The Annals of Statistics
2001, Vol. 29, No. 3, 624–647

NONPARAMETRIC KERNEL REGRESSION SUBJECT TO MONOTONICITY CONSTRAINTS

BY PETER HALL AND LI-SHAN HUANG

Australian National University and CSIRO and Australian National University

We suggest a method for monotonizing general kernel-type estimators, for example local linear estimators and Nadaraya–Watson estimators. Attributes of our approach include the fact that it produces smooth estimates, indeed with the same smoothness as the unconstrained estimate. The method is applicable to a particularly wide range of estimator types, it can be trivially modified to render an estimator strictly monotone and it can be employed after the smoothing step has been implemented. Therefore, an experimenter may use his or her favorite kernel estimator, and their favorite bandwidth selector, to construct the basic nonparametric smoother and then use our technique to render it monotone in a smooth way. Implementation involves only an off-the-shelf programming routine. The method is based on maximizing fidelity to the conventional empirical approach, subject to monotonicity. We adjust the unconstrained estimator by tilting the empirical distribution so as to make the least possible change, in the sense of a distance measure, subject to imposing the constraint of monotonicity.

1. Introduction. We suggest a method for "monotonizing" a general kernel-type estimator of a regression mean. It applies to Gasser–Müller, Nadaraya–Watson, Priestley–Chao and local linear estimators, as well as many of the modified forms of these types that have been proposed [e.g., by Hougaard (1988), Hougaard, Plum and Ribel (1989), Müller and Song (1993), Mammen and Marron (1997), Müller (1997)]. It may be implemented rapidly and without difficulty, using standard software. It involves tilting the empirical distribution by the least possible amount, subject to the constraint being enforced. Its roots lie partly in biased-bootstrap techniques suggested by Hall and Presnell (1999). Monotone estimates are of course required in many practical applications, where physical considerations suggest that a response should be monotone in the dosage or the explanatory variable. These include analysis of dose-response in pharmokinetics, and many specific practical problems mentioned in the literature cited below.

Competing methods include those of Friedman and Tibshirani (1984), based on isotonic regression, and Bloch and Silverman (1997), where the cost of misclassification is minimized. Mammen (1991) discusses, among other matters, the theoretical performance of Friedman and Tibshirani's (1984) technique.

Received March 1999; revised December 2000.

AMS 2000 *subject classifications*. Primary 62G07; secondary 62G20.

Key words and phrases. Bandwidth, biased bootstrap, Gasser–Müller estimator, isotonic regression, local linear estimator, Nadaraya–Watson estimator, order restricted inference, power divergence, Priestley–Chao estimator, weighted bootstrap.

In common with other methods based on projecting an unconstrained estimator onto a constrained subspace, these approaches reduce the smoothness of the estimator with which they started. In the just-mentioned examples, the projected estimator is so unsmooth as to commonly have jump discontinuities. In related techniques the lack of smoothness can be in the form of a discontinuous derivative; then cusps in a graph of the smoother are visible. We sought a method which, by way of contrast, would produce a curve estimator that enjoyed the same level of smoothness (i.e., the same number of derivatives) as its unconstrained counterpart. At the same time it should be applicable to general kernel methods, modifying them principally in regions where they are nonmonotone, and differing in only trivial respects in other places. Furthermore, it should require no more additional computational labor (relative to that needed for the unconstrained estimator) than the constraining step of, say, a spline estimator, and it should use only standard computing routines and software.

Our method has all these features, and in addition is readily able to enforce strict monotonicity, for example by constraining the minimum gradient to be at least a given value $\varepsilon > 0$. Moreover, it employs a smoothing parameter which can be chosen prior to and separately from the monotonizing step. In particular, the experimenter may use a conventional kernel-type method, and his or her favorite bandwidth selection rule, to construct a nonparametric estimator and can then apply our technique to make it monotone in a smooth way. Furthermore, our approach can be applied locally to monotonize a conventional kernel estimator in regions where there are "dips" and "bumps," and then trivially extended monotonically to the entire curve, the resulting estimator being equivalent to that obtained by a global application. We do not discuss optimality issues, but they are of interest. In particular it is possible that one approach or another to monotonizing a regression mean has theoretical performance advantages from a minimax viewpoint.

Ramsay (1988), Kelly and Rice (1990), Turlach (1997) and Mammen and Thomas-Agnan (1999) have considered constrained spline-based methods for constructing monotone nonparametric estimators of regression means. These, and new projection-based techniques suggested by Mammen, Marron, Turlach and Wand (1999a, b) [see also Marron, Turlach and Wand (1997)] are arguably the principal competitors with our approach, in that they produce smooth estimators and are in the same class as familiar estimator families (spline and kernel methods, respectively). The fact that our technique has roots in conventional kernel estimation is one reason why it is appealing to users of those methods; spline smoothing does not hold the same attraction for them.

Alternative monotonizing techniques include those of Mukerjee (1988), who modified maximum likelihood methods of Brunk (1955) to construct a monotone function defined at the response or design variables and then smoothed and interpolated this function using a kernel with special log-concave properties. General methodology for inference under order restrictions, developed by Bartholomew (1959) and Barlow, Bartholomew, Bremner and Brunk (1972), among others, should also be mentioned in this regard.

Ramsay (1998) proposed methods based on solving differential equations that define twice-differentiable monotone functions. Statistical tests for monotonicity of a regression mean include those of Schlee (1982), Bowman, Jones and Gijbels (1998) and Hall and Heckman (1998).

Section 2 will introduce our method and discuss its main features. Numerical and theoretical properties will be addressed in Sections 3 and 4, respectively. Technical arguments will be summarized in Section 5. In Section 4 we shall show that our constrained estimator is well-defined under very general conditions. Moreover, we shall prove that, away from places where the unconstrained estimator is non-monotone, the constrained and unconstrained estimators are very close, generally differing only to second order.

The methods that we suggest have application to a host of related problems, including monotonization of hazard rate estimators [see, e.g., Patil, Wells and Marron (1994) and González-Manteiga, Cao and Marron (1996) for more conventional nonparametric methods in that context], to rendering nonparametric estimators convex [see Fisher, Hall, Turlach and Watson (1997) for examples where nonparametric techniques are important in that setting] and many other areas. Excitingly, they can be used to impose shape constraints along with other conditions, for example, to impose the constraint that a density estimate produce a hazard rate having a certain shape in a given region, and at the same time that the associated smoothed distribution estimate have the same first k moments as the conventional empirical distribution, and/or have a given probability mass in a specified interval.

Throughout we use the terms "increasing" and "decreasing" to mean "nondecreasing" and "nonincreasing," respectively.

2. Methodology. Our method has several different versions, of which we give only one here. It involves applying the weights primarily to the response variables and is chosen because it is generic to a large class of linear estimators. An alternative approach, where the weights are applied to the data pairs, is briefly discussed in the Appendix. It has the advantage that it is fully translation invariant, but since this is achieved at the expense of greater computational labor, and since the simpler, alternative approach is virtually translation invariant in practice, we present only the latter in detail.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ denote a sample of pairs of explanatory and response variables, and write (X, Y) for a general data pair. We wish to estimate the mean response, $g(x) \equiv E(Y|X = x)$, given that the explanatory variable equals x. Conventional linear estimators of $g(x)$ may be expressed in the form

$$\hat{g}(x) = n^{-1} \sum_{i=1}^{n} A_i(x)\, Y_i, \tag{2.1}$$

where the weight functions A_i depend only on the X_i's, not on the Y_i's. For example, in the case of regularly spaced design, without loss of generality on the interval $[0, 1]$, $A_i(x) = K_i(x) \equiv h^{-1}K\{(x - X_i)/h\}$, where the kernel

function K is generally a bounded, symmetric, compactly supported probability density, and h is a bandwidth. For the Nadaraya–Watson estimator, appropriate for irregular design, $A_i = nK_i/(\sum_j K_j)$; for the Priestley–Chao estimator, $A_i = n(X_i - X_{i-1})\,K_i$, where here and in the next example it is assumed that the pairs (X_i, Y_i) have been ordered so that $X_1 \le \cdots \le X_n$; for the Gasser–Müller estimator, $A_i(x) = n\int_{\mathscr{I}(i)} K_i(x-u)\,du$, where $\mathscr{I}(i)$ denotes the interval $[Z_{i-1}, Z_i]$ and $Z_i = (X_i + X_{i+1})/2$; and for the local linear estimator,

$$A_i(x) = \frac{S_2(x) - \{(x - X_i)/h\}\,S_1(x)}{S_0(x)\,S_2(x) - S_1(x)^2}\,K_i(x),$$

where $S_k(x) = n^{-1}\sum_i \{(x - X_i)/h\}^k\,K_i(x)$. See Wand and Jones [(1995), Chapter 5] for discussion. In the cases of Gasser–Müller and Priestley–Chao methods it is common to assume that design points are sampled randomly from a continuous distribution supported on a known interval $[a, b]$, and to take $X_0 = a$ and $X_{n+1} = b$.

We suggest generalizing the definition of $\hat{g}(x)$ at (2.1) to

$$\hat{g}(x|p) = \sum_{i=1}^{n} p_i\,A_i(x)\,Y_i, \tag{2.2}$$

where $p = (p_1, \ldots, p_n)$ is a probability distribution on the set $\{X_1, \ldots, X_n\}$. To impose the condition that $\hat{g}(\cdot|p)$ is increasing on an interval $\mathscr{I}$ we propose choosing $p = \hat{p}$ to minimize the distance $D(p)$ from p to the uniform distribution, $p_{\text{unif}} = (1/n, \ldots, 1/n)$, subject to $\hat{g}'(\cdot|p) \ge 0$ on $\mathscr{I}$. More generally we might ask that $\hat{g}'(\cdot|p) \ge \varepsilon$, for a given positive number ε. Of course, additional constraints, $\sum_i p_i = 1$ and $\min p_i \ge 0$, are required by virtue of the fact that p is a probability measure. The first of these must be imposed explicitly, but the second is often implied by the distance measure and then does not need to be considered explicitly.

Suitable distance measures were introduced by Cressie and Read (1984):

$$D_\rho(p) = \frac{1}{\rho\,(1-\rho)}\left\{n - \sum_{i=1}^{n}(np_i)^\rho\right\}$$

for $-\infty < \rho < \infty$ and $\rho \ne 0, 1$, and

$$D_0(p) = -\sum_{i=1}^{n}\log(np_i), \qquad D_1(p) = \sum_{i=1}^{n} p_i\log(np_i).$$

The latter are both Kullback–Leibler divergences. If $0 \le \rho \le 1$, which is the range that we shall address in detail, then $D_\rho(p)$ is not well defined unless each p_i is nonnegative, or strictly positive in the case $\rho = 0$. Therefore, positivity does not need to be imposed.

Implementation is straightforward, and (in our experience) without vices, using an off-the-shelf quadratic programming routine such as E04UCF in the NAG library. If $\hat{g}(\cdot|p)$ is constrained to be monotone on a grid of N points, then the algorithm for computing p to a given level of accuracy involves

only $O(N \log N)$ iterations, the constant increasing logarithmically as a function of the required accuracy. Of course, quadratic programming is routinely used to construct other constrained nonparametric smoothers, such as those based on splines; see for example Turlach (1997). Monotonicity of a function may be verified to hold, or to fail, in numerical terms by checking the signs of differences between function values at adjacent grid points.

Our algorithm need only be applied on those intervals (and a little beyond) where the basic estimator $\hat{g}(\cdot \mid p_{\text{unif}})$ is not monotone; it does not have to be implemented on the entire interval of estimation. During this step the other p_i's may be taken identical to one another, at a value determined by the constraint $\sum_i p_i = 1$. This is a consequence of the following properties:

$$\begin{aligned} &\text{if } \hat{g}'(\cdot \mid p) > 0 \text{ on an interval } \mathscr{I}, \text{ then the sign of } \hat{g}'(\cdot \mid p) \geq 0 \\ &\text{on } \mathscr{I} \text{ does not change if all those values of } p_i \text{ for which} \\ &A_i'(x) \neq 0 \text{ (for } x \in \mathscr{I}) \text{ are multiplied by a fixed constant} \end{aligned} \tag{2.3}$$

and

$$\begin{aligned} &\text{if } \mathscr{A} \text{ and } \mathscr{B} \text{ are complementary subsets of the integers } 1, \ldots, n, \\ &\text{and if } p_i, \text{ for } i \in \mathscr{A}, \text{ are fixed, then the values of } p_j, \text{ for } j \in \mathscr{B}, \\ &\text{that minimise } D_\rho(p) \text{ are identical, and are uniquely determined} \\ &\text{by the constraint that their sum should equal } 1 - \sum_{i \in \mathscr{A}} p_i. \end{aligned} \tag{2.4}$$

Result (2.3) follows immediately from the definition of $\hat{g}(\cdot \mid p)$ at (2.2), and (2.4) follows via an application of the calculus of variations to the distance measure D_ρ.

One implication of (2.3) and (2.4) is that, for values of X_i that lie away from dips or bumps in the unconstrained estimator, the corresponding value of p_i equals a constant that is close to $1/n$. As a result, the constrained estimator equals a constant multiple (close to 1) of the unconstrained estimator, in places that are not near to dips or bumps. A characteristic of a graph of an estimator produced by our method is that it corrects for nonmonotonicity in dips or bumps in the unconstrained estimator, and returns quickly to the graph of the unconstrained estimator on either side of those places. These properties will be expanded upon in theoretical work in Section 4. Illustrations of (2.3) and (2.4) in action will be given in Section 3.

3. Numerical illustrations. We present a comparative simulation study (Example 1) and two real-data examples. We originally implemented our method for Nadaraya–Watson estimators without experiencing any difficulty, although local linear estimators are chosen for illustration here. We use the biweight kernel in Example 1, and the Gaussian kernel in Examples 2 and 3.

A more extensive simulation study, comparing qualitative properties of smoothers that result from different simulated datasets, is available from the authors. Among other matters, this work makes it clear that varying ρ usually makes little difference to either the curve estimates or the weights $\hat{p}_i$. In none

of the cases that we considered, involving either simulated data or real data, was there any difficulty carrying out the monotonizing step.

EXAMPLE 1.

$$Y = -X^3 + 3X + \varepsilon, \quad X \sim \text{Uniform}[-1, 1], \quad \varepsilon \sim N(0, 0.4^2).$$

This cubic regression function is increasing on $\mathscr{I} = [-0.9, 0.9]$. When either imposing the constraints or plotting the curves we took 100 equally spaced grid points on $\mathscr{I}$. We used sample size $n = 50$ and employed bandwidth $h = 0.25$, in order that fluctuations might occur despite the underlying function being monotone. Forty out of 250 simulations produced increasing estimates without manipulation. Figure 1(a) shows the simulated dataset corresponding to the largest divergence distance, $D_0(\hat{p}) = 1.02$, among the remaining 210. Also depicted are the true function (dotted line), the original local linear estimate (short-dashed line), and the constrained-to-be monotone estimates when $\rho = 0$ and 1 (unbroken line and long-dashed line, respectively). Estimates for other values of ρ lie between the latter two.

As can be seen, varying ρ makes little difference to the constrained estimates, although $D_1(\hat{p}) = 0.015$. Both constrained curves correct for the small bump near $x = 0$. Though the constrained estimates may not be visually very smooth, this is an artifact of the small bandwidth used; the estimators have as many derivatives as the kernel. The weights $\hat{p}_i$ that yielded $D_0(\hat{p}) = 1.02$ are plotted in Figure 2(a). Values of $\hat{p}_i$ for $\rho = 1$ are connected by dashed lines.

A comparison with other methods for constrained estimation is insightful. We implemented the algorithm suggested by Friedman and Tibshirani [(1984), Section 3.2], obtaining, for the dataset used in Figure 1(a), the estimate depicted in Figure 1(b). [For ease of comparison we have repeated from panel (a) the constrained local linear estimate with $\rho = 0$.] Discontinuities caused by the "pool adjacent violators" algorithm are evident. Monotone spline estimates were kindly provided by Turlach (1997) with the smoothing parameter $\alpha = 10^{-5}$ and 10^{-3}. The spline curves show several smooth steps; this tendency to increase in steps is less marked for our algorithm.

These qualitative differences between the three estimator types, that is, the Friedman and Tibshirani (1984) estimator showing discontinuous jumps, the Turlach (1997) estimator tending to increase in smooth jumps, and our estimator tending to increase more gradually after constraint, are typical of a much larger number of simulations, not given here. It is not really possible to give a quantitative comparison of these effects, however, since the smoothing parameters for the different techniques [particularly that of Turlach (1997) and our own] are not comparable.

Figure 2(b)–(d) gives pointwise squared bias, variance, and mean squared error, respectively, of the constrained estimators and their unconstrained counterpart. (These quantities were computed by averaging over the 250 simulated curve estimates.) The most obvious feature is a marked decrease in bias

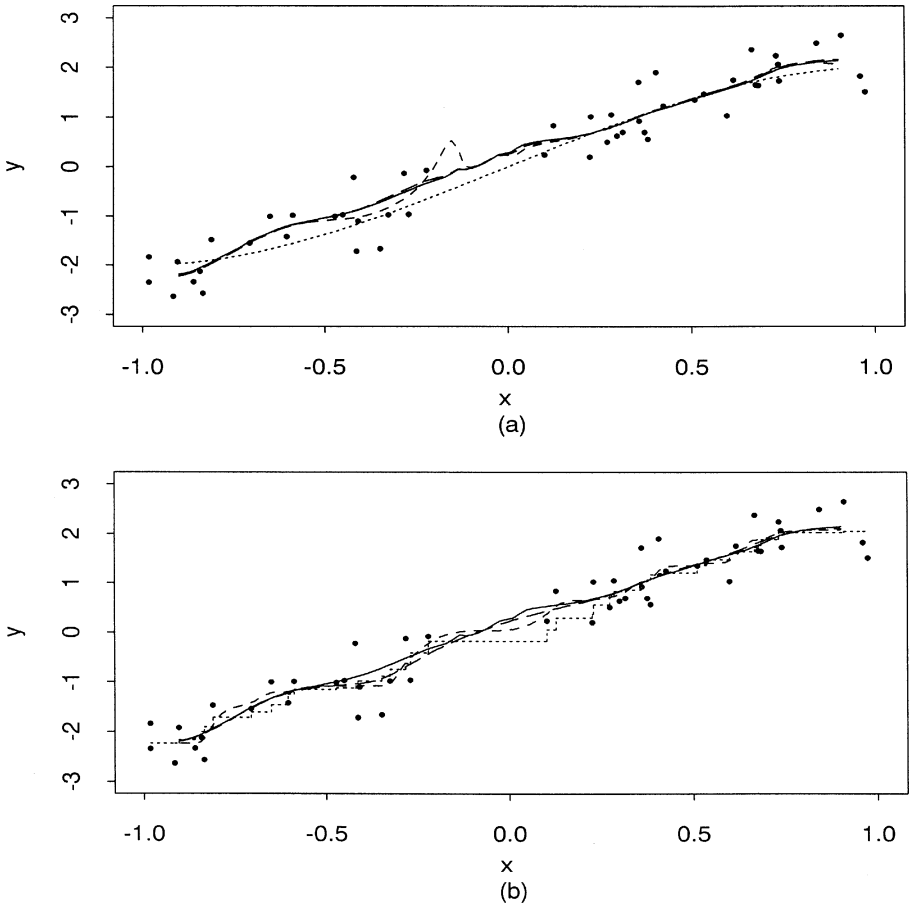

FIG. 1. *Performance of monotonization for data simulated from model in Example* 1. *In panel* (a) *the dotted line represents the true regression function, the dashed line shows the unconstrained local linear estimate, the solid line shows the constrained local linear estimate for* $\rho = 0$ *and the long-dashed line shows the constrained estimate for* $\rho = 1$. *In panel* (b) *the solid line is as in* (a), *the jump function shown by the dotted line represents the estimate proposed by Friedman and Tibshirani* (1984), *and the dashed and long-dashed lines are the spline estimates of Turlach* (1997) *with smoothing parameter* $\alpha = 10^{-5}$ *and* 10^{-3}, *respectively.*

in most parts of $\mathscr{I}$, virtually the same variance for the constrained estimator relative to its unconstrained form, and an overall smaller mean squared error (MSE). The slightly larger variance near the origin results from frequent need to enforce monotonicity there, using relatively large weights. In this respect, among others, the weights shown in Figure 2(a) are typical. The weights employed in this region are not as large when $\rho = 1$. Average mean integrated squared errors of the unconstrained estimator, and the constrained forms with $\rho = 0$ and $\rho = 1$, are 0.089, 0.084 and 0.084, respectively.

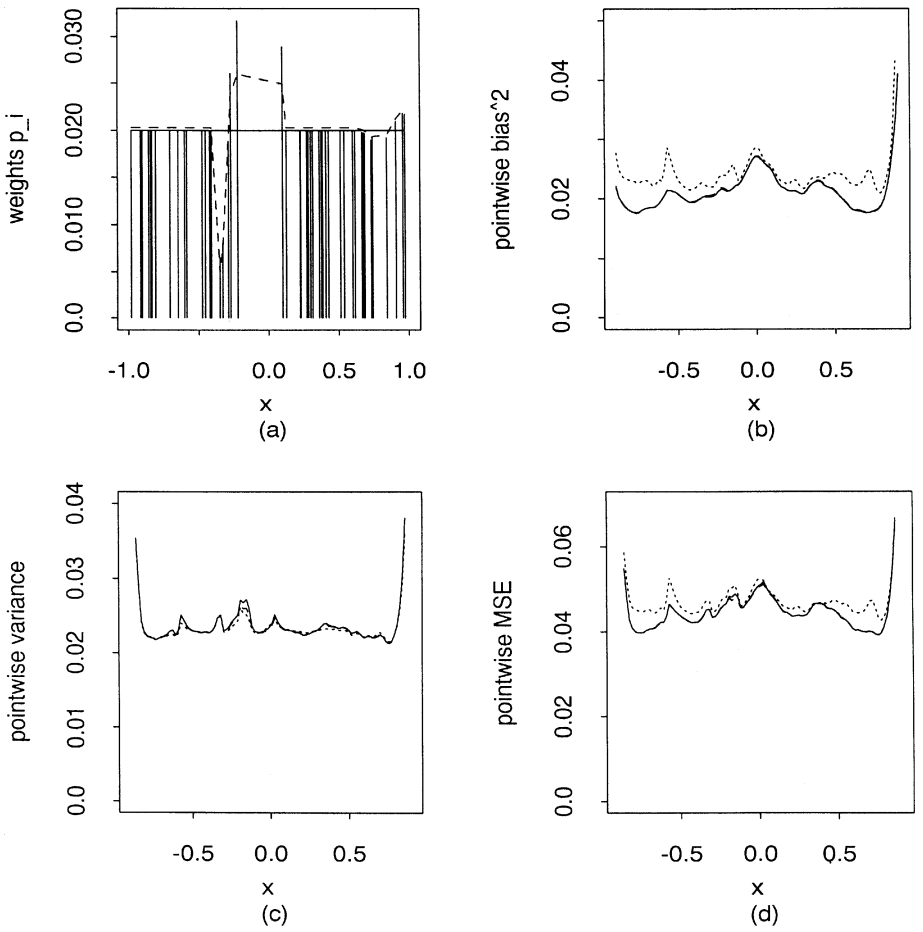

FIG. 2. *Performance of monotonization for data simulated from model in Example* 1. *In panel* (a) *the value of* $\hat{p}_i$, *for the dataset from Figure* 1 *and after the constraint has been achieved with* $\rho = 0$, *is plotted against the data values. The corresponding result for* $\rho = 1$ *is indicated by the dashed line. Panels* (b)–(d) *give pointwise squared bias, variance and mean squared error, respectively. The solid line, dashed line, and dotted line represent, respectively, the constrained estimators with* $\rho = 0$ *and* 1 *and the unconstrained local linear estimator.* (*The cases* $\rho = 0$ *and* 1 *are virtually indistinguishable on the scale of the figure.*)

EXAMPLE 2 (Radiocarbon data). These data were published by Pearson and Qua (1993), and a subset analyzed by Bowman and Azzalini (1997). We use the same subset, and bandwidth $h = 30$ as suggested by Bowman and Azzalini (1997). The variables are radiocarbon age, predicted from the radiocarbon dating process, and calendar age, that is, the true calendar age. In this example we also tried constraining the minimum gradient to be at least $\varepsilon = 0.25$.

Figure 3 shows the data, the local linear estimate with Gaussian kernel and the estimates under the "increasing" constraints when $\varepsilon = 0$. The curve

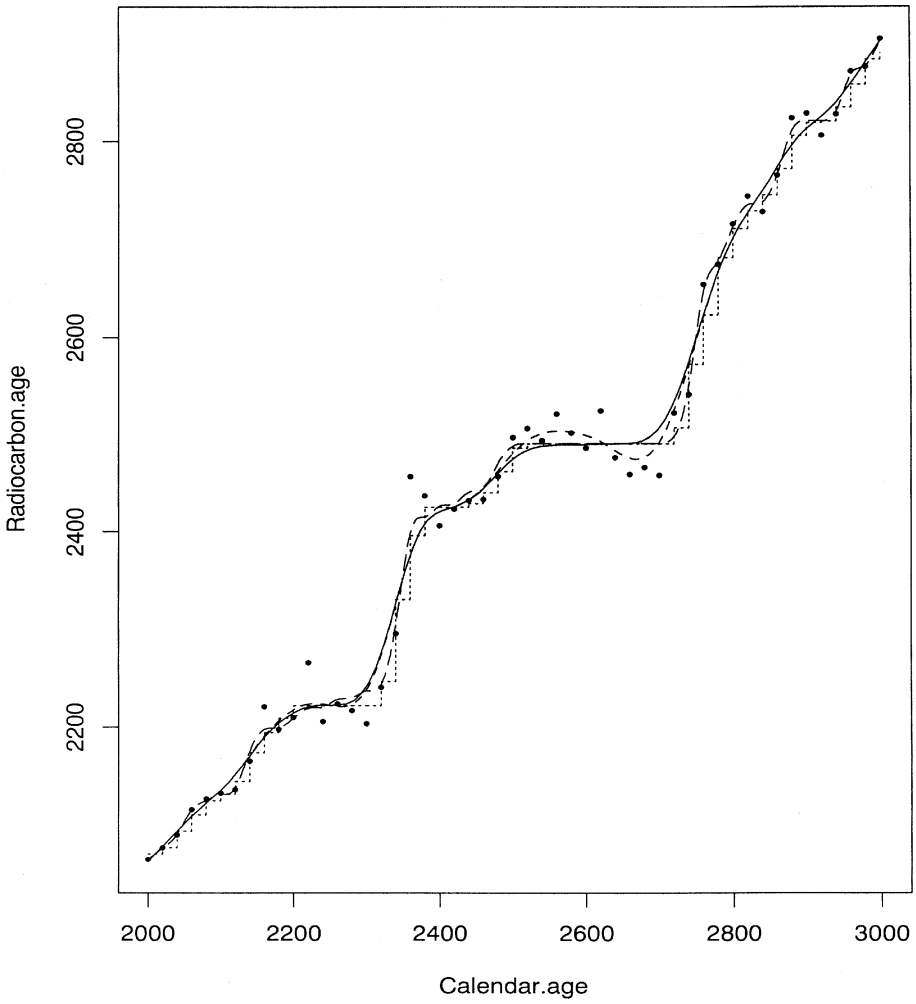

FIG. 3. *Radiocarbon data. The figure shows the unconstrained local linear estimator* (*dashed line*), *its monotonically constrained form with* $\varepsilon = 0$ (*solid line*), *the jump function* (*dotted line*) *of Friedman and Tibshirani* (1984), *and the spline estimate* (*long-dashed line*) *of Turlach* (1997) *with* $\alpha = 100$.

with $\varepsilon = 0.25$ is barely distinguishable on the scale of Figure 3(a). Curves obtained using larger values of ε depart more markedly from the scatterplot and seem not to be such satisfactory estimates. The divergence corresponding to $\varepsilon = 0$ is $D_0(\hat{p}) = 0.00029$.

Estimates of Friedman and Tibshirani (1984) and Turlach (1997), with $\alpha = 100$, are also included. They show the same qualitative features revealed in Figure 1, with the Friedman and Tibshirani (1984) estimator having marked jumps and the Turlach (1997) estimator being smooth but less gradually increasing than our own.

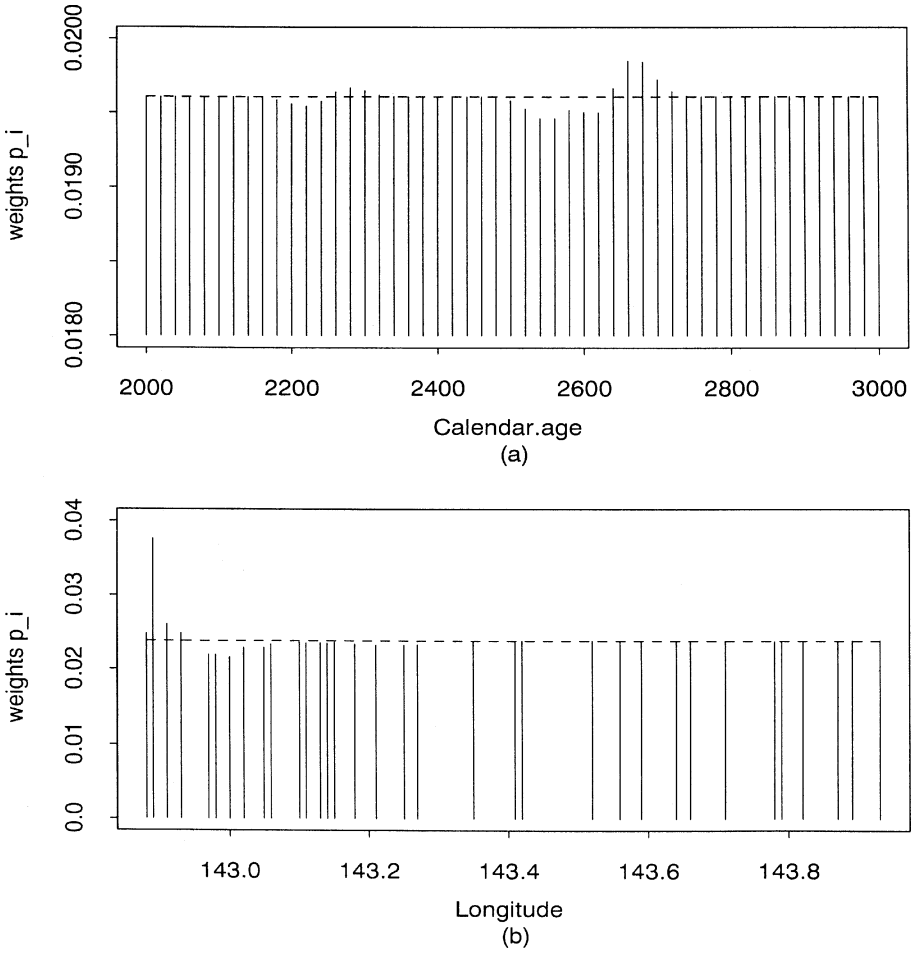

FIG. 4. *Weights $\hat{p}_i$ for real datasets. Panel* (a) *shows the weights $\hat{p}_i$ for the radiocarbon data, plotted against values of X_i, in the case $\varepsilon = 0$. Panel* (b) *presents the same information for the Great Barrier Reef data.*

The weights $\hat{p}_i$ when $\varepsilon = 0$ are shown in Figure 4(a). The plot there has been truncated so that the differences between the $\hat{p}_i$'s and the uniform weights is more clearly visible. It is clear that $\hat{p}_i$'s that correspond to X_i's that are not in the immediate vicinity of dips or bumps in the unconstrained estimate, are constant and very close to n^{-1}. This was predicted by arguments in Section 2 [see in particular (2.3) and (2.4)], and will be corroborated by theoretical results in Section 4; see Theorem 4.3. It explains why the constrained and unconstrained estimates are so close in such places.

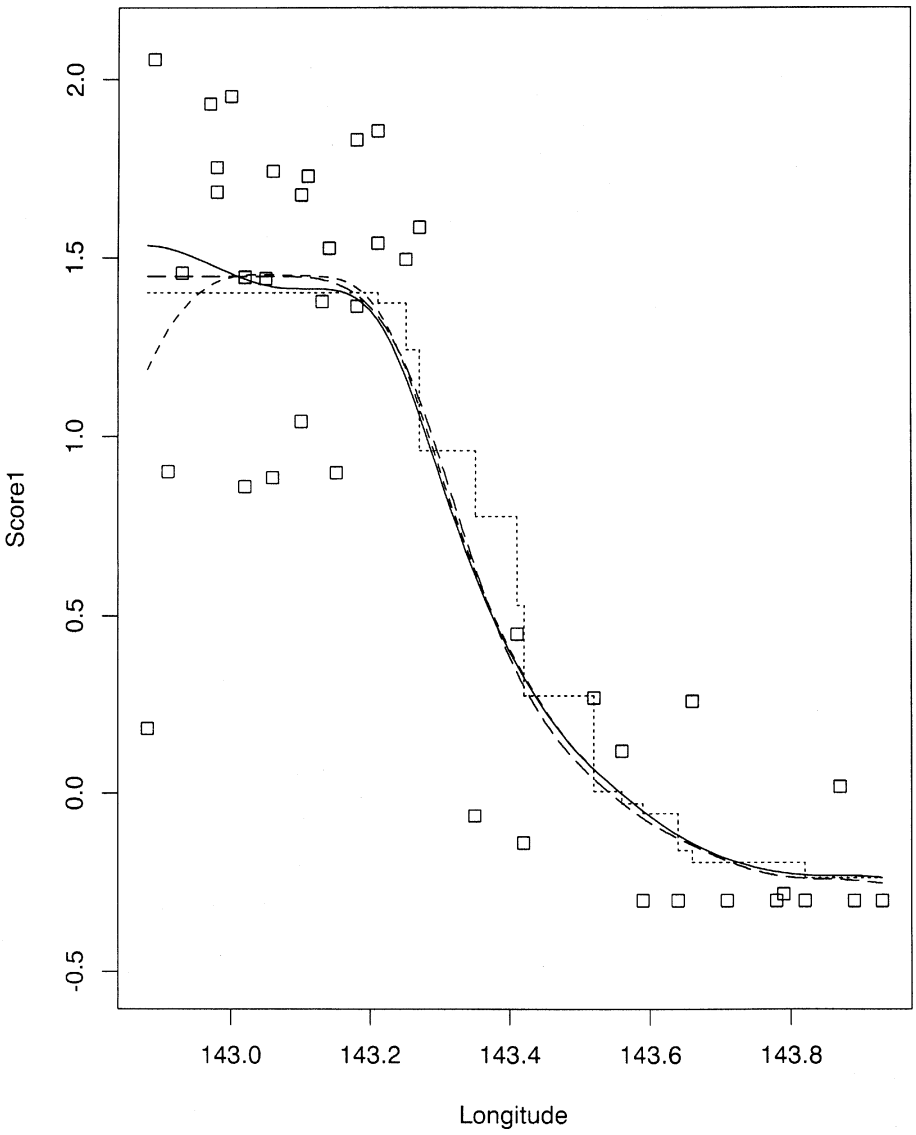

FIG. 5. *Great Barrier Reef data. Panel descriptions and legends are as for Figure* 3, *except that now the constraint was that the curve be monotone decreasing and* $\alpha = 0.001$ *for the spline estimate.*

EXAMPLE 3 (Great Barrier Reef data). These data derive from a survey of fauna on the seabed between the coast of northern Queensland and the Great Barrier Reef. They were analyzed by Poiner et al. (1997) and Bowman and Azzalini (1997). We used the data subset suggested by Bowman and Azzalini [(1997), Figure 5.4]. The variables are longitude and catch score 1. We employed the bandwidth $h = 0.1$ used by Bowman and Azzalini (1997).

Figure 5 plots the data, the local linear estimate with Gaussian kernel and uniform weights, the estimate under the "decreasing" constraint with $\varepsilon = 0$, and the estimates by Friedman and Tibshirani (1984) and Turlach (1997) with $\alpha = 0.001$. The corresponding power divergence distance is $D_0(\hat{p}) = 0.1499$. Taking $\varepsilon = -0.5$ changes the curve slightly, and produces a larger value of $D_\rho(\hat{p})$. Larger negative values of ε produce curves that arguably depart too far from the scatterplot. The weights $\hat{p}_i$ in the case $\varepsilon = 0$ are shown in Figure 4(b). Again it is seen that, away from nonmonotone parts of the unconstrained curve, the $\hat{p}_i$'s are constant and virtually identical to n^{-1}. That explains why constrained and unconstrained estimates in Figure 5 are virtually identical in places away from dips and bumps.

Incidentally, applying translations of up to several thousand to the radiocarbon data, and then estimating g and correcting back to the original location, produces results that are indistinguishable on the scale of Figure 3. Similar results are observed in many other examples, in particular Example 2. Thus, for practical purposes the method proposed in Section 2 is translation equivariant. A fully translation equivariant approach will be noted in the Appendix.

4. Theoretical properties. If the weight functions A_i are approximately bell-shaped; if for each x there is an index i such that $A_i'(x) > 0$, and another index j such that $A_j'(x) < 0$; and if the response variables are all of the one sign then monotonization in either direction is always possible. This is true despite the fact that the unconstrained estimator $\hat{g}(\cdot|p_{\text{unif}})$ might oscillate wildly. Our first theorem formalizes this result under explicit conditions. It requires the weight functions to be monotone increasing only in their far left-hand tails, and imposes those conditions on only some of the weights. It treats the case where we wish to construct a monotone increasing estimator, and the response variables are positive, but by reversing the direction of the axis on which the explanatory variables X_i are plotted, and/or changing the sign of the response variables Y_i, one obtains the more general result.

THEOREM 4.1. *Assume that the set* $\{1, \ldots, n\}$ *contains a sequence* $i_1, \ldots, i_r$ *with the properties*.

(a) *For each* k, *the function* A_{i_k}' *is strictly positive and continuous on* (U_{i_k}, V_{i_k}), *and vanishes on* $(-\infty, U_{i_k}]$, *where the differences* $V_{ik} - U_{i_k}$ *are strictly positive* (*but may be infinite*, *if* $U_{ik} = -\infty$).

(b) *Each* $x \in \mathscr{I} = [a, b]$ *is contained in at least one interval* (U_{i_k}, V_{i_k}).

(c) *For* $1 \le i \le n$, A_i' *is continuous on* $(-\infty, \infty)$.

(d) *Each* $Y_{i_k} > 0$.

Then there exists a probability measure $p = (p_1, \ldots, p_n)$ *such that each* $p_i > 0$ *and* $\hat{g}'(x|p) > 0$ *for all* $x \in \mathscr{I}$.

Next we shall show that under general conditions, not requiring such detailed assumptions about the weight functions or the responses, or indeed monotonicity of g, there exists $\tilde{p}$ such that $\hat{g}(\cdot|\tilde{p})$ is asymptotically linear,

either increasing or decreasing. For simplicity we assume that the regression function g is bounded away from 0—without loss of generality it is positive—although a longer argument, involving an additional regularity condition, can be used to address other cases.

Let $\mathscr{I}$ denote a compact interval. Suppose the estimator, the distribution of (X, Y) and the kernel K satisfy

(4.1) the estimator is of Gasser–Müller, Nadaraya–Watson, Priestley–Chao or local linear type; the data (X_i, Y_i) are generated by the model $Y_i = g(X_i) + \varepsilon_i$, where the ε_i's are independent and identically distributed with zero mean, and are independent of the X_i's, which are either regularly spaced on a compact interval $\mathscr{J} = [c, d]$, or independent random variables coming from a distribution whose density f is continuous and nonvanishing on $\mathscr{J}$; K is a symmetric, compactly supported density with a Hölder-continuous derivative, $\mathscr{I} \subseteq \mathscr{J}$, and, in the Gasser–Müller, Nadaraya–Watson and Priestley–Chao cases, $\mathscr{I} \subseteq [c + \delta, d - \delta]$ for some $\delta > 0$

and

$$\inf_{x \in \mathscr{I}} g(x) \geq B_1 > 0, \qquad E(|\varepsilon_i|^t) \text{ is bounded}, \tag{4.2}$$

where $t > 0$. A longer argument, involving an additional regularity condition, can be used to remove the assumption at (4.2) that g is positive.

THEOREM 4.2. *Assume* (4.1) *and* (4.2), *that* $t > 0$ *is chosen sufficiently large in* (4.2), *and that* $h = h(n) \to 0$ *and* $n^{1-\delta}h^3 \to \infty$ *for some* $\delta > 0$. *Then for either choice of the* $\pm$ *signs*, *taken respectively*, *with probability* 1 *there exists a probability measure* $\tilde{p}_\pm = (\tilde{p}_{\pm,1}, \ldots, \tilde{p}_{\pm,n})$ *on* $\{1, \ldots, n\}$, *and a constant* $B_\pm > 0$, *such that* (a) $\tilde{p}_{\pm,i} > 0$ *for each* i, *and* (b) $\hat{g}'(x|\tilde{p}_\pm) \to \pm B_\pm$ *uniformly in* $x \in \mathscr{I}$.

The value of t in (4.2) depends on the exponent of Hölder continuity of K', among other aspects of the conditions.

In Theorems 4.1 and 4.2 we did not need to specify the distance D, but it is required for our next result, where we establish consistency and rates of convergence in cases where the true regression mean is monotone. Define $\tilde{g} \equiv \hat{g}(\cdot|p_{\text{unif}})$ to be the unconstrained estimator. In Theorem 4.3 we constrain $\hat{g}(\cdot|\hat{p})$ to be monotone increasing and treat the case where bandwidth is chosen in a way that would be "asymptotically optimal" for $\tilde{g}$. That is, we take $h \asymp h^{-1/5}$.

Let us paraphrase Theorem 4.3 before we state it. Part (a) notes that if g is strictly monotone on $\mathscr{I}$, then asymptotically, the estimate $\tilde{g}$ is also strictly monotone. Part (b) shows that if g' has an isolated zero, and if g is locally quadratic there, then the tilted empirical distribution will have nonconstant

weights only within an $O(h)$ neighborhood of the zero point, and the constrained estimate will be equal to its unconstrained counterpart, to first order, outside the neighborhood. Moreover, the ratio of the two estimators will be constant there. Part (c) shows that the same is true if g' has an isolated zero where it is locally cubic, with degenerate quadratic component, except that the neighborhood is now of radius $O(h^{1/2})$ rather than $O(h)$.

THEOREM 4.3. *Assume* (4.1) *and* (4.2), *that* $t > 0$ *is chosen sufficiently large in* (4.2), *that* f *and* g *have two continuous derivatives on* $\mathscr{I} = [a, b]$, *that* $h \asymp n^{-1/5}$ *as* $n \to \infty$, *and that* $D = D_\rho$ *where* $0 \le \rho \le 1$.

(a) *If* $g' > 0$ *on* $\mathscr{I}$ *then with probability* 1, $\hat{p} = p_{\text{unif}}$ *for all sufficiently large* n. *Hence*, $\hat{g}(\cdot|\hat{p}) \equiv \tilde{g}$ *on* $\mathscr{I}$, *for all sufficiently large* n.

(b) *If* $g' > 0$ *except at a single point* $x_0 = a$ *or* b, *where* $g' = 0$ *and* $g'' \neq 0$, *then* $|\hat{g}(\cdot|\hat{p}) - \tilde{g}| = O_p(h^2)$ *uniformly on* $\mathscr{I}$, *and* $\sup_i \hat{p}_i = O_p(n^{-1})$. *Furthermore*, *there exist random variables* $\Delta = \Delta(n)$ *and* $Z_1 = Z_1(n) \ge 0$, *satisfying* $\Delta = O_p(h^{5/2})$ *and* $Z_1 = O_p(1)$, *and such that* $\hat{g}(x|\hat{p}) = (1 + \Delta)\,\tilde{g}(x)$ *uniformly in* $x \in \mathscr{I}$ *such that* $|x - x_0| > Z_1 h$. *The latter property reflects the fact that*, *for a random variable* $Z_2 = Z_2(n) \ge 0$ *satisfying* $Z_2 = O_p(1)$, *we have* $\hat{p}_i = n^{-1}(1 + \Delta)$ *for all indices* i *such that both* $|X_i - x_0| > Z_2 h$ *and* $A_i(x) \neq 0$ *for some* $x \in \mathscr{I}$.

(c) *If* $g' > 0$ *except at a single point* x_0 *interior to* $\mathscr{I}$, *in the neighborhood of which* g *has three continuous derivatives*, *with* $g''(x_0) = 0$ *and* $g'''(x_0) \neq 0$, *then* $|\hat{g}(\cdot|\hat{p}) - \tilde{g}| = O_p(h^{5/4})$ *uniformly on* $\mathscr{I}$. *Furthermore*, *there exist random variables* $\Delta = \Delta(n)$ *and* $Z_1 = Z_1(n) \ge 0$, *satisfying* $\Delta = o_p(h^2)$ *and* $Z_1 = O_p(1)$, *and such that* $\hat{g}(x|\hat{p}) = (1 + \Delta)\,\tilde{g}(x)$ *uniformly in* $x \in \mathscr{I}$ *such that* $|x - x_0| > Z_1 h^{1/2}$. *The latter property reflects the fact that*, *for a random variable* $Z_2 = Z_2(n) \ge 0$ *satisfying* $Z_2 = O_p(1)$, *we have* $\hat{p}_i = n^{-1}(1 + \Delta)$ *for all indices* i *such that both* $|X_i - x_0| > Z_2 h^{1/2}$ *and* $A_i(x) \neq 0$ *for some* $x \in \mathscr{I}$.

In part (b) of the theorem, a longer argument shows that $\Delta = O_p(h^3)$ rather than simply $\Delta = O_p(h^{5/2})$. We believe that the assertion $\Delta = o_p(h^2)$ in part (c) can likewise be strengthened to $\Delta = O_p(h^3)$, but we do not have a proof. On the other hand, it is straightforward to extend the theorem to the case where a finite number of points $x_1, \ldots, x_m$, with the properties assumed of x_0 in parts (b) or (c), is distributed within the interval. In this case the results in part (c) continue to hold, provided the quantities $|x - x_0|$ and $|X_i - x_0|$ there are replaced by $\inf_j |x - x_j|$ and $\inf_j |X_i - x_j|$, respectively.

5. Technical arguments.

5.1. *Proof of Theorem* 4.1. Let $\mathscr{I} = [a, b]$. Without loss of generality, the V_{i_k}'s are arranged in increasing order, and $i_1 < \cdots < i_r$. Let j_1 denote the largest i_k such that $U_{i_k} < a$. Taking $x = a$ in condition (b) we see that j_1 is well-defined and $V_{j_1} > a$. Given j_ℓ, with $a < V_{j_\ell} \le b$, let $j_{\ell+1}$ be the largest i_k such that $U_{i_k} < V_{j_\ell}$. Using condition (b), with $x = V_{j_\ell}$, we see that

$j_{\ell+1}$ exists and $j_{\ell+1} > j_\ell$. Terminate the sequence $j_1, \ldots, j_s$ when $V_{j_s} > b$. Put $\Delta = \min_k(V_{j_k} - U_{j_{k+1}}) > 0$, and let $0 < \varepsilon < \min(V_{j_1} - a, V_{j_s} - b, \Delta)$. In view of properties (a) and (b), the sequence $j_1, \ldots, j_s$ is well-defined and strictly increasing. Take $q_{j_1} = 1$, and note that by definition of j_1, $q_{j_1} A'_{j_1} Y_{j_1} > 0$ on $[a, V_{j_1})$.

Suppose we have constructed a sequence $q_{j_1}, \ldots, q_{j_\ell}$, where $\ell < s$, with the property that $\sum_{k \le \ell} q_{j_k} A'_{j_k} Y_{j_k} > 0$ on $[a, V_{j_\ell} - \varepsilon]$. Since (i) $U_{j_{\ell+1}} < V_{j_\ell} - \varepsilon$ [by virtue of the definition of $j_{\ell+1}$ and our choice of ε], (ii) $A'_{j_{\ell+1}} > 0$ on $(U_{j_{\ell+1}}, V_{j_{\ell+1}})$ [by property (a)], and (iii) $A'_{j_k}(x)$ is bounded away from $-\infty$, uniformly in $k \le \ell$ and in $-\infty < x < \infty$ [by property (c)], then if $q_{j_{\ell+1}}$ is sufficiently large positive we shall have $\sum_{k \le \ell+1} q_{j_k} A'_{j_k} Y_{j_k} > 0$ on $[a, V_{j_{\ell+1}} - \varepsilon]$.

Using induction over ℓ, this proves the existence of integers $j_1 < \cdots < j_s$, and of positive numbers $q_{j_1}, \ldots, q_{j_s}$, such that $\{j_1, \ldots, j_s\} \subseteq \{1, \ldots, n\}$ and $\sum_k q_{j_k} A'_{j_k} Y_{j_k} > 0$ on $[a, V_{j_s} - \varepsilon]$. In view of the definition of ε, the latter interval contains $\mathscr{I}$. Taking $q_i = 0$ if i is not in the sequence $j_1, \ldots, j_s$, and defining $p_i = q_i / (\sum_j q_j)$, we see that $\hat{g}'(\cdot \mid p) > 0$ on $\mathscr{I}$. In view of property (c), we may take the vanishing p_i's to be very small but strictly positive without invalidating this result. This gives a vector p for which $\hat{g}'(\cdot \mid p) > 0$ on $\mathscr{I}$ and each p_i is strictly positive. □

5.2. *Proof of Theorem* 4.2. In the proofs of this theorem and the next we shall suppose that a ridge parameter has been incorporated into the denominators in definitions of A_i and A'_i for Nadaraya–Watson and local linear estimators, so that their moments are all well-defined. The ridge will be taken to equal $n^{-\lambda}$, where $\lambda \ge 3$; note that λ can be taken arbitrarily large without affecting second- or third-order properties of the nonridged estimator. A simple subsidiary argument enables the theorem to be derived without the ridge, once it has been established with the ridge.

The conditions imposed in the theorem are sufficient to ensure the existence of constants $B_2, B_3, B_4 > 0$ such that

$$\sup_{x \in \mathscr{I}} \sum_{i=1}^n |A'_i(x)|^r = O\bigl(n^{(1-B_2)r}\bigr) \quad \text{almost surely for } 2 \le r \le r_0, \tag{5.1}$$

$$\sup_{x \in \mathscr{I}} \left\{ \left| n^{-1} \sum_{i=1}^n A'_i(x) \right| + \left| n^{-1} \sum_{i=1}^n X_i A'_i(x) - 1 \right| \right\} = o(1) \qquad \text{almost surely,} \tag{5.2}$$

$$\sup_{1 \le i \le n} \sup_{x, y \in \mathscr{I}} |x - y|^{-B_3} \, |A'_i(x) - A'_i(y)| = O\bigl(n^{B_4}\bigr) \quad \text{almost surely,} \tag{5.3}$$

where the value of r_0 may be taken arbitrarily large by choosing t in (4.2) sufficiently large. We need only (4.2), (5.1), (5.2) and (5.3).

Without loss of generality, $\mathscr{I} = [0, 1]$. Put $q_i = (1 \pm \frac{1}{2} X_i)/g(X_i)$, $Q = n^{-1} \sum_i q_i$ and $\tilde{p}_{\pm,i} = (nQ)^{-1} q_i$. Then each $\tilde{p}_{\pm,i} > 0$, $\sum_i \tilde{p}_{\pm,i} = 1$ and $\hat{g}'(\cdot \mid \tilde{p}_\pm) = Q^{-1}(\bar{g}_1 + \bar{g}_2)$, where $\bar{g}_1 = n^{-1} \sum_i (1 \pm \frac{1}{2} X_i) A'_i$ and $\bar{g}_2 = n^{-1} \sum_i (1 \pm$

$\frac{1}{2}X_i)\ \{Y_i\, g(X_i)^{-1} - 1\}\, A_i'$. Let E' denote expectation conditional on the explanatory variables X_i. By (5.2), $\sup|\bar{g}_1 - (\pm\frac{1}{2})| \to 0$ almost surely, where here and below, suprema are over $\mathscr{I}$; by (4.2), (5.1) and Rosenthal's inequality, $\sup E'|\bar{g}_2|^{2r} = O(n^{1-2B_2 r})$ almost surely for $1 \le r \le \frac{1}{2}r_0$, and from this result, (5.3) and the Borel–Cantelli lemma, $\sup|\bar{g}_2| \to 0$ almost surely. (The Borel–Cantelli lemma is applied conditional on the explanatory variables and along a sequence $X_1, X_2, \ldots$ that arises with probability 1.) Hence, $\sup|Q\,\hat{g}'(\cdot|\tilde{p}_\pm) - (\pm\frac{1}{2})| \to 0$ almost surely. Since $Q \to E\{(1 \pm \frac{1}{2}X)/g(X)\} > 0$, where the convergence is almost sure, then the theorem is proved. □

5.3. *Proof of Theorem* 4.3. *Part* (a). Under the conditions of part (a), $\tilde{g}' = g' + o(1)$ uniformly on $\mathscr{I}$, with probability 1. Therefore, with probability 1, for all sufficiently large n, $\tilde{g}'$ satisfies the monotonicity constraint throughout the interval $\mathscr{I}$, without manipulation. Result (a) is immediate.

Part (b). Without loss of generality, x_0 is the left-hand endpoint of $\mathscr{I} = [a, b]$, in which case $g' > 0$ on $(a, b]$. Recall that $\tilde{g} = \hat{g}(\cdot|p_{\text{unif}})$. The proof for part (b) is divided into six steps.

Step (i) [Size of $D(\hat{p})$]. In Step (iv) below we shall show that for each $\delta > 0$ there exists a multinomial distribution $\tilde{p} = \tilde{p}(\delta)$ satisfying

$$P\{\hat{g}'(x|\tilde{p}) > 0 \text{ for all } x \in \mathscr{I}\} > 1 - \delta \tag{5.4}$$

for all sufficiently large n and

$$D(\tilde{p}) = O_p(1). \tag{5.5}$$

Since $p = \hat{p}$ denotes the multinomial distribution that minimizes $D(p)$ subject to nonnegativity of $\hat{g}'(\cdot|p)$ on $\mathscr{I}$, then (5.4) implies that for all sufficiently large n,

$$P\{D(\hat{p}) \le D(\tilde{p})\} > 1 - \delta.$$

This result and (5.5) imply that

$$D(\hat{p}) = O_p(1). \tag{5.6}$$

It may be proved that $D(p) = D_\rho(p)$ is bounded below by a constant multiple of $\max\{S(p), S(p)^2\}$, where $S(p) \equiv \sup_i |np_i - 1|$ and the constant depends only on ρ. It follows from this property and (5.6) that $S(\hat{p}) = O_p(1)$, and hence that

$$\sup_{1 \le i \le n} \hat{p}_i = O_p\big(n^{-1}\big), \tag{5.7}$$

which is one of the results noted in part (b) of the theorem.

Step (ii) [Bounds on $\hat{g}(\cdot|\hat{p}) - \tilde{g}$ and its derivative]. By the Cauchy–Schwarz inequality and for $j = 0, 1$,

$$\begin{aligned}\bigl|\hat{g}^{(j)}(x|\hat{p}) - \tilde{g}^{(j)}(x)\bigr| &\le \Bigl\{n^{-1}\sum_{i=1}^{n}(n\hat{p}_i - 1)^2\Bigr\}^{1/2}\Bigl\{n^{-1}\sum_{i=1}^{n}A_i^{(j)}(x)^2\Bigr\}^{1/2} \\ &= O_p\bigl(h^{2-j}\bigr),\end{aligned} \tag{5.8}$$

uniformly in $x \in \mathscr{I}$. Here we have used, in addition to (5.6) and (5.7), the following properties:

$$C_1(C,\rho)\,D_\rho(p) \le \sum_{i=1}^{n}(np_i - 1)^2 \le C_2(C,\rho)\,D_\rho(p) \tag{5.9}$$

uniformly in p such that $\sup_i np_i \le C$ [where $C_j(C,\rho)$ depends only on the indicated arguments] and

$$\sup_{x\in\mathscr{I}} n^{-1}\sum_{i=1}^{n}A_i^{(j)}(x)^2 = O_p\bigl(h^{-(2j+1)}\bigr).$$

Taking $j = 0$ in (5.8) we deduce that $\hat{g}(\cdot|\hat{p}) - \tilde{g} = O_p(h^2)$, uniformly on $\mathscr{I}$, which is one part of result (b) of the theorem. It also follows from (5.8) that

$$\sup_{x\in\mathscr{I}}\bigl|\hat{g}'(\cdot|\hat{p}) - \tilde{g}'\bigr| = O_p(h). \tag{5.10}$$

Step (iii) [Properties of $\hat{p}$, and thence $\hat{g}(\cdot|\hat{p})$]. In Step (v) we shall show that for an absolute constant $C_1 > 0$, and for each $\delta > 0$, there exists $C_2 = C_2(\delta) > 0$ such that, for all sufficiently large n,

$$P\bigl\{\tilde{g}'(x) > C_1\,g'(x) \text{ for all } x \ge a + C_2h\bigr\} > 1 - \delta. \tag{5.11}$$

We may assume that K is supported on the interval $[-1, 1]$. For the estimator types that we are considering, this means that

$$A_i(x) = 0 \qquad \text{if } |x - X_i| \ge h. \tag{5.12}$$

If η is a random variable, and if

$$p_i = n^{-1}(1 + \Delta) \qquad \text{for all } i \text{ with } X_i \ge \eta, \tag{5.13}$$

where Δ is a random variable not depending on i and satisfying $-1 < \Delta < \infty$, then in view of (5.12), $x \ge \eta + h$ implies that $p_i = n^{-1}(1 + \Delta)$ for each i such that $A_i(x) \ne 0$. Hence, using the definition of $\hat{g}(x|p)$ at (2.2), we see that $x \ge \eta + h$ implies that $\hat{g}(x|p) = (1 + \Delta)\,\tilde{g}(x)$. Therefore it follows from (5.11) that if p satisfies (5.13) then

$$\begin{aligned}P\bigl\{\hat{g}'(x|p) > C_1(1 + \Delta)\,g'(x) > 0 \text{ for all} \\ x \ge \max(a + C_2h,\, \eta + h)\bigr\} > 1 - \delta.\end{aligned} \tag{5.14}$$

From (2.4), (5.10) and (5.14) we may deduce that if Z_2h is defined to equal the infimum of $V > 0$ such that $\hat{p}_i$ is identically constant for all i such that $|x_0 - X_i| \ge V$, then $Z_2 = O_p(1)$.

That is, $\hat{p}_i$ is identically equal to $n^{-1}(1+\Delta)$, for some random variable Δ and for all i's such that $|X_i - x_0| \geq Z_2 h$, where $Z_2 = O_p(1)$. This establishes one aspect of part (b) of the theorem. To bound Δ we note that the contribution to $D(\hat{p})$ from those $\hat{p}_i$'s for which the index i satisfies $|X_i - x_0| > Z_2 h$ is, by (5.9), bounded below by a constant multiple of the number, N, of such i's [which satisfies $n = O_p(N)$] multiplied by Δ^2. [It is to be understood, here and below, that we are referring only to i's such that $A_i(x) \neq 0$ for some $x \in \mathscr{I}$.] Therefore, $n\Delta^2 = O_p\{D(\hat{p})\}$. It now follows from (5.6) that $\Delta = O_p(h^{5/2})$, which is the result in part (b) of the theorem.

Result (5.12), and the properties of $\hat{p}$ noted in the previous paragraph, imply that the probability that the random variable Z_1, defined by

$$Z_1 h = \inf\big\{z \geq 0\text{: for all } x \in \mathscr{I} \text{ for which } x \geq a + z,\\ A_i(x) = 0 \text{ whenever } \hat{p}_i \neq n^{-1}(1+\Delta)\big\},$$

is well-defined, converges to 1 as $n \to \infty$, and that conditional on Z_1 being well-defined, it satisfies $Z_1 = O_p(1)$. Moreover, $\hat{g}(x|\hat{p}) = (1+\Delta)\,\tilde{g}(x)$ for all x's with the property that $A_i(x) = 0$ whenever $\hat{p}_i \neq n^{-1}(1+\Delta)$. It follows that $\hat{g}(x|\hat{p}) = (1+\Delta)\,\tilde{g}(x)$ uniformly in values $x \in \mathscr{I}$ for which $x \geq a + Z_1 h$. This is the last result in part (b) of the theorem that remains to be proved.

Step (iv) [Derivation of (5.4) and (5.5)]. Define

$$\tilde{p}_i = n^{-1}\left\{1 + \Delta + g(X_i)^{-1} h^2 L\left(\frac{x_0 - X_i}{h}\right)\right\}, \tag{5.15}$$

where L is a fixed, compactly supported, twice-differentiable function and Δ is defined by $\sum_i \tilde{p}_i = 1$. [The variable Δ plays a role similar to Δ at (5.14), and so is expressed in the same notation, but it is generally different from that quantity.] Then,

$$\hat{g}'(x|\tilde{p}) = (1+\Delta)\tilde{g}'(x) + h^2\,\Delta_1(x) + h^2\Delta_2(x), \tag{5.16}$$

where

$$\Delta_1(x) = n^{-1}\sum_{i=1}^{n} L\left(\frac{x_0 - X_i}{h}\right)A_i'(x),$$

$$\Delta_2(x) = n^{-1}\sum_{i=1}^{n} g(X_i)^{-1} L\left(\frac{x_0 - X_i}{h}\right)A_i'(x)\varepsilon_i.$$

Note that

$$n^{-1}\sum_{i=1}^{n} g(X_i)^{-1} L\left(\frac{x_0 - X_i}{h}\right) = O_p(h).$$

Therefore, solving the equation $\sum_i \tilde{p}_i = 1$ for Δ, we find that

$$\Delta = O_p\big(h^3\big). \tag{5.17}$$

More simply, it may be proved that $\sup_{x\in\mathscr{I}} |\tilde{g}(x)| = O_p(1)$.

Under the conditions imposed in the theorem it may be shown that

$$\sup_{x\in\mathscr{I}} |\Delta_2(x)| = O_p\{(nh^3)^{-1/2}\} = O_p(h).$$

[Note that $\Delta_2(x)$ vanishes if x is distant more than $O(h)$ from x_0.] Likewise it may be proved that

$$\Delta_1(x) = -h^{-1}\int L'\Big(\frac{x_0 - x}{h} + y\Big)K(y)\,dy + O_p(1)$$

uniformly in $x \in \mathscr{I}$. Combining the results from (5.16) down we see that we may write

$$\hat{g}'(x|\tilde{p}) = \tilde{g}'(x) - h\int L'\Big(\frac{x_0 - x}{h} + y\Big)\,K(y)\,dy + \Delta_3(x), \tag{5.18}$$

where $\Delta_3(x) = O_p(h^2)$ uniformly in $x \in \mathscr{I}$. (The last two displayed formulas differ slightly in the local linear case, but the subsequent argument is identical.)

Given $\delta > 0$, we may choose $C > 0$ so large that for all sufficiently large n,

$$P\{\tilde{g}'(x) > 3h \quad \text{for all } x \in [a + Ch, b]\} > 1 - \tfrac{1}{3}\delta.$$

Given both C and δ we may choose L to be linearly decreasing, at a sufficiently fast rate, on a sufficiently wide interval containing the origin, and returning sufficiently slowly to 0 on either side of the interval where it is decreasing, such that

$$P\left\{\tilde{g}'(x) - h\int L'\Big(\frac{x_0 - x}{h} + y\Big)K(y)dy > 2h \text{ for all } x \in [a, a + Ch]\right\} > 1 - \tfrac{1}{3}\delta,$$

$$\int L'\Big(\frac{x_0 - x}{h} + y\Big)K(y)dy < 1 \qquad \text{for all } x \in [a, b].$$

Furthermore, for all $C, \delta > 0$, and all sufficiently large n,

$$P\{\Delta_3(x) \geq -h \text{ for all } x \in [a, b]\} > 1 - \tfrac{1}{3}\delta.$$

Combining (5.18) and the displayed results in this paragraph we deduce that (5.4) holds for all sufficiently large n. Result (5.5) follows from (5.9), (5.15), (5.17) and the property

$$n^{-1}\sum_{i=1}^{n} g(X_i)^{-2}L\left(\frac{x_0 - X_i}{h}\right)^2 = O_p(h).$$

Step (v) [Derivation of (5.11)]. Define $\mathscr{X} = \{X_1, \ldots, X_n\}$ and $\mu = E(\tilde{g}|\mathscr{X})$. (Thus, μ is a random function unless the X_i's are regularly spaced.) Let

$$A_n = \sup\{x \in \mathscr{I}\colon \mu'(x) < \tfrac{2}{3}\,g'(x)\}, \tag{5.19}$$

$$B_n = \sup\{x \in \mathscr{I}\colon |\tilde{g}'(x) - \mu'(x)| > \tfrac{1}{3}g'(x)\} \tag{5.20}$$

and $C_n = \max(A_n, B_n)$. Then,

$$\tilde{g}'(x) \geq \tfrac{1}{3}g'(x) > 0 \quad \text{for } x \geq C_n. \tag{5.21}$$

Therefore, (5.11) (with $C_1 = \frac{1}{3}$) would follow from the properties

$$A_n - a = O_p(h) \quad \text{and} \quad B_n - a = O_p(h). \tag{5.22}$$

These results may both be derived as follows. Assume (without loss of generality) that $x_0 = 0$ and $\mathscr{I} = [0, 1]$. Consider the stochastic processes $\gamma_1 = \mu' - g'$, $\gamma_2 = \tilde{g}' - \mu'$ and $\gamma_{kj}(u) = h^{-1}\,\gamma_k\{(j+u)h\}$, the latter defined for $0 \le j \le h^{-1} - 1$ and $u \in [0, 1]$. Using Taylor expansion and the Hungarian embedding it may be shown that on $\mathscr{I}_j \equiv [0, \min(1, h^{-1} - j)]$, and writing λ_j for either γ_{1j} or γ_{2j}, we have

$$\sup_j P\left\{\sup_{u \in \mathscr{I}_j} |\lambda_j(u)| > v\right\} \le C_1(1+v)^{-2}$$

for all $v \in [0, C_2 h^{-1}]$, uniformly in $j \le J$, where the constants $C_1, C_2 > 0$ do not depend on n, and J denotes the integer part of $h^{-1} - 1$. Therefore, defining $w_j \equiv (3h)^{-1} \inf_{u \in \mathscr{I}_j} g'\{(j+u)h\}$ and $v_j \equiv \min(w_j, C_2 h^{-1})$, we deduce that

$$P\big\{|\gamma(t)| > \tfrac{1}{3}\, g'(t) \text{ for some } t \in [j_0 h, 1]\big\} \le \sum_{j=j_0}^{J} C_1(1+v_j)^{-2}, \tag{5.23a}$$

where γ represents either γ_1 or γ_2. The conditions imposed on g in the theorem imply that $v_j \ge C_3 \min(j, C_4 h^{-1})$, where $C_3, C_4 > 0$ do not depend on n. It now follows from (5.23) that

$$\lim_{C \to \infty} \limsup_{n \to \infty} P\big\{|\gamma(t)| > \tfrac{1}{3} g'(t) \text{ for some } t \in [Ch, 1]\big\} = 0.$$

Applying this result to the respective versions of γ we obtain (5.22).

Part (c). The method of proof is similar to that in part (b), and so we do no more than give an outline of places where the arguments differ. The main difference is that the weights $\hat{p}_i$ and $\tilde{p}_i$ are now nonconstant for indices i such that X_i lies in a neighborhood of radius $O(h^{1/2})$, rather than $O(h)$, of x_0. In particular, analogously to (5.15), we define

$$\tilde{p}_i = n^{-1}\left\{1 + \Delta + g(X_i)^{-1} h^{3/2} L\left(\frac{x_0 - X_i}{h^{1/2}}\right)\right\}. \tag{5.23b}$$

In place of (5.16) one proves that $\hat{g}'(\cdot\,|\,\tilde{p}) = (1+\Delta)\,\tilde{g}' + h^{3/2}\,\Delta_1 + h^{3/2}\,\Delta_2$, where Δ_1, Δ_2 are the analogues of the respective quantities introduced in Step (iv) and satisfy $\sup_{x \in \mathscr{I}} |\Delta_2(x)| = O_p\{h(\log n)^{1/2}\}$ and $\Delta_1(x) = -h^{-1/2}\,L'\{(x_0 - x)/h^{1/2}\} + O_p(1)$ uniformly in $x \in \mathscr{I}$ and $\Delta = O_p(h^2)$ [instead of (5.17)]. Arguing as in Step (iv) it may be shown that (5.4) is satisfied, and in place of (5.5),

$$D(\tilde{p}) = O_p\big(nh^{7/2}\big). \tag{5.24}$$

On this occasion (5.7) does not necessarily hold. That result was used in the first paragraph of Step (ii), and in the present setting the argument there

should be replaced by

$$\begin{aligned}\bigl|\hat g^{(j)}(x|\hat p) - \tilde g^{(j)}(x)\bigr| &\le \Biggl\{n^{-1}\sum_{i:|n\hat p_i-1|\le 2}(n\hat p_i-1)^2\Biggr\}^{1/2}\Biggl\{n^{-1}\sum_{i=1}^n A_i^{(j)}(x)^2\Biggr\}^{1/2}\\ &\quad + n^{-1}\sum_{i:|n\hat p_i-1|>2}|n\hat p_i-1|\bigl|A_i^{(j)}(x)\bigr| \qquad (5.25)\\ &= O_p\Bigl[\bigl\{n^{-1}D(\hat p)h^{-(2j+1)}\bigr\}^{1/2} + n^{-1}D(\hat p)h^{-j-1}\Bigr]\\ &= O_p\bigl(h^{(5/4)-j}\bigr),\end{aligned}$$

uniformly in $x \in \mathscr{I}$. Here we have used (5.24) and the fact that

$$D(p) \asymp \sum_{i=1}^n \min\bigl\{|np_i-1|, (np_i-1)^2\bigr\}. \qquad (5.26)$$

[In the case $D = D_\rho$ with $\rho = 1$, an additional factor $\{1+(\log|np_i-1|)\,I(np_i>2)\}$ should be adjoined to the right-hand side of (5.26). This does not influence (5.25), however.] Taking $j = 0$ in (5.25) we obtain one of the results in part (c).

On this occasion the weights $\hat p_i$ are are identical to $n^{-1}(1+\Delta)$ [for a random variable which plays much the same role as Δ at (5.23), but is generally different from that quantity] for indices i that are distant $O_p(h^{1/2})$ from x_0. To derive a bound for Δ, let there be just N indices i such that $\hat p_i = n^{-1}(1+\Delta)$, and let $\mathscr{A}$ denote the set of the other $N_1 \equiv n - N$ indices. The fact that $\sum_i \hat p_i = 1$ implies that

$$N\Delta + \sum_{i\in\mathscr{A}}(n\hat p_i - 1) = 0, \qquad (5.27)$$

from which, since $N_1 = O_p(nh^{1/2})$, it follows that

$$\begin{aligned}|\Delta| &\le N^{-1}\sum_{i\in\mathscr{A}}|n\hat p_i - 1|\\ &\le N^{-1}\Biggl[\Biggl\{N_1\sum_{i\in\mathscr{A}:|n\hat p_i-1|\le 2}(n\hat p_i-1)^2\Biggr\}^{1/2} + \sum_{i\in\mathscr{A}:|n\hat p_i-1|>2}|n\hat p_i-1|\Biggr]\\ &= O_p(n^{-1})\Bigl[\bigl\{nh^{1/2}D(\hat p)\bigr\}^{1/2} + D(\hat p)\Bigr] = O_p(h^2),\end{aligned}$$

using (5.24) to obtain the last line. Under the assumptions for part (c), g is locally cubic in a neighborhood of x_0, and the cubic has vanishing first and second derivatives at x_0. It may be shown from this approximation that $\sum_{i\in\mathscr{A}}(n\hat p_i - 1) = o_p(nh^2)$ rather than simply $O_p(nh^2)$. Therefore, $\Delta = o_p(h^2)$. □

APPENDIX

Pairwise weighting. In the definition of $\hat{g}(\cdot|p)$ at (2.2) the weights p_i are applied primarily to the response variables. An alternative approach is to apply them to the data pairs (X_i, Y_i). This is not really feasible for Priestley–Chao and Gasser–Müller methods, but for Nadaraya–Watson and local linear methods it is potentially attractive. The pairwise-weighted versions of these estimators are, respectively,

$$\hat{g}_{\mathrm{NW}}(x|p) = \frac{\sum_i p_i K_i(x) Y_i}{\sum_i p_i K_i(x)},$$

$$\hat{g}_{\mathrm{LL}}(x|p) = \frac{S_2(x|p)T_0(x|p) - S_1(x|p)T_1(x|p)}{S_2(x|p)S_0(x|p) - S_1(x|p)^2},$$

where

$$S_k(x|p) = \sum_{i=1}^{n} p_i \{(x - X_i)/h\}^k K_i(x),$$

$$T_k(x|p) = \sum_{i=1}^{n} p_i \{(x - X_i)/h\}^k K_i(x) Y_i.$$

As before we would choose p to minimise $D(p)$ subject to the constraint of monotonicity.

Implementation is straightforward in the case of Nadaraya–Watson estimators, but for local linear estimators it is made awkward by the relative complexity of the second derivative. This is the main reason we did not develop the pairwise-weighting method further in this paper. Its main advantage is that it is equivariant under translation. However, as reported in Section 3, the simpler approach suggested at (2.2) is virtually translation-equivariant in practice; even very large translations do not have a perceptible influence on the shape of the estimate.

Acknowledgments. The authors are grateful to two referees, an Associate Editor and an Editor for helpful comments. Thanks are due too to Berwin Turlach for his help in implementing the algorithm of Turlach (1997).

REFERENCES

BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference Under Order Restrictions*. Wiley, New York.

BARTHOLOMEW, D. J. (1959). A test of homogeneity for ordered alternatives. *Biometrika* **46** 36–48.

BLOCH, D. A. and SILVERMAN, B. W. (1997). Monotone discriminant functions and their applications in rheumatology. *J. Amer. Statist. Assoc.* **92** 144–153.

BOWMAN, A. W. and AZZALINI, A. (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford Univ. Press.

BOWMAN, A. W., JONES, M. C. and GIJBELS, I. (1998). Testing monotonicity of regression. *J. Comput. Graph. Statist.* **7** 489–500.

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

CRESSIE, N. A. C. and READ, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B* **46** 440–464.

HALL, P. and HECKMAN, N. (2000). Testing for monotonicity of a regression mean by calibrating for linear functionals. *Ann Statist.* **28** 20–39.

HALL, P. and PRESNELL, B. (1999). Intentionally biased bootstrap methods. *J. Roy. Statist. Soc. Ser. B* **61** 143–158.

FISHER, N. I., HALL, P., TURLACH, B. A. and WATSON, G. S. (1997). On the estimation of a convex set from noisy data on its support function. *J. Amer. Statist. Assoc.* **92** 84–91.

FRIEDMAN, J. H. and TIBSHIRANI, R. J. (1984). The monotone smoothing of scatterplots. *Technometrics* **26** 243–250.

GONZÁLEZ-MANTEIGA, W., CAO, R. and MARRON, J. S. (1996). Bootstrap selection of the smoothing parameter in nonparametric hazard rate estimation. *J. Amer. Statist. Assoc.* **91** 1130–1140.

HOUGAARD, P. (1988). A boundary modification of kernel function smoothing, with application to insulin absorption kinetics. In *Compstat Lectures* 31–36. Physica, Vienna.

HOUGAARD, P., PLUM, A. and RIBEL, U. (1989). Kernel function smoothing of insulin absorption kinetics. *Biometrics* **45** 1041–1052.

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

MAMMEN, E. (1991). Estimating a smooth 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. and MARRON, J. S. (1997). Mass centred kernel smoothers. *Biometrika* **84** 765–777.

MAMMEN, E., MARRON, J. S., TURLACH, B. A. and WAND, M. P. (1999a). A general framework for constrained smoothing. Unpublished manuscript.

MAMMEN, E., MARRON, J. S., TURLACH, B. A. and WAND, M. P. (1999b). Monotone local polynomial smoothers. Unpublished manuscript.

MARRON, J. S., TURLACH, B. A. and WAND, M. P. (1997). Local polynomial smoothing under qualitative constraints. In *Graph-Image-Vision* (L. Billard and N. I. Fisher, eds.) 647–652. Interface Foundation of North America, Fairfax Station, VA.

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

MÜLLER, H.-G. (1997). Density adjusted kernel smoothers for random design nonparametric regression. *Statist. Probab. Lett.* **36** 161–172.

MÜLLER, H.-G. and SONG, K.-S. (1993). Identity reproducing multivariate nonparametric regression. *J. Multivariate Anal.* **46** 237–253.

PATIL, P., WELLS, M. T. and MARRON, J. S. (1994). Some heuristics of kernel based estimators of ratio functions. *J. Nonparametr. Statist.* **4** 203–209.

PEARSON, G. W. and QUA, F. (1993). High precision ^{14}C measurement of Irish oaks to show the natural^{14} variations from AD 1840–5000 BC: a correction. *Radiocarbon* **35** 105–123.

POINER, I. R., BLABER, S. J. M., BREWER, D. T., BURRIDGE, C. Y., CAESAR, D., CONNELL, M., DENNIS, D., DEWS, G. D., ELLIS, A. N., FARMER, M., FRY, G. J., GLAISTER, J., GRIBBLE, N., HILL, B. J., LONG, B. G., MILTON, D. A., PITCHER, C. R., PROH, D., SALINI, J. P., THOMAS, M. R., TOSCAS, P., VERONISE, S., WANG, Y. G. and WASSENBERG, T. J. (1997). The effects of prawn trawling in the far northern section of the Great Barrier Reef. Final report to CBRMPA and FRDC on 1991–96 research. CSIRO Division of Marine Research, Queensland Dept. Primary Industries.

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.

SCHLEE, W. (1982). Nonparametric tests of the monotony and convexity of regression. In *Nonparametric Statistical Inference II* (B. V. Gnedenko, M. L. Puri and I. Vincze, eds.) 823–836. North-Holland, Amsterdam.

TURLACH, B. A. (1997). Constrained smoothing splines revisited. Technical report SSR97-008, Australian National Univ., Centre for Mathematics and Its Applications.
WAND, M. P. and JONES, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

CENTRE FOR MATHEMATICS
AND ITS APPLICATIONS
AUSTRALIAN NATIONAL UNIVERSITY
CANBERRA, ACT 0200
AUSTRALIA
E-MAIL: halpstat@pretty.anu.edu.au
huang@maths.anu.edu.au