EMPIRICAL LIKELIHOOD CONFIDENCE BANDS IN CURVE ESTIMATION

BY

PETER HALL AND ART B. OWEN

TECHNICAL REPORT NO. 324
JULY 1989

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT DMS86-00235

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
EMPIRICAL LIKELIHOOD CONFIDENCE BANDS IN CURVE ESTIMATION

BY

PETER HALL AND ART B. OWEN

TECHNICAL REPORT NO. 324
JULY 1989

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT DMS86-00235

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
EMPIRICAL LIKELIHOOD CONFIDENCE BANDS IN CURVE ESTIMATION

by

Peter Hall
Australian National University

Art B. Owen
Stanford University

SUMMARY. Empirical likelihood methods are developed for constructing confidence bands in problems of nonparametric density estimation and nonparametric regression. These techniques have an advantage over more conventional methods in that the shape of the bands is determined solely by the data. We show how to construct an empirical likelihood functional, rather than a function, and contour it to produce confidence bands for functions. Analogues of Wilks' theorem are established in this infinite parameter setting, and may be used to select the appropriate contour. An alternative contouring method, based on bootstrap calibration, is also suggested. Large sample theory is developed to show that the bands have asymptotically correct coverage, and numerical examples are presented to demonstrate that the techniques are practicable.

KEY WORDS. Bootstrap, confidence bands, curve estimation, empirical likelihood, hypothesis test, nonparametric density estimation, nonparametric regression, Wilks' theorem.

SHORT TITLE. Confidence bands.
1. INTRODUCTION

The method of empirical likelihood was proposed by Owen (1988, 1990) as a technique for constructing confidence regions or hypothesis tests in finite parameter problems. As Owen (1990) pointed out, a major advantage of empirical likelihood is that it involves no predetermined assumptions about the shape which the confidence region should assume. It produces regions which reflect emphasis in the observed data set (Hall 1989). This property has considerable potential in the problem of confidence bands for curve estimators. Like bivariate confidence regions, confidence bands are intrinsically two-dimensional, and it is necessary to develop a technique for determining their shape.

To better appreciate this problem, consider using the bootstrap to construct confidence bands from a large collection of simulated curve estimates (e.g. Efron and Tibshirani 1986). There are a great many "95% confidence bands" which contain, as envelopes, precisely 95% of the simulated estimates. A natural choice is to find the value $\alpha > .95$ for which bootstrap pointwise $\alpha$ confidence bands have simultaneous coverage $.95$. This is an infinite dimensional analogue of the projection method used in Owen (1990) to construct bootstrap confidence regions for a vector mean. It requires the resampled curves to do double duty: they must determine the shape of a family of confidence bands and then pick out the 95% band. Unless a large number of curves is resampled one can find that more than 5% of the resampled curves are, at some point in their domain, the largest or smallest of the resampled curves and that in general, many fewer confidence levels are available because of ties. Moreover, the resulting bands are not very smooth. Empirical likelihood, as described and illustrated below, avoids both of these problems. Our aim in this paper is to show that empirical likelihood is a practicable method for drawing confidence bands in infinite parameter problems, such as nonparametric density estimation and nonparametric regression.

The method of applying empirical likelihood is similar in the finite and infinite parameter cases. First construct the empirical loglikelihood functional—a function in the finite parameter case. Secondly, contour it to form a confidence set. Thirdly, select a contour to serve as a confidence set for the desired level of coverage. The level sets of the loglikelihood, as infinite dimensional objects, are not well suited to display. One therefore projects them onto confidence bands, that is, infinite dimensional rectangles of the form $\{f(x) \mid g(x) \leq f(x) \leq h(x)\}$. The shape of the confidence regions then reduces to the choice of curves $g(x)$ and $h(x)$. As in the bootstrap case we choose them with the goal of getting equal pointwise coverage levels. In the finite parameter case Owen (1988,1990) suggested selecting the contour by using an empirical
version of Wilks' theorem. An analogue of Wilks' theorem exists for infinite parameter problems, where it is basically a limit theorem for extremes of a Gaussian process. It is known that the convergence rate of Gaussian extremes can be particularly slow (Hall 1979), and so in an infinite parameter problem it might not be advisable to select the appropriate contour by appealing to Wilks' theorem. Therefore we develop an alternative method based on bootstrap calibration, which employs a bootstrap approximation to the distribution of empirical loglikelihood. That is, we use empirical likelihood to set the shape of the confidence bands, and use the bootstrap to set the level.

Our adaptation of empirical likelihood to the infinite parameter setting is based on kernel estimates. The kernel function does introduce a degree of bias into the problem, with the result that our confidence bands are really for the expected value of the function estimator rather than for the function itself. However this problem is readily overcome, either by making an explicit bias correction or by undersmoothing so that bias is negligible.

Section 2 will develop methodology and basic theory, including an infinite parameter version of Wilks' theorem, for empirical likelihood methods in curve estimation. Section 3 will describe a bootstrap calibration approach to selecting the appropriate contour. In both these sections we develop ideas first for density estimation, and then discuss the relatively minor modifications which are appropriate for nonparametric regression. Section 4 presents numerical examples, Section 5 discusses our main conclusions, and Section 6 gives proofs of results from Section 2.

The present paper represents a first step in developing empirical likelihood methods in an infinite parameter setting, as alternatives to more conventional techniques suggested by Bickel and Rosenblatt (1973), Knafl, Sacks and Ylvisaker (1985), Härdle and Bowman (1988) and Hall and Titterington (1988). We shall not attempt to develop second-order refinements of asymptotic theory for empirical likelihood, of the type studied by DiCiccio, Hall and Romano (1989a, 1989b) and Hall and LaScala (1989) in the finite parameter case. Nevertheless it is worth remarking that in finite parameter problems empirical likelihood has been shown to yield confidence regions which are very close to those of likelihood-based statistics (Hall 1989). This demonstrates that the shape of an empirical likelihood region conveys important information, and indicates that the empirical likelihood method is worthy of attention in the more complex infinite parameter case.
2. METHODOLOGY

2.1. Nonparametric density estimation

The empirical likelihood of a distribution $F$ based on a sample $X_1, \ldots, X_n$ is defined to be $L(F) = \Pi F\{X_i\}$ where $F\{x\}$ is the probability that $F$ places on the singleton $\{x\}$. We use $p_i = p_i(F)$ to denote $F\{X_i\}$ below. Where a parameter $\theta$ is the subject of interest, the profile empirical likelihood of a candidate value $\theta_1$ for $\theta$ is defined to be the maximum of $L(F)$ over all distributions $F$ with $\theta(F) = \theta_1$. We abuse notation and denote this profile $L(\theta_1)$, the argument of $L$ being clear from context. It is usual to restrict attention to distributions $F$ for which $\sum p_i = 1$, although an exception is made in Section 4. For example, when $\theta$ is $E(X_1)$ then $L(\theta_1)$ equals the maximum of $\Pi p_i$ subject to each $p_i \geq 0$, $\sum p_i = 1$ and $\sum p_i X_i = \theta_1$.

Consider a sample $X_1, \ldots, X_n$ of real values from a density $f_0$. The kernel density estimate (e.g. Silverman 1986, Section 2.4) of $f_0$ is

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

where $K$ is a kernel function, often a probability density, and $h > 0$ is a bandwidth. At a given point $x$ therefore $\hat{f}(x)$ is simply the mean of the random variables $K_i = h^{-1} K\{(x-X_i)/h\}$. We shall use $\mu_0$ to denote $E(K_i)$. Dependence of $K_i$ upon $x$ will be suppressed for notational convenience.

The value of the empirical likelihood for a candidate value $f_1(x)$ of $\mu_0(x)$ is readily shown to be $L(f_1) = \Pi \hat{p}_i$, where $\hat{p}_i = n^{-1}\{1 + \lambda(K_i - f_1)\}^{-1}$ and the function $\lambda$, which is proportional to a Lagrange multiplier in the constrained maximization problem, is determined by

$$\sum_{i=1}^{n} (K_i - f_1) \{1 + \lambda(K_i - f_1)\}^{-1} = 0.$$ 

This holds provided only that $\min K_i < f_1 < \max K_i$. The maximum value of $L(f_1)$ is $n^{-n}$, and occurs when $f_1 = \hat{f}$.

Following the usual prescription for either parametric or empirical likelihood (e.g. Owen 1988), define the empirical loglikelihood ratio to be

$$l(f_1) = -2 \log \left\{ \frac{L(f_1)}{L(\hat{f})} \right\} = 2 \sum_{i=1}^{n} \log \{1 + \lambda(K_i - f_1)\},$$

and signed root empirical loglikelihood to be

$$s(f_1) = \text{sgn}(\hat{f} - f_1) l(f_1)^{\frac{1}{2}}.$$
For each \( x \) the signed root evaluated at \( E(\mu_0) \) is asymptotically standard normal, and hence can be used to make inferences on \( \mu_0(x) \), the latter being a "blurred" version of \( f_0(x) \). Since \( h \) is usually taken of order \( n^{-1/5} \) then Theorem 1 of Owen (1988) does not directly apply, but to cover this case one can replace the central limit theorem there by one based on triangular arrays. Similarly one can obtain a joint confidence region for \( \mu_0 \) at any finite set of \( x \) values.

Our idea is that confidence bands for the unknown true density \( f_0 \) may be obtained from uniform bounds on the empirical loglikelihood or the signed root empirical loglikelihood. Given an interval \([a, b]\) over which we wish to construct the bands, and numbers \( c, c_1, c_2 > 0 \), define the classes of densities

\[
F(c) = \{ f : l(f)(x) \leq c \text{ for } a \leq x \leq b \},
\]

\[
F(c_1, c_2) = \{ f : c_1 \leq s(f)(x) \leq c_2 \text{ for } a \leq x \leq b \}.
\]

We take these classes, for appropriate values of \( c, c_1 \) and \( c_2 \) and with bias corrections where necessary, to represent confidence bands on \([a, b]\) for the unknown true density \( f_0 \). The set \( F(c) \) is a level set of the empirical loglikelihood functional \( L(f) = \sup_{a \leq x \leq b} l(f)(x) \) mentioned in the introduction.

If \( \alpha \) is the prescribed coverage level then we might wish to choose \( c, c_1 \) and \( c_2 \) such that \( \Pr \{ f_0 \in F(c) \} \) and \( \Pr \{ f_0 \in F(c_1, c_2) \} \) are both approximately equal to \( \alpha \). The present infinite parameter problem differs from finite parameter problems in two important ways. Firstly, the loglikelihood ratio is evaluated at the expected value of \( \hat{f} \) under the true density, not at the true density itself. This indicates that a bias correction of the confidence band will usually be in order. Secondly, the asymptotic distribution of the supremum of the loglikelihood ratio is a Type III extreme value distribution (e.g. David 1981, p. 259), not chi squared. Compare Bickel and Rosenblatt (1973). In practice it would be advisable to use a penultimate approximation to the extreme value limit, such as the bootstrap approximation described in Section 3, but the extreme value limit theorem is nevertheless of importance in describing first-order properties of empirical likelihood.

The following conditions on \( f_0, K \) and \( h \) will be needed in our infinite parameter version of Wilks' theorem. To simplify notation, we suppose here and below that the interval \([a, b]\) over which we are estimating \( f_0 \) is actually the interval \([0, 1]\). Assume that \( f_0 \) and \( f_0' \) are uniformly continuous, and \( f_0 > 0 \) on \([0, 1]\); that \( (1 + z^2)|K(z)| \) is bounded, \( K' \) exists and is bounded, and

\[
\int |z| \{ |K(z)| + |K'(z)| \} \, dz < \infty;
\]

and that \( (nh)^{-1} + h \langle \log n \rangle^4 \to 0 \) as \( n \to \infty \). The latter condition is very mild, since consistency of \( \hat{f} \) already demands that \( (nh)^{-1} + h \to 0 \). Define \( a_h = (2\log h^{-1})^{1/3} \) and \( b_h = a_h + a_h^{-1} \log C_0 \), where
$C_0 = \{ \int (K')^2 \}^{\frac{1}{2}} (\int K^2)^{-\frac{1}{2}} (2\pi)^{-1}$. Recall that $\mu_0 = E_{f_0}(\hat{f})$, and put $M_1 = \sup s(\mu_0)$ or $-\inf s(\mu_0)$, and $M_2 = \sup l(\mu_0)^{\frac{1}{2}}$, where the suprema and infimum are taken over the interval $[0, 1]$.

THEOREM 2.1. Under the above conditions and for $j = 1, 2$ we have

$$P \{ a_h (M_j - b_h) \leq y \} \rightarrow \exp(-y e^{-y}), \quad -\infty < y < \infty,$$

as $n \rightarrow \infty$.

This theorem may be used to select values of $c, c_1$ and $c_2$ such that the confidence bands $F(c)$ and $F(c_1, c_2)$, defined in (2.2) and (2.3), have asymptotic coverage $\alpha$ when viewed as confidence bands for $\mu_0$. For example, applying the theorem in the case $j = 2$ we see that if $y_0$ is chosen so that $\exp(-2e^{-y_0}) = \alpha$, and if we take $c = (a_h^{-1} y_0 + b_h)^2$, then we shall have

$$\Pr\{\mu_0 \in F(c)\} \rightarrow \alpha. \quad (2.4)$$

Applying the theorem for $j = 1$, choosing $y_1, y_2$ such that $\exp(-e^{-y_2}) - \exp(-e^{-y_1}) = \alpha$, and taking $c_k = a_h^{-1} y_k + b_h$ for $k = 1, 2$, we shall have

$$\Pr\{\mu_0 \in F(c_1, c_2)\} \rightarrow \alpha. \quad (2.5)$$

Our proof of Theorem 2.1 demonstrates that the signed root loglikelihood ratio $s$, evaluated at $\mu_0$, is asymptotic to

$$s_0 = (nh)^{\frac{1}{2}} (\hat{f} - \mu_0) / \left( f_0 \int K^2 \right)^{\frac{1}{2}},$$

in the sense that $s(\mu_0)/s_0$ converges in probability to unity uniformly on the interval $[0, 1]$. This property is typical of the first-order behavior of empirical likelihood, for in a finite dimensional setting, signed root empirical loglikelihood evaluated at the true parameter value $\theta_0$ is usually asymptotic to $(\hat{\theta} - \theta_0)/(\text{var} \hat{\theta})^{\frac{1}{2}}$; see DiCiccio, Hall and Romano (1989b). Note that $s_0$ is simply $\hat{f}$ standardized for location and scale; the asymptotic variance of $\hat{f}$ is $\text{var}(\hat{f}) = (nh)^{-1} f_0 \int K^2$.

Let $c, c_1, c_2$ be the constants appearing in (2.4) and (2.5). Bearing in mind the comments in the previous paragraph we see that the confidence bands $F(c)$ and $F(c_1, c_2)$ are closely approximated by the bands

$$G(c) = \left\{ f : -c_1^{1/2} \text{var}(\hat{f})^{1/2} \leq f - \hat{f} \leq c_1^{1/2} \text{var}(\hat{f})^{1/2} \text{ on } [0, 1] \right\}, \quad (2.6)$$

$$G(c_1, c_2) = \left\{ f : -c_2 \text{var}(\hat{f})^{1/2} \leq f - \hat{f} \leq -c_1 \text{var}(\hat{f})^{1/2} \text{ on } [0, 1] \right\}. \quad (2.7)$$
respectively. Indeed, the ideas in our proof of Theorem 2.1 may be employed to show that

\[ \text{pr}\{\mu_0 \in F(c)\Delta G(c)\} \to 0, \quad \text{pr}\{\mu_0 \in F(c_1, c_2)\Delta G(c_1, c_2)\} \to 0, \]

(2.8)

where \( \Delta \) denotes symmetric difference. Compare (2.4) and (2.5).

We have shown that Wilks’ theorem allows us to construct confidence bands for \( \mu_0 \). To convert these into confidence bands for \( f_0 \) we should correct for bias, \( \beta = \mu_0 - f_0 \). The amount of bias correction which is necessary may be gauged from the sensitivity of the asymptotic coverage level \( \alpha \) to changes in the bands. Formulae (2.6) and (2.7) show that the asymptotic coverage of the bands is affected by adjustments of size \((nh)^{-\frac{1}{2}}(\log h^{-1})^{-\frac{1}{2}}\) to the width of the bands. Any perturbation of the bands which is of smaller order than \((nh)^{-\frac{1}{2}}(\log h^{-1})^{-\frac{1}{2}}\) will not affect the asymptotic coverage accuracy of the bands, when viewed as confidence bands for \( \mu_0 \). It follows that if \( \hat{\beta} \) is an estimator of \( \beta \) satisfying

\[ \sup_{0 \leq x \leq 1} |\hat{\beta}(x) - \beta(x)| = o_p\left\{(nh)^{-\frac{1}{2}}(\log h^{-1})^{-\frac{1}{2}}\right\}, \]

(2.9)

if \( H \) denotes any one of the bands \( F(c), F(c_1, c_2), G(c) \) and \( G(c_1, c_2) \), and if we bias-correct \( H \) to

\[ H_{\text{corr}} = H - \hat{\beta} = \{f - \hat{\beta} : f \in H\}, \]

then

\[ \text{pr}(f_0 \in H_{\text{corr}}) \to \alpha. \]

(2.10)

In checking this result, note (2.4), (2.5) and (2.8).

For the numerical work in Section 4 we shall use an explicit bias correction based on the asymptotic formula

\[ \beta(x) = \kappa h^2 f''_0(x) + o(h^2), \]

valid with \( \kappa = \frac{1}{2} \int z^2 K(z)dz \) when \( K \) is a symmetric probability density. The bias correction is not always necessary. For example, if the estimator is constructed so that

\[ \sup_{0 \leq x \leq 1} |\beta(x)| = o\left\{(nh)^{-\frac{1}{2}}(\log h^{-1})^{-\frac{1}{2}}\right\}, \]

which will be the case if \( h = o\left\{(n \log n)^{-1/5}\right\} \), then (2.9) holds with \( \hat{\beta} \equiv 0. \)
2.2. Nonparametric regression

In the problem of nonparametric regression we wish to estimate the conditional mean \( f(x) = E(Y|X = x) \) from pairs \((X_1, Y_1), \ldots, (X_n, Y_n)\), without imposing structural assumptions on \( f \). We base our method on the kernel regression estimate \( \hat{f} = \sum K_i Y_i / \sum K_i \) where \( K_i = h^{-1} K((x - X_i)/h) \) as before. When the observations are reweighted by distribution \( F \) with atoms \( p_i \) the regression estimate becomes \( \hat{f}(F) = \sum p_i K_i Y_i / \sum p_i K_i \) or equivalently \( 0 = \sum p_i K_i (Y_i - \hat{f}(F)) \). The profile empirical likelihood at the value \( f_1 \) is found by maximizing \( \Pi p_i \) subject to

\[
\sum_{i=1}^{n} p_i K_i (Y_i - f_1) = 0. \tag{2.11}
\]

It is usual in regression problems to regard the \( X_i \)'s as fixed, either by design or by conditioning, and we shall indicate that we take this view by writing \( x_i \) instead of \( X_i \) in all the work which follows. We take our regression model to be of the form

\[
Y_i = f(x_i) + \epsilon_i, \quad 1 \leq i \leq n, \tag{2.12}
\]

where the \( \epsilon_i \)'s are independent identically distributed random variables with zero means. Analogues of all our results may be developed for the so-called correlation model, in which the pairs are regarded as proper bivariate random variables. The justification for the reweighting above is clearer in the correlation model, for then the components \( K_i (Y_i - f_1) \) are independent and identically distributed. Under our model they are independent but differ in distribution.

Our model assumes homoscedasticity. More generally, if there are grounds for suspecting heteroscedasticity then we might replace \( \epsilon_i \) by \( \sigma(x_i) \epsilon_i \) in (2.12), where \( \sigma^2(x) \) denotes the variance of \( Y \) when \( X = x \) and the \( \epsilon_i \)'s are once again assumed independent and identically distributed with zero mean. In practice we would estimate \( \sigma^2(x) \) by \( \hat{\sigma}^2(x) \) say. An appropriate modification of empirical likelihood is to replace the constraint (2.11) by

\[
\sum_{i=1}^{n} p_i K_i (Y_i - f_1) \hat{\sigma}(x_i)^{-2} = 0 \tag{2.13}
\]

wherein the observations are weighted in inverse proportion to their estimated variance.

Maximization of \( \Pi p_i \) subject to each \( p_i \geq 1, \sum p_i = 1 \) and (2.13), produces the empirical loglikelihood \( L(f_1) = \Pi \hat{p}_i \), where \( \hat{p}_i = n^{-1} \{1 + \lambda K_i (Y_i - f_1) \hat{\sigma}(x_i)^{-2}\} \) and the function \( \lambda \) is determined by

\[
\sum_{i=1}^{n} K_i (Y_i - f_1) \hat{\sigma}(x_i)^{-2} \{1 + \lambda K_i (Y_i - f_1) \hat{\sigma}(x_i)^{-2}\}^{-1} = 0.
\]
The maximum likelihood estimate is now the weighted kernel regression estimator:

\[
\hat{f}(x) = \left[ \sum_{i=1}^{n} Y_i \hat{\sigma}(x_i)^{-2} K\{(x - x_i)/h\} \right] / \left[ \sum_{i=1}^{n} \hat{\sigma}(x_i)^{-2} K\{(x - x_i)/h\} \right].
\]  (2.14)

Empirical loglikelihood evaluated at \( f_1 \) is given by

\[
l(f_1) = -2 \log\{L(f_1)/L(\hat{f})\} = 2 \sum_{i=1}^{n} \log(1 + \lambda(Y_i - f_1)\hat{\sigma}(x_i)^{-2}K_i),
\]  (2.15)

and its signed root is

\[
s(f_1) = \text{sgn}(\hat{f} - f_1)l(f_1)^{\frac{1}{2}}.
\]  (2.16)

We shall assume that \( \sigma^2(x) \) is known up to a constant factor, that is the ratio \( \hat{\tau}^2 = \hat{\sigma}^2(x)/\sigma^2(x) \) does not depend on \( x \). It is possible to relax this restriction, but rigorous theory is tedious to derive or even describe without it. Under the restriction it is permissible to remove the circumflex from \( \hat{\sigma} \) in (2.14) and (2.15), since the effect of \( \hat{\tau} \) cancels. In the case of (2.15), this is equivalent to replacing \( \lambda \) by \( \lambda/\hat{\tau} \). Thus, the expected value of \( \hat{f} \) under the true model \( Y_i = f_0(x_i) + \sigma(x_i)\epsilon_i \), is

\[
\mu_0 = E_{f_0}(\hat{f}) = \left\{ \sum_{i=1}^{n} f_0(x_i)\sigma(x_i)^{-2}K_i \right\} / \left\{ \sum_{i=1}^{n} \sigma(x_i)^{-2}K_i \right\}.
\]

Once again we may define confidence bands \( F(c) \) and \( F(c_1, c_2) \) by (2.2) and (2.3) respectively, where of course \( l(f) \) and \( s(f) \) are given by the definitions in the present section rather than those in Section 2.1. With the exception that regularity conditions needed for Wilks’ theorem are slightly different, the results and discussion in the paragraph containing (2.2) and (2.3), and in the ensuing three paragraphs, apply as they stand to the present case of nonparametric regression. In particular, if we define \( a_h = (2 \log h^{-1})^{\frac{1}{2}} \) and \( b_h = a_h^{-1} \log C_0 \) where \( C_0 = \{ f(K')^2 \}^{\frac{1}{2}}(f(K^2)^{-\frac{1}{2}}(2\pi)^{-1} \), if we choose \( y_0, y_1 \) and \( y_2 \) such that \( \exp(-2\epsilon^{-y_0}) = \alpha \) and \( \exp(-e^{-y_2}) - \exp(-e^{-y_1}) = \alpha \), and if we put \( c = (a_h^{-1}y_0 + b_h)^2 \) and \( c_k = a_h^{-1}y_k + b_h \) for \( k = 1, 2 \), then we shall have

\[
\text{pr}\{\mu_0 \in F(c)\} = \alpha \quad \text{and} \quad \text{pr}\{\mu_0 \in F(c_1, c_2)\} = \alpha.
\]

Furthermore, if \( \beta = \mu_0 - f_0 \) denotes bias, if the bias estimate \( \hat{\beta} \) satisfies

\[
\sup_{0 \leq x \leq 1} |\beta(x) - \hat{\beta}(x)| = o_p\left\{ (nh)^{-\frac{1}{2}}(\log h^{-1})^{-\frac{1}{2}} \right\},
\]

if \( H \) denotes any one of the bands \( F(c) \) and \( F(c_1, c_2) \) just considered, and if we bias-correct \( H \) to \( H_{\text{corr}} = H - \hat{\beta} \), then \( \text{pr}(f_0 \in H_{\text{corr}}) = \alpha \).
We conclude this section by stating the new version of Wilks' theorem. For simplicity, let the interval $[a, b]$ over which we are estimating $f$ be $[0, 1]$. Take the regression model to be $Y_i = f(x_i) + \sigma(x_i)\epsilon_i$, and let $\hat{f}, \hat{F}$ and $s$ be given by (2.14)-(2.16), with $\hat{\sigma}^2(x)/\sigma^2(x)$ not depending on $x$. Assume that $f$ and $f'$ are continuous on $[-\epsilon, 1+\epsilon]$ for some $\epsilon > 0$; that $K$ and $K'$ are bounded and compactly supported; that the real-line may be partitioned into a finite number of intervals each of which $K$ is monotone; that for some $\gamma > 2$, $E(\epsilon_1 \log |\epsilon_1|^\gamma) < \infty$ and $n^{(\gamma-2)/\gamma}h$ is bounded away from zero; and that $(nh^2)^{-1}(\log n)^6 + h(\log n)^4 \to 0$.

We also need a well-behaved design sequence and so we assume that the design variables $x_i$ are drawn randomly from a population with distribution function $G$ whose first two derivatives $g = G'$ and $g'$ are bounded on $[-\epsilon, 1+\epsilon]$, and $g > 0$ on $[0, 1]$. The $x_i$ are fixed by conditioning in our analysis. We say that a property holds "with $x$-probability one" if it holds for almost all sequences $x_1, x_2, \ldots$ drawn randomly from the distribution $G$. Define $M_1 = \sup s(\mu_0)$ or $-\inf s(\mu_0)$, and $M_2 = \sup l(\mu_0)^\frac{1}{2}$, where the suprema and infimum are taken over $[0, 1]$. An infinite parameter version of Wilks' theorem for nonparametric regression is as follows.

**THEOREM 2.2.** Under the above conditions, for sequences $x_1, x_2, \ldots$ having $x$-probability one, and for $j = 1$ and 2, we have

$$P\{a_h(M_j - b_h) \leq y\} \to \exp(-ye^{-y}), \quad -\infty < y < \infty,$$

as $n \to \infty$.

This result is also available for non-randomly generated design. In particular it holds if we take $x_i = x_{ni} = i/n$ for $1 \leq i \leq n$, and only minimal changes to the proof are necessary.

**3. Bootstrap Calibration**

**3.1. Nonparametric density estimation**

We showed in Section 2.1 that in principle, the constants $c, c_1$ and $c_2$ required to construct our density confidence bands $F(c)$ and $F(c_1, c_2)$, defined at (2.2) and (2.3), may be obtained from an infinite parameter version of Wilks' theorem. However, that result is essentially a limit theorem for extreme values of a Gaussian process, and it is well-known (e.g. Hall 1979; Leadbetter, Lindgren and Rootzén 1983, pp. 39-40) that distributions of Gaussian extremes converge very slowly to their extreme value limits. Therefore it is
advisable to consider a method which better approximates the distribution of empirical loglikelihood. In this section we suggest bootstrap estimates.

We wish to construct a confidence band for the value taken by the true density \( f_0 \) on the interval \([0, 1]\). Draw a resample \( \mathcal{X}^* = \{X_1^*, \ldots, X_n^*\} \) from the sample \( \mathcal{X} = \{X_1, \ldots, X_n\} \), using random sampling with replacement. Construct the bootstrap empirical loglikelihood,

\[
I^*(f_1) = 2 \sum_{i=1}^{n} \log \{1 + \lambda^* (K_i^* - f_1)\},
\]

where \( K_i^*(x) = h^{-1}K((x - X_i^*)/h) \) and \( \lambda^* \) is the solution of \( \sum (K_i^* - f_1) (1 + \lambda^* (K_i^* - f_1))^{-1} = 0 \). Compare (2.1) and (3.1). The bootstrap signed root empirical loglikelihood is

\[
s^*(f_1) = \text{sgn}(\hat{f}^* - f_1) I^*(f_1)^{1/2},
\]

where \( \hat{f}^* = (nh)^{-1} \sum K_i^* \).

Note particularly that \( E(\hat{f}^* | \mathcal{X}) = \hat{f} \), so that in the resampling argument \( \hat{f} \) should play the role of \( \mu_0 = E_{\mu_0}(\hat{f}) \). Therefore the conditional bootstrap distributions of \( I^*(\hat{f}) \) and \( s^*(\hat{f}) \) are approximations to the unconditional distributions of \( I(\mu_0) \) and \( s(\mu_0) \), respectively. In particular, if we calculate \( \hat{c}, \hat{c}_1 \) and \( \hat{c}_2 \) such that

\[
\text{pr}\{I^*(\hat{f})(x) \leq \hat{c} \text{ for } 0 \leq x \leq 1 | \mathcal{X}\} = \alpha,
\]

\[
\text{pr}\{\hat{c}_1 \leq s^*(\hat{f})(x) \leq \hat{c}_2 \text{ for } 0 \leq x \leq 1 | \mathcal{X}\} = \alpha
\]

then these quantities should be good estimates of \( c, c_1 \) and \( c_2 \) such that

\[
\text{pr}\{I(\mu_0) \leq c \text{ for } 0 \leq x \leq 1\} = \alpha,
\]

\[
\text{pr}\{c_1 \leq s(\mu_0) \leq c_2 \text{ for } 0 \leq x \leq 1\} = \alpha.
\]

Our bootstrap confidence bands are simply \( F(\hat{c}) \) and \( F(\hat{c}_1, \hat{c}_2) \).

We shall report on applications of this method in Section 4. There we shall select \( \hat{c}_1 \) and \( \hat{c}_2 \) so that the band \( F(\hat{c}_1, \hat{c}_2) \) is equal-tailed, in the sense that

\[
\text{pr}\{\hat{c}_1 \leq s^*(\hat{f})(x) \text{ for } 0 \leq x \leq 1 | \mathcal{X}\} = \text{pr}\{s^*(\hat{f})(x) \leq \hat{c}_2 \text{ for } 0 \leq x \leq 1 | \mathcal{X}\} = \frac{1}{2}(1 + \alpha).
\]

The reason for the approximate equality at the end of (3.5) is that there is positive probability that a realization of \( s^*(\hat{f})(x) \) will be smaller than \( \hat{c}_1 \) at some \( x \) and larger than \( \hat{c}_2 \) at another \( x \). Equality of tail
errors is a convenient way of fixing \( \hat{c}_1 \) and \( \hat{c}_2 \) in (3.4), and is in line with most contemporary work on the bootstrap. However, there are alternatives. For example one could choose \( \hat{c}_1 \) and \( \hat{c}_2 \) so as to minimize the average width of the band over the interval \([0, 1]\), subject to condition (3.4). Of course, there is no such latitude in the choice of \( \hat{c} \) in (3.3).

When defining \( \hat{c} \), \( \hat{c}_1 \) and \( \hat{c}_2 \) we have used the common percentile method; see for example Efron (1979). There is no need to entertain pivotal methods, such as percentile-\( t \), since empirical loglikelihood is pivoted internally. This may be seen from the fact that \( s(\mu_0) \) is asymptotic to the standardized statistic \((\hat{f} - E\hat{f})/(\text{var}\hat{f})^{1/2}\), as noted during our discussion in Section 2.1.

We contend that our bootstrap approximations to the distributions of \( l(\mu_0) \) and \( s(\mu_0) \) are better than the extreme value approximations developed in Section 2.1. When the bootstrap approximations are assessed for fixed \( x \), their good performance may be demonstrated theoretically by a rather lengthy argument of the type in Singh (1981) and Hall (1988). For example, it will be found that our bootstrap approximation to the distribution of \( s(\mu_0)(x) \) is second-order correct, in that it correctly captures the effect of skewness. Therefore it improves on the normal approximation, which ignores skewness. We should point out that in this nonparametric problem the effect of skewness is of size \( (nh)^{-1/2} \), the effect of kurtosis of size \( (nh)^{-1} \), and so on; in a parametric problem these influences are of sizes \( n^{-1/2} \) and \( n^{-1} \), respectively.

The theorem below describes first-order properties of the bootstrap, and demonstrates that our bootstrap method does yield confidence bands which enjoy asymptotically correct coverage.

Define \( \hat{c} \), \( \hat{c}_1 \) and \( \hat{c}_2 \) as in (3.3) and (3.4). In the following theorem it is not necessary to specify whether \( \hat{c}_1 \) and \( \hat{c}_2 \) are chosen to yield equal-tailed regions, or whether they are selected by some other criterion.

**THEOREM 3.1.** Assume the conditions of Theorem 2.1, and in addition that \( (nh^2)^{-1}(\log n)^6 \rightarrow 0 \). Then

\[
\Pr(\mu_0 \in F(\hat{c})) \rightarrow \alpha \quad \text{and} \quad \Pr(\mu_0 \in F(\hat{c}_1, \hat{c}_2)) \rightarrow \alpha
\]

as \( n \rightarrow \infty \).

Theorem 3.1 describes the performance of \( F(\hat{c}) \) and \( F(\hat{c}_1, \hat{c}_2) \) as confidence bands for \( \mu_0 \), rather than for \( f_0 \). A bias correction may be necessary if we are to use these bands for \( f_0 \). Denote bias by \( \beta = \mu_0 - f_0 \), and let \( \hat{H} \) be either \( F(\hat{c}) \) or \( F(\hat{c}_1, \hat{c}_2) \). If the bias estimate \( \hat{\beta} \) satisfies (2.9), and if the corrected band is \( \hat{H}_{\text{corr}} = \hat{H} - \hat{\beta} = \{f - \hat{\beta} : f \in \hat{H}\} \), then

\[
\Pr(f_0 \in \hat{H}_{\text{corr}}) \rightarrow \alpha.
\]
This result may be proved in the same manner as (2.10).

In practice the bandwidth \( h \) would always be chosen to be of larger order than \( n^{\epsilon - \frac{1}{2}} \), for some \( \epsilon > 0 \). Therefore the condition \( (nh^2)^{-1}(\log n)^6 \rightarrow 0 \) in Theorem 3.1 is assuredly satisfied. A proof longer than that in Section 6.3 shows that this condition may be relaxed considerably.

3.2. Nonparametric regression

The ideas and methods described in Section 3.1 go over with only minor changes in the case of nonparametric regression. For the sake of brevity we shall restrict ourselves to itemizing these changes, with only minimal discussion.

Recall from Section 2.2 that our model is

\[
Y_i = f(x_i) + \sigma(x_i)\epsilon_i,
\]

where the \( x_i \)'s are regarded as fixed, the \( \epsilon_i \)'s are independent and identically distributed with zero mean, and the function \( \sigma^2(x) \) is estimated by \( \hat{\sigma}^2(x) \) with \( \hat{\sigma}^2(x)/\sigma^2(x) \) not depending on \( x \). Define \( \hat{f} \) by (2.14). The \( \epsilon_i \)'s comprise the independent disturbances driving the model, and so we should resample the residuals \( \hat{\epsilon}_i = (Y_i - \hat{f}(x_i))/\hat{\sigma}(x_i) \) rather than the pairs in the sample \( \mathcal{X} = \{(x_1, Y_1), \ldots, (x_n, Y_n)\} \). However, the \( \epsilon_i \)'s do not have zero mean:

\[
\hat{\epsilon}' = n^{-1} \sum_{i=1}^{n} \hat{\epsilon}_i = \bar{Y} - n^{-1} \sum_{i=1}^{n} w_i Y_i
\]

where

\[
\hat{w}_i = \hat{\sigma}(x_i)^{-2} \left( \sum_{j=1}^{n} \hat{\sigma}(x_k)^{-2} K\{(x_j - x_k)/h\} \right)^{-1} K\{(x_j - x_i)/h\}.
\]

Therefore we work with the centered residuals, \( \hat{\epsilon}_i = \hat{\epsilon}_i' - \bar{\epsilon}' \).

Let \( \{\epsilon^*_1, \ldots, \epsilon^*_n\} \) denote a sample drawn randomly, with replacement, from \( \{\epsilon_1, \ldots, \epsilon_n\} \). Put \( Y^*_i = \hat{f}(x_i) + \epsilon^*_i \) and write \( \mathcal{X}^* = \{(x^*_1, Y^*_1), \ldots, (x^*_n, Y^*_n)\} \) for the bootstrap resample. Note that

\[
E(\epsilon^*_i | \mathcal{X}) = n^{-1} \sum_{i=1}^{n} \hat{\epsilon}_i = 0.
\]

Construct the bootstrap empirical loglikelihood

\[
\hat{l}(f_1) = 2 \sum_{i=1}^{n} \log \left\{ 1 + \lambda^*(Y^*_i - f_1)\hat{\sigma}(x_i)^{-2}K_1 \right\},
\]

(3.6)
where \( K_i = h^{-1} K \{(x - x_i)/h\} \) is exactly as in Section 2.2, and the function \( \lambda^* \) is determined by
\[
\sum_{i=1}^{n} (Y_i^* - f_1) \hat{\sigma}(x_i)^{-2} \left\{ 1 + \lambda^*(Y_i^* - f_1) \hat{\sigma}(x_i)^{-2} K_i \right\}^{-1} = 0.
\]

Compare (2.15) and (3.6). The bootstrap signed root empirical log likelihood is
\[
s^*(f_1) = \text{sgn}(\hat{f}^* - f_1) l^*(f_1)^{1/2}
\]
where \( \hat{f}^* = \{\sum_{i=1}^{n} Y_i^* \hat{\sigma}(x_i)^{-2} K_i\}/\{\sum \hat{\sigma}(x_i)^{-2} K_i\} \).

Define \( \tilde{f} = \{\sum_{i=1}^{n} \tilde{f}(x_i) \hat{\sigma}(x_i)^{-2} K_i\}/\{\sum \hat{\sigma}(x_i)^{-2} K_i\} \) and note that \( E(\hat{f}^*|\mathcal{X}) = \tilde{f} \). Hence in the resampling argument, \( \tilde{f} \) should play the role of \( \mu_0 = E_{f_0}(\hat{f}) \). In particular, the conditional bootstrap distributions of \( l^*(\tilde{f}) \) and \( s^*(\tilde{f}) \) are approximations to the unconditional distributions of \( l(\mu_0) \) and \( s(\mu_0) \), respectively.

Therefore if we compute \( \hat{c}, \hat{c}_1 \) and \( \hat{c}_2 \) such that
\[
\text{pr}\{l^*(\tilde{f})(x) \leq \hat{c} \text{ for } 0 \leq x \leq 1|\mathcal{X}\} = \alpha
\]
and
\[
\text{pr}\{\hat{c}_1 \leq s^*(\tilde{f})(x) \leq \hat{c}_2 \text{ for } 0 \leq x \leq 1|\mathcal{X}\} = \alpha
\]
then \( F(\hat{c}), F(\hat{c}_1, \hat{c}_2) \) are bootstrap confidence bands for \( f \) with nominal coverage \( \alpha \). In the case of the latter band, \( \hat{c}_1 \) and \( \hat{c}_2 \) would typically be chosen so that the band has equal tails:
\[
\text{pr}\{\hat{c}_1 \leq s^*(\tilde{f})(x) \text{ for } 0 \leq x \leq 1|\mathcal{X}\} = \text{pr}\{s^*(\tilde{f})(x) \leq \hat{c}_2 \text{ for } 0 \leq x \leq 1|\mathcal{X}\} = \frac{1}{2}(1 + \alpha).
\]

Compare (3.5). An analogue of Theorem 3.1 may be proved, showing that these bands do indeed have asymptotic coverage \( \alpha \).

4. EXAMPLE

This section presents an example of our method applied to density estimation. A comparison is made to a bootstrap method for choosing both the family of bands and their level. It is possible for the sample density estimate \( \hat{f} \) to exceed all or most of the resampled \( K_i \)'s, leading to some large values of \( l^*(\hat{f})(x) \). We begin this section by describing a convenient remedy to this problem.

4.1 Empirical Likelihood for Bounded Random Variables

We shall assume that the kernel \( K \) is a bounded density. Then the random variables \( K_i \) are bounded below by 0 and above by \( K_0 = h^{-1} \sup_{x} K(x) \). It is convenient to exploit this in the definition of the
profile empirical likelihood for the mean of the $K_i$. For $f_1 \in [0, K_0]$ define $L(f_1)$ to be the maximum of $L(F)$ over distributions $F$ supported in $[0, K_0]$ with expectation $f_1$. The maximizing $F$ will have support in \{0\} \cup \{K_1, \ldots, K_n\} \cup \{K_0\}$ so the problem is still finite dimensional.

Suppose that $f_1 > K = n^{-1} \sum K_i$. Then the maximizing $F$ will have support in $\{K_1, \ldots, K_n\} \cup \{K_0\}$. Let $p_i = F\{K_i\}$ for $i = 0, \ldots, n$. Then $L(F) = \Pi_{i=1}^n p_i$ and $\sum_{i=0}^n p_i K_i = f_1$. If one knew the maximizing value of $p_0$, the other values of $p_i$ could be found as before by maximizing the usual empirical likelihood of the sample at the value $g(p_0) = (f_1 - p_0 K_0)/(1 - p_0)$. Let $H(\mu)$ maximize $\log \Pi_{i=0}^n p_i$ over $p_i \geq 0$, $\sum_{i=1}^n p_i = 1$ and $\sum_{i=1}^n p_i K_i = \mu$. That is $H(\mu) = -\frac{1}{2} l(\mu)$. Then

$$L(f_1) = \max_{p_0} (1 - p_0)^n \exp\{H\{g(p_0)\}\}.$$ 

We may assume $p_0 \geq \max\{0, (f_1 - K(n))/(K_0 - K(n))\}$ because $H(\mu)$ is only defined for $\mu < K(n)$. Also the maximizing $p_0$ must satisfy $g(p_0) \geq K$, so that $p_0 \leq (f_1 - K)/(K_0 - K)$. Now

$$\log L(f_1) = \max_{p_0} n \log(1 - p_0) + H\{g(p_0)\}$$

(4.1) is convex because both terms in (4.1) are. This is trivial for the first term; the second derivative of the second term is $H''\{g(p_0)\} g'\{g(p_0)\}^2 + H'\{g(p_0)\} g''\{g(p_0)\}$ which is negative because $H$ is concave, $g$ is convex and $H'$ is negative for $g(p_0) > K$. Convexity in $p_0$ makes it easy to find the maximum numerically. If $p_0 = 0$ is feasible and $-n + H'(f_1)(f_1 - K_0) < 0$ then the maximizing $p_0$ is 0, so it is possible to determine when it is necessary to search for $p_0$.

In the example below this modification is made to empirical likelihood. It has the advantage of damping the bootstrap signed root process $l^*(\hat{f})(x)$, especially near the ends of the range of interest. Without the modification the suprema of $l^*(\hat{f})$ are much larger. There was no need to introduce an extra atom on 0 because a great many $K_i$ are near zero.

4.2. Example

Silverman (1986) lists 107 eruption lengths, in minutes, for the Old Faithful geyser. We consider density estimation using for $K$ the standard Gaussian density. Least squares cross-validation (Silverman 1986) suggests $h = .15$ is a reasonable bandwidth. We take our interval of interest to be $[1.67, 4.93]$ with endpoints matching the shortest and longest eruptions. Our asymptotics do not account for dependence of the interval upon the data, but in practice one chooses the interval of interest after seeing the data.
Mapping our interval to the unit interval would change the bandwidth \( h \) to \( \frac{15}{(4.93 - 1.67)} \approx .046 \). This leads to \( a_h \approx 2.48 \) and \( b_h \approx 1.60 \). For simultaneous 80%, 90% and 95% confidence bands the corresponding values of \( y \) in Theorem 2.1 are 2.19, 2.94 and 3.66 respectively and this leads to critical signed roots of 2.48, 2.79 and 3.08 respectively. For example \( F(3.08^2) \) in (2.2) is the asymptotically justified 95% confidence band and \( F(-3.08, 3.08) \) in (2.3) is one choice of 95% confidence band. For \( b = 1, \ldots, 500 \) with \( l_b^* \) the empirical loglikelihood based on the \( b \)th resample

\[
m_b = \sup_{1.67 \leq x \leq 4.93} \left\{ l_b^*(\hat{f}(x)) \right\}^{1/2}
\]

was computed. Of these \( m_{(400)} = 3.51 \), \( m_{(480)} = 4.81 \) and \( m_{(470)} = 5.55 \). This suggests that the critical values of the signed root computed using Theorem 2.1 are too small. The corresponding equal tailed regions chosen by bootstrap calibration are \( F(-4.81, 2.46) \), \( F(-5.55, 2.80) \) and \( F(-6.87, 2.95) \). The first of these contains 405, not 400 resampled densities for the reason given after (3.5); the other bands differed still less than this from their nominal levels. Note that in \( F(c_1, c_2) \), it is \( c_1 \) that determines the upper bound for the densities and \( c_2 \) the lower. The equal tailed intervals extend farther above the density estimate than they do below it, in terms of signed root empirical loglikelihood.

Figure 1 shows the first 6 realizations of the bootstrap signed root process \( l^*(\hat{f})(x) \) plotted on a grid of 128 points extending a distance \( 2h \) beyond the sample extremes. The asymptotic theory treats this process as a realization of a stationary continuous Gaussian process with mean 0 and covariance given after equation (6.6). The process is smooth as it is supposed to be, but is negatively skewed.

Figure 2 shows the density estimate and equal tailed bootstrap-calibrated empirical likelihood confidence bands at the 81%, 90% and 95% levels for this data. Of the resampled densities, 10% cross the upper bound of the 81% band, 10% cross the lower bound and 1% cross both. The dotted curve is the bias-corrected density estimate and the solid curves are the bias-corrected confidence bands. The bias correction was made as described in Section 2.1. No unimodal density fits between the 90% bands but unimodal densities can fit between the 95% bands. The use of confidence bands is not a very powerful way to test modality. For instance the bands must allow that \( f_0 \) might be as high as 0.5 say for any \( x \) between 2.3 and 3.1. What they do not convey is that it is unlikely that \( f_0 \) is so high at many points in that range.

The bumps in the upper bounds that appear between 2.3 and 3.1 correspond to specific observation values: two at 2.93, one at each of 2.72 and 2.5 and three at points between 2.25 and 2.33.

Because of the bias correction, some points of the lower bounds are negative. One could dispense with
the bias correction and consider the bands to be for \( f_0 \) convolved with \( h^{-1}K(\cdot/h) \). The advantage of a Gaussian choice for \( K \) in that case is that the number of modes in \( f \) or its derivatives is not increased by convolution (Silverman 1981). If \( f \) is smooth enough to make the bias correction then we could use a higher order kernel to construct \( \hat{f} \) and reduce the error rate. We choose instead to use the lower order kernel since it is then possible to construct bands for \( f \) instead of \( f \) convolved with the higher order kernel.

Of 500 bootstrap density estimates, 34 were extreme at one point or other of the range. Thus it was necessary to use more resamples. Of the resamples \( \hat{f}_b^*(x) \) for \( b = 1, \ldots, 2000 \) there were 58 that were extreme at one or more points of the domain. A band formed by pointwise intervals \([\hat{f}_b(0.01)(x), \hat{f}_b(0.99)(x)]\) contained 1892 resampled curves and the band \([\hat{f}_b(0.05)(x), \hat{f}_b(0.95)(x)]\) contained 1850 resampled curves. Figure 3 presents bootstrap bands based on pointwise intervals \([\hat{f}_k^*(x), \hat{f}_k(2000-k+1)]\) for \( k = 1, 4, 10 \) which strictly contain 1942, 1812 and 1622 resampled curves respectively, for nominal coverages of 97.1%, 90.6% and 81.1%. The same bias correction was applied to these bands as was applied to the empirical likelihood ones and as in Figure 2 the dotted line represents the bias corrected density estimate. The bands are rough because of the discreteness of their construction. The most extreme bands contain no unimodal densities, but we suspect that these bands are too narrow, as they are analogues of the percentile bootstrap method which typically produces confidence sets that are too small. In one dimension there are usually only 2 bootstrap resamples that determine the boundary of the confidence region. In higher dimensions as we see here, the number of resamples on the boundary of the region may be comparable to the number of resamples that are outside the region, making it more difficult to decide what region to use for a given coverage level. The same phenomena arise when one uses the bootstrap to determine the shape of a region in the plane (Owen 1990).

5. CONCLUSIONS

We have shown that empirical likelihood can be used to construct confidence bands for density estimates and regressions. A version of Wilks’ theorem holds for empirical likelihood in these infinite dimensional settings, though we have reason to expect that bootstrap calibration of the empirical likelihood bands is superior. Using the bootstrap to set the shape of the confidence bands as well as their level, runs into difficulties in high dimensional settings.

We have developed a version of empirical likelihood for bounded random variables that may be of independent interest. In it, the bounds for the random variables are included as support points of the multinomial along with the observations.
Extensions to other curve estimates, such as survival functions with censored data, hazard functions and conditional statistics $T\{F(Y|X = z)\}$ can be made. Berk and Jones (1979) construct confidence bands for the distribution function $F(x) = \Pr(X \leq x)$ using a method that is similar to ours, and they also obtain an extreme value limit.

In nonparametric regression, consider $X$ or $Y$ taking values in $\mathbb{R}^2$. The empirical likelihood procedure will produce smooth confidence sandwiches or tubes with shape determined by the data. In the latter case selecting the shape of the tubal cross sections could be quite difficult by other means.

6. PROOFS

6.1. Proof of Theorem 2.1

Write $f$ for the true density $f_0$, and $E$ for expectation under $f$. As noted in Section 2.1, the result of maximizing $\Pi p_i$ subject to $p_i \geq 0$, $\sum p_i = 1$ and $\sum p_i K_i = f$, is to produce $\bar{p}_i = n^{-1} (1 + \lambda(K_i - f_i))^{-1}$, where the function $\lambda$ is determined by $\sum(K_i - f_i)(1 + \lambda(K_i - f_i))^{-1} = 0$. It may be shown from this result, after some algebra, that under the conditions of Theorem 2.1,

\[
\begin{align*}
\ell(E\hat{f}) = -2 & \sum_{i=1}^{n} \log(n\hat{p}_i) \left\{ \sum_{i=1}^{n} (K_i - E\hat{f}) \right\}^{-1} \left\{ \sum_{i=1}^{n} (K_i - E\hat{f}) \right\}^{-1} + o_p \{(\log n)^{-2}\} \\
& = n h \left\{ f(x) \int K^2 \right\}^{-1} \left\{ \hat{f}(x) - E\hat{f}(x) \right\}^{-1} + o_p \{(\log n)^{-2}\},
\end{align*}
\]

where each of the remainder terms is of the stated order uniformly in $0 \leq z \leq 1$. To obtain this uniformity in $z$ it is helpful to note that $K$ is Hölder continuous, so that methods based on smoothness of $K$ and Bernstein's inequality (e.g. Stone 1984) may be employed.

Therefore it suffices to prove that under the conditions of Theorem 2.1, and defining

\[
M_{i}^{[1]} = (nh)^{\frac{1}{2}} \left( \int K^2 \right)^{-\frac{1}{2}} \sup_{0 \leq z \leq 1} \{ \hat{f}(x) - E\hat{f}(x) \} f(x)^{-\frac{1}{2}},
\]

\[
M_{2}^{[1]} = (nh)^{\frac{1}{2}} \left( \int K^2 \right)^{-\frac{1}{2}} \sup_{0 \leq z \leq 1} |\hat{f}(x) - E\hat{f}(x)| f(x)^{-\frac{1}{2}},
\]

we have for $j = 1, 2$ and $k = 1$,

\[
\Pr \left\{ a_k(M_j^{[k]} - b_k) \leq y \right\} = \exp(-je^{-y}), \quad -\infty < y < \infty.
\]

The case where sup in (6.2) is replaced by $-\inf$ may be treated similarly.
Write $\hat{F}$ for the empirical distribution function of $X_1, \ldots, X_n$. By Theorem 3 of Komlós, Major and Tusnády (1975) there exists a standard Brownian bridge $W^0$, whose almost sure structure relative to $\hat{F}$ depends on $n$, such that the distribution of the random variable

$$V_1 = n (\log n)^{-1} \sup_{-\infty < z < \infty} |\hat{F}(x) - F(x) - n^{-\frac{1}{2}} W^0(F(x))|$$

satisfies $\Pr(V_1 > x) \leq C_1 \exp(-C_2 x)$ for all $x > 0$ and $n \geq 2$. Here $C_1$, $C_2$ are positive absolute constants.

Write $W^0(t) = W(t) - tW(1)$ where $W$ is a standard Wiener process, and put $C_3 = (\sup f) \int |K|$ and $V_2 = C_3 |W(1)|$. Then

$$\hat{f}(x) - E\hat{f}(x) = -n^{-\frac{1}{2}} \int \{\hat{F}(x - hz) - F(x - hz)\} dK(z)$$

$$= -n^{-\frac{1}{2}} h^{-1} \int W^0(F(x - hz)) dK(z) + (nh)^{-1} (\log n) R_1(x)$$

$$= -n^{-\frac{1}{2}} h^{-1} \int W(F(x - hz)) dK(z) + (nh)^{-1} (\log n) R_1(x) + n^{-\frac{1}{2}} R_2(x), \quad (6.4)$$

where $\sup |R_1| \leq V_1 \int |K'|$ and $\sup |R_2| \leq V_2$.

The processes $\xi_1(x) = -\int W\{F(x - hz)\} dK(z)$, $\xi_2(x) = \int K((x - y)/h) f(y) \frac{1}{2} dW(y)$ have identical mean and covariance functions, and so have identical distributions. Defining $\xi_3(x) = \int K((x - y)/h) dW(y)$, $\xi_4(x) = f(x)^{\frac{1}{2}} \xi_3(x)$ and $\xi_5(x) = \xi_2(x) - \xi_3(x)$ we may prove that $E\{\xi_5(x)\}^2 \leq C_4 h^2$ and $E\{\xi_5(x_1) - \xi_5(x_2)\}^2 \leq C_5 h^2 |x_1 - x_2|$ uniformly in $h$ and in $x, x_1, x_2 \in [0, 1]$. It therefore follows by Fernique's lemma (Leadbetter, Lindgren and Rootzén 1983, p. 219) that $|\xi_4(x)| = O_p(h \log n)$ uniformly in $x \in [0, 1]$. Combining the results from (6.5) down we conclude that there exists a stationary Gaussian process $\eta(x)$ with the same distribution as $(h \int K'^2)^{-\frac{1}{2}} \xi_3(hz)$, such that, uniformly in $x$,

$$(nh)^{\frac{1}{2}} \left\{ f(x) \int K^2 \right\}^{-\frac{1}{2}} \{\hat{f}(x) - E\hat{f}(x)\} = \eta(x/h) + O_p \left[\left( (nh)^{-\frac{1}{2}} + h^\frac{3}{2} \right) \log n \right]. \quad (6.6)$$

The process $\eta$ has zero mean and covariance function

$$r(t) = E\{\eta(x)\eta(x + t)\} = 1 + \left( \int K^2 \right) \int K(z) \{K(z + t) - K(z)\} dz = 1 - Ct^2 + o(t^2)$$

as $t \to 0$, where $C = \int (K')^2/(2 \int K^2)$. Therefore, putting

$$M_1^{[2]} = \sup_{0 \leq x \leq h^{-1}} \eta(x) \quad \text{and} \quad M_2^{[2]} = \sup_{0 \leq x \leq h^{-1}} |\eta(x)|,$$

and applying Theorem A1 of Bickel and Rosenblatt (1973), we obtain result (6.3) for $j = 1, 2$ and $k = 2$. It now follows from (6.6) that if $\{(nh)^{\frac{1}{2}} + h^\frac{3}{2}\}(\log n)^2 \to 0$ then (6.3) holds for $j = 1, 2$ and $k = 1$.  

18
6.2. Proof of Theorem 2.2

It may be proved that under the conditions of Theorem 2.2,

\[ l(E\bar{f}) = nh \left\{ g(x) \int K^2 \right\}^{-1} \{ \bar{f}(x) - E\bar{f}(x) \}^2 + o_p \left\{ (\log n)^{-2} \right\}, \tag{6.7} \]

where the remainder term is of the stated order uniformly in \( 0 \leq x \leq 1 \), and we have assumed without loss of generality that \( E(e^2) = 1 \). Compare (6.1) and (6.7). The method of proof involves truncating the \( e_i \)'s at \( n^{1/\gamma}(\log n)^{-1} \), noting that \( P(\max_{1 \leq i \leq n} |e_i| > n^{1/\gamma}(\log n)^{-1}) \to 0 \), recentering the truncated \( e_i \)'s at expectations, using Bernstein's inequality to derive (6.7) when \( e_i \) is replaced by

\[ e'_i = e_i I\{|e_i| \leq n^{1/\gamma}(\log n)^{-1} \} - E \left[ e_i I\{|e| \leq n^{1/\gamma}(\log n)^{-1} \} \right], \]

and showing that this implies the original version of (6.7).

We shall start from (6.7). Therefore it suffices to prove that under the conditions of Theorem 2.2, and defining

\[ M'_1 = (nh)^{1/2} \left( \int K^2 \right)^{-1/2} \sup_{0 \leq x \leq 1} \{ \bar{f}(x) - E\bar{f}(x) \} g(x)^{1/2}, \]

\[ M'_2 = (nh)^{1/2} \left( \int K^2 \right)^{-1/2} \sup_{0 \leq x \leq 1} |\bar{f}(x) - E\bar{f}(x)| g(x)^{1/2}, \]

we have for \( j = 1, 2, \)

\[ \Pr\{a_h(M'_j - b_h) \leq y\} \to \exp(-je^{-y}), \quad -\infty < y < \infty. \tag{6.8} \]

Write \( \{(x_{(i)}, e_{(i)}), 1 \leq i \leq n\} \) for the collection of pairs \( \{(x_i, e_i), 1 \leq i \leq n\} \) whose first co-ordinate has been ranked in order of increasing size: \( x_{(1)} \leq \ldots \leq x_{(n)} \). Note particularly that the \( e_{(i)} \)'s represent a deterministic re-ordering of the \( e_i \)'s, since we regard the \( x_i \)'s as fixed. Put \( S_j = \sum_{i \leq j} e_{(i)}. \) Since \( E(\|e\|^\gamma) < \infty \) then there exists an increasing, continuous function \( H \) satisfying the conditions of Theorem 2.6.6 of Csörgő and Révész (1981, p. 108) and such that \( E\{H(|e|)\} < \infty \) and \( H(x)/x^{1/\gamma} \uparrow \infty \) as \( x \uparrow \infty \). Note that \( E(e^2) = 1 \). Hence by Theorem 2.6.7 of Csörgő and Révész (1981, p. 110), there exists a sequence of independent standard normal variables \( N_1, \ldots, N_n \), whose almost sure structure relative to \( e_1, \ldots, e_n \) depends on \( n \), such that with \( T_j = \sum_{i \leq j} N_i, \)

\[ \sup_{1 \leq j \leq n} |S_j - T_j| = o_p \left\{ n^{1/\gamma}(\log n)^{-1} \right\}. \]
Define $K(i)(x) = K(((x - x(i))/h)$ for $1 \leq i \leq n$, $K(n+1) = 0$ and $\tilde{g} = (nh)^{-1} \sum K(i)$. Then

$$nh(\tilde{f} - E\tilde{f})\tilde{g} = \sum_{i=1}^{n} e(i)K(i) = \sum_{i=1}^{n}(S_i - S_{i-1})K(i)$$

$$= \sum_{i=1}^{n} S_i(K(i) - K(i+1)) = \sum_{i=1}^{n} T_i(K(i) - K(i+1)) + R_1$$

where $|R_1| \leq \left(\sup |S_j - T_j|\right) \sum |K(i) - K(i+1)|$. Since the real line may be broken into $m$, say, intervals on each of which $K$ is monotone, then

$$\sup_{-\infty < x < \infty} \sum_{i=1}^{n} |K(i)(x) - K(i+1)(x)| \leq 2m \sup_{-\infty < x < \infty} |K(i)(x)|.$$ 

Hence $\sup |R_1| \leq 2m(\sup |K|)(\sup |S_j - T_j|) = o_p\{n^{-1/2}(\log n)^{-1}\}$. Therefore, defining

$$\xi_1(x) = n^{-\frac{1}{2}} \sum_{i=1}^{n} T_i \{K(i)(x) - K(i+1)(x)\} = n^{-\frac{1}{2}} \sum_{i=1}^{n} N_iK(i)(x),$$

we have

$$\sup_{-\infty < x < \infty} \{nh\tilde{g}(x)\{\tilde{f}(x) - E\tilde{f}(x)\} - n^{\frac{1}{2}}\xi_1(x)\} = o_p\{n^{1/2}(\log n)^{-1}\}. \tag{6.9}$$

Routine methods from the theory of density estimation (e.g. Prakasa Rao 1983, p. 38 ff) may be used to prove that under the conditions of Theorem 2.2, and with $x$-probability one, $\tilde{g} - g = o\{(\log n)^{-\frac{1}{2}}\}$ uniformly on the interval [0,1]. It will follow from the work below that $(\tilde{f} - E\tilde{f})\tilde{g} = O_p\{(nh)^{-\frac{1}{2}}(\log n)^{\frac{1}{2}}\}$ uniformly on [0,1]. Result (6.9) now implies that

$$\sup_{0 \leq x \leq 1} \{nh\tilde{g}(x)\{\tilde{f}(x) - E\tilde{f}(x)\} - n^{\frac{1}{2}}\xi_1(x)\} = o_p\left\{(n^{1/2} + (nh)^{\frac{1}{2}})(\log n)^{-1}\right\} \tag{6.10}$$

Let $W$ denote a standard Wiener process on $(-\infty, \infty)$, with $W(0) = 0$. Define processes $\xi_2$ through $\xi_7$ via:

$$\xi_2(x) = \sum_{i=1}^{n} K\{(x - x(i))/h\} \{W(i/n) - W((i - 1)/n)\},$$

$$\xi_3(x) = \sum_{i=1}^{n} K\{[x - G^{-1}(i/n)]/h\} \{W(i/n) - W((i - 1)/n)\},$$

$$\xi_4(x) = \int K\{[x - G^{-1}(t)]/h\} dW(t),$$

$$\xi_5(x) = \int K\{(x - y)/h\} g(y)^{\frac{1}{2}} dW(y),$$

$$\xi_6(x) = \int K\{(x - y)/h\} dW(y),$$

$$\xi_7(x) = g(x)^{\frac{1}{2}} \xi_6(x).$$
Tedious calculation of covariances, and application of Fernique's lemma, demonstrates that under the conditions of Theorem 2.2, and with \( \mathcal{X} \)-probability one,

\[
\sup_{0 \leq x \leq R} |\xi_j(x) - \xi_k(x)| = o_p \left\{ h^{\frac{1}{4}} (\log n)^{-1} \right\}
\]

for \((j, k) = (2, 3), (3, 4)\) and \((5, 7)\). The Gaussian processes \( \xi_1 \) and \( \xi_2 \) have the same distribution, as do \( \xi_4 \) and \( \xi_5 \). It therefore follows from (6.10) that there exists a stationary Gaussian process \( \eta(x) \) with the same distribution as \( (h \int K^2)^{-\frac{1}{4}} \xi_0 (hx) \), such that, uniformly in \( x \),

\[
\sup_{0 \leq x \leq R} |nhg(x) \{ \tilde{f}(x) - E \tilde{f}(x) \} - \left\{ nhg(x) \int K^2 \right\}^{\frac{1}{2}} \eta(x/h)| = o_p \left\{ n^{1/\gamma} + (nh)^{\frac{1}{2}} \right\} (\log n)^{-1}. \tag{6.11}
\]

By assumption, \( n^{(\gamma - 2)/\gamma} h \) is bounded away from zero, which implies that \( n^{1/\gamma} / (nh)^{\frac{1}{2}} \) is bounded. Therefore the right-hand side of (6.11) equals \( o_p \{ (nh)^{\frac{1}{2}} (\log n)^{-1} \} \).

Formula (6.11) is our analogue of (6.6). It leads directly to the desired result (6.8), as may be seen by noting the argument in the paragraph immediately following (6.6).

6.3. Proof of Theorem 3.1

Write \( \hat{F}^* \) for the empiric distribution functions of \( X_1^*, \ldots, X_n^* \). By the argument leading to (6.4) we may prove that conditional on \( \mathcal{X} \) there exists a Brownian bridge \( W_0^* \) such that

\[
\hat{f}^*(x) - \hat{f}(x) = -n^{-\frac{1}{2}} h^{-1} \int W_0^* \{ \hat{F}(x - hz) \} dK(z) + (nh)^{-1} (\log n) R^*_1(x),
\]

where \( \sup |R^*_1| \leq V^*_1 \int |K'| \) and \( \Pr(V^*_1 > x|\mathcal{X}) \leq C_1 \exp(-C_2 x) \) for the same constants \( C_1, C_2 \) as in Section 6.1. Application of Fernique's lemma to the process

\[
\xi^*_1(x) = \int \left[ W_0^* \{ \hat{F}(x - hz) \} - W_0^* \{ F(x - hz) \} \right] dK(z),
\]

conditional on \( \mathcal{X} \), allows that for \( C > 0 \) sufficiently large,

\[
\Pr \left\{ \sup_{0 \leq x \leq R} |\xi^*_1(x)| > Cn^{-\frac{1}{4}} (\log n)^{\frac{1}{2}} |\mathcal{X} \right\} \rightarrow 0
\]

in probability. Therefore

\[
\hat{f}^*(x) - \hat{f}(x) = -n^{-\frac{1}{2}} h^{-1} \int W_0^* \{ F(x - hz) \} dK(z) + R^*_1(x), \tag{6.12}
\]

21
where for $C > 0$ sufficiently large, and $k = 2$,

$$\text{pr} \left\{ \sup_{0 \leq x \leq 1} |R^*_z(x)| > Cn^{-\frac{1}{4}}h^{-\frac{1}{2}}(\log n)^{\frac{1}{2}}|\mathcal{X}| \right\} \to 0$$

in probability.

The Gaussian process on the right-hand side of (6.12) has a distribution which is identical to that of the process on the right-hand side of (6.5), both conditional on $\mathcal{X}$ and unconditionally. The argument which follows the latter result may therefore be used to prove that there exists a Gaussian process $\eta^*$, whose distribution conditional on $\mathcal{X}$ is identical to that of the unconditional distribution of the process $\eta$ appearing in (6.6), and such that

$$(nh)^{\frac{1}{2}} \{f(x) \int K^2 \}^{-\frac{1}{2}} \{\hat{f}^*(x) - \hat{f}(x)\} = \eta^*(x/h) + R^*_q(x),$$

(6.13)

where for $C > 0$ sufficiently large,

$$\text{pr} \left\{ \sup_{0 \leq x \leq 1} |R^*_q(x)| > Cn^{-\frac{1}{4}}h^{-\frac{1}{2}}(\log n)^{\frac{1}{2}}|\mathcal{X}| \right\} \to 0$$

(6.14)

in probability.

The argument which formerly lead to result (6.1) in the proof of Theorem 2.1 may be used to show that

$$I^*(\hat{f})(x) = nh \left\{f(x) \int K^2 \right\}^{-1} \{\hat{f}^*(x) - \hat{f}(x)\}^2 + R^*_q(x),$$

(6.15)

where for each $\varepsilon > 0$,

$$\text{pr} \left\{ \sup_{0 \leq x \leq 1} |R^*_q(x)| > \varepsilon(\log n)^{-2}|\mathcal{X}| \right\} \to 0$$

(6.16)

in probability. Combining (6.13)–(6.16) we conclude that if $n^{-\frac{1}{4}}h^{-\frac{1}{2}}(\log n)^{\frac{1}{2}} \to 0$ then $s(\hat{f}^*)(x) = \eta^*(x/h) + R^*_q(x)$, where for each $\varepsilon > 0$,

$$\text{pr} \left\{ \sup_{0 \leq x \leq 1} |R^*_q(x)| > \varepsilon(\log n)^{-1}|\mathcal{X}| \right\} \to 0$$

in probability. Theorem 3.1 follows easily form this result and Theorem 2.1, on noting the properties of the Gaussian process $\eta$ mentioned in the last paragraph of the proof of Theorem 2.1.

ACKNOWLEDGEMENTS

The authors wish to thank Bond University for sponsoring a workshop which lead directly to this work, and the University of Chicago for the use of their computers in manuscript preparation and computation.
REFERENCES


CAPTION FOR FIGURE 1: Shown are six realizations of the bootstrapped signed root empirical loglikelihood process, given by equation (3.2).

CAPTION FOR FIGURE 2: The dotted curve is a kernel density estimate based on the Old Faithful eruption lengths. Surrounding it are bootstrap-calibrated empirical likelihood confidence bands at coverage levels 81%, 90% and 95%. The estimate and the bands have been bias-corrected as described in Section 2.1.

CAPTION FOR FIGURE 3: The dotted curve is a kernel density estimate based on the Old Faithful eruption lengths. Surrounding it are bootstrap confidence bands formed by adjusting the levels of pointwise bootstrap confidence intervals in order to attain simultaneous coverage. The nominal coverages levels are 97.1%, 90.6% and 81.1%. The estimate and the bands have been bias-corrected, as described in Section 2.1.
Figure 1: Bootstrap signed root processes