Visual Error Criteria for Qualitative Smoothing
Author(s): J. S. Marron and A. B. Tsybakov
Source: *Journal of the American Statistical Association*, Vol. 90, No. 430 (Jun., 1995), pp. 499-507
Published by: American Statistical Association
Stable URL: http://www.jstor.org/stable/2291060
Accessed: 27/01/2009 08:37

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/action/showPublisher?publisherCode=astata.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit organization founded in 1995 to build trusted digital archives for scholarship. We work with the scholarly community to preserve their work and the materials they rely upon, and to build a common research platform that promotes the discovery and use of these resources. For more information about JSTOR, please contact support@jstor.org.

American Statistical Association is collaborating with JSTOR to digitize, preserve and extend access to *Journal of the American Statistical Association.*

http://www.jstor.org

Visual Error Criteria for Qualitative Smoothing

J. S. MARRON and A. B. TSYBAKOV*

An important gap, between the classical mathematical theory and the practice and implementation of nonparametric curve estimation, is due to the fact that the usual norms on function spaces measure something different from what the eye can see visually in a graphical presentation. Mathematical error criteria that more closely follow "visual impression" are developed and analyzed from both graphical and mathematical viewpoints. Examples from wavelet regression and kernel density estimation are considered.

KEY WORDS: Bandwidth selection; Nonparametric curve estimation; Wavelet regression.

1. INTRODUCTION

Smoothing methods for nonparametric curve etimation, including density estimation (see Silverman 1986 for a good introduction) and regression (see Eubank 1988, Härdle 1991, Müller 1988, and Wahba 1991, for example), provide effective tools for graphical data analysis. Although there is a large theoretical literature on this topic, much of it is not particularly useful for understanding issues important to applications. One reason for this is that mathematical theory is typically based on error criteria that measure distance between curves (e.g., a "true curve" and an "estimate"), in terms of classical mathematical norms on function spaces (e.g., L^1, L^2, or L^∞). This can be quite inappropriate from a graphical viewpoint, because the eye does not work in this way, so one is led away from a *visual notion of distance between curves.* A dramatic example of this was provided by Kooperberg and Stone (1991), where a picture similar to Figure 1 is presented.

The estimators in Figure 1 have not been constructed from data but instead are carefully chosen Normal mixture densities. Table 1 shows some of the classical L^p distances between the true curve in Figure 1 and each of the estimates.

Note that with respect to any of the usual integral norms, estimate 1 is closer to the true curve (because the norms "feel" both spikes when comparing estimate 2 to the true curve). But in terms of visual impression, estimate 2 is much more appealing. The reason is that estimate 2 captures the important *qualitative* feature of the "bump" on the right side, although the location of the bump is not quite correct. The conventional mathematical approach to measuring error in curve estimation clearly gives the wrong answer concerning which estimate is "best." An exception to this is situations where peak locations are crucial (e.g., in spectroscopy), as pointed out by L. Tierney. But in the spectral analysis of time series, the presence and size of peaks can be a more important issue than their precise location.

Ideas of the type dramatized in Figure 1 have motivated the young field of "qualitative smoothing." This is a mathematical theory that attempts to study curve estimation, not through norms on function space but instead through qualitative features; for example, defining a "good" estimator to be one that has the correct number of modes (see Mammen 1991) or inflection points (see Cuevas and Gonzales Mantiega 1992). These are good first attempts at mathematically quantifying visual impression as to comparison of curve estimators, but they do not go far enough. In particular, there are many possible estimates with the right number of modes (or inflection points), some of which will be visually closer than others, because the modes are more nearly in the correct locations with more accurate heights. A deeper mathematical quantification that addresses this issue is needed.

In Section 2 we develop some nonstandard mathematical error criteria for curve estimation, which in our opinion follow visual impression much more closely in assessing the distance between two curves. In particular we define "symmetric error criteria," SE_1, SE_2, and SE_∞, which give performance that is much closer to "what the eye sees" in the example of Figure 1 (see Table 2).

Note that these error criteria all show that estimate 2 is closer to the true curve, which is the opposite conclusion from Table 1.

In Section 3 we focus on the specific problem of smoothing parameter selection (which is fundamental to graphical applications of curve estimation). We consider several challenging examples of curve estimation and compare the behavior of estimates with the smoothing parameter minimizing certain criteria. By "challenging," we mean the examples where L^p criteria do not give "visually" adequate quantification of the error. We illustrate there reasons why we prefer SE_2 to SE_1 or SE_∞. We also see that there are some situations where SE_2 also does not match "visual impression," especially that of an experienced data analyst. In particular, in situations where there is not information present in the data to recover all features of the true underlying curve, an *asymmetric* error criterion is preferable.

Hence our final recommendation for error criteria comes in two parts:

1. When the goal is "the estimate should capture as many of the qualitative features of the data as possible," we prefer the criterion SE_2.
2. For the *different* goal of "duplicate the choice of an experienced data analyst," we prefer an asymmetric criterion discussed in Section 3.

Section 4 provides some asymptotic analysis of some of these new error criteria, including description of their rela-

* J. S. Marron is Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27599. A. B. Tsybakov is Professor of Statistics, University of Paris VI, 75252 Paris Cedex 05, France. Tsybakov was at C.O.R.E. and the Institute of Statistics at the Catholic University of Louvain, Louvain-la-Neuve, Belgium and the Institute for Problems in Information Transmission, Moscow, Russia, when this research was done. This research was partially supported by National Science Foundation Grant DMS 9203135. The authors are very grateful for many helpful comments made at the meeting "Curves, Images, and Massive Computation" at Oberwolfach.

© 1995 American Statistical Association

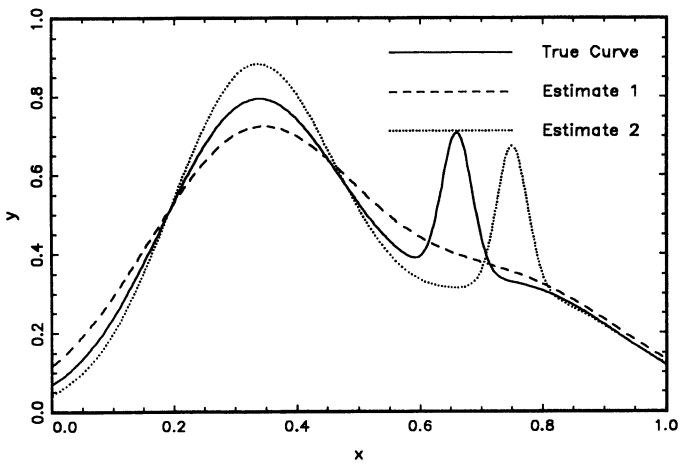

Figure 1. Kooperberg Stone Example, Showing That Classical Distances are Different From "Visual Distance."

tionship to conventional asymptotic analysis. It is seen that both suggested error criterion are rather tame in nature, being analogous to using a weighted version of L^2.

Although these criteria work well in a visual sense, they are not ideal for all statistical problems. In particular, if a regression is to be used solely for prediction purposes, then the L^2 fit has important optimality properties and is preferable to those presented here.

Approaches to related problems may be found in the work of Bookstein (1986), Ripley (1986), Kendall (1989), Nielsen and Foley (1989), and Baddeley (1993).

2. SOME NONSTANDARD ERROR CRITERIA

Figure 2a gives insight as to why the usual norms are inappropriate in comparing the two estimates presented in Figure 1. Note that L^1, L^2, and L^∞ are all based on the *vertical* distances between the curves (represented by the thin vertical lines). But the eye uses *both horizontal and vertical* information. A mathematical method for capturing this is to treat the curves not as functions of a single variable but instead as sets of points in the plane. For example, a given continuous function $f : [a, b] \to \mathbb{R}$ can be represented by its "graph,"

$$G_f = \{(x, y) : x \in [a, b], y = f(x)\} \subset \mathbb{R}^2$$

(G_f is the set in the plane-shaded black in a plot of the function $f(x)$). This allows replacement of the vertical distance between the true curve $f(x)$ and an estimate $\hat{f}(x)$, by some planar distance between the sets of points G_f and $G_{\hat{f}}$.

As "distances between sets" may not be such a familiar notion, we first review the standard Hausdorff distance (see, for example, Sendov 1990). The basis of this is the notion of the distance from a point to a set:

$$d((x, y), G) = \inf_{(x',y')\in G} \|(x, y) - (x', y')\|_2;$$

that is, the shortest distance from the given point (x, y) to any point in the closed set G, where $\|\cdot\|_2$ denotes the usual euclidean distance (chosen over other possibilities because this is *visual* distance in the plane). Distances from the points in the set G_1 to the set G_2 can then be combined into the set of distances

$$\mathcal{D}(G_1, G_2) = \{d((x, y), G_2) : (x, y) \in G_1\}.$$

Table 1. Some Classical Distances Between the True Curve and the Two Estimates in Figure 1

	Estimate 1	Estimate 2
L^1	.048	.069
L^2	.073	.115
L^∞	.308	.393

Table 2. New Measures of the Distances Between the True Curve and the Two Estimates in Figure 1

	Estimate 1	Estimate 2
SE_1	.0536	.0506
SE_2	.0576	.0516
SE_∞	.0527	.0408

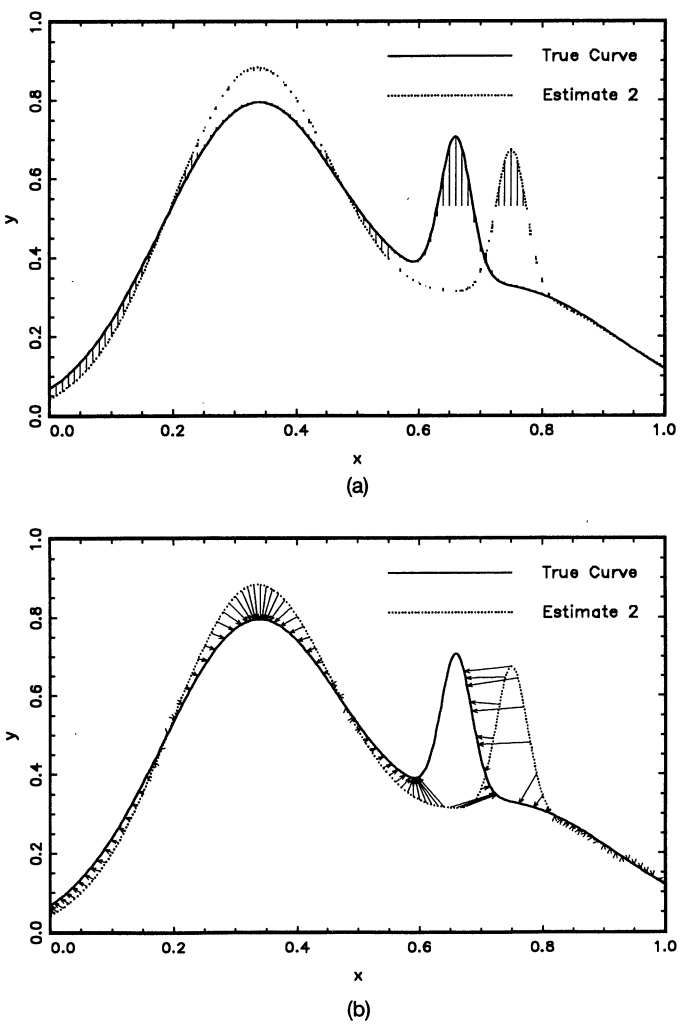

Figure 2. The Difference Between (a) "Vertical Distance," as Classically Used, and (b) "Visual Distance," as Quantified Here.

Some of the distances in $\mathcal{D}(G_{\text{Estimate 2}}, G_{\text{True Curve}})$ are represented by the arrows in Figure 2b (the representation is not exact because of discretization, as discussed at the end of this section). These distances are then combined to give the Hausdorff distance as

$$d_H(G_1, G_2) = \max\{\sup(\mathcal{D}(G_1, G_2)), \sup(\mathcal{D}(G_2, G_1))\}.$$

The Hausdorff distance, $d_H(G_{\hat{f}}, G_f)$, is an improvement over the L^p norms, in the sense that it accounts for both "vertical" and "horizontal" information about the deviation of a curve estimate $\hat{f}$ from the true curve f. But it has the disadvantage that it inherits the sometimes unappealing characteristics of sup type norms, that it only measures error at the worst location and totally ignores error elsewhere. (The fact that this can be quite different from what the eye sees" is illustrated clearly in Figure 5.) Hence alternate methods of basing error criteria on the sets $\mathcal{D}(G_1, G_2)$ and $\mathcal{D}(G_2, G_1)$, which provide a more complete summary of the deviations, are considered as well. In our applications, G_1 and G_2 are the graphs of some functions, say f_1 and f_2. A class of summaries of $\mathcal{D}(G_1, G_2)$, which yield *asymmetric* "visual error" criteria, are

$$\text{VE}_i(f_1 \to f_2) = \left[\int_a^b d((x, f_1(x)), G_{f_2})^i \, dx\right]^{1/i},$$

where $i = 1, 2, \infty$ (replacing the integral by the sup norm for $i = \infty$), which represent the various integral norms of the thin arrows in Figure 2b.

For some (but as seen in the next section, not all) purposes, it is useful to work with "symmetrized versions" of the VE_i. In particular, the symmetric error criteria in Table 2 are

$$\text{SE}_1(f_1, f_2) = \text{VE}_1(f_1 \to f_2) + \text{VE}_1(f_2 \to f_1),$$

$$\text{SE}_2(f_1, f_2) = [\text{VE}_2(f_1 \to f_2)^2 + \text{VE}_2(f_2 \to f_1)^2]^{1/2},$$

and

$$\text{SE}_\infty(f_1, f_2) = d_H(G_{f_1}, G_{f_2}) = \max\{\text{VE}_\infty(f_1 \to f_2), \text{VE}_\infty(f_2 \to f_1)\}.$$

Many other error criteria can also be constructed from the VE_i; for example, by taking various sums and suprema, most of which will give better visual assessment of the performance of the curve estimators than the classical L^1, L^2, and L^∞ norms. But observe that these are different from most (except d_H) of the conventional notions of distance between sets, because heavy use is made of the fact that G_1 and G_2 are the graphs of *functions* (in particular, each x appears once and only once). For this reason SE_1 and SE_2 are not directly applicable to the apparently related problem of measuring distances between arbitrary sets in the plane, although d_H is well known. An approach to this problem is the distance proposed by Baddeley (1993), which is another integrated alternative to the Hausdorff distance.

Note that SE_1 and SE_2 are *not* "distances" on the space of functions, because they do not satisfy the "triangle inequality." For example, let $[a, b] = [0, 1], f_1(x) \equiv -1, f_2(x) = \sin(2\pi kx)$, and $f_3(x) \equiv 1$. Consider first $i = 1$. The intuition behind this example is that each point of both f_1 and also f_3 are close to some point of f_2, but that f_1 and f_3 are "very far from each other." Note that $\text{VE}_1(f_1 \to f_3) = \text{VE}_1(f_3 \to f_1) = 2$, and that by taking k large enough, $\text{VE}_1(f_1 \to f_2) = \text{VE}_1(f_3 \to f_2)$ can be made arbitrarily small. But

$$\text{VE}_1(f_2 \to f_1) = \text{VE}_1(f_3 \to f_1) = \int_0^1 [1 \pm \sin(2\pi kx)] \, dx = 1.$$

Hence

$$\text{SE}_1(f_1, f_3) = 4 > 2 = 1 + 1 \approx \text{SE}_1(f_1, f_2) + \text{SE}_1(f_2, f_3).$$

With slightly more work, this same example also shows that SE_2 also does not satisfy the triangle inequality.

To cover the case of discontinuous f (important, for example, in image processing), it is convenient to assume that f is a set-valued function. For example, when f has only finitely many discontinuities, it is sensible to define the "value" of f at a discontinuity point x to be the *set*

$$\{z \in \mathbb{R} : \min(f(x-0), f(x+0)) \le z \le \max(f(x-0), f(x+0))\}.$$

In other words, we suppose that $f : [a, b] \to \mathcal{B}(\mathbb{R})$, where $\mathcal{B}(\mathbb{R})$ is the collection of all Borel subsets of $\mathbb{R}$, and define

$$G_f = \{(x, y) : x \in [a, b], y \in f(x)\}.$$

The other definitions then carry over in a straightforward way if we take $d((x, f_1(x)), G_2)$ to mean $\sup_{y \in f_1(x)} d((x, y), G_2)$.

F. Götze and others have pointed out that our distances are similar in spirit to the Prokhorov distances between cumulative distribution functions. But this is different in that the distances in $\mathcal{D}(G_1, G_2)$ and $\mathcal{D}(G_2, G_1)$ are essentially measured along normal vectors versus the 45-degree vectors that are implicit in Prokhorov distance.

It is often useful to think of graphical smoothing methods in the case of continuous functions, but practical implementation requires discretization. Here the common practice of constructing curves for plotting on an equally spaced (with grid spacing Δx, say), compact grid of x locations, say $\mathcal{X}$, is followed. Denote the discretized version of a graph G_f of a function $f(x)$ by $G_f^{\text{discr}} = \{(x, f(x)) : x \in \mathcal{X}\}$. The discretized version of $\text{VE}_i(f_1 \to f_2)$ is given by

$$\text{VE}_i^{\text{discr}}(f_1 \to f_2) = \left[\Delta x \sum_{x \in \mathcal{X}} d((x, f_1(x)), G_2^{\text{discr}})^i\right]^{1/i}.$$

Such discretized measures are used in all examples in this article, where $\mathcal{X}$ is an equally spaced grid of 400. It is because of this discretization (and because of the aspect ratio in the plots) that the arrows in Figure 2b are not exactly perpendicular to the curves.

The error criteria VE_i and SE_i, $i = 1, 2, \infty$, all depend heavily on the relative units of x and y. For example, in Figure 2 the units shown will give results different from "visual impression," because d will put much more weight in the x direction. For this reason we recommend working with rescalings of x and y. The goal of visual impression is best served by something in the spirit of linearly transforming x and y so that both $[a, b]$ and $[\inf_{x \in \mathcal{X}} f(x), \sup_{x \in \mathcal{X}} f(x)]$

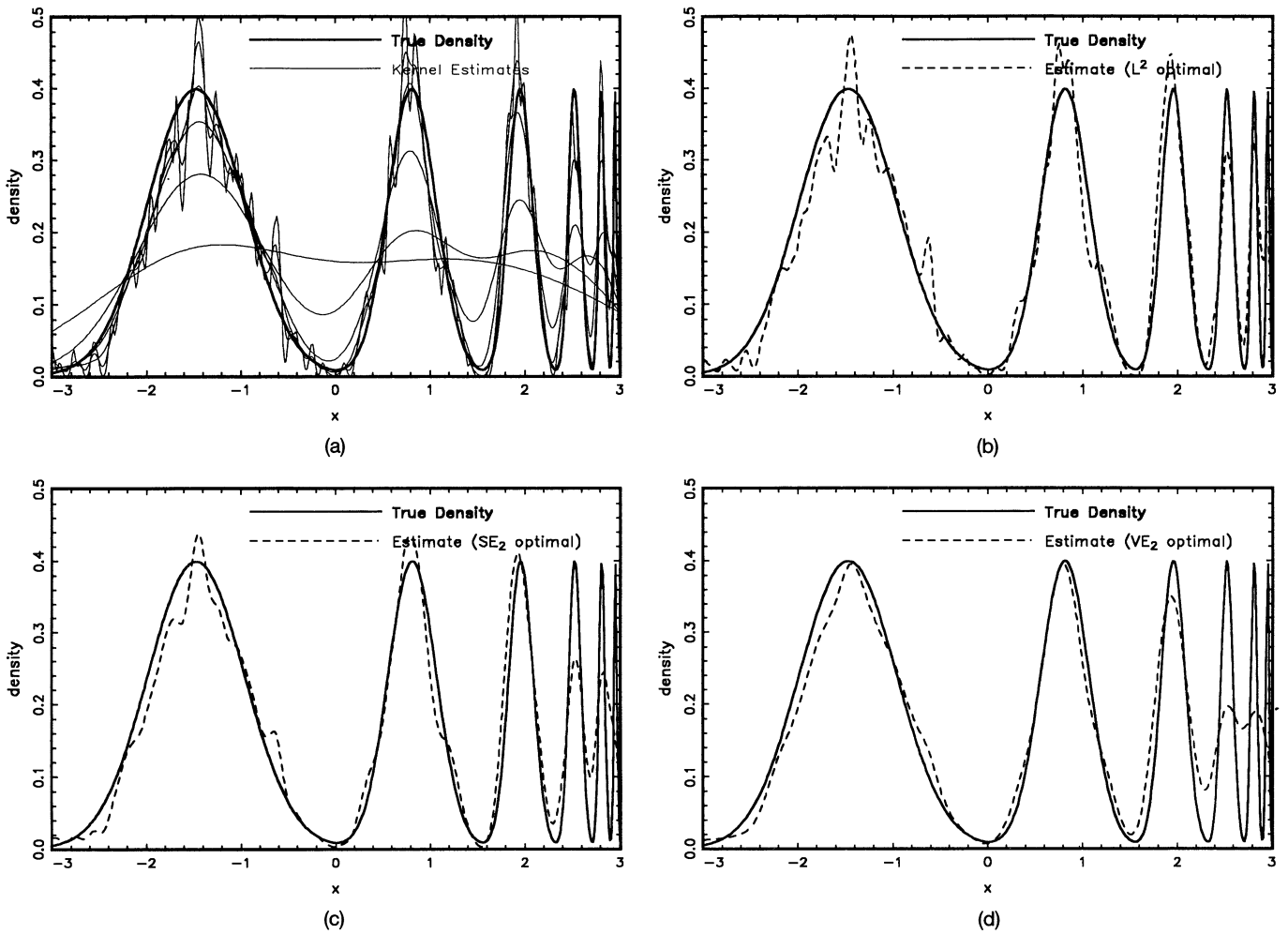

Figure 3. Kernel Density Estimation, With Different Bandwidths Applied to Same Pseudodata Sets, Showing That "Visual Choice" is Different From "Optimal With Respect to the Usual Norms." Figure 3a shows a family of estimates, indexed by the bandwidth, from undersmoothed to oversmoothed. Remaining figures show estimates using various optimal bandwidths.

are mapped to [0, 1]. Note that the L^p errors also depend on relative units, except in the special case of L^1 when the estimate $\hat{f}$ and the true curve f are both probability densities (see Devroye and Györfi 1985).

All examples here are given for one-dimensional curve estimates. But all these ideas can be generalized in a straightforward way to estimation in higher dimensions.

3. VISUAL ERROR IN SMOOTHING PARAMETER SELECTION

Practical use of any smoother (i.e., nonparametric curve estimator) requires choice of some smoothing parameter. There has been much work done on methods that use the data in this choice (see, for example, Jones, Marron, and Sheather 1992 and Marron 1988). But a trial-and-error visual fit (especially by an experienced analyst) still remains a very effective method of choosing the smoothing parameter. Nearly all of the theoretical work on smoothing parameter selection is based on assessing error through norms such as L^1, L^2, or L^∞, which is a drawback when these behave differently from visual choice. Other common criteria, such as Hellinger and (entropy-based) Kullback–Leibler distances do not overcome this difficulty, and in fact are often even less suitable, because they place too much weight on tails, as noted by Hall (1987).

To address this problem, we have experimented with the aforementioned measures of visual fit, in a number of different examples, in the particular cases of kernel density estimation and of wavelet regression estimation, but it is clear the ideas apply generally. Our first guess was that the symmetrized criteria, SE_i, $i = 1, 2, \infty$, would be most sensible, but we found that in many situations our "visual favorite" was more along the lines of the asymmetric VE_i.

In Figure 3 we consider kernel density estimation (see Silverman 1986 or Sec. 4 for a formal definition) using a Gaussian kernel function, for different bandwidths, based on a single simulated data set of size $n = 1{,}000$. The underlying true density is Normal Mixture #14 from Marron and Wand (1992). Note that the smaller bandwidths h give estimates that are rougher and more wiggly because they are too strongly influenced by sampling variability, and the larger ones give results that are smoother although at the expense of smoothing away some of the features of the true underlying curve. A range of such estimates is shown in Figure 3a. As

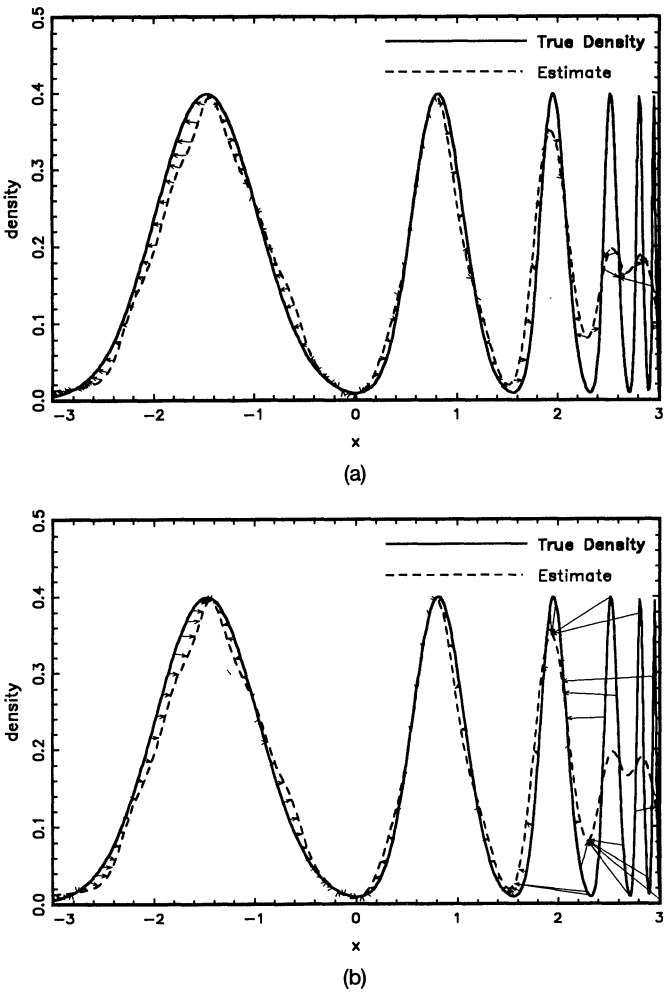

Figure 4. Asymmetry of Visual Error Criteria. Distances in $\mathcal{D}(G_{\hat{f}}, G_f)$ and $\mathcal{D}(G_f, G_{\hat{f}})$ are represented by small arrows in (a) and (b).

noted earlier, choice of the amount of smoothing (i.e., the bandwidth h), is a well-known hurdle in the practical use of this and other smoothing methods. Bandwidth selection is especially challenging in this case, because for best results in estimating this particular density, one should use an estimator with "location varying smoothing." (Clearly, a relatively larger bandwidth is desirable on the left side, and smaller one on the right side.) But before using such a complicated method, one needs to see the need for it, based on the simpler global bandwidth smoother shown here so it is important to study global bandwidth selection even in such challenging settings.

To see how "visual impression" can be asymmetric in nature, compare the curve estimates shown in Figure 3, b–d. Figure 3b presents the estimator with the L^2 optimal bandwidth, which clearly is visually inappropriate. In particular the resulting estimate is "too rough" on the first two peaks (which represent three-quarters of the picture!), because the L^2 distance is dominated by the behavior at the last three peaks and the valleys between. The SE_2 optimal bandwidth choice in Figure 3c is an improvement, but is still rougher than most data analysts would choose in our opinion. An estimator based on a bandwidth that we believe more closely models the choice of an experienced data analyst is that shown in Figure 3d, where the bandwidth is the minimizer of the $VE_2(\hat{f} \to f)$ criterion. Although the SE_2 optimum does a fine job of recovering the spikes on the right side, note that this is quite different from one's intuitive desires, because with only $n = 1{,}000$ observations there is a limit as to how much of these smaller spikes can be recovered from the data. The $VE_2(\hat{f} \to f)$-optimal estimate is less efficient at recovering the thin spikes, while sensibly trading off this aspect of the picture for good overall behavior.

The asymmetry in these VE is demonstrated in Figure 4. Figure 4a graphically illustrates the distances $\mathcal{D}(G_{\hat{f}}, G_f)$. Note that the magnitude of the (difficult to recover) thin spikes in $f(x)$ are not important factors and thus do not contribute to $VE_2(\hat{f} \to f)$. On the other hand, the distances $\mathcal{D}(G_f, G_{\hat{f}})$, shown in Figure 4b, are strongly affected by the (visually irrelevant, for moderate sample sizes) heights of the narrow spikes, which is why $VE_2(f \to \hat{f})$ is clearly inappropriate in this case. But $VE_2(\hat{f} \to f)$ does not solve all problems, and in particular its analog of Table 2 shows that in Figure 1, estimate 1 is closer to the true curve in this sense (although estimate 2 is closer in the sense of $VE_2(f \to \hat{f})$). Our preference on the basis of this experience is for the criterion $VE_2(\hat{f} \to f)$ in situations where it is desirable to duplicate the choice of an experienced data analyst in "recovering only those features of the true curve that can be well obtained from the data at hand." On the other hand, we prefer SE_2 in the much different situation where "the estimate should reflect as many qualitative features of the true curve as possible."

Figure 3c also shows that the symmetrized criterion $SE_2(\hat{f}, f)$ behaves more like the inappropriate L^2 optimal than like $VE_2(\hat{f} \to f)$. Pictures for SE_1 and SE_2 contain a similar lesson.

The reason that we prefer VE_2 and SE_2 to VE_1 and SE_1 is that the former summarize the distances in $\mathcal{D}(G_{\hat{f}}, G_f)$ (and in $\mathcal{D}(G_f, G_{\hat{f}})$) in a way which is closer to visual impression, because VE_2 "feels the larger distances more strongly." For example, note that in Table 2, the SE_1 distances tell the visually clear story less strongly than do their SE_2 analogs.

The reason that we prefer VE_2 and SE_2 to VE_∞ and SE_∞ is demonstrated in Figure 5. Each part of Figure 5 shows a regression problem, with true underlying regression curve, $f(x)$, a step function. A sample, (X_i, Y_i), $i = 1, \ldots, 100$, where $X_i \sim U(0, 1)$, and $Y_i \mid X_i \sim N(f(X_i), (.1)^2)$, was generated. We then attempted to recover $f(x)$ from the data using a Haar wavelet regression estimator. This estimator is a particular type of "orthogonal series" estimator, with interesting properties, especially when $f(x)$ has jumps, as here. (See Donoho and Johnstone 1992 for interesting discussion and references to statistical aspects of this estimator and (1.1.16) in Chui 1992 for the Haar basis.) Because wavelet estimators require equally spaced "design points," x_i, we "binned" the data to 64 equally spaced points, by averaging the Y_i in each bin. The result of this binning is shown by dotted-line segments that interpolate the data. Figure 5a shows the estimator, where only coefficients below a certain frequency are used (and is the best of that type). Figure 5b shows the estimator, using the "hard thresholding" idea of Donoho and Johnstone (1992), and their automatic thresh-

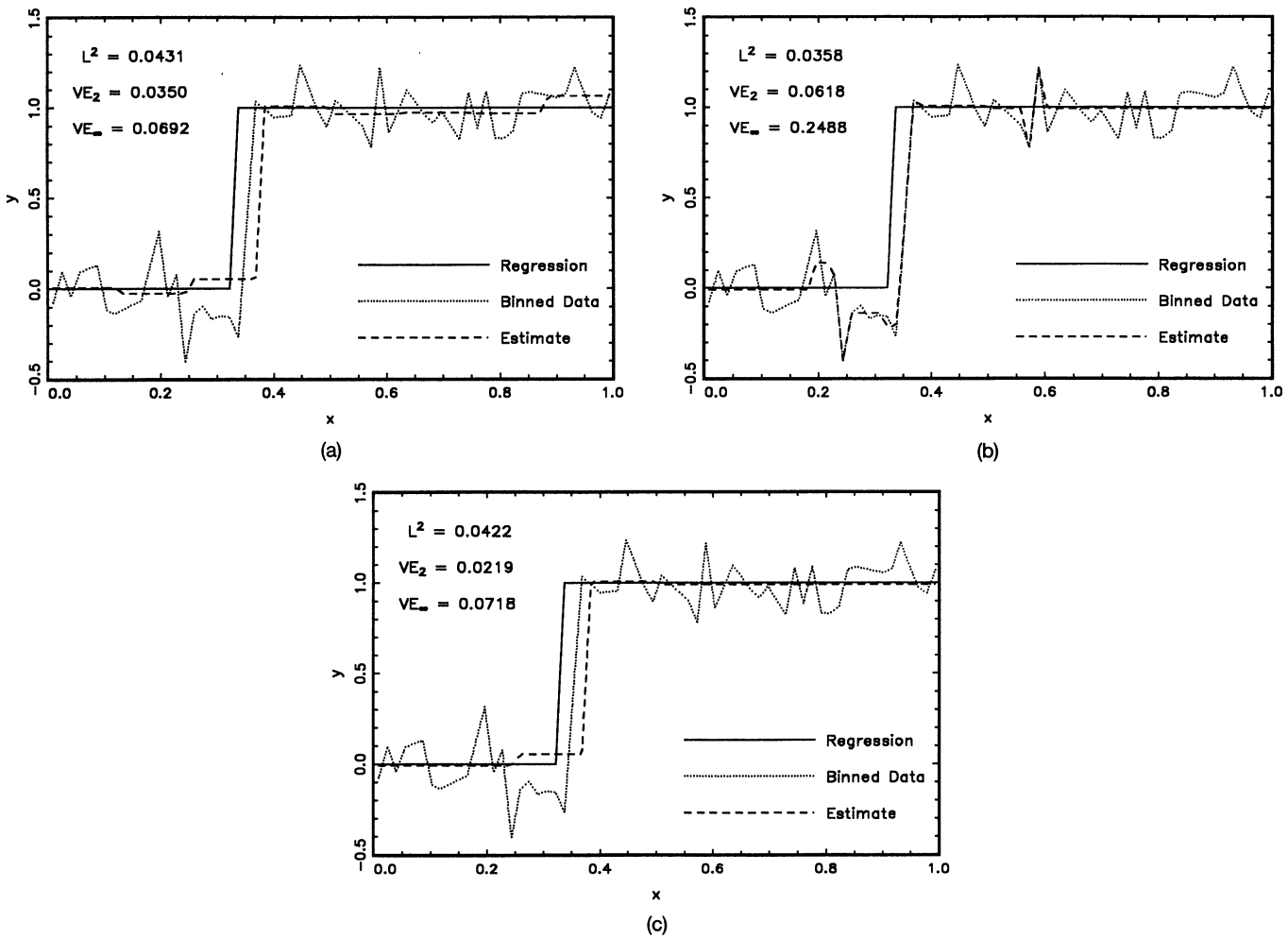

Figure 5. Haar Wavelet Regression Example, With Estimators. (a) Naive low-frequency reconstruction without the top 3 rows; (b) Donoho–Johnstone thresholded; (c) four times the Donoho–Johnstone threshold.

old method "wave shrink." The visual fit was not as good as we had hoped, so we tried other thresholds as well, with a good result achieved in Figure 5c using four times the suggested threshold.

The estimate in Figure 5c is clearly a visually superior fit to that in Figure 5b, but the L^2 error is actually worse. This is because the unsightly blips in Figure 5b have less effect on L^2 than the rightward shift in the location of the jump in the Figure 5c estimates. On the other hand, $\mathrm{VE}_2(\hat{f} \to f)$ again gives the correct (from a visual point of view) comparison of the two estimates. The smoothing parameter selector "wave shrink" has some excellent asymptotic properties with respect to L^2, so it is not surprising that it works well in that sense. But as L^2 is clearly different from visual impression in this case, it is of interest to develop smoothing parameter selectors that optimize criteria related to $\mathrm{VE}_2(\hat{f} \to f)$.

Figure 5, a and c, addresses two issues. The first is the fact that various "thresholding" methods can give better performance than simply deleting all high-frequency terms. (This is intuitively clear and was well discussed in Donoho and Johnstone 1992.) The second is the choice of $\mathrm{VE}_2(\hat{f} \to f)$, as opposed to $\mathrm{VE}_\infty(\hat{f} \to f)$. Although it is visually clear that the estimate in Figure 5c is superior to that in Figure 5a, note that $\mathrm{VE}_\infty(\hat{f} \to f)$ gives the reverse ordering. The reason is that $\mathrm{VE}_2(\hat{f} - f)$ feels *only* the worst distance in $\mathcal{D}(G_{\hat{f}}, G_f)$ (at the lower part of the jump in both cases), thus ignoring the fact that the estimate in Figure 5a is inferior at most other locations. But $\mathrm{VE}_2(\hat{f} \to f)$ corrects this by including error from all locations in its integral.

C. J. Stone has pointed out a situation in which the effectiveness of SE_2 in "capturing qualitative features" needs careful interpretation. The essence of this idea is demonstrated in Figure 6, which again shows a "true curve" and some "estimates" (again constructed only from shifts and scales of Normal mixture densities and not from actual estimates based on real data).

Note that in terms of any of our visual error criteria, estimate 1 is closer to the true curve. But if one cared *only* about qualitative features, such as the number of modes, then estimate 2 could be viewed as the better estimate. This shows that our preference for SE_2, when qualitative features are of interest, in fact represents some compromise between qualitative features and "goodness of fit."

We view the criteria proposed in this article only as starting points and believe that many improvements are possible.

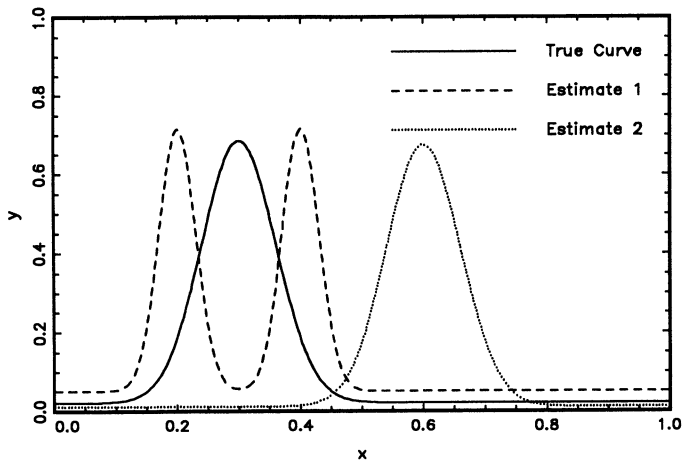

Figure 6. Stone Example Showing That the Symmetrized Visual Error Criterion SE_2 Does not Always Capture Strictly Qualitative Features, But Instead Provides a Compromise Between Qualitative Features and Good Fit.

D. Donoho has proposed making these criteria more qualitative (e.g., addressing the problem raised in Fig. 6) by "not allowing reuse of arrow heads." For example, in Figure 2b note that arrows from both sides of the small bump in the estimate go to essentially the same points on the left side of the small bump in the true curve. Qualitative improvement may be obtained by not allowing this, by some method. In conversation with F. Natterer, another possibility was discussed, which was to apply these same visual criteria not just to the curves, but also to the derivative curves (i.e., to get visual analogs of Sobolev distance). This also could address the issue raised in Figure 6, because the derivative of estimate 1 is far different from that of the true curve.

4. MATHEMATICAL ANALYSIS

In this section we explicitly analyze the criterion VE in the context of kernel density estimation, but very similar ideas apply also to the other criteria discussed here and to other curve estimation settings as well. Many extensions (e.g., weakening of the assumptions) are straightforward but are not pursued here to avoid obscuring our main points. The kernel density estimator is defined by $\hat{f}_n(x) = n^{-1} \sum_{i=1}^n K_h(x - X_i)$, where $X_1, \ldots, X_n$ are a random sample from $f(x)$, and $K_h(\cdot) = K(\cdot/h)/h$, for a symmetric probability density K, and a "bandwidth" $h = h_n > 0$.

For simple asymptotics we use the following technical assumptions:

A1. K is twice continuously differentiable and has compact support.
A2. The second derivative K'' is Lipschitz continuous on $\mathbb{R}$.
A3. $\int K(u)\,du = 1$, $\int uK(u)\,du = 0$, $\int u^2K(u)\,du \neq 0$.
A4. $h_n \to 0$, $nh_n \to \infty$ as $n \to \infty$.
A5. The density $f(x)$ is twice continuously differentiable on $\mathbb{R}$.

Under A1–A5, we have, for every x_0 (see Silverman 1986, for example),

$$\left(\frac{nh_n}{f(x_0)(\int K(u)^2\,du)}\right)^{1/2} \times \left(\hat{f}_n(x_0) - f(x_0) - \frac{1}{2}h_n^2 f''(x_0)\left(\int u^2K(u)\,du\right)\right) \xrightarrow{\mathcal{L}} N(0,1) \tag{1}$$

as $n \to \infty$, where $\xrightarrow{\mathcal{L}}$ denotes the convergence in distribution. This shows that a small bandwidth results in high variance, quantifying the roughness in the "undersmoothed" part of Figure 2, and that a large bandwidth results in high bias (worse for more curvature in f), which quantifies the tendency to smooth away features of f seen in the "oversmoothed" part of Figure 2. The convergence rate of $\hat{f}_n$ to f is optimized if

A6. $h_n = h_0 n^{-1/5}$, for a constant $h_0 > 0$.

Under A1–A6, we have

$$n^{2/5}(\hat{f}_n(x_0) - f(x_0)) \xrightarrow{\mathcal{L}} N(b(x_0), \sigma^2(x_0)), \tag{2}$$

where

$$b(x_0) = \frac{h_0^2}{2} f''(x_0) \int u^2K(u)\,du$$

and

$$\sigma^2(x_0) = \frac{1}{h_0} f(x_0) \int K(u)^2\,du.$$

Note that the asymptotic distribution results (1) and (2) work only "vertically," measuring error as in Figure 2a. Here we analyze the asymptotic behavior of the more visual error criteria, $\mathrm{VE}_2(\hat{f}_n \to f)$ and $\mathrm{VE}_2(f \to \hat{f}_n)$. This is essentially determined by the asymptotics of $d((x, f(x)), G_{\hat{f}_n})$ and $d((x, \hat{f}_n(x)), G_f)$.

For fixed $x_0 \in \mathbb{R}$, consider the random variables

$$d_n(x_0) \triangleq d((x_0, f(x_0)), G_{\hat{f}_n})$$

and

$$d_n^*(x_0) \triangleq d((x_0, \hat{f}_n(x_0)), G_f),$$

whose distributions are related to the limiting distributions in (2) through the following proposition.

Proposition 1. Assume A1–A6. Then

$$n^{2/5}\left(d_n(x_0) - \frac{|\hat{f}_n(x_0) - f(x_0)|}{\sqrt{1 + (f'(x_0)^2)}}\right) \xrightarrow{p} 0 \tag{3}$$

and

$$n^{2/5}\left(d_n^*(x_0) - \frac{|\hat{f}_n(x_0) - f(x_0)|}{\sqrt{1 + (f'(x_0)^2)}}\right) \xrightarrow{p} 0, \tag{4}$$

as $n \to \infty$.

Proof of this proposition is given in the Appendix.

Using (1) and (4), we find that the mean visual squared error $E(\mathrm{VE}_2(\hat{f}_n \to f)^2)$ is

$$E(\mathrm{VE}_2(\hat{f}_n \to f)^2) = \int d((x, \hat{f}_n(x)), G_f)^2 \, dx$$
$$= \int E([d_n^*(x)]^2)$$
$$\approx \int \frac{E((\hat{f}_n(x) - f(x))^2)}{1 + (f'(x))^2} \, dx$$
$$\approx n^{-4/5} \int \frac{b^2(x) + \sigma^2(x)}{1 + (f'(x))^2} \, dx,$$

which gives useful intuitive information. For example, as seen in Figure 3, there is substantially less error (both variance and bias) at x locations where $|f'(x)|$ is large (i.e., f is steep), but roughly the usual error at locations where $f'(x) \approx 0$ (i.e., f is flat).

Note that the final expression in the foregoing display is the standard asymptotic representation of the *weighted* mean integrated squared error (MISE), with weight function

$$w(x) = \frac{1}{1 + (f'(x))^2}.$$

Quite similarly, using (2) and (3), we obtain

$$E(\mathrm{VE}_2(f \to \hat{f}_n)^2) \approx \int \frac{E((\hat{f}_n(x) - f(x))^2)}{1 + (f'(x))^2} \, dx.$$

Hence both $E(\mathrm{VE}_2(\hat{f}_n \to f)^2)$ and $E(\mathrm{VE}_2(f \to \hat{f}_n)^2)$ converge to the same weighted version of MISE. Thus it is not surprising that similar arguments show that SE_2^2 converges to twice this same weighted MISE. It is important to keep in mind, though, that these relationships are only asymptotic, as $\mathrm{VE}_2(\hat{f}_n \to f)$. In many situations, such as those illustrated in the figures used in this article, the behavior of $\mathrm{VE}_2(f \to \hat{f}_n)$ is substantially different from MISE. But this does show the relationship between $\mathrm{VE}_2(f \to \hat{f}_n)$ and the classical asymptotic theory and demonstrates that much of the usual theory can be fairly simply adapted to the new error criteria introduced here.

Although the asymptotic expressions for

$$E(\mathrm{VE}_2(\hat{f}_n \to f)^2)$$

and

$$E(\mathrm{VE}_2(f \to \hat{f}_n)^2)$$

are the same, the proof of the proposition indicates that the "asymptotics should take effect sooner" for $\mathrm{VE}_2(\hat{f}_n \to f)$, because there is one less approximation step in the proof of (4). This probably reflects different finite sample behavior of both error criteria discussed above.

Interesting future work involves trying to find data-based smoothing parameter selectors that attempt to optimize $\mathrm{VE}_2(\hat{f}_n \to f)$. One approach would be to use "plug in" ideas (as in Sheather and Jones 1991), based on the foregoing asymptotic representation. Another, which may track the performance of $\mathrm{VE}_2(\hat{f}_n \to f)$ more closely in small-sample situations, is a bootstrap estimate of $\mathrm{VE}_2(\hat{f}_n \to f)$ in a straightforward extension of the usual ideas (see, for example, Marron 1992).

5. CONCLUSION

We have investigated methods for measuring differences between curves that correspond much more closely to visual choice than to the usual norms on function space. Our final recommendation for error criteria comes in two parts:

1. When the goal is "the estimate should capture as many of the qualitative features of the data as possible," we prefer the criterion SE_2.
2. For the *different* goal of "duplicate the choice of an experienced data analyst," we prefer the criterion $\mathrm{VE}_2(\hat{f} \to f)$.

APPENDIX: PROOF OF THE PROPOSITION

This proof is based on analyzing the value x_n, the element of $G_{\hat{f}_n}$ that is closest to $(x_0, f(x_0))$, and the value x_n^*, the member of G_f that is closest to $(x_0, \hat{f}_n(x_0))$. In other words, define

$$x_n = \arg\min_x \|(x_0, f(x_0)) - (x, \hat{f}_n(x))\|_2$$

and

$$x_n^* = \arg\min_x \|(x, f(x)) - (x_0, \hat{f}_n(x_0))\|_2.$$

The key to the proof of (3) is to show that the point $(x_n, \hat{f}_n(x_n))$ lies in the shaded region in Figure A.1.

Let $s_n = |\hat{f}_n(x_0) - f(x_0)|$ denote the usual "vertical distance." Because $\hat{f}_n$ converges to f, we will show that it is enough to approximate the curve $\hat{f}_n(x)$ by the line $y = \hat{f}_n(x_0) + \hat{f}'_n(x_0)(x - x_0)$ for points near x_0. The error in this approximation, over points in the large circle in Figure A.1, is quantified by

$$\Delta_n = \sup_{x:|x-x_0| \le s_n} |\hat{f}_n(x) - \hat{f}_n(x_0) - \hat{f}'_n(x_0)(x - x_0)|. \quad (A.1)$$

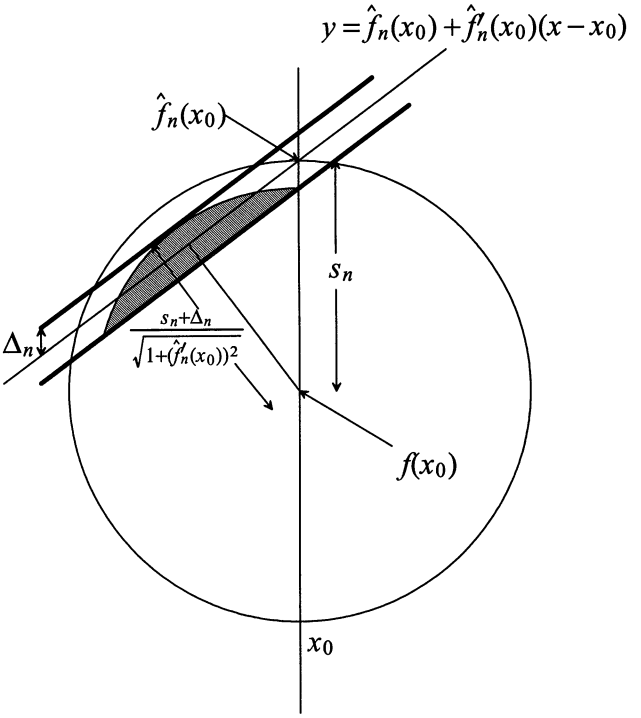

Figure A.1. Visual Representation of the Main Idea Behind the Proof of 3. The thin diagonal line is $y = \hat{f}_n(x_0) + \hat{f}'_n(x_0)(x - x_0)$. The radius of the shaded sector is $(s_n + \Delta_n)/\sqrt{1 + (\hat{f}'_n(x_0))^2}$.

The fact that the error Δ_n is asymptotically much smaller than the vertical distance, s_n, is demonstrated in the following lemma.

Lemma 1. Under assumptions A1–A6,

$$\Delta_n = O_p(s_n^2),$$

as $n \to \infty$.

The proof of the lemma is provided at the end of the Appendix.

Note that

$$|x_0 - x_n| \le |\hat{f}_n(x_0) - f(x_0)| = s_n,$$

from which it follows that $(x_n, \hat{f}_n(x_n))$ is inside the large circle in Figure A.1. Hence by the approximation (A.1), $(x_n, \hat{f}_n(x_n))$ must be between the heavy lines in Figure A.1. Now by a straightforward trigonometric argument,

$$d_n(x_0) \le \frac{s_n + \Delta_n}{\sqrt{1 + (\hat{f}'_n(x_0))^2}}, \tag{A.2}$$

and so $(x_n, \hat{f}_n(x_n))$ must be inside the shaded region in Figure A.1. It follows from this that

$$\frac{s_n - \Delta_n}{\sqrt{1 + (\hat{f}'_n(x_0))^2}} \le d_n(x_0). \tag{A.3}$$

Thus from the standard fact (see, for example, Stone 1980) that

$$\hat{f}'_n(x_0) \xrightarrow{p} f'(x_0) \quad \text{as} \quad n \to \infty,$$

it follows that with probability tending to 1 as $n \to \infty$,

$$\left| \frac{s_n}{\sqrt{1 + (f'(x_0))^2}} - d_n(x_0) \right| \le \frac{\Delta_n}{\sqrt{1 + (f'(x_0))^2}}, \tag{A.4}$$

with probability tending to 1 as $n \to \infty$. Hence, using (2), we see that $\Delta_n = o_p(s_n) = o_p(n^{-2/5})$, which entails (3).

To prove (4), we use a quite similar argument. Interchanging $\hat{f}_n$ and f, we find that (A.2) and (A.3) can be replaced by

$$\frac{s_n - \Delta_n^*}{\sqrt{1 + (f'(x_0))^2}} \le d_n^*(x_0) \le \frac{s_n + \Delta_n^*}{\sqrt{1 + (f'(x_0))^2}}, \tag{A.5}$$

where

$$\Delta_n^* = \sup_{x: |x - x_0| \le s_n} |f(x) - f(x_0) - f'(x_0)(x - x_0)|$$

$$= O(s_n^2) = o_p(n^{-2/5}),$$

as $n \to \infty$. This leads to (4). The weak indication that the convergence in (3) is faster than that in (4) comes from the fact that no approximation of the denominators is needed in (A.5) of the type that was used to go from (A.2) and (A.3) to (A.4).

Proof of the Lemma

Using (2), note that

$$\Delta_n = \sup_{x: |x - x_0| \le s_n} s_n^2 \left| \frac{1}{2} \int_0^1 \hat{f}''_n(x_0 + \theta(x - x_0))\, d\theta \right|$$

$$\le \frac{1}{2} s_n^2 \sup_{x: |x - x_0| \le s_n} |\hat{f}''_n(x) - f''(x)| + \frac{1}{2} s_n^2 \sup_{x: |x - x_0| \le s_n} |f''(x)|$$

$$= \frac{1}{2} s_n \sup_{x: |x - x_0| \le s_n} |\hat{f}''_n(x) - f''(x)| + O(s_n^2),$$

as $n \to \infty$. The lemma follows from the fact that

$$p_n = P\{n^{-2/5} \sup_{x: |x - x_0| \le Cn^{-2/5}} |\hat{f}''_n(x) - f''(x)| \ge \varepsilon / C^2\} \to 0$$

for every $\varepsilon > 0$ and $C > 0$, which is easily shown by standard methods.

[Received July 1993. Revised September 1994.]

REFERENCES

Baddeley, A. J. (1993), "Errors in Binary Images and an L^p Version of the Hausdorff Metric," unpublished manuscript.

Bookstein, F. L. (1986), "Size and Shape–Specific Landmark Data in Two Dimensions" (with discussion), *Statistical Science,* 1, 181–242.

Chui, C. K. (1992), *An Introduction to Wavelets,* New York: Academic Press.

Cuevas, A., and Gonzales Mantiega, W. (1992), "Data-Driven Smoothing Based on Convexity Properties," in *Nonparametric Functional Estimation and Related Topics,* NATO ASI Series C, Vol. 335 ed. Roussas, Amsterdam: Kluwer, pp. 225–240.

Devroye, L., and Györfi, L. (1985), *Nonparametric Density Estimation: The L^1 View,* New York: John Wiley.

Donoho, D. L., and Johnstone, I. M. (1992), "Ideal Spatial Adaptation by Wavelet Shrinkage," unpublished manuscript.

Eubank, R. L. (1988), *Spline Smoothing and Nonparametric Regression,* New York: Marcel Dekker.

Härdle, W. (1991), *Applied Nonparametric Regression,* Boston: Cambridge University Press.

Hall, P. (1987), "On Kullback–Leibler Loss and Density Estimation," *The Annals of Statistics,* 15, 1491–1519.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1992), "Progress in Data-Based Bandwidth Selection for Kernel Density Estimation," unpublished manuscript.

Kendall, D. G. (1989), "A Survey of the Statistical Theory of Shape" (with discussion), *Statistical Science,* 4, 87–120.

Kooperberg, C., and Stone, C. J. (1991), "A Study of Log Spline Density Estimation," *Computational Statistics and Data Analysis,* 12, 327–347.

Mammen, E. (1991), "Nonparametric Regression Under Qualitative Smoothness Assumptions," *The Annals of Statistics,* 19, 741–759.

Marron, J. S. (1988), "Automatic Smoothing Parameter Selection: A Survey," *Empirical Economics,* 13, 187–208.

——— (1992), "Bootstrap Bandwidth Selection," in *Exploring the Limits of Bootstrap,* eds. LePage and Billard, New York: John Wiley.

Marron, J. S., and Wand, M. P. (1992), "Exact Mean Integrated Squared Error," *The Annals of Statistics,* 20, 712–736.

Müller, H. G. (1988), *Nonparametric Analysis of Longitudinal Data,* Berlin: Springer-Verlag.

Nielsen, G. M., and Foley, T. A. (1989), "A Survey of Applications of an Affine-Invariant Norm," in *Mathematical Methods in Computer-Aided Geometric Design,* eds. T. Lyche and L. Schumaker, Boston: Academic Press.

Ripley, B. D. (1986), "Statistics, Images and Pattern Recognition," *Canadian Journal of Statistics,* 14, 83–111.

Sendov, B. (1990), *Hausdorff Approximations* (*Mathematics and Its Applications,* Vol. 50) Dordrecht: Kluwer.

Sheather, S. J., and Jones, M. C. (1991), "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society,* Ser. B, 53, 683–690.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis,* New York: Chapman and Hall.

Stone, C. J. (1980), "Optimal Rates of Convergence for Nonparametric Estimators," *The Annals of Statistics,* 8, 1348–1360.

Wahba, G. (1991), *Spline Models for Observational Data,* Philadelphia: SIAM.