0, 1])$. If m is a strictly increasing function on the interval [0, 1] with positive derivative, K_d is a kernel function and h_d a bandwidth, then

$$\frac{1}{Nh_d} \sum_{i=1}^{N} K_d\left(\frac{m(U_i) - u}{h_d}\right)$$

is the classical kernel estimate of the density $(m^{-1})'(u) I_{[m(0), m(1)]}(u)$ of the random variable $m(U_1)$. Consequently,

$$\frac{1}{Nh_d} \int_{-\infty}^{t} \sum_{i=1}^{N} K_d\left(\frac{m(U_i) - u}{h_d}\right) \mathrm{d}u \tag{1.1}$$

is a consistent estimate of the function m^{-1} at the point t. In the context of nonparametric regression $m(X) = \mathrm{E}[Y|X]$ is the regression of Y with respect to X and the function m can be estimated by any standard method (kernel type, local polynomial, series or spline estimator), which yields an estimate of the inverse of the strictly increasing function m. The corresponding estimate of m is finally obtained by inversion of this estimate. Thus the new monotone smoother is constructed in three steps and uses two smoothing parameters. It starts with an unconstrained estimate of the regression function, say $\hat{m}$. In a second step a density estimate of the observations $\hat{m}(U_i)$ is calculated, which is integrated to obtain an estimate of the inverse of the regression function. The final step is the inversion of this estimate.

The estimate is carefully described in Section 2, where we also discuss some of its main properties as a monotone approximation of a given function. In Section 3 we study some of its statistical properties and prove asymptotic normality if kernel type or local polynomial estimators are used for the preliminary estimation of the regression function. In particular, we show that for local linear estimators the bandwidths in the procedure can be chosen such that the new estimate is asymptotically first-order equivalent to the unconstrained estimate. The choice of the smoothing parameters is also investigated from an asymptotic point of view. In Section 4 we discuss the finite-sample properties of the new estimator by means of a simulation study. Finally, some of the technical details are given in the Appendix. The main advantages of the new procedure are its simplicity (because it does not require any constrained optimization techniques) and its asymptotic equivalence to the unconstrained estimate. The new estimator is also asymptotically first-order equivalent to the estimates based on smoothing an isotone regression (or vice versa) as considered in Mammen (1991) and to a tilting estimate proposed by Hall and Huang (2001). A finite-sample comparison shows slight advantages of the new method with respect to the mean squared error (MSE) criterion.

We finally note that isotone regression has been criticized because practitioners do not believe in all those flat spots. Ramsay (1998) proposed a procedure for estimating a smooth and strictly increasing function which is computationally convenient. However, this method

is semiparametric in the sense that it requires the regression function to satisfy a specific second-order differential equation. The solution of this equation is of the form $m(x) = c_0 + c_1 \int \exp(\int w(x)\mathrm{d}x)\mathrm{d}x$, where c_0, c_1 are arbitrary constants and w is a square-integrable unconstrained function. As a consequence the procedure of Ramsay (1998) is not consistent in general (for example, for the regression functions $m(x) = x^p$, $x \in [0, 1]$, $p > 0$). An important contribution of this paper is that it provides a simple smooth, strictly monotone and generally consistent nonparametric estimator of the regression function. Furthermore, R code for the new estimator is available. All these properties should make the new method particularly attractive to users of conventional kernel methods.

2. Monotone smoothing by inversion

Consider the nonparametric regression model

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i, \qquad i = 1, \ldots, n, \tag{2.1}$$

where $\{(X_i, Y_i)\}_{i=1}^n$ is a bivariate sample of i.i.d. observations such that X_i has a positive twice continuously differentiable density f with compact support, say $[0, 1]$. We further assume that the random variables ε_i are i.i.d. with $\mathrm{E}[\varepsilon_i] = 0$, $\mathrm{E}[\varepsilon_i^2] = 1$ and finite fourth moment. The variance function $\sigma : [0, 1] \to \mathbb{R}^+$ and the regression function $m : [0, 1] \to \mathbb{R}$ are assumed to be continuous and twice continuously differentiable, respectively. Throughout this paper we restrict ourselves to the case of an isotone regression function. Corresponding results for the antitone case are very similar and obtained by the same reasoning. If there is evidence that the regression function m is (strictly) increasing we define, for $N \in \mathbb{N}$,

$$\hat{m}_I^{-1}(t) := \frac{1}{Nh_d} \sum_{i=1}^{N} \int_{-\infty}^{t} K_d\left(\frac{\hat{m}(i/N) - u}{h_d}\right)\mathrm{d}u \tag{2.2}$$

as an estimate of $m^{-1}(t)$, where

$$\hat{m}(x) = \frac{\sum_{i=1}^n K_r((X_i - x)/h_r)Y_i}{\sum_{i=1}^n K_r((X_i - x)/h_r)}$$

is the classical Nadaraya–Watson estimate, K_d and K_r denote symmetric kernels with compact support, say $[-1, 1]$, and finite second moment, and h_d, h_r are the corresponding bandwidths converging to 0 with increasing sample size n. We assume that K_d is twice continuously differentiable on its support and that the kernel K_r has been appropriately modified at the boundary; see Müller (1985). For the sake of transparency we restrict ourselves to the Nadaraya–Watson estimate, but it is notable that all results in this paper remain valid (subject to an appropriate modification of constants) for other types of kernel estimators such as the Gasser–Müller estimator (see Gasser and Müller 1979) or local polynomials (Wand and Jones 1995; Fan and Gijbels 1996).

Note that the indices r and d correspond to 'regression' and 'density' because we combine a regression with a density estimate to define the estimator in (2.2). Comparing this estimate with the motivation in equation (1.1), we see that the uniformly distributed

random variables have been replaced by an equidistant design. It is not necessary (and in many cases not desirable) that the number N of design points coincides with the sample size n. Finally, we note that the estimate $\hat{m}_I^{-1}$ is isotone if the kernel K_d is positive, which will be assumed throughout this paper. In this case an isotone estimate of the regression function $\hat{m}_I$ is simply obtained by reflection of the function $\hat{m}_I^{-1}$ in the line $y = x$. Note that the estimator $\hat{m}_I^{-1}$ is equal to 1 if $t > \max_{i=1}^N \hat{m}(i/N) + h_d$ and to 0 if $t < \min_{i=1}^N \hat{m}(i/N) - h_d$, and that the inverse of the function $\hat{m}_I$ is calculated only for $t \in [\min_{i=1}^N \hat{m}(i/N), \max_{i=1}^N \hat{m}(i/N)]$.

It is heuristically clear that the estimate $\hat{m}_I^{-1}$ is in some sense close to the function

$$m_N^{-1}(t) = \frac{1}{Nh_d} \int_{-\infty}^{t} \sum_{i=1}^{N} K_d\left(\frac{m(i/N) - u}{h_d}\right) \mathrm{d}u = \int_0^1 I\{m(x) \leqslant t\} \mathrm{d}x + o(1) \tag{2.3}$$

(note that the kernel K_d has compact support). In other words, the statistic $\hat{m}_I^{-1}(t)$ is a consistent estimate of the quantity $\int_0^1 I\{m(x) \leqslant t\} \mathrm{d}x$ and this property does not depend on the particular consistent estimate $\hat{m}$ used in the regression step. The leading term on the right-hand side of equation (2.3) is equal to $m^{-1}(t) = \inf\{u | m(u) > t\}$ if the regression function is increasing, and the following lemma gives the precise order of this approximation.

Lemma 2.1. *If the regression function is strictly increasing and the assumptions stated at the beginning of this section are satisfied, then we have for any* $t \in (m(0), m(1))$ *with* $m'(m^{-1}(t)) > 0$,

$$m_N^{-1}(t) = m^{-1}(t) + \kappa_2(K_d) h_d^2 (m^{-1})''(t) + o(h_d^2) + O\left(\frac{1}{Nh_d}\right),$$

where the constant $\kappa_2(K)$ *is given by*

$$\kappa_2(K) = \frac{1}{2} \int_{-1}^{1} v^2 K(v) \mathrm{d}v. \tag{2.4}$$

It is easy to see that the functions m_N^{-1} and $\hat{m}_I^{-1}$ are strictly increasing, if $\max_{i=1}^{N-1} v_{(i+1)} - v_{(i)} < 2h_d$, where $v_i = m(i/N)$ and $v_i = \hat{m}(i/N)$, respectively. If the sample size n and number of design points N are chosen sufficiently large, this inequality is satisfied because of the continuity of the regression function m and the estimate $\hat{m}$. Throughout this paper m_N denotes the inverse of the function m_N^{-1}. Because m_N^{-1} is expected to be an approximation of the function m^{-1}, it is intuitively clear that the inverse m_N of m_N^{-1} is an approximation of the function m. The following lemma makes this statement precise and is proved in the Appendix.

Lemma 2.2. *If the regression function* m *is strictly increasing and the assumptions stated at the beginning of this section are satisfied, then we have for any* $t \in (0, 1)$ *with* $m'(t) > 0$,

$$m_N(t) = m(t) + \kappa_2(K_d) h_d^2 \frac{m''(t)}{(m'(t))^2} + o(h_d^2) + O\left(\frac{1}{Nh_d}\right).$$

If m is not necessarily increasing, the function $g : t \to \int_0^1 I\{m(x) \leqslant t\}\mathrm{d}x$ or its approximation

$$g_{h_d} : t \to \int_0^1 \int_{-\infty}^{t} \frac{1}{h_d} K_d\left(\frac{m(x) - u}{h_d}\right)\mathrm{d}u\,\mathrm{d}x$$

is still well defined and non-decreasing. Note that the function g is not necessarily differentiable (see the examples presented below) and g_{h_d} can be considered as a smooth version of g which converges to g if $h_d \to 0$. The (generalized) inverse $g_{h_d}^{-1}$ can be considered as an approximation of the function m by a non-decreasing smooth function. Some mathematical properties of a related function have been discussed in the context of measure-preserving transformations and non-decreasing rearrangements, and the interested reader is referred to the work of Ryff (1965, 1970) or Bennett and Sharpley (1988), among others. Further properties of this function will be briefly described in the following. For the sake of brevity we restrict ourselves to the function g and mention that the properties of g_{h_d} are similar.

If for a fixed t_0 the set $m^{-1}(\{t_0\}) = \{x_0\}$ is a singleton and $m'(x_0) > 0$, then, obviously, $g(t_0) = x_0$ and $g^{-1}(x_0) = m(x_0)$. Now let $x_0 \in [0, 1]$ denote the infimum of all points such that there exists a t_0 with this property (note that the case $x_0 = 0$ is not excluded) and define $x_1 \geqslant 0$ as the maximal point such that this property is satisfied for all $x \in (x_0, x_1)$ with corresponding value $t_1 = m(x_1)$. In this case we have for all $t = m(x) \in [t_0, t_1]$ the representation $g(t) = x_0 + (x - x_0) = x$, which proves $g^{-1}(x) = m(x)$ for all $x \in [x_0, x_1]$. If $x_1 < 1$, the function m is decreasing in a neighbourhood $(x_1, x_1 + \varepsilon)$ and there may exist a second interval, say (x_2, x_3), such that m is strictly increasing on (x_2, x_3) and such that for all $t \in (m(x_2), m(x_3))$ the set $m^{-1}(\{t\})$ is a singleton. For this interval the same argument shows $g^{-1}(x) = m(x)$ for all $x \in [x_2, x_3]$. The repetition of this argument shows that on any interval $[a, b]$, where m is strictly increasing such that $m^{-1}(\{a\})$ and $m^{-1}(\{b\})$ are singletons the inverse of the function g coincides with the regression function m.

Example 2.1. Consider the function

$$m(x) = \frac{11}{3}x - 8x^2 + \frac{16}{3}x^3,$$

which is strictly increasing on the intervals $[0, (6 - \sqrt{3})/12]$, $[(6 + \sqrt{3})/12, 1]$. The functions m, g and g^{-1} are depicted in Figure 2.1(a), where the function g^{-1} coincides with the function m whenever $m^{-1}(\{t\})$ is a singleton. Note that there are two points where the function g is not differentiable. Figure 2.1(b) illustrates the approximation of the oscillating function

$$m(x) = x + \frac{1}{4}\sin(4\pi x)$$

by the monotone increasing function g^{-1}.

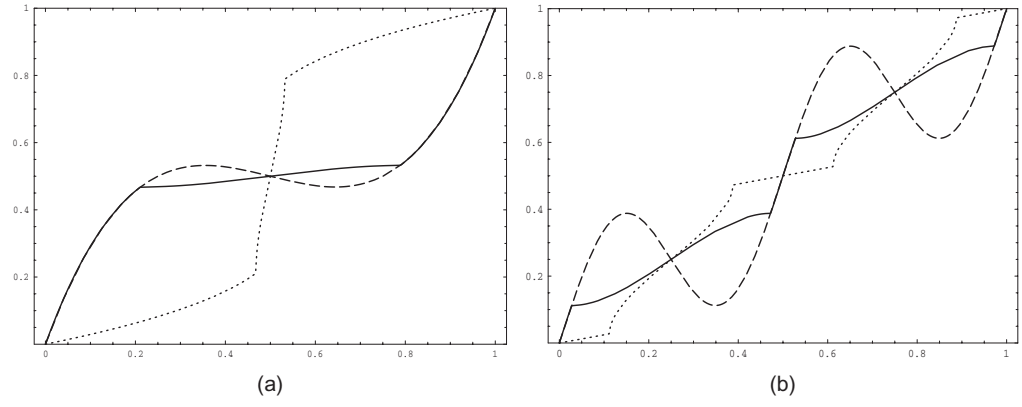

Figure 2.1. Approximation of a non-monotone regression function m (dashed line) by its monotone approximation g^{-1} (solid line) for (a) $m(x) = \frac{11}{3}x - 8x^2 + \frac{16}{3}x^3$, (b) $m(x) = x + \frac{1}{4}\sin(4\pi x)$. The figures show also the function $g(t) = \int_0^1 I\{m(x) \leqslant t\}\mathrm{d}x$ (dotted line).

3. Main results – asymptotic behaviour

In this section we investigate some of the asymptotic properties of the estimates $\hat{m}_{\mathrm{I}}^{-1}$ and $\hat{m}_I$. It turns out that both estimates (appropriately centred) are asymptotically normally distributed, where the standardizations depend on the limit $\lim_{h_r,h_d\to 0} h_r/h_d =: c \in [0, \infty]$ of the ratio of the smoothing parameters. In the case $c = \infty$ we show that the new monotone estimate $\hat{m}_I$ is first-order asymptotically equivalent to the unconstrained estimate $\hat{m}$, if the Nadaraya–Watson estimator or a local linear estimator is used for the estimation of the regression function.

3.1. Asymptotic normality

We assume that the smoothness conditions regarding the density, variance and regression function stated at the beginning of Section 2 are satisfied. For the bandwidths h_r and h_d in the regression and density estimate we require $h_r \to 0$, $h_d \to 0$, $nh_r \to \infty$, $nh_d \to \infty$ and additionally

$$nh_r^5 = O(1), \qquad n = O(N), \tag{3.1}$$

$$\frac{\log h_r^{-1}}{nh_r h_d^3} = o(1). \tag{3.2}$$

Note that for the 'optimal' rate in regression estimation $h_r = \gamma n^{-1/5}$ with respect to the MSE the latter assumption reduces to $h_d n^{4/15}/(\log n)^{1/3} \to \infty$.

Theorem 3.1. *If the assumptions* (3.1) *and* (3.2) *are satisfied,* $\lim_{n\to\infty} h_r/h_d = c \in [0, \infty)$

exists and m is strictly increasing, then it follows that, for all $t \in (m(0), m(1))$ with $m'(m^{-1}(t)) > 0$,

$$\sqrt{nh_d}\left(\hat{\boldsymbol{m}}_I^{-1}(t) - m_N^{-1}(t) + \kappa_2(K_r)h_r^2\left(\frac{m''f + 2m'f'}{fm'}\right)(m^{-1}(t))\right) \xRightarrow{\mathcal{D}} \mathcal{N}(0,\ r^2(t)),$$

where the constant $\kappa_2(K_r)$ is defined in (2.4) *and the asymptotic variance is given by*

$$r^2(t) = \frac{\sigma^2(m^{-1}(t))}{m'(m^{-1}(t))f(m^{-1}(t))} \times \int\int\int K_d(w + cm'(m^{-1}(t))(v-u))K_d(w)K_r(u)K_r(v)\mathrm{d}w\,\mathrm{d}u\,\mathrm{d}v. \tag{3.3}$$

If $\lim_{n\to\infty} h_r/h_d = \infty$, then we have, for all $t \in (m(0), m(1))$ with $m'(m^{-1}(t)) > 0$,

$$\sqrt{nh_r}\left(\hat{\boldsymbol{m}}_I^{-1}(t) - m_N^{-1}(t) + \kappa_2(K_r)h_r^2\left(\frac{m''f + 2m'f'}{fm'}\right)(m^{-1}(t))\right) \xRightarrow{\mathcal{D}} \mathcal{N}(0,\ \tilde{r}^2(t)),$$

where the asymptotic variance is given by

$$\tilde{r}^2(t) = \frac{\sigma^2(m^{-1}(t))}{\{m'(m^{-1}(t))\}^2 f(m^{-1}(t))}\int K_r^2(u)\mathrm{d}u.$$

Note that for sufficiently large n and N the functions $\hat{\boldsymbol{m}}_{\mathrm{I}}^{-1}$ and m_N^{-1} are strictly increasing independent of the monotonicity of the 'true' regression function m. The following result shows that the corresponding inverse functions $\hat{\boldsymbol{m}}_I$ and m_N also satisfy an asymptotic normal law.

Theorem 3.2. *Assume that the assumptions of Theorem 3.1 are satisfied and let $\hat{\boldsymbol{m}}_I$ and m_N denote the inverse functions of the functions $\hat{\boldsymbol{m}}_I^{-1}$ and m_N^{-1} defined by* (2.2) *and* (2.3)*, respectively. If $\lim_{n\to\infty} h_r/h_d = c \in [0, \infty)$ exists, then we have, for every $t \in (0, 1)$ with $m'(t) > 0$,*

$$\sqrt{nh_d}\left(\hat{\boldsymbol{m}}_I(t) - m_N(t) - \kappa_2(K_r)h_r^2\left(\frac{m''f + 2m'f'}{f}\right)(t)\right) \xRightarrow{\mathcal{D}} \mathcal{N}(0,\ s^2(t)),$$

where the asymptotic variance is given by

$$s^2(t) = \frac{\sigma^2(t)m'(t)}{f(t)}\int\int\int K_d(w + cm'(t)(v-u))K_d(w)K_r(u)K_r(v)\mathrm{d}w\,\mathrm{d}u\,\mathrm{d}v. \tag{3.4}$$

If $\lim_{n\to\infty} h_r/h_d = \infty$ it follows that, for every $t \in (0, 1)$ with $m'(t) > 0$,

$$\sqrt{nh_r}\left(\hat{\boldsymbol{m}}_I(t) - m_N(t) - \kappa_2(K_r)h_r^2\left(\frac{m''f + 2m'f'}{f}\right)(t)\right) \xRightarrow{\mathcal{D}} \mathcal{N}(0,\ \tilde{s}^2(t)),$$

where the asymptotic variance is given by

$$\tilde{s}^2(t) = \frac{\sigma^2(t)}{f(t)} \int K_r^2(u)\mathrm{d}u.$$

Remark 3.1. For a regression function m with $m'(m^{-1}(t)) = 0$, Theorem 3.1 (and 3.2) are no longer valid – see (A.11) and (A.13) in the proof of Theorem 3.1 in the Appendix. Note also that in this case the variance and bias in Theorem 3.1 are undefined and that the proof of Theorem 3.2 is based on Theorem 3.1 – see (A.14) in the proof of Theorem 3.2.

Remark 3.2. It follows from Lemma 2.2 and the second part of Theorem 3.2 that in the case $h_d = o(h_r)$ the monotone estimator $\hat{\boldsymbol{m}}_I$ exhibits the same first-order asymptotic behaviour as the unconstrained estimate $\hat{\boldsymbol{m}}$. A similar property was observed by Mammen (1991) for the L^2-projection of the Nadaraya–Watson estimate onto the space of all increasing functions. It is also notable that Theorems 3.1 and 3.2 are applicable for the optimal bandwidth with respect to uniform convergence; see Mack and Silverman (1982). In this case bandwidths satsfying $h_r = o(h_d)$ (i.e. $c = 0$) have to be applied.

3.2. Bandwidth selection

The choice of the two bandwidths is essential for the performance of the new smoothing procedure. While the bandwidth h_r for the regression estimate $\hat{\boldsymbol{m}}$ can be chosen by standard methods, the choice of the bandwidth h_d in the second step of the density estimate is less clear. In the following we will demonstrate that bandwidths satisfying $h_d = o(h_r)$ should be preferred from an asymptotic point of view if the MSE is used to compare estimates obtained from different choices for the bandwidth h_d. For this we assume that $f \equiv 1$ (or that a local linear estimate is used as the unconstrained estimate $\hat{\boldsymbol{m}}$) and note that by Theorem 3.2 the leading term of the bias of the estimate $\hat{\boldsymbol{m}}_I$ is given by

$$\Gamma_I(h_d, h_r) = \kappa_2(K_d)\frac{m''(t)}{(m'(t))^2} h_d^2 + \kappa_2(K_r)m''(t)h_r^2.$$

We choose the bandwidth $h_d = \gamma\, m'(t)h_r$ in the estimate $\hat{\boldsymbol{m}}_I$, for some constant $\gamma > 0$. As a consequence of Theorem 3.2, the estimate $\hat{\boldsymbol{m}}_I$ is asymptotically normal distributed with bias

$$[\kappa_2(K_r) + \gamma^2\kappa_2(K_d)]m''(t)h_r^2$$

and variance

$$\frac{\sigma^2(t)}{nh_r f(t)}\mu_K^2(\gamma),$$

where

$$\mu_K^2(\gamma) = \int \left(\int K_r(u)K_r(u + \gamma w)\mathrm{d}u \right) \left(\int K_d(v)K_d(v + w)\mathrm{d}v \right) \mathrm{d}w.$$

Note that it is easy to see that these calculations also include the case $\gamma = 0$ (corresponding to the second part of Theorem 3.2), if this is interpreted as $\lim_{h_r, h_d \to 0} h_r/h_d = \infty$. Numerical

results show that for the commonly used kernels the function μ_K^2 is decreasing with γ. Therefore the variance of the statistic $\hat{\boldsymbol{m}}_I$ is decreasing with γ, while the converse holds for the bias. Heuristically, the choice $\gamma = 0$ may have particular advantages if the standard error is small compared to the bias, while values as $\gamma = 0.5$ or $\gamma = 1$ may be appropriate for a small bias and larger standard errors.

In the rest of this section we study the effect of the choice of γ on the rule $h_d = \gamma\, m'(t) h_r$ if the local optimal bandwidth

$$h_r = \left(\frac{\mu_K^2(0)\sigma^2(t)}{4f(t)(m''(t))^2 \kappa_2^2(K_r) n} \right)^{1/5} \tag{3.5}$$

with respect to the MSE criterion is used for the estimation of the regression function. A standard calculation shows that for this choice the first-order approximation of the MSE is given by

$$\mathrm{mse}(\gamma) = \left(\frac{\mu_K^2(0)\sigma^2(t)}{4nf(t)} \right)^{4/5} (m''(t)\kappa_2(K_r))^{2/5} \left\{ \left(1 + \gamma^2 \frac{\kappa_2(K_d)}{\kappa_2(K_r)} \right)^2 + \frac{4}{\mu_K^2(0)} \mu_K^2(\gamma) \right\}.$$

The corresponding MSE for the unconstrained estimate is given by mse(0), which gives

$$e(\gamma) = \frac{\mathrm{mse}(\gamma)}{\mathrm{mse}(0)} = \frac{(1 + \gamma^2 \kappa_2(K_d)/\kappa_2(K_r))^2 + (4/\mu_K^2(0))\mu_K^2(\gamma)}{5}. \tag{3.6}$$

Figure 3.1 shows the function e for the cases where $K_r = K_d$ is the Epanechnikov and rectangular kernel. We see that for these kernels the optimal choice (minimizing $e(\gamma)$ with respect to the parameter γ) is $\gamma = 0$, which corresponds to the case $\lim_{h_r, h_d \to 0} h_r / h_d = \infty$, while for the rectangular kernel the choice $\gamma \approx 0.3$ yields the smallest efficiency. In this case the bandwidths h_d and h_r should be chosen of the same order according to the rule $h_d = \gamma\, m'(t) h_r$ with (3.5). We investigated several other kernels (including the beta family) and conclude that the situation displayed in Figure 3.1(b) for the Epanechnikov kernel is quite typical. All kernels except the rectangular yield the same picture for the efficiency as displayed in Figure 3.1(b) for the Epanechnikov kernel. These results indicate that from an asymptotic point of view the bandwidths h_d and h_r should not be of the same order, but bandwidths satisfying $h_d = o(h_r)$ should be preferred for the monotone estimator.

4. Finite-sample properties

In this section we illustrate the behaviour of the new monotone estimator for finite sample sizes. We consider the nonparametric regression model (2.1) with a uniform design and normally distributed errors. As a preliminary estimate $\hat{\boldsymbol{m}}$ we use a local linear estimator, while the density estimate is based on $N = 100$ design points. The bandwidth h_r of the unconstrained estimators is chosen as $h_r = (\hat{\sigma}^2/n)^{1/5}$, where

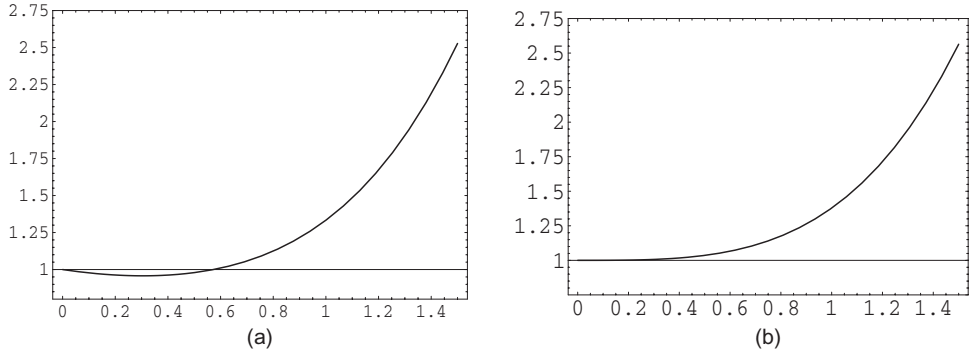

Figure 3.1. The function e defined in (3.6) for (a) the rectangular and (b) the Epanechnikov kernel, where $K_d = K_r$.

$$\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{[i+1]} - Y_{[i]})^2$$

is Rice's (1984) estimator; here $Y_{[1]}, \ldots, Y_{[n]}$ denote the observations ordered with respect to their corresponding X-values. The Epanechnikov kernel is used for K_d and K_r.

In order to illustrate the performance of the new procedure we show in Figure 4.1 the estimate $\hat{m}_I$ based on $n = 100$ observations. The standard deviation of the errors is $\sigma = 0.1$, while the bandwidth is chosen as $h_d = h_r^3$. Figure 4.1(a) corresponds to the regression function $m(x) = \frac{1}{2}(2x-1)^3 + \frac{1}{2}$ and Figure 4.1(b) to $m(x) = \sin(\frac{1}{2}\pi x)$. Three monotone estimates obtained from different simulations are displayed. These correspond to the 5th percentile (solid line), 50th percentile (dashed line) and 95th percentile (dotted line) with respect to the integrated squared error performance (based on 1000 simulation runs).

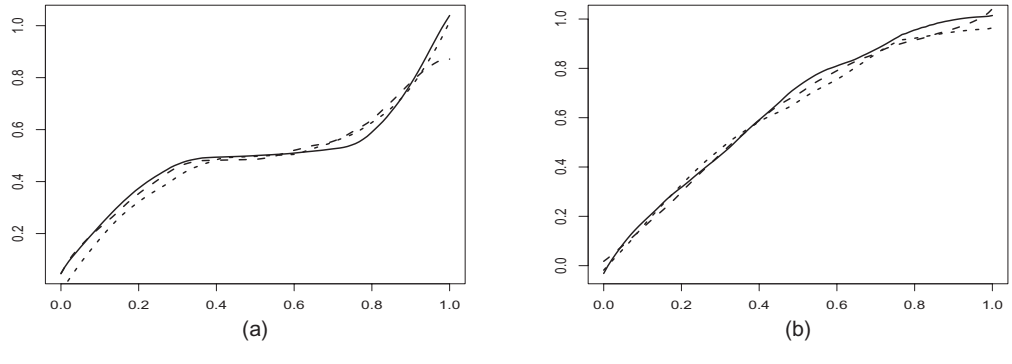

Figure 4.1. Three monotone estimates obtained from different simulations ($n = 100$ observations and $\sigma = 0.1$): (a) $m(x) = \frac{1}{2}(2x-1)^3 + \frac{1}{2}$; (b) $m(x) = \sin(\frac{1}{2}\pi x)$.

4.1. Bandwidth choice

In the second part of our simulation study we investigate the effect of the choice of the bandwidth h_d in more detail. Two choices of h_d are considered, $h_d = \frac{1}{2}h_r$ and $h_d = h_r^3$. In Figure 4.2 we show the simulated MSE, squared bias and variance of the monotone estimates for $n = 100$ observations with standard deviation $\sigma = 1$. Four cases for the regression function are considered in our study, namely $m(x) = x$ (first row), $m(x) = x^2$ (second row), $m(x) = \sin(\frac{\pi}{2}x)$ (third row), $m(x) = \frac{1}{2} + \frac{1}{2}(2x-1)^3$ (fourth row). The monotone estimates with bandwidths $h_d = h_r^3$ and $h_d = \frac{1}{2}h_r$ are represented by the solid and dashed line, respectively. The dotted lines show the unconstrained local linear estimate. In all cases we observe a smaller variance and a larger bias for the constrained estimates. Also the bandwidth $h_d = \frac{1}{2}h_r$ yields a smaller variance but a larger bias than the choice $h_d = h_r^3$. This corresponds to the asymptotic theory presented in Section 3.

The effect of the choice of the bandwith on the MSE depends on the size of the variance and the size of $m''(t)$. For regression functions with a large value of $|m''(t)|$ the bias dominates the MSE. Consider, for example, the function $m(x) = \frac{1}{2}(2x-1)^3 + \frac{1}{2}$. At the point $x = \frac{1}{2}$ we have $m'(\frac{1}{2}) = m''(\frac{1}{2}) = 0$ and the larger bandwidth $h_d = \frac{1}{2}h_r$ for the density estimate yields a smaller MSE than the choice $h_d = h_r^3$ (see the fourth row in Figure 4.2). On the other hand, if $x = \frac{1}{4}$ or $x = \frac{3}{4}$ we have $|m''(\frac{1}{4})| = |m''(\frac{3}{4})| = 6$ and the effect of the bias is visible such that a smaller bandwidth h_d in the density estimate is appropriate. Based on our numerical results we recommend the choice $h_d = h_r^3$ or $h_d = h_r^2$ for the bandwidth in the density estimation step, where the particular alternative depends on the desired smoothness of the monotone estimate. This choice has the additional advantage that the regions where boundary effects affect the density estimate are very small and that the first-order asymptotic behaviour of the monotone estimate coincides with that of the local linear estimate.

4.2. A brief comparison with other estimators

In this section we briefly compare the new estimator with two procedures for monotone estimation which are most similar in spirit to the method proposed in this paper. A detailed comparison can be found in Dette and Pilz (2004). Note that our asymptotic results in Section 3 suggest the use of bandwidths satisfying $h_d = o(h_r)$. For this choice the new estimate is asymptotically first-order equivalent to the classical smoothed isotone estimate of Brunk (1958) (see Mammen 1991) and to the tilting method proposed by Hall and Huang (2001), which will be denoted by $\hat{m}_{IS}$ and $\hat{m}_{HH}$ in the following. Note that the estimate $\hat{m}_{IS}$ is not necessarily monotone increasing, but it can be shown that the estimate $\hat{m}_{IS}$ is asymptotically first-order equivalent to the estimate obtained by projecting a smooth curve on the space of monotone functions; see Mammen (1991). Moreover, this author compared the finite-sample performance of the two isotone regression estimates obtained by interchanging the order of smoothing and isotonizing and concluded that the estimate $\hat{m}_{IS}$ usually yields a smaller MSE than the estimate obtained by projecting a non-monotone curve on the space of monotone functions; see Mammen (1991, Table 1) for more details.

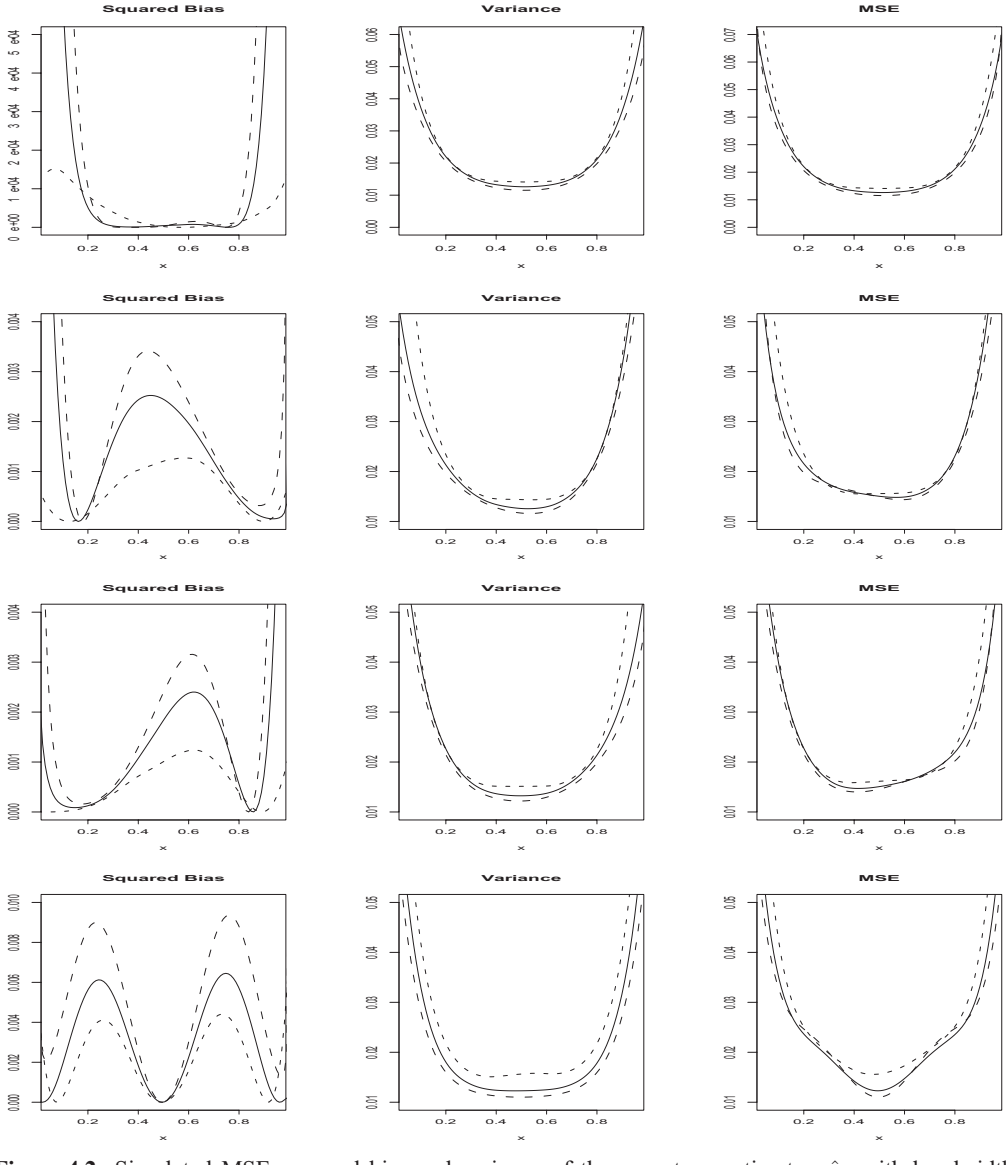

Figure 4.2. Simulated MSE, squared bias and variance of the monotone estimator $\hat{m}_I$ with bandwidth $h_d = h_r^3$ (solid line), bandwidth $h_d = 0.5h_r$ (dashed line) and the local linear estimator $\hat{m}$ (dotted line) for various regression functions and a uniform design: $m(x) = x$ (row 1), $m(x) = x^2$ (row 2), $m(x) = \sin(\frac{1}{2}\pi x)$ (row 3), $m(x) = \frac{1}{2}(2x-1)^3 + \frac{1}{2}$ (row 4). The estimates are calculated from $n = 100$ observations with standard deviation $\sigma = 1$.

Thus we included the more efficient method of these two monotone regression estimates in our numerical study.

The following example has two purposes. On the one hand, we wish to compare the three different estimation procedures. On the other hand, we would like to see the performance of the new estimate at points where the derivative of the regression function vanishes. For this reason we consider the regression function

$$m(x) = \begin{cases} 2x, & \text{if } x \in [0, \frac{1}{4}], \\ \frac{1}{2}, & \text{if } x \in [\frac{1}{4}, \frac{3}{4}], \\ 2x - 1, & \text{if } x \in [\frac{3}{4}, 1], \end{cases}$$

and display the curves for the squared bias, variance and MSE in Figure 4.3 for a normally distributed error with standard deviation 0.2 and sample size $n = 80$. These results are based on 1000 simulation runs. Note that all monotone estimation techniques require a preliminary nonparametric (unconstrained) estimate of the regression function with corresponding smoothing parameter. For the sake of comparison we use the same regression estimate $\hat{m}$ for all three methods in the first step, namely a local linear estimator (see Wand and Jones 1995) with Epanechnikov kernel. The additional bandwidth for the density estimation step in the calculation of the estimate $\hat{m}_I$ was chosen as $h_d = h_r^3$. We observe that all three estimators exhibit quite similar behaviour (as predicted by the asymptotic theory), where the new estimate $\hat{m}_I$ proposed in this paper has some advantages with respect to MSE and the worst performance is observed for the isotone estimate obtained by the tilting method. Moreover, all estimates seem to be consistent even at points where the derivative of the regression function vanishes. Further simulation results, which show a very similar picture, are available in Dette and Pilz (2004). Thus the new estimate has at least a similar finite-sample behaviour to that of its closest competitors and leads in many cases to a smaller MSE.

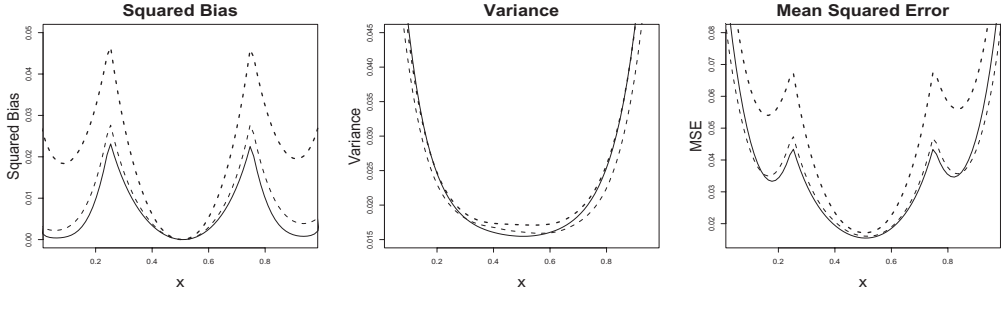

Figure 4.3. Simulated squared bias, variance and MSE of the smoothed isotonized estimate $\hat{m}_{IS}$ (dashed line), the estimator $\hat{m}_{HH}$ obtained by the tilting method (dotted line) and the estimator $\hat{m}_I$ (solid line). The sample size is $n = 80$, the standard deviation is $\sigma = 0.2$.

Appendix: Proofs

Throughout this section we assume without loss of generality that the function m has a positive derivative on the interval $[0, 1]$. The general case can easily be obtained by considering a subinterval, for which this property is satisfied (note that m' is continuous). Moreover, we assume for the sake of a transparent notation that the number of design points N in the estimate $\hat{m}_I$ equals the sample size n and write m_n instead of m_N.

Proof of Lemma 2.1. Obviously, we have

$$m_n^{-1}(t) = \int_0^1 \int_{-\infty}^t K_d\left(\frac{m(x)-u}{h_d}\right)\frac{1}{h_d}\,\mathrm{d}u\,\mathrm{d}x \cdot \left(1 + O\left(\frac{1}{nh_d}\right)\right)$$

and, observing that the support of the kernel K_d is given by the interval $[-1, 1]$, the leading term on the right-hand side is estimated as follows:

$$\begin{aligned} A(h_d) &= \int_0^1 \int_{-\infty}^t K_d\left(\frac{m(x)-u}{h_d}\right)\frac{\mathrm{d}u}{h_d}\,\mathrm{d}x = \int_0^{m^{-1}(t+h_d)} \int_{m(x)-h_d}^t K_d\left(\frac{m(x)-u}{h_d}\right)\frac{\mathrm{d}u}{h_d}\,\mathrm{d}x \\ &= m^{-1}(t-h_d) \\ &\quad + \int_0^1 I\{m^{-1}(t-h_d) \leqslant x \leqslant m^{-1}(t+h_d)\}\int_{m(x)-h_d}^t K_d\left(\frac{m(x)-u}{h_d}\right)\frac{\mathrm{d}u}{h_d}\,\mathrm{d}x \\ &= m^{-1}(t-h_d) + h_d\int_{(m(0)-t)/h_d}^{(m(1)-t)/h_d} I\{-1 \leqslant z \leqslant 1\}(m^{-1})'(t+zh_d)\int_z^1 K_d(v)\mathrm{d}v\,\mathrm{d}z. \end{aligned}$$

If $t \in (m(0), m(1))$ is fixed, we obtain from the identity $\int_{-1}^1 \int_z^1 K_d(v)\mathrm{d}v\,\mathrm{d}z = 1$ (note that K_d is symmetric and has compact support $[-1, 1]$) and a Taylor expansion,

$$\begin{aligned} A(h_d) &= m^{-1}(t-h_d) + h_d\int_{-1}^1 (m^{-1})'(t+zh_d)\int_z^1 K_d(v)\mathrm{d}v\,\mathrm{d}z \\ &= m^{-1}(t) + h_d^2(m^{-1})''(t)\left\{\frac{1}{2} + \int_{-1}^1 z\int_z^1 K_d(v)\mathrm{d}v\,\mathrm{d}z\right\} + o(h_d^2) \\ &= m^{-1}(t) + \kappa_2(K_d)h_d^2(m^{-1})''(t) + o(h_d^2) \end{aligned}$$

as $h_d \to 0$, where the last identity follows from $\int_{-1}^1 z\int_z^1 K_d(v)\mathrm{d}v\,\mathrm{d}z = \frac{1}{2}\int_{-1}^1 v^2 K_d(v)\mathrm{d}v - \frac{1}{2}$. □

For a proof of Lemma 2.2 and Theorem 3.2 it is necessary to understand the operator which maps a non-decreasing function m to its 'quantile' $m^{-1}(t)$. Consider a fixed $t \in \mathbb{R}$, and let $\mathcal{M}$ denote the set of all functions $H \in C^2[0, 1]$ with positive derivative on the interval $[0, 1]$, which contain t in the interior of their image, that is, $t \in \operatorname{int} H([0, 1])$. Consider the functional

$$\Phi : \begin{cases} \mathcal{M} \mapsto [0, 1] \\ H \mapsto H^{-1}(t) \end{cases}$$

and define for $H_1, H_2 \in \mathcal{M}$ the function

$$Q : \begin{cases} [0, 1] \mapsto \mathbb{R} \\ \lambda \mapsto \Phi(H_1 + \lambda(H_2 - H_1)). \end{cases} \tag{A.1}$$

Note that in the case of existence $Q'(0)$ is the Gatéaux derivative of the functional Φ at H_1 in the direction of $H_2 - H_1$. The following result shows that this derivative exists and also gives the second derivative.

Lemma A.1. *The mapping* $Q : [0, 1] \to \mathbb{R}$ *defined by* (A.1) *is twice continuously differentiable with*

$$Q'(\lambda) = -\frac{H_2 - H_1}{h_1 + \lambda(h_2 - h_1)} \circ (H_1 + \lambda(H_2 - H_1))^{-1}(t), \tag{A.2}$$

$$Q''(\lambda) = Q'(\lambda)\left\{\frac{-2(h_2 - h_1)}{h_1 + \lambda(h_2 - h_1)} + \frac{(H_2 - H_1)(h'_1 + \lambda(h'_2 - h'_1))}{\{h_1 + \lambda(h_2 - h_1)\}^2}\right\} \circ Q(\lambda), \tag{A.3}$$

where h_1, h_2 *denote the derivatives of* H_1, H_2, *respectively.*

Proof. Let $F(x, y) = (H_1 + x(H_2 - H_1))(y) - t$, then $Q(\lambda)$ is determined by the equation $F(\lambda, Q(\lambda)) = 0$. It is easy to see that the domain of the function Q can be extended in a neighbourhood of the interval [0, 1], and by the implicit function theorem it follows that Q is differentiable with derivative

$$Q'(\lambda) = -\frac{(H_2 - H_1) \circ Q(\lambda)}{h_1 \circ Q(\lambda) + \lambda(h_2 - h_1) \circ Q(\lambda)},$$

which proves (A.2). The calculation of the second derivative now follows by a straightforward application of the chain rule, which gives

$$\begin{aligned} Q''(\lambda) = {} & \frac{(H_2 - H_1)(h_2 - h_1)}{\{h_1 + \lambda(h_2 - h_1)\}^2} \circ Q(\lambda) \\ & - Q'(\lambda) \cdot \frac{(h_2 - h_1)(h_1 + \lambda(h_2 - h_1)) - (H_2 - H_1)(h'_1 + \lambda(h'_2 - h'_1))}{\{h_1 + \lambda(h_2 - h_1)\}^2} \circ Q(\lambda), \end{aligned}$$

and an application of (A.2) yields the representation (A.3). □

Proof of Lemma 2.2. By a Taylor expansion we have from Lemma A.1 (with $H_1 = m^{-1}$, $H_2 = m_n^{-1}$)

$$m_n(t) - m(t) = \Phi(m_n^{-1}) - \Phi(m^{-1}) \;=\; Q(1) - Q(0) = Q'(\lambda^*)$$

for some $\lambda^* \in [0, 1]$ (see Serfling 1980), where

$$Q'(\lambda^*) = -\frac{m_n^{-1} - m^{-1}}{(m^{-1} + \lambda^*(m_n^{-1} - m^{-1}))'} \circ (m^{-1} + \lambda^*(m_n^{-1} - m^{-1}))^{-1}(t). \tag{A.4}$$

Note that $(m^{-1} + \lambda^*(m_n^{-1} - m^{-1})) \to m^{-1}$ by Lemma 2.1 and that for $t_n = (m^{-1} + \lambda^*(m_n^{-1} - m^{-1}))^{-1}(t)$ we have $t_n \to m(t)$. For the numerator in (A.4) we obtain

$$(m_n^{-1} - m^{-1})(t_n) - (m_n^{-1} - m^{-1})(m(t)) = (m_n^{-1} - m^{-1})'(\eta_n) \cdot (t_n - m(t)) \tag{A.5}$$

for some η_n with $\eta_n - m(t) \leqslant t_n - m(t)$. For the first factor in (A.5) we have by a standard argument

$$(m_n^{-1} - m^{-1})'(\eta_n) = \int_0^1 K_d\left(\frac{m(x) - \eta_n}{h_d}\right)\frac{\mathrm{d}x}{h_d} - (m^{-1})'(\eta_n) + O\left(\frac{1}{nh_d}\right) = O(h_d^2) + O\left(\frac{1}{nh_d}\right),$$

and as a consequence it follows from (A.4) and (A.5) that

$$Q'(\lambda^*) = -\frac{(m_n^{-1} - m^{-1}) \circ m(t)}{(m^{-1})'(m(t))} + o(h_d^2) + o\left(\frac{1}{nh_d}\right).$$

The assertion of Lemma 2.2 is now obtained from Lemma 2.1 and (A.4) – note that $(m^{-1})''(m(t)) = -m''(t)/\{m'(t)\}^3$. □

Proof of Theorem 3.1. We only prove the first part of the theorem; the second assertion follows by similar arguments. We use the decomposition

$$\hat{m}_I^{-1}(t) = \frac{1}{nh_d}\int_{-\infty}^{t}\sum_{i=1}^{n} K_d\left(\frac{\hat{m}(i/n) - u}{h_d}\right)\mathrm{d}u = m_n^{-1}(t) + \Delta_n(t), \tag{A.6}$$

where m_n^{-1} is defined in (2.3) and Δ_n is given by

$$\Delta_n(t) = \frac{1}{nh_d}\sum_{i=1}^{n}\int_{-\infty}^{t}\left\{K_d\left(\frac{\hat{m}(i/n) - u}{h_d}\right) - K_d\left(\frac{m(i/n) - u}{h_d}\right)\right\}\mathrm{d}u. \tag{A.7}$$

For the latter term it follows that

$$\Delta_n(t) = \Delta_n^{(1)}(t) + \frac{1}{2}\Delta_n^{(2)}(t), \tag{A.8}$$

where

$$\Delta_n^{(1)}(t) = \frac{1}{nh_d^2}\sum_{i=1}^{n}\int_{-\infty}^{t} K_d'\left(\frac{m(i/n) - u}{h_d}\right)\left\{\hat{m}\left(\frac{i}{n}\right) - m\left(\frac{i}{n}\right)\right\}\mathrm{d}u,$$

$$\Delta_n^{(2)}(t) = \frac{1}{nh_d^3}\sum_{i=1}^{n}\int_{-\infty}^{t} K_d''\left(\frac{\xi_i - u}{h_d}\right)\left\{\hat{m}\left(\frac{i}{n}\right) - m\left(\frac{i}{n}\right)\right\}^2\mathrm{d}u,$$

with $|\xi_i - m(i/n)| < |\hat{m}(i/n) - m(i/n)|$ $(i = 1, \ldots, n)$. A straightforward calculation shows that

$$\Delta_n^{(2)}(t) = \frac{1}{h_d^2}\left|\frac{1}{n}\sum_{i=1}^{n} K_d'\left(\frac{\xi_i - t}{h_d}\right)\left\{\hat{m}\left(\frac{i}{n}\right) - m\left(\frac{i}{n}\right)\right\}^2\right|$$

$$= \frac{1}{h_d^2}\left|\int_0^1 K_d'\left(\frac{m(x)-t}{h_d}\right)\{\hat{m}(x) - m(x)\}^2\,\mathrm{d}x\right| \cdot (1 + o_p(1)).$$

If we assume that the kernel K_r has been appropriately modified near the boundaries (see Müller 1985) it follows that this term is of order $O(\{1/nh_r + h_r^4\}/h_d)$. This implies that

$$\sqrt{nh_d}\,\Delta_n^{(2)}(t) = o_p(1), \tag{A.9}$$

and a combination of (A.6), (A.8) and (A.9) shows that the assertion of Theorem 3.1 can be proved, establishing the weak convergence

$$\sqrt{nh_d}\left(\Delta_n^{(1)}(t) + \kappa_2(K_r)h_r^2\left(\frac{m''f + 2m'f'}{fm'}\right)(m^{-1}(t))\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\, r^2(t)). \tag{A.10}$$

For this we use the decomposition

$$\Delta_n^{(1)}(t) = \left(\Delta_n^{(1.1)}(t) + \Delta_n^{(1.2)}(t)\right)(1 + o_p(1))$$

with

$$\Delta_n^{(1.1)}(t) = \frac{-1}{n^2 h_d h_r}\sum_{i,j=1}^{n} K_d\left(\frac{m(i/n)-t}{h_d}\right)K_r\left(\frac{X_j - i/n}{h_r}\right)\frac{m(X_j) - m(i/n)}{f(i/n)}$$

$$\Delta_n^{(1.2)}(t) = \frac{-1}{n^2 h_d h_r}\sum_{i,j=1}^{n} K_d\left(\frac{m(i/n)-t}{h_d}\right)K_r\left(\frac{X_j - i/n}{h_r}\right)\sigma(X_j)\frac{\varepsilon_j}{f(i/n)}.$$

For the first term we obtain

$$E[\Delta_n^{(1.1)}(t)] = -\frac{1 + o(1)}{h_r h_d}\int_0^1\int_0^1 K_d\left(\frac{m(x)-t}{h_d}\right)K_r\left(\frac{y-x}{h_r}\right)f(y)\frac{m(y)-m(x)}{f(x)}\,\mathrm{d}y\,\mathrm{d}x$$

$$= -h_r^2\kappa_2(K_r)\int_0^1 \frac{1}{h_d}K_d\left(\frac{m(x)-t}{h_d}\right)\left\{m''(x) + \frac{2m'(x)f'(x)}{f(x)}\right\}\mathrm{d}x \cdot (1 + o(1)) \tag{A.11}$$

$$= -h_r^2\kappa_2(K_r)\left(\frac{m''f + 2m'f'}{fm'}\right)(m^{-1}(t)) \cdot (1 + o(1)),$$

while the variance of $\Delta_n^{(1.1)}(t)$ is given by

$$\begin{aligned}\operatorname{var}\bigl(\Delta_n^{(1.1)}(t)\bigr) &= \frac{1}{n^3 h_d^2 h_r^2} \operatorname{var}\left(\sum_{i=1}^n K_d\left(\frac{m(i/n)-t}{h_d}\right) K_r\left(\frac{X_j - i/n}{h_r}\right) \frac{m(X_j)-m(i/n)}{f(i/n)}\right) \\ &\leqslant \frac{1}{nh_d^2 h_r^2} \mathrm{E}\left[\left(\int_0^1 K_d\left(\frac{m(x)-t}{h_d}\right) K_r\left(\frac{X_j - x}{h_r}\right) \frac{m(X_j)-m(x)}{f(x)}\,\mathrm{d}x\right)^2\right](1+o(1)) \\ &= o\left(\frac{1}{nh_d}\right).\end{aligned}$$

This implies, using assumption (3.1), that

$$\Delta_n^{(1.1)}(t) + h_r^2 \kappa_2(K_r)\left(\frac{m''f + 2m'f'}{fm'}\right)(m^{-1}(t)) = o_p\left(\frac{1}{\sqrt{nh_d}}\right),$$

and consequently (A.10) follows from

$$\sqrt{nh_d}\,\Delta_n^{(1.2)}(t) \xRightarrow{\mathcal{D}} \mathcal{N}(0,\, r^2(t)). \tag{A.12}$$

For a proof of this relation we note that $\mathrm{E}[\Delta_n^{(1.2)}(t)] = 0$ and calculate the variance

$$\begin{aligned}\operatorname{var}(\sqrt{nh_d}\,\Delta_n^{(1.2)}(t)) &= \frac{1}{n^3 h_d h_r^2} \sum_{j=1}^n \operatorname{var}\left(\sum_{i=1}^n \frac{\sigma(X_j)\varepsilon_j}{f(i/n)} K_d\left(\frac{m(i/n)-t}{h_d}\right) K_r\left(\frac{X_j - i/n}{h_r}\right)\right) \\ &= \frac{1}{h_d h_r^2}\int_0^1 \sigma^2(x)\left[\int_0^1 K_d\left(\frac{m(y)-t}{h_d}\right) K_r\left(\frac{x-y}{h_r}\right)\frac{dy}{f(y)}\right]^2 f(x)\mathrm{d}x \cdot (1+o(1)) \\ &= \frac{1}{h_d h_r^2}\int_0^1 K_d\left(\frac{m(z)-t}{h_d}\right)\int_0^1 K_d\left(\frac{m(y)-t}{h_d}\right)\frac{1}{f(y)f(z)} \\ &\quad \times \int_0^1 \sigma^2(x) K_r\left(\frac{x-y}{h_r}\right) K_r\left(\frac{x-z}{h_r}\right) f(x)\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z \cdot (1+o(1)) \\ &= \frac{\sigma^2(m^{-1}(t))}{m'(m^{-1}(t))^2 f(m^{-1}(t))}\frac{h_d}{h_r}\int\int\int K_d(w) K_d(v) K_r(u) \\ &\quad \times K_r\left(\frac{m^{-1}(t+h_d v) - m^{-1}(t+h_d w)}{h_r} + u\right)\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}w \cdot (1+o(1)) \\ &= \frac{\sigma^2(m^{-1}(t))}{m'(m^{-1}(t)) f(m^{-1}(t))}\int\int\int K_d\left(w + \frac{h_r}{h_d} m'(m^{-1}(t))(v-u)\right) \\ &\quad \times K_d(w) K_r(u) K_r(v)\mathrm{d}w\,\mathrm{d}u\,\mathrm{d}v \cdot (1+o(1)),\end{aligned} \tag{A.13}$$

where we have applied the substitution $v \to \{m(m^{-1}(t + h_d w) + h_r(v-u)) - t\}/h_d$ and the last identity uses the relation

$$\lim_{\substack{h_r\to 0, h_d\to 0\\ h_r/h_d\to c}} K_d\left(\frac{m(m^{-1}(t+h_d w)+h_r(v-u))-t}{h_d}\right) = K_d(w+cm'(m^{-1}(t))(v-u)).$$

This proves the representation of the asymptotic variance in (3.3). For a proof of the asymptotic normality we calculate by similar arguments

$$\sum_{j=1}^{n} \mathrm{E}\left[\left\{\frac{\sigma(X_j)}{n^{3/2}h_d^{1/2}h_r}\varepsilon_j \sum_{i=1}^{n} K_d\left(\frac{m(i/n)-t}{h_d}\right)K_r\left(\frac{X_j-i/n}{h_r}\right)\frac{1}{f(i/n)}\right\}^4\right]$$

$$= \frac{\mathrm{E}[\varepsilon_1^4]}{nh_d^2h_r^4}\int\left\{\prod_{j=1}^{4}\int K_d\left(\frac{m(x_j)-t}{h_d}\right)K_r\left(\frac{x-x_j}{h_r}\right)\frac{\mathrm{d}x_j}{f(x_j)}\right\}\sigma^4(x)\mathrm{d}x\cdot(1+o(1))$$

$$= \frac{\sigma^4(m^{-1}(t))\mathrm{E}[\varepsilon_1^4]}{nh_d}\frac{(m^{-1})'(t)}{\{f(m^{-1}(t))\}^4}\int\int\left\{\prod_{j=2}^{4}\int K_d\left(\tilde{x}+\frac{h_r}{h_d}m'(m^{-1}(t))(y_j-y_1)\right)K_r(y_j)dy_j\right\}$$

$$\times K_d(\tilde{x})K_r(y_1)\mathrm{d}y_1\,\mathrm{d}\tilde{x}\cdot(1+o(1))$$

$$= O\left(\frac{1}{nh_d}\right) = o(1),$$

and the asymptotic normality in (A.12) follows from the central limit theorem of Lyapunov. □

Proof of Theorem 3.2. We only prove the first part of the theorem; the second assertion follows by exactly the same arguments. From Lemma A.1 we obtain the Taylor expansion

$$H_2^{-1}(t) - H_1^{-1}(t) = Q(1) - Q(0) = Q'(0) + \frac{1}{2}Q''(\lambda^*)$$

for some $\lambda^* \in [0, 1]$ (see Serfling 1980), which will now be applied for the functions $H_2 = \hat{m}_I^{-1}$, $H_1 = m_n^{-1}$. This gives for the estimator $\hat{m}_I$ and the quantity m_n at the point t the representation

$$\hat{m}_I(t) - m_n(t) = A_n + \frac{1}{2}B_n, \tag{A.14}$$

where

$$A_n = -\frac{\hat{m}_I^{-1} - m_n^{-1}}{(\mathrm{m}_n^{-1})'} \circ m_n(t),$$

$$B_n = \frac{2(\hat{m}_I^{-1} - m_n^{-1})(\hat{m}_I^{-1} - m_n^{-1})'}{\{(m_n^{-1} + \lambda^*(\hat{m}_I^{-1} - m_n^{-1}))'\}^2} \circ \left(m_n^{-1} + \lambda^*(\hat{m}_I^{-1} - m_n^{-1})\right)^{-1}(t)$$

$$- \frac{(\hat{m}_I^{-1} - m_n^{-1})^2(m_n^{-1} + \lambda^*(\hat{m}_I^{-1} - m_n^{-1}))''}{\{(\hat{m}_I^{-1} + \lambda^*(\hat{m}_I^{-1} - m_n^{-1}))'\}^3} \circ (m_n^{-1} + \lambda^*(\hat{m}_I^{-1} - m_n^{-1}))^{-1}(t).$$

At the end of this proof we will show the estimates

$$A_n = -\frac{\hat{m}_I^{-1} - m_n^{-1}}{(m^{-1})'} \circ m(t) + o_p\left(\frac{1}{\sqrt{nh_d}}\right), \tag{A.15}$$

$$B_n = o_p\left(\frac{1}{\sqrt{nh_d}}\right), \tag{A.16}$$

then the first assertion of Theorem 3.2 can be obtained as follows. From (A.15), (A.16) and (A.14) we have

$$\begin{aligned}
&\sqrt{nh_d}\left(\hat{m}_I(t) - m_n(t) - \kappa_2(K_r)h_r^2\left(\frac{m''f + 2m'f'}{f}\right)(t)\right) \\
&\quad = -\sqrt{nh_d}\frac{(\hat{m}_I^{-1} - m_n^{-1}) \circ m(t) + \kappa_2(K_r)h_r^2((m''f + 2m'f')/f)(t) \cdot (m^{-1})' \circ m(t)}{(m^{-1})' \circ m(t)} + o_p(1) \\
&\quad = -m'(t)\sqrt{nh_d}\left\{(\hat{m}_I^{-1} - m_n^{-1}) \circ m(t) + \kappa_2(K_r)h_r^2\left(\frac{m''f + 2m'f'}{m'f}\right)(t)\right\} + o_p(1) \\
&\quad \stackrel{\mathcal{D}}{\Longrightarrow} \mathcal{N}(0, s^2(t)),
\end{aligned}$$

where $s^2(t)$ is defined in (3.4) and we have used the first part of Theorem 3.1 in the last step.

For a proof of the estimate (A.15) we consider the difference

$$\begin{aligned}
D_n &= (\hat{m}_I^{-1} - m_n^{-1}) \circ m_n(t) - (\hat{m}_I^{-1} - m_n^{-1}) \circ m(t) \\
&= (\hat{m}_I^{-1} - m_n^{-1})'(\xi_n)(m_n(t) - m(t)),
\end{aligned} \tag{A.17}$$

where $|\xi_n - m(t)| \leqslant |m_n(t) - m(t)|$. The first factor can be estimated as follows (recall the definition of Δ_n in (A.7)):

$$\begin{aligned}
\Delta_n'(\xi_n) = (\hat{m}_I^{-1} - m_n^{-1})'(\xi_n) &= \frac{1}{nh_d}\sum_{i=1}^{n}\left\{K_d\left(\frac{\hat{m}(i/n) - \xi_n}{h_d}\right) - K_d\left(\frac{m(i/n) - \xi_n}{h_d}\right)\right\} \\
&= \frac{1}{nh_d^2}\sum_{i=1}^{n} K_d'\left(\frac{\eta_{i,n} - \xi_n}{h_d}\right)\left\{\hat{m}\left(\frac{i}{n}\right) - m\left(\frac{i}{n}\right)\right\},
\end{aligned}$$

where $|\eta_{i,n} - m(i/n)| \leqslant |\hat{m}(i/n) - m(i/n)| = O(R_n)$ almost surely, with

$$R_n = \left(\frac{\log h_r^{-1}}{nh_r}\right)^{1/2};$$

see Mack and Silverman (1982, Theorem B). This yields

$$\Delta_n'(\xi_n) = \frac{1}{nh_d^2}\sum_{i=1}^{n} K_d'\left(\frac{m(i/n)-\xi_n}{h_d}\right)\left\{\hat{m}\left(\frac{i}{n}\right) - m\left(\frac{i}{n}\right)\right\} + O\left(\frac{R_n^2}{h_d^3}\right) \quad \text{a.s.}$$

$$= \frac{1}{h_d^2}\int K_d'\left(\frac{m(x)-m(t)}{h_d}\right)\{\hat{m}(x)-m(x)\}\mathrm{d}x + O\left(R_n + \frac{R_n^2}{h_d^3} + \frac{1}{nh_d}\right) \quad \text{a.s.}$$

$$= O\left(\frac{R_n}{h_d} + \frac{R_n^2}{h_d^3} + \frac{1}{nh_d}\right) \quad \text{a.s.}$$

As a consequence, we obtain from (A.17) and Lemma 2.2,

$$D_n = O\left(R_n h_d + \frac{R_n^2}{h_d} + \frac{h_d}{n}\right) = o\left(\frac{1}{\sqrt{nh_d}}\right) \quad \text{a.s.}$$

The estimate (A.15) now follows from the fact that $(m_n^{-1})'(t) = (m^{-1})'(t) + o(1)$; (see the proof of Lemma 2.1).

The second estimate (A.16) is proved similarly and we only indicate the main steps. First, we decompose $B_n = 2B_{n1} - B_{n2}$, where

$$B_{n1} = \frac{(\hat{m}_I^{-1} - m_n^{-1})(\hat{m}_I^{-1} - m_n^{-1})'(t_n)}{\{m_n^{-1} + \lambda^*(\hat{m}_I^{-1} - m_n^{-1})\}^2(t_n)},$$

$$B_{n2} = \frac{(\hat{m}_I^{-1} - m_n^{-1})^2(m_n^{-1} + \lambda^*(\hat{m}_I^{-1} - m_n^{-1}))''(t_n)}{\{\hat{m}_I^{-1} + \lambda^*(\hat{m}_I^{-1} - m_n^{-1})\}^3(t_n)}$$

and $t_n = (m_n^{-1} + \lambda^*(\hat{m}_I^{-1} - m_n^{-1}))^{-1}(t)$. Note that $t_n m(t)$, $(m_n^{-1} + \lambda^*(\hat{m}_I^{-1} - m_n^{-1})) \xrightarrow{P} m^{-1}$. In view of Theorem 3.1, we therefore obtain from (3.2),

$$B_{n1} = O_p\left(\frac{1}{\sqrt{nh_d}} \cdot \left(\frac{R_n}{h_d} + \frac{R_n^2}{h_d^3}\right)\right) = o_p\left(\frac{1}{\sqrt{nh_d}}\right),$$

$$B_{n2} = O_p\left(\frac{1}{nh_d}\right),$$

which proves (A.16). □

Acknowledgements

The authors are grateful to Isolde Gottschlich who typed numerous versions of this paper with considerable technical expertise and to E. Mammen and W. Polonik for useful discussions and some help with the references. The work of the authors was supported by the Sonderforschungsbereich 475, Komplexitätsreduktion in multivariaten Datenstrukturen. The authors are also grateful to three unknown referees and an associate editor for their constructive comments on an earlier version of this manuscript.

References

Brunk, H.D. (1955) Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.*, **26**, 607–616.

Bennett, C. and Sharpley, R. (1988) *Interpolation of Operators*, Pure Appl. Math. 129. Boston: Academic Press.

Cheng, K.F. and Lin, P.E. (1981) Nonparametric estimation of a regression function. *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, **57**, 223–233.

Delecroix, M. and Thomas-Agnan, C. (2000) Spline and kernel regression under shape restrictions. In M.G. Schimek (ed.), *Smoothing and Regression. Approaches, Computation and Application*. New York: Wiley.

Dette, H. and Pilz, K.F. (2004) A comparative study of monotone nonparametric kernel estimates. Technical report. http://www.ruhr-uni-bochum.de/mathematik3/preprint.htm

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

Friedman, J. and Tibshirani, R. (1984) The monotone smoothing of scatterplots. *Technometrics*, **26**, 243–250.

Gasser, T. and Müller, H.G. (1979) Kernel estimates of regression functions. In T. Gasser and M. Rosenblatt (eds), *Smoothing Techniques for Curve Estimation*, Lecture Notes in Math. 757. Berlin: Springer-Verlag.

Gijbels, I. (2005) Monotone regression. In N. Balakrishnan, S. Kotz, C.B. Read and B. Vadakovic (eds), *The Encyclopedia of Statistical Sciences*, 2nd edition. Hoboken, NJ: Wiley.

Hall, P. and Huang, L.-S. (2001) Nonparametric kernel regression subject to monotonicity constraints. *Ann. Statist.*, **29**, 624–647.

Kelly, C. and Rice, J. (1990) Monotone smoothing with application to dose response curves and the assessment of synergism. *Biometrics*, **46**, 1071–1085.

Mack, Y.P. and Silverman, B.W. (1982) Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, **61**, 405–415.

Mammen, E. (1991) Estimating a smooth monotone regression function. *Ann. Statist.*, **19**, 724–740.

Mammen, E. and Thomas-Agnan, C. (1999) Smoothing splines and shape restrictions. *Scand. J. Statist.*, **26**, 239–252.

Mammen, E., Marron, J.S., Turlach, B.A. and Wand, M.P. (2001) A general projection framework for constrained smoothing. *Statist. Sci.*, **16**, 232–248.

Müller, H.G. (1985) Kernel estimators of zeros and of location and size of extrema of regression functions. *Scand. J. Statist.*, **12**, 221–232.

Mukerjee, H. (1988) Monotone nonparametric regression. *Ann. Statist.*, **16**, 741–750.

Ramsay, J.O. (1988) Monotone regression splines in action (with comments). *Statist. Sci.*, **3**, 425–461.

Ramsay, J.O. (1998) Estimating smooth monotone functions. *J. Roy. Statist. Soc. Ser. B*, **60**, 365–375.

Rice, J. (1984) Bandwidth choice for nonparametric regression. *Ann. Statist.*, **12**, 1215–1230.

Ryff, J.V. (1965) Orbits of L^1-functions under doubly stochastic transformations. *Trans. Amer. Math. Soc.*, **117**, 92–100.

Ryff, J.V. (1970) Measure preserving transformations and rearrangements. *J. Math. Anal. Appl.*, **31**, 449–458.

Serfling, R.J. (1980) *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

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

Wright, F.T. (1981) The asymptotic behavior of monotone regression estimates. *Ann. Statist.*, **9**, 443–448.

Received September 2004 and revised September 2005