AISM (2006) 58: 327–340
DOI 10.1007/s10463-005-0014-8

Young Kyung Lee · Byeong U. Park

Estimation of Kullback–Leibler divergence by local likelihood

Received: 7 June 2004 / Revised: 3 February 2005 / Published online: 3 June 2006

Abstract Motivated from the bandwidth selection problem in local likelihood density estimation and from the problem of assessing a final model chosen by a certain model selection procedure, we consider estimation of the Kullback–Leibler divergence. It is known that the best bandwidth choice for the local likelihood density estimator depends on the distance between the true density and the 'vehicle' parametric model. Also, the Kullback–Leibler divergence may be a useful measure based on which one judges how far the true density is away from a parametric family. We propose two estimators of the Kullback-Leibler divergence. We derive their asymptotic distributions and compare finite sample properties.

Keywords Kernel smoothing · Local likelihood density estimation · Bandwidth · Kullback–Leibler divergence

1 Introduction

Local likelihood methods hold considerable promise in density estimation. They offer unmatched flexibility and adaptivity as the resulting density estimators inherit both of the best properties of nonparametric approaches and parametric inference. They operate with a locally weighted log-likelihood where the local weights are determined by a kernel function and a bandwidth and are applied to the likelihood of the selected parametric model $\mathcal{F}$. When the true density is far from the parametric model, they have the best properties of nonparametric approach if the bandwidth is taken small. On the other hand, when the true density is close to the parametric

Research of Young Kyung Lee was supported by the Brain Korea 21 Projects in 2004. Byeong U. Park's research was supported by KOSEF through Statistical Research Center for Complex Systems at Seoul National University.

Y.K. Lee · B.U. Park (✉)
Department of Statistics, Seoul National University, Seoul 151-747, South Korea
E-mail: bupark@stats.snu.ac.kr

model, they give the best performance of parametric inference, too, if the bandwidth is chosen large. The latter property is not possessed by the conventional smoothing techniques such as the ordinary kernel density estimation. See, among others, Eguchi and Copas (1998), Hjort and Jones (1996), Park et al. (2002) and Park et al. (2006) for more details. Thus, estimating a measure of distance between the true density and the selected parametric model provides an important clew to determine a proper size of the bandwidth so that these potential advantages of local likelihood methods can be realized.

In this paper, we consider estimation of the Kullback–Leibler (KL) divergence between the true density and a selected parametric model. The KL divergence (Kullback and Leibler 1951) has been widely studied in statistical literature as a central index measuring qualitative similarity between two probability distributions. It is closely related to the notion of entropy in the context of statistical physics. For given two probability distributions with density functions g and f, the KL divergence of f from g is defined by

$$D(g, f) = \int g(x) \log \frac{g(x)}{f(x)} \mathrm{d}x. \tag{1}$$

Let g denote the true density function and $\mathcal{F} = \{f(\cdot, \theta) : \theta \in \Omega\}$ be a selected statistical model for the data distribution g, where Ω is a subset of $\mathbb{R}^p$. When g actually belongs to $\mathcal{F}$, the minimal value, $\min_{\theta \in \Omega} D(g, f(\cdot, \theta))$, of the KL divergence is zero. On the other hand, if g is separate from $\mathcal{F}$ with detectable or undetectable degree of model misspecification, the minimal KL divergence is strictly positive. In the model selection context, only the first term on the right hand side of

$$D(g, f(\cdot, \theta)) = -\int \{\log f(x, \theta)\} g(x) \, \mathrm{d}x + \int \{\log g(x)\} g(x) \, \mathrm{d}x$$

is relevant since the second term does not depend on the model. See Akaike (1973, 1974) and Konishi and Kitagawa (1996) for a detailed account of the problem of estimating $\int \{\log f(x, \widehat{\theta})\} g(x) \, \mathrm{d}x$ in the context of model selection, where $\widehat{\theta}$ is an estimator of θ based on the model $\mathcal{F}$. However, the neglected term $\int \{\log g(x)\} g(x) \, \mathrm{d}x$ is an important counterpart of the KL divergence. When a statistical model $\mathcal{F}$ is chosen by some model selection criterion, it would be of much interest to estimate the minimal value of $D(g, f(\cdot, \theta))$ over $f(\cdot, \theta) \in \mathcal{F}$ as it provides a useful tool for goodness-of-fit tests for the chosen model $\mathcal{F}$.

We propose two estimators of the minimal KL divergence. They are based on the local likelihood method in density estimation. We derive their asymptotic distributions. These asymptotic results are presented in the usual smoothing context of the bandwidth tending to zero as the sample size tends to infinity. Along with the theoretical properties of the estimators, we present some numerical results which compare the finite sample performance of the two estimators. We note that working on large bandwidth asymptotics is not relevant in the context of estimating the minimal KL divergence. When the bandwidth tends to infinity, the local likelihood density estimator converges to a member of the parametric family which minimises $D(g, f(\cdot, \theta))$. Thus, the resulting estimator of the minimal KL divergence, which is obtained by substituting the density estimator for the true g in $\min_{\theta \in \Omega} D(g, f(\cdot, \theta))$, always converges to zero.

This paper is organised as follows. In Sect. 2, the two estimators of the minimal KL divergence are introduced and their theoretical properties are presented. In Sect. 3, the two estimators are compared through a simulation study. Technical proofs are deferred to the Appendix.

2 Estimation of Kullback–Leibler divergence

Let $D(g, f)$ denote the KL divergence between the two density functions g and f, defined at Eq. (1). Let $\mathcal{F} = \{f(\cdot, \theta) : \theta \in \Omega\}$ with $\Omega \subset \mathbb{R}^p$ be a selected parametric model, where

$$f(z, \theta) = \mu(z) \exp\{t(z)^{\mathrm{T}}\theta - \psi(\theta)\} \tag{2}$$

for some fixed function $\mu(z) \geq 0$. We choose to work with the exponential family Eq. (2) as the parametric model for simplicity of the presentation. Our methods and theory are still valid for more general parametric families. Write $f(\cdot, \theta^G)$ for the best parametric approximation to the true density in the sense of minimizing the KL divergence $D(g, f(\cdot, \theta))$. We note that θ^G satisfies

$$\int \{t(z) - \psi'(\theta)\} g(z) \mathrm{d}z = 0. \tag{3}$$

In this section, we introduce two estimators of the minimal KL divergence $D\left(g, f(\cdot, \theta^G)\right)$, and provide their asymptotic distributions.

2.1 Definition of estimators

The minimal KL divergence involves the true density g and the parameter value θ^G, which we need to estimate. For estimation of g, we consider the general class of local likelihood introduced by Eguchi and Copas (1998). Let $u(z, \theta) = (\partial/\partial\theta) \log f(z, \theta)$. Define

$$\begin{aligned} \ell_n(x, \theta) = & \frac{1}{n} \sum_{i=1}^{n} u(X_i, \theta) K\left(\frac{X_i - x}{h}\right) \\ & - \frac{1}{n} \sum_{i=1}^{n} \xi\left[K\left(\frac{X_i - x}{h}\right), E_\theta K\left(\frac{X_i - x}{h}\right)\right] \\ & \times E_\theta u(X_1, \theta) K\left(\frac{X_1 - x}{h}\right). \end{aligned} \tag{4}$$

Here and below, E_θ denotes the expectation with respect to $f(\cdot, \theta)$. Also, K is the kernel function which is usually a probability density function, and h is a positive constant called the bandwidth. For the function $\xi(\cdot, \cdot)$, we follow Park et al. (2002) by assuming that $\xi(u, v) = \beta(v) + \gamma(v)(u/v)$ for sufficiently smooth β and γ near $v = 0$, and requiring that $\gamma(0) \neq 1$. We also assume

$$E_\theta\left\{\xi\left[K\left(\frac{X_1 - x}{h}\right), E_\theta K\left(\frac{X_1 - x}{h}\right)\right]\right\} = 1$$

for consistency of the local likelihood procedure when the true density g actually belongs to the parametric family $\mathcal{F}$. Examples of ξ that satisfy these conditions include $\xi(u, v) \equiv 1$ of Hjort and Jones (1996) and $\xi(u, v) = (1-u)/(1-v)$ of Copas (1995).

Incorporating the specific form of the working parametric model given at Eq. (2), the local likelihood equation at (4) is given by

$$\begin{aligned}\ell_n(x,\theta) = &\frac{1}{n}\sum_{i=1}^{n}\{t(X_i)-\psi'(\theta)\}K\left(\frac{X_i-x}{h}\right)\\ &-\frac{1}{n}\sum_{i=1}^{n}\xi\left[K\left(\frac{X_i-x}{h}\right), E_\theta K\left(\frac{X_i-x}{h}\right)\right]\\ &\times E_\theta\left\{t(X_1)-\psi'(\theta)\right\}K\left(\frac{X_1-x}{h}\right). \end{aligned} \tag{5}$$

Let $\widehat{\theta}_n^L(x)$ denote a solution of the equation $\ell_n(x,\theta) = 0$. The local likelihood estimator of g is then given by

$$\widehat{g}_n(x) = f(x, \widehat{\theta}_n^L(x)).$$

The estimator $\widehat{\theta}_n^L(x)$ aims at the solution of the equation $E\ell_n(x,\theta) = 0$, which we denote by $\theta_h^L(x)$. The function $f(\cdot, \theta_h^L(x))$ is the best 'local' approximation, near x, among the working parametric family $\mathcal{F}$ to the true density g. The bandwidth h determines the degree of the local approximation. If h tends to infinity, the function $\ell_n(x,\theta)$ converges to $n^{-1}\sum_{i=1}^n\{t(X_i)-\psi'(\theta)\}$ so that $\theta_h^L(x)$ goes to the 'global' approximant θ^G regardless of x. On the other hand, if h tends to zero, it converges to the solution of the equation $\lim_{h\to 0} E\ell_n(x,\theta) = 0$, which is the value of θ such that $g(x) = f(x,\theta)$. The latter property follows from the fact that

$$\lim_{h\to 0} Eh^{-1}\ell_n(x,\theta) = \{t(x)-\psi'(\theta)\}\{1-\gamma(0)\}\{g(x)-f(x,\theta)\}.$$

Next, let $\widehat{\theta}_n^G$ denote the parametric maximum likelihood estimator of θ^G. It is given by the maximizer of the full likelihood

$$\mathcal{L}_n(\theta) = \sum_{i=1}^{n}\log f(X_i,\theta)$$

based on $\mathcal{F}$. It satisfies the full likelihood equation

$$\psi'(\theta) = \frac{1}{n}\sum_{i=1}^{n} t(X_i).$$

The best local approximation $\theta_h^L(x)$ depends on x, so does its estimator $\widehat{\theta}_n^L(x)$, while the 'global' or parametric best approximation θ^G and its estimator $\widehat{\theta}_n^G$ do not depend on x.

Now, we define the two estimators of $D\left(g, f(\cdot, \theta^G)\right)$. The first estimator is defined by $\widehat{D}_{1n} = D\left(\widehat{g}_n, f(\cdot, \widehat{\theta}_n^G)\right)$, i.e.

$$\begin{aligned}\widehat{D}_{1n} &= -\int \log\left\{\frac{f(x, \widehat{\theta}_n^G)}{\widehat{g}_n(x)}\right\}\widehat{g}_n(x)\,\mathrm{d}x \\ &= \int \left[t(x)^T\{\widehat{\theta}_n^L(x) - \widehat{\theta}_n^G\} - \{\psi(\widehat{\theta}_n^L(x)) - \psi(\widehat{\theta}_n^G)\}\right]\widehat{g}_n(x)\,\mathrm{d}x. \qquad (6)\end{aligned}$$

For the definition of our second estimator, let $\widehat{\theta}_{n,-i}^L(x)$ be the leave-one-out (or cross-validatory) version of $\widehat{\theta}_n^L(x)$. It is the solution of the estimating equation $\ell_{n,-i}(x, \theta) = 0$, where $\ell_{n,-i}$ is the leave-one-out version of ℓ_n defined at Eq. (5) with the i-th observation deleted. Define $\widehat{g}_{n,-i}(x) = f(x, \widehat{\theta}_{n,-i}^L(x))$. Now, let $\widehat{\theta}_{n,-i}^G$ denote the leave-one-out version of the parametric maximum likelihood estimator $\widehat{\theta}_n^G$. It satisfies the equation $\psi'(\theta) = (n-1)^{-1}\sum_{j\neq i} t(X_j)$. The second estimator is defined by

$$\widehat{D}_{2n} = -\frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{f(X_i, \widehat{\theta}_{n,-i}^G)}{\widehat{g}_{n,-i}(X_i)}\right\}. \qquad (7)$$

It is obtained by plugging the leave-one-out estimators $\widehat{g}_{n,-i}$ and $\widehat{\theta}_{n,-i}^G$ into the empirical version $-n^{-1}\sum_{i=1}^n \log\{f(X_i, \theta^G)/g(X_i)\}$ of the minimal KL divergence.

Remark 2.1 In the definition of the second estimator, one may use $\widehat{\theta}_n^G$ and $\widehat{g}_n$ instead of the leave-one-out estimators $\widehat{\theta}_{n,-i}^G$ and $\widehat{g}_{n,-i}$, respectively. This yields $\widetilde{D}_{2n} = -n^{-1}\sum_{i=1}^n \log\left\{f(X_i, \widehat{\theta}_n^G)/\widehat{g}_n(X_i)\right\}$. However, the latter estimator would produce some 'overfitting' bias since the empirical distribution corresponds more closely to $\widehat{\theta}_n^G$ and $\widehat{g}_n$ than does the true distribution g. In fact, Akaike's AIC is a clever device to correct this overfitting bias of $\widehat{\theta}_n^G$ for estimating $E_g \log f(X, \theta^G)$. Also, it is well known that the leave-one-out version $n^{-1}\sum_{i=1}^n \log f(X_i, \widehat{\theta}_{n,-i}^G)$ is asymptotically equivalent to the AIC (Stone 1977). See Konishi and Kitagawa (1996) for more details. Jackknifing is an alternative procedure for correcting the overfitting bias of the estimator $\widetilde{D}_{2n}$. A jackknife estimator of the bias is given by $(n-1)(\widetilde{D}_{2n(\cdot)} - \widetilde{D}_{2n})$ where $\widetilde{D}_{2n(\cdot)} = -\{n(n-1)\}^{-1}\sum\sum_{i\neq j}\log\left\{f(X_j, \widehat{\theta}_{n,-i}^G)/\widehat{g}_{n,-i}(X_j)\right\}$ and the resulting bias-corrected estimator equals $n\widetilde{D}_{2n} - (n-1)\widetilde{D}_{2n(\cdot)}$. The jackknife estimator is more complicated than our leave-one-out cross-validatory estimator $\widehat{D}_{2n}$. See Efron (1982) or Ripley (1996) for a general comparison between the cross-validation and jackknifing.

2.2 Theoretical properties

We provide some theoretical properties of the estimators defined at Eqs. (6) and (7), in the usual smoothing context of the bandwidth, h, tending to zero as the sample size tends to infinity. To describe their asymptotic properties, let N be the $p \times p$ matrix which has $\int y^{i+j}K(y)\,\mathrm{d}y$ as its (i, j)-th component $(i, j = 0, \ldots, p-1)$,

and write η for the p-dimensional vector with its i-th element being $\int y^{p+i} K(y)\,\mathrm{d}y$. Let e_0 be the p-dimensional unit vector $(1, 0, \ldots, 0)$. Define

$$c_1 = e_0^{\mathrm{T}} N^{-1} \eta \int \mathcal{B}_p(x)\{1 - \log[f(x, \theta^G)/g(x)]\}\,\mathrm{d}x,$$
$$\mathcal{B}_p(x) = (p!)^{-1}\left[(\partial^p/\partial y^p)\{g(y) - f(y, \theta_0^L(x))\}\right]_{y=x},$$

where $\theta_0^L(x)$ denotes the solution of the equation $\lim_{h\to 0} E\ell_n(x, \theta) = 0$.

For the first estimator $\widehat{D}_{1n}$, it may be proved under the assumptions stated in the Appendix that as $n \to \infty$

$$\widehat{D}_{1n} = -\int \log\left\{\frac{f(x, \theta^G)}{g(x)}\right\} \widehat{g}_n(x)\,\mathrm{d}x + h^p\, e_0^{\mathrm{T}} N^{-1} \eta \int \mathcal{B}_p(x)\,\mathrm{d}x + o_p(h^p + n^{-1/2}). \tag{8}$$

Also, it can be shown under the assumptions stated in the Appendix that as $n \to \infty$

$$-\int \log\left\{\frac{f(x, \theta^G)}{g(x)}\right\} \widehat{g}_n(x)\,\mathrm{d}x = D - h^p\, e_0^{\mathrm{T}} N^{-1} \eta \int \mathcal{B}_p(x) \log\left\{\frac{f(x, \theta^G)}{g(x)}\right\} \mathrm{d}x + n^{-1/2} Z_n + o_p(h^p + n^{-1/2}), \tag{9}$$

where $D \equiv D\Big(g,\, f(\cdot, \theta^G)\Big)$ and Z_n is asymptotically normal with mean zero and variance given by

$$\sigma_1^2 = e_0^{\mathrm{T}} N^{-1} e_0\, E \log^2\left\{\frac{f(X_1, \theta^G)}{g(X_1)}\right\} - \left\{E \log \frac{f(X_1, \theta^G)}{g(X_1)}\right\}^2. \tag{10}$$

Thus, we obtain the following theorem. Proofs of Eqs. (8), (9) and (10) will be given in the Appendix.

Theorem 2.1 *Suppose that $h \to 0$ and $nh^2/(\log n)^2 \to \infty$ as n tends to infinity. Then, under the assumptions stated in the Appendix, we have*

$$\sqrt{n}\left(\widehat{D}_{1n} - D - c_1 h^p + o_p(h^p)\right) \longrightarrow N\left(0, \sigma_1^2\right).$$

Next, for the second estimator $\widehat{D}_{2n}$, we find

$$\frac{1}{n}\sum_{i=1}^{n} \log\left\{\frac{f(X_i, \widehat{\theta}_{n,-i}^L(X_i))}{g(X_i)}\right\} = c_2 h^p + o_p(h^p + n^{-1/2}), \tag{11}$$

where $c_2 = e_0^{\mathrm{T}} N^{-1} \eta \int \mathcal{B}_p(x)\,\mathrm{d}x$. Also,

$$n^{-1}\sum_{i=1}^{n} \log f(X_i, \widehat{\theta}_{n,-i}^G) = n^{-1}\sum_{i=1}^{n} \log f(X_i, \theta^G) + O_p(n^{-1}). \tag{12}$$

A sketch of a proof for Eq. (11) will be also given in the Appendix. From Eqs. (11) and (12), we can deduce that $\widehat{D}_{2n}$ is also asymptotically normal with mean $c_2 h^p$ and variance now given by

$$\sigma_2^2 = \mathrm{var}\left[\log\left\{\frac{f(X_1, \theta^G)}{g(X_1)}\right\}\right].$$

Theorem 2.2 *Under the same assumptions of Theorem 2.1, we have*

$$\sqrt{n}\left(\widehat{D}_{2n} - D - c_2 h^p + o_p(h^p)\right) \longrightarrow N\left(0, \sigma_2^2\right).$$

The theorems imply that the first-order asymptotic properties of the two estimators do not depend on the function ξ, which appears in the definition of the local likelihood equation, except for the requirement that $\gamma(0) \neq 1$. This has been already seen by Park et al. (2002) in the context of density function estimation.

For the variance terms, it can be shown that $e_0^{\mathrm{T}} N^{-1} e_0 \geq 1$. In fact, if K is symmetric and $\int K = 1$, then $e_0^{\mathrm{T}} N^{-1} e_0 = 1$ for $p = 1$ and 2. For $p = 3$ and 4, it equals $\mu_4/(\mu_4 - \mu_2^2) = \left\{1 - (\mu_2^2/\mu_4)\right\}^{-1}$ which is greater than one since $\mu_4 > \mu_2^2$ by Liapounov's inequality. Here, μ_j denotes the j-th moment of the kernel K. In general for $p \geq 3$, $e_0^{\mathrm{T}} N^{-1} e_0 = \{1 - N_{12} N_{22} N_{21}\}^{-1}$ where $N_{21} = \left(\mu_1, \ldots, \mu_{p-1}\right)^{\mathrm{T}}$, $N_{12} = N_{21}^{\mathrm{T}}$, and N_{22} denotes the $(p-1) \times (p-1)$ matrix whose components are μ_{i+j} for $i, j = 1, \ldots, p-1$. Since N is positive definite and so is N^{-1}, $e_0^{\mathrm{T}} N^{-1} e_0 > 0$, i.e., $N_{12} N_{22} N_{21} < 1$. Furthermore, since N_{22} is also positive definite, $N_{12} N_{22} N_{21} > 0$. This establishes $e_0^{\mathrm{T}} N^{-1} e_0 > 1$. Thus, we observe $\sigma_1^2 \geq \sigma_2^2$ for all p. The bias terms c_1 and c_2 for the two estimators are not comparable in general. But, c_2 has a simpler formula than c_1, and our simulation study in the next section suggests that c_2, the bias factor for the second estimator, is less than c_1 in the simulation settings.

The above theorems can be used to define confidence intervals for D. Estimators of the bias factors c_i could be built in by using estimators of the derivatives of $g - f(\cdot, \theta_0^L(x))$ at every x. But this would considerably complicate the procedure. The construction is greatly simplified if one does 'under-smoothing'. Suppose that h tends to zero faster than $n^{-1/(2p)}$. Then, the bias terms are negligible compared to the variance. In this case, asymptotic confidence intervals for D can be constructed by estimating σ_i^2 only. For example, σ_2^2 may be estimated by

$$\widehat{\sigma}_2^2 = n^{-1} \sum_{i=1}^{n} \left[\log\left\{\frac{f(X_i, \widehat{\theta}_{n,-i}^G)}{\widehat{g}_{n,-i}(X_i)}\right\} - (-\widehat{D}_{2n})\right]^2.$$

Confidence intervals for D may be used for goodness-of-fit tests for the parametric model $\mathcal{F}$. Note that $D = 0$ when $g \in \mathcal{F}$. If a confidence interval for D contains zero, one may accept the hypothesis that the actual distribution belongs to the chosen model $\mathcal{F}$.

Remark 2.2 When the true density g belongs to $\mathcal{F}$, both σ_1^2 and σ_2^2 are zero, and so are the bias factors c_1 and c_2. In view of Theorems 2.1 and 2.2, this means $\widehat{D}_{1n}$ and $\widehat{D}_{2n}$ converge to D at a rate faster than $n^{-1/2}$. A higher order asymptotic analysis is required to obtain non-degenerate limit distributions in this case. Derivation of the limit distribution with the exact rate of convergence when $g \in \mathcal{F}$ is an interesting future research problem.

3 Numerical properties

We compare the small sample performance of the two estimators $\widehat{D}_{1n}$ and $\widehat{D}_{2n}$. We considered $N(\theta, 1)$ as the parametric model $f(\cdot, \theta)$. The true density was taken to be

$$g(x) \equiv g_\beta(x) = 2\phi(x)\Phi(\beta x),$$

where ϕ and Φ are the standard normal density and its distribution function. This is the so-called skewed normal distribution of Azzalini (1985), and was also considered by Eguchi and Copas (1998). Here, β acts as a discrepancy parameter. When $\beta = 0$, the density g is identical to ϕ. As $|\beta|$ increases, it becomes increasingly skewed. In this setting, we find

$$\theta^G = EX = \sqrt{\frac{2}{\pi}} \frac{\beta}{\sqrt{1+\beta^2}}.$$

The minimal KL divergence $D(g, f(\cdot, \theta^G))$ is a symmetric function of β.

We generated 500 pseudo samples of size $n = 400$ from g_β for $\beta = 0$ and 1. The standard normal density was taken for the kernel function K. We took $\xi(\cdot, \cdot) \equiv 1$, which corresponds to the U-version of Hjort and Jones (1996). Table 1 shows the squared biases and variances of the two estimators when the optimal bandwidths were used. For each β and for each estimator, the optimal bandwidth was obtained to minimize the Monte Carlo approximation of the mean squared error which is the average of 500 values of the squared error.

Figure 1 compares the quantile plots of the squared errors of $\widehat{D}_{1n}$ and $\widehat{D}_{2n}$ which employ the optimal bandwidths. For this, we computed 500 values of $\widehat{D}_{1n}$ and $\widehat{D}_{2n}$, and calculated 500 values of the squared errors $d_1 = (\widehat{D}_{1n} - D)^2$ and $d_2 = (\widehat{D}_{2n} - D)^2$. Then, we arranged them in increasing order. Write $d_1^{(1)} \le d_1^{(2)} \le \cdots \le d_1^{(500)}$ and $d_2^{(1)} \le d_2^{(2)} \le \cdots \le d_2^{(500)}$, respectively, for the arranged squared deviations. Figure 1 shows the quantile plots $\{(i/500, d_1^{(i)})\}_{i=1}^{500}$ and $\{(i/500, d_2^{(i)})\}_{i=1}^{500}$ for each β. Figure 2 depicts the densities of the centered estimators $\widehat{D}_{1n} - D$ and $\widehat{D}_{2n} - D$. From the table and figures, we find that $\widehat{D}_{2n}$ performs always better than $\widehat{D}_{1n}$.

Table 1 *Squared biases and variances of the estimators, multiplied by* 10^4

β	$\widehat{D}_{1n}$		$\widehat{D}_{2n}$	
	Sq. bias	Variance	Sq. bias	Variance
0	5.953	4.982	0.810	5.480
1	13.104	6.212	0.462	3.689

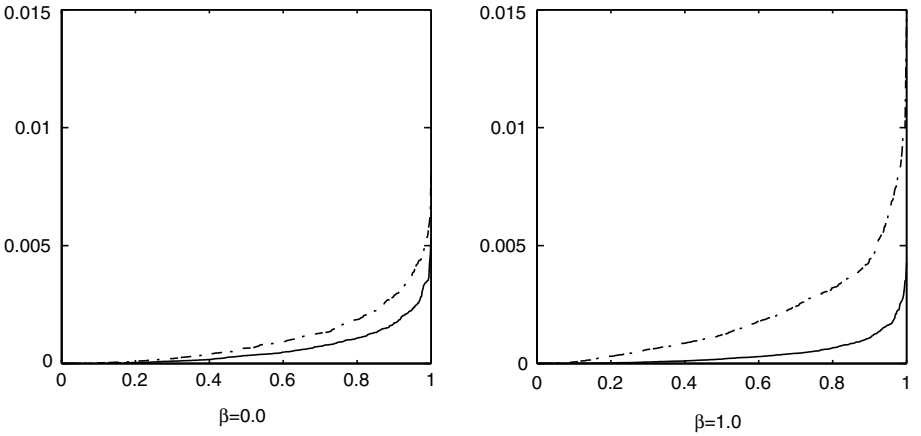

Fig. 1 Quantile functions of the squared errors of the estimators. *Dot-dashed* curves correspond to $\widehat{D}_{1n}$, and *solid curves* are for $\widehat{D}_{2n}$

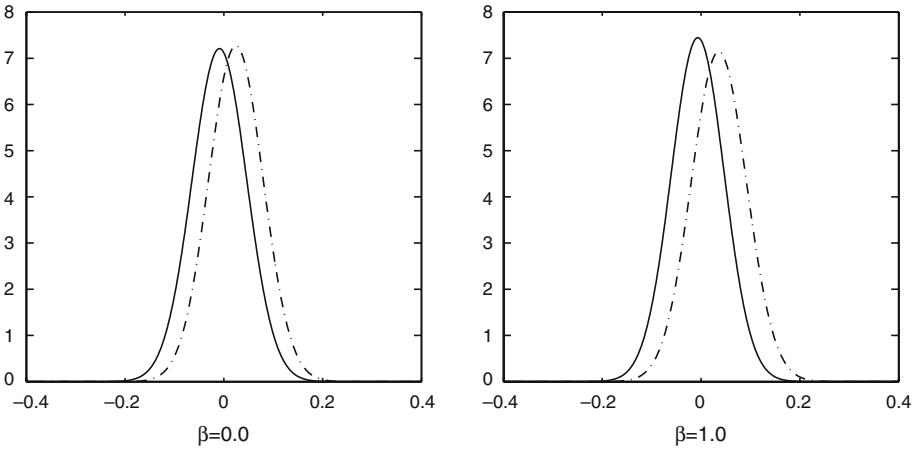

Fig. 2 Density plots of the estimators recentered at the true value D. Line types are the same as in Fig. 1

4 Appendix

A.1 Assumptions

Define $u_0(x, y) = \lim_{h\to 0} u\left(x, \theta_h^L(y)\right)$ and

$$U(x, y) = \left(u_0(x, y), \frac{\partial}{\partial x}u_0(x, y), \ldots, \frac{1}{(p-1)!}\frac{\partial^{p-1}}{\partial x^{p-1}}u_0(x, y)\right).$$

We make use of the following assumptions in the proofs of the main results in this paper.

(C1) The functions $\mu(\cdot)$ and $g(\cdot)$ are supported on a compact set $\mathcal{X}$;

(C2) Ω is a compact set;
(C3) The equations $E\ell_n(x,\theta) = 0$ and $\lim_{h\to 0} E\ell_n(x,\theta) = 0$ have the unique solutions $\theta_h^L(x)$ and $\theta_0^L(x)$, respectively;
(C4) $E\ell_n(x,\theta)$ converges to $\lim_{h\to 0} E\ell_n(x,\theta)$ uniformly on $\mathcal{X}\times\Omega$, and $\lim_{h\to 0} E\ell_n(x,\theta)$ is continuous on $\mathcal{X}\times\mathcal{H}$;
(C5) ψ is three times continuously partially differentiable on Ω;
(C6) The densities g and $f(\cdot,\theta)$ have p continuous derivatives for all $\theta\in\Omega$;
(C7) $U(x,y)$ is invertible for all $x\in\mathcal{X}$ and all y in a neighborhood of x;
(C8) β and γ have two continuous derivatives at zero;
(C9) The kernel K is twice continuously differentiable, nonnegative, bounded and supported on a compact set with non-empty interior;
(C10) The bandwidth h tends to zero as the sample size n goes to infinity.

A.2 Proof of (8)

We decompose $\widehat{D}_{1n}$ by

$$\widehat{D}_{1n} = \int \log\left\{\frac{g(x)}{f(x,\theta^G)}\right\}\widehat{g}_n(x)\,\mathrm{d}x + \int \log\left\{\frac{\widehat{g}_n(x)}{g(x)}\right\}\widehat{g}_n(x)\,\mathrm{d}x - \int \log\left\{\frac{f(x,\widehat{\theta}_n^G)}{f(x,\theta^G)}\right\}\widehat{g}_n(x)\,\mathrm{d}x. \tag{13}$$

By (C3) the second term in the decomposition Eq. (13) can be written as

$$\int \{t(x)-\psi'(\theta_h^L(x))\}^{\mathrm{T}}\{\widehat{\theta}_n^L(x)-\theta_h^L(x)\}\widehat{g}_n(x)\,\mathrm{d}x + \int \{t(x)-\psi'(\theta_0^L(x))\}^{T}\{\theta_h^L(x)-\theta_0^L(x)\}\widehat{g}_n(x)\,\mathrm{d}x + R_{n1}, \tag{14}$$

where R_{n1} has a faster order of convergence than the two preceding integrals. Call the two integrals at Eq. (14), J_1 and J_2. Write $I_n(x,\theta) = -h^{-1}E\{(\partial/\partial\theta)\ell_n(x,\theta)\}$, and let

$$S_n(x) = \dot{f}(x,\theta_h^L(x))^{\mathrm{T}} I_n(x,\theta_h^L(x))^{-1}h^{-1}\ell_n(x,\theta_h^L(x))$$

From (4.1) of Park, Kim and Jones (2002), it follows that

$$\widehat{g}_n(x) = f(x,\theta_h^L(x)) + S_n(x) + O_p\left(\frac{\log n}{nh}\right) \tag{15}$$

uniformly for $x\in\mathcal{X}$. This follows from the fact that

$$\widehat{\theta}_n^L(x) - \theta_h^L(x) = I_n(x,\theta_h^L(x))^{-1}h^{-1}\ell_n(x,\theta_h^L(x)) + O_p\left(\frac{\log n}{nh}\right) \tag{16}$$

uniformly for $x\in\mathcal{X}$. Furthermore, by a similar argument for the proof of Theorem 6 of Eguchi, Kim and Park (2003), we may deduce that uniformly for $x\in\mathcal{X}$

$$\left\{t(x)-\psi'(\theta_0^L(x))\right\}^{\mathrm{T}}\left\{\theta_h^L(x)-\theta_0^L(x)\right\} = \{g(x)\}^{-1}e_0^{\mathrm{T}}N^{-1}\eta\mathcal{B}_p(x)h^p + o(h^p). \tag{17}$$

Plugging Eq. (17) into the second integral at (14) yields

$$J_2 = h^p\, e_0^{\mathrm{T}} N^{-1} \eta \int \mathcal{B}_p(x)\,\mathrm{d}x + o_p(h^p).$$

We will prove $J_1 = o_p(n^{-1/2})$. Let $K_h(u) = h^{-1}K(h^{-1}x)$. Define

$$\begin{aligned}
&\phi_n(x, y) \\
&= \left\{t(x) - \psi'(\theta_h^L(x))\right\}^{\mathrm{T}} I_n(x, \theta_h^L(x))^{-1} \left[\{t(y) - \psi'(\theta_h^L(x))\}K_h(y-x)\right. \\
&\quad \left. - \xi\left(hK_h(y-x), hE_{\theta_h^L(x)}K_h(X_1-x)\right)E_{\theta_h^L(x)}\left\{t(X_1) - \psi'(\theta_h^L(x))\right\}K_h(X_1-x)\right],
\end{aligned}$$

and write $\widetilde{\phi}_n(y) = E\phi_n(X_1, y)$. Then, by using (16) we obtain

$$\begin{aligned}
J_1 = \frac{1}{n}\sum_{j=1}^{n}\widetilde{\phi}_n(X_j) + \frac{1}{n}\sum_{j=1}^{n}\int \phi_n(x, X_j)\{\widehat{g}_n(x) - f(x, \theta_h^L(x))\}\,\mathrm{d}x \\
+\frac{1}{n}\sum_{j=1}^{n}\int \phi_n(x, X_j)\{f(x, \theta_h^L(x)) - g(x)\}\,\mathrm{d}x + O_p\left(\frac{\log n}{nh}\right). \qquad (18)
\end{aligned}$$

Note that $E\{\widetilde{\phi}_n(X_1)\} = 0$ from the definition of $\theta_h^L(x)$, i.e. from the fact that $E\phi_n(x, X_1) = 0$ for all x. Also, it may be proved that $\mathrm{var}(\widetilde{\phi}_n(X_1)) = o(1)$. This follows from the fact that $\widetilde{\phi}_n(x) = o(1)$ for x in the interior of $\mathcal{X}$. Thus, the first term on the right hand side of Eq. (18) equals $o_p(n^{-1/2})$. Now, by Eq. (15) the second term in the expansion Eq. (18) can be written as

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int \phi_n(x, X_i)\phi_n(x, X_j) f(x, \theta_h^L(x))\,\mathrm{d}x + O_p\left(\frac{\log n}{nh}\right).$$

The above double summation equals $n^{-2}\sum_{i=1}^{n}\int \phi_n^2(x, X_i) f(x, \theta_h^L(x))\,\mathrm{d}x$, which has an order of magnitude $O_p(n^{-1}h^{-1})$, plus a degenerate U-statistic whose variance is of order $O(n^{-2}h^{-1})$. Thus, the second term in the expansion of J_1 has the rate $O_p\,(\log n/(nh))$. From the fact that $\theta_h^L(x) - \theta_0^L(x) = O(h^p)$ uniformly for $x \in \mathcal{X}$, it immediately follows that the third term in the expansion equals $O_p(n^{-1/2}h^p)$. This shows

$$\int \log\left\{\frac{\widehat{g}_n(x)}{g(x)}\right\}\widehat{g}_n(x)\,\mathrm{d}x = h^p\, e_0^{\mathrm{T}} N^{-1}\eta \int \mathcal{B}_p(x)\,\mathrm{d}x + o_p(h^p + n^{-1/2}). \qquad (19)$$

Next, we treat the third integral in the decomposition (13). It can be written as

$$(\widehat{\theta}_n^G - \theta^G)^{\mathrm{T}} \int \{t(x) - \psi'(\theta^G)\}\widehat{g}_n(x)\,\mathrm{d}x + R_{n2}, \qquad (20)$$

where R_{n2} is of lower order than the first term. By using Eq. (3) we may show that the integral at Eq. (20) has the order of magnitude $O_p\{n^{-1/2} + h^p + \log n/(nh)\}$. Since $\widehat{\theta}_n^G - \theta^G = O_p(n^{-1/2})$, we obtain

$$\int \log\left\{\frac{f(x, \widehat{\theta}_n^G)}{f(x, \theta^G)}\right\}\widehat{g}_n(x)\,\mathrm{d}x = O_p\left(\frac{1}{n} + \frac{h^p}{\sqrt{n}} + \frac{\log n}{n^{3/2}h}\right). \qquad (21)$$

Combining Eqs. (13), (19) and (21) completes the proof of (8).

A.3 Proofs of (9) and (10)

By Eqs. (15) and (17), it is enough to show that

$$n \operatorname{var}\left[\int \log\left\{\frac{g(x)}{f(x,\theta^G)}\right\} S_n(x)\,\mathrm{d}x\right] = n\int\int \log\left\{\frac{g(x)}{f(x,\theta^G)}\right\}\log\left\{\frac{g(y)}{f(y,\theta^G)}\right\} E\{S_n(x)S_n(y)\}\,\mathrm{d}x\,\mathrm{d}y \quad (22)$$

converges to σ_1^2. Write $p(v) = (1, v, \ldots, v^{p-1})^{\mathrm{T}}$. Define

$$\begin{aligned} L_1(x,y,h) &= L_h^{(2,2)}(x-y)D_hU(x,y)^{\mathrm{T}}g(x) \\ &\quad - \gamma(0)L_h^{(2,1)}(x-y)D_hU(y,y)^{\mathrm{T}}g(x) \\ &\quad - \gamma(0)L_h^{(1,2)}(x-y)D_hU(y,y)^{\mathrm{T}}g(y) \\ &\quad + \gamma(0)^2L_h^{(1,1)}(x-y)D_hU(y,y)^{\mathrm{T}}g(x), \\ L_2(x,y) &= \{1-\gamma(0)\}^2\Big(\int p(t)K(t)\,\mathrm{d}t\Big)\Big(\int p(t)^{\mathrm{T}}K(t)\,\mathrm{d}t\Big) \\ &\quad \times D_hU(y,y)^{\mathrm{T}}g(x)g(y), \end{aligned}$$

where $L_h^{(i,j)}(w) = h^{-1}L^{(i,j)}(h^{-1}w)$ for $i, j = 1, 2$ and

$$\begin{aligned} L^{(1,1)}(w) &= \Big(\int p(t)K(t)\,\mathrm{d}t\Big)\Big(\int p(t)^{\mathrm{T}}K(t)\,\mathrm{d}t\Big)\int K(t)K(t+w)\,\mathrm{d}t, \\ L^{(1,2)}(w) &= \Big(\int p(t)K(t)\,\mathrm{d}t\Big)\int p(t)^{\mathrm{T}}K(t)K(t-w)\,\mathrm{d}t, \\ L^{(2,1)}(w) &= \int p(t)K(t)K(t+w)\,\mathrm{d}t\Big(\int p(t)^{\mathrm{T}}K(t)\,\mathrm{d}t\Big), \\ L^{(2,2)}(w) &= \int p(t)p(t)^{\mathrm{T}}K(t)K(t+w)\,\mathrm{d}t. \end{aligned}$$

Let $\zeta_i = \int y^i K(y)\,\mathrm{d}y$. Write N_1 for the $p \times p$ matrix whose (i, j)-th entry equals $\zeta_{i+j} - \gamma(0)\zeta_i\zeta_j$. Also, write M_1 for the $p \times p$ matrix whose (i, j)-th entry equals $\zeta_{i+j} - 2\gamma(0)\zeta_i + \gamma(0)^2\zeta_i\zeta_j$. Let D_h denote the $p \times p$ diagonal matrix whose i-th diagonal element equals h^i $(i = 0, \ldots, p-1)$. Then, as in the proof for Theorem 2 of Park et al. (2002), it can be shown that

$$\begin{aligned} &n\,E\{S_n(x)S_n(y)\} \\ &\quad = e_0^{\mathrm{T}}N_1^{-1}\{L_1(x,y,h) - L_2(x,y)\}\{U(y,y)^{\mathrm{T}}\}^{-1}D_h^{-1}N_1^{-1}e_0\{1+O(h)\}. \end{aligned} \quad (23)$$

By plugging Eq. (23) into Eq. (22) and using the fact

$$\int \{L^{(2,2)}(w) - \gamma(0)L^{(2,1)}(w) - \gamma(0)L^{(1,2)}(w) + \gamma(0)^2L^{(1,1)}(w)\}\,\mathrm{d}w = M_1,$$

we obtain

$$n \operatorname{var}\left[\int \log\left\{\frac{g(x)}{f(x,\theta^G)}\right\} S_n(x)\,\mathrm{d}x\right] \longrightarrow e_0^{\mathrm{T}} N_1^{-1} M_1 N_1^{-1} e_0\, E \log^2\left\{\frac{g(X_1)}{f(X_1,\theta^G)}\right\} - \left[E \log\left\{\frac{g(X_1)}{f(X_1,\theta^G)}\right\}\right]^2.$$

As noted in Park et al. (2002), $e_0^{\mathrm{T}} N_1^{-1} M_1 N_1^{-1} e_0 = e_0^{\mathrm{T}} N^{-1} N N^{-1} e_0 = e_0^{\mathrm{T}} N^{-1} e_0$. This completes the proofs of (9) and (10).

A.4 Proof of (11)

By (C3) again, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \log\left\{\frac{f(X_i, \widehat{\theta}^L_{n,-i}(X_i))}{g(X_i)}\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n} \left\{t(X_i) - \psi'(\theta_h^L(X_i))\right\}^{\mathrm{T}} \left\{\widehat{\theta}^L_{n,-i}(X_i) - \theta_h^L(X_i)\right\} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \left\{t(X_i) - \psi'(\theta_0^L(X_i))\right\}^{T} \left\{\theta_h^L(X_i) - \theta_0^L(X_i)\right\} + R_{n3}, \qquad (24)
\end{aligned}$$

where R_{n3} has a faster order of convergence than the other terms in the expansion.

First, we treat the first term in the expansion. By Eq. (16), it equals

$$\frac{1}{n(n-1)} \sum\sum_{i\neq j} \phi_n(X_i, X_j) + O_p\left(\frac{\log n}{nh}\right). \qquad (25)$$

We find the projection of the U-statistic at Eq. (25) onto the space of sums of independent random variables, and then decompose it into two orthogonal terms. For this, define $\phi_n^*(x, y) = \phi_n(x, y) - \widetilde{\phi}_n(y)$. Then, we may write the U-statistic at Eq. (25) by

$$\frac{1}{n}\sum_{i=1}^{n} \widetilde{\phi}_n(X_i) + \frac{1}{n(n-1)} \sum\sum_{i\neq j} \phi_n^*(X_i, X_j). \qquad (26)$$

The first term of Eq. (26) equals $o_p(n^{-1/2})$ as is shown in the proof of (8). The second term of Eq. (26) has mean zero and the variance of order $O(n^{-2}h^{-1})$. This follows from the fact that

$$E\left\{\phi_n^*(X_1, X_2)|X_1\right\} = E\left\{\phi_n^*(X_1, X_2)|X_2\right\} = 0.$$

Thus, the first term in the expansion Eq. (24) equals $o_p(n^{-1/2})$. By Eq. (17), the second term in the expansion Eq. (24) equals $h^p\, e_0^{\mathrm{T}} N^{-1} \eta \int \mathcal{B}_p(x)\,\mathrm{d}x + o_p(h^p)$. This completes the proof of (11).

Acknowledgements The authors thank two referees for their helpful comments on the earlier version of the paper.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: B. N. Petrov., F. Csaki (Eds.) 2nd International symposium on information Theory. pp 267–81. Budapest: Akademiai Kiado. (Reproduced (1992) In: S. Kotz & N. L. Johnson (Eds.) Breakthroughs in Statistics *1*, Sringer-Verlag, New York, pp. 610–624.

Akaike, H. (1974). A new look at the statistical model selection. *IEEE Transactions on Automatic Control 19*, 716–723.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statististics 12*, 171–178.

Copas, J. B. (1995). Local likelihood based on kernel censoring. *Journal of the Royal Statistical Society, Series B 57*, 221–235.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: SIAM.

Eguchi, S., Copas, J. B. (1998). A class of local likelihood methods and near-parametric asymptotics. *Journal of the Royal Statistical Society, Series B 60*, 709–724.

Eguchi, S., Kim, T. Y., Park, B. U. (2003). Local likelihood method: a bridge over parametric and nonparametric regression. *Journal of Nonparametric Statistics 15*, 665–683.

Hjort, N. L., Jones, M. C. (1996). Locally parametric nonparametric density estimation. *The Annals of Statistics 24*, 1619–1647.

Konishi, S., Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika 83*, 875–890.

Kullback, S., Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics 22*, 79–86.

Park, B. U., Kim, W. C., Jones, M. C. (2002). On local likelihood density estimation. *The Annals of Statistics 30*, 1480–1495.

Park, B. U., Lee, Y. K., Kim, T. Y., Park, C., Eguchi, S. (2006). On local likelihood density estimation when the bandwidth is large. *Journal of Statistical Planning and Inference*, *136*, 839–859.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B 39*, 44–47.