Chapter 4
Function Estimation

Hans-Georg Müller

4.1 Introduction to Three Papers on Nonparametric Curve Estimation

4.1.1 Introduction

The following is a brief review of three landmark papers of Peter Bickel on theoretical and methodological aspects of nonparametric density and regression estimation and the related topic of goodness-of-fit testing, including a class of semiparametric goodness-of-fit tests. We consider the context of these papers, their contribution and their impact. Bickel's first work on density estimation was carried out when this area was still in its infancy and proved to be highly influential for the subsequent wide-spread development of density and curve estimation and goodness-of-fit testing.

The first of Peter Bickel's contributions to kernel density estimation was published in 1973, nearly 40 years ago, when the field of nonparametric curve estimation was still in its infancy and was poised for the subsequent rapid expansion, which occurred later in the 1970s and 1980s. Bickel's work opened fundamental new perspectives, that were not fully developed until much later. Kernel density estimation was formalized in Rosenblatt (1956) and then developed further in Parzen (1962), where bias expansions and other basic techniques for the analysis of these nonparametric estimators were showcased.

Expanding upon an older literature on spectral density estimation, this work set the stage for substantial developments in nonparametric curve estimation that began in the later 1960s. This earlier literature on curve estimation is nicely surveyed in Rosenblatt (1971) and it defined the state of the field when Peter Bickel made

H.-G. Müller (✉)
Department of Statistics, University of California, Davis, CA, USA
e-mail: hgmueller@ucdavis.edu

J. Fan et al. (eds.), *Selected Works of Peter J. Bickel*, Selected Works in Probability and Statistics 13, DOI 10.1007/978-1-4614-5544-8_4,

the first highly influential contribution to nonparametric curve estimation in Bickel and Rosenblatt (1973). This work not only connected for the first time kernel density estimation with goodness-of-fit testing, but also did so in a mathematically elegant way.

A deep study of the connection between smoothness and rates of convergence and improved estimators of functionals of densities, corresponding to integrals of squared derivatives, is the hallmark of Bickel and Ritov (1988). Estimation of these density functionals has applications in determining the asymptotic variance of nonparametric location statistics. Functional of this type also appear as a factor in the asymptotic leading bias squared term for the mean integrated squared error. Thus the estimation of these functional has applications for the important problem of bandwidth choice for nonparametric kernel density estimates.

In the third article covered in this brief review, Bickel and Li (2007) introduce a new perspective to the well-known curse of dimensionality that affects any form of smoothing and nonparametric function estimation in high dimension: It is shown that for local linear smoothers in a nonparametric regression setting where the predictors at least locally lie on an unknown manifold, the curse of dimensionality effectively is not driven by the ostensible dimensionality of the predictors but rather by the dimensionality of the predictors, which might be much lower. In the case of relatively low-dimensional underlying manifolds, the good news is that the curse would then not be as severe as it initially appears, and one may obtain unexpectedly fast rates of convergence.

The first two papers that are briefly discussed here create a bridge between density estimation and goodness-of-fit. The goodness-of-fit aspect is central to Bickel and Rosenblatt (1973), while a fundamental transition phenomenon and improved estimation of density functionals are key aspects of Bickel and Ritov (1988). Both papers had a major impact in the field of nonparametric curve estimation. The third paper (Bickel and Li 2007) creates a fresh outlook on nonparametric regression and will continue to inspire new approaches. Some remarks on Bickel and Rosenblatt (1973) can be found in Sect. 2, on Bickel and Ritov (1988) Sect. 3, and on Bickel and Li (2007) in Sect. 4.

4.1.2 Density Estimation and Goodness-of-Fit

Nonparametric curve estimation originated in spectral density estimation, where it had been long known that smoothing was mandatory to improve the properties of such estimates (Daniell 1946; Einstein 1914). The smoothing field expanded to become a major field in nonparametric statistics around the time the paper Bickel and Rosenblatt (1973) appeared. At that time, kernel density estimation and other basic nonparametric estimators of density functions such as orthogonal least squares (Čencov 1962) were established. While many results were available in 1973 about local properties of these estimates, there had been no in-depth investigation yet of their global behavior.

This is where Bickel's influential contribution came in. Starting with the Rosenblatt-Parzen kernel density estimator

$$f_n(x) = \frac{1}{nb(n)} \sum_{i=1}^{n} w\left(\frac{x - X_i}{b(n)}\right) = \int \frac{1}{b(n)} w\left(\frac{x-u}{b(n)}\right) dF_n(u), \tag{4.1}$$

where $b(n)$ is a sequence of bandwidths that converges to 0, but not too fast, w a kernel function and dF_n stands for the empirical measure, Bickel and Rosenblatt (1973) consider the functionals

$$D_1 = \sup_{a_1 \le x \le a_2} |f_n(x) - f(x)|/(f(x))^{1/2}, \tag{4.2}$$

$$D_2 = \int_{a_1}^{a_2} \frac{[f_n(x) - f(x)]^2}{f(x)}. \tag{4.3}$$

The asymptotic behavior of these two functionals proves to be quite different. Functional D_1 corresponds to a maximal deviation on the interval, while functional D_2 is an integral and can be interpreted as a weighted integrated absolute deviation. While D_2, properly scaled, converges to a Gaussian limit, D_1 converges to an extreme value distribution. Harnessing the maximal deviation embodied in D_1 was the first serious attempt to obtain global inference in nonparametric density estimation. As Bickel and Rosenblatt (1973) state, *the statistical interest in this functional is twofold, as (i) a convenient way of getting a confidence band for f. (ii) A test statistic for the hypothesis $H_0 : f = f_0$*. They thereby introduce the goodness-of-fit theme, that constitutes one major motivation for density estimation and has spawned much research to this day. Motivation (i) leads to Theorem 3.1, and (ii) to Theorem 3.2 in Bickel and Rosenblatt (1973).

In their proofs, Bickel and Rosenblatt (1973) use a strong embedding technique, which was quite recent at the time. Theorem 3.1 is a remarkable achievement. If one employs a rectangular kernel function $w = 1_{[-\frac{1}{2},\frac{1}{2}]}$ and a bandwidth sequence $b(n) = n^{-\delta}$, $0 < \delta < \frac{1}{2}$, then the result in Theorem 3.1 is for centered processes

$$P\left[(2\delta \log n)^{1/2}\left([nb(n)f^{-1}(t)]^{1/2} \sup_{a_1,a_2}[f_n(t) - E(f_n(t))] - d_n\right) < x\right] \to e^{-2e^{-x}},$$

where

$$d_n = \rho_n - \frac{1}{2}\rho_n^{-1}[\log(\pi + \delta) + \log\log n], \quad \rho_n = (2\delta \log n)^{1/2}.$$

The slow convergence to the limit that is indicated by the rate $(\log n)^{1/2}$ is typical for maximal deviation results in curve estimation, of which Theorem 3.1 is the first. A multivariate version of this result appeared in Rosenblatt (1976).

A practical problem that has been discussed by many authors in the 1980s and 1990s has been how to handle the bias for the construction of confidence intervals and density-estimation based inference in general. This is a difficult problem. It is also related to the question how one should choose bandwidths when constructing confidence intervals, even pointwise rather than global ones, in relation to choosing the bandwidth for the original curve estimate for which the confidence region is desired (Hall 1992; Müller et al. 1987). For instance, undersmoothing has been advocated and also other specifically designed bias corrections. This is of special relevance when the maximal deviation is to be constructed over intervals that include endpoints of the density, where bias is a particularly notorious problem.

For inference and goodness-of-fit testing, Bickel and Rosenblatt (1973), based on the deviation D_2 as in (4.3), propose the test statistic

$$T_n = \int [f_n(x) - E(f_n(x))]^2 a(x)\, dx$$

with a weight function a for testing the hypothesis H_0. Compared to classical goodness-of-fit tests, this test is shown to be better than the χ^2 test and incorporates nuisance parameters as needed. This Bickel-Rosenblatt test has encountered much interest; an example is an application for testing independence (Rosenblatt 1975).

Recent extensions and results under weaker conditions include extensions to the case of an error density for stationary linear autoregressive processes that were developed in Lee and Na (2002) and Bachmann and Dette (2005), and for GARCH processes in Koul and Mimoto (2010). A related L^1-distance based goodnes-of-fit test was proposed in Cao and Lugosi (2005), while a very general class of semiparametric tests targeting composite hypotheses was introduced in Bickel et al. (2006).

4.1.3 Estimating Functionals of a Density

Kernel density estimators (4.1) require specification of a kernel function w and of a bandwidth or smoothing parameter $b = b(n)$. If one uses a kernel function that is a symmetric density, this selection can be made based on the asymptotically leading term of mean integrated squared error (MISE),

$$\frac{1}{4} b(n)^4 \int w(u) u^2\, du \int [f^{(2)}(x)]^2\, dx + [nb(n)]^{-1} \int w(u)^2\, du,$$

which leads to the asymptotically optimal bandwidth

$$b^*(n) = c \left(n \int [f^{(2)}(x)]^2\, dx \right)^{-1/5},$$

where c is a known constant. In order to determine this optimal bandwidth, one is therefore confronted with the problem of estimating integrated squared density derivatives

$$\int [f^{(k)}(x)]^2\,dx, \tag{4.4}$$

where cases $k > 2$ are of interest when choosing bandwidths for density estimates with higher order kernels. These have faster converging bias at the cost of increasing variance but are well known to have rates of convergence that are faster in terms of MISE, if the underlying density is sufficiently smooth and optimal bandwidths are used. Moreover, the case $k = 0$ plays a role in the asymptotic variance of rank-based estimators (Schweder 1975).

The relevance of the problem of estimating density functionals of type (4.4) had been recognized by various authors, including Hall and Marron (1987), at the time the work Bickel and Ritov (1988) was published. The results of Bickel and Ritov however are not a direct continuation of the previous line of research; rather, they constitute a surprising turn of affairs. First, the problem is positioned within a more general semiparametric framework. Second, it is established that the $\sqrt{n}$ of convergence that one expects for functionals of type (4.4) holds if $f^{(m)}$ is Hölder continuous of order α with $m+\alpha > 2k+\frac{1}{4}$, and, with an element of surprise, that it does not hold in a fairly strong sense when this condition is violated.

The upper bound for this result is demonstrated by utilizing kernel density estimates (4.1), employing a kernel function of order $\max(k, m-k)+1$ and then using plug-in estimators. However, straightforward plug-in estimators suffer from bias that is severe enough to prevent optimal results. Instead, Bickel and Ritov employ a clever bias correction term (that appears in their equation (2.2) after the plug-in estimator is introduced) and then proceed to split the sample into two separate parts, combining two resulting estimators.

An amazing part of the paper is the proof that an unexpected and surprising phase transition occurs at $\alpha = 1/4$. This early example for such a phase transition hinges on an ingenious construction of a sequence of measures and the Bayes risk for estimating the functional. For less smooth densities, where the transition point has not been reached, Bickel and Rosenblatt (1973) provide the optimal rate of convergence, a rate slower than $\sqrt{n}$. The arguments are connected more generally with semiparametric information bounds in the precursor paper Bickel (1982).

Bickel and Ritov (1988) is a landmark paper on estimating density functionals that inspired various subsequent works by other authors. These include further study of aspects that had been left open, such as adaptivity of the estimators (Efromovich and Low 1996), extensions to more general density functionals with broad applications (Birgé and Massart 1995) and the study of similar problems for other curve functionals, for example integrated second derivative estimation in nonparametric regression (Efromovich and Samarov 2000).

4.1.4 Curse of Dimensionality for Nonparametric Regression on Manifolds

It has been well known since Stone (1980) that all nonparametric curve estimation methods, including nonparametric regression and density estimation, suffer severely in terms of rates of convergence in high-dimensional or even moderately dimensioned situations. This is born out in statistical practice, where unrestricted nonparametric curve estimation is known to make little sense if moderately sized data have predictors with dimensions say $D \geq 4$. Assuming the function to be estimated is in a Sobolev space of smoothness p, optimal rates of convergence of Mean Squared Error and similar measures are $n^{-2p/(2p+D)}$ for samples of size n. To circumvent the curse of dimensionality, alternatives to unrestricted nonparametric regression have been developed, ranging from additive, to single index, to additive partial linear models. Due to their inherent structural constraints, such approaches come at the cost of reduced flexibility with the associated risk of increased bias.

The cause of the curse of dimensionality is the trade-off between bias and variance in nonparametric curve estimation. Bias control demands to consider data in a small neighbourhood around the target predictor levels $\mathbf{x}$, where the curve estimate is desired, while variance control requires large neighbourhoods containing many predictor-response pairs. For increasing dimensions, the predictor locations become increasingly sparse, with larger average distances between predictor locations, moving the variance-bias trade-off and resulting rate of convergence in an unfavorable direction.

Using an example where $p = 2$ and the local linear regression method, Bickel and Li (2007) analyze what happens if the predictors are in fact not only located on a compact subset of $\mathscr{R}^D$, where D is potentially large, but in fact are, at least locally around $\mathbf{x}$, located on a lower-dimensional manifold with intrinsic dimension $d < D$. They derive that in this situation, one obtains the better rate $n^{-2p/(2p+d)}$, where the manifold is assumed to satisfy some local regularity conditions, but otherwise is unknown. This can lead to dramatic gains in rates of convergence, especially if $d = 1, 2$ while D is large.

This nice result can be interpreted as a consequence of the denser packing of the predictors on the lower-dimensional manifold with smaller average distances as compared to the average distances one would expect for the ostensible dimension D of the space, when the respective densities are not degenerate. A key feature is that knowledge of the manifold is not needed to take advantage of its presence. The data do not even have to be located precisely on the manifold, as long as their deviation from the manifold becomes small asymptotically. Bickel and Li (2007) also provide thoughtful approaches to bandwidth choices for this situation and for determining the intrinsic dimension of the unknown manifold, and thus the rate of effective convergence that is determined by d.

This approach likely will play an important role in the ongoing intensive quest for flexible yet fast converging dimension reduction and regression models. Methods for variable selection, dimension reduction and for handling collinearity among predictors, as well as extensions to "large p, small n" situations are in high demand. The idea of exploiting underlying manifold structure in the predictor space for these purposes is powerful, as has been recently demonstrated in Mukherjee et al. (2010) and Aswani et al. (2011). These promising approaches define a new line of research for high-dimensional regression modeling.

Acknowledgements Supported in part by NSF grant DMS-1104426.

References

Aswani A, Bickel P, Tomlin C (2011) Regression on manifolds: estimation of the exterior derivative. Ann Stat 39:48–81

Bachmann, D, Dette H (2005) A note on the Bickel-Rosenblatt test in autoregressive time series. Stat Probab Lett 74:221–234

Bickel P (1982) On adaptive estimation. Ann Stat 10:647–671

Bickel P, Li B (2007). Local polynomial regression on unknown manifolds. In: Complex datasets and inverse problems: tomography, networks and beyond. IMS lecture notes-monograph series, vol 54. Institute of Mathematical Statistics, Beachwood, pp 177–186

Bickel P, Ritov Y (1988) Estimating integrated squared densiuty derivatives: sharp best order of convergence estimates. Sankhya Indian J Stat Ser A 50:381–393

Bickel P, Rosenblatt M (1973) On some global measures of the deviations of density function estimates. Ann Stat 1:1071–1095

Bickel P, Ritov Y, Stoker T (2006) Tailor-made tests for goodness of fit to semiparametric hypotheses. Ann Stat 34:721–741

Birgé L, Massart P (1995) Estimation of integral functionals of a density. Ann Stat 23:11–29

Cao R, Lugosi G (2005) Goodness-of-fit tests based on kernel density estimator. Scand J Stat 32:599–616

Čencov N (1962) Evaluation of an unknown density from observations. Sov Math 3:1559–1562

Daniell P (1946) Discussion of paper by M.S. Bartlett. J R Stat Soc Suppl 8:88–90

Efromovich S, Low M (1996) On Bickel and Ritov's conjecture about adaptive estimation of the integral of the square of density derivative. Ann Stat 24:682–686

Efromovich S, Samarov A (2000) Adaptive estimation of the integral of squared regression derivatives. Scand J Stat 27:335–351

Einstein A (1914) Méthode pour la détermination de valeurs statistiques d'observations concernant des grandeurs soumises à des fluctuations irrégulières. Arch Sci Phys et Nat Ser 4 37:254–256

Hall P (1992) Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. Ann Stat 20:675–694

Hall P, Marron J (1987) Estimation of integrated squared density derivatives. Stat Probab Lett 6:109–115

Koul H, Mimoto N (2010) A goodness-of-fit test for garch innovation density. Metrika 71:127–149

Lee S, Na S (2002) On the Bickel-Rosenblatt test for first order autoregressive models. Stat Probab Lett 56:23–35

Mukherjee S, Wu Q, Zhou D (2010) Learning gradients on manifolds. Bernoulli 16:181–207

Müller H-G, Stadtmüller U, Schmitt T (1987) Bandwidth choice and confidence intervals for derivatives of noisy data. Biometrika 74:743–749

Parzen E (1962) On estimation of a probability density function and mode. Ann Math Stat 33:1065–1076

Rosenblatt M (1956) Remarks on some nonparametric estimates of a density function. Ann Math Stat 27:832–837

Rosenblatt M (1971) Curve estimates. Ann Stat 42:1815–1842

Rosenblatt M (1975) A quadratic measure of deviation of two-dimensional density estimates and a test of independence. Ann Stat 3:1–14

Rosenblatt M (1976) On the maximal deviation of k-dimensional density estimates. Ann Probab 4:1009–1015

Schweder T (1975) Window estimation of the asymptotic variance of rank estimators of location. Scand J Stat 2:113–126

Stone CJ (1980) Optimal rates of convergence for nonparametric estimators. Ann Stat 10:1040–1053

The Annals of Statistics
1973, Vol. 1, No. 6, 1071-1095

ON SOME GLOBAL MEASURES OF THE DEVIATIONS OF DENSITY FUNCTION ESTIMATES

BY P. J. BICKEL[1] AND M. ROSENBLATT[2]

University of California, Berkeley;
University of California, San Diego

We consider density estimates of the usual type generated by a weight function. Limt theorems are obtained for the maximum of the normalized deviation of the estimate from its expected value, and for quadratic norms of the same quantity. Using these results we study the behavior of tests of goodness-of-fit and confidence regions based on these statistics. In particular, we obtain a procedure which uniformly improves the chi-square goodness-of-fit test when the number of observations and cells is large and yet remains insensitive to the estimation of nuisance parameters. A new limit theorem for the maximum absolute value of a type of nonstationary Gaussian process is also proved.

1. Introduction. Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed random variables with continuous density function $f(x)$. By now there are a goodly number of papers on estimation of the density function (see [13] for a bibliography). The class of estimates $f_n(x)$ that we consider are determined by a bounded integrable weight function w,

$$f_n(x) = \frac{1}{nb(n)} \sum_{j=1}^{n} w\left(\frac{x - X_j}{b(n)}\right) \tag{1.1}$$
$$= \int \frac{1}{b(n)} w\left(\frac{x - s}{b(n)}\right) dF_n(s) .$$

In formula (1.1), F_n is the sample distribution function. Also $b(n)$ is a bandwidth that tends to zero as $n \to \infty$ but is such that $n^{-1} = o(b(n))$.

The local properties of such estimates have been discussed extensively. Our object will be to get global measures of how good $f_n(x)$ is as an estimate of $f(x)$. In particular, the asymptotic distribution of the functionals

$$\max_{0 \leq x \leq 1} |f_n(x) - f(x)|/(f(x))^{\frac{1}{2}}$$

and

$$\int_0^1 \frac{[f_n(x) - f(x)]^2}{f(x)}\, dx$$

are evaluated under appropriate conditions as $n \to \infty$.

Received November 1971; revised March 1973.

[1] This work was performed while the author was a fellow of the John Simon Guggenheim Memorial Foundation and supported in part by the Office of Naval Research under Contract N00014-69-A0200-1038.

[2] This work was carried out while the author was a fellow of the John Simon Guggenheim Memorial Foundation and supported in part by the Office of Naval Research.

Key words and phrases. Density function estimates, weight function, bandwidth, global measure of deviation, asymptotic distribution, functionals of stochastic process, Gaussian process.

We shall state two results, one concerned with absolute deviation of the estimate $f_n(x)$ from $f(x)$ and the other with integrated quadratic deviation. They will give some insight into the type of result that is obtained. However, in order to give the result on absolute deviation it is convenient to introduce at this point certain convenient assumptions which we shall refer to as A1, A2, A3, and A4.

A1. The weight function w also assigns mass one to the line and either (a) vanishes outside an interval $[-A, A]$ and is absolutely continuous on $[-A, A]$ with derivative w' or (b) is absolutely continuous on $(-\infty, \infty)$ with derivative w' such that $\int |w'(t)|^k \, dt < \infty$, $k = 1, 2$.

A2. The density f is continuous, positive and bounded.

A3. The function $f^{\frac{1}{2}}$ is absolutely continuous and its derivative $\frac{1}{2}f'/f^{\frac{1}{2}}$ is bounded in absolute value. Moreover,

$$\int_{[|z| \geqq 3]} |z|^{\frac{3}{2}}[\log \log |z|]^{\frac{1}{2}}[|w'(z)| + |w(z)|] \, dz < \infty .$$

A4. The second derivative f'' of f exists and is bounded. Moreover w is symmetric (about 0) and $z^2 w(z)$ is integrable.

We shall simply state a corollary of a main result on absolute deviations which is appealing because it is phrased in a form that is convenient if one wishes to set up a confidence band for the density function.

COROLLARY. *Let assumptions* A1—A4 *be satisfied with* $b(n) = n^{-\delta}$, $\frac{1}{5} < \delta < \frac{1}{2}$. *Then*

$$\lim_{n \to \infty} P\Big[f_n(x) - \left(\frac{f_n(x)\lambda(w)}{nb(n)}\right)^{\frac{1}{2}} \left(\frac{z}{(2\delta \log n)^{\frac{1}{2}}} + d_n\right)$$

$$\leqq f(x) \leqq f_n(x) + \left(\frac{f_n(x)\lambda(w)}{nb(n)}\right)^{\frac{1}{2}} \left(\frac{z}{(2\delta \log n)^{\frac{1}{2}}} + d_n\right) \ \textit{for all} \ \ 0 \leqq x \leqq 1 \Big] \tag{1.2}$$

$$= e^{-2e^{-z}}$$

where

$$\lambda(w) = \int w^2(t) \, dt$$

and

$$d_n = (2\delta \log n)^{\frac{1}{2}} + \frac{1}{(2\delta \log n)^{\frac{1}{2}}} \left\{ \log\left(\frac{K_1(w)}{\pi^{\frac{1}{2}}}\right) + \tfrac{1}{2}[\log \delta + \log \log n] \right\}$$

if (a) *of* A1 *holds and*

$$K_1(w) = \frac{w^2(A) + w^2(-A)}{2} \Big/ \lambda(w) > 0 ,$$

and otherwise

$$d_n = (2\delta \log n)^{\frac{1}{2}} + \frac{1}{(2\delta \log n)^{\frac{1}{2}}} \left[\log \frac{1}{\pi} \left(\frac{K_2(w)}{2}\right)^{\frac{1}{2}} \right]$$

with

$$K_2(w) = \tfrac{1}{2} \int_{-\infty}^{\infty} [w'(t)]^2 \, dt / \lambda(w) .$$

The following result for a quadratic functional is also of some interest. The function $a(x)$ used in the theorem is assumed to be a bounded piece-wise smooth integrable function.

THEOREM. *Let* A1—A4 *hold. Then if* $b(n) = o(n^{-\frac{2}{9}})$, $n^{-\frac{1}{4}}(\log n)^{\frac{1}{2}}(\log\log n)^{\frac{1}{4}} = o(b(n))$ *as* $n \to \infty$,

$$b(n)^{-\frac{1}{2}}[nb(n) \int [f_n(x) - f(x)]^2 a(x)\, dx - \int f(x)a(x)\, dx \int w^2(z)\, dz]$$

is asymptotically normally distributed with mean zero and variance

$$2 \int (\int w(x+y)w(x)\, dx)^2\, dy \int a^2(x)f^2(x)\, dx$$

as $n \to \infty$.

The basic technique in obtaining the results is that of approximating the normalized and centered sample distribution function by an appropriate Brownian motion process on a convenient probability space by using a Skorohod-like imbedding due to Brillinger and Breiman. The details of this approximation and remarks on approximation of other functionals are given in Section 2. The asymptotic theory of the maximal deviation and that of quadratic deviations are developed in Sections 3 and 4 respectively. Some computations on the power of these procedures are also carried out. In particular, we show that a goodness-of-fit test based on a quadratic functional is strictly better than the χ^2 test. There is also an appendix on the asymptotic distribution of the maximal deviation for a type of nonstationary Gaussian process.

2. Approximations. As has been indicated in the introduction our technique is to consider the statistics of interest as functionals of certain stochastic processes on the interval [0, 1] and then to substitute Gaussian processes with the same covariance structure for the latter where possible.

It is convenient to introduce $Z_n^0(\cdot)$ given by

$$Z_n^0(t) = n^{\frac{1}{2}}(F_n^*(t) - t)\,, \qquad 0 \leq t \leq 1 \tag{2.1}$$

where $F_n^* = F_n(F^{-1})$ is the empirical distribution of $F(X_1), \cdots, F(X_n)$. This will be approximated by $Z^0(\cdot)$, the Brownian bridge, given by

$$Z^0(t) = Z(t) - tZ(1) \tag{2.2}$$

where Z is a standard Wiener process on [0, 1].

The process $[nb(n)f^{-1}(t)]^{\frac{1}{2}}(f_n(\cdot) - E(f_n(\cdot)))$ is central to our discussion. It can be written as

$$Y_n(t) = b^{-\frac{1}{2}}(n)f^{-\frac{1}{2}}(t) \int_{-\infty}^{\infty} w\left(\frac{t-s}{b(n)}\right) dZ_n^0(F(s))\,. \tag{2.3}$$

Approximations ${}_0Y_n$ and ${}_1Y_n$ to this process are obtained by substituting $Z^0(F(\cdot))$ and $Z(F(\cdot))$ respectively for the random measure in (2.3). The resulting processes are well defined, at least if $\int_{-\infty}^{\infty} w^2(t)\, dF(t) < \infty$.

Two other processes which also arise naturally are given by

$${}_2Y_n(t) = [b(n)f(t)]^{-\frac{1}{2}} \int w\left(\frac{t-s}{b(n)}\right) (f(s))^{\frac{1}{2}}\, dZ(s) \tag{2.4}$$

and

$$(2.5)\qquad {}_3Y_n(t) = [b(n)]^{-\frac{1}{2}} \int w\left(\frac{t-s}{b(n)}\right) dZ(s)$$

where Z is a two-sided Wiener process on $(-\infty, \infty)$ (dZ is Wiener measure). The process ${}_3Y_n$ is well defined if $\int w^2(t)\, dt < \infty$, and all the integrals with respect to $dZ^0(F(\cdot))$, $dZ(F(\cdot))$, $dZ(\cdot)$, $dZ^0(\cdot)$ are taken in the L^2 sense (cf. Doob [6] page 426). For convenience, suppose all our processes are realized as random elements taking their values in the space $D[0, 1]$ (cf. [3]). For $x \in D[0, 1]$ let $||x|| = \sup\{|x(t)| : 0 \leqq t \leqq 1\}$. Our approximations rest on the following theorem of Brillinger (1969). (A similar argument appeared simultaneously in Breiman 1969).)

THEOREM. *There exists a probability space* (Ω, A, P) *on which one can construct versions of* Z_n^0 *and* Z *such that*

$$(2.6)\qquad ||Z_n^0 - Z^0|| = O_p(n^{-\frac{1}{4}}(\log n)^{\frac{1}{2}}(\log\log n)^{\frac{1}{4}})\,.$$

From this we can derive

PROPOSITION 2.1. *If the processes* Z_n^0, Z^0 *are constructed as above and* A1 *and* A2 *hold, then*

$$(2.7)\qquad ||Y_n - {}_0Y_n|| = O_p(b^{-\frac{1}{2}}(n)n^{-\frac{1}{4}}(\log n)^{\frac{1}{2}}(\log\log n)^{\frac{1}{4}})\,.$$

PROOF. Write, using A1,

$$\begin{aligned}(2.8)\qquad Y_n(q) = {} & [b(n)f(q)]^{-\frac{1}{2}}\{-w(A)Z_n^0(F(q - Ab(n))) \\ & + w(-A)Z_n^0(F(q + Ab(n)))\} \\ & + b^{-\frac{3}{2}}(n)f^{-\frac{1}{2}}(q) \int_{-\infty}^{\infty} Z_n^0(F(s))w'\left(\frac{q-s}{b(n)}\right) ds\,.\end{aligned}$$

(The first two terms inside the curly brackets are taken to be 0 in the event A1(b) holds but A1(a) does not.) A similar representation is valid for ${}_0Y_n$ and (2.7) follows.

PROPOSITION 2.2. *If* A2 *holds then*

$$(2.9)\qquad ||{}_0Y_n - {}_1Y_n|| = O_p(b^{\frac{1}{2}}(n))\,.$$

If A2 *and* A3 *hold then*

$$(2.10)\qquad ||{}_2Y_n - {}_3Y_n|| = O_p(b^{\frac{1}{2}}(n))\,.$$

PROOF. From the representation (2.2),

$$\begin{aligned}(2.11)\qquad |{}_0Y_n(q) - {}_1Y_n(q)| = {} & |Z(1)|[b(n)f(q)]^{-\frac{1}{2}} \\ & \int w\left(\frac{q-s}{b(n)}\right) f(s)\, ds = |Z(1)|O(b^{\frac{1}{2}}(n))\,.\end{aligned}$$

Applying (2.8) and its analogues, if A1(a) holds,

$$\begin{aligned} &|{}_2Y_n(q) - {}_3Y_n(q)| \\ &\quad \leqq b^{-\frac{1}{2}}(n)\{[|Z(Ab(n) + q)||[f(Ab(n) + q)/f(q)]^{\frac{1}{2}} - 1| \\ &\qquad + |Z(-Ab(n) + q)||[f(-Ab(n) + q)/f(q)]^{\frac{1}{2}} - 1|] \sup_t |w(t)| \\ &\qquad + \textstyle\int |Z(sb(n) + q)||[f(q + sb(n))/f(q)]^{\frac{1}{2}} - 1||w'(s)|\, ds \\ &\qquad + \tfrac{1}{2}(b(n)) \textstyle\int |Z(sb(n) + q)||f'(q + sb(n))|[f(q)f(q + sb(n))]^{-\frac{1}{2}}|w(s)|\, ds\} \\ &\quad = O_p(b^{\frac{1}{2}}(n)) \end{aligned} \tag{2.12}$$

by using A3 and the law of the iterated logarithm for the Wiener process. If A1(b) holds the first two terms vanish and the same argument applies.

To apply these propositions we make the elementary

Remark. If $\{g_n\}$ is a sequence of functionals on $D[0, 1]$ satisfying Lipschitz conditions such that

$$|g_n(x) - g_n(y)| \leqq M_n||x - y|| \tag{2.13}$$

and A_n, B_n are stochastic processes realizable in D such that $||A_n - B_n|| = o_p(1/M_n)$, then $g_n(A_n)$ converges in law if and only if $g_n(B_n)$ does, and to the same limit.

We shall apply this proposition in the next two sections to the functionals

$$\text{I} \qquad (2|\log b(n)|)^{\frac{1}{2}}\left[\max\left\{\frac{|Y_n(t)|}{(\lambda(w))^{\frac{1}{2}}} : 0 \leqq t \leqq 1\right\} - B([b(n)]^{-1})\right]$$

where B is defined in Theorem A1 and,

$$\text{II} \qquad b^{-\frac{1}{2}}(n)\left[\textstyle\int_{-\infty}^{\infty} Y_n^2(t)f(t)a(t)\, dt - \int_{-\infty}^{\infty} w^2(t)\, dt\right]$$

where a is an integrable weight function. Evidently, since ${}_1Y_n$ and ${}_2Y_n$ have the same joint laws, we can substitute ${}_3Y_n$ for Y_n in I if A1—A3 hold and

$$o\left(\frac{b(n)}{|\log b(n)|}\right) = n^{-\frac{1}{2}} \log n(\log \log n)^{\frac{1}{2}} \tag{2.14}$$

and ${}_0Y_n$ can be substituted for Y_n in II if A1 and A2 hold and,

$$o(b(n)) = n^{-\frac{1}{4}}(\log n)^{\frac{1}{2}}(\log \log n)^{\frac{1}{4}} . \tag{2.15}$$

Although we do not pursue this it is clear that the same technique can be applied to other functionals, e.g., a normalized version of the total time in [0, 1] spent by Y_n above a high level (cf. Berman (1971) [2]).

3. The maximum absolute deviation. The first measure of global deviation that we consider is $\tilde{M}_n = \max\{|Y_n(t)| : 0 \leqq t \leqq 1\}$. (There is no loss in considering [0, 1] rather than any other interval on which the density is bounded away from 0 and ∞.) The statistical interest of this functional is twofold as

(i) A convenient way of getting a confidence band for f.

(ii) A test statistic for the hypothesis $H: f = f_0$.

Under (ii) we shall also consider the possibility of testing composite hypotheses, for example, that f is Gaussian. The asymptotic theorem we need to discuss (i), and the behavior of (ii) under the null hypothesis is a consequence of our remarks in Section 2 and Theorem A1 of the appendix.

THEOREM 3.1. *Let* w *satisfy assumptions* A1—A3 *and*

$$b(n) = n^{-\delta}, \qquad 0 < \delta < \tfrac{1}{2}.$$

Then,

$$P\left[(2\delta \log n)^{\frac{1}{2}} \left(\frac{\tilde{M}_n}{(\lambda(w))^{\frac{1}{2}}} - d_n\right) < x\right] \to e^{-2e^{-x}}, \tag{3.1}$$

where

$$\lambda(w) = \int w^2(t)\, dt \tag{3.2}$$

and

$$d_n = (2\delta \log n)^{\frac{1}{2}} + \frac{1}{(2\delta \log n)^{\frac{1}{2}}} \left\{\log \frac{K_1(w)}{\pi^{\frac{1}{2}}} - \tfrac{1}{2}[\log \delta + \log\log n]\right\} \tag{3.3}$$

where

$$K_1(w) = \frac{w^2(A) + w^2(-A)}{2}\Big/\lambda(w),$$

if $K_1(w) > 0$, *and otherwise*

$$d_n = (2\delta \log n)^{\frac{1}{2}} + \frac{1}{(2\delta \log n)^{\frac{1}{2}}}\left[\log \frac{1}{\pi} \frac{K_2(w)}{2}\right]$$

where

$$K_2(w) = \tfrac{1}{2}\left[\int [w'(t)]^2\, dt\right]/\lambda(w).$$

REMARK 1. The natural weight function $w(t) = \frac{1}{2}$, $|t| \leqq 1$, $= 0$ otherwise, falls under the first case, while the "optimal" weight function of Epanechnikov (1969) $w(t) = 3/(4(5)^{\frac{1}{2}})(1 - (v^2/5))$ if $|v| \leqq 5^{\frac{1}{2}}$, $= 0$ otherwise, falls under the second.

REMARK 2. A similar result holds if one considers the maximum deviation (rather than absolute deviation) of a density function estimate as in Rosenblatt (1971). However, since one-sided deviations for density functions are unnatural the present result seems more interesting.

REMARK 3. The techniques of proof of this result may readily be adapted to prove limit theorems such as that of Woodroofe (1967) for the maximum deviation observed at an increasing finite number of points.

PROOF. It follows from Propositions 2.1 and 2.2 and the following remark that the limiting behavior of $(2\delta \log n)^{\frac{1}{2}}[(\tilde{M}_n/(\lambda(w))^{\frac{1}{2}}) - d_n]$ is the same as that of $(2 \log b(n))^{\frac{1}{2}}(\max\{|{}_2Y_n(t)|/(\lambda(w))^{\frac{1}{2}} : 0 \leqq t \leqq 1\} - d_n)$. By the similarity transform for the Wiener process, the law

$$L({}_3Y_n(t) : 0 \leqq t \leqq 1) = L\left(\int w\left(\frac{t}{b(n)} - s\right) dZ(s) : 0 \leqq t \leqq 1\right). \tag{3.4}$$

Since $1/(\lambda(w))^{\frac{1}{2}} \int w(t-s)\, dZ(s)$ is a stationary Gaussian process with mean 0 and covariance

$$r(t) = \frac{\int w(s+t) w(s)\, ds}{\lambda(w)}, \tag{3.5}$$

Theorem 3.1 follows from Corollary A.1 provided we show that r satisfies condition (v) and (vi) of Theorem A1 with $\alpha = 1, 2$. That (v) is satisfied is equivalent to Theorem B1. Moreover,

$$\int r^2(t)\, dt = \frac{1}{2\pi} \int |\hat{r}(t)|^2\, dt = \frac{1}{2\pi\lambda^2(w)} \int |\hat{w}(t)|^4\, dt \tag{3.6}$$

where $\hat{}$ denotes Fourier transformation. Since w is integrable and bounded $\hat{w}$ is square integrable and bounded and (vi) must hold.

Applications. (i) To obtain a confidence band for f that is simple and explicit it is natural to consider δ such that $E(f_n)$ can be replaced by f. This is true if $\delta > \frac{1}{5}$ and A4 holds. Then,

$$\frac{1}{b(n)} \int w\left(\frac{t-s}{b(n)}\right) f(s)\, ds = f(t) + O(b^2(n)) \tag{3.7}$$

with 0 independent of t. If we now define $Y_n{}^*$ by replacing $E(f_n(t))$ with $f(t)$ we conclude that

$$||Y_n - Y_n{}^*|| = O([nb^5(n)]^{\frac{1}{2}}) . \tag{3.8}$$

Using the usual approximations we conclude that $\max\{|Y_n{}^*(t)| : 0 \leqq t \leqq 1\}$ behaves like $\tilde{M}_n$ if A4 holds and $\delta > \frac{1}{5}$. In this case inverting as usual we obtain the confidence band

$$\begin{aligned} f &\leqq f_n + \left(\frac{f_n}{nb(n)}\right)^{\frac{1}{2}} c(\alpha)\left(1 + \frac{c^2(\alpha)}{4nb(n)f_n}\right)^{\frac{1}{2}} + \frac{c^2(\alpha)}{2nb(n)} \\ f &\geqq f_n - \left(\frac{f_n}{nb(n)}\right)^{\frac{1}{2}} c(\alpha)\left(1 + \frac{c^2(\alpha)}{4nb(n)f_n}\right)^{\frac{1}{2}} + \frac{c^2(\alpha)}{2nb(n)} \end{aligned} \tag{3.9}$$

where $c(\alpha)$ is given by (3.11). A simpler band is obtained if we further substitute f_n for f in the denominator of Y_n. The resulting process $Y_n{}^{**}$ (say) has

$$\begin{aligned} ||Y_n{}^* - Y_n{}^{**}|| &= O_p\left(\frac{||Y_n{}^*||^2}{(nb(n))^{\frac{1}{2}}}\, ||f_n{}^{-\frac{1}{2}}||\right) \\ &= O_p\left(\frac{\log n}{(nb(n))^{\frac{1}{2}}}\right) \end{aligned} \tag{3.10}$$

if A1—A4 hold and $\frac{1}{5} < \delta < \frac{1}{2}$. The approximate confidence band obtained by looking at the maximum of $|Y_n{}^{**}|$ is given in the introduction (1.2).

There is no choice of δ which asymptotically makes this simple band as thin as possible, i.e. one should choose δ as small as possible. This of course ignores the obvious—the speed with which bias disappears asymptotically depends on δ as does the speed of convergence to the asymptote. However, for fixed n there is an optimal $\delta(n)$ (depending on α) > 0 which for moderate n and small α may be the right thing to use if the choice of bandwidth is free.

(ii) To test $H: f = f_0$ it is natural to compute $\tilde{M}_n$ with $f = f_0$ and reject for large values of the statistic. According to the theorem to obtain approximate level α we should use as cutoff point,

$$c(\alpha) = -[\log|\log(1-\alpha)| - \log 2]\frac{(\lambda(w))^{\frac{1}{2}}}{(2\delta\log n)^{\frac{1}{2}}} + d_n(\lambda(w))^{\frac{1}{2}}\,. \tag{3.11}$$

Under some assumptions the same cutoff point may be used for testing composite hypotheses of the form $H: f = f_0(\cdot, \theta)$ where θ is an unknown vector parameter by using $\tilde{M}_n$ with an estimate $\hat{\theta}$ substituted for the unknown parameter θ. We need the following assumption.

A5. The estimate $\hat{\theta}$ is such that if $\theta = \theta_0$, for every θ_0,

$$\begin{aligned}\sup\{|\textstyle\int [f_0(t+sb(n),\hat{\theta}) - f_0(t+sb(n),\theta_0)]w(s)\,ds| : 0 \leqq t \leqq 1\}& \\ = o_p([nb(n)\log b(n)]^{-\frac{1}{2}})&\end{aligned} \tag{3.12}$$

and

$$\|f_0(\cdot,\theta_0) - f_0(\cdot,\hat{\theta})\| = o_p(|\log b(n)|^{-1})\,.$$

Typically for maximum likelihood and method of moments estimates

$$|\hat{\theta} - \theta_0| = O_p(n^{-\frac{1}{2}})\,. \tag{3.13}$$

If, moreover, $\theta = (\theta^{(1)}, \cdots, \theta^{(k)})$, $\partial f_0/\partial\theta^{(j)}$ is bounded for θ in a neighborhood of θ_0, all x, and $1 \leqq j \leqq k$, it is easy to see that A5 holds. To see that A5 is the needed assumption again introduce a process $\bar{Y}_n$ with $E_\theta(f_n)$ replaced by $E_{\hat{\theta}}(f_n)$ and $(f(\cdot,\theta))^{\frac{1}{2}}$ replaced in the denominator of Y_n by $(f(\cdot,\hat{\theta}))^{\frac{1}{2}}$. Then

$$\|Y_n - \bar{Y}_n\| = o_p([\log b(n)]^{-\frac{1}{2}}) \tag{3.14}$$

and the result follows.

To make local power calculations on the test of the simple hypothesis described above we need to consider the behavior of $\tilde{M}_n$ (calculated under f_0) for a sequence of alternatives of the form,

$$g_n(x) = f_0(x) + \gamma_n\eta(x) + o(\gamma_n) \tag{3.15}$$

where g_n satisfy A2—A3 uniformly in n, $\gamma_n \downarrow 0$ at a suitable rate, and $o(\gamma_n)$ is uniform in x on $[0, 1]$. (Note that η must be continuous on $[0, 1]$.) Denote probabilities calculated under g_n by P_n. Our basic result is,

THEOREM 3.2. *Suppose that g_n are as above. Let w satisfy* A1—A3 *and define $\tilde{M}_n$ in terms of f_0. Let*

$$\gamma_n = n^{-\frac{1}{2}+\delta/2}[2\delta\log n]^{-\frac{1}{2}}\,.$$

Then,

$$P_n\left[(2\delta\log n)^{\frac{1}{2}}\left(\frac{\tilde{M}_n}{(\lambda(w))^{\frac{1}{2}}} - d_n\right) < x\right] \to \exp[-s(\eta)e^{-x}] \tag{3.16}$$

where

$$s(\eta) = \textstyle\int_0^1 \{\exp[\eta(t)/(f_0(t)\lambda(w))^{\frac{1}{2}}] + \exp[-\eta(t)/(f_0(t)\lambda(w))^{\frac{1}{2}}]\}\,dt\,. \tag{3.17}$$

This result follows from Theorem A1 quite readily.

One interesting consequence of this formula is that our test is asymptotically strictly unbiased for such alternatives. The reason is that $s(\eta) \geqq 2$ with $s(\eta) > 2$ unless $\eta = 0$ and the family of distributions $e^{-\theta e^{-x}}$ is an exponential family in θ.

Unfortunately these tests are asymptotically inadmissible (have Pitman efficiency 0) when compared to the test based on the quadratic functional of the next section based on the same w and $b(n)$. The reason is that alternatives there may be permitted to come in to f_0 at rate $n^{-\frac{1}{2}+\delta/4}$ rather than $n^{-\frac{1}{2}+\delta/2}$. However, this test for moderate sample sizes and some alternatives may well be preferable.

In analogy to the confidence band situation it would appear that maximum power is achieved by taking δ as small as possible. However, consideration of the approximation arguments suggests that $s(\eta_n)$ is a better measure of the "true shift" than $s(\eta)$ where,

$$\eta_n = (g_n - f_0)(2nb(n) \log b(n))^{\frac{1}{2}} . \tag{3.18}$$

Of course, $s(\eta_n)$ may well be maximized for $\delta > 0$. In all of these questions it would be desirable to have some small sample Monte Carlo explorations.

4. Quadratic functionals. We are interested in the behavior of the functional,

$$T_n = nb(n) \int_{-\infty}^{\infty} [f_n(x) - E(f_n(x))]^2 a(x)\, dx = \int_{-\infty}^{\infty} L_n^2(x) a(x)\, dx , \tag{4.1}$$

where $L_n = f^{\frac{1}{2}} Y_n$ and a is integrable. We have already remarked that if A1 and A2 hold and (say) $b_n = n^{-\delta}$, $\delta < \frac{1}{4}$, then.

$$|T_n - \int {}_0L_n^2(x) a(x)\, dx| = o(b^{\frac{1}{2}}(n)) . \tag{4.2}$$

Moreover, if a is bounded as well as integrable and w and f are bounded, we can replace ${}_0L_n$ by ${}_1L_n = f^{\frac{1}{2}}{}_1Y_n$ and hence by ${}_2L_n = f^{\frac{1}{2}}{}_2Y_n$. To see this note that,

$$\begin{aligned} &|\int ({}_1L_n^2(x) - {}_0L_n^2(x)) a(x)\, dx| \\ &\quad = \Big|\int \frac{1}{b(n)} \Big\{\Big(Z(1) \int w\Big(\frac{t-s}{b(n)}\Big) f(s)\, ds\Big)^2 \\ &\qquad - 2Z(1) \int w\Big(\frac{t-s}{b(n)}\Big) dZ(F(s)) \int w\Big(\frac{t-s}{b(n)}\Big) f(s)\, ds\Big\}\, a(t)\, dt\Big| \\ &\quad \leqq Z(1)^2 b(n) \sup_x |f(x)| \int |a(t)\, dt| \\ &\qquad + 2|Z(1)|b(n)| \int (\int w(y) c(s + b(n)y) a(s + b(n)y)\, dy)\, dZ(F(s))| \end{aligned} \tag{4.3}$$

where

$$c(t) = \int w(y) f(t - b(n)y)\, dy .$$

But,

$$\begin{aligned} &E(\int (\int w(y) c(s + b(n)y) a(s + b(n)y)\, dy)\, dZ(F(s)))^2 \\ &\qquad = \int (\int w(y) c(s + b(n)y) a(s + b(n)y)\, dy)^2\, dF(s) \end{aligned} \tag{4.4}$$

is bounded.

By (4.3) and (4.4),

$$|T_n - \int {}_2L_n^2(x) a(x)\, dx| = O_p(b(n)) . \tag{4.5}$$

(The infinite range poses no problem since we are approximating L_n rather than the normalized Y_n.)

The following lemma lets us determine the characteristic function of a quadratic functional

$$Z = \int Y(x)^2 a(x)\,dx \tag{4.6}$$

of a Gaussian process $Y(x)$ under appropriate conditions.

LEMMA 4.1. *Let $Y(x)$, $EY(x) \equiv 0$, be a Gaussian process with bounded, uniformly continuous covariance function $r(x, y)$. If $a(x)$ is a piecewise smooth integrable function, the quadratic functional* (4.6) *has characteristic function formally given by*

$$E(e^{itZ}) = \exp\{\textstyle\sum_{k=1}^{\infty} 2^{k-1}(it)^k c_k/k\} \tag{4.7}$$

with

$$c_k = \int \cdots \int r(x_1, x_2)r(x_2, x_3) \cdots r(x_k, x_1)a(x_1)a(x_2) \cdots a(x_k)\,dx_1 \cdots dx_k\,.$$

The representation (4.6) is valid for $|t| < 1/2M$ where $M = ||r|| \int |a(t)|\,dt$. The quantities $(k-1)!\,2^{k-1}c_k$ are of course the cumulants of (4.6).

The lemma is obtained by considering the form

$$\textstyle\sum_{j=1}^{n} \bar{Y}_j^2 a_j \tag{4.8}$$

in jointly Gaussian random variables $\bar{Y}_j$, $E\bar{Y}_j \equiv 0$ with the a_j's constants. Let R be the covariance matrix of the $\bar{Y}_j$'s with A the diagonal matrix with diagonal entries a_j. The characteristic function of (4.8) is then

$$|1 - 2itRA|^{-\frac{1}{2}} = \textstyle\prod_{j=1}^{n}(1 - 2\lambda_j it)^{-\frac{1}{2}} = \exp\{\sum_{k=1}^{\infty} 2^{k-1}(it)^k \operatorname{tr}(RA)^k/k\}\,,$$

at least if $|t| < 1/2 \operatorname{tr}(RA)$.

Here $\operatorname{tr}(M)$ denotes the trace of M, $|M|$ its determinant and $\lambda_1, \cdots, \lambda_n$ are the eigenvalues of RA. Lemma 4.1 is then obtained by going through an appropriate limiting operation.

The covariance function of the Gaussian process ${}_2L_n(x)$ can be written

$$\begin{aligned} r(x, y) &= \int w(z)w(\alpha + z)f(x - b(n)z)\,dz \\ &= f(x) \int w(z)w(\alpha + z)\,dz + O(b(n)) \end{aligned} \tag{4.9}$$

where

$$\alpha = (y - x)/(b(n))$$

and $O(b(n))$ is independent of x if f is bounded and has a uniformly bounded derivative and $w^2(z)(1 + |z|)$ is integrable. Then

$$E(\int {}_2L_n(x)^2 a(x)\,dx) = \int f(x)a(x)\,dx \int w(z)^2\,dz + O(b(n))\,. \tag{4.10}$$

Similarly if a is bounded as well as integrable and w is bounded and f is as above, the variance of $\int {}_2L_n{}^2(x)a(x)\,dx$ is $2b(n) \int [w * \bar{w}(u)]^2\,du \int a^2(x)f^2(x)\,dx$ to first order as $n \to \infty$, where $\bar{w}(t) = w(-t)$ and $*$ denotes convolution. A similar argument shows that under the same conditions the kth cumulant of $\int {}_2L_n{}^2(x)a(x)\,dx$ equals to first order $(k-1)!\,2^{k-1}b^{k-1}(n)[w * \bar{w}]^{(k)}(0) \int a^k(x)f^k(x)\,dx$ as $n \to \infty$ where the superscript (k) indicates that $w * \bar{w}$ is convoluted with itself k times. As a result we have the following theorem which actually holds under the weaker assumptions indicated above.

THEOREM 4.1. *Let* A1—A3 *hold and suppose that a is integrable piecewise continuous and bounded. Suppose moreover that* (2.16) *holds. Then* $b^{-\frac{1}{2}}(n)(T_n - (\int f(x)a(x)\,dx) \int w^2(z)\,dz)$ *is asymptotically normally distributed with mean* 0 *and variance* $2(w * \bar{w})^{(2)}(0) \int a^2(x)f^2(x)\,dx$ *as* $n \to \infty$.

A particular case of interest for the application of the theorem is that in which as in Section 3, $a(x)$ vanishes off an interval, say $[0, 1]$, and one sets $a(x) = f(x)^{-1}$ on $[0, 1]$. In this case under A1—A3, T_n is asymptotically Gaussian with mean $\int w^2(z)\,dz$ and variance $2b(n)(w * \bar{w})^{(2)}(0)$.

The statistic

$$\tilde{T}_n = nb(n) \int [f_n(x) - f(x)]^2 a(x)\,dx \tag{4.11}$$

is probably of greater interest than that considered in Theorem 4.1. However, let us expand $\tilde{T}_n$ in the form

$$\begin{aligned} nb(n)\{&\int [f_n(x) - Ef_n(x)]^2 a(x)\,dx \\ &+ 2 \int [f_n(x) - Ef_n(x)][Ef_n(x) - f(x)]a(x)\,dx \\ &+ \int [Ef_n(x) - f(x)]^2 a(x)\,dx\}. \end{aligned} \tag{4.12}$$

Let w be positive and symmetric about zero with

$$c = \int w(u)u^2\,du < \infty\,. \tag{4.13}$$

Then if $n^{-1} = O(b(n))$, $b(n) \to 0$ as $n \to \infty$, A1 holds and f has a continuous bounded second derivative, the second term of (4.12) may, by the usual approximation arguments, be shown to be asymptotically normal with mean zero and variance

$$n^{-1}b(n)^4 c^2 \int f''(x)^2 a(x)^2 f(x)\,dx \tag{4.14}$$

to the first order. Also, under the same conditions, the last term of (4.12) can be shown to be

$$b(n)^4 c^2 \int f''(x)^2 a(x)\,dx \tag{4.15}$$

to the first order. Then $[b(n)]^{-\frac{1}{2}}[\tilde{T}_n - T_n] = o_p(1)$ if and only if $b(n) = o(n^{-\frac{2}{9}})$. (The term (4.14) is then negligible.) The theorem quoted in the introduction follows.

APPLICATIONS. An explicit confidence band is hard to obtain from Theorem 4.1 and the theorem of the introduction. However we can test $H: f = f_0$ at (approximate) level α by calculating T_n for $f = f_0$ and rejecting when $T_n \geqq d(\alpha)$ where by Theorem 4.1

$$\begin{aligned} d(\alpha) = &[\int f_0(x)a(x)\,dx][\int w^2(z)\,dz] \\ &+ b^{\frac{1}{2}}(n)\Phi^{-1}(1 - \alpha)/[2(w * \bar{w})^{(2)}(0) \int a^2(x)f_0^2(x)\,dx]^{\frac{1}{2}}\,. \end{aligned} \tag{4.16}$$

As in Section 3 it is easy to see that in testing $H: f = f_0(\cdot, \theta)$ where θ is an unknown vector parameter we may use T_n with f replaced by $f_0(\cdot, \hat{\theta})$ and $d(\alpha)$ with f_0 replaced by $f_0(\cdot, \hat{\theta})$, provided that A6 below holds.

A6. For each θ_0, $(\partial^2 f(x, \theta)/\partial\theta^{(i)}\partial\theta^{(j)})$ is bounded in absolute value for all θ in a neighborhood of θ_0 and all x, i, j. Moreover, if θ_0 is true,

$$|\hat{\theta} - \theta_0| = o_p([nb(n)]^{-\frac{1}{2}}) . \tag{4.17}$$

To see this, taking $k = 1$ for simplicity, expand as in (4.12) and note that it suffices to show that

$$\int [f_n(x) - E_{\theta_0}(f_n(x))][E_{\theta_0}(f_n(x)) - E_{\hat{\theta}}(f_n(x))]a(x)\,dx = o_p([nb^{\frac{1}{2}}(n)]^{-1}) \tag{4.18}$$

and

$$\int [E_{\theta_0}(f_n(x)) - E_{\hat{\theta}}(f_n(x))]^2 a(x)\,dx = o_p([nb^{\frac{1}{2}}(n)]^{-1}) . \tag{4.19}$$

Taylor expanding the integral in (4.18) about θ_0 we obtain a first term

$$(\hat{\theta} - \theta_0) \int [f_n(x) - E_{\theta_0}(f_n(x))]\left[\int \frac{\partial f(x + b(n)z, \theta)}{\partial\theta}\bigg|_{\theta=\theta_0} w(z)\,dz\right] a(x)\,dx$$

which is $O_p(|\hat{\theta} - \theta_0|n^{-\frac{1}{2}})$, and a second term which is $O_p([nb(n)]^{-\frac{1}{2}}(\hat{\theta} - \theta_0)^2)$, and (4.18) follows. A similar argument yields (4.19).

To make local power calculations we again suppose g_n is as in (3.15) with g_n satisfying A2—A3 uniformly in n and $o(\gamma_n)$ uniform in x and η is bounded.

THEOREM 4.2. *Let g_n be as above, w satisfy* A1—A4, *a be integrable piecewise continuous and bounded, $b(n) = n^{-\delta}$, $\delta < \frac{1}{4}$, $\gamma_n = n^{-\frac{1}{2}+\delta/4}$. Define T_n in terms of f_0. Then,*

$$b^{-\frac{1}{2}}(n)(T_n - [\textstyle\int f_0(x)a(x)\,dx] \int w^2(z)\,dz) \tag{4.20}$$

is asymptotically normally distributed with mean $\int \eta^2(x)a(x)\,dx$ and variance

$$2(w * \tilde{w})^{(2)}(0) \int a^2(x)f_0^2(x)\,dx .$$

The proof is straightforward. As in Section 3 it follows that the test which rejects when T_n is $\geqq d(\alpha)$ is locally strictly unbiased if $a(x) > 0$ for all x.

Also as before the asymptotics lead to choosing δ as large as possible and again this conclusion is shaken if one uses the better approximation to the asymptotic mean, $\int \eta_n{}^2(x)a(x)\,dx$ where

$$\eta_n(x) = \int w(z)[g_n(x + b(n)z) - f_0(x + b(n)z)]\,dz . \tag{4.21}$$

It is also clear that for fixed δ we can let $\lambda_n \to 0$ more quickly than for the sup functional and still get power. Thus the Pitman efficiency of the T_n test to the $\tilde{M}_n$ test for the same δ is ∞.

Suppose that f_0 is the uniform density on [0, 1] an effect we can always achieve by applying the probability integral transformation to our observations before making the test. Let $a(x) = 1$ on [0, 1] and 0 otherwise, w be the uniform density on $[-\frac{1}{2}, \frac{1}{2}]$. Neglecting fringe effects we may then write

$$T_n = \int_0^1 \frac{(N[t - \frac{1}{2}b(n), t + \frac{1}{2}b(n)] - b(n))^2}{nb(n)}\, dt \tag{4.22}$$

where $N[x, y]$ is the number of observations falling in the interval $[x, y]$. A related statistic for testing uniformity on the circle was considered by Watson in [15]. This is, of course, very similar to the χ^2 statistic for the problem based on the cells $[0, b(n)]$, $[b(n), 2b(n)]$, $\cdots$, $[(K - 1)\frac{1}{2}b(n), (K + 1)\frac{1}{2}b(n)]$ given by,

$$\chi_n^2 = \sum_{k=1}^{*K} \frac{(N[\frac{1}{2}kb(n) - \frac{1}{2}b(n), \frac{1}{2}kb(n) + \frac{1}{2}b(n)] - nb(n))^2}{nb(n)} \tag{4.23}$$

where $(K + 1)\frac{1}{2}b(n) \leqq 1 < (K + 2)\frac{1}{2}b(n)$ and $\sum^*$ is a sum over odd index.

Now we can write,

$$\chi_n^2/K = nb(n) \int_0^1 (f_n(t) - E(f_n(t)))^2\, dA_n(t) \tag{4.24}$$

where A_n places mass $1/K$ at each of the points $\frac{1}{2}b(n)$, $\cdots$, $K\frac{1}{2}b(n)$. It is easy to see that the arguments leading to Theorem 4.2 apply to functionals of this type also and that under the conditions of that theorem, if $b(n) = n^{-\delta}$, $\delta < \frac{1}{2}$, χ_n^2/K is asymptotically normal with the natural parameters $E(\chi_n^2/K)$ and $\mathrm{Var}\,(\chi_n^2/K)$.

This result is, of course, known. A rigorous proof under milder conditions but using a different method may be found in Steck (1957). Now

$$\begin{aligned} E\left(\frac{\chi_n^2}{K}\right) &= 1 + \frac{1}{K} \sum_{j=1}^{*K} nb(n) \left(1 - \frac{1}{b(n)} \int_{(j-1)b(n)/2}^{(j+1)b(n)/2} g_n(x)\, dx\right)^2 \\ &= 1 + nb(n)\gamma_n^2 \frac{1}{K} \sum_{j=1}^{*K} \left[\frac{1}{b(n)} \int_{(j-1)b(n)/2}^{(j+1)b(n)/2} \eta(x)\, dx\right]^2 + o(nb(n)\gamma_n^2) \end{aligned} \tag{4.25}$$

$$\mathrm{Var}\left(\frac{\chi_n^2}{K}\right) = \frac{1}{K} \mathrm{Var}\left(\frac{N^2[0, b(n)]}{nb(n)}\right) + o\left(\frac{1}{K}\right) = \frac{2}{K} + o\left(\frac{1}{K}\right). \tag{4.26}$$

Thus if we take $\gamma_n = n^{-\frac{1}{2}+\delta/4}$ as in Theorem 4.2, under g_n the statistics

$$W_n = b^{-\frac{1}{2}}(n) \left(\frac{\chi_n^2}{K} - 1\right) \tag{4.27}$$

have a limiting Gaussian distribution with mean $\int_0^1 \eta^2(x)\, dx$ and variance 2. Under the same circumstances the asymptotic mean of $b^{-\frac{1}{2}}(n)(T_n - 1)$ with T_n given by (4.22) is also $\int_0^1 \eta^2(x)\, dx$ while its asymptotic variance is,

$$2w^{(4)}(0) = 2 \int_{-1}^{1} (1 - |t|)^2\, dt = \tfrac{4}{3}\,. \tag{4.28}$$

The Pitman efficiency of the tests based on T_n to those based on W_n is thus by the usual calculations,

$$e(T_n, W_n) = (\tfrac{3}{2})^{\frac{1}{2}-\delta} \tag{4.29}$$

and thus at least $(\frac{3}{2})^{\frac{1}{2}} = 1.217$ on the range $\delta > 0$. For the Mann-Wald (1942) prescription $\delta = \frac{2}{5}$ we get an efficiency of 1.292.

Although as we have seen these asymptotic calculations are to be taken with a grain of salt we feel that the procedure T_n has promise as a competitor to the χ^2 test, at least for moderate sample sizes.

Acknowledgment. We are grateful to S. Berman and J. Pickands, III for providing us with preprints of some of the papers cited in the list of references.

5. Appendix A. On the extrema of some nonstationary Gaussian processes.

Let $Y_T(\cdot)$ be a sequence of separable Gaussian processes with mean $\mu_T(\cdot)$ such that $Y_T(\cdot) - \mu_T(\cdot)$ is stationary. Let $r(\cdot)$ be the covariance function of Y_T,

$$M_T = \max\{Y_T(t) : 0 \leqq t \leqq T\}, \qquad m_T = \min\{Y_T(t) : 0 \leqq t \leqq T\}.$$

Let $b_T(t) = \mu_T(t)(2 \log T)^{\frac{1}{2}}$.

THEOREM A1. *Suppose that,*

(i) $b_T(t)$ *is uniformly bounded in* t *and* T *on* $[0, T]$ *as* $T \to \infty$.

(ii) $b_T(t) \to b(t)$ *uniformly on* $[0, T]$ *as* $T \to \infty$.

(iii) $T^{-1}\lambda[t : b(t) \leqq x, 0 \leqq t \leqq T] \to \eta(x)$ *the* cdf *of a probability measure as* $T \to \infty$. *(λ as usual denotes Lebesgue measure.)*

(iv) $b(\cdot)$ *is uniformly continuous on* R.

(v) $r(t) = 1 - C|t|^\alpha + o(|t|^\alpha)$, $0 < \alpha \leqq 2$, *as* $t \to \infty$.

(vi) $\int_0^\infty r^2(t)\,dt < \infty$.

Let

$$B(t) = (2 \log t)^{\frac{1}{2}} + \frac{1}{(2 \log t)^{\frac{1}{2}}} \times \left\{\left(\frac{1}{\alpha} - \tfrac{1}{2}\right)\log\log t + \log(2\pi)^{-\frac{1}{2}}(C^{1/\alpha}H_\alpha 2^{(2-\alpha)/2\alpha})\right\}$$

where

$$H_\alpha = \lim_{T\to\infty} \frac{1}{T} \int_0^\infty e^s P[\sup_{0\leqq t\leqq T} Y(t) > s]\,ds$$

and Y *is a Gaussian process with,*

$$E(Y(t)) = -|t|^\alpha, \qquad \operatorname{Cov}(Y(t_1), Y(t_2)) = |t_1|^\alpha + |t_2|^\alpha - |t_1 - t_2|^\alpha. \tag{A.1}$$

Then,

$$U_T = (2 \log T)^{\frac{1}{2}}(M_T - B(T)) \qquad \textit{and} \qquad V_T = -(2 \log T)^{\frac{1}{2}}(m_T + B(T))$$

are asymptotically independent with,

$$P[U_T < z] \to e^{-\lambda_1 e^{-z}}, \qquad P[V_T < z] \to e^{-\lambda_2 e^{-z}}; \tag{A.2}$$

where,

$$\lambda_1 = \int e^z\,d\eta(z), \qquad \lambda_2 = \int e^{-z}\,d\eta(z). \tag{A.3}$$

An immediate consequence of Theorem A1 is, Corollary A1. *If $\tilde{M}_T = \max\{|Y_T(t)| : 0 \leqq t \leqq T\}$ then under the conditions of the theorem,*

$$(\text{A.4}) \qquad P[(2\log T)^{\frac{1}{2}}(\tilde{M}_T - B(T)) < x] \to \exp[-(\lambda_1 + \lambda_2)e^{-x}]\,.$$

Note. $\lambda_1 + \lambda_2 \geqq 2$ with strict inequality unless η concentrates at 0.

Corollary A2. *Let $Y_0(t) - \mu(t)$ be a stationary mean* 0 *Gaussian process with covariance function $r(t)$ satisfying the conditions of the theorem. Suppose that $b(t) = (2\log(t+2)^{\frac{1}{2}}\mu(t)$ is a bounded uniformly continuous function of t and that $b(\cdot)$ satisfies condition* (iii) *of the theorem. Then,*

$$(\text{A.5}) \qquad P[(2\log T)^{\frac{1}{2}}(\max\{Y_0(s) : 0 \leqq s \leqq T\} - B(T)) < x] \to e^{-\lambda_1 e^{-x}}\,.$$

Similar assertions hold about the independence of maximum and minimum and the asymptotic distribution of the minimum.

This corollary may be viewed as complementing Theorem 4.1 of Qualls and Watanabe (1971) which deals with the extrema of a mean 0 process whose covariance function is asymptotically locally approximated by that of a stationary process while we deal with a process which is stationary when centered and asymptotically stationary.

The constants H_1 and H_2 are the only ones known explicitly. They are given by $H_1 = 1$, $H_2 = \pi^{-\frac{1}{2}}$ (cf. [11]).

Proof of Corollary A2. Define,

$$(\text{A.6}) \qquad \begin{aligned} Y_T(t) &= Y_0(t) \qquad \text{on} \quad [\varepsilon(T), T] \\ &= Y_0(t) + ([\log(t+2)/\log(T+2)]^{\frac{1}{2}} - 1)\mu(t) \quad \text{otherwise} \end{aligned}$$

where $\varepsilon(T) = o(T)$, $\log \varepsilon(T) \sim \log T$. Evidently, $(2\log T)^{\frac{1}{2}}E(Y_T(t)) \to b(t)$ uniformly and

$$(\text{A.7}) \qquad \begin{aligned} &\left| P\left[\max\{Y_T(s) : 0 \leqq s \leqq T\} < \frac{x}{(2\log T)^{\frac{1}{2}}} + B(T)\right] \right. \\ &\qquad \left. - P\left[\max\{Y_0(s) : 0 \leqq s \leqq T\} < \frac{x}{(2\log T)^{\frac{1}{2}}} + B(T)\right]\right| \\ &\qquad \leqq 2P\left[\max\{Y_0(s) - E(Y_0(s)) : 0 \leqq s \leqq \varepsilon(T)\} \geqq \frac{x}{(2\log T)^{\frac{1}{2}}} \right. \\ &\qquad\qquad \left. - K + B(T)\right] \end{aligned}$$

where $K = \max\{\mu(t) : 0 \leqq t \leqq \varepsilon(t)\}$. Since $B(\varepsilon(T)) - B(T) \to -\infty$ the term on the right of (A.7) tends to 0 by the theorem. □

Proof of Theorem A1. The theorem is argued much as Theorem 3.1 of Pickands (1969). We refer the reader to this paper and Berman (1971) for the details of the argument.

LEMMA A1. *Let $\psi(x) = \phi(x)/x$ where ϕ is the standard normal density. Let $C = 1$, $x = x(T) = B(T) + z_1/(2 \log T)^{\frac{1}{2}}$. Then for $a > 0$,*

$$\begin{aligned} &P[\max\{Y_T(t + akx^{-2/\alpha}), 0 \leqq k \leqq n\} > x] \\ \text{(A.8)} \qquad &\qquad = \psi(x)e^{b(t)}H_\alpha(n, \alpha) + o(\psi(x)) \\ &P[\min\{Y_T(t + kax^{-2/\alpha}), 0 \leqq k \leqq n\} < -x] \\ &\qquad = \psi(x)e^{-b(t)}H_\alpha(n, a) + o(\psi(x)) \end{aligned}$$

as $T \to \infty$ uniformly in $0 \leqq t \leqq T$ where

$$H_\alpha(n, a) = \textstyle\int_{-\infty}^{\infty} e^s P[\max\{Y(ka) : 0 \leqq k \leqq n\} > s]\, ds\,. \tag{A.9}$$

Moreover, if $y = y(T) = B(T) + z_2/(2 \log T)^{\frac{1}{2}}$ then

$$\begin{aligned} \text{(A.10)} \qquad P[\max\{Y_T(t + kax^{-2/\alpha}) : 0 \leqq k \leqq n\} > x,& \\ \min\{Y_T(t + kax^{-2/\alpha}) : 0 \leqq k \leqq n\} < -y]& \\ = o(\psi(x)) = o(\psi(y))\,,& \end{aligned}$$

uniformly in $0 \leqq t \leqq T$. (Throughout, k may take on integer values only.)

PROOF. As in [11] consider the "local" process

$$\tilde{Y}_T(s) = x(Y_T(t + sx^{-2/\alpha}) - \mu_T(t) - x)\,. \tag{A.11}$$

$$\begin{aligned} \text{(A.12)} \qquad &P[\max\{Y_T(t + akx^{-2/\alpha}) : 0 \leqq k \leqq n\} < x] \\ &\quad = \textstyle\int_{-\infty}^{\infty} \gamma(z) P[\max\{\tilde{Y}_T(ka) : 0 \leqq k \leqq n\} > -x\mu_T(t) \mid \tilde{Y}_T(0) = z]\, dz \end{aligned}$$

where γ is the density of $\tilde{Y}_T(0)$,

$$\gamma(z) = \frac{1}{x}\phi\left(x + \frac{z}{x}\right) = \psi(x)\exp[-z - z^2/2x^2]\,. \tag{A.13}$$

It is easy to see using (ii) and (iv) that the finite dimensional conditional distributions of $\tilde{Y}_T(s)$ given $\tilde{Y}_T(0) = z$ converge uniformly in t to those of the process $Y(s) + z$ where Y is given by (A.1). Arguing as in [11] the first part of (A.8) follows since $x\mu_T(t) \to b(t)$ uniformly as required. By considering $-Y_T$ we obtain the second part. To prove (A.10) let A be the event whose probability is being estimated. Then,

$$\begin{aligned} &P\left(A, Y_T(t) > x - \frac{1}{x^{\frac{1}{2}}} + \mu_T(t)\right) \\ &\qquad \leqq \textstyle\int_{-x^{\frac{1}{2}}}^{\infty} \gamma(z) P[\min\{\tilde{Y}_T(ka) : 0 \leqq k \leqq n\} - z \\ &\qquad \leqq -z - x(y + x + \mu_T(t)) \mid \tilde{Y}_T(0) = z]\, dz \\ \text{(A.14)} \qquad &\qquad \leqq \psi(x) \textstyle\int_{-\infty}^{x^{\frac{1}{2}}} e^z P[\min\{\tilde{Y}_T(ka) + z : 0 \leqq k \leqq n\} \\ &\qquad < z - x(y + x + \mu_T(t)) \mid \tilde{Y}_T(0) = -z]\, dz \\ &\qquad \leqq \psi(x) \textstyle\sum_{k=0}^{n} \{\int_{-\infty}^{0} P[\tilde{Y}_T(ka) + z < z \\ &\qquad\qquad - x(y + x + \mu_T(t)) \mid \tilde{Y}_T(0) = -z]\, dz \\ &\qquad\qquad + x^{\frac{1}{2}} \exp x^{\frac{1}{2}} \max\{P[\tilde{Y}_T(ka) + z \\ &\qquad < x^{\frac{1}{2}} - x(y + x + \mu_T(t)) \mid \tilde{Y}_T(0) = -z] : 0 \leqq z \leqq x^{\frac{1}{2}}\}\}]\,. \end{aligned}$$

Applying the usual estimate $\Phi(z) \leqq \psi(|z|)$ for $z \leqq 0$ we conclude that the left-hand side of (A.14) is $o(\psi(x))$. Similarly,

$$P\left(A,\, Y_T(t) - \mu_T(t) < -y + \frac{1}{y^{\frac{1}{2}}}\right) = o(\psi(y))\,. \tag{A.15}$$

Finally,

$$P\left(A,\, -y + \frac{1}{y^{\frac{1}{2}}} \leqq Y_T(t) - \mu_T(t) \leqq x - \frac{1}{x^{\frac{1}{2}}}\right)$$

$$\begin{aligned} &\leqq \textstyle\int_{-\infty}^{-x^{\frac{1}{2}}} \gamma(z) P[\max\{\tilde{Y}_T(ka) : 0 \leqq k \leqq n\} > -x\mu_T(t) \mid \tilde{Y}(0) = z]\, dz \\ &\leqq \psi(x) \textstyle\int_A^{\infty} e^z P[\max\{\tilde{Y}_T(ka) : 0 \leqq k \leqq n\} > z]\, dz \end{aligned} \tag{A.16}$$

for every $A < \infty$.

The final statement of the lemma follows.

LEMMA A2. *The assertion of Lemma* A1 *remains valid if* $a = 1$, k *is permitted to range over all values in* $[0, n]$ *and* $H_\alpha(n, a)$ *is replaced by*

$$\bar{H}_\alpha(n) = \cdot \textstyle\int_{-\infty}^{\infty} e^t P[\max\{Y(s) : 0 \leqq s \leqq n\} > t]\, dt \tag{A.17}$$

PROOF. We prove the analogue of (A.8); the other assertions follow similarly. We need to check that uniformly in T,

(a) The conditional distributions of the continuous processes $\tilde{Y}_T(t) - z$ given $\tilde{Y}_T(0) = z$ converge weakly (in the sense of Prohorov) to that of $Y(\cdot)$,

(b) $\qquad P[\max\{\tilde{Y}_T(k) : 0 \leqq k \leqq n\} > x\mu_T(t) \mid \tilde{Y}_T(0) = z] \leqq g(z)$

where $\int e^{-z} g(z)\, dz < \infty$.

To see that (a) holds it suffices to note that,

$$\operatorname{Var}[(\tilde{Y}_T(s_1) - \tilde{Y}_T(s_2)) \mid \tilde{Y}_T(0) = z] \leqq C|s_1 - s_2|^\alpha \tag{A.18}$$

and then apply Billingsley [3] page 95. To see that (b) is valid use the estimate of Fernique (1970) given below on the tails of $\max\{|\tilde{Y}_T(k)| : 0 \leqq k \leqq n\}$.

LEMMA. *Let* $Z(\cdot)$ *be a Gaussian process on* $(0, 1)$. *Let* a *be such that* $P[\|Z\| \leqq a] \geqq \frac{3}{4}$, $P[\|Z\| \geqq a] \geqq \frac{1}{4}$. *Then, for* $z \geqq a$

$$P[\|Z\| > z] \leqq \exp\left\{-\frac{z^2}{24a^2} \log 3\right\}.$$

LEMMA A3. *Fix* $t > 0$ *such that* $\inf\{s^{-\alpha}(1 - r(s)) : 0 \leqq s \leqq t\} \geqq A(t) > 0$. *Define* x *and* y *as before. Let,*

$$H_\alpha(a) = \lim_{n\to\infty} \frac{H_\alpha(n, a)}{n}\,. \tag{A.19}$$

$$0 < H_\alpha = \lim_{a\to 0} \frac{H_\alpha(a)}{a} = \lim_{n\to\infty} \frac{\bar{H}_\alpha(n)}{n}\,. \tag{A.20}$$

(See the note at the end of the lemma.)

Then,

$$P\left[\max\left\{Y_T(v + kax^{-2/\alpha}) : 0 \leqq k \leqq \left[\frac{x^{2/\alpha}}{a}t\right]\right\} > x\right] = x^{2/\alpha}\psi(x)\,\frac{H_\alpha(a)}{a}\int_v^{v+t} \exp b(s)\,ds + o(x^{2/\alpha}\psi(x)) , \tag{A.21}$$

$$P[\max\{Y_T(v + s) : 0 \leqq s \leqq t\} > x] = x^{2/\alpha}\psi(x)\left[\int_v^{v+t} \exp b(s)\,ds\right]H_\alpha + o(x^{2/\alpha}\psi(x)) , \tag{A.22}$$

uniformly in $0 \leqq v \leqq T$. *Similar assertions hold for* $P[\min\{Y_T(v + s) : 0 \leqq s \leqq t\} < -x]$ *with* $-b$ *replacing* b. *Finally,*

$$P[\max\{Y_T(v + s) : 0 \leqq s \leqq t\} > x, \min\{Y_T(v + s) : 0 \leqq s \leqq t\} < -y] = o(x^{2/\alpha}\psi(x)) . \tag{A.23}$$

Note. The existence of the limit in (A.19) was first proved in [11]. An incorrect proof of (A.20) was also given. Subsequently, a correct proof was communicated to the author by J. Pickands and another is included in [12]. We provide yet a third in Appendix B.

Proof. We prove (A.22); (A.21) is argued similarly. Begin by bounding the left-hand side of (A.22) from above by,

$$\sum_{k=0}^{M} P[\max\{Y_T(v + knx^{-2/\alpha} + s) : 0 \leqq s \leqq nx^{-2/\alpha}\} > x] \tag{A.24}$$

where $M = [tx^{2/\alpha}/n]$. By Lemma A2 the expression above is asymptotic to

$$\frac{t\bar{H}_\alpha(n)}{n}\,x^{2/\alpha}\psi(x)\left[\frac{1}{M+1}\sum_{k=0}^{M} \exp b(v + knx^{-2/\alpha})\right] = \frac{\bar{H}_\alpha(n)}{n}\,x^{2/\alpha}\psi(x)\left[\int_v^{v+t} \exp b(s)\,ds + o(1)\right] \tag{A.25}$$

since b is assumed uniformly continuous and bounded. On the other hand we can bound from below by the left-hand side of (A.21) which in turn is bounded from below by,

$$\sum_{r=0}^{M_a} P(A_r) - \sum_{0 \leqq r \neq s \leqq M_a} P(A_r A_s) \tag{A.26}$$

where $A_r = [\{\max\{Y_T(v + kax^{-2/\alpha}), rn \leqq k < (r + 1)n\} > x]$, $M_a = [x^{-2/\alpha}t/na]$. If we apply Lemma A.1 to the first term on the right of (A.26) we obtain that,

$$\sum_{r=0}^{M_a} P(A_r) \sim \frac{H_\alpha(n, a)}{na}\,x^{2/\alpha}\psi(x)\left[\int_v^{v+t} e^{b(s)}\,ds\right] . \tag{A.27}$$

Finally,

$$P(A_r A_s) \leqq P(C_r C_s) \tag{A.28}$$

where

$$C_r = \left[\max\{Y_T(kax^{-2/\alpha} + v) - \mu_T(kax^{-2/\alpha}) : rn \leqq k < (r + 1)n\} > x - \frac{K}{(2\log T)^{\frac{1}{2}}}\right] \tag{A.29}$$

where $K = \sup\{(2 \log T)^{\frac{1}{2}}|\mu_T(t)| : 0 \leqq t \leqq T\}$. Now applying Lemma 2.3 of [11] and arguing as in Lemma 2.5 of the same paper we see that,

$$\sum P(C_r C_s) = o(x^{2/\alpha}\psi(x)) . \tag{A.30}$$

Applying (A.20) we see that (A.22) follows. To prove (A.23) it suffices to show that,

$$\begin{aligned} &P\{[\max\{Y_T(v+s) : 0 \leqq s \leqq t\} > x] \\ &\qquad \cup [\min\{Y_T(v+s) : 0 \leqq s \leqq t\} < -y]\} \\ &\qquad = x^{2/\alpha}\psi(x)H_\alpha \textstyle\int_v^{v+t} [\exp b(\xi)]\, d\xi \\ &\qquad\qquad + y^{2/\alpha}\psi(y)H_\alpha \textstyle\int_v^{v+t} [\exp -b(\xi)]\, d\xi + o(x^{2/\alpha}\psi(x)) . \end{aligned} \tag{A.31}$$

But we can bound the expression on the left of (A.31) from above by

$$P[\max\{Y_T(v+s) : 0 \leqq s \leqq t\} > x] + P[\min\{Y_T(v+s) : 0 \leqq s \leqq t\} < -y]$$

and from below as in (A.26) where we add $A_{M_a+1}, \cdots, A_{2M_a+1}$ with $A_{M_a+j} = \{\min\{Y_T(v + kax^{-2/\alpha}) : (j-1)n \leqq k < jn\}\} < -y\}$. Now by (A.10)

$$\frac{1}{n} \textstyle\sum_{j=0}^{M_a} P(A_j A_{M_a+j+1}) = o(x^{2/\alpha}\psi(x)) . \tag{A.32}$$

Finally, again arguing as for the previous case,

$$\begin{aligned} &\frac{1}{n} \textstyle\sum_{0 \leqq j \neq k \leqq M_a} P(A_j A_k) , \quad \frac{1}{n} \textstyle\sum_{1 \leqq j \neq k \leqq M_a+1} P(A_{j+M_a} A_{k+M_a}) \quad \text{and} \\ &\qquad \frac{1}{n} \textstyle\sum_{0 \leqq j \neq k \leqq M_a} P(A_j A_{M_a+k+1}) \quad \text{are all} \quad o(x^{2/\alpha}\psi(x)) . \ \square \end{aligned} \tag{A.33}$$

The rest of the proof goes much as in Berman [1]. Neglecting fringe effects break the interval $[0, T]$ up into $2N$ intervals of which half, $W_1, \cdots, W_N$ are of length t and the others $V_1, \cdots, V_N$ of length ε so that V_i follows W_i which follows V_{i-1}, $i = 2, \cdots, N$. Of course, $N \sim T/(t + \varepsilon)$. Define x and y as in Lemma A1 and note that,

$$x^{2/\alpha}\psi(x)H_\alpha \sim \frac{1}{T} e^{-z_1} . \tag{A.34}$$

Then, by Lemma A3,

$$\begin{aligned} P[\max\{Y_T(\tau) : \tau \in \textstyle\bigcup_{j=1}^N V_j\} \geqq x] &\leqq \textstyle\sum_{j=1}^N P[\max\{Y_T(\tau) : \tau \in V_j\} \geqq x] \\ &\sim [\textstyle\sum_{j=1}^N \int_{V_j} \exp b(s)\, ds] \frac{e^{-z_1}}{T} \\ &= \varepsilon O\left(\frac{N}{T}\right) = \varepsilon O(1) \end{aligned} \tag{A.35}$$

where the O term is independent of ε and the V_j. A similar assertion holds for $\min\{Y_T(\tau) : \tau \in \bigcup_{j=1}^N V_j\}$ and hence we need only show that,

$$\begin{aligned} &\lim_{\varepsilon \to 0} \overline{\lim}_{T \to \infty} P[\max\{Y_T(\tau) : \tau \in \textstyle\bigcup_{j=1}^N W_j\} \leqq x, \\ &\qquad \min\{Y_T(\tau) : \tau \in \textstyle\bigcup_{j=1}^N W_j\} \geqq -y] \\ &\qquad\qquad = \exp -\{\lambda_1 e^{-z_1} + \lambda_2 e^{-z_2}\} , \end{aligned} \tag{A.36}$$

where λ_1, λ_2 are defined in (A.3) and the bars above and below the limit sign indicate lim sup and lim inf respectively. Next choose $a > 0$. If $W_j = [a_j, a_j + t)$, $j = 1, \cdots, N$.

$$\begin{aligned} &\Big| P[\max\{Y_T(\tau) : \tau \in \textstyle\bigcup_{j=1}^{N} W_j\} \leqq x] \\ &\qquad - P\left[Y_T(a_j + kax^{-2/\alpha}) \leqq x : 0 \leqq k \leqq \left[\frac{tx^{2/\alpha}}{a}\right], 1 \leqq j \leqq N\right]\Big| \\ &\leqq \textstyle\sum_{j=1}^{N} \Big| P[\max\{Y_T(\tau) : \tau \in W_j\} \leqq x] \\ &\qquad - P\left[\max\left\{Y_T(a_j + kax^{-2/\alpha}) : 0 \leqq k \leqq \left[\frac{tx^{2/\alpha}}{a}\right]\right\} \leqq x\right]\Big| \\ &\sim [\textstyle\sum_{j=1}^{N} \int_{W_j} \exp b(s)\, ds] x^{2/\alpha}\psi(x)\left[H_\alpha - \frac{H_\alpha(a)}{a}\right] e^{-z_1}, \end{aligned} \tag{A.37}$$

by Lemma A3.

A similar argument holds for $P[\min\{Y_T(\tau) : \tau \in \bigcup_{j=1}^{N} W_j\} \geqq -y]$ and by simple probability manipulations it follows that to prove the theorem we need only show,

$$\begin{aligned} &\lim_{a\to 0} \lim_{\varepsilon\to 0} \overline{\underline{\lim}}_T\, P\Big[-y \leqq Y_T(a_j + kax^{-2/\alpha}) \leqq x : 1 \leqq j \leqq N, \\ &\qquad 0 \leqq k \leqq \left[\frac{tx^{2/\alpha}}{a}\right]\Big] \\ &\qquad\qquad = \exp\{-[\lambda_1 e^{-z_1} + \lambda_2 e^{-z_2}]\}. \end{aligned} \tag{A.38}$$

Now in view of Lemma A3 it is easy to show that,

$$\begin{aligned} &\underline{\lim}_T \textstyle\sum_{j=1}^{N} \left(1 - P\left[-y \leqq Y_T(a_j + kax^{-2/\alpha}) \leqq x : 0 \leqq k \leqq \left[\frac{tx^{2/\alpha}}{a}\right]\right]\right) \\ &\qquad = \frac{H_\alpha(a)}{aH_\alpha} \underline{\lim} \frac{1}{T} \textstyle\sum_{j=1}^{N} \int_{W_j} \{\exp[b(s) - z_1] + \exp -[b(s) + z_2]\}\, ds. \end{aligned} \tag{A.39}$$

Since, by the boundedness of b, $T^{-1}[\sum_{j=1}^{N} \int_{W_j} \exp b(s)\, ds - \int_0^T \exp b(s)\, ds] = O(\varepsilon)$ uniformly in T it follows from (A.39) and (A.20) that

$$\begin{aligned} &\lim_{a\to 0} \lim_{\varepsilon\to 0} \underline{\lim}_T \textstyle\sum_{j=1}^{N} \Big(1 - P\Big[-y \leqq Y_T(a_j + kax^{-2/\alpha}) \\ &\qquad \leqq x : 0 \leqq k \leqq \left[\frac{tx^{2/\alpha}}{a}\right]\Big]\Big) \\ &\qquad\qquad = \lambda_1 e^{-z_1} + \lambda_2 e^{-z_2}. \end{aligned} \tag{A.40}$$

Let E_j, $j = 1, \cdots, N$ be the events whose probabilities are being summed in (A.40). The assertion (A.38) corresponds to a limiting statement about $P(E_1 \cdots E_N)$. If the E_j were independent assertion (A.38) would follow readily from (A.40). Let $\tilde{P}$ be the measure which makes the vectors $(Y_T(a_1), Y_T(a_1 + ax^{-2/\alpha}), \cdots, Y_T(a_1 + ax^{-2/\alpha}[tx^{2/\alpha}/a]))$, $(Y_T(a_2), \cdots, Y_T(a_2 + ax^{-2/\alpha}[tx^{2/\alpha}/a]))$, $\cdots$, $(Y_T(a_N), \cdots, Y_T(a_N + ax^{-2/\alpha}[tx^{2/\alpha}/a])$ independent and otherwise agrees with P.

To conclude the proof of the theorem we need to show that.

$$(A.41) \qquad \lim_{\varepsilon\to 0} \overline{\lim}_T |(P - \tilde{P})(E_1 \cdots E_N)| = 0 .$$

To do this apply the following modification of Lemma 4.1 of [1].

LEMMA A4. *Let*

$$(A.42) \qquad \phi(x, y, p) = \frac{1}{2\pi(1-p^2)^{\frac{1}{2}}} \exp -\frac{(x^2 - 2pxy + y^2)}{2(1-p^2)} .$$

Let $\Sigma_1 = |r_{ij}|$, $\Sigma_2 = |s_{ij}|$ *be* $k \times k$ *nonnegative semi-definite matrices with* $r_{ii} = s_{ii} = 1$ *for all* i. *Let* $\mathbf{X} = (X_1, \cdots, X_k)$ *be a mean* 0 *Gaussian vector with covariance matrix* Σ_1 *or* Σ_2. *Let* $u_1, \cdots, u_k$ *be nonnegative numbers and* $u = \min_j u_j$. *Then,*

$$(A.43) \qquad |P_{\Sigma_1}[X_j \leqq u_j, 1 \leqq j \leqq k] - P_{\Sigma_2}[X_j \leqq u_j, 1 \leqq j \leqq k]| \\ \leqq 4 \textstyle\sum_{i,j} |\int_{s_{ij}}^{r_{ij}} \phi(u, u; \lambda)\, d\lambda| .$$

PROOF. By the usual argument (see [1] page 931) the left-hand side of (A.43) is bounded by, $4 \sum_{i,j} |\int_{s_{ij}}^{r_{ij}} \phi(u_i, u_j; \lambda)\, d\lambda|$. But, by an elementary inequality

$$(A.44) \qquad x^2 - 2pxy + y^2 \geqq \frac{(1-p)}{2} (x+y)^2 .$$

Thus,

$$(A.45) \qquad \phi(u_i, u_j, \lambda) \leqq \phi\left(\frac{u_i + u_j}{2}, \frac{u_i + u_j}{2}, \lambda\right) \leqq \phi(u, u, \lambda) . \qquad \square$$

Take $X_1 = Y_T(a_1) - \mu_T(a_1)$, $X_2 = -Y_T(a_1) + \mu_T(a_1)$ etc., $k = 2N[tx^{2/\alpha}/a]$, $|r_{ij}|$ corresponding to the distribution of $\mathbf{X}$ under P, $|s_{ij}|$ corresponding to $\tilde{P}$, $u_1 = x - \mu_T(a_1)$, $u_2 = y + \mu_T(a_1)$ etc. Evidently,

$$(A.46) \qquad u = (2 \log T)^{\frac{1}{2}} + O((\log T)^{-\frac{1}{2}}) .$$

It is clear now that we can apply to the bound of (A.43) exactly the same analysis as that given by Berman on pages 933-936 of [1] to arrive at the conclusion of the theorem.

Note. By applying the more refined analysis of Pickands [11] pages 64-72 we can show that the conclusion of the theorem also holds if (vi) is replaced by,

$$(A.47) \qquad \lim_{t\to\infty} r(t) \log t = 0 .$$

Unfortunately, the analysis of Berman appears to only yield the conclusion under the stronger

$$(A.48) \qquad r(t)[\log t]^{2/\alpha} \to 0 .$$

We do not enter into this further since (vi) is what we need for Theorems 1.1 and 1.2.

5. Appendix B. Miscellanea.

THEOREM B1. *Let* w *be an absolutely continuous square integrable function with* *a square integrable derivative w'. Let,*

$$r(t) = \int w(t+s)w(s)\,ds\,. \tag{B.1}$$

Then r is twice differentiable and

$$r''(t) = -\int w'(t+s)w'(s)\,ds\,. \tag{B.2}$$

PROOF. We first show that

$$r'(t) = \int w'(t+s)w(s)\,ds = \int w(s-t)w'(s)\,ds\,. \tag{B.3}$$

Let $\hat{w}$, $\hat{w}'$ be the Fourier transforms of w, w'. Then by Parseval,

$$\frac{r(t+h)-r(t)}{h} = \frac{1}{2\pi}\int \frac{(e^{-i(t+h)u} - e^{-itu})}{h}\,|\hat{w}(u)|^2\,du\,. \tag{B.4}$$

Applying the dominated convergence theorem we obtain the existence of r' given by

$$r'(t) = -\frac{i}{2\pi}\int e^{-itu}u|\hat{w}(u)|^2\,du = \int w'(t+s)w(s)\,ds\,.$$

Similarly

$$\begin{aligned}\frac{r'(t+h)-r'(t)}{h} &= \int \frac{w(s-t-h)-w(s-t)}{h}\,w'(s)\,ds\\ &= \frac{i}{2\pi}\int \left(\frac{e^{i(t+h)u} - e^{itu}}{h}\right)u|\hat{w}(u)|^2\,du\\ &\to -\frac{1}{2\pi}\int e^{itu}u^2|\hat{w}(u)|^2\,du = -\int w'(s-t)w'(s)\,ds\,.\end{aligned} \tag{B.5}$$

The theorem follows. Note that $r'(0) = 0$ from (B4) since $|\hat{w}|$ is symmetric.

THEOREM B2. *Let w be absolutely continuous on $[-A, A]$ and 0 otherwise. Then r has left and right derivatives at 0 and*

$$r_+'(0) = -r_-'(0) = -\tfrac{1}{2}(w^2(A) + w^2(-A))\,. \tag{B.6}$$

PROOF. Write, for $h > 0$,

$$\begin{aligned}&\int \frac{w(s+h)-w(s)}{h}\,w(s)\,ds\\ &= \int_{-A}^{A-h}\left[\frac{1}{h}\int_s^{s+h} w'(z)\,dz\right]w(s)\,ds - \frac{1}{h}\int_{A-h}^{A} w^2(s)\,ds\\ &\to \int_{-A}^{A} w'(s)w(s)\,ds - w^2(A) = -\tfrac{1}{2}(w^2(A) + w^2(-A))\end{aligned} \tag{B.7}$$

by arguing as in Theorem A1 and using Lebesgue's theorem. Since $r(-t) = r(t)$ the result follows.

THEOREM B3. (Pickands) *If $H_\alpha(n, a)$, $\bar{H}_\alpha(n)$ are defined as in* (A.17), (A.19) *then* (A.20) *holds.*

PROOF. Suppose first that $0 < \alpha < 2$. Let for $\gamma > 0$,

$$\bar{H}_\alpha(n, \gamma) = \int_{-\infty}^{\infty} e^s[\max_{0\leq t\leq n} Y(t) > s + \gamma]\,ds = e^{-\gamma}\bar{H}_\alpha(n)\,. \tag{B8}$$

Then

$$\begin{aligned} \frac{1}{n}\,&|H_\alpha(n, a) - \bar{H}_\alpha(na, \gamma)| \\ &\leqq \frac{1}{n}\,[\textstyle\int_{-\infty}^{\infty} e^s P[\max_{0\leqq t\leqq na} Y(t) > s + \gamma,\ \max_{0\leqq k\leqq n} Y(ka) \leqq s]\, ds \\ &\qquad + \textstyle\int_{-\infty}^{\infty} e^s P[s < \max_{0\leqq t\leqq na} Y(t) \leqq s + \gamma]\, ds] \\ &\leqq \frac{1}{n} \textstyle\sum_{k=0}^{n-1} \int_{-\infty}^{\infty} e^s P[Y(ka) \leqq s,\ \max_{ka\leqq t\leqq (k+1)a} Y(t) > s + \gamma]\, ds \\ &\qquad + \frac{1}{n}\,[\bar{H}_\alpha(na) - \bar{H}_\alpha(na, \gamma)]\,. \end{aligned} \tag{B.9}$$

If the summands on the right of the first term of (B.9) are denoted by $A(k, \gamma, a)$ then,

$$\begin{aligned} A(k, \gamma, a) = \textstyle\int_{-\infty}^{\infty} e^s \int_{-\infty}^{s} \tau(z, ka) \\ \times P[\max_{0\leqq t\leqq a} Y(t + ka) > s + \gamma \mid Y(ka) = z]\, dz\, ds \end{aligned} \tag{B.10}$$

where $\tau(z, ka)$ is the density of $Y(ka)$. After some manipulation we obtain

$$\begin{aligned} A(k,\gamma,a) = \textstyle\int_{-\infty}^{\infty} \phi(w) \int_0^\infty e^s P[\max_{0\leqq t\leqq a} (Y(t+ka) - Y(ka)) > s + \gamma \mid Y(ka) \\ = w + (ka)^\alpha]\, ds\, dw\,. \end{aligned} \tag{B.11}$$

As $k \to \infty$, the finite dimensional conditional distributions of $Y(t + ka) - Y(ka)$ given $Y(ka) = w + (ka)^\alpha$ tend for each w to those of $Y(t)$, $0 \leqq t \leqq a$. Arguing as in Lemma A1 we conclude that,

$$\lim_k A(k, \gamma, \alpha) = A(\gamma, \alpha) = \textstyle\int_0^\infty e^s P[\max_{0\leqq t\leqq a} Y(t) > s + \gamma]\, ds\,. \tag{B.12}$$

Let $Y^*(t) = Y(t) + |t|^\alpha$. Then,

$$\begin{aligned} A(\gamma, \alpha) &\leqq \textstyle\int_0^\infty e^s P[\max_{0\leqq t\leqq a} Y^*(t) > s + \gamma]\, ds \\ &= \textstyle\int_0^\infty e^s P[\max_{0\leqq t\leqq 1} Y^*(t) > (s + \gamma)a^{-\alpha/2}]\, ds \\ &= a^{\alpha/2} e^{-\gamma} \textstyle\int_{\gamma a^{-\alpha/2}}^{\infty} e^{wa^{\alpha/2}} P[\max_{0\leqq t\leqq 1} Y^*(t) > w]\, dw\,. \end{aligned} \tag{B.13}$$

Applying Fernique's estimate the right-hand side of (B.13) is $O(\exp -a^{-\alpha/2})$ for every $\gamma > 0$. We conclude that,

$$\begin{aligned} \limsup_a \limsup_n \frac{1}{na}\,&|H_\alpha(n, a) - \bar{H}_\alpha(na, \gamma)| \\ &\leqq (1 - e^{-\gamma}) \limsup_a \limsup_n \frac{\bar{H}_\alpha(na)}{na} \end{aligned} \tag{B.14}$$

for every $\gamma > 0$. Since,

$$\begin{aligned} P[\max_{0\leqq t\leqq n} & Y(t) > s] \\ &\leqq \textstyle\sum_{k=0}^{n-1} P[\max_{k\leqq t\leqq k+1} Y(t) > s] \\ &\leqq \textstyle\sum_{k=0}^{n-1} \{P[Y(k) \leqq s,\ \max_{k\leqq t\leqq k+1} Y(t) > s] + P[Y(k) > s]\}\,, \end{aligned} \tag{B.15}$$

it is easy to see that,

$$\text{(B.16)} \qquad \sup_{x \geq 1} \frac{\bar{H}_\alpha(x)}{x} < \infty \, .$$

Hence,

$$\text{(B.17)} \qquad \lim_a \limsup_n \frac{1}{na} |H_\alpha(n, a) - \bar{H}_\alpha(na)| = 0 \, .$$

But from the argument of Lemma A3 it is clear that for every $a > 0$,

$$\text{(B.18)} \qquad \limsup_n \frac{H_\alpha(n, a)}{na} \leqq \liminf_n \frac{\bar{H}_\alpha(na)}{na} \, .$$

The theorem follows for $0 < \alpha < 2$. For $\alpha = 2$ we can use the representation $Y(t) = 2^{\frac{1}{2}}tZ - t^2$ where Z is a standard normal deviate. Evidently,

$$\text{(B.19)} \qquad \max_{0 \leqq s \leqq na/2^{\frac{1}{2}}} Y(s) = \frac{Z^2}{2} \quad \text{if} \quad 0 \leqq Z < \frac{na}{2^{\frac{1}{2}}}$$

$$= naZ - \frac{n^2a^2}{2} \quad \text{otherwise} \, .$$

It follows that,

$$\text{(B.20)} \qquad \frac{1}{na} \left| \bar{H}_2\left(\frac{na}{2^{\frac{1}{2}}}\right) - H_2\left(n, \frac{a}{2^{\frac{1}{2}}}\right) \right|$$

$$\leqq \frac{1}{na} \int_0^{n^2a^2} e^{s/2} P[s^{\frac{1}{2}} < Z < (s + a^2)^{\frac{1}{2}}] \, ds \sim 2(1 - e^{-a^2/2})$$

by standard arguments. The theorem now follows generally.

REFERENCES

[1] Berman, S. N. (1971). Asymptotic independence of the numbers of high and low level crossings of stationary Gaussian processes. *Ann. Math. Statist.* **42** 927-946.

[2] Berman, S. N. (1971). Maxima and high level excursions of stationary Gaussian processes. *Trans. Amer. Math. Soc.* To appear.

[3] Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

[4] Breiman, L. (1969). *Probability.* Addison-Wesley, Reading.

[5] Brillinger, D. (1969). An asymptotic representation of the sample distribution function. *Bull. Amer. Math. Soc.* **75** 545-547.

[6] Doob, J. L. (1953). *Stochastic Processes.* Wiley, New York.

[7] Epanechnikov, V. A. (1969). Nonparametric estimates of a multivariate probability density. *Theor. Probability Appl.* **14** 153-158.

[8] Fernique, X. (1970). Intégrabilité des vecteurs gaussiens. *C.R. Acad. Sci. Paris* **270** A 1698-99.

[9] Hajek, J. and Sidak, Z. (1967). *Theory of Rank Tests.* Academic Press, New York.

[10] Mann, H. B. and Wald, A. (1942). On the choice of the number of class intervals in the application of the χ^2 test. *Ann. Math. Statist.* **13** 306-317.

[11] Pickands, J. S. III (1960). Upcrossing probabilities for Gaussian processes. *Trans. Amer. Math. Soc.* **145** 51-73.

[12] Qualls, C. and Watanabe, H. (1971). Asymptotic properties of Gaussian processes. Tech. Report No. 736, Univ. of North Carolina, Chapel Hill.

[13] ROSENBLATT M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815–1842.

[14] STECK, G. P. (1957). Limit theorems for conditional distributions. *Univ. California Publ. Statist.* **2** 237–284.

[15] WATSON, G. S. (1967). Some problems in the statistics of directions. *Bull. I.S.I.* **17** 374–385.

[16] WOODROOFE, M. (1967). On the maximum deviation of the sample density. *Ann. Math. Statist.* **38** 475–481.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA
LA JOLLA, CALIFORNIA 92037

Sankhyā : The Indian Journal of Statistics
1988, Volume **50**, Series A, Pt. 3, pp. 381-393.

ESTIMATING INTEGRATED SQUARED DENSITY DERIVATIVES : SHARP BEST ORDER OF CONVERGENCE ESTIMATES*

By P. J. BICKEL
University of California at Berkeley
and
Y. RITOV
The Hebrew University of Jerusalem

SUMMARY. Estimation of the integral of the square of a derivative of the probability density function is considered. The estimators we propose and their properties are a function of the amount of smoothness assumed. The rate of convergence of the appropriate estimator is shown to be optimal given the amount of smoothness assumed. In particular the appropriate estimator achieves the information bound when estimation at an $n^{-1/2}$ rate is possible.

1. Introduction

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d., each with distribution function F. Let $f(.)$ be the probability density function of F, $f^{(k)}$ its k-th derivative and $\theta_k(F) = \int \{f^{(k)}(x)\}^2\, dx$. These functionals appear in the asymptotic variance of the Wilcoxon statistic and in the asymptotics of the integrated M.S.E. for kernel density estimates. Discussion of the estimation of θ_k and similar parameters appear in Schweder (1975), Hasminskii and Ibragimov (1978), Pfanzagl (1982), Prakasa Rao (1983), Donoho and Liu (1987) and Hall and Marron (1987).

Ritov and Bickel (1987) show that the standard semiparametric information bound for the estimation of $\theta_0(F)$ fails to give an achievable rate of convergence. In fact, the information is strictly positive when f is bounded, promising that the $n^{-1/2}$ rate is achievable. Nevertheless, there is no rate that can be achieved uniformly in small compact neighborhoods (in the total variation norm) of a given distribution. Moreover, even if the uniformity requirement is dropped then for any sequence of estimates $\{\hat{\theta}_k\}$ there exists an (unknown) point F such that $n^{\gamma}(\hat{\theta}_k - \theta_k(F))$ doesn't converge to 0 for any $\gamma > 0$.

In this paper we consider classes of F which satisfy Hölder conditions on $f^{(m)}$ for suitable m. We establish the rate achievable under these condi-

*Research supported by Office of Naval Research N00014-80-C-0163.

AMS (1980) *subject classification :* 62G05, 62G20,

Key words and phrases : estimation, density derivatives, semiparametric information bound, sharp rate convergence.

*An invited paper to Commemorate the 50-th volume of Sankhyā.

tions and exhibit estimators that achieve these rates. Our estimators converge uniformly and when improvement is possible faster than similar estimators suggested by Schweder (1975), Hasminskii and Ibragimov (1978), and Hall and Marron (1987). In particular we need to assume weaker Hölder conditions to obtain $n^{-1/2}$ rates and efficient estimators.

We believe that our proof of the best achievable rates is novel in that it cannot be reduced to considering a sequence of simple vs. simple testing problems and in effect requires the use of composite hypotheses of growing size. Note that θ_k can be estimated at the $n^{-1/2}$ rate in any fixed regular finite dimensional submodel.

2. Main results : the estimators and their properties

Let $\theta_k(F) = \int \{f^{(k)}(x)\}^2\, dx$ where f is the (continuous) density of the distribution F. (In general we denote distribution functions by F or F_n and their densities by f or f_n respectively.) Let $\alpha > 0$, m be a nonnegative integer and $g(\cdot) \in L_2 \bigcap L_\infty$. Suppose $X_1, \ldots, X_n$ is a random sample from F. How well can $\theta_k(F)$ be estimated if it is known a priori only that $F \in \boldsymbol{F}_{m,\alpha,g}$ where $\boldsymbol{F}_{m,\alpha,g} = \{F : |f^{(m)}(x) - f^{(m)}(x+\xi)| \leqslant g(x)|\xi|^\alpha$ for all x real $|\xi| < 1\}$?

We begin by suggesting a family of estimators. Let $h_\sigma(x) = \sigma^{-1} h(x/\sigma)$ where h is a kernel with the following properties :

h is symmetric about zero,

$h(x) = 0$ for $|x| > 1$,

$\int h(x)dx = 1,$

$\int x^i h(x)dx = 0, \quad i = 1, 2, \ldots, \max\{k, m-k\}$

and h has $2k+1$ derivatives.

Divide the sample into two subsamples $X_1, \ldots, X_{n_1}$ and $X_{n_1+1}, \ldots, X_n$ with comparable sizes (i.e. n_1/n is bounded away from 0 and 1). Let $\hat{F}_1$ and $\hat{F}_2$ be the empirical distribution functions of each subsample respectively. Define, $\hat{f}_i(x) = \int h_\sigma(x-y)d\hat{F}_i(y)$, $i = 1, 2$. The dependence of $\hat{f}_i$ on σ is left implicit. Consider the following estimator of θ_0.

$$\hat{\theta}^*_0(X_1, \ldots, X_n\,; \sigma) = \frac{n_1}{n}\,\hat{\theta}^*_{01} + \frac{n_2}{n}\,\hat{\theta}^*_{02} \qquad \ldots \quad (2.1)$$

where $n = n_1 + n_2$

$$\begin{aligned}&\hat{\theta}^*_{01}(X_1, \ldots, X_n; \sigma)\\ &= \int \hat{f}_2^2(x)dx + 2n_1^{-1} \sum_{i=1}^{n_1} (\hat{f}_2(X_i) - \int \hat{f}_2^2(x)dx) + \frac{1}{n_2} \int h_\sigma^2(x)dx\\ &= 2\int h_\sigma(x-t)d\hat{F}_1(t)d\hat{F}_2(x) - n_2^{-2} \sum_{n_1+1 \leqslant i \neq j \leqslant n}{}' \int h_\sigma(x-X_i)h_\sigma(x-X_j)dx\end{aligned} \qquad \ldots \quad (2.2)$$

and $\hat{\theta}^*_{02}$ is obtained by interchanging the roles of the two subsamples in $\hat{\theta}^*_{01}$. The first two terms of $\hat{\theta}^*_{01}$ can be recgonized as Hasminskii and Ibragimov's estimate of this parameter which they show is efficient in $\mathrm{F}_{0,\alpha,M}$ if $\alpha > 1/2$. This is the, by now, familiar one step estimate (see Bickel, 1982 ; Schick, 1986) using the estimated influence function $2(\hat{f}_2 - \int \hat{f}_2^2(x)dx)$. The last term in (2.2) removes the pure known bias component, $n_2^{-2} \sum_{i=n_1+1}^{n} \int h_\sigma^2(x-X_i)dx$ from

$$\int \hat{f}_2^2(x)dx = n_2^{-2} \sum_{i,j} \int h_\sigma(x-X_i)h_\sigma(x-X_j)dx. \qquad \ldots \quad (2.3)$$

Curiously enough this simple debiasing leads to efficient estimation in $\boldsymbol{F}_{0,\alpha,M}$ for $\alpha > 1/4$ and (uniformly) $\sqrt{n}$ consistent estimation on $\boldsymbol{F}_{0,1/4,M}$. Moreover, $\sqrt{n}$ consistent estimation is shown to be impossible for $\alpha < 1/4$. More generally, if f has $2k$ continuous derivatives,

$$\begin{aligned} \theta_k(F) &= (-1)^k \int f^{(2k)}(x)f(x)dx \\ &= (-1)^k E_F(f^{(2k)}(X)). \end{aligned}$$

This suggests, by the same process as above, estimates $\hat{\theta}^*_{k1}$, $\hat{\theta}^*_{k2}$ and $\hat{\theta}^*_k$. For convenience we replace $\hat{\theta}^*_{01}$ by $\hat{\theta}_{01}$ where n_2^{-2} in (2.2) is replaced by $[n_2(n_2-1)]^{-1}$ and similar replacements are made in $\hat{\theta}^*_{02}$ and more generally $\hat{\theta}^*_k$. So the estimate we study is

$$\begin{aligned} \hat{\theta}_k(X_1, \ldots, X_n; \sigma) &= 2(-1)^k \int h_\sigma^{(2k)}(x-t)d\hat{F}_1(t)d\hat{F}_2(x) \\ &-n_2[n\, n_1(n_1-1)]^{-1} \sum_{1 \leqslant i < j \leqslant n_1} \int h_\sigma^{(k)}(x)-X_i)h_\sigma^{(k)}(x-X_j)dx \\ &-n_1[n\, n_2(n_2-1)]^{-1} \sum_{n_1+1 \leqslant i < j \leqslant n} \int h_\sigma^{(k)}(x-X_i)h_\sigma^{(k)}(x-X_j)dx. \qquad \ldots \quad (2.4) \end{aligned}$$

Our main results are summarized in the following two theorems. In the first we describe the performance of $\hat{\theta}_k$ in terms of the assumed family $\boldsymbol{F}_{m,\alpha,g}$. The rate of convergence of $\hat{\theta}_k$ to $\theta_k(F)$ is a function of $m+\alpha$ and $\hat{\theta}_k$ is "efficient" when $m+\alpha > 2k+1/4$. In the second theorem we show that the rates given in the first theorem are, essentially, the best possible.

Theorem 1 : *Let* $\{F_1, F_2, \ldots\} \subseteq \boldsymbol{F}_{m,\alpha,g}$ *where* $0 \leqslant \alpha < 1$, $m+\alpha > k$ *and* $g \in L_2 \cap L_\infty$. *Let* $X_{n1}, \ldots, X_{nn}$ *be i.i.d.,* $X_{n1} \sim F_n$ *and let* $\hat{\theta}_k = \hat{\theta}_k(X_{n1}, \ldots, X_{nn}; \sigma_n)$ *where* $\sigma_n = n^{-2/(1+4m+4\alpha)}$.

(i) *If* $m+\alpha > 2k+1/4$ *then*

$$\sqrt{n}\left[\, \hat{\theta}_k - \theta_k(F_n) - \frac{2}{n} \sum_{i=1}^{n} \{(-1)^k f_n^{(2k)}\ (X_{ni}) - \theta_k(F_n)\}\right] \longrightarrow 0. \qquad \ldots \quad (2.5)$$

Let $I_k(F_n) = [Var\{f_n^{(2k)}(X_{n1})\}]^{-1}$. *Then,* $n\, I_k(F_n)E\,\{\hat{\theta}_k - \theta_k(F_n)\}^2 \rightarrow 1$ *and* $L\{\sqrt{n}\, I_k^{1/2}(F_n)\,(\hat{\theta}_k - \theta_k(F_n))\} \rightarrow N(0, 1)$ *provided* $\limsup_n I_k(F_n) < \infty$.

(ii) *If* $k < m+\alpha \leqslant 2k+1/4$ *then* $n^{2\gamma}E\{\hat{\theta}_k-\theta_k(F_n)\}^2$ *is bounded when* $\gamma = 4(m+\alpha-k)/(1+4m+4\alpha)$.

We conjecture, but have not checked the details, that it is possible to estimate σ by cross validation to obtain an estimate $\hat{\theta}_k^* = \hat{\theta}_k(X_{n1}, \ldots, X_{nn} ; \hat{\sigma}_n)$ which does not depend on m and α but is equivalent to $\hat{\theta}_k$ which does so depend through σ_n given in the statement of Theorem 1.

Theorem 2 : (i) *The information bound (in the sense of Khoshevnik and Levit* (1976)) *for non parametric estimation of* $\theta_k(F)$, $F \in \boldsymbol{F}_{2k,\alpha,g}$ *is given by* $I_k(F)$ *as defined in Theorem* 1.

(ii) *Suppose* $k < m+\alpha \leqslant 2k+1/4$. *Then there is a small compact set* $\boldsymbol{F}^* \subsetneq \boldsymbol{F}_{m,\alpha,g}$ *such that for any* $c_n \to \infty$ *and any sequence of estimators* $T_1, T_2, \ldots, T_n = T_n(X_1, \ldots, X_n)$, $X_1, X_2, \ldots, X_n$ *iid*, $X_1 \sim F$:

$$\liminf_{n} \sup_{F \in \boldsymbol{F}^*} P_F\{c_n n^{\gamma} |T_n - \theta_k(F)| \geqslant 1\} = 1 \qquad \ldots \quad (2.6)$$

where $\gamma = 4(m+\alpha-k)/(1+4m+4\alpha)$. *Moreover* $\boldsymbol{F}^*$ *can be constructed so that its only accumulation point is any specified* $F_0 \in \boldsymbol{F}_{m,\alpha,g}$.

The proof of the first part of Theorem 2 is quite standard and follows essentially the discussion in Hasminskii and Ibragimov (1978). The proof of the second part of the Theorem is an extension of the ideas presented in Ritov and Bickel (1987). In our problem, θ_0 can be estimated at the $n^{-1/2}$ rate in any one dimensional sub model of $\boldsymbol{F}_{m,\alpha,g}$ and the information bound of Theorem 2i) is the best bound that can be achieved using these techniques. Yet for $m+\alpha < 2k+1/4$ this bound is unachievable by uniformly $n^{1/2}$ consistent estimates. In fact, for $m+\alpha < 2k+1/4$ no uniformly $n^{1/2}$ consistent estimate exists. Even uniformity can be dropped—see Ritov and Bickel (1987), Theorem 1. Our proof is based on the demonstration of a sequence of difficult multiparameter Bayesian problems.

3. Proofs

We begin the proofs with the following technical lemma whose own proof is postponed to the end of the section.

Lemma 1 : *Let* α, m *and* g *be such that* $\alpha > 0$ $m \geqslant 0$ *and* $g \in L_\infty$. *Then* $sup\{|f^{(i)}(x)| : x, F \in \boldsymbol{F}_{m,\alpha,g}\} < \infty$, $i = 0, 1, \ldots, m$.

Proof of Theorem 1 : Evidently to establish Theorem 1 it is enough to consider the asymmetric estimate

$$\hat{\theta}_{k2} = 2(-1)^k \int \int h_\sigma^{(2k)} (x-t)\, d\hat{F}_1(t)\, d\hat{F}_2(x).$$

$$- 2\{n_1(n_1-1)\}^{-1} \sum_{1 \leqslant i < j \leqslant n_1} \int h_\sigma^{(k)} (x-X_{ni})\, h_\sigma^{(k)} (x-X_{nj})\, dx.$$

We begin by estimating the conditional bias

$$E(\hat{\theta}_{k2} \,|\, \hat{F}_1)-\theta_k(F_n) = 2(-1)^k \int \hat{f}_1^{(2k)} (x)\, f_n(x)\, dx$$

$$-2\, \{n_1(n_1-1)\}^{-1} \sum_{i=1}^{n_1} \sum_{j=1}^{i-1} \int h_\sigma^{(k)} (x-X_{ni})\, h_\sigma^{(k)} (x-X_{nj})\, dx - \int \{f_n^{(k)}(x)\}^2\, dx.$$

But

$$(-1)^k \int \hat{f}_1^{(2k)}(x) f_n(x) dx = \int \hat{f}_1^{(k)}(x) f_n^{(k)}(x) dx$$

$$= n_1^{-1} \sum_{i=1}^{n_1} \int h_\sigma^{(k)}(x-X_{ni}) f_n^{(k)}(x) dx$$

$$= \{n_1(n_1-1)\}^{-1} \sum_{i=1}^{n_1} \sum_{1 \leqslant j \neq i \leqslant n_1} \int h_\sigma^{(k)}(x-X_{ni}) f_n^{(k)}(x) dx.$$

Hence

$$E(\hat{\theta}_{k2} \,|\, \hat{F}_1)-\theta_k(F_n) = -2\{n_1(n_1-1)\}^{-1} \sum_{i=1}^{n_1} \sum_{j=1}^{i-1} \int \{h_\sigma^{(k)}(x-X_{ni})-f_n^{(k)}(x)\}$$

$$\{h_\sigma^{(k)}(x-X_{nj})-f_n^{(k)} (x)\} dx. \qquad \ldots \quad (3.1)$$

We obtain from (3.1) that

$$E\, \hat{\theta}_{k2}-\theta_k(F_n) = \int \{f_{n\sigma}^{(k)}(x)-f_n^{(k)}(x)\}^2 dx \qquad \ldots \quad (3.2)$$

where $f_{n\sigma} = f_n * h_\sigma$.

But

$$f_{n\sigma}^{(k)}(x)-f_n^{(k)}(x) = \int h(t)\, \{f_n^{(k)}(x+\sigma t)-f_n^{(k)}(x)\} dt$$

$$= \int h(t) \left\{ \sum_{i=1}^{m-k-1} \frac{f_n^{(k+i)}(x)}{i\,!}\, \sigma^i t^i \right\} dt \qquad \ldots \quad (3.3)$$

$$+ \int h\,(t)\, \frac{1}{(m-k)!}\, \{f_n^{(m)}(x+\sigma^* t)-f_n^{(m)}(x)\} \sigma^{m-k} t^{m-k} dt,$$

where $0 \leqslant \sigma^* \leqslant \sigma$. The first term in the RHS of (3.3) is null by the construction of h. Since $F_n \in \boldsymbol{F}_{m,a,g}$ we can bound the integrand in the second term and obtain :

$$|f_{n\sigma}^{(k)}(x)-f_n^{(k)}(x)| \leqslant g\,(x) \sigma^{m+a-k} \int |t|^{m+a-k} |h(t)|\, dt. \qquad \ldots \quad (3.4)$$

Combine (3.2) and (3.4) to conclude that

$$|E\,\hat{\theta}_{k2}-\theta_k(F_n)| \leqslant \|g\|_2^2\, n^{-4(m+\alpha-k)/(1+4m+4\alpha)}\left(\int |t|^{m+\alpha-k}\,|h(t)|\,dt\right)^2. \quad \ldots \quad (3.5)$$

Next we estimate var $(E(\hat{\theta}_{k2}\,|\,\hat{F}_1))$. Note that $E(\hat{\theta}_{k2}\,|\,\hat{F}_1)$ was written in (3.1) as a U-statistic, $E(\hat{\theta}_{k2}\,|\,F_1)-\theta_k(F_n) = 2\{n_1(n_1-1)\}^{-1}\sum_{i=1}^{n_1}\sum_{j=1}^{i-1} U(X_{ni},\, X_{nj}{}')$ say.

By standard U-statistic theory,

$$\text{var}\,\{E(\hat{\theta}_{k2}\,|\,\hat{F}_1)\} = n^{-1}\{O(\text{var}[E\,(U\,(X_{n1},\, X_{n2})\,|\,X_{n1}])$$

$$+O\,(n^{-1}\ \text{var}\ U(X_{n1},\, X_{n2}))\}. \quad \ldots \quad (3.6)$$

Now

$$E\,U\,(x,\, X_{n2}) = \int \{h_\sigma^{(k)}(t-x)-f_n^{(k)}(t)\}\,\{f_{n\sigma}^{(k)}(t)-f_n^{(k)}(t)\}\,dt$$

$$= \int \delta(t)\,\{h_\sigma^{(k)}(t-x)-f_n^{(k)}(t)\}dt,$$

say. Hence,

$$\text{var}\ [E\{U(X_{n1},\, X_{n2})\,|\,X_{n1}\}] = E[\int\delta(x)\,\{h_\sigma^{(k)}(x-X_{n1})-f_{n\sigma}^{(k)}(x)\}\,dx]^2$$

$$= E\int\int \delta(y)\,\delta(x)\,\{h_\sigma^{(k)}(y-X_{n1})-f_{n\sigma}^{(k)}(y)\}\,\{h_\sigma^{(k)}(x-X_{n1})-f_{n\sigma}^{(k)}(x)\}\,dx\,dy$$

$$\leqslant \int\int \delta(y)\,\delta(x)\,\int h_\sigma^{(k)}(y-t)h_\sigma^{(k)}(x-t)f_n\,(t)\,dt\,dx\,dy$$

$$= \int \{\int\delta(x)h_\sigma^{(k)}(x-t)dx\}^2 f_n\,(t)\,dt$$

$$\leqslant \|\delta\|_\infty^2\,\sigma^{-2k}\,\{\int|h^{(k)}\,(x)\,|\,dx\}^2 = O(\sigma^{2(m+\alpha-2k)}) \quad \ldots \quad (3.7)$$

by (3.4). At the same time, the random variable $\int h_\sigma^{(k)}(x-X_{n1})h_\sigma^{(k)}(x-X_{n2})dx$ is bounded by $\sigma^{-2k-1}\,\|h^{(k)}\|_2^2$ and is equal to zero unless $|X_{n1}-X_{n2}| \leqslant 2\sigma$. Since f_n is bounded this last event has probability of the same order as σ.

Hence

$$\text{var}\quad \{\int h_\sigma^{(k)}(x-X_{n1})h_\sigma^{(k)}(x-X_{n2})dx\} = O\,(\sigma.\,\sigma^{-4k-2}).$$

Since $|\int f_n^{(k)}(x)\,.\,h_\sigma^{(k)}(x-X_{n1})dx| \leqslant \|f_n^{(k)}\|_\infty\,\sigma^{-k}\int\,|h^{(k)}(x)|\,dx$ we conclude that

$$\text{var}\,\{U\,(X_{n1},\, X_{n2})\}$$

$$= \text{var}[\int\{h_\sigma^{(k)}(x-X_{n1})h_\sigma^{(k)}(x-X_{n2})-f_n^{(k)}(x)h_\sigma^{(k)}(x-X_{n1})-f_n^{(k)}(x)h_\sigma^{(k)}(x-X_{n2})\}\,dx)]$$

$$= O(\sigma^{-4k-1}). \quad \ldots \quad (3.8)$$

We obtain from (3.1), (3.4), (3.6), (3.7), and (3.8) that

$$\text{var}\,\{E\,(\hat{\theta}_{k2}\,|\,\hat{F}_1)\} = O(n^{-1}\sigma^{2(m+\alpha-2k)}+n^{-2}\sigma^{-4k-1})$$

$$= O(n^{-8(m+\alpha-k)/(1+4m+4\alpha)})$$

for σ given in the statement of Theorem 1. Hence (3.5) implies that

$$E\{E\,(\hat{\theta}_{k2}\,|\,\hat{F}_1)-\theta_k(F_n)\}^2 = O\,(n^{-8(m+\alpha-k)/(1+4m+4\alpha)}). \qquad \dots \quad (3.9)$$

We have proved that $E\,(\hat{\theta}_{k2}\,|\,\hat{F}_1)-\theta_k\,(F_n)$ is of the right order (in particular it is $o_P(n^{-1/2})$ if $m+\alpha > 2k+1/4$). We turn to the investigation of the behaviour of $\hat{\theta}_{k2}-E\,(\hat{\theta}_{k2}\,|\,\hat{F}_1)$. This will be carried on separately for the two cases : $2k+1/4 < m+\alpha$ and $k < m+\alpha \leqslant 2k+1/4$.

(i) Suppose $2k+1/4 < m+\alpha$. In the light of (3.9) we need only to consider the conditional variance of $\hat{\theta}_{k2}$ given the first sub sample. But, given $X_{n1}, \dots, X_{nn_1}$, $\hat{\theta}_{k2}$ is just a sum of *i.i.d.* random variables, hence

$$\mathrm{var}\left\{\hat{\theta}_{k2}-\frac{2(-1)^k}{n-n_1}\sum_{i=n_1+1}^{n} f_n^{(2k)}\,(X_{ni})+\theta_k(F_n)\,|\,\hat{F}_1\right\}$$

$$\leqslant \frac{4}{n-n_1}\int\{\hat{f}_1^{(2k)}\,(x)-f_n^{(2k)}\,(x)\}^2 f_n\,(x)\,dx.$$

So $$E\ \mathrm{var}\,\{\hat{\theta}_{k2}-2\,(-1)^k\int f_n^{(2k)}\,(x)\,d\hat{F}_2(x)+\theta_k(F_n)\,|\,\hat{F}_1\}$$

$$\leqslant \frac{4}{n-n_1}\int\{f_{n\sigma}^{(2k)}\,(x)-f_n^{(2k)}\,(x)\}^2 f_n(x)dx+\frac{4}{n-n_1}\int\{\mathrm{var}\,\hat{f}_1^{(2k)}\,(x)\}dx.$$

$$= o_P(n^{-1}). \qquad \dots \quad (3.10)$$

Now (3.9) and (3.10) imply the validity of (2.5). Since by Lemma 1, f_n is uniformly bounded, the first part of Theorem 1 follows.

(ii) Suppose $k < m+\alpha \leqslant 2k+1/4$. We separate into two cases, $2k \leqslant m$, $2k > m$. If $2k \leqslant m$,

$$\begin{aligned} |E\hat{f}_1^{(2k)}\,(x)-f_n^{(2k)}\,(x)| &= |\int h_\sigma^{(2k)}\,(x-t)f_n\,(t)dt-f_n^{(2k)}\,(x)| \\ &= |\int f_n^{(2k)}\,(x-t)h_\sigma(t)dt-f_n^{(2k)}\,(x)| \\ &= |\int(f_n^{(2k)}\,(x-\sigma t)-f_n^{(2k)}\,(x))h(t)dt| \\ &= O(1) \end{aligned}$$

so that

$$E\hat{f}_1^{(2k)}\,(x) = O(1). \qquad \dots \quad (3.11)$$

Also,

$$\mathrm{var}\,\{\hat{f}_1^{(2k)}\,(x)\} \leqslant \frac{1}{n_1}\int\{h_\sigma^{(2k)}\,(x-t)\}^2 f_n(t)dt$$

$$\leqslant \frac{1}{n_1}\,\|f_n\|_\infty\,\sigma^{-4k-1}\,\|h^{(2k)}\|_2^2. \qquad \dots \quad (3.12)$$

Then,

$$E \text{ var } (\hat{\theta}_{k2} \,|\, \hat{F}_1) \leqslant \frac{1}{n-n_1} \int E[\{f_1^{(2k)}(x)\}^2] f_n(x)dx$$
$$= O\,(n^{-2}\,\sigma^{-4k-1}+n^{-1})$$
$$= O\,(n^{-8(m+\alpha-k)/(1+4m+4\alpha)}). \qquad \ldots \quad (3.13)$$

If $2k > m$ we compute,

$$|E\hat{f}_1^{(2k)}(x)| = |\int h_\sigma^{(2k)}\,(x-t)f_n(t)dt\,|$$
$$= |\int h_\sigma^{(2k-m)}\,(x-t)f_n^{(m)}(t)dt\,|$$
$$= \sigma^{-2k+m}\,|\int h^{(2k-m)}(t)f_n^{(m)}(x-\sigma t)dt\,|$$
$$= \sigma^{-2k+m}\,|\int h^{(2k-m)}(t)\,\{f_n^{(m)}(x-\sigma t)-f_n^{(m)}(x)\}dt\,|$$
$$\leqslant g(x)\,\sigma^{m+\alpha-2k} \int |\,h^{(2k-m))}(t)\,|\,dt \qquad \ldots \quad (3.14)$$

Again, by (3.12) and (3.14)

$$E \text{ var } (\hat{\theta}_{k2} \,|\, \hat{F}_1) = O\,(n^{-2}\,\sigma^{-(4k+1)}+n^{-1}\,\sigma^{m+\alpha-2k})$$
$$= O\,(n^{-8(m+\alpha-k)(1+4m+4\alpha)}) \qquad \ldots \quad (3.15)$$

The result follows by (3.13), (3.15) and (3.9). □

Proof of Theorem 2 : (i) Let $\{F_\nu\}$ be a sequence of distributions with densities f_ν and square root of densities s_ν. Suppose $\|s_\nu - s_0\|_2^2 \rightarrow 0$ and $\int \{f_\nu^{(2k)}(x) - f_0^{(2k)}(x)\}^2 f_0(x)dx \rightarrow 0$.

Write, with some abuse of notation, $\theta_k(s_\nu) = \theta_k(F_\nu)$. Then,

$$\theta_k(s_\nu) = \int \{f_0^{(k)}(x)\}^2 dx + 2 \int f_0^{(k)}(x)\{f_\nu^{(k)}(x) - f_0^{(k)}(x)\}dx + \int\{f_\nu^{(k)}(x) - f_0^{(k)}(x)\}^2\, dx. \qquad \ldots \quad (3.16)$$

Now

$$\int f_0^{(k)}(x)\,\{f_\nu^{(k)}(x)-f_0^{(k)}(x)\}dx = (-1)^k \int f_0^{(2k)}(x)f_\nu(x)dx - \theta_k(s_0)$$
$$= \int\{(-1)^k f_0^{(2k)}(x) - \theta_k(s_0)\}f_\nu(x)dx, \qquad \ldots \quad (3.17)$$

and

$$\int\{f_\nu^{(k)}(x)-f_0^{(k)}(x)\}^2dx$$
$$= (-1)^k \int\{f_\nu(x)-f_0(x)\}\,\{f_\nu^{(2k)}(x)-f_0^{(2k)}(x)\}dx$$
$$= (-1)^k \int\{s_\nu(x)-s_0(x)\}^2\,\{f_\nu^{(2k)}(x)-f_0^{(2k)}(x)\}dx$$
$$+2(-1)^k \int s_0(x)\{s_\nu(x)-s_0(x)\}\,\{f_\nu^{(2k)}(x)-f_0^{(2k)}(x)\}dx$$
$$\leqslant \|f_0^{(2k)}+f_\nu^{(2k)}\|_\infty \|s_\nu-s_0\|_2^2 + 2\|s_\nu-s_0\|_2\,[\int\{f_\nu^{(2k)}(x)-f_0^{(2k)}(x)\}^2 f_0(x)dx]^{1/2}$$
$$= o(\|s_\nu-s_0\|_2). \qquad \ldots \quad (3.18)$$

(3.16), (3.17) and (3.18) imply that

$$\theta_k(s_\nu) = \theta_k(s_0) + 2\int\{(-1)^k f_0^{(2k)}(x) - \theta_k(F_0)\}\,f_\nu(x)dx + O(\|s_\nu - s_0\|_2).$$

This means that $\theta_k(s)$ is Fréchet differentiable along such paths with derivative $4\{(-1)^k f_0^{(2k)}-\theta_k(F_0)\}s_0$ and the result follows by standard theory.

(ii) Here, as in Ritov and Bickel (1987) we prove the assertion by presenting a sequence of Bayes problems. In the nth problem we observe $X_1,\ldots,X_n$ iid, $X_1 \sim F \epsilon \boldsymbol{F}_{m,\alpha,g}$. The loss function is $\boldsymbol{L}_n(\theta, d) = 1_{\{|\theta-d|> c_n^{-1}n^{-\gamma}\}}$. F is picked according to a measure Π_ν to be described next. Note that the sequence $\Pi_1, \Pi_2,\ldots$ is constructed such that the union of their supports $\boldsymbol{F}^*$ is compact with F_0 its only accumulation point. Let $F_0 \epsilon \boldsymbol{F}_{m,\alpha,g}$ be arbitrary. Clearly, f_0 is bounded away from zero on some interval. For simplicity we take this interval to be [0, 1]. To simplify the notation we assume also that $\sup\limits_{x\epsilon[0,1]} g(x) \leqslant 1$.

We now describe Π_ν. Let h_i, $i = 0, 1,\ldots, \nu-1$ be a sequence of functions such that $\int\limits_0^1 h_i(x)dx = 0$, $h_i^{(j)}(0) = h_i^{(j)}(1) = 0$, $j = 0,\ldots, m+1$, $\int\{h_i^{(k)}(x)\}^2\, dx = 1$ and $\int\limits_{i/\nu}^{(i+1)/\nu} h_i^{(k)}(\nu x-i)f_0^{(k)}(x)dx = 0$. Let β equal $0, 1, \ldots, r-1$ with probability $1/r$ and let $\Delta_0,\ldots, \Delta_{\nu-1}$ be iid, independent of β and each equal to ± 1 with probability 1/2. Let F be the random measure with density

$$f(x) = f_0(x)+\beta\nu^{-(m+\alpha)}\, \Delta_i h_i(\nu x-i) \text{ on } [i/\nu, (i+1)/\nu).$$

The measure that governs the selection of F is Π_ν. Clearly, for any F in the support of Π_ν by our assumptions of h_i,

$$\theta_k(F) = \theta_k(F_0)+\beta^2\nu^{-2(m+\alpha)+2k}.$$

That is $\theta_k(F)$ equals $\theta_k(F_0)+j\nu^{-2(m+\alpha-k)}$ if $\beta = j$.

We show that if

$$n^2\nu^{-(4m+4\alpha+1)} \to 0 \qquad \ldots \quad (3.19)$$

then the variational distance between the probability measures of $X_1, \ldots, X_n$ under $\beta = i$ and $\beta = j$ tends to 0. Assume that this is the case and F is distributed according to Π_ν, Π_ν satisfies (3.19) and

$$\nu^{-2(m+\alpha-k)}c_n n^\gamma \to \infty \qquad \ldots \quad (3.20)$$

where

$$\gamma = 4(m+\alpha-k)/(1+4m+4\alpha).$$

This is possible if $k < m+\alpha$. If

$$A_{nj} = \{|T_n-\theta_k(F_j)| < [c_n n^\gamma]^{-1}\}$$

then by construction for n sufficiently large the A_{nj} are disjoint. The Bayes risk for estimating $\theta_k\ (F)$ using our loss function is

$$R_n = \frac{1}{r} \sum_{j=1}^{r} P_j^{(n)}\,(A_{nj}^C)$$

$$= 1 - \frac{1}{r} \sum_{j=1}^{r} P_j^{(n)}\,(A_{nj}).$$

But, by the equivalence of $P\,[\,.\,|\beta = i]$ and $P\,[\,.\,|\beta = j]$ we have observed

$$P_j^{(n)}\,(A_{nj}) - P_0^{(n)}\,(A_{nj}) \to 0 \text{ for each } j.$$

So,

$$\underline{\lim}_n R_n \geqslant 1 - \frac{1}{r}\ \overline{\lim} \sum_{j=1}^{r} P_0^{(n)}\,(A_{nj})$$

$$= 1 - \frac{1}{r}\ \overline{\lim}\ P_0^{(n)} \left(\bigcup_{j=1}^{r} A_{nj} \right) \geqslant 1 - \frac{1}{r}\,.$$

Finally

$$\inf_{T_n} \sup_{F \epsilon F^*} P_F\,[c_n\, n^\gamma\, |T_n - \theta_k(F)| \geqslant 1] \geqslant R_n.$$

Hence, since r is arbitrary,

$$\underline{\lim}\ \inf_{T_n} \sup_{F \epsilon F^*} P_F[c_n\, n^\gamma\, |T_n - \theta_k(F)| \geqslant 1] = 1$$

as advertised. This combines ideas of Hasminskii (1979) and Stone (1983).

We turn to the proof that (3.22) implies convergence of the variational distance. Let N_i, $i = 0, \ldots, \nu - 1$ be the number of X's in $[i/\nu, (i+1)/\nu)$ and let $X_{i1}, \ldots, X_{iN_i}$ be the set of observations in that interval. Note that the random vector $(N_0, \ldots, N_{\nu-1})$ is independent of β and $(\Delta_0, \ldots, \Delta_{\nu-1})$, and that the blocks $(X_{i1}, \ldots, X_{iN_i})$ and $(X_{j1}, \ldots, X_{jN_j})$, $i \neq j$ are independent given N_i and N_j. Without loss of generality consider $\beta = 0$ and $\beta = 1$.

The likelihood ratio of $\beta = 1$ to $\beta = 0$ is $L = \prod_{i=0}^{\nu-1} L_i$ where

$$L_i = 1/2 \prod_{j=1}^{N_i} \{1 + \nu^{-(m+\alpha)} h_i(U_{ij})/f_0(U_{ij})\} + 1/2 \prod_{j=1}^{N_i} \{1 - \nu^{-(m+\alpha)} h_i(U_{ij})/f_0(U_{ij})\}$$

$$= 1 + \sum_{l=1}^{[N_i/2]} \nu^{-2(m+\alpha)l} \sum_{\substack{j_1, \ldots, j_{2l} \\ \text{all different}}} \!\!\cdots \sum \frac{h_i(U_{ij_1})}{f_0(U_{ij_1})} \cdots \frac{h_i(U_{ij_{2l}})}{f_0(U_{ij_{2l}})}$$

where $U_{ij} = \nu X_{ij} - i$ and $[x]$ is the greatest integer not larger than x.

Note that, $f_i(x) := \left[\nu\left\{F_0\left(\frac{i+1}{\nu}\right)-F_0\left(\frac{i}{\nu}\right)\right\}\right]^{-1} f_0\left(\frac{i+x}{\nu}\right)$ is the density of U_{ij} under f_0. We show that $L \xrightarrow{P} 1$ under F_0, which implies that the variational distance between the two conditional distribution tends to 0.

Since $\int_0^1 h_i(x)dx = 0$,

$$E(L_i-1\,|\,N_i) = 0. \qquad \ldots \quad (3.21)$$

Since $\|f_0\| < \infty$ by the lemma and the infimum of f_j on [0, 1] is > 0 by construction we obtain

$$\int_0^1 \frac{h_i^2(u)}{f_0^2(u)} f_i(u)du = \int_0^1 \frac{h_i^2(u).}{f_0(u)} \left[\nu\left\{F_0\left(\frac{i+1}{\nu}\right)-F_0\left(\frac{i}{\nu}\right)\right\}\right]^{-1} \leqslant \frac{1}{[\inf\limits_{x\in[0,1]} f_0(x)]^2} < \infty.$$

Let $A = \sup\limits_i \int_0^1 f_0^{-2}(u)\, f_i(u)\, h_i^2(u)\, du$. Then

$$\text{var}\,(L_i-1\,|\,N_i) \leqslant \sum_{l=1}^{[N_i/2]} \nu^{-4(m+\alpha)l} \binom{N_i}{2l} A^{2l},$$

and

$$\text{var}\left\{\sum_{i=0}^{\nu-1}(L_i-1)\right\} = E\left\{\sum_{i=0}^{\nu-1}(L_i-1)^2\right\} \leqslant E\sum_{i=0}^{\nu-1}\sum_{l=1}^{[N_i/2]} \nu^{-4(m+\alpha)l}\binom{N_i}{2l}A^{2l}. \quad \ldots \quad (3.22)$$

Let $p_i = F_0((i+1)/\nu)-F_0(i/\nu)$. Straightforward calculations give

$$E\sum_{l=1}^{[N_i/2]} \nu^{-4(m+\alpha)l}\binom{N_i}{2l}A^{2l}$$

$$= \sum_{j=2}^{n}\binom{n}{j} p_i^j(1-p_i)^{n-j}\sum_{l=1}^{[j/2]}(A\nu^{-2(m+\alpha)})^{2l}\binom{j}{2l}$$

$$= \sum_{l=1}^{[n/2]}(A\nu^{-2(m+\alpha)})^{2l}\frac{n\,!}{(2l)!}\sum_{j=2l}^{n}\frac{1}{(j-2l)\,!\,(n-j)\,!}\,p_i^j(1-p_i)^{n-j}$$

$$= \sum_{l=1}^{[n/2]}(A\nu^{-2(m+\alpha)})^{2l}\binom{n}{2l}\sum_{j=0}^{n-2l}\frac{(n-2l)\,!}{j\,!\,(n-2l-j)\,!}\,p_i^{j+2l}(1-p_i)^{n-2l-j}$$

$$= \sum_{l=1}^{[n/2]}(A\,p_i\,\nu^{-2(m+\alpha)})^{2l}\binom{n}{2l}$$

$$\leqslant \sum_{l=1}^{[n/2]}\frac{1}{(2l)\,!}(nA\,p_i\,\nu^{-2(m+\alpha)})^{2l} \leqslant \exp\{nAp_i\,\nu^{-2(m+\alpha)}\}^2-1 \qquad \ldots \quad (3.23)$$

$$= (1+o(1))A^2n^2p_i^2\,\nu^{-4(m+\alpha)} = O(n\,\nu^{-(2m+2\alpha+1)})^2$$

since $\nu\,p_i < \|f_0\|_\infty$.

We obtain from (3.22) and (3.23) that

$$\operatorname{var}\left\{\sum_{i=0}^{\nu-1}(L_i-1)\right\} = O(n^{-2}\nu^{-(4m+4\alpha+1)})$$

Therefore, from (3.19) and (3.21) we obtain :

$$\sum_{i=0}^{\nu-1}(L_i-1) = o_P(1) \text{ and } \sum_{i=0}^{\nu-1}(L_i-1)^2 = o_P(1)$$

both under F_0. Hence

$$\log L = \sum_{i=0}^{\nu-1}(L_i-1)+O\left(\sum_{i=0}^{\nu-1}(L_i-1)^2\right) \xrightarrow{P} 0$$

under F_0 proving the assertion. □

Proof of Lemma 1: It is enough to prove that for any $\alpha_i > 0$ and $d_i < \infty$,

$$\sup_{\substack{x,y \\ 0<|x-y|\leqslant 1}} \{|f^{(i)}(x)-f^{(i)}(y)|/|x-y|^{\alpha_i}\} \leqslant d_i \qquad \dots \quad (3.24)$$

implies that

$$\|f^{(i)}\|_\infty \leqslant c_i \qquad \dots \quad (3.25)$$

where $c_i < \infty$ is a function of α_i and d_i only. Suppose (3.24) implies (3.25) then

$$|f^{(i-1)}(x)-f^{(i-1)}(y)| = |f^{(i)}(x^*)|\,|x-y| \leqslant c_i|x-y| \text{ for } 0<|x-y|\leqslant 1$$

and the lemma follows by backward induction from m.

Suppose (3.1) holds. Let b_i be an arbitrary number lying in (0, 1] and assume that $f^{(i)}(x) \geqslant d_i(b_i/2)^{\alpha i}$ for a point $x \in R$. Then

$$f^{(i)}(y) \geqslant a_i = f^{(i)}(x)-d_i(b_i/2)^{\alpha i} \geqslant 0 \qquad \dots \quad (3.26)$$

for all $y \in [x-b_i/2, x+b_i/2] \equiv J_i$.

Then $f^{(i-1)}(u)$ is monotone on J_i and $|f^{(i-1)}(y)|$, $y \in J_i$, can be smaller than $a_{i-1} \equiv 1/4a_ib_i$ only on an interval of length smaller than $1/2b_i$. This leaves an interval J_{i-1} of length $b_{i-1} \geqslant 1/4b_i$ on which either $\inf_{y\in J_{i-1}}\{f^{(i-1)}(y)\} \geqslant a_{i-1}$ or $\sup_{y\in J_{i-1}}\{f^{(i-1)}(y)\} \leqslant -a_{i-1}$. Continue this line of argument inductively and obtain that (3.26) entails that $f(y) \geqslant a_0 \geqslant a_ib_i^i/2^{i(i+1)}$ on the interval J_0 whose length is $b_0 \geqslant 4^{-i}\, b_i$. But $f(\cdot)$ is a probability density function and hence

$$1 \geqslant a_0b_0 \geqslant 2^{-i(i+3)}\, a_i\, b_i^{i+1}.$$

Therefore,

$$\begin{aligned} f^{(i)}(x) &= a_i + d_i(b_i/2)^{\alpha_i} \\ &\leqslant 2^{i(i+3)} b_i^{-(i+1)} + d_i(b_i/2)^{\alpha_i}. \end{aligned}$$

Hence $f^{(i)}$ is bounded and the lemma follows. □

Acknowledgment. P. Hall and S. Marron pointed out a gap in our original proof of Theorem 2 which we have corrected.

References

Bickel, P. J. (1982): On adaptive estimation. *Ann. Statist.*, **10**, 647-671.

Donoho, D. L. and Liu. R. C. (1987): Geometrizing rates of convergence I–III (manuscript).

Farrel, R. H. (1972): On the best obtainable asymptotic rates of convergence of a density function at a point. *Ann. Math. Stat.*, **43**, 170-180.

Hall, P. and Marron, J. S. (1987): Estimation of integrated squared density derivatives. To appear in *Statistical and Probability Letters.*

Hasminskii, R. Z. (1979): Lower bound for the risks of nonparametric estimates of the mode. *Contributions to Statistics*, (J. Hájek Memorial Volume). *Academia, Prague*, 91-97.

Hasminskii, R. Z. and Ibragimov, I. A. (1978): On the non parametric estimation of functionals. *Symposium in Asymptotic Statistics, Prague,* 41-52.

Ibragimov, I. A. and Hasminskii, R. Z. (1981): *Statistical Estimation*, Springer Verlag, New York 237-240.

Khoshevnik, Yu, A. and Levit, B.Ya. (1976): On a non-parametric analogue of the information matrix. *Theor. Probab. Appl.*, **21**, 738-753.

Pfanzagl, J. (1982): Contributions to a general asymptotic statistical theory. *Lecture Notes in Statistics*, **13**, Springer-Verlag, New York.

Prakasa Rao, B. L. S. (1983): *Nonparametric Functional Estimation*, Academic Press, New York.

Ritov, Y. and Bickel, P. J. (1987): Achieving information bounds in non and semi-parametric models. To appear in *Ann. Statist.*

Schick, A. S. (1986): On asymptotically efficient estimators in semiparametric models. *Ann. Statist.*, **14**, 1139-1151.

Schweder, T. (1975): Window estimation of the asymptotic variance of rank estimators of location. *Scandinavian J. of Statist.*, **2**, 113-126.

Stone, C. J. (1983): Optimal uniform rate of convergence for nonparametric estimators of a density function and its derivatives. *Recent Advances in Statistics*, Paper in Honor of H. Chernoff. Academic Press, New York.

Paper received : September, 1988.

IMS Lecture Notes–Monograph Series
Complex Datasets and Inverse Problems: Tomography, Networks and Beyond
Vol. 54 (2007) 177–186

DOI: 10.1214/074921707000000148

Local polynomial regression on unknown manifolds

Peter J. Bickel[1] **and Bo Li**[2]

University of California, Berkeley and Tsinghua University

Abstract: We reveal the phenomenon that "naive" multivariate local polynomial regression can adapt to local smooth lower dimensional structure in the sense that it achieves the optimal convergence rate for nonparametric estimation of regression functions belonging to a Sobolev space when the predictor variables live on or close to a lower dimensional manifold.

1. Introduction

It is well known that worst case analysis of multivariate nonparametric regression procedures shows that performance deteriorates sharply as dimension increases. This is sometimes refered to as the curse of dimensionality. In particular, as initially demonstrated by [19, 20], if the regression function, $m(x)$, belongs to a Sobolev space with smoothness p, there is no nonparametric estimator that can achieve a faster convergence rate than $n^{-\frac{p}{2p+D}}$, where D is the dimensionality of the predictor vector X.

On the other hand, there has recently been a surge in research on identifying intrinsic low dimensional structure from a seemingly high dimensional source, see [1, 5, 15, 21] for instance. In these settings, it is assumed that the observed high-dimensional data are lying on a low dimensional smooth manifold. Examples of this situation are given in all of these papers — see also [14]. If we can estimate the manifold, we can expect that we should be able to construct procedures which perform as well as if we know the structure. Even if the low dimensional structure obtains only in a neighborhood of a point, estimation at that point should be governed by actual rather than ostensible dimension. In this paper, we shall study this situation in the context of nonparametric regression, assuming the predictor vector has a lower dimensional smooth structure. We shall demonstrate the somewhat surprising phenomenon, suggested by Bickel in his 2004 Rietz lecture, that the procedures used with the expectation that the ostensible dimension D is correct will, with appropriate adaptation not involving manifold estimation, achieve the optimal rate for manifold dimension d.

Bickel conjectured in his 2004 Rietz lecture that, in predicting Y from X on the basis of a training sample, one could automatically adapt to the possibility that the apparently high dimensional X that one observed, in fact, lived on a much smaller dimensional manifold and that the regression function was smooth on that manifold. The degree of adaptation here means that the worst case analyses for prediction are governed by smoothness of the function on the manifold and not on

[1] 367 Evans Hall, Department of Statistics, University of California, Berkeley, CA, 94720-3860, USA, e-mail: bickel@stat.berkeley.edu

[2] S414 Weilun Hall, School of Economics and Management, Tsinghua University, Beijing, 100084, China, e-mail: libo@em.tsinghua.edu.cn

AMS 2000 subject classifications: primary 62G08, 62H12; secondary 62G20.

Keywords and phrases: local polynomial regression, manifolds.

the space in which X ostensibly dwells, and that purely data dependent procedures can be constructed which achieve the lower bounds in all cases.

In this paper, we make this statement precise with local polynomial regression. Local polynomial regression has been shown to be a useful nonparametric technique in various local modelling, see [8, 9]. We shall sketch in Section 2 that local linear regression achieves this phenomenon for local smoothness $p = 2$, and will also argue that our procedure attains the global IMSE if global smoothness is assumed. We shall also sketch how polynomial regression can achieve the appropriate higher rate if more smoothness is assumed.

A critical issue that needs to be faced is regularization since the correct choice of bandwidth will depend on the unknown local dimension $d(x)$. Equivalently, we need to adapt to $d(x)$. We apply local generalized cross validation, with the help of an estimate of $d(x)$ due to [14]. We discuss this issue in Section 3. Finally we give some simulations in Section 4.

A closely related technical report, [2] came to our attention while this paper was in preparation. Binev et al consider in a very general way, the construction of nonparametric estimation of regression where the predictor variables are distributed according to a fixed completely unknown distribution. In particular, although they did not consider this possibility, their method covers the case where the distribution of the predictor variables is concentrated on a manifold. However, their method is, for the moment, restricted to smoothness $p \leq 1$ and their criterion of performance is the integral of pointwise mean square error with respect to the underlying distribution of the variables. Their approach is based on a tree construction which implicitly estimates the underlying measure as well as the regression. Our discussion is considerably more restrictive by applying only to predictors taking values in a low dimensional manifold but more general in discussing estimation of the regression function at a point. Binev et al promise a further paper where functions of general Lipschitz order are considered.

Our point in this paper is mainly a philosophical one. We can unwittingly take advantage of low dimensional structure without knowing it. We do not give careful minimax arguments, but rather, partly out of laziness, employ the semi heuristic calculations present in much of the smoothing literature.

Here is our setup. Let $(X_i, Y_i), (i = 1, 2, \ldots, n)$ be i.i.d $\Re^{D+1}$ valued random vectors, where X is a D-dimensional predictor vector, Y is the corresponding univariate response variable. We aim to estimate the conditional mean $m_0(x) = E(Y|X = x)$ nonparametrically. Our crucial assumption is the existence of a local *chart*, i.e., each small patch of $\mathcal{X}$ (a neighborhood around x) is isomorphic to a ball in a d-dimensional Euclidean space, where $d = d(x) \leq D$ may vary with x. Since we fix our working point x, we will use d for the sake of simplicity. The same rule applies to other notations which may also depend on x.) More precisely, let $\mathcal{B}^d_{z,r}$ denote the ball in $\Re^d$, centered at z with radius r. A similar definition applies to $\mathcal{B}^D_{x,R}$. For small $R > 0$, we consider the neighborhood of x, $\mathcal{X}_x := \mathcal{B}^D_{x,R} \cap \mathcal{X}$ within $\mathcal{X}$. We suppose there is a continuously differentiable bijective map $\phi : \mathcal{B}^d_{0,r} \mapsto \mathcal{X}_x$. Under this assumption with $d < D$, the distribution of X degenerates in the sense that it does not have positive density around x with respect to Lebesgue measure on $\Re^D$. However, the induced measure $\mathbb{Q}$ on $\mathcal{B}^d_{0,r}$ defined below, can have a non-degenerate density with respect to Lebesgue measure on $\Re^d$. Let $\mathcal{S}$ be an open subset of $\mathcal{X}_x$, and $\phi^{-1}(\mathcal{S})$ be its preimage in $\mathcal{B}^{d(x)}_{0,r}$. Then $\mathbb{Q}(Z \in \phi^{-1}(\mathcal{S})) = \mathbb{P}(X \in \mathcal{S})$. We assume throughout that $\mathbb{Q}$ admits a continuous positive density function $f(\cdot)$. We proceed to our main result whose proof is given in the Appendix.

2. Local linear regression

[17] develop the general theory for multivariate local polynomial regression in the usual context, i.e., the predictor vector has a D dimensional compact support in $\Re^D$. We shall modify their proof to show the "naive" (brute-force) multivariate local linear regression achieves the "oracle" convergence rate for the function $m(\phi(z))$ on $\mathcal{B}^d_{0,r}$.

Local linear regression estimates the population regression function by $\hat{\alpha}$, where $(\hat{\alpha}, \hat{\beta})$ minimize

$$\sum_{i=1}^{n} \left(Y_i - \alpha - \beta^T(X_i - x)\right)^2 K_h(X_i - x).$$

Here $K_h(\cdot)$ is a $D-$variate kernel function. For the sake of simplicity, we choose the same bandwidth h for each coordinate. Let

$$X_x = \begin{bmatrix} 1 & (X_1 - x)^T \\ \vdots & \vdots \\ 1 & (X_n - x)^T \end{bmatrix}$$

and $W_x = diag\{K_h(X_1 - x), \ldots, K_h(X_n - x)\}$. Then the estimator of the regression function can be written as

$$\hat{m}(x, h) = e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x Y$$

where e_1 is the $(D+1) \times 1$ vector having 1 in the first entry and 0 elsewhere.

2.1. Decomposition of the conditional MSE

We enumerate the assumptions we need for establishing the main result. Let M be a canonical finite positive constant,

(i) The kernel function $K(\cdot)$ is continuous and radially symmetric, hence bounded.

(ii) There exists an $\epsilon (0 < \epsilon < 1)$ such that the following asymptotic irrelevance conditions hold.

$$E\Big[K^\gamma(\frac{X - x}{h}) w(X) 1\big(X \in (\mathcal{B}^D_{x,h^{1-\epsilon}} \cap \mathcal{X})^c\big)\Big] = o(h^{d+2})$$

for $\gamma = 1, 2$ and $|w(x)| \leq M(1 + |x|^2)$.

(iii) $v(x) = Var(Y|X = x) \leq M$.

(iv) The regression function $m(x)$ is twice differentiable, and $\|\frac{\partial^2 m}{\partial x_a x_b}\|_\infty \leq M$ for all $1 \leq a \leq b \leq D$ if $x = (x_1, \ldots, x_D)$.

(v) The density $f(\cdot)$ is continuously differentiable and strictly positive at 0 in $\mathcal{B}^d_{0,r}$.

Condition (ii) is satisfied if K has exponential tails since if $V = \frac{X-x}{h}$, the conditions can be written as

$$E\big[K^\gamma(V) w(x + hV) 1(V \in (\mathcal{B}^D_{0,h^{1-\epsilon}})^c\big] = o(h^{d+2}).$$

Theorem 2.1. *Let x be an interior point in $\mathcal{X}$. Then under assumptions (i)-(v), there exist some $J_1(x)$ and $J_2(x)$ such that*

$$E\{\hat{m}(x,h) - m(x)|X_1,\dots,X_n\} = h^2 J_1(x)(1+o_P(1)),$$

$$Var\{\hat{m}(x,h) - m(x)|X_1,\dots,X_n\} = n^{-1}h^{-d}J_2(x)(1+o_P(1)).$$

Remark 1. The predictor vector doesn't need to lie on a perfect smooth manifold. The same conclusion still holds as long as the predictor vector is "close" to a smooth manifold. Here "close" means the noise will not affect the first order of our asymptotics. That is, we think of $X_1,\dots,X_n$ as being drawn from a probability distribution P on $\Re^D$ concentrated on the set

$$\mathcal{X} = \{y : |\phi(u) - y| \le \epsilon_n \text{ for some } u \in \mathcal{B}^d_{0,r}\}$$

and $\epsilon_n \to 0$ with n. It is easy to see from our arguments below that if $\epsilon_n = o(h)$, then our results still hold.

Remark 2. When the point of interest x is on the boundary of the support $\mathcal{X}$, we can show that the bias and variance have similar asymptotic expansions, following the Theorem 2.2 in [17]. But, given the extra complication of the embedding, the proof would be messier, and would not, we believe, add any insight. So we omit it.

2.2. Extensions

It's somewhat surprising but not hard to show that if we assume the regression function m to be p times differentiable with all partial derivatives of order p bounded ($p \ge 2$, an integer), we can construct estimates $\hat{m}$ such that,

$$E\{\hat{m}(x,h) - m(x)|X_1,\dots,X_n\} = h^p J_1(x)(1+o_P(1)),$$

$$Var\{\hat{m}(x,h) - m(x)|X_1,\dots,X_n\} = n^{-1}h^{-d}J_2(x)(1+o_P(1))$$

yielding the usual rate of $n^{-\frac{2p}{2p+d}}$ for the conditional MSE of $\hat{m}(x,h)$ if h is chosen optimal, $h = \lambda n^{-\frac{1}{2p+d}}$. This requires replacing local linear regression with local polynomial regression with a polynomial of order $p-1$. We do not need to estimate the manifold as we might expect since the rate at which the bias term goes to 0 is derived by first applying Taylor expansion with respect to the original predictor components, then obtaining the same rate in the lower dimensional space by a first order approximation of the manifold map. Essentially all we need is that, locally, the geodesic distance is roughly proportionate to the Euclidean distance.

3. Bandwidth selection

As usual this tells us, for $p = 2$, that we should use bandwidth $\lambda n^{-\frac{1}{4+d}}$ to achieve the best rate of $n^{-\frac{2}{4+d}}$. This requires knowledge of the local dimension as well as the usual difficult choice of λ. More generally, dropping the requirement that the bandwidth for all components be the same we need to estimate d and choose the constants corresponding to each component in a simple data determined way.

There is an enormous literature on bandwidth selection. There are three main approaches: plug-in ([7, 16, 18], etc); the bootstrap ([3, 11, 12], etc) and cross validation ([6, 10, 22], etc). The first has always seemed logically inconsistent to us since it requires higher order smoothness of m than is assumed and if this higher order smoothness holds we would not use linear regression but a higher order polynomial. See also the discussion of [23].

We propose to use a blockwise cross-validation procedure defined as follows. Let the data be $(X_i, Y_i), 1 \leq i \leq n$. We consider a block of data points $\{(X_j, Y_j) : j \in \mathcal{J}\}$, with $|\mathcal{J}| = n_1$. Assuming the covariates have been standardized, we choose the same bandwidth h for all the points and all coordinates within the block. A leave-one-out cross validation with respect to the block while using the whole data set is defined as following. For each $j \in \mathcal{J}$, let $\hat{m}_{-j,h}(X_j)$ be the estimated regression function (evaluated at X_j) via local linear regression with the whole data set except X_j. In contrast to the usual leave-one-out cross-validation procedure, our modified leave-one-out cross-validation criterion is defined as $mCV(h) = \frac{1}{n_1}\sum_{j\in\mathcal{J}}(Y_j - \hat{m}_{-j,h}(X_j))^2$. Using a result from [23], it can be shown that

$$mCV(h) = \frac{1}{n_1}\sum_{j\in\mathcal{J}} \frac{(Y_j - \hat{m}_h(X_j))^2}{(1 - S_h(j,j))^2}$$

where $S_h(j,j)$ is the diagonal element of the smoothing matrix S_h. We adopt the GCV idea proposed by [4] and replace the $S_h(j,j)$ by their average $atr_{\mathcal{J}}(S_h) = \frac{1}{n_1}\sum_{j\in\mathcal{J}} S_h(j,j)$. Thereby our modified generalized cross-validation criterion is,

$$mGCV(h) = \frac{1}{n_1}\sum_{j\in\mathcal{J}} \frac{(Y_j - \hat{m}_h(X_j))^2}{(1 - atr_{\mathcal{J}}(S_h))^2}.$$

The bandwidth h is chosen to minimize this criterion function.

We give some heuristics for the justifying the (blockwise homoscedastic) mGCV. In a manner analogous to [23], we can show

$$S_h(j,j) = e_1^T(X_x^T W_x X_x)^{-1} e_1 K_h(0)|_{x=X_j}.$$

In view of (A.2) in the Appendix, we see $S_h(j,j) = n^{-1}h^{-d}K(0)(A_1(X_j) + o_p(1))$. Thus as $n^{-1}h^{-d} \to 0$,

$$\begin{aligned} atr_{\mathcal{J}}(S_h) &= n^{-1}h^{-d}K(0)(n_1^{-1}\sum_{j\in\mathcal{J}} A_1(X_j) + o_p(1)) \\ &= O_p(n^{-1}h^{-d}) = o_p(1). \end{aligned}$$

Then, as is discussed in [22], using the approximation $(1-x)^{-2} \approx 1 + 2x$ for small x, we can rewrite $mGCV(h)$ as

$$mGCV(h) = \frac{1}{n_1}\sum_{j\in\mathcal{J}}(Y_j - \hat{m}_h(X_j))^2 + \frac{2}{n_1} tr_{\mathcal{J}}(S_h)\frac{1}{n_1}\sum_{j\in\mathcal{J}}(Y_j - \hat{m}_h(X_j))^2.$$

Now regarding $\frac{1}{n_1}\sum_{j\in\mathcal{J}}(Y_j - \hat{m}_h(X_j))^2$ in the second term as an estimator of the constant variance for the focused block, the mGCV is approximately the same as the C_p criterion, which is an estimator of the prediction error up to a constant.

In practice, we first use [14]'s approach to estimate the local dimension d, which yields a consistent estimate $\hat{d}$ of d. Based on the estimated intrinsic dimensionality $\hat{d}$, a set of candidate bandwidths $\mathcal{CB} = \{\lambda_1 n^{-\frac{1}{\hat{d}+4}}, \ldots, \lambda_B n^{-\frac{1}{\hat{d}+4}}\}$ $(\lambda_1 < \cdots < \lambda_B)$ are chosen . We pick the one minimizing the $mGCV(h)$ function.

4. Numerical experiments

The data generating process is as following. The predictor vector $X = (X_{(1)}, X_{(2)}, X_{(3)})$, where $X_{(1)}$ will be sampled from a standard normal distribution, $X_{(2)} = X^3_{(1)} + sin(X_{(1)}) - 1$, and $X_{(3)} = \log(X^2_{(1)} + 1) - X_{(1)}$. The regression function $m(x) = m(x_{(1)}, x_{(2)}, x_{(3)}) = cos(x^{(1)}) + x_{(2)} - x^2_{(3)}$. The response variable Y is generated via the mechanism $Y = m(X) + \varepsilon$, where ε has a standard normal distribution. By definition, the 3-dimensional regression function $m(x)$ is essentially a 1-dimensional function of $x_{(1)}$. $n = 200$ samples are drawn. The predictors are standardized before estimation. We estimate the regression function $m(x)$ by both the "oracle" univariate local linear (ull) regression with a single predictor $X_{(1)}$ and our blind 3-variate local linear regression with all predictors $X_{(1)}, X_{(2)}, X_{(3)}$.

We focus on the middle block with 100 data points, with the number of neighbor parameter k, needed for Levina and Bickel's estimate, set to be 15. The intrinsic dimension estimator is $\hat{d} = 1.023$, which is close to the true dimension, $d = 1$. We use the Epanechnikov kernel in our simulation. Our proposed modified GCV procedure is applied to both the ull and mll procedures. The estimation results are displayed in Figure 1. The $x - axis$ is the standardized $X_{(1)}$. From the right panel, we see the blind mll indeed performs almost as well as the "oracle" ull.

Next, we allow the predictor vector to only lie close to a manifold. Specifically, we sample $X_{(1)} = X'_{(1)} + \epsilon'_1, X_{(2)} = X'^3_{(1)} + sin(X'_{(1)}) - 1 + \epsilon'_2, X_{(3)} = \log(X'^2_{(1)} + 1) - X'_{(1)} + \epsilon'_3$, where $X'_{(1)}$ is sampled from a standard normal distribution, and ϵ'_1, ϵ'_2 and ϵ'_3 are sampled from $\mathcal{N}(0, \sigma'^2)$. The noise scale is hence governed by σ'. In our experiment, σ' is set to be $0.02, 0.04, \ldots, 0.18, 0.20$ respectively. The predictor vector samples are visualized in the left panel of Figure 2 with $\sigma' = 0.20$. In the maximum noise scale case, the pattern of the predictor vector is somewhat vague. Again, a blind "mll" estimation is done with respect to new data generated in the aforementioned way. We plot the MSEs associated with different noise scales in the right panel of Figure 2. The moderate noise scales we've considered indeed don't have a significant influence on the performance of the "mll" estimator in terms of MSE.

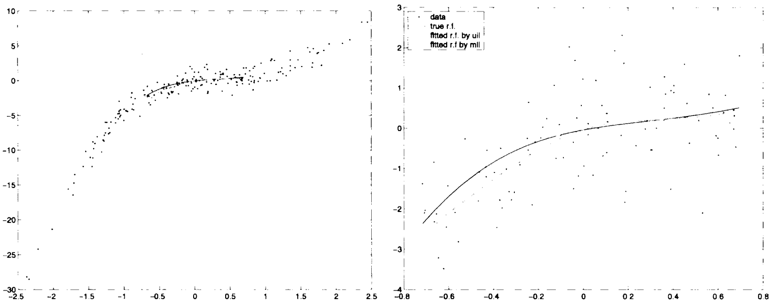

FIG 1. *The case with perfect embedding. The left panel shows the complete data and fitting of the middle block by both univariate local linear (ull) regression and multivariate local linear (mll) regression with bandwidths chosen via our modified GCV. The focused block is amplified in the right panel.*

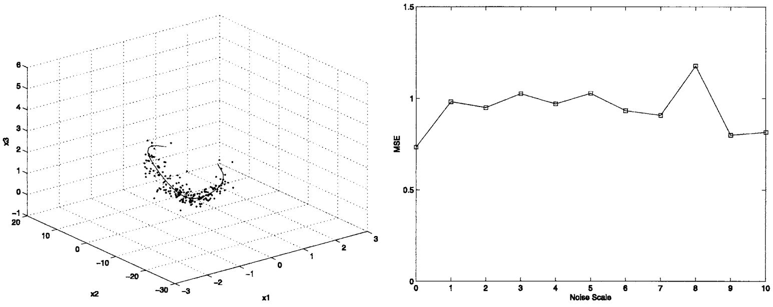

FIG 2. *The case with "imperfect" embedding. The left panel shows the predictor vector in a 3-D fashion with the noise scale $\sigma' = 0.2$. The right panel gives the MSEs with respect to increasing noise scales.*

Appendix

Proof of Theorem 2.1. Using the notation of [17], $\mathcal{H}_m(x)$ is the $D \times D$ Hessian matrix of $m(x)$ at x, and

$$Q_m(x) = [(X_1 - x)^T \mathcal{H}_m(x)(X_1 - x), \cdots, (X_n - x)^T \mathcal{H}_m(x)(X_n - x)]^T.$$

Ruppert and Wand have obtained the bias term.

$$\begin{aligned}(\text{A.1}) \qquad & E(\hat{m}(x,h) - m(x)|X_1, \cdots, X_n) \\ & = \frac{1}{2} e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x \{Q_m(x) + R_m(x)\}\end{aligned}$$

where if $|\cdot|$ denotes Euclidean norm, $|R_m(x)|$ is of lower order than $|Q_m(x)|$. Also we have

$$\begin{aligned} & n^{-1} X_x^T W_x X_x \\ & = \begin{bmatrix} n^{-1}\sum_{i=1}^n K_h(X_i - x) & n^{-1}\sum_{i=1}^n K_h(X_i - x)(X_i - x)^T \\ n^{-1}\sum_{i=1}^n K_h(X_i - x)(X_i - x) & n^{-1}\sum_{i=1}^n K_h(X_i - x)(X_i - x)(X_i - x)^T \end{bmatrix}. \end{aligned}$$

The difference in our context lies in the following asymptotics.

$$\begin{aligned} EK_h(X_i - x) &= E\big[K_h(X_i - x)1\big(X_i \in \mathcal{B}^D_{x,h^{1-\epsilon}} \cap \mathcal{X}\big)\big] \\ & \quad + E\big[K_h(X_i - x)1\big(X_i \in \big(\mathcal{B}^D_{x,h^{1-\epsilon}} \cap \mathcal{X}\big)^c\big)\big] \\ & \overset{(ii)}{=} h^{-D}\Big(\int_{N^d_{0,h^{1-\epsilon}}} K\big(\frac{\phi(z') - \phi(0)}{h}\big) f(z')\mathrm{d}z' + o_P(h^d)\Big) \\ &= h^{d-D}\Big(f(0)\int_{\Re^d} K(\nabla\phi(0)u)\mathrm{d}u + o_P(1)\Big) \\ &= h^{d-D}\big(A_1(x) + o_P(1)\big). \end{aligned}$$

Thus, by the LLN, we have

$$n^{-1}\sum_{i=1}^n K_h(X_i - x) = h^{d-D}\big(A_1(x) + o_P(1)\big).$$

Similarly, there exist some $A_2(x)$ and $A_3(x)$ such that

$$n^{-1}\sum_{i=1}^{n} K_h(X_i-x)(X_i-x) = h^{2+d-D}\big(A_2(x)+o_P(1)\big)$$

and

$$n^{-1}\sum_{i=1}^{n} K_h(X_i-x)(X_i-x)(X_i-x)^T = h^{2+d-D}\big(A_3(x)+o_P(1)\big)$$

where we used assumption (i) to remove the term of order h^{1+d-D} in deriving the asymptotic behavior of $n^{-1}\sum_{i=1}^{n} K_h(X_i-x)(X_i-x)$. Invoking Woodbury's formula, as in the proof of Lemma 5.1 in [13], leads us to

$$\text{(A.2)} \qquad \big(n^{-1}X_x^TW_xX_x\big)^{-1} = h^{D-d}\begin{bmatrix} A_1(x)^{-1}+o_P(1) & O_P(1) \\ O_P(1) & h^{-2}O_p(1)\end{bmatrix}$$

On the other hand,

$$\begin{aligned} &n^{-1}X_xW_xQ_m(x) \\ &= \begin{bmatrix} n^{-1}\sum_{i=1}^{n} K_h(X_i-x)(X_i-x)^T\mathcal{H}_m(x)(X_i-x) \\ n^{-1}\sum_{i=1}^{n}\{K_h(X_i-x)(X_i-x)^T\mathcal{H}_m(x)(X_i-x)\}(X_i-x)\end{bmatrix}. \end{aligned}$$

In a similar fashion, we can deduce that for some $B_1(x), B_2(x)$,

$$n^{-1}\sum_{i=1}^{n} K_h(X_i-x)(X_i-x)^T\mathcal{H}_m(x)(X_i-x) = h^{2+d-D}\big(B_1(x)+o_P(1)\big)$$

and

$$n^{-1}\sum_{i=1}^{n}\{K_h(X_i-x)(X_i-x)^T\mathcal{H}_m(x)(X_i-x)\}(X_i-x) = h^{3+d-D}\big(B_2(x)+o_P(1)\big).$$

We have

$$\text{(A.3)} \qquad n^{-1}X_xW_xQ_m(x) = h^{d-D}\begin{bmatrix} h^2\big(B_1(x)+o_P(1)\big) \\ h^3\big(B_2(x)+o_P(1)\big)\end{bmatrix}.$$

It follows from (A.1),(A.2) and (A.3) that the bias admits the following approximation.

$$\text{(A.4)} \qquad E(\hat{m}(x,h)-m(x)|X_1,\dots,X_n) = h^2A_1(x)^{-1}B_1(x)+o_P(h^2).$$

Next, we move to the variance term.

$$\text{(A.5)} \qquad \begin{aligned} &Var\{\hat{m}(x,h)|X_1,\dots,X_n\} \\ &= e_1^T(X_x^TW_xX_x)^{-1}X_x^TW_xVW_xX_x(X_x^TW_xX_x)^{-1}e_1. \end{aligned}$$

The upper-left entry of $n^{-1}X_x^TW_xVW_xX_x$ is

$$n^{-1}\sum_{i=1}^{n} K_h(X_i-x)^2v(X_i) = h^{d-2D}C_1(x)(1+o_P(1)).$$

The upper-right block is

$$n^{-1}\sum_{i=1}^{n} K_h(X_i-x)^2(X_i-x)^Tv(X_i) = h^{1+d-2D}C_2(x)(1+o_P(1))$$

and the lower-right block is

$$n^{-1}\sum_{i=1}^{n} K_h(X_i - x)^2 (X_i - x)(X_i - x)^T v(X_i) = h^{2+d-2D} C_3(x)(1 + o_P(1)).$$

In light of (A.2), we arrive at

$$Var\{\hat{m}(x,h)|X_1,\dots,X_n\} = n^{-1}h^{-d}A_1(x)^{-2}C_1(x)(1 + o_P(1)). \tag{A.6}$$

The proof is complete. □

Acknowledgment. We thank Ya'acov Ritov for insightful comments.

References

[1] Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* **15** 1373–1396.

[2] Binev, P., Cohen, A., Dahmen, W., DeVore, R. and Temlyakov, V. (2004). Universal algorithms for learning theory part i: piecewise constant functions. IMI technical reports, SCU.

[3] Cao-Abad, R. (1991). Rate of convergence for the wild bootstrap in nonparametric regression. *Ann. Statist.* **19** 2226–2231. MR1135172

[4] Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* **31** 377–403. MR516581

[5] Donoho, D. and Grimes, C. (2003). Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci. USA* **100** 5591–5596 (electronic).

[6] Dudoit, S. and van der Laan, M. (2005). Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Stat. Methodol.* **2** 131–154. MR2161394

[7] Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser. B* **57** 371–394. MR1323345

[8] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London. MR1383587

[9] Fan, J. and Gijbels, I. (2000). Local polynomial fitting. In *Smoothing and Regression. Approaches, Computation and Application* (M.G. Schimek, ed.) 228–275. Wiley, New York.

[10] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer, New York. MR1920390

[11] Hall, P., Lahiri, S. and Truong, Y. (1995). On bandwidth choice for density estimation with dependent data. *Ann. Statist.* **23** 2241–2263. MR1389873

[12] Härdle, W. and Mammen, E. (1991). Bootstrap methods in nonparametric regression. In *Nonparametric Functional Estmation and Related Topics (Spetses, 1990)* **335** 111–123. Kluwer Acad. Publ., Dordrecht. MR1154323

[13] Lafferty, J. and Wasserman, L. (2005). Rodeo: Sparse nonparametric regression in high dimensions. Technical report, CMU.

[14] Levina, E. and Bickel, P. J. (2005). Maximum likelihood estimation of intrinsic dimension. *Advances in NIPS* **17**. MIT Press.

[15] ROWEIS, S. AND SAUL, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290** 2323–2326.

[16] RUPPERT, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Amer. Statist. Assoc.* **22** 1049–1062. MR1482136

[17] RUPPERT, D. AND WAND, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22** 1346–1370. MR1311979

[18] RUPPERT, D., SHEATHER, S. J. AND WAND, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90** 1257–1270. MR1379468

[19] STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360. MR594650

[20] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR673642

[21] TENENBAUM, J. B., DE SILVA, V. AND LANGFORD, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319–2323.

[22] WANG, Y. (2004). Model selection. In *Handbook of Computational Statistics* 437–466. Springer, New York. MR2090150

[23] ZHANG, C. (2003). Calibrating the degrees of freedom for automatic data smoothing and effective curve checking. *J. Amer. Statist. Assoc.* **98** 609–628. MR2011675