Austral. J. Statist. **39**(1), 1997, 47–56

EMPIRICAL LIKELIHOOD-BASED KERNEL DENSITY ESTIMATION

SONG XI CHEN[1]

La Trobe University

Summary

This paper considers the estimation of a probability density function when extra distributional information is available (e.g. the mean of the distribution is known or the variance is a known function of the mean). The standard kernel method cannot exploit such extra information systematically as it uses an equal probability weight n^{-1} at each data point. The paper suggests using empirical likelihood to choose the probability weights under constraints formulated from the extra distributional information. An empirical likelihood-based kernel density estimator is given by replacing n^{-1} by the empirical likelihood weights, and has these advantages: it makes systematic use of the extra information, it is able to reflect the extra characteristics of the density function, and its variance is smaller than that of the standard kernel density estimator.

Key words: Density estimation; empirical likelihood; extra information; kernel method.

1. Introduction

The kernel method has been a popular tool for the nonparametric estimation of the probability density function (p.d.f.) f on the basis of an independent and identically distributed (i.i.d.) sample $X_1, \ldots, X_n$ from a continuous distribution. A kernel density estimator for f at an arbitrary point x is

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

where K is a kernel function and h is a smoothing parameter that controls the smoothness of the fit.

In some statistical applications, additional information about f is available: the mean of a distribution may be known or the variance may be a known function of the mean, as occurs with estimating equations. As an example, in an aerial line transect survey for estimating the abundance of Southern Bluefin

Received January 1996; revised August 1996; accepted November 1996.

[1]School of Statistical Science, La Trobe University, VIC 3083.

Acknowledgments. The author thanks a referee for beneficial comments which improved the presentation of the paper, and Mrs Diana Hiller for proofreading it.

Tuna (Chen, 1996a), two spotters make sightings of tuna schools on both sides of randomly allocated transect lines. The distribution of perpendicular sighting distances, of the detected tuna schools to the transect lines, should have mean value zero. However, as the detection patterns of the two spotters need not be identical, we cannot assume that the distribution is symmetric about zero. This additional information usually can be expressed as

$$\mathrm{E}_X\{g_\ell(X)\} = 0 \qquad (\ell = 1, \ldots, q). \tag{2}$$

where g_ℓ are some known real functions.

The kernel estimator (1) is unable to make systematic use of such extra distributional information; instead, it is reflected passively through the data. One situation where the kernel method can handle extra information well is when the underlying density is known to be symmetric about a known point. By data reflection, the kernel method produces symmetric density estimates.

This paper uses empirical likelihood, in conjunction with the kernel method, to provide a systematic approach for capturing the extra data information. The kernel estimator (1) uses an equal probability weight n^{-1} at every data point assuming no extra information is available. However, when extra information is available, the probability weights should be constructed in such a way as to reflect the extra knowledge. Suppose the extra information can be formulated as (2). Then, an empirical likelihood-based kernel density estimator is constructed by replacing n^{-1} in (1) with the empirical likelihood weights p_i under (2). This new kernel density estimator is a bona fide probability density, provided that the kernel K is itself a density; it reflects the extra characteristic of the density function better than does the kernel density estimator.

We show that the variance of the empirical likelihood-based density estimate $\hat{f}$ is smaller than that of the standard kernel estimate $\tilde{f}$, confirming the general belief that empirical likelihood reduces the variance of an estimator in the presence of extra information (see Owen, 1991; Chen & Qin, 1993; Zhang, 1995). However, in density estimation the reduction in the variance occurs in the second order term rather than in the dominant term, as the above authors show in other situations. This is reasonable because empirical likelihood achieves a smaller variance by its use of unequal weights, which offers more flexibility than estimators using equal weights n^{-1}. However, the kernel smoothing negates the first order effect of using unequal weights.

The empirical likelihood-based kernel estimator also is flexible enough to reflect the extra characteristics of f, as represented by (2), better than the kernel estimator does. For instance, the mean of the standard kernel estimate for a density with a known mean value is not necessarily equal to that value. By imposing a zero mean constraint, the empirical likelihood-based density estimate achieves the mean exactly.

Section 2 introduces the empirical likelihood-based kernel density estimator. Section 3 examines the ability of both the empirical likelihood-based estimator

and the kernel estimator to reflect the extra characteristics of the density function. Section 4 compares the bias and variance of the empirical likelihood-based kernel estimator with those of the ordinary kernel estimator. Section 5 analyses two datasets and presents some simulation results.

2. Empirical Likelihood-based Estimator

Empirical likelihood, introduced by Owen (1988, 1990), is a computer intensive statistical method, as is the bootstrap. However, instead of applying an equal probability weight n^{-1} to all data values, empirical likelihood chooses the weights, say p_i on the ith data value X_i, by profiling a multinomial likelihood under a set of constraints. The constraints reflect either the meaning of the parameters of interest or some extra distributional knowledge. Empirical likelihood has already been used in kernel density estimation. Chen (1996b) shows that empirical likelihood can be used to construct confidence intervals for $f(x)$, which have better coverage and are shorter in length than those of the bootstrap.

If extra distributional information is available and is expressed as (2), the empirical likelihood determines the p_i by maximising a multinomial likelihood $\prod_1^n p_i$ subject to

$$\sum p_i = 1 \qquad \text{and} \qquad \sum p_i g_\ell(X_i) = 0 \qquad (\ell = 1, \ldots, q).$$

Let $\lambda_1, \ldots, \lambda_q$ be Lagrange multipliers corresponding to the q constraints. Define $\lambda = (\lambda_1, \ldots, \lambda_q)^T$ and $g(X_i) = \{g_1(X_i), \ldots, g_q(X_i)\}^T$. The optimal weights are

$$p_i = n^{-1}\{1 + \lambda^T g(X_i)\}^{-1} \qquad (i = 1, \ldots n), \tag{3}$$

where λ is the solution of

$$\sum_i \frac{g_\ell(X_i)}{1 + \lambda^T g(X_i)} = 0 \qquad (\ell = 1, \ldots, q). \tag{4}$$

An empirical likelihood-based kernel density estimator is obtained by replacing n^{-1} with the p_i at (3) in the kernel density estimator (1), i.e.

$$\hat{f}_{el}(x) = \frac{1}{h} \sum_1^n p_i K\Big(\frac{x - X_i}{h}\Big). \tag{5}$$

3. Density Estimators and Extra Information

If the extra information (2) is available, it is natural to ask a density estimate, say $\tilde{f}(x)$, constructed from the data, to satisfy (2), such that

$$\int g_\ell(x)\tilde{f}(x)\,dx = 0. \tag{6}$$

Suppose the data are from a distribution with known mean μ_0. Then, $q = 1$, $g_1(x) = x - \mu_0$ and (2) becomes $\mathrm{E}(X - \mu_0) = 0$. It is easy to show that for the kernel estimator $\hat{f}(x)$,

$$\int x\hat{f}(x)\,dx = \bar{X}.$$

Therefore, whether $\hat{f}(x)$ satisfies (2) depends entirely on the data at hand. For a finite sample, there is no guarantee that (2) will be satisfied, though $\bar{X} \to \mu_0$ in probability as $n \to \infty$.

If the mean of the distribution is known to be μ_0, the empirical likelihood chooses

$$p_i = n^{-1}\{1 + \lambda(X_i - \mu_0)\}^{-1}$$

where λ is the solution of

$$\sum_i \frac{X_i - \mu_0}{1 + \lambda(X_i - \mu_0)} = 0.$$

The empirical likelihood-based kernel density estimator is

$$\hat{f}_{el}(x) = \frac{1}{nh}\sum_{i=1}^{n} \frac{1}{1 + \lambda(X_i - \mu_0)} K\Big(\frac{x - X_i}{h}\Big). \tag{7}$$

If the kernel K is a probability density itself and is symmetric about zero, then

$$\begin{aligned}\int x\hat{f}_{el}(x)\,dx &= \frac{1}{nh}\sum \frac{1}{1 + \lambda(X_i - \mu_0)} \int xK\Big(\frac{x - X_i}{h}\Big)\,dx \\ &= \frac{1}{n}\sum \frac{X_i}{1 + \lambda(X_i - \mu_0)} = \mu_0.\end{aligned}$$

Thus the empirical likelihood-based kernel estimator $\hat{f}_{el}$ preserves the population mean.

In general, if the data are known to be from a distribution satisfying the general constraint (2), then $\hat{f}_{el}(x)$ is given by (5) with p_i given by (3). We note that for $1 \le \ell \le q$,

$$\begin{aligned}\int g_\ell(x)\hat{f}_{el}(x)\,dx &= \frac{1}{nh}\sum \frac{1}{1 + \lambda^T g(X_i)} \int g_\ell(x)K\Big(\frac{x - X_i}{h}\Big)\,dx \\ &= \frac{1}{n}\sum \frac{1}{1 + \lambda^T g(X_i)} \int g_\ell(X_i + ht)K(t)\,dt. \end{aligned}\tag{8}$$

Assume

(i) for $\ell = 1, \ldots, q$, g_ℓ are smooth functions with enough derivatives;

(ii) $\mathrm{E}\{g_\ell^{(k)}(X)\} < \infty$ for nonnegative integer $k \le 4$;

(iii) K is symmetric about zero and is a probability density.

Expanding $g_\ell(X_i + ht)$ around X_i gives, with $\sigma_k^2 = \int t^2 K(t)\,dt$,

$$\int g_\ell(X_i + ht)K(t)\,dt = g_\ell(X_i) + \tfrac{1}{2}g_\ell^{(2)}(X_i)h^2\sigma_K^2 + O(h^4). \tag{9}$$

Substituting (8) into (9) and recalling that λ is the solution of equation (4),

$$\begin{aligned}\int g_\ell(x)\hat{f}_{el}(x)\,dx &= \frac{1}{n}\sum \frac{1}{1+\lambda^T g(X_i)}\{g_\ell(X_i) + \tfrac{1}{2}g_\ell^{(2)}(X_i)h^2\sigma_K^2 + O(h^4)\} \\ &= \tfrac{1}{2}h^2\frac{\sigma_K^2}{n}\sum \frac{1}{1+\lambda^T g(X_i)}g_\ell^{(2)}(X_i) + O(h^4) \\ &= \tfrac{1}{2}h^2\frac{\sigma_K^2}{n}\sum g_\ell^{(2)}(X_i) + O_p(h^2 n^{-1/2} + h^4). \end{aligned} \tag{10}$$

In the last of the above equations, we use the fact that $\lambda_\ell = O_p(n^{-1/2})$, which can be obtained by a method similar to that of Owen (1990). It may be shown in a similar manner that

$$\int g_\ell(x)\hat{f}(x)\,dx = \frac{1}{n}\sum g_\ell(X_i) + \tfrac{1}{2}h^2\frac{\sigma_k^2}{n}\sum g_\ell^{(2)}(X_i) + O(h^4).$$

In kernel density estimation it is required that $h \to 0$ as $n \to \infty$. As the extra information (2) makes $n^{-1}\sum_i g_\ell(X_i)$ of order $n^{-1/2}$, both $\hat{f}_{el}(x)$ and $\hat{f}(x)$ satisfy the constraint (2) in probability. In fact, $\int g_\ell(x)\hat{f}_{el}(x)\,dx$ and $\int g_\ell(x)\hat{f}(x)\,dx$ both have the same dominant term of order h^2 if we use a bandwidth h at an order larger than $n^{-1/4}$. But $\int g_\ell(x)\hat{f}(x)\,dx$ has an extra second order term $n^{-1}\sum_i g_\ell(X_i)$ which can be quite large for a finite sample. Note that the use of the empirical likelihood weights leads to the disappearance of a similar term $n^{-1}\sum_i g_\ell(X_i)/[1+\lambda^T g(X_i)]$ in $\int g_\ell(x)\hat{f}(x)\,dx$. Thus, the constraint (2) is better met by the empirical likelihood- based kernel estimator $\hat{f}_{el}(x)$ than the kernel estimator.

4. Bias and Variance

In this section we investigate the bias and variance of the empirical likelihood-based kernel density estimator and compare them with those of the kernel density estimator. Using the fact that $\lambda = O_p(n^{-1/2})$, we have

$$\begin{aligned}\hat{f}_{el}(x) &= \frac{1}{nh}\sum \frac{1}{1+\lambda^T g(X_i)}K\Big(\frac{x-X_i}{h}\Big) \\ &= \frac{1}{nh}\sum[1 - \lambda^T g(X_i) + O_p(n^{-1})]K\Big(\frac{x-X_i}{h}\Big) \\ &= \hat{f}(x) - \lambda^T T_1 + \lambda^T T_2 \lambda + O_p(n^{-1}),\end{aligned}$$

where the vector T_1 and $q \times q$ matrix T_2 are defined by

$$T_1 = \frac{1}{nh}\sum g(X_i)K\Big(\frac{x - X_i}{h}\Big), \qquad T_2 = \frac{1}{nh}\sum g(X_i)g(X_i)^T K\Big(\frac{x - X_i}{h}\Big).$$

A Taylor expansion for λ, similar to those given in DiCiccio *et al.* (1988) and Chen (1994), is

$$\lambda = \Sigma^{-1}\frac{1}{n}\sum g(X_i) + O_p(n^{-1}), \tag{11}$$

where $\Sigma = \big(\operatorname{cov}\big(g_\ell(X), g_m(X)\big)\big)$. It may be shown that $\mathrm{E}(\lambda) = O(n^{-2})$ by a more detailed expansion than the one given in (11).

From (11) we see that $\mathrm{E}(\lambda^T T_1)$ and $\mathrm{E}(\lambda^T T_2 \lambda)$ agree in the first order term, i.e. $\mathrm{E}(\lambda^T T_1) = g(x)^T\Sigma^{-1}g(x)f(x)n^{-1} + o(n^{-1}) = \mathrm{E}(\lambda^T T_2\lambda)$. Thus, by the delta method,

$$\mathrm{E}\{\hat f_{el}(x)\} = \mathrm{E}\{\hat f(x)\} + o(n^{-1}). \tag{12}$$

To derive the variance of $\hat f_{el}(x)$, we notice that $\hat f_{el}(x)^2$ equals

$$\frac{1}{(nh)^2}\sum_{ij}\frac{1}{[1 + \lambda^T g(X_i)]\,[1 + \lambda^T g(X_j)]}K\Big(\frac{x - X_i}{h}\Big)K\Big(\frac{x - X_j}{h}\Big)$$
$$= \hat f(x)^2 - 2\lambda^T T_1\hat f(x) + 2\lambda^T T_2\lambda\hat f(x) + \lambda^T T_1 T_1^T\lambda + o_p(n^{-1}).$$

Using the Taylor expansion for λ and the delta method again, we can show that

$$\mathrm{E}\{\lambda^T T_1\hat f(x)\} = 2g(x)^T\Sigma^{-1}g(x)f^2(x)\,n^{-1} + o(n^{-1}),$$
$$\mathrm{E}\{\lambda^T T_2\hat f(x)\} = g(x)^T\Sigma^{-1}g(x)f^2(x)\,n^{-1} + o(n^{-1}) = \mathrm{E}(\lambda^T T_1 T_1^T\lambda).$$

Thus, $\mathrm{E}\{\hat f_{el}(x)^2\} = \mathrm{E}\{\hat f(x)^2\} - g(x)^T\Sigma^{-1}g(x)f^2(x)\,n^{-1} + o(n^{-1})$. This and (12) imply that

$$\operatorname{var}\{\hat f_{el}(x)\} = \operatorname{var}\{\hat f(x)\} - g(x)^T\Sigma^{-1}g(x)f^2(x)\,n^{-1} + o(n^{-1}). \tag{13}$$

As the coefficient of n^{-1} is always negative, there is an $O(n^{-1})$ reduction in the variance of $\hat f_{el}(x)$ even if its dominant variance term is the same as that of $\hat f(x)$. For small to medium samples, the extent of the reduction in the variance can be substantial, as we show by a simulation study in the next section. This reduction in variance is due to the use of the extra information by the empirical likelihood. However, the smoothing makes the size of the reduction a second order effect. In contrast, the difference between $\mathrm{E}\{\hat f_{el}(x)\}$ and $\mathrm{E}\{\hat f(x)\}$ is $o(n^{-1})$, and thus negligible compared with the difference between the variances. Combining (12) and (13), we immediately have

$$\mathrm{MISE}(\hat f_{el}) = \mathrm{MISE}(\hat f) - \frac{1}{n}\int g(x)^T\Sigma^{-1}g(x)f^2(x)\,dx + o(n^{-1}).$$

So, there is a reduction in the mean integrated square error due to the use of the extra information by the empirical likelihood-density estimate.

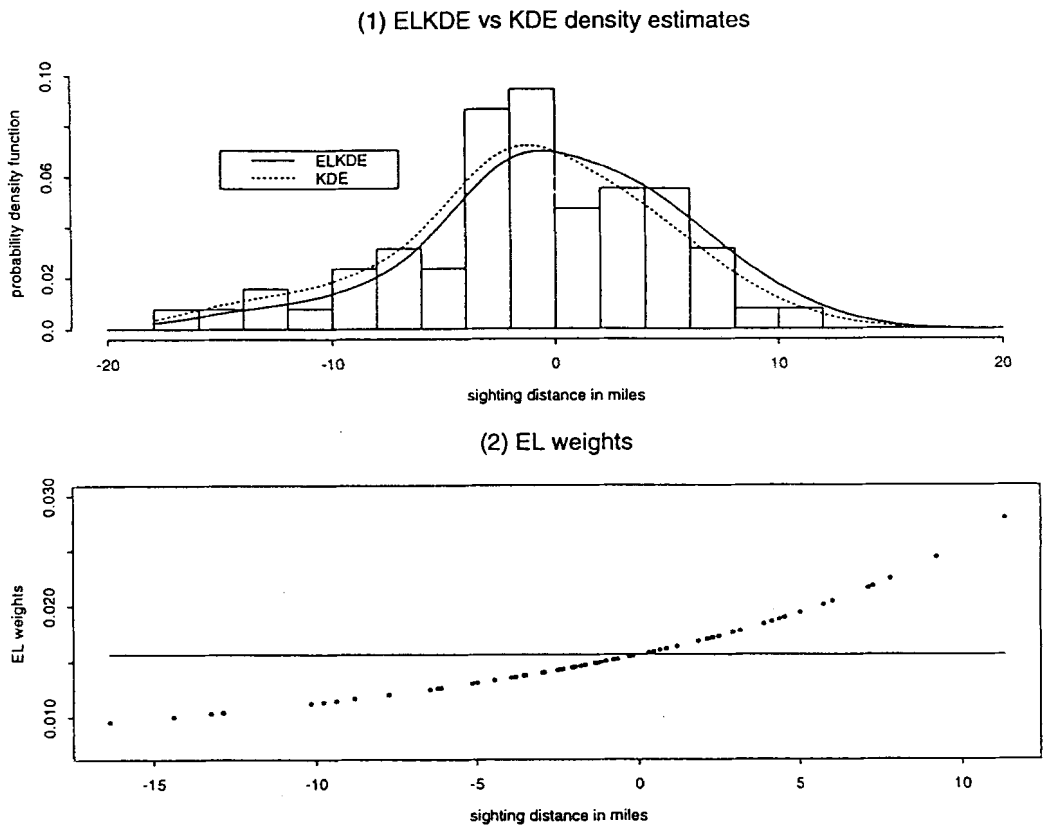

Fig. 1. — Empirical likelihood-based kernel density estimate and the kernel density estimate for the tuna dataset

5. Some Empirical Results

We have conducted density estimation using both the kernel and empirical likelihood kernel density estimators for two datasets designed to examine the performance of the empirical likelihood method. The first dataset is from the tuna aerial survey; the second is simulated from the standard normal distribution.

A line transect survey (Buckland *et al.*, 1993) was used to estimate tuna abundance in the Great Australian Bight in summer when the tuna tend to stay on the surface. One measure of tuna abundance is $D = N/A$ where N is the total number of surface schools in the Bight and A is the total survey area. To estimate D, a light aircraft with two tuna spotters on board flies along randomly allocated transect lines to detect tuna schools. Each school sighted is counted and its perpendicular distance to the transect is measured by a satellite-based Global Positioning System (GPS). Suppose n independent schools are detected after flying a distance L with $X_1, \ldots, X_n$ being the perpendicular sighting distances; X_i is negative/positive if the ith detected school is on the left/right of the transect line. Let f denote the p.d.f. of the sighting distances. Standard line transect theory shows that $D = L^{-1}\mathrm{E}(n)f(0)$. Therefore, density

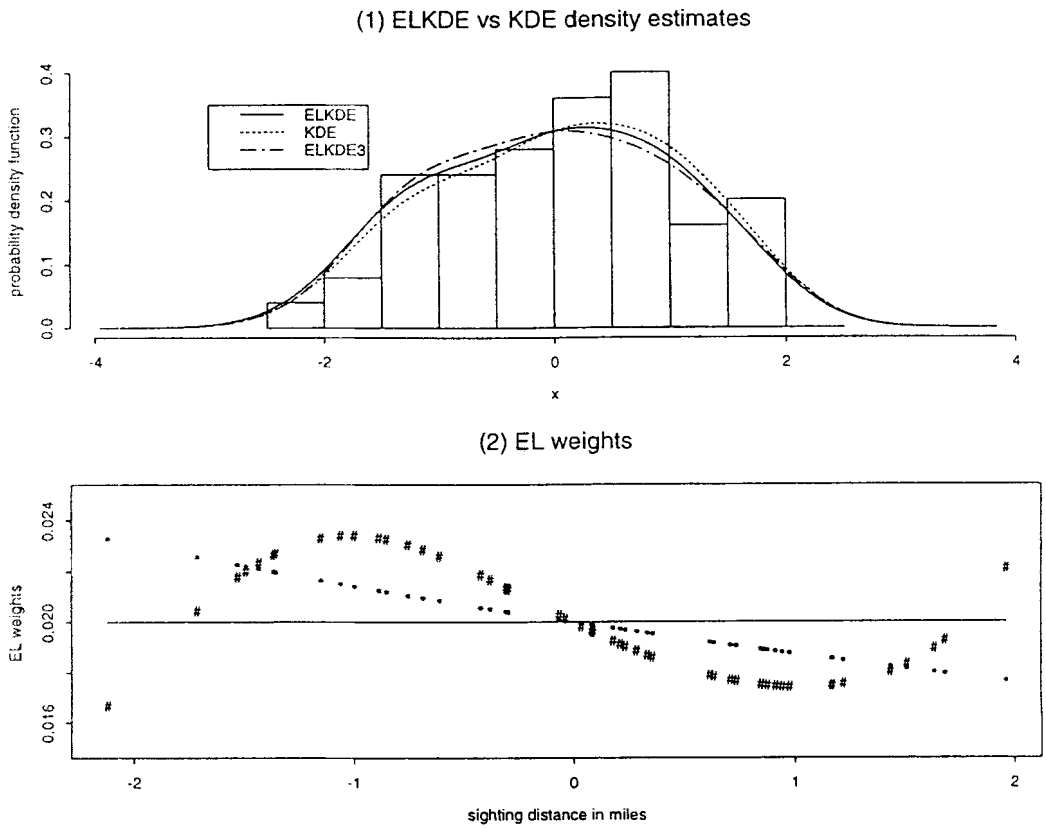

Fig. 2. — Empirical likelihood-based kernel density estimates and the kernel density estimate for a simulated $N(0,1)$ dataset

estimation plays a crucial role in a line transect survey.

The tuna dataset contains 64 perpendicular sighting distances obtained from the second replicate of the 1993 survey. As the sample mean is -1.21 and its standard error is 0.712, a zero mean hypothesis is consistent with the data. The sample skewness coefficient is -1.491 with standard error 0.26 calculated by the bootstrap. This indicates a certain discrepancy between the detection patterns of the two spotters. Therefore, only the zero mean constraint has been used to choose the weights.

Figure 1 shows the empirical likelihood-based kernel estimate $\hat{f}_{el}(x)$ and the kernel estimate $\hat{f}(x)$, together with a histogram of the tuna data and a plot of the empirical likelihood weights p_i. The empirical likelihood weights p_i used by $\hat{f}_{el}(x)$ are determined by $p_i = n^{-1}\{1 + \lambda X_i\}$ after obtaining a value (-0.039) for the λ. In the construction of both estimates, we choose the smoothing bandwidth $h = \hat{\sigma} n^{-1/5}$ with the sample standard deviation $\hat{\sigma} = 5.697$.

There is a clear shift in the two density estimates. The kernel estimate has a mode near the sample mean (-1.21). Its entire body is shifted towards the right by the empirical likelihood to such an extent that the empirical likelihood-

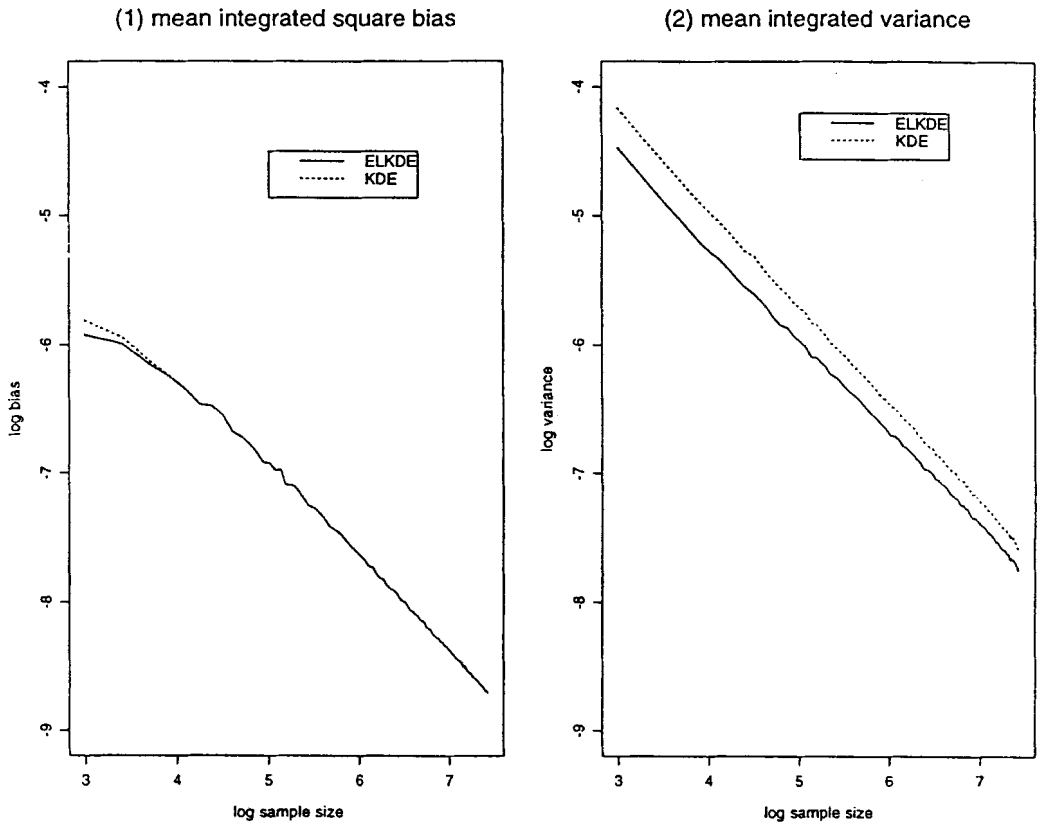

Fig. 3. — Mean integrated square bias and variance of the empirical likelihood-based kernel density estimator and the kernel density estimator

based density curve is centered at zero. The density curve has a rightwards shift because of the increasing empirical likelihood weights p_i as X_i increases. Note also that the p_i are quite different from $n^{-1} = 0.016$; this leads to the significant difference in the two density estimates.

The second dataset contains 50 standard normal random variables generated by using the routine 'gasdev' in Press *et al.* (1992) with a seed value -103. The sample has a mean value 0.07 and the standard deviation 1.022. The skewness coefficient is -0.1004 with a standard error 0.199 obtained by the bootstrap.

Figure 2 presents two empirical likelihood-based density estimates for the second dataset together with the kernel estimate. One (ELKDE) is based on the zero mean constraint only; the other (ELKDE3) is constructed by assuming both the zero mean and the zero third moment. We observe that the original kernel estimate has a mode near $x = 0.4$. By taking the zero mean constraint, the empirical likelihood density estimate shifts the mode to near $x = 0.27$. By adding the zero third moment constraint, we see that the body of the density is further shifted to the left (the mode is now around $x = 0.1$), and that the empirical likelihood weights change from a nearly linear pattern to a pattern like

that of the reciprocal of a cubic polynomial.

We used a simulation study to evaluate the mean integrated square bias and the mean integrated variance of the kernel and empirical likelihood-based kernel density estimates, from 2000 $N(0,1)$ random samples of sizes ranging from 10 to 1000, generated using the routine in Press *et al.* (1992). The empirical likelihood-based kernel estimates are all based on a single zero mean constraint.

Figure 3 presents the mean integrated square bias and the mean integrated variance for both the kernel and empirical likelihood kernel density estimates on a natural logarithm scale. We see that there is not much difference in the mean integrated square bias between the two estimates, as predicted by the theory in (12) which says that the difference is only $o(n^{-1})$. There is a substantial reduction in the mean integrated variance by the empirical likelihood estimate for all the sample sizes considered, as indicated in (13). The exact amount of reduction in the variance depends on the coefficient of the n^{-1} term in (13). Only the zero mean constraint has been used and the data are from the standard normal distribution, so $q = 1$, $g(x) = x$ and $\Sigma = 1$, and because we calculated the mean integrated variance, the coefficient is

$$\int g(x)^T \Sigma^{-1} g(x) f^2(x) dx = \int x^2 f^2(x) dx = \frac{1}{4\sqrt{\pi}} \approx 0.141.$$

For a sample size of 1000 the second order term should be 0.000141, which is about the level of variance reduction shown in the simulation results. Note that the natural logarithm scale was used in the plots. We conclude that our theoretical findings in Section 4 are confirmed by the simulation study.

References

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

CHEN, J. & QIN, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika* **80**, 107–116.

CHEN, S.X. (1994). Comparing empirical likelihood and bootstrap hypothesis tests. *J. Multivariate Anal.* **51**, 277–293.

— (1996a). A kernel estimate for the density of a biological population by using line transect sampling. *Appl. Statist.* **44**, 135–150.

— (1996b). Empirical likelihood confidence intervals for nonparametric density estimation. *Biometrika* **83**, 329–341.

DICICCIO, T.J., HALL, P. & ROMANO, J.P. (1988). Bartlett adjustment for empirical likelihood. Research Report No. 298. Department of Statistics, Stanford University.

OWEN, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249.

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

— (1991). Empirical likelihood for linear models. *Ann. Statist.* **19**, 1725–1747.

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

ZHANG, B. (1995). M-estimation and quantile estimation in the presence of auxiliary information. *J. Statist. Plann. Inference* **44**, 77–94.