Biometrika (1996), **83**, 2, pp. 329–341
Printed in Great Britain

Empirical likelihood confidence intervals for nonparametric density estimation

BY SONG XI CHEN

School of Statistics, La Trobe University, Victoria 3083, Australia

SUMMARY

We suggest using empirical likelihood in conjunction with the kernel method to construct confidence intervals for the value of a probability density f at a point x. This suggestion arises from a simulation study which shows that confidence intervals produced by the kernel-based percentile-t bootstrap do not have the coverage claimed by the theory. This coverage discrepancy is due to a conflict between the prescribed undersmoothing and the explicit variance estimate needed by the percentile-t method. Empirical likelihood avoids this conflict by studentising internally. We show that the empirical likelihood produces confidence intervals having theoretical coverage accuracy of the same order of magnitude as the bootstrap, and which are also empirically more accurate.

Some key words: Bartlett correction; Confidence interval; Empirical likelihood; Density estimation; Kernel method; Percentile-t bootstrap; Smoothing bandwidth.

1. INTRODUCTION

Let $X_1, \ldots, X_n$ be a random sample from an unknown continuous distribution with density f. A kernel estimate for f can be viewed as a smoothed version of a histogram. However, instead of counting the number of data values falling into bins, we weight each data point by a smooth kernel function K centred at the data point. A kernel density estimator for f at an arbitrary fixed x is

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

where the kernel K determines the shape of the 'bumps' centred at each data point, and h is the smoothing bandwidth that controls the smoothness of the 'bumps'. See Silverman (1986) and Scott (1992) for comprehensive reviews of the kernel density estimation. Sometimes confidence intervals for $f(x)$ are required, for instance, to test a hypothesis about $f(x)$. An application is in a line transect survey, where confidence intervals for wildlife abundance are required.

The bootstrap has been used to construct confidence intervals for $f(x)$. Hall (1991) studied the coverage accuracy of the bootstrap confidence intervals by developing an Edgeworth expansion for the kernel density estimator. As pointed out by Hall, due to bias in the kernel estimate, the confidence intervals are not directly applicable to $f(x)$. To convert them into confidence intervals for $f(x)$, a smaller smoothing bandwidth of order

$n^{-1/(r+1)}$ should be used for two-sided percentile-t confidence intervals if an rth order kernel K is used (Hall, 1991; Hall, 1992, Ch. 4).

However, a simulation study given in § 6 shows that the percentile-t confidence intervals have poor empirical coverage. The reason for the poor coverage is a conflict between the prescribed undersmoothing and the explicit variance estimation needed by the percentile-t method. To appreciate this point, we note that $\hat{f}(x)$ only uses local information around x. The 'effective sample size' for kernel density estimation is nh rather than n, as the number of data points within interval $(x-h, x+h)$ is of size nh. Therefore, a smaller bandwidth h further reduces the 'effective sample size' as reflected by a large number of ties in the 'derived sample' $K\{h^{-1}(x-X_1)\}, \ldots, K\{h^{-1}(x-X_n)\}$. As only the 'derived sample' is used in the variance estimation for the percentile-t method, undersmoothing consequently produces erratic variance estimates. Here we have a conflicting situation where, on the one hand, we need to use a smaller h in order to convert the percentile-t confidence intervals into ones for $f(x)$ and, on the other hand, a smaller h causes erratic variance estimates and reduces the quality of the percentile-t method. Thus, the kernel density estimate, like the correlation coefficient and ratio of means, is a case where the percentile-t confidence interval does not have good empirical coverage.

To overcome the conflict faced by the percentile-t bootstrap, we consider using empirical likelihood to construct confidence intervals for $f(x)$. Empirical likelihood was introduced by Owen (1988, 1990) as an alternative to the bootstrap for constructing nonparametric confidence intervals. It has similar sampling properties to the bootstrap. However, instead of using equal weights it chooses the weights by profiling a multinomial likelihood supported on the sample. To illustrate, an empirical likelihood for a population mean μ assigns notional weight p_i to X_i, and is defined as

$$L(\mu) = \sup \prod_{i=1}^{n} p_i$$

subject to $\sum p_i = 1$ and $\sum p_i X_i = \mu$, where the latter reflects the fact that μ is the mean. The empirical likelihood is calculated via an optimisation procedure to find the optimal p_i. Confidence intervals for the mean are constructed by contouring the empirical likelihood. No explicit variance estimation is required in the construction of the empirical likelihood confidence intervals. This means that the empirical likelihood conducts the studentising internally.

We show in this paper that this 'internal studentising' adopted by empirical likelihood can overcome the problem faced by the bootstrap, and produces better confidence intervals. Hall & Owen (1993) used empirical likelihood to construct confidence bands for f, by establishing an analogue of Wilks' Theorem based on extreme value type limiting distributions. Only the first order properties were discussed and no comparison with the bootstrap was made.

The concept of empirical likelihood for density estimation is introduced in § 2. In § 3, we show that, by properly selecting the smoothness bandwidth, a nonparametric version of Wilks' Theorem, which gives us asymptotic correct confidence intervals for $f(x)$, is valid. The coverage accuracy is assessed in § 4. We show in § 5 that the coverage accuracy can be improved by a Bartlett correction. In § 6 we present some simulation results where a comparison with the percentile-t is made. In § 7, we use empirical likelihood to construct confidence intervals for $f(0)$ based on a data set from a line transect aerial survey of Southern Bluefin Tuna.

2. Empirical likelihood for density estimation

Let K be an rth order kernel. That is, for some integer $r \geqslant 2$ and constant $\kappa \neq 0$,

$$\int u^j K(u)\,du = \begin{cases} 1 & \text{if } j = 0, \\ 0 & \text{if } 1 \leqslant j \leqslant r-1, \\ \kappa & \text{if } j = r. \end{cases} \tag{1}$$

Throughout this paper, we assume that

(i) K is bounded, has compact support and satisfies (1);
(ii) f has continuous derivatives up to the rth order in a neighbourhood of x, and $f(x) > 0$;
(iii) $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

Define $K_i(x) = h^{-1}K\{h^{-1}(x - X_i)\}$ for $i = 1, \ldots, n$, and $u(x) = E\{K_i(x)\}$. For brevity of notation, we write $K_i(x)$ as K_i. It is well known that the kernel estimator $\hat{f}(x)$ is biased, as shown by

$$u(x) = f(x) + (r!)^{-1}\kappa f^{(r)}(x)h^r + o(h^r), \tag{2}$$

where $f^{(r)}$ denotes the rth derivative of f. We first introduce empirical likelihood for $u(x)$ as the mean of K_i, which will give us confidence intervals for $u(x)$. Then, by properly selecting h, we convert them into confidence intervals for $f(x)$.

Let $p_1, \ldots, p_n$ be nonnegative numbers adding to unity. The empirical likelihood for $u(x)$ is defined as

$$L\{u(x)\} = \sup_{\sum p_i K_i = u(x)} \prod_{i=1}^{n} p_i.$$

After using a Lagrange multiplier to find the optimal p_i, the log empirical likelihood ratio for $u(x)$ is

$$l\{u(x)\} = -2\log[L\{u(x)\}n^n] = 2\sum \log[1 + \lambda\{K_i - u(x)\}],$$

where λ satisfies

$$\sum_{i=1}^{n} \{K_i - u(x)\}[1 + \lambda\{K_i - u(x)\}]^{-1} = 0. \tag{3}$$

Put $\bar{s}_j = n^{-1}\sum \{K_i - u(x)\}^j$ for $j = 1$ and 2. Since $\bar{s}_2 = O_p(h^{-1})$, we may show in a manner similar to Owen (1990) that, under conditions (i), (ii) and (iii), $\lambda = O_p(n^{-\frac{1}{2}}h)$. By solving for λ in (3) we have $\lambda = \bar{s}_2^{-1}\bar{s}_1 + O_p(n^{-1}h)$. Substituting into $l\{u(x)\}$, we obtain

$$l\{u(x)\} = n\bar{s}_2^{-1}\bar{s}_1^2 + O_p(n^{-\frac{1}{2}}h).$$

Because $n^{\frac{1}{2}}\bar{s}_2^{-\frac{1}{2}}\bar{s}_1$ is asymptotically standard normal, $l\{u(x)\}$ is asymptotically χ_1^2 distributed. Thus, an empirical likelihood confidence interval for $u(x)$ with nominal coverage α is $I_\alpha = \{u(x) \mid l\{u(x)\} \leqslant c_\alpha\}$, where $\operatorname{pr}(\chi_1^2 < c_\alpha) = \alpha$.

However, what we want is a confidence interval for $f(x)$. Using I_α as a confidence interval for $f(x)$ without correcting for the bias will create a coverage discrepancy. This coverage discrepancy can be reduced by reducing the bias of $u(x)$. Indeed, as pointed out by Hall (1991) for the bootstrap, there are two approaches leading to confidence intervals for $f(x)$. One is implicitly to correct the bias by undersmoothing. Another is explicitly to correct the bias by shifting I_α by $(r!)^{-1}\kappa f^{(r)}(x)h^r$, the dominant bias term in (2). In this paper, we consider only the first approach.

The coverage of I_α for $f(x)$ is

$$\mathrm{pr}\{f(x) \in I_\alpha\} = \mathrm{pr}[l\{f(x)\} \leqslant c_\alpha].$$

To avoid confusion, we denote by $l\{f(x)\}$ the log empirical likelihood ratio for the mean of K_i, evaluated at a candidate value which happens to be $f(x)$.

3. ASYMPTOTIC COVERAGE

We show how to choose h such that I_α has correct asymptotic coverage for $f(x)$ in this section. Using the same method used to derive $l\{u(x)\}$,

$$l\{f(x)\} = 2\sum \log[1 + \lambda\{K_i - f(x)\}], \tag{4}$$

where λ satisfies

$$\sum_{i=1}^{n} \{K_i - f(x)\}[1 + \lambda\{K_i - f(x)\}]^{-1} = 0. \tag{5}$$

As analytic solutions are not obtainable for (4) and (5), expansions must be sought. Define $w_i = K_i - f(x)$, $\mu_j = E(w_i^j)$ for integer exponent $j \geqslant 1$, and $\bar{w}_j = n^{-1}\sum w_i^j$.

A Taylor expansion for $l\{f(x)\}$, whose derivation is given in Appendix 1, gives

$$l\{f(x)\} = n w_2^{-1} \bar{w}_1^2 + O_p\{nh(n^{-\frac{1}{2}} + h^r)^3\}.$$

Putting $Z = n^{\frac{1}{2}}(\bar{w}_1 - \mu_1)\mu_2^{-\frac{1}{2}}$, and noticing that $\bar{w}_2 = \mu_2 + o_p(1)$, we have

$$l\{f(x)\} = (Z + n^{\frac{1}{2}}\mu_2^{-\frac{1}{2}}\mu_1)^2\{1 + o_p(1)\} + O_p\{nh(n^{-\frac{1}{2}} + h^r)^3\}. \tag{6}$$

Therefore, $l\{f(x)\}$ is asymptotically χ_1^2 if and only if $n^{\frac{1}{2}}\mu_2^{-\frac{1}{2}}\mu_1 \to 0$. Put $\rho_j(K) = \int K^j(t)\,dt$ for integer j. Note that

$$\mu_1 = (r!)^{-1}\kappa f^{(r)}(x)h^r + o(h^r), \quad \mu_2 = h^{-1}f(x)\rho_2(K) - f^2(x) + O(h).$$

Thus, a sufficient condition for $l\{f(x)\}$ to be asymptotically χ_1^2 is that $n^{\frac{1}{2}}h^{r+\frac{1}{2}} \to 0$ as $n \to \infty$. This condition is also necessary if $f^{(r)}(x) \neq 0$. Thus, for an rth order kernel K and $f^{(r)}(x) \neq 0$,

$$\mathrm{pr}\{f(x) \in I_\alpha\} = \alpha + o(1) \tag{7}$$

if and only if $h = o\{n^{-1/(2r+1)}\}$. Note that $n^{-1/(2r+1)}$ is the order of magnitude normally obtained in density estimation by minimising the mean integrated square error. Thus, (7) implies that, to ensure I_α has correct asymptotic coverage for $f(x)$, a smaller than usual bandwidth should be used. The prescription $h = o\{n^{-1/(2r+1)}\}$ coincides with that of the bootstrap (Hall, 1991). We see from (6) that the length of I_α is asymptotically the same as the confidence interval based on the normal approximation.

Empirical likelihood does not require explicit variance estimation in the construction of I_α. Rather, as revealed by (6), it uses an optimisation procedure to obtain a profile likelihood which studentises implicitly by the true variance. This implies that the empirical likelihood actually carries out the studentising internally via the optimisation procedure. In contrast, the percentile-t method tries to approximate the distribution of a studentised statistic $\{\hat{f}(x) - \mu(x)\}/\hat{\sigma}(x)$ by generating a large number of bootstrap resamples, where $\hat{\sigma}^2(x)$ is an estimate for $\mathrm{var}\{\hat{f}(x)\}$. The performance of the percentile-t method depends heavily on the quality of the variance estimate $\hat{\sigma}^2(x)$. Therefore, when the variance estimate is of low quality, the percentile-t method performs poorly. This is exactly what happens for the percentile-t method in kernel density estimation.

4. Coverage accuracy

To study the coverage accuracy of I_α for $f(x)$, the following two conditions are used in addition to conditions (i), (ii) and (iii):

(iv) $h = o\{n^{-1/(2r+1)}\}$,

(v) $nh(\log n)^{-1} \to \infty$ as $n \to \infty$.

Condition (iv), as shown in (7), gives I_α asymptotically correct coverage for $f(x)$. Condition (v) is required to develop an Edgeworth expansion for the empirical likelihood ratio.

The results of this section are based on the following Edgeworth expansion, whose derivation is outlined in Appendix 2:

$$\begin{aligned}\mathrm{pr}\{f(x)\in I_\alpha\} &= \mathrm{pr}[l\{f(x)\} \leqslant c_\alpha]\\ &= \alpha - n^{-1}\{\mu_2^{-1}(n\mu_1)^2 + \tfrac{1}{2}\mu_2^{-2}\mu_4 - \tfrac{1}{3}\mu_2^{-3}\mu_3^2\}c_\alpha^{\frac{1}{2}}\phi(c_\alpha^{\frac{1}{2}})\\ &\quad + O(nh^{3r+1} + h^{2r}) + O\{(nh)^{-1}h^r + (nh)^{-2}\}. \qquad (8)\end{aligned}$$

We see from (8) that the coverage error for $f(x)$ is dominated by two terms. One is $n^{-1}\mu_2^{-1}(n\mu_1)^2$ of order nh^{2r+1}, which is a result of the bias μ_1. The remaining part is of order $(nh)^{-1}$, which usually exists in the Edgeworth expansion of empirical likelihood.

Simple algebra shows that $\mu_j = h^{-(j-1)}f(x)\rho_j(K) + O(h^{-(j-2)})$ for $j \geqslant 2$. Define

$$a(x, K, f) = \{(r!)^{-1}\kappa f^{(r)}(x)\}^2\{f(x)\rho_2(K)\}^{-1},$$

$$b(x, K, f) = f(x)^{-1}[\tfrac{1}{2}\rho_4(K)\{\rho_2(K)\}^{-2} - \tfrac{1}{3}\rho_3(K)^2\rho_2(K)^{-3}].$$

Then (8) can be rewritten as

$$\mathrm{pr}\{f(x)\in I_\alpha\} = \alpha - \{a(x, K, f)nh^{2r+1} + b(x, K, f)(nh)^{-1}\}c_\alpha^{\frac{1}{2}}\phi(c_\alpha^{\frac{1}{2}}) + o\{nh^{2r+1} + (nh)^{-1}\}. \qquad (9)$$

Therefore, if $f^{(r)}(x) \neq 0$, the optimal bandwidth which minimises the second order term in (9) is

$$h = (2r+1)^{-1}a(x, K, f)^{-1}b(x, K, f)n^{-1/(r+1)}. \qquad (10)$$

Substituting the above optimal h into (9) and defining $v = 2r + 1$, we have

$$\begin{aligned}\mathrm{pr}\{f(x)\in I_\alpha\} &= \alpha - n^{-r/(r+1)}\{b(x, K, f)^v v^{-v} a(x, K, f)^{-2r} + va(x, K, f)\}c_\alpha^{\frac{1}{2}}\phi(c_\alpha^{\frac{1}{2}})\\ &\quad + o\{n^{-r/(r+1)}\}. \qquad (11)\end{aligned}$$

Thus the optimal coverage error is of the exact order of $n^{-r/(r+1)}$, which is a larger order than n^{-1} for usual confidence intervals. This is because the effective sample size is $nh = n^{r/(r+1)}$. The optimal bandwidth prescribed in (10) may be used as a guideline for choosing h if $f^{(r)}(x) \neq 0$. As the ρ_j's are known after a kernel is chosen, the optimal h may be estimated by estimating $f^{(r)}(x)$.

In practice, we may first obtain kernel estimates for $f(x)$ and $f^{(r)}(x)$ with bandwidths h_0 and h_r respectively. As we are only interested in the density and its derivative at x, h_0 and h_r are chosen locally by minimising $E\{\hat{f}^{(l)}(x) - f^{(l)}(x)\}^2$ for $l = 0$ and 2 respectively. It can be shown that for $l = 0$ and 2,

$$h_l = \{(2l+1)\rho_2(K^{(l)})f(x)\}^{1/(2l+5)}\{\kappa^2 f^{(l+2)}(x)\}^{-2/(2l+5)}n^{-1/(2l+5)}. \qquad (12)$$

The bandwidths may be obtained by referring to a standard distribution to obtain values for $f(x)$ and $f^{(r+2)}(x)$, as done by Silverman (1986, p. 45) for choosing global bandwidth. A more effective adaptive kernel method can be used to obtain better estimates of $f(x)$

and $f^{(r)}(x)$. Having obtained these estimates, we compute the dominant bias term in (2) and divide it by $\hat{f}(x)$ to obtain the relative bias. If this is small, say less than 0·1 for a bandwidth, the same bandwidth can be used to construct confidence intervals for $f(x)$. If the relative bias is larger, we may use (10) to choose h. This issue is further discussed in § 7 when we analyse some data.

5. BARTLETT CORRECTION

We show in this section that Bartlett correction is valid for the current density estimation case. That empirical likelihood confidence intervals are Bartlett correctable has been proved valid in many cases: see DiCiccio, Hall & Romano (1991) for smooth functions of means, Chen & Hall (1993) for quantiles, and Chen (1993, 1994) for linear regression.

We first calculate $E[l\{f(x)\}]$. From (A1·3) in the Appendix, it can be shown that

$$E[l\{f(x)\}] = 1 + \beta_0 n^{-1} + o\{nh^{2r+1} + (nh)^{-1}\},$$

where

$$\beta_0 := \mu_2^{-1}(n\mu_1)^2 + \tfrac{1}{6}(3\mu_2^{-2}\mu_4 - 2\mu_2^{-3}\mu_3^2) \tag{13}$$

is the coefficient of the dominant coverage error term in (8). Therefore, $l\{f(x)\}$ and χ_1^2 have different means to the second order.

Now

$$\mathrm{pr}\{\chi_1^2 \leqslant c_\alpha(1 + \beta_0 n^{-1})\} = \alpha + \beta_0 c_\alpha^{\frac{1}{2}}\phi(c_\alpha^{\frac{1}{2}})n^{-1} + O[\{nh^{2r+1} + (nh)^{-2}\}^2].$$

Applying the Edgeworth expansion (8) to the mean adjusted empirical likelihood ratio $l\{f(x)\}/(1 + \beta_0 n^{-1})$, and choosing h of order $n^{-1/(r+1)}$ as suggested by (10), we have

$$\begin{aligned}\mathrm{pr}[l\{f(x)\} \leqslant c_\alpha(1 + \beta_0 n^{-1})] &= \mathrm{pr}\{\chi_1^2 \leqslant c_\alpha(1 + \beta_0 n^{-1})\} - \beta_0 n^{-1} c_\alpha^{\frac{1}{2}}\phi(c_\alpha^{\frac{1}{2}}) + O(n^{-2r/(r+1)}) \\ &= \alpha + O(n^{-2r/(r+1)}).\end{aligned} \tag{14}$$

This leads us to define a Bartlett corrected confidence interval:

$$J_{\alpha f} := \{t \mid l\{f(x)\} \leqslant c_\alpha(1 + \beta_0 n^{-1})\}.$$

Comparing (14) and (11), we see that the dominant error term $\beta_0 n^{-1} c_\alpha^{\frac{1}{2}}\phi(c_\alpha^{\frac{1}{2}})$ is removed by the Bartlett correction. Thus, $J_{\alpha f}$ will have better asymptotic coverage than I_α.

To implement the Bartlett correction practically, β_0 has to be estimated. From (13) the moment estimates for μ_j can be established for $j \geqslant 2$. However, the density derivative $f^{(r)}(x)$ has to be estimated in order to estimate μ_1, as the moment estimate for μ_1 is always zero. To avoid estimating the density derivative, we introduce a partial Bartlett corrected confidence interval

$$J_{\alpha p} := \{t \mid l\{f(x)\} < c_\alpha(1 + \beta_1 n^{-1})\},$$

where $\beta_1 = \frac{1}{6}(3\mu_2^{-2}\mu_4 - 2\mu_2^{-3}\mu_3^2)$. The idea is to do a partial mean adjustment $l\{f(x)\}/(1 + \beta_1 n^{-1})$ such that the coverage error term $n^{-1}\beta_1 c_\alpha^{\frac{1}{2}}\phi(c_\alpha^{\frac{1}{2}})$ in (8) can be removed, and the $n^{-1}\mu_2^{-1}(n\mu_1)^2$ term becomes negligible by choosing h properly. A derivation similar to that leading to (14) gives

$$\begin{aligned}\mathrm{pr}\{f(x) \in J_{\alpha p}\} &= \mathrm{pr}[l\{f(x)\} < c_\alpha(1 + \beta_1 n^{-1})] \\ &= \alpha + n^{-1}\mu_2^{-1}(n\mu_1)^2 c_\alpha^{\frac{1}{2}}\phi(c_\alpha^{\frac{1}{2}}) + o(nh^{2r+1}) + O\{(nh)^{-2}\}.\end{aligned}$$

If we choose h such that n^3h^{2r+3} is bounded, we readily have

$$\mathrm{pr}\{f(x) \in J_{\alpha p}\} = \alpha + O\{n^{-4r/(2r+3)}\}.$$

Thus, in order to use the partial Bartlett correction β_1, an even smaller h, of order $n^{-3/(2r+3)}$, should be used.

6. Simulation results

In this section we present some simulation results designed to examine the performance of the empirical likelihood confidence intervals proposed, and to compare them with their percentile-t counterparts. We chose the standard normal distribution as the true density f and constructed confidence intervals for $f(x)$ at $x = 0$. We used the second order biweight kernel,

$$K(u) = \tfrac{15}{16}(1 - u^2)^2 I(|u| \leqslant 1),$$

where I is the indicator function. Five thousand samples were generated from f with sample size $n = 20$, 30, 40 and 50 using the routines provided by Press et al. (1992, Ch. 7). For each simulated sample, we generated 499 bootstrap resamples. We assessed the coverage of four confidence intervals with nominal coverage levels 0·95. The four intervals are the empirical likelihood, I_α, the partial Bartlett corrected, $J_{\alpha p}$, the full Bartlett corrected, $J_{\alpha f}$, and the percentile-t. When constructing the full Bartlett corrected interval $J_{\alpha f}$, we used a bandwidth of $2{\cdot}26\hat{\sigma}n^{-1/9}$ to estimate $f^{(2)}(x)$ as suggested by (12) for the biweight kernel, where $\hat{\sigma}$ is the sample standard deviation.

The coverage of a kernel-based confidence interval is directly influenced by the bias, where the latter is controlled by the bandwidth h. To appreciate this, we chose four levels of bandwidth: $h = 2{\cdot}5n^{-a}$, where $a = \frac{3}{7}, \frac{1}{3}, \frac{1}{4}$ and $\frac{1}{5}$. The coefficient 2·5 and the four levels of a were selected as they produce kernel estimates with different levels of bias as shown in Table 3. Table 1 contains the simulated coverages of the four intervals for the four bandwidth levels at $x = 0$. Table 2 presents the average lengths of the empirical likelihood and percentile-t confidence intervals. We did not supply the lengths of the two Bartlett corrected confidence intervals as they were very close to those of the empirical likelihood. Table 3 gives the average bias of the kernel estimate for each h level used; these can be used to explain the coverages given in Table 1.

The simulation results can be summarised as follows. The three empirical likelihood confidence intervals perform well. Interval I_α had a much better coverage and shorter length than its percentile-t counterpart. The coverage of I_α can be further improved by the Bartlett correction. The coverage of the partial Bartlett corrected interval $J_{\alpha p}$ was close to that of the full Bartlett corrected interval when h or the bias was small. The advantage of $J_{\alpha p}$ is that it avoids density derivative estimation. However, when the bias is large, the full Bartlett corrected interval should be used. We also calculated confidence intervals by the percentile bootstrap and the usual normal approximation. The percentile-t confidence interval had the poorest coverage and greatest length. It was even worse than the percentile confidence interval as the latter does not need a variance estimate and is more stable.

We observe a positive correlation between the coverage and the bias in Tables 1 and 3. In the case of $h = 2{\cdot}5n^{-1/5}$, poor coverage for the empirical likelihood intervals is observed at $x = 0$. This is not surprising as this level of h produces a larger bias in the density estimation which makes the converting of I_α difficult. We actually did a simulation at $x = 1$, where we observed good coverages of the empirical likelihood intervals for all the four levels of h including $h = 2{\cdot}5n^{-1/5}$. This is because the second derivative of f at $x = 1{\cdot}0$ is zero, which means that the bias is one order of magnitude smaller than at $x = 0$ and consequently allows a larger bandwidth without causing severe bias.

Table 1. *Simulated coverages of confidence intervals and, in square brackets,* $100\times$ *their standard errors, for* 95% *empirical likelihood* (EL), *partial Bartlett corrected* (PBCEL), *full Bartlett corrected* (FBCEL), *and percentile-t* (Percent.-t) *confidence intervals for* $f(0)$ *with* $h = 2{\cdot}5n^{-a}$

n	EL	PBCEL	FBCEL	Percent.-t
		$a=\frac{3}{7}$		
20	0·938[0·34]	0·945[0·32]	0·947[0·32]	0·888[0·44]
30	0·942[0·33]	0·945[0·33]	0·945[0·32]	0·898[0·42]
40	0·945[0·32]	0·949[0·31]	0·950[0·31]	0·904[0·42]
50	0·944[0·33]	0·947[0·31]	0·948[0·31]	0·905[0·42]
		$a=\frac{1}{3}$		
20	0·934[0·35]	0·939[0·34]	0·946[0·32]	0·893[0·44]
30	0·936[0·34]	0·939[0·34]	0·943[0·33]	0·901[0·42]
40	0·940[0·34]	0·944[0·33]	0·946[0·32]	0·905[0·41]
50	0·937[0·34]	0·940[0·34]	0·940[0·34]	0·904[0·42]
		$a=\frac{1}{4}$		
20	0·911[0·40]	0·917[0·39]	0·924[0·37]	0·889[0·44]
30	0·923[0·37]	0·927[0·37]	0·934[0·35]	0·899[0·43]
40	0·916[0·39]	0·919[0·39]	0·927[0·37]	0·892[0·44]
50	0·922[0·38]	0·924[0·37]	0·929[0·36]	0·896[0·43]
		$a=\frac{1}{5}$		
20	0·858[0·49]	0·866[0·48]	0·873[0·47]	0·871[0·47]
30	0·873[0·47]	0·877[0·46]	0·883[0·45]	0·869[0·48]
40	0·868[0·48]	0·873[0·47]	0·881[0·46]	0·856[0·50]
50	0·885[0·45]	0·886[0·45]	0·893[0·44]	0·867[0·48]

Table 2. *Average length of confidence intervals and, in square brackets, standard errors of* 95% *empirical likelihood* (EL) *and percentile-t* (Percent.-t) *confidence intervals for* $f(0)$ *with* $h = 2{\cdot}5n^{-a}$

n	EL	Percent.-t	EL	Percent.-t
	$a=\frac{3}{7}$		$a=\frac{1}{3}$	
20	0·415[0·051]	0·551[0·575]	0·326[0·029]	0·387[0·044]
30	0·396[0·046]	0·466[0·112]	0·307[0·025]	0·343[0·030]
40	0·378[0·044]	0·426[0·046]	0·290[0·023]	0·315[0·027]
50	0·363[0·042]	0·400[0·043]	0·277[0·021]	0·296[0·026]
	$a=\frac{1}{4}$		$a=\frac{1}{5}$	
20	0·255[0·019]	0·296[0·028]	0·215[0·016]	0·249[0·024]
30	0·236[0·014]	0·260[0·021]	0·197[0·012]	0·216[0·017]
40	0·221[0·012]	0·237[0·017]	0·183[0·009]	0·195[0·014]
50	0·210[0·011]	0·221[0·016]	0·173[0·008]	0·181[0·012]

We also observe that as the bandwidth increases the length of the confidence intervals decreases. A smaller bandwidth produces empirical likelihood confidence intervals with good coverage but greater length. This reflects the trade-off between the bias and variance of kernel density estimation, where the bias affects the coverage, and variance affects the length of the intervals.

Table 3. *Average bias and, in square brackets, standard errors of the kernel density estimate for $f(0)$ with $h = 2{\cdot}5n^{-a}$*

n	$a=\frac{3}{7}$	$a=\frac{1}{3}$	$a=\frac{1}{4}$	$a=\frac{1}{5}$
20	0·015[0·114]	0·024[0·089]	0·037[0·069]	0·048[0·058]
30	0·010[0·105]	0·017[0·082]	0·029[0·063]	0·040[0·052]
40	0·007[0·102]	0·014[0·077]	0·026[0·059]	0·036[0·048]
50	0·006[0·098]	0·013[0·073]	0·023[0·056]	0·033[0·046]

7. An example

We compute the empirical likelihood confidence intervals on a data set from a line transect aerial survey of Southern Bluefin Tuna. The line transect survey (Buckland et al., 1993) has been used to estimate the tuna abundance in the Great Australian Bight in summer when the tuna tend to stay on the surface. The abundance is measured by $D = N/A$, where N is the total number of surface schools in the Bight and A is the survey area. To estimate D, an aircraft with two spotters on board is used to fly randomly allocated transect lines to detect tuna schools. Each school sighted from the transect is counted and its perpendicular distance to the transect is measured. Assume n independent schools are detected after flying a distance L. Let f be the probability density function of the sighting distances. Standard line transect theory shows that the population density $D = (2L)^{-1}E(n)f(0)$. Parametric estimators were proposed to estimate $f(0)$. However, they are not robust against changing detection conditions. Therefore, a robust nonparametric estimate should be used. The kernel method has been shown to be robust for analysing line transect data (Chen, 1996).

One way of constructing confidence intervals for D is to multiply confidence intervals for $f(0)$ by $(2L)^{-1}n$, assuming a stable sighting rate n/L. Therefore, confidence intervals for $f(0)$ are needed to construct confidence intervals for the abundance measure D. Table 4 presents a data set of 64 perpendicular sighting distances collected from one replicate of the 1993 survey. Figure 1 gives a histogram and a kernel density estimate for the data.

It is well known that kernel estimation is subject to boundary bias. In fact the asymptotic expectation of $\tilde{f}(0)$ is only $f(0)/2$. To correct for this, we double the sample size by reflecting each data value x_i to $-x_i$, and apply the kernel estimator to the combined sample. This data reflection is very natural in the line transect survey as the sightings are made on the both sides of the transect, and the density function f has zero derivative at $x = 0$ (Buckland et al., 1993).

We first estimate $f^{(2)}(0)$ which will give us some indication of the bias level of the kernel estimate. The histogram of the data suggests that the underlying distribution is close to

Table 4. *The tuna data set: perpendicular sighting distances of detected tuna schools to the transect lines*

0·19	0·28	0·29	0·45	0·64	0·65	0·78	0·85
1·00	1·16	1·17	1·29	1·31	1·34	1·55	1·60
1·83	1·91	1·97	2·05	2·10	2·17	2·28	2·41
2·46	2·51	2·89	2·89	2·90	2·92	3·03	3·19
3·48	3·79	3·83	3·94	3·95	4·11	4·14	4·19
4·36	4·53	4·97	5·02	5·13	5·75	6·03	6·19
6·19	6·45	7·13	7·35	7·77	7·80	8·81	9·22
9·29	9·78	10·15	11·32	13·21	13·27	14·39	16·26

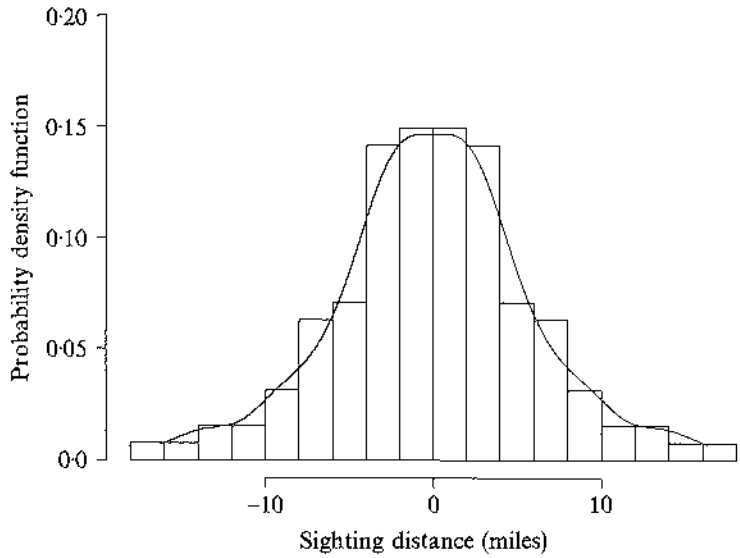

Fig. 1. Histogram and kernel density estimate for the reflected tuna data. The S-PLUS function 'ksmooth' was used with the Gaussian kernel and $h = 4{\cdot}0$.

a normal distribution after reflection. Therefore, we may just use (12) with f the density of $N(0, \sigma^2)$ and $x = 0$ in determining the bandwidths h_0 and h_2. For the Biweight kernel used, (12) implies $h_0 = 2{\cdot}45\hat{\sigma}n^{-1/5}$ and $h_2 = 2{\cdot}26\hat{\sigma}n^{-1/9}$. This gives an estimate of $-0{\cdot}0019$ for $f^{(2)}(0)$. This suggests a very small $f^{(2)}(0)$. This is no surprise, as the data show a very flat shoulder near the origin. Therefore, we cannot use the optimal bandwidth (10) as $f^{(2)}(0)$ is too small. However, as suggested by both theory and simulation, the empirical likelihood can produce accurate confidence intervals for a wide range of bandwidth values when $f^{(2)}(0)$ is small. We may just choose h as h_0 in constructing confidence intervals for $f(0)$.

This choice of h gives a kernel estimate of 0·149 for $f(0)$ and an estimated bias of −0·002. The relative bias is less than 0·2%, which is very small indeed. We obtain the following 95% confidence intervals for $f(0)$: the empirical likelihood (0·108, 0·1952), the partial and full Bartlett corrected intervals (0·1078, 0·1955) and (0·1076, 0·1957) respectively, and the percentile-t interval (0·0946, 0·1955). The empirical likelihood interval has the shortest length (0·087), whereas the percentile-t has the longest (0·101). The two Bartlett corrected intervals almost coincide each other. This is because the two Bartlett factors β_0 and β_1 are basically the same when $f^{(2)}(0)$ is small.

8. CONCLUSION

We have shown that empirical likelihood produces more accurate confidence intervals for density estimation than the percentile-t bootstrap. As shown by the simulation, the empirical likelihood confidence intervals are shorter than their percentile-t counterparts. This is due to the fact that empirical likelihood conducts the studentising internally via an optimisation procedure, whereas the percentile-t method has to use erratic variance estimates. Therefore, empirical likelihood should be used for constructing confidence intervals for density estimation.

In general, empirical likelihood relies on good and robust optimisation procedures, and is more computationally involved than bootstrap. The computation of empirical likelihood for density estimation is very regular. However, in situations when the number of structure constraints exceeds the number of parameters, as treated by Qin & Lawless (1994), computational issues arise as the convergence to the optimum can be very slow. Art Owen has developed a computer software for computing empirical likelihood for various applications. However, more work is certainly needed.

Acknowledgement

This research was carried out while the author was employed by CSIRO Division of Fisheries and Biometrics unit. The author would like to thank Dr M. J. Buckley, and Professors Peter Hall and Steve Marron for beneficial comments, and a referee and the Associate Editor for comments leading to improvements in the presentation of this paper.

Appendix 1

Taylor expansion of $l\{f(x)\}$

Along with the derivation given by Owen (1990, p. 101), (5) implies that there exists a constant $d>0$, such that $\lambda(|\bar{w}_2|-d|\bar{w}_1|)\leqslant|\bar{w}_1|$. This implies $\lambda=O_p(n^{-\frac{1}{2}}h+h^{r+1})$ as $\bar{w}_1=O_p(n^{-\frac{1}{2}}+h^r)$ and $\bar{w}_2=O(h^{-1})$. From (5),

$$\bar{w}_1-\lambda\bar{w}_2+\ldots+(-\lambda)^j n^{-1}\sum w_i^{j+1}(1+\lambda w_i)^{-1}=0. \tag{A1·1}$$

Solving equation (A1·1) for λ gives

$$\begin{aligned}\lambda=\;&\bar{w}_2^{-1}\bar{w}_1+\bar{w}_2^{-3}\bar{w}_3\bar{w}_1^2+(2\bar{w}_2^{-5}\bar{w}_3^2-\bar{w}_2^{-4}\bar{w}_4)\bar{w}_1^3+(5\bar{w}_2^{-7}\bar{w}_3^3-5\bar{w}_2^{-6}\bar{w}_3\bar{w}_4+\bar{w}_2^{-5}\bar{w}_5)\bar{w}_1^4\\&+\sum_{k=5}^{j}R_{1k}\bar{w}_1^k+O_p\{(n^{-\frac{1}{2}}+h^r)^j h\},\end{aligned} \tag{A1·2}$$

where R_{lk} denote polynomials in $\bar{w}_2,\ldots,\bar{w}_{k+1}$ with constant coefficients multiplied by $\bar{w}_2^{-(2k-1)}$. Substituting (A1·2) into (4), we obtain the following Taylor expansion of $l\{f(x)\}$:

$$\begin{aligned}l\{f(x)\}=\;&n\{\bar{w}_2^{-1}\bar{w}_1^2+\tfrac{2}{3}\bar{w}_2^{-3}\bar{w}_3\bar{w}_1^3+(\bar{w}_2^{-5}\bar{w}_3^2-\tfrac{1}{2}\bar{w}_2^{-4}\bar{w}_4)\bar{w}_1^4+8\bar{w}_2^{-6}\bar{w}_3\bar{w}_4\bar{w}_1^5\\&-(8\bar{w}_2^{-7}\bar{w}_3^3=\tfrac{8}{5}\bar{w}_2^{-5}\bar{w}_5)\bar{w}_1^5\}+O_p\{nh(n^{-\frac{1}{2}}+h^r)^6\}.\end{aligned} \tag{A1·3}$$

Appendix 2

Derivation of the Edgeworth expansion (8)

The derivation of the Edgeworth expansion (8) uses the Edgeworth expansion for the kernel density estimate developed by Hall (1991), and follows the same approach used in the proof of Theorem 3.2 of Chen & Hall (1993). Thus, only a brief outline is given here.

According to (A1·3), for $j\geqslant 5$, $l\{f(x)\}$ can be decomposed by taking the signed square root as

$$l\{f(x)\}=(n^{\frac{1}{2}}R)^2,$$

where

$$\begin{aligned}R&=\bar{w}_2^{-\frac{1}{2}}\Big\{\bar{w}_1+\tfrac{1}{3}\bar{w}_2^{-2}\bar{w}_3\bar{w}_1^2+(\tfrac{4}{9}\bar{w}_2^{-4}\bar{w}_3^2-\tfrac{1}{4}\bar{w}_3^{-3}\bar{w}_4)\bar{w}_1^3\\&\qquad+(-\tfrac{112}{27}\bar{w}_2^{-6}\bar{w}_3^3+\tfrac{49}{12}\bar{w}_2^{-5}\bar{w}_3\bar{w}_4-\tfrac{4}{5}\bar{w}_2^{-4}\bar{w}_5)\bar{w}_1^4+\sum_{k=5}^{j}T_k\bar{w}_1^k\Big\}+U_j\\&=R_j+U_j,\end{aligned}$$

say, where $U_j = O_p\{h^{\frac{1}{2}}(n^{-\frac{1}{2}} + h^r)^{j+1}\}$, and T_k denotes $\bar{w}_2^{-2(k-1)}$ multiplied by a polynomial in $\bar{w}_2, \ldots, \bar{w}_k$ with constant coefficients.

Put $V_k = \bar{w}_k - \mu_k$, $V = (V_1, \ldots, V_j)$ and

$$d_{k_1 \ldots k_m} = \left(\prod_{l=1}^{m} \partial/\partial u_{k_l} \right) R_j(u_1, \ldots, u_j) \bigg|_{u=\mu}.$$

A Taylor expansion of R_j at $\mu = (\mu_1, \ldots, \mu_j)$ has the form

$$T(V) = R_j(\mu) + \sum_{m=1}^{6} (m!)^{-1} \sum_{k_1, \ldots, k_m \in \{1, \ldots, j\}} d_{k_1 \ldots k_m} V_{k_1} \ldots V_{k_m}.$$

By proper choice of j, we have $R_j = T(V) + U_{2j}$, where $U_{2j} = O_p(n^{-3})$.

Let k_l denote the ith cumulant of $n^{\frac{1}{2}}T(V)$ for $l \geqslant 1$. After computing the partial derivatives $d_{k_1, \ldots, k_m}$ and the joint cumulants of V, the formulae given by James & Mayne (1962) are used to calculate k_l. We obtain

$$k_1 = n^{\frac{1}{2}} R_j(\mu) - \tfrac{1}{6}\mu_2^{-\frac{3}{2}}\mu_3 n^{-\frac{1}{2}} + O\{(nh)^{-\frac{1}{2}}\mu_1\},$$

$$k_2 = \sigma^2 + (\tfrac{1}{2}\mu_2^{-2}\mu_4 - \tfrac{13}{36}\mu_2^{-3}\mu_3^2)n^{-1} + O\{(nh)^{-1}\mu_1 + (nh)^{-2}\},$$

$$k_3 = O\{(nh)^{-\frac{1}{2}}\mu_1\}, \quad k_4 = O\{(nh)^{-1}\}, \quad k_l = O\{(nh)^{-(l-2)/2}\} \quad (l \geqslant 5),$$

where

$$\sigma_2 = \sum_{k_1} \sum_{k_2} d_{k_1} d_{k_2} E\{(w_1^{k_1} - \mu_{k_1})(w_1^{k_2} - \mu_{k_2})\}$$
$$= 1 + \tfrac{1}{3}\mu_2^{-2}\mu_1\mu_3 + (\tfrac{7}{9}\mu_2^{-4}\mu_3^2 - \tfrac{1}{4}\mu_2^{-1} - \tfrac{7}{12}\mu_2^{-3}\mu_4)\mu_1^2 + O(h^{3r}).$$

Based on the above expression of the cumulants and assuming $nh^{2r+1} \to 0$, a formal Edgeworth expansion for the distribution of $n^{\frac{1}{2}}T(V)$ can be formulated:

$$\begin{aligned} \mathrm{pr}\{n^{\frac{1}{2}}T(V) \leqslant x\} = \Phi(x) &- n^{-1}\{\tfrac{1}{2}\mu_2^{-1}(n\mu_1)^2 + \tfrac{1}{4}\mu_2^{-2}\mu_4 - \tfrac{1}{6}\mu_2^{-3}\mu_3^2\}x\phi(x) \\ &+ (\text{even polynomial in } x) \times \phi(x) + O(nh^{3r+1} + h^{2r}) \\ &+ O\{(nh)^{-1}h^r + (nh)^{-2}\}. \end{aligned} \qquad (\mathrm{A2{\cdot}1})$$

Using the techniques of Hall (1991), an Edgeworth expansion for the distribution of $n^{\frac{1}{2}}V$ can be established by proving an analogue of the Cramér condition based on conditions (i)–(v). Notice that $n^{\frac{1}{2}}T(V)$ is a smooth function of V. Using the technique of Bhattacharya & Ghosh (1978) or Skovgaard (1981), it can be shown that the Edgeworth expansion for the distribution of $n^{\frac{1}{2}}V$ may be transformed to produce the Edgeworth expansion given in (A2·1). Expansion (8) can be obtained from (A2·1).

REFERENCES

BHATTACHARYA, R. N. & GHOSH, J. K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.* **6**, 434–51.

BUCKLAND, S. T., ANDERSON, D. R., BURNHAM, K. P. & LAAKE, J. L. (1993). *Distance Sampling*. London: Chapman and Hall.

CHEN, S. X. (1993). On the coverage accuracy of empirical likelihood confidence regions for linear regression model. *Ann. Inst. Statist. Math.* **45**, 621–37.

CHEN, S. X. (1994). Empirical likelihood confidence intervals for linear regression coefficients. *J. Mult. Anal.* **49**, 24–40.

CHEN, S. X. (1996). Kernel estimate for density of biological populations using line transect sampling. *Appl. Statist.* **45**. To appear.

CHEN, S. X. & HALL, P. (1993). Smoothed empirical likelihood confidence intervals for quantiles. *Ann. Statist.* **21**, 1166–81.

DICICCIO, T. J., HALL, P. & ROMANO, J. P. (1991). Empirical likelihood is Bartlett correctable. *Ann. Statist.* **19**, 1053–61.

HALL, P. (1991). Edgeworth expansions for nonparametric density estimators, with applications. *Statistics* **22**, 215–32.

HALL, P. (1992). *The Bootstrap and Edgeworth Expansion.* New York: Springer-Verlag.

HALL, P. & OWEN, A. (1993). Empirical likelihood confidence bands in density estimation. *J. Computat. Graph. Statist.* **2**, 273–89.

JAMES, G. S. & MAYNE, A. J. (1962). Cumulants of functions of random variables. *Sankhyā* A **24**, 47–54.

OWEN, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–49.

OWEN, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18**, 90–120.

PRESS, W. H., FLANNERY, B. F., TEUKOLSKY, S. A. & VETTERLING, W. T. (1992). *Numerical Recipes: the Art of Scientific Computing.* Cambridge: Cambridge University Press.

QIN, J. & LAWLESS, J. (1994). Empirical likelihood and general estimating functions. *Ann. Statist.* **22**, 300–25.

SCOTT, D. W. (1992). *Multivariate Density Estimation.* New York: Wiley.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

SKOVGAARD, I. B. M. (1981). Transformation of an Edgeworth expansion by a sequence of smooth functions. *Scand. J. Statist.* **8**, 207–17.

[*Received November* 1994. *Revised October* 1995]