Bandwidth-based Nonparametric Inference

David B. Hitchcock

University of South Carolina*

July 12, 2006

Abstract

Nonparametric curve estimation is an extremely common statistical procedure. While its primary purpose has been exploratory, some advances in inference have been made. This paper provides a critical review of inferential tests that make fundamental use of a key element of nonparametric smoothing, the bandwidth, to determine the significance of certain features. A major focus is on two important problems that have been tackled using bandwidth-based inference: testing for the multimodality of a density and testing for the monotonicity of a regression curve. Early research in bandwidth-based inference is surveyed, as well as recent theoretical advances. Possible future directions in bandwidth-based inference are discussed.

Key words and phrases: Bump hunting; Density estimation; Monotone non-

*David B. Hitchcock is Assistant Professor, Department of Statistics, University of South Carolina, Columbia, South Carolina 29208 (e-mail: hitchcock@stat.sc.edu).

parametric regression; Multimodality; Smoothing.

1 Introduction

A major research area that has blossomed with the advent of the computer age is the study of automated smoothing methods, such as density estimation and nonparametric regression. These methods have primarily been viewed as descriptive in nature, able to summarize characteristics of the sample data in a flexible and innovative manner, free from many restrictions imposed by parametric approaches. The nature of nonparametric smoothing, however, renders formal inferential procedures difficult.

Some progress has been made in developing confidence bands for nonparametric regression curves. See, for example, Eubank and Speckman [12] and Xia [55] and the many references therein, as well as an extended discussion on the limitations of such methods in Chaudhuri and Marron [3]. On the other hand, hypothesis tests about nonparametric curve estimates often involve questions quite separate from those resolved using confidence bands. Two major questions of interest come to mind.

When we estimate a density, it is important to determine whether the density is multimodal, and in particular, we may wish to determine how many modes, or "bumps," the density has. Aside from the mathematical interest in the relative maxima of the density, the various bumps often represent distinct subpopulations within the population under study. Whether these bumps reflect genuine groupings or simply sampling artifacts is an important data-analytic question. (While the terms *mode* and *bump* are used interchangeably here to mean a relative maximum, note that

some authors include "shoulders"—points that are not relative maxima but where that function has slope zero—in their definition of *bump*.)

In a regression problem, we wish to estimate the relationship between the mean of a response variable Y and an explanatory variable X. In nonparametric regression, this relationship is typically represented by a smooth curve. Note that if such a curve has no bumps, it may (loosely) be labeled as monotone. So it is natural that a test for modes, or bumps, could be adapted to a test for the monotonicity of a regression relationship. Again, this is often an issue of interest: in dose-response studies, the response Y may be expected to vary monotonically with a dosage level X of a drug, say, and testing for deviations from monotonicity may be tantamount to testing whether the drug is affecting recipients abnormally.

Statistical researchers have offered various methods to solve these problems. For example, Cox [7] and Good and Gaskins [21] proposed tests for multimodality of densities. Since then, many similar tests have been proposed. In this paper we focus on inferential methods that rely on a fundamental aspect of many nonparametric curve estimation techniques: namely, the bandwidth.

The bandwidth plays a prominent role in the "kernel" methods of density estimation or regression. The kernel density estimate (kde) of a (univariate) density f_X, based on n sample observations $X_1, \ldots, X_n$, is defined pointwise for each point t:

$$\hat{f}(t; h) = (nh)^{-1} \sum_{i=1}^{n} K\{h^{-1}(t - X_i)\},$$

where the *kernel function* $K(\cdot)$ is often chosen to be a common symmetric density function. (See Schucany [48] for an excellent detailed introduction to kernel smoothing.) At any point, the ordinate of the density estimate is a sum of several densities'

ordinates. The bandwidth h (also called the window width) represents the spread of these kernel densities. Note that a large h implies that an observation x_j can have a nonnegligible effect on the kde value at a point quite distant from x_j. For large h, more observations contribute meaningfully to the kde value at each t. This leads to lower pointwise variability of the density estimator and tends to produce a smoother estimate. On the other hand, a small bandwidth leads to a "bumpier" estimate which is more reflective of the vagaries of the noisy data.

The same principle holds for the kernel regression estimate (kre) based on paired data $x_1, \ldots, x_n$ and $Y_1, \ldots, Y_n$, defined pointwise as:

$$\hat{m}(t; h) = \sum_{i=1}^{n} Y_i h^{-1}(x_i - x_{i-1}) K\{h^{-1}(t - x_i)\}.$$

(This is the Priestley-Chao [45] kernel estimator; related kre's include that of Nadaraya [44] and Watson [53] and that of Gasser and Müller [17].) Again, a large bandwidth tends to produce a smooth, regular regression curve, while a small h leads to a wiggly curve that closely follows the raw data when overlain on a scatterplot.

Once we recognize the role the bandwidth plays in determining the shape of the kernel estimate, we see how it could be used for inference. If the major question is, *Is the true nature of the data "bumpy" or "smooth"?*, then one way to answer this is to determine how much an ostensibly appropriate bandwidth must be increased to produce a "smooth" curve estimate. The applications of bandwidth-based inference have each used variations on this idea.

Thinking broadly, a major principle in bandwidth-based inference is that, in a hypothesis-testing context, a particular bandwidth value defines the boundary of a null region. (The particular definition of that null region depends on the situation

being analyzed.) The important idea is that varying bandwidths lead to data models that imply different states of nature. The extremity of the particular bandwidth that puts the model in the null region says something, loosely, about the likelihood that the state of nature defined by the null hypothesis is correct. Therefore, bandwidth-based inference, while it has been focused on two major data-analytic situations, could be applied more generally to any situation in which the bandwidth value influences our perception of the state of nature.

2 Inference for Multimodality

The seminal paper on bandwidth-based inference was Silverman's [50] short article proposing a test for multimodality. Silverman noted that if the kernel chosen was a normal density function, then the number of modes in the kde was a right-continuous decreasing function of h. Hence, in testing the null hypothesis that the true density f has at most k modes against the alternative that f has more than k modes, one could use what Silverman called the *critical bandwidth* h_{crit}: the smallest bandwidth yielding a kde with at most k modes. (This is well-defined if the number of modes is a monotone function of the bandwidth, as is the case if the kernel is normal.) Note that h_{crit}, then, is the bandwidth that places the density estimate exactly on the boundary between the null and alternative regions. Since the number of modes decreases as h increases, then $\hat{f}(\cdot; h)$ has more than k modes if and only if $h < h_{crit}$. This makes the search for h_{crit} a simple matter; one performs a binary search across a grid of h values until finding the "boundary" bandwidth yielding $\hat{f}$ having k modes, but such that any smaller bandwidth would yield $k + 1$ modes.

The critical bandwidth serves as a test statistic. A large value of h_{crit} favors the alternative, since this implies a great deal of smoothing is needed to reduce the number of modes to k. Let h'_{crit} be the particular value of the critical bandwidth that is obtained from the sample data. Employing the bootstrap, one might approximate the null distribution of h_{crit} by generating bootstrap samples based on the "null estimate" $\hat{f}(\cdot; h'_{crit})$, and calculating the bootstrap-data h^*_{crit} each time. Then the p-value of the test would be the proportion of the h^*_{crit} values that exceed h'_{crit}.

One can cleverly avoid the computational expense of finding h^*_{crit} for each of the many bootstrap data sets by defining the p-value as follows. Apply the density estimator with bandwidth h'_{crit} to each of the bootstrap data sets and let the p-value be the proportion of the resulting density estimates with more than k modes. This is equivalent to finding the proportion of h^*_{crit} values exceeding h'_{crit} because, for a particular bootstrap sample, $P(h^*_{crit} > h'_{crit})$ equals the probability that h'_{crit} is not large enough to force $\hat{f}(t; h)$ to have k or fewer modes when fit to that bootstrap data set. Hence a bootstrap data set requires $h^*_{crit} > h'_{crit}$ if and only if $\hat{f}(t; h'_{crit})$ has more than k modes when applied to that bootstrap data set.

Silverman used the now-familiar principle of the bootstrap—which at that time had been fairly recently proposed by Efron [11]—to approximate the sampling distribution of the test statistic and thus determine the significance of h'_{crit}. When one obtains the bootstrap samples, Silverman recommended rescaling the critical-bandwidth kde $\hat{f}(t; h'_{crit})$ to have the same variance as the sample variance of the data. He gave the formula for deriving the "scaled" bootstrap data as:

$$Y^*_i = [1 + (h'_{crit}/s)^2]^{-1/2}(X^*_i + h'_{crit}\epsilon_i)$$

where the X_i^*'s are sampled with replacement from $X_1, \ldots, X_n$, s^2 is the sample variance of the data, and ϵ_i are independent standard normal random variables. Silverman noted that this rescaling yielded a *smoothed bootstrap* procedure that sampled from a density on the "boundary" of the composite null hypothesis "f has at least k modes" and thus honestly assessed the significance of h'_{crit}.

Formally, the null hypothesis in Silverman's test specifies a specific number of modes k. In practice, however, the investigator often does not have a clear idea how many modes a density has and may sequentially run the test for several k values. The sequence of null hypotheses might be $H_0 : k = 1$; $H_0 : k \leq 2$; and so on, with the procedure stopping when one of the null hypotheses is not rejected. (For example, if $H_0 : k = 1$ is rejected, and then $H_0 : k \leq 2$ is not rejected, the conclusion would be that the population is bimodal.) The resulting p-values do not necessarily form a monotone sequence, however, and the issue of a precise stopping rule is unresolved, as are any pertinent multiple-testing concerns.

Izenman and Sommer [35] employed Silverman's test to determine the number of modes of a distribution of stamp thickness measurements. In addition to presenting this data analysis, the paper includes an extensive, detailed discussion of Silverman's test, including incisive comments on the issue of a stopping rule when considering a sequence of k values. In the stamp-thickness analysis, they suggested adjusting Silverman's procedure via an adaptive-bandwidth method. For the stamp-thickness data, several small, possibly spurious, modes in the tails of the density—where few data values were observed—were declared significant by the Silverman test. To reduce the impact of regions where data were sparse, Izenman and Sommer suggested varying

h, making h small in regions dense with data and large in regions sparse with data. In that case, some type of average critical bandwidth $\bar{h}_{crit}$ could serve as the test statistic. Exactly how this would be implemented would depend on the adaptive procedure, and it remains an area for further research.

Some other applications of Silverman's test include Segal and Wiemels [49], in a clustering problem, and Bianchi [1], in an analysis of econometric data.

Though the Silverman test is heavily cited and has been called "ingenious" ([15], p. 499), some drawbacks have been pointed out. The test seems to be conservative—possible reasons for which were discussed by Silverman [51]—and the test makes no distinction between tiny modes and large, "important" modes in determining modal significance. Fisher and Marron [15] suggested an alternative method that addressed these drawbacks, and was more resistant to sample outliers. They modified the critical bandwidth concept, synthesizing Silverman's approach with another proposed test for multimodality ([43]; see also [32]), which did not rely on a bandwidth as a test statistic.

To identify "minor" modes, Fisher and Marron used a measure of "continuous excess mass," an idea modified from Müller and Sawitzki [43]. The excess mass E_j, for each mode j in the density estimate, is the area under the curve bounded vertically by the peak of the bump and the nearest local minimum. When any of a pair of adjacent modes has excess mass below a threshold m_0, they should be combined, forming one mode with larger excess mass. In addition, small isolated bumps can be eliminated if their excess mass is less than another threshold λ_0. To test the hypothesis

$$H_0 : m \leq k \text{ vs. } H_1 : m > k \tag{1}$$

where m is the true number of modes of f, an appropriate test statistic is based on S_k, the sum of the excess masses other than the k largest E_j's. Beginning with the bandwidth large enough to yield a unimodal kde and then decreasing h, Fisher and Marron defined a generalization of Silverman's critical bandwidth:

$$h_k = \sup[h : S_k > 0].$$

Testing a particular k and for a given m_0 and λ_0, this is the largest bandwidth such that S_k is positive, and reduces to Silverman's h_{crit} when $m_0 = 0$ and $\lambda_0 = 0$.

The parameters m_0 and λ_0 affect the test's sensitivity to small bumps. A value of m_0 near 0 will be sensitive to small modes adjacent to other modes and more often declare them "significant"; a value of λ_0 near 0 will be sensitive to small isolated modes—often associated with sample outliers—and declare them significant. Increasing m_0 or λ_0 reduces the respective form of sensitivity and typically results in fewer modes being deemed significant. A proper choice of parameter values depends on the investigator's needs: Is the identification of minor subpopulations or outlying observations a goal of interest? Furthermore, while increasing m_0 can lead to a more accurate size, it may also decrease power since H_0 is less likely to be rejected. Fisher and Marron provide some guidelines for choosing m_0 and λ_0.

While h_k could serve as a test statistic, Fisher and Marron recommend adjusted statistics which are more resistant to outliers. Significance is determined via a bootstrap approach. In addition, they propose a variation of the test statistic designed to test for multimodality of circular data.

In a similar vein, the method of Minnotte [41] is something of a mix of two approaches: It uses the critical bandwidth idea of Silverman [50], but the test statistic

is based on probability masses of bumps in the density estimate, rendering it closer in spirit to Fisher and Marron [15] and Müller and Sawitzki [43].

Fisher, Mammen and Marron [14] took issue with Silverman's rescaling of the bootstrap sample values. For example, for multimodal f, this rescaling alters the relative location of the modes. Another means of rescaling follows from the property that the kernel estimator is scale-invariant; hence "rescaling the bootstrap observations is equivalent to rescaling the bandwidth," noted Fisher, Mammen and Marron ([14], p. 503). Thus an appropriate p-value is the proportion of bootstrap-sample critical bandwidth values exceeding Rh'_{crit}, where R is a general rescaling factor. Fisher, Mammen and Marron provide guidelines for choosing R so that Rh'_{crit} provides the same relative amount of smoothing as the h_{crit} value for the bootstrap sample.

3 Inference for Monotonicity

As mentioned previously, bumps (departures from monotonicity) in regression curves are analogous to bumps (departures from unimodality) in density curves. An analogue of Silverman's test was introduced by Bowman, Jones and Gijbels [2] to test monotonicity of a regression function. As with the Silverman procedure, the test statistic is a critical bandwidth h_{crit} that puts the curve estimate at the border of the null region (here this means monotonicity), and the p-value is calculated via a bootstrap method that approximates the null distribution of the test statistic.

When we use a kernel regression estimator such as the Gasser-Müller or Priestley-Chao estimator with a normal kernel, the property of monotonicity is in fact a monotone function of the bandwidth: As h decreases, the number of bumps in the estimated

curve can only increase. Bowman et al., however, employ the local linear estimator

$$\hat{m}(t;h) = \hat{\beta}_0,$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize

$$LSSE = \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1(x_i - t)]h^{-1}K[(x_i - t)/h],$$

since it has desirable behavior at the boundaries of the X region, and since as $h \to \infty$, the local linear estimator approaches a straight line, the quintessential monotone function estimate. Since the local linear estimator does not share the "monotonicity of monotonicity" property (such departures are infrequent), a slightly altered definition of the critical bandwidth is needed: h_{crit} is here the smallest bandwidth that yields a monotone regression curve, *even if larger bandwidths yield a nonmonotone curve.* This produces a slightly conservative test.

The nature of the bootstrap procedure in the regression situation is somewhat different in that the bootstrapped values are here residuals. Initially a reasonable set of residuals (from which to resample) must be obtained, and this problem reduces to finding a reasonable estimate of the underlying regression curve. Bowman et al. select the local linear curve estimator with the plug-in bandwidth selector of Ruppert, Sheather and Wand [47]. Hence, their algorithm is: (1) find the critical bandwidth h_{crit}; (2) Obtain error estimates $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$ where $\hat{\epsilon}_i = Y_i - \hat{m}(X_i; h_0)$ and h_0 is the plug-in bandwidth; Generate the bootstrap sample $\hat{\epsilon}_1^*, \ldots, \hat{\epsilon}_n^*$ by resampling from $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$, and form the bootstrap data set by adding the bootstrap residuals to the curve estimate that is *just* monotone:

$$Y_i^* = \hat{m}(X_i; h_{crit}) + \hat{\epsilon}_i^*, \quad i = 1, \ldots, n;$$

(4) Calculate $\hat{m}(\cdot|h_{crit})$ based on $\{(X_i, Y_i^*), i = 1, \ldots, n\}$ and observe whether the resulting curve estimate is monotone (in practice this is done by discretizing $\hat{m}$ along a fine grid of x-values); (5) Repeat (3) and (4) many times, with the p-value being the proportion of nonmonotone results.

As with Silverman's test, one is saved from having to compute h_{crit} at each bootstrap iteration to determine the proportion of times $h_{crit}^* > h_{crit}^{'}$: This event holds if $h_{crit}^{'}$ is not large enough to make $\hat{m}$ monotone when fit to the bootstrap data, which means that $\hat{m}(\cdot|h_{crit})$ based on $\{(X_i, Y_i^*), i = 1, \ldots, n\}$ is nonmonotone.

Bowman et al. provide a neat heuristic argument for the consistency of the test, in the sense that the power $\to 1$ (under H_1) as $n \to \infty$: Under H_1 ("m is nonmonotone"), h_{crit} converges to a positive number as $n \to \infty$, since to produce a monotone $\hat{m}$ from a nonmonotone m, we need a positive amount of smoothing (i.e., $h_{crit} > 0$). Assuming H_0 for the purpose of generating the bootstrap data, the theoretical "best" bandwidth (approximated by the plug-in h_0) will asymptotically produce a monotone $\hat{m}(\cdot)$. Then by its definition, $h_{crit}^* \leq h_0$; and since as $n \to \infty$, the optimal bandwidth tends to 0, then as $n \to \infty$, $h_{crit}^* \to 0$. Because h_{crit} converges to a positive number, $P[h_{crit}^* \leq h_{crit}] \to 1$ as $n \to \infty$ under H_1, implying consistency.

The test for monotonicity has been naturally extended by Harezlak and Heckman [30], who called their methodology CriSP (short for Critical Smoothing Parameter), to a test of the number of bumps in a regression curve. The null hypothesis that there are k or fewer bumps in the true regression function $m(x)$ (or possibly a derivative of m) is tested against the alternative that the number of bumps exceeds k. In practice, the test would be successively carried out for $k = 0, 1, 2 \ldots$. Like Silverman, Harezlak

and Heckman do not offer precise significance level adjustments to account for this sequential testing, but do suggest that performing the tests with α between 0.10 and 0.20 gives good results.

While Bowman et al. focused on local linear estimation of m, Harezlak and Heckman consider both local linear and L-spline (a generalization of smoothing splines; see below) estimation methods. They also address in uncommon detail the issue of exactly what constitutes a "bump." The algorithm of Bowman et al. judged a curve estimate as having a bump if there were any deviations from monotonicity along a grid of 50 points, but Harezlak and Heckman define an l-bump to be a point on a discretized curve larger than any points plus or minus l positions away, a more stringent definition of a bump.

On the other hand, bandwidth-based tests for a regression curve's monotonicity have their detractors: Hall and Heckman [24] pointed out that such methods can have undesirable properties when the true density is flat or nearly flat. For an underlying regression function having a small dip, even as $n \to \infty$, the bandwidth-based test is not asymptotically guaranteed to detect this nonmonotonicity. Like Fisher and Marron [15] in the density estimation problem, Hall and Heckman introduce a monotonicity test related to Müller and Sawitzki [43] in that it does not directly employ the bandwidth for inference.

Rather than testing a null hypothesis of monotonicity, Huang [34] proposed a test for whether the regression function is linear. Although linearity is a specific case of monotonicity, Huang's test used a slightly different bandwidth-based method than Bowman et al. Here, the critical bandwidth is the h producing a regression curve

which is on the boundary between H_0^* and H_1^*, where these are the hypotheses of the F-test for linearity of Hastie and Tibshirani [33]:

$$H_0^* : m(x) \text{ linear vs. } H_1^* : m(x) \text{ is a smooth nonlinear function.} \tag{2}$$

That is, h_{crit} by definition is the smallest bandwidth that produces an estimate whose p-value in the F-test is α. (The modifier "smallest" is required, again, because this p-value is not necessarily a monotone function of bandwidth.) Though similar in conception to the Bowman et al. test, there is a notable difference in implementation, since the critical bandwidth need not be determined via a grid-check along the estimated curve. Again, the null distribution of the test statistic is determining by a bootstrap method. As with the tests of Silverman and Bowman et al., calculation of the test statistic for every bootstrap sample is unnecessary; one merely checks at each bootstrap iteration whether the original h'_{crit} yields an F-statistic larger when applied to the bootstrap sample than when applied to the original sample. Simulations show that, like the other two procedures, Huang's test is slightly conservative.

A novel bandwidth-based approach, notable for innovative color maps that summarized the inferential conclusions about the curve's important features, was introduced by Chaudhuri and Marron [3]. Their SiZer method (the name being a contraction of "Significant Zero Crossing of Derivatives"), is appropriate for both density estimation and nonparametric regression. Rather than focusing on inference regarding a "true" underlying curve, Chaudhuri and Marron consider a plethora of visions of curves, depending on the bandwidth. A small bandwidth produces a small-scale, zoomed-in picture of the curve, while a large bandwidth yields a broad-scale "distant" image. Whether features (modes of densities or turning points in regression

curves) are significant depends not only on the data but also through which bandwidth the data are "viewed." The SiZer map is a two-dimensional pictorial array, with the horizontal axis paralleling the X (location) axis of the curve and the vertical axis representing a range of bandwidths. The map is color-coded so that at each X-location, and for each bandwidth, the display is red when the curve has been judged to be significantly decreasing, blue when the curve is significantly increasing and purple when the slope is not significantly different from zero. (Gray denotes regions where the data are too sparse for any sort of judgment.) Thus the purple areas represent points of interest: density modes or turning points.

Aside from the innovative presentation, SiZer requires some methodological determination of whether particular features are significant. While pointing out certain weaknesses of confidence-bound methods in nonparametric regression, some of which are "grossly invalid because of bias problems" (p. 811), Chaudhuri and Marron use a confidence-limit procedure as a basis for significance tests. The main difference is that while other intervals purport to estimate (the derivative of) the underlying curve $f(t)$—typically using a biased curve estimator—the SiZer intervals estimate the expected value of (the derivative of) the bandwidth-specific curve estimator $E[\hat{f}'(t;h)]$. By disregarding the notion of a true regression curve and making everything depend on h, the problem of bias is sidestepped. The Chaudhuri-Marron confidence limits are

$$\hat{f}'(t;h) \pm q \cdot \widehat{SD}[\hat{f}'(t;h)]$$

where q is a quantile chosen by a normal approximation or bootstrap method as appropriate for simultaneous inference. The standard deviation term is approximated

through a binning procedure of Fan and Marron [13].

One advantage of SiZer is that it explicitly identifies feature locations rather than simply the number of features as previous tests did; however, a user of, say, Silverman's test could fairly easily identify the "significant" feature locations with a quick plot. The more important novelty of the SiZer approach is that it presents information about significant features for many different bandwidths at once. The idea that different bandwidths yield different, possibly equally valid, views of the data, instead of forcing the initial choice of one "correct" bandwidth, is an interesting one. The development of SiZer has spurred several recent contributions, including [27], [28], [37], and [39].

The SiZer graph is a close relative of the *mode tree* of Minnotte and Scott [42], which also plots, in a treelike hierarchical graph, mode locations along a horizontal axis, with bandwidth values on a vertical axis. In fact, some of the varying-bandwidth perspective of Chaudhari and Marron had been expressed by Minnotte and Scott; the mode tree also contains some of SiZer's inferential flavor. Whereas SiZer uses different colors to identify significant or nonsignificant modes, Minnotte and Scott place filled circles and open circles, respectively, at the "nodes" of the mode tree to indicate such. Significance is determined through a test similar to that of [41].

Note that these methods could be, in principle, applied to smoothing methods that employ a continuous smoothness parameter other than a bandwidth. Two well-known examples are the *span*, a key element in lowess and loess methods [5, 6], and the smoothing parameter of the smoothing-spline method described in detail in [52] and [22], which is a type of global nonparametric regression method. The span is the

fraction of the data used at each point to estimate the curve; the larger the span, the smoother (less wiggly) the curve estimate. The smoothing-spline parameter λ penalizes the roughness of a curve estimate; in the regression situation the fitted curve minimizes the penalized residual sum of squares

$$P = \sum_{j=1}^{n} [y(t_j) - \hat{m}(t_j)]^2 + \lambda \int [\hat{m}''(s)]^2 \, ds.$$

A higher value of λ yields a smoother estimated curve.

Some work has been done in this area: Harezlak and Heckman [30] investigated the use of a critical λ-value as the test statistic, finding that their test performed about as well empirically as it did when the kernel-based bandwidth was used. Wong [54] proposed a relative of Silverman's test that used a kth-nearest-neighbor density estimator; the critical k value (k is the span times n) serves as the test statistic. While these quantities have similar purposes as bandwidths, it is probably more difficult to develop theoretical properties of these alternative tests, and the main theoretical advances have all focused on a "critical bandwidth" in traditional kernel-based curve estimators.

Other approaches to testing for monotonicity, which rely on more advanced stochastic processes theory, include those of Ghosal, Sen and van der Vaart [18] and Dümbgen and Spokoiny [10], the latter of which also addresses multimodality testing. While the test statistics in these methods involve bandwidths, they use the bandwidth much less directly than do the aforementioned articles.

4 Theoretical Investigations of Bandwidth-based Inference

Since bandwidth-based inference is such a computationally intensive procedure, the amount of investigation of the theoretical properties of Silverman's test and its successors serves as an indication of their elegance and attractiveness.

One of the major theoretical contributions in the general area of modality testing was that of Donoho [9]. His result dealt with a general class of functionals of a density, one example of which was the number of modes. Donoho proved that the only reasonable confidence statements one could make about such functionals were one-sided; furthermore, upper confidence bounds on, say, the number of modes were not possible. This had the effect of specifying the relevant test of modality to be (1); testing $H_0 : m \geq k$ against $H_1 : m < k$ would be unreasonable.

The initial theoretical investigation of the method of Silverman was given in [51] and concerned the rate of convergence of h_{crit} as the sample size n tends to ∞. Silverman proved that under the null hypothesis, h_{crit} converges in probability to zero, a convergence that does not occur under the alternative, implying that the test is consistent.

Mammen, Marron and Fisher [40] derived an asymptotic formula for the expected number of modes of a kernel density estimator. They thus proved the rate of convergence of Silverman's test statistic h_{crit} (as n tends to ∞). Under regularity conditions mirroring those of [51], they showed h_{crit} is of order $n^{-1/5}$, correcting a result given in [51]. The bootstrap-sample critical bandwidth h^*_{crit} is also of order $n^{-1/5}$, suggesting

that the probability of rejecting the null is asymptotically bounded away from 0 [40]. While [40] assumed a "regular" f having neither flat intervals nor "shoulders," others have considered a more general case, as we now explain.

In mode-testing, as in so many decision-theoretic setups, there is a fine line between the null hypothesis and the alternative. While one can easily conceptually distinguish between a unimodal density and a bimodal density, what about the case in which f has a "shoulder"? While this f falls into the "null" class of unimodal f, it only just does; Cheng and Hall [4] consider the performance of Silverman's test (and that of [43]) in this "boundary" situation. When the true f is on the boundary of the null, the critical bandwidth is of order $n^{-1/7}$. Hence theory and methods designed for the classic situation, argue Cheng and Hall, may fail when f is on the boundary, or even close to it. Numerical studies show that "classic" methods are anticonservative, rejecting H_0 too often, when the true density is on the boundary; on the other hand, methods calibrated for the boundary case are conservative if f is classically unimodal [4].

Silverman [51] noted that the exact size of his modality test is typically lower than its nominal level (even as $n \to \infty$), and Hall and York [26] investigated this conservatism both theoretically and numerically. Deriving the null distribution of the critical bandwidth, they used this result to suggest a calibration to correct the size. Rather than taking the p-value to be the proportion of times that h^*_{crit} exceeds h'_{crit}, they proposed using the proportion of times that $h'_{crit}/h^*_{crit} \leq \lambda_\alpha$, where the number λ_α depends on the level α rather than identically equaling 1.

Because of its convenient property that the number of modes is a monotone func-

tion of the bandwidth, the normal kernel is the customary choice when applying Silverman's test. Compactly supported kernels, however, are a common choice in curve estimation, and Hall, Minnotte and Zhang [25] showed that although the monotonicity property fails in the most popular compact kernels, the impact on the modality test is often minor. In particular, they examined the Epanechnikov (uniweight), biweight, and triweight kernels. Both the biweight and triweight kernels are "safe" in that nonmonotonicity of the number of modes does not tend to appear near h_{crit}, although the Epanechnikov kernel does lead to nonmonotonicity near h_{crit}. Furthermore, the level of the test (which is known to be asymptotically conservative) is relatively unaffected by the choice of kernel [25].

While many of these theoretical discoveries at first glance seem to diminish or poke holes in the seminal method of [50] and its subsequent parallels, that is an unfair perspective. Silverman's test was a cleverly innovative, approximately correct procedure that has desirable properties in terms of consistency, though not necessarily optimality. While further research has pointed out deficiencies and made improvements in the Silverman test, the fundamental idea remains unassailable. The volume of scrutiny afforded the procedure ultimately compliments it.

5 The Future of Bandwidth-based Inference

With the plethora of methodological and theoretical articles already published on bandwidth-based inference, one reasonable question to ask is whether the research area has run its course. Perhaps in terms of dealing with the two main problems of interest, the modality and monotonicity questions, we are nearing that point. Recent

approaches to testing for monotonicity [19, 20] have foregone the bandwidth-based methods and even stated this lack of reliance on a critical smoothing parameter as an advantage of their new methods. On the other hand, a potential influx of new researchers could solve new variations on these problems using bandwidth-based techniques. More broadly, the next step in bandwidth-based inference is to move beyond the two familar problems and endeavor to find solutions to more complex problems using bandwidth-based approaches. Some possible directions are outlined in this section.

Extensions of bandwidth-based testing to other problems depend on the ability to frame statistical questions in a particular way. As mentioned in the introduction, null and alternative regions must be specified such that different choices of bandwidth yield pictures of the data reflecting varying degrees of support of (or opposition to) the null state of nature. This methodology has been carefully developed in the modality and monotocity issues, but it could be applied to several other statistical problems.

In fact, the previously mentioned work of Huang [34] indeed employs this methodology to test for linearity. An immediate extension would be to test for other specified functional forms for $m(x)$, or even for the additivity of a nonparametric regression function (see [8]). Estimating the functions of additive models may involve several bandwidths, complicating the problem, but bandwidth-based solutions may have potential.

Another obvious extension of the work detailed in this paper is to consider the multidimensional analogues of the two problems. In the regression context, we might consider the detection of (and even the rigorous definition of) nonmonotonicity for a

regression function of multiple independent variables. Testing for the multimodality of the density of a multidimensional random vector has important practical considerations. With data mining techniques such as model-based clustering via mode-finding growing in popularity, bandwidth-based methods could provide an opportunity for formal inference in that direction. While the well-known "curse of dimensionality" limits the effectiveness of kernel-based techniques for high-dimensional data, statisticians are increasingly encountering practical problems with large numbers of variables. Discovering computationally feasible ways to adapt bandwidth-based inference to high-dimensional data would certainly expand its utility.

Most, if not all, of the development of bandwidth-based inference has dealt with a single density curve or regression curve: that is, there is assumed to be one sample of interest. In recent years, the field of *functional data analysis* (FDA) has grown rapidly; typically in FDA, the data set at hand consists of many curves (which may be densities or regression curves, depending on the application). In FDA, nonparametric curve estimation is of fundamental importance, since the observed data curves are generally smoothed via some nonparametric method. Often such smoothing methods involve a bandwidth (or other smoothing parameter) so that each smoothed curve in the functional data set has an associated bandwidth. Since the collection of these bandwidths characterizes features of the functional data sample, the procedures that have aided single-curve analyses may, in future years, be extended to inferential issues arising in multiple-curve analyses.

Gasser, Hall and Presnell [16] discuss the concept of a mode of a collection of curves. The bandwidth plays an important role in the identification of the "modal

curve," and potential inference about the mode might involve the bandwidth value. Jones and Rice ([36], p. 141), presenting a principal components approach to identify representative curves in a sample of functional data, note that "certain features of [the curves] might be of particular interest and an ordering, and hence a selection of curves, might be based on such features (e.g., roughness of density estimates)." Whether or not the authors have inference in mind here, it is obvious the the bandwidths exemplify the type of informative features of the curves that could be used for further analysis. Since Jones and Rice point out the effect that varying the bandwidth has on the principal component scores of the data set (and these scores determine the "representative" curves), it is not a great stretch to envision inference about these representative curves being based on a set of bandwidths. In another recent example, Harezlak, Naumova and Laird [31] have extended the CriSP method [30] to detect local extrema in the mean regression curve for longitudinal data.

Tests for the equality of two regression curves have appeared in the literature [23, 29, 38]. Ramsay and Silverman ([46], chapter 9) suggest functional ANOVA-type procedures for testing the equality of curves from several groups. While much work is needed to rigorously apply bandwidth-based methods to even these simple multi-curve situations, a broad strategy for many of these FDA situations would be to define a null state that unifies the entire set of curves in some sense. Then the set of default bandwidths arising from fitting each curve separately may be compared with a kind of critical bandwidth resulting from fitting the curves under the imposed unifying constraint.

On a related note, much new research in nonparametric curve estimation has

foregone a single-bandwidth approach in favor of adaptive schemes that allow the bandwidth to vary in different regions of the curve. Typically the bandwidth varies according to the density of the data or the steepness of the curve in each region. Bandwidth-based inference must account for these new approaches, since the traditional notion of a single critical bandwidth would not apply in the varying-bandwidth scheme. Izenman and Sommer's suggestion of a critical "average bandwidth" is a step in this direction, but much further work is needed.

Because nonparametric curve estimation provides a flexible framework for data analysis, the desire for related inferential methods will likely continue. Bandwidth-based methods have provided useful avenues for such inference, many of which are discussed in this paper. Ultimately, the future of bandwidth-based inference depends on the ability of researchers to adapt these approaches to the complex issues arising in modern smoothing and functional data analysis.

Acknowledgments

The author is grateful to an associate editor and two referees for their helpful comments which improved this paper.

References

[1] M. Bianchi, Testing for convergence: Evidence from non-parametric multimodality tests. J. Appl. Econometrics 12 (1997) 393-409.

[2] A.W. Bowman, M.C. Jones, I. Gijbels, Testing monotonicity of regression. J. Comput. Graph. Statist. 7 (1998) 489-500.

[3] P. Chaudhuri, J.S. Marron, SiZer for exploration of structures in curves. J. Amer. Statist. Assoc. 94 (1999) 807-823.

[4] M.-Y. Cheng, P. Hall, Mode testing in difficult cases. Ann. Statist. 27 (1999) 1294-1315.

[5] W.S. Cleveland, Robust locally weighted regression and smoothing scatterplots. J. Amer. Statist. Assoc. 74 (1979) 829-836.

[6] W.S. Cleveland, S.J. Devlin, Locally weighted regression: An approach to regression analysis by local fitting. J. Amer. Statist. Assoc. 83 (1988) 596-610.

[7] D.R. Cox, Notes on the analysis of mixed frequency distributions. Brit. J. Math. Statist. Psychol. 19 (1966) 39-47.

[8] S. Derbort, H. Dette, A. Munk, A test for additivity in nonparametric regression. Ann. Inst. Statist. Math. 54 (2002) 60-82.

[9] D.L. Donoho, One-sided inference about functionals of a density. Ann. Statist. 16 (1988) 1390-1420.

[10] L. Dümbgen, V.G. Spokoiny, Multiscale testing of qualitative hypotheses. Ann. Statist. 29 (2001) 124-152.

[11] B. Efron, Bootstrap methods—another look at the jack-knife. Ann. Statist. 7 (1979) 1-26.

[12] R.L. Eubank, P.L. Speckman, Confidence bands in nonparametric regression. J. Amer. Statist. Assoc. 88 (1993) 1287-1301.

[13] J. Fan, J.S. Marron, Fast implementations of nonparametric curve estimators. J. Comput. Graph. Statist. 3 (1994) 35-56.

[14] N.I. Fisher, E. Mammen, J.S. Marron, Testing for multimodality. Comput. Statist. Data Anal. 18 (1994) 499-512.

[15] N.I. Fisher, J.S. Marron, Mode testing via the excess mass estimate. Biometrika 88 (2001) 499-517.

[16] T. Gasser, P. Hall, B. Presnell,Nonparametric estimation of the mode of a distribution of random curves. J. Roy. Statist. Soc. Ser. B. 60 (1998) 681-691.

[17] T. Gasser, H.-G. Müller, Kernel estimation of regression functions. Smoothing Techniques for Curve Estimation. Lecture Notes in Math. 757 (1979) 23-68. Springer, Heidelberg.

[18] S. Ghosal, A. Sen, A.W. Van Der Vaart,Testing monotonicity of regression. Ann. Statist. 28 (2000) 1054-1082.

[19] I. Gijbels, P. Hall, M.C. Jones, I. Koch, Tests for monotonicity of a regression mean with guaranteed level. Biometrika. 87 (2000) 663-673.

[20] I. Gijbels, N.E. Heckman, Nonparametric testing for a monotone hazard function via normalized spacings. J. Nonparametr. Statist. 16 (2004) 463-477.

[21] I.J. Good, R.A. Gaskins, Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. J. Amer. Statist. Assoc. 75 (1980) 42-56.

[22] P.J. Green, B.W. Silverman, Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. Chapman and Hall, London, 1994.

[23] P. Hall, J.D. Hart, Bootstrap test for difference between means in nonparametric regression. J. Amer. Statist. Assoc. 85 (1990) 1039-1049.

[24] P. Hall, N.E. Heckman, Testing for monotonicity of a regression mean by calibrating for linear functions. Ann. Statist. 28 (2000) 20-39.

[25] P. Hall, M.C. Minnotte, C. Zhang, Bump hunting with non-Gaussian kernels. Ann. Statist. 32 (2004) 2124-2141.

[26] P. Hall, M. York, On the calibration of Silverman's test for multimodality. Statist. Sinica. 11 (2001) 515-536.

[27] J. Hannig, T.C.M. Lee, Robust SiZer for exploration of regression structures and outlier detection. J. Comput. Graph. Statist. 15 (2006) 101-117.

[28] J. Hannig, J.S. Marron, Advanced distribution theory for SiZer. J. Amer. Statist. Assoc. 101 (2006) 484-499.

[29] W. Härdle, J.S. Marron, (1990). Semiparametric comparison of regression curves. Ann. Statist. 18 (1990) 63-89.

[30] J. Harezlak, N.E. Heckman, CriSP: A tool for bump hunting. J. Comput. Graph. Statist. 10 (2001) 713-729.

[31] J. Harezlak, E. Naumova, N.M. Laird, LongCriSP: A test for bump hunting in longitudinal data. To appear, Stat. Med. (2006).

[32] J.A. Hartigan, P.M. Hartigan, The DIP test of unimodality. Ann. Statist. 13 (1985) 70-84.

[33] T.J. Hastie, R.J. Tibshirani, Generalized Additive Models. Monographs on Statistics and Applied Probability. 43 Chapman and Hall, Ltd., London, 1990.

[34] L.-S. Huang, Testing the adequacy of a linear model *via* critical smoothing. J. Stat. Comput. Simul. 68 (2001) 281-294.

[35] A.J. Izenman, C. Sommer, Philatelic mixtures and multimodal densities. J. Amer. Statist. Assoc. 83 (1988) 941-953.

[36] M.C. Jones, J.A. Rice, Displaying the important features of large collections of similar curves. Amer. Statist. 46 (1992) 140-145.

[37] C.S. Kim, J.S. Marron, SiZer for jump detection. J. Nonparametr. Statist. 18 (2006) 13-20.

[38] K.B. Kulasekera, J. Wang, Smoothing parameter selection for power optimality in testing of regression curves. J. Amer. Statist. Assoc. 92 (1997) 500-511.

[39] R. Li, J.S. Marron, Local likelihood SiZer map. Sankhyā Ser. A. 67 (2005) 476-498.

[40] E. Mammen, J.S. Marron, N.I. Fisher, Some asymptotics for multimodality tests based on kernel density estimates. Probab. Theory Related Fields. 91 (1992) 115-132.

[41] M.C. Minnotte, Nonparametric testing of the existence of a mode. Ann. Statist. 25 (1997) 1646-1660.

[42] M.C. Minnotte, D.W. Scott, The mode tree: A tool for visualization of nonparametric density features. J. Comput. Graph. Statist. 2 (1993) 51-68.

[43] D.W. Müller, G. Sawitzki, Excess mass estimates and tests for multimodality. J. Amer. Statist. Assoc. 86 (1991) 738-746.

[44] E.A. Nadaraya, On nonparametric estimates of density functions and regression curves. Theory Probab. Appl. 10 (1965) 186-190.

[45] M.B. Priestley, M.T. Chao, Nonparametric function fitting. J. Roy. Statist. Soc. Ser. B. 34 (1972) 385-392.

[46] J.O. Ramsay, B.W. Silverman, Functional Data Analysis. Springer-Verlag Inc., New York, 1997.

[47] D. Ruppert, S.J. Sheather, M.P. Wand, An effective bandwidth selector for local least squares regression. J. Amer. Statist. Assoc. 90 (1995) 1257-1270.

[48] W.R. Schucany, Kernel smoothers: An overview of curve estimators for the first graduate course in nonparametric statistics. Statist. Sci. 19 (2004) 663-675.

[49] M.R. Segal, J.L. Wiemels, Clustering of translocation breakpoints. J. Amer. Statist. Assoc. 97 (2002) 66-76.

[50] B.W. Silverman, Using kernel density estimates to investigate multimodality. J. Roy. Statist. Soc. Ser. B. 43 (1981) 97-99.

[51] B.W. Silverman, Some properties of a test for multimodality based on kernel density estimates. In: J. F. C. Kingman and G. E. H. Reuter (Eds.), Probability, Statistics and Analysis, Cambridge University Press, Cambridge, 1983, pp. 248-259.

[52] G. Wahba, Spline Models for Observational Data. Society for Industrial and Applied Mathematics, Philadelphia, 1990.

[53] G.S. Watson, Smooth regression analysis. Sankhyā Ser. A. 26 (1964) 359-372.

[54] M.A. Wong, A bootstrap testing procedure for investigating the number of subpopulations. J. Stat. Comput. Simul. 22 (1985) 99-112.

[55] Y. Xia, Bias-corrected confidence bands in nonparametric regression. J. Roy. Statist. Soc. Ser. B. 60 (1998) 797-811.