Bernoulli **10**(4), 2004, 583–604

Semiparametric density estimation under a two-sample density ratio model

K.F. CHENG[1] and C.K. CHU[2]

[1]*Institute of Statistics, National Central University, Chungli, Taiwan 320.*
E-mail: kfcheng@cc.ncu.edu.tw
[2]*Department of Applied Mathematics, National Donghua University, Hualien, Taiwan 974.*
E-mail: chu@server.am.ndhu.edu.tw

A semiparametric density estimation is proposed under a two-sample density ratio model. This model, arising naturally from case–control studies and logistic discriminant analyses, can also be regarded as a biased sampling model. Our proposed density estimate is therefore an extension of the kernel density estimate suggested by Jones for length-biased data. We show that under the model considered the new density estimator not only is consistent but also has the 'smallest' asymptotic variance among general nonparametric density estimators. We also show how to use the new estimate to define a procedure for testing the goodness of fit of the density ratio model. Such a test is consistent under very general alternatives. Finally, we present some results from simulations and from the analysis of two real data sets.

Keywords: asymptotic relative efficiency; biased sampling problem; case–control data; density estimation; goodness-of-fit test; logistic regression; semiparametric maximum likelihood estimation

1. Introduction

A basic characteristic describing the behaviour of a random variable X is its probability density function $g(x)$. Knowledge of the density function helps us in many respects. For instance, the density tells us when observations cluster and occur more frequently. By looking at the graph of a density function we may also say 'the distribution of X is skewed' or 'the distribution is multimodal'. Many more structural elements and features of X can be seen just by interpreting and analysing the density. The estimation of the unknown density g thus provides a way of understanding and representing the behaviour of a random variable. For example, Zhao *et al*., (1996) considered a case–control study focusing analyses on the association of colon cancer with energy, alcohol, and fibre intake among 238 male cases and 223 male controls. In the preliminary analyses, they used the method described by Silverman (1986, Section 2.4) to show that for total energy, the shapes of the nonparametric density estimates for cases and controls do not differ substantially. For alcohol intake, the estimated densities appear to be bimodal, clustering near zero and 19 g/day. Controls were more likely to be non-drinkers, and among drinkers controls consumed less alcohol than cases. For fibre, the estimated densities among cases and

1350–7265

controls are similar. Such preliminary results are helpful in studying the relationships between colon cancer and other potential risk factors.

Let $\{X_1, \ldots, X_{n_0}\}$ be a random sample from a population with density function $g(t)$. Independent of these X_j, let $\{Z_1, \ldots, Z_{n_1}\}$ be another random sample from a population with density function $h(t)$. If the densities g and h are not related in any way, then the ordinary estimated density functions are

$$\hat{g}(t) = n_0^{-1} \sum_{j=1}^{n_0} K_b(t - X_j) \quad \text{and} \quad \hat{h}(t) = n_1^{-1} \sum_{k=1}^{n_1} K_b(t - Z_k),$$

where K is a probability density function supported on $[-1, 1]$, b is the smoothing parameter, and $K_b(\cdot) = b^{-1}K(\cdot/b)$; see Silverman (1986, Section 2.4). Sometimes, however, the two density functions are related in some way, and hence when estimating the densities g and h, we can take the 'information' contained in both samples $\{X_1, \ldots, X_{n_0}\}$ and $\{Z_1, \ldots, Z_{n_1}\}$ into account. In this paper, we shall consider estimating $g(t)$ and $h(t)$ under the following two-sample density ratio model in which the two density functions are related by

$$h(t) = g(t)\exp\{\alpha + r(t)\beta\}, \tag{1.1}$$

where $r(t) = \{r_1(t), \ldots, r_d(t)\}$ is a known d-dimensional row vector of functions of t, $\beta = (\beta_1, \ldots, \beta_d)^{\mathrm{T}}$ is a d-dimensional column vector of parameters, and α is a normalizing parameter that makes $h(t)$ integrate to 1.

The two-sample density ratio model arises naturally from the logistic regression analysis of case–control data. For example, let Y be a two-state response taking value $Y = 1$ for an individual who becomes a 'case', and $Y = 0$ otherwise (that is, for a 'control'). Let T be the explanatory variable such as the potential risk factor for a disease. Then the usual prospective logistic regression model relating T to Y is given by

$$\Pr(Y = 1|T = t) = \frac{\exp(\alpha^{\#} + \beta t)}{1 + \exp(\alpha^{\#} + \beta t)}.$$

Suppose $h(t) = f(t|Y = 1)$, the conditional density of T given $Y = 1$, and $g(t) = f(t|Y = 0)$, the conditional density of T given $Y = 0$; then Bayes's theorem gives

$$h(t) = g(t)\exp(\alpha + t\beta),$$

where $\alpha = \alpha^{\#} + \log\{\Pr(Y = 0)/\Pr(Y = 1)\}$. Thus we have two-sample density ratio model (1.1) with $r(t) = t$. In view of this discussion, suppose we can find better estimates of $g(t)$ and $h(t)$ based on model (1.1) and the combined sample of $\{X_1, \ldots, X_{n_0}\}$ and $\{Z_1, \ldots, Z_{n_1}\}$; then the conclusions in Zhao *et al.* (1996) may be improved. Further, since the improved semiparametric density estimate is derived from model (1.1), the 'difference' between this density estimate and the ordinary kernel density estimate can serve as a statistic to test the validity of model (1.1). Under case–control data, such a statistic can be used to test the goodness of fit of the logistic regression model.

Note that model (1.1) can also be viewed as a biased sampling model. For example, if $n_0 = 0$, $\beta = 1$ and $r(t) = \ln(t)$, then the resulting model (1.1) is the one-sample biased sampling model studied by Jones (1991). Also, if $n_0 > 0$ and $n_1 > 0$, then model (1.1)

becomes the two-sample biased sampling model considered by Vardi (1982; 1985) and Jones (1991). Hence our new density estimate is an extension of the kernel density estimate proposed by Jones (1991) for length-biased data. Other discussions of the biased sampling model and related problems can be found in Zhang (2000).

In this paper, we are interested in the problem of density estimation under model (1.1). In Section 2, a simple motivation and a formal definition of our semiparametric density estimate are given. Using the Cauchy–Schwarz inequality, our proposed semiparametric density estimator not only is consistent but also has the 'smallest' asymptotic variance among general nonparametric density estimators. Theoretical properties of the new density estimator and its relative asymptotic efficiency with respect to the ordinary kernel density estimator are presented in Section 3. An application for testing the adequacy of model (1.1) using the L_2 norm of the difference between the semiparametric and the nonparametric density estimators is described in Section 4. Some examples and simulation studies are presented in Section 5 to show the finite-sample behaviour of the new methods. A concluding remark and some related topics are given in Section 6. Finally, proofs of the main theoretical results are given in the Appendix.

Before closing this section, we remark that model (1.1) is also related to the exponential family of densities considered by Efron and Tibshirani (1996). They used a single sample to estimate

$$h(t) = g(t)\exp\{\alpha + s(t)\beta\},$$

where $s(t)$ is a known d-dimensional row vector of sufficient statistics, α and β are as in (1.1), and $g(t)$ is a carrier density. If $r(t)$ in (1.1) is taken as $s(t)$, then our model (1.1) becomes Efron and Tibshirani's model. To estimate $h(x)$, their method is to first use a nonparametric smoother to estimate $g(t)$ and then fit a parametric family. This is the reverse of the order of the estimation procedures suggested by Hjort and Glad (1995): first fit a parametric family to the data and then fit a nonparametric smoother to the residuals from the parametric estimator. Our approach is more like that of Hjort and Glad: first estimate parameters (α, β) and then calculate the density estimate, except that here we consider a two-sample model. Efron and Tibshirani's method has also been extended to investigate density differences in multisample situations. They used the exponential family model for the different densities with a shared carrier. While the present paper was under review, Fokianos (2002) also considered the density estimation of $h(t)$ under the same multisample model discussed by Efron and Tibshirani. His estimate of $h(t)$ is an extension of ours, which is for the two-sample situation. However, he only discussed the asymptotic bias and variance property of the estimator. In the present paper, we give a more insightful motivation to derive the new estimator, showing that it has the smallest variance among general density estimates. Moreover, we also consider the problem of testing model (1.1) based on the L_2 norm of the difference between the semiparametric and nonparametric density estimators. Approximate p-values of the test can be obtained using the normal distribution. The multisample model has many other applications. Here we only remark that this model has recently been applied to suggest an approach which generalizes the classical normal-based one-way analysis of variance; see Fokianos *et al*. (2001).

2. Motivation and formal definitions

Let $\{T_1, \ldots, T_n\}$ denote the combined sample $\{X_1, \ldots, X_{n_0}, Z_1, \ldots, Z_{n_1}\}$ with $n = n_0 + n_1$. We first consider the estimation of the density function $g(t)$. A similar definition can be applied for estimating the density function $h(t)$. To fully use the information contained in the combined sample, we consider a general density estimate

$$\tilde{g}^*(t) = \sum_{i=1}^{n} u_i(T_i) K_b(t - T_i).$$

Here $u_i(T_i)$ is a random weight attached to T_i, for each $i = 1, \ldots, n$. If $u_i(T_i) = n_0^{-1}$ for $i = 1, \ldots, n_0$ and $u_i(T_i) = 0$ for $i = n_0 + 1, \ldots, n$, then clearly $\tilde{g}^*(t) = \hat{g}(t)$, the ordinary kernel density estimate. On the other hand, if we have $u_i(T_i)$ independent of T_i, that is, $u_i(T_i) = u_i$, then we have non-random weights.

Under model (1.1), to determine the optimal random weights $u_i(T_i)$, we first derive the following results by assuming $b = b_n \to 0$ as $n \to \infty$, and Lipschitz continuity of K, g and u_i:

$$\mathrm{E}\{u_i(T_i) K_b(t - T_i)\} = \begin{cases} u_i(t) g(t) + O(b), & \text{for } 1 \leqslant i \leqslant n_0, \\ u_i(t) g(t) w(t) + O(b), & \text{for } n_0 + 1 \leqslant i \leqslant n, \end{cases},$$

$$\mathrm{var}\{u_i(T_i) K_b(t - T_i)\} = \begin{cases} u_i(t)^2 g(t) b^{-1} \kappa_S \{1 + o(1)\}, & \text{for } 1 \leqslant i \leqslant n_0, \\ u_i(t)^2 g(t) w(t) b^{-1} \kappa_S \{1 + o(1)\}, & \text{for } n_0 + 1 \leqslant i \leqslant n, \end{cases}$$

where the function $w(t) = \exp\{\alpha + r(t)\beta\}$ and $\kappa_S = \int_{-1}^{1} K(u)^2 \, \mathrm{d}u$. As a consequence, if $n \to \infty$,

$$\mathrm{E}\{\tilde{g}^*(t)\} = \left\{ \sum_{i=1}^{n_0} u_i(t) + \sum_{i=n_0+1}^{n} u_i(t) w(t) \right\} g(t) + O(b),$$

$$\mathrm{var}\{\tilde{g}^*(t)\} = \left\{ \sum_{i=1}^{n_0} u_i(t)^2 + \sum_{i=n_0+1}^{n} u_i(t)^2 w(t) \right\} g(t) b^{-1} \kappa_S \{1 + o(1)\}.$$

Suppose we wish $\tilde{g}^*(t)$ to be asymptotically unbiased for $g(t)$, like the ordinary kernel density estimate. Then the optimal choice of $\{u_i(T_i)\}$ can be obtained by solving

$$\min \left\{ \sum_{i=1}^{n_0} u_i(t)^2 + \sum_{i=n_0+1}^{n} u_i(t)^2 w(t) \right\},$$

subject to

$$\sum_{i=1}^{n_0} u_i(t) + \sum_{i=n_0+1}^{n} u_i(t) w(t) = 1, \qquad u_i(t) \geqslant 0, \quad \text{for } i = 1, \ldots, n. \tag{2.1}$$

A straightforward calculation using the Cauchy–Schwarz inequality gives

$$\min_{\{u_i(t) \text{ subject to } (2.1)\}} \left\{ \sum_{i=1}^{n_0} u_i(t)^2 + \sum_{i=n_0+1}^{n} u_i(t)^2 w(t) \right\} = \{n_0 + n_1 w(t)\}^{-1},$$

and such a minimum value is arrived at by choosing

$$u_i(t) = \{n_0 + n_1 w(t)\}^{-1} \equiv p(t).$$

Note that these optimal random weights $p(T_i)$ depend on the unknown regression parameters α and β. In practice, we replace (α, β) by the semiparametric maximum likelihood estimate $(\tilde{\alpha}, \tilde{\beta})$ obtained as the solution of the score equation:

$$\sum_{k=1}^{n_1} \{1,\ r(Z_k)\}^{\mathrm{T}} - \sum_{i=1}^{n} n_1 \{1,\ r(T_i)\}^{\mathrm{T}} \frac{\exp\{\alpha + r(T_i)\beta\}}{n_0 + n_1 \exp\{\alpha + r(T_i)\beta\}} = 0; \tag{2.2}$$

see Prentice and Pyke (1979) or Qin (1998). Therefore, the formal definition of the new semiparametric density estimate is given by

$$\tilde{g}(t) = \sum_{i=1}^{n} \tilde{p}(T_i) K_b(t - T_i),$$

where $\tilde{p}(t) = [n_0 + n_1 \exp\{\tilde{\alpha} + r(t)\tilde{\beta}\}]^{-1}$.

Note also that our semiparametric density estimate $\tilde{g}(t)$ is a direct development of the kernel density estimates $\hat{f}(t)$ and $\hat{f}_2(t)$ in Jones (1991) proposed respectively for the one-sample and the two-sample biased sampling models. The latter paper considers the density ratio model (1.1) with $\beta = 1$ and $r(t) = \ln(t)$. From the above arguments, we can conclude that each $\hat{f}(t)$ and $\hat{f}_2(t)$ has the 'smallest' asymptotic variance among general non-parametric density estimators.

Note, further, that our optimal semiparametric density estimate $\tilde{g}(t)$ is related to the semiparametric maximum likelihood estimate of the cumulative distribution function $G(t)$ of $g(t)$. According to Qin (1998), the semiparametric maximum likelihood estimate of $G(t)$ is

$$\tilde{G}(t) = \sum_{i=1}^{n} \tilde{p}(T_i) I(T_i \leqslant t).$$

Thus one can show that

$$\tilde{g}(t) = \int_{-\infty}^{\infty} K_b(t - u) \mathrm{d}\tilde{G}(u).$$

In contrast, the ordinary kernel density estimate is

$$\hat{g}(t) = \int_{-\infty}^{\infty} K_b(t - u) \mathrm{d}\hat{G}(u),$$

where $\hat{G}(t) = n_0^{-1} \sum_{j=1}^{n_0} I(X_j \leqslant t)$ is the ordinary nonparametric maximum likelihood estimate of $G(t)$.

Finally, we remark that the two-sample density ratio model (1.1) can be rewritten as

$$X_1, \ldots, X_{n_0} \sim_{\text{i.i.d.}} g(t) = \exp\{-\alpha - r(t)\beta\} h(t),$$

and

$$Z_1, \ldots, Z_{n_1} \sim_{\text{i.i.d.}} h(t).$$

Thus the semiparametric maximum likelihood estimate for the cumulative distribution function $H(t)$ of $h(t)$ is given by

$$\tilde{H}(t) = \sum_{i=1}^{n} \tilde{p}(T_i)\exp\{\tilde{\alpha} + r(T_i)\tilde{\beta}\} I[T_i \leqslant t].$$

As a consequence, the optimal semiparametric density estimator of $h(t)$ is given by

$$\tilde{h}(t) = \sum_{i=1}^{n} \tilde{p}(T_i)\exp\{\tilde{\alpha} + r(T_i)\tilde{\beta}\} K_b(t - T_i).$$

3. Asymptotic mean square errors and asymptotic relative efficiency

In this section, we study the asymptotic bias and variance of the semiparametric density estimator. For this purpose, the following assumptions are required:

(B1) The probability density function g is positive on $\mathbb{R}$ and has two Lipschitz continuous derivatives.
(B2) The kernel function K is a Lipschitz continuous and symmetric probability density function with support $[-1, 1]$.
(B3) The value of b is selected from the interval $[\delta n^{-1+\delta}, \delta^{-1} n^{-\delta}]$, where δ is an arbitrarily small positive constant.
(B4) $n_0/n \to \zeta \in (0, 1)$ as $n \to \infty$, and the value of b satisfies $nb^4 \to \infty$ as $n \to \infty$.

Set $g^{(2)}$ as the second derivative of g, and recall $\kappa_S = \int_{-1}^{1} K(u)^2 \, \mathrm{d}u$ and $\kappa_j = \int_{-1}^{1} u^j K(u) \mathrm{d}u$, for $j \geqslant 0$. The following theorem gives the asymptotic bias and variance of the semiparametric density estimator $\tilde{g}(t)$. The corresponding results for the nonparametric density estimator $\hat{g}(t)$ are also provided for the purpose of comparison. The proof of Theorem 3.1 will be given in the Appendix.

Theorem 3.1. *If model* (1.1) *and assumptions* (B1)–(B4) *are satisfied, then we have the following results for asymptotic bias and variance as* $n \to \infty$*:*

$$\text{bias}\{\tilde{g}(t)\} = \tfrac{1}{2} b^2 g^{(2)}(t)\kappa_2 + o(b^2), \tag{3.1}$$

$$\text{bias}\{\hat{g}(t)\} = \tfrac{1}{2} b^2 g^{(2)}(t)\kappa_2 + o(b^2), \tag{3.2}$$

$$\text{var}\{\tilde{g}(t)\} = n^{-1} b^{-1} \{\zeta + (1 - \zeta) w(t)\}^{-1} g(t)\kappa_S + o(n^{-1} b^{-1}), \tag{3.3}$$

$$\text{var}\{\hat{g}(t)\} = n^{-1} b^{-1} \zeta^{-1} g(t)\kappa_S + o(n^{-1} b^{-1}), \tag{3.4}$$

for $t \in \mathbb{R}$.

Remark 3.1. From Theorem 3.1 we see that, up to first order, the asymptotic bias of $\tilde{g}(t)$ is the same as that of $\hat{g}(t)$, independent of the regression parameters (α, β), and dependent only on the unknown factor $g^{(2)}(t)$. However, the dominant term of the asymptotic variance of $\tilde{g}(t)$ is smaller than that of $\hat{g}(t)$. The magnitude of the difference between these two asymptotic variances increases as ζ, the proportion of the control data, decreases. On the other hand, by (A.2) and (A.3) in the Appendix, the dominant terms of the asymptotic bias and variance of $\tilde{g}(t)$ using estimated values $(\tilde{\alpha}, \tilde{\beta})$ are the same as those of $\tilde{g}(t)$ using the true values (α, β). The same remark can be applied to $\tilde{g}(t)$ using other estimated values of (α, β) with $n^{1/2}$ consistency.

Remark 3.2. Suppose we take $b = cn^{-a}$, where a and c are two positive constants. Then, from Theorem 3.1, we can derive the 'optimal' $a^* = \frac{1}{5}$ and $c^* = [\{\zeta + (1-\zeta)w(t)\}^{-1} g(t)\kappa_S g^{(2)}(t)^{-2}\kappa_2^{-2}]^{1/5}$, in order to minimize the asymptotic mean square error for $\tilde{g}(t)$ over b. Let $\tilde{g}_{\mathrm{opt}}(t)$ be such a semiparametric density estimator with $b = c^* n^{-a^*}$. Then the corresponding asymptotic mean square error for $\tilde{g}_{\mathrm{opt}}(t)$ is

$$\mathrm{AMSE}\{\tilde{g}_{\mathrm{opt}}(t)\} = \tfrac{5}{4}\{g^{(2)}(t)\kappa_2\{\zeta + (1-\zeta)w(t)\}^{-2} g(t)^2\kappa_S^2\}^{2/5} n^{-4/5} + o(n^{-4/5}).$$

We can define $\hat{g}_{\mathrm{opt}}(t)$ similarly, and the corresponding asymptotic mean square error is

$$\mathrm{AMSE}\{\hat{g}_{\mathrm{opt}}(t)\} = \tfrac{5}{4}\{g^{(2)}(t)\kappa_2\zeta^{-2} g(t)^2\kappa_S^2\}^{2/5} n^{-4/5} + o(n^{-4/5}).$$

From this argument, it is seen that the asymptotic relative efficiency of $\hat{g}(t)$ with respect to $\tilde{g}(t)$ can be naturally defined as

$$e\{\hat{g}(t), \tilde{g}(t)\} = \lim_{n\to\infty} \frac{\mathrm{AMSE}\{\tilde{g}_{\mathrm{opt}}(t)\}}{\mathrm{AMSE}\{\hat{g}_{\mathrm{opt}}(t)\}} = \left[\frac{\zeta}{\zeta + (1-\zeta)w(t)}\right]^{4/5}.$$

Clearly, $e\{\hat{g}(t), \tilde{g}(t)\}$ depends on ζ, regression parameters (α, β), and t. However, since $w(t)$ is a positive function of t, the asymptotic relative efficiency $e\{\hat{g}(t), \tilde{g}(t)\}$ is always less than or equal to one. Figure 1 shows some results for $e\{\hat{g}(t), \tilde{g}(t)\}$ when $\alpha = 0$, $\beta = 1$ and $r(t) = t$. Clearly, the asymptotic relative efficiency of $\hat{g}(t)$ with respect to $\tilde{g}(t)$ decreases as ζ decreases, and, for each fixed ζ value, $e\{\hat{g}(t), \tilde{g}(t)\}$ is a monotonic decreasing function of t. Therefore, $\tilde{g}(t)$ performs better than $\hat{g}(t)$, especially for large t values.

Remark 3.3. From Theorem 3.1 and the discussions in Epanechnikov (1969), we can see that, in order to minimize the asymptotic mean square error, the optimal K for constructing $\tilde{g}(t)$ is the Epanechnikov kernel $K(u) = \frac{3}{4}(1-u^2)I_{[-1,1]}(u)$. Further, in practice, one can consider the idea of least-squares cross-validation (Silverman 1986) to determine b. The practical choice of b will be $\tilde{b}_{CV}$, which is the minimizer of

$$\begin{aligned} CV(b; \tilde{g}) &= \int_{-\infty}^{\infty} \tilde{g}(t)^2 \,\mathrm{d}t - 2n_0^{-1}\sum_{i=1}^{n_0} \tilde{g}_i(T_i) \\ &= b^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{p}(T_i)\tilde{p}(T_j) K * K\{(T_i - T_j)/b)\} - 2n_0^{-1}\sum_{i=1}^{n_0} \tilde{g}_i(T_i), \end{aligned}$$

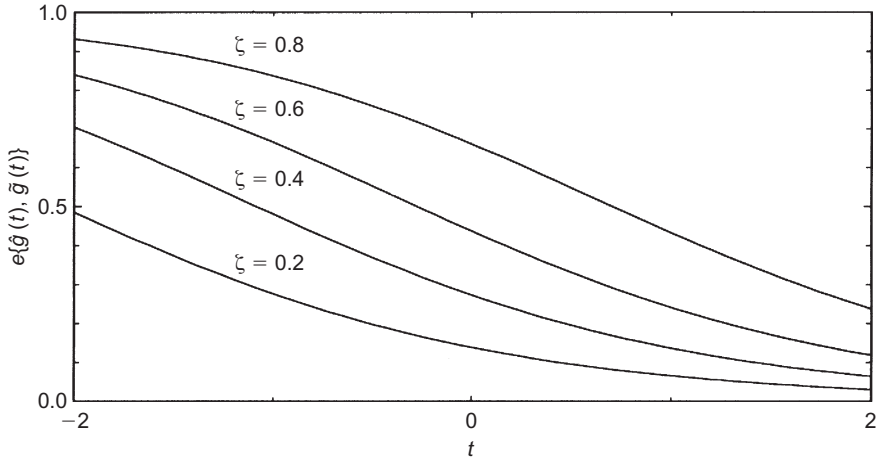

Figure 1. Plot of the asymptotic relative efficiency $e\{\hat{g}(t), \tilde{g}(t)\}$ when $\alpha = 0$, $\beta = 1$ and $r(t) = t$.

where the symbol $*$ denotes convolution and $\tilde{g}_i(t)$ is $\tilde{g}(t)$ with T_i dropped from the combined data. Specifically, if K is the Epanechnikov kernel, then

$$K * K(u) = \tfrac{3}{160}\{32 - 40|u|^2 + 20|u|^3 - |u|^5\} I_{[-2,2]}(u).$$

Remark 3.4. Under the conditions of Theorem 3.1, we can also derive similar asymptotic results for $\tilde{h}(t)$ and $\hat{h}(t)$:

$$\begin{aligned}
\text{bias}\{\tilde{h}(t)\} &= \tfrac{1}{2} b^2 h^{(2)}(t)\kappa_2 + o(b^2), \\
\text{bias}\{\hat{h}(t)\} &= \tfrac{1}{2} b^2 h^{(2)}(t)\kappa_2 + o(b^2), \\
\text{var}\{\tilde{h}(t)\} &= n^{-1}b^{-1}\{\zeta + (1-\zeta)w(t)\}^{-1} h(t) w(t) \kappa_S + o(n^{-1}b^{-1}), \\
\text{var}\{\hat{h}(t)\} &= n^{-1}b^{-1}(1-\zeta)^{-1} h(t)\kappa_S + o(n^{-1}b^{-1}),
\end{aligned}$$

for $t \in \mathbb{R}$. Similar conclusions to those given in Remarks 3.1–3.3 for $\tilde{g}(t)$ can also be drawn for $\tilde{h}(t)$.

4. An application to testing goodness of fit

In this section, we shall discuss how to use the semiparametric density estimator $\tilde{g}(t)$ and nonparametric density estimator $\hat{g}(t)$ to define a test statistic for testing the adequacy of model (1.1). Conceptually speaking, if model (1.1) is valid, then $\tilde{g}(t)$ and $\hat{g}(t)$ estimate the same density function $g(t)$. Otherwise, $\hat{g}(t)$ estimates density function $g(t)$, but $\tilde{g}(t)$, still a density estimator because of score equation (2.2), may estimate some other density function.

Theorem 4.1 below gives the general asymptotic bias and variance of $\tilde{g}(t)$ in the

situation where model (1.1) is not necessarily valid. Its proof will be given in the Appendix. We require the following additional assumptions:

(B5) The probability density function h is positive on $\mathbb{R}$ and has two Lipschitz continuous derivatives.
(B6) The function r has two Lipschitz continuous partial derivatives.
(B7) The equation $\int\{1, r(u)\}^{\mathrm{T}}\{h(u) - g(u)w(u)\}\{\zeta + (1-\zeta)w(u)\}^{-1}\,\mathrm{d}u = 0$ has unique solution (α^*, β^*).

Theorem 4.1. *If assumptions* (B1)–(B7) *are satisfied, then we have the following results for asymptotic bias and variance as* $n \to \infty$*:*

$$\mathrm{bias}\{\tilde{g}(t)\} = \{N(t)/D^*(t) - g(t)\} + \tfrac{1}{2}b^2\kappa_2[N(t)/D^*(t)]^{(2)} + o(b^2),$$
$$\mathrm{var}\{\tilde{g}(t)\} = n^{-1}b^{-1}\kappa_S N(t)/D^*(t)^2 + o(n^{-1}b^{-1}),$$

where

$$N(t) = \zeta g(t) + (1-\zeta)h(t), \qquad D^*(t) = \zeta + (1-\zeta)w^*(t), \qquad w^*(t) = \exp\{\alpha^* + r(t)\beta^*\}.$$

Note that $N(t)/D^*(t)$ is a density function because of (B7). Generally, $\tilde{g}(t)$ estimates $N(t)/D^*(t)$; it will be reduced to $g(t)$ when model (1.1) is satisfied.

We next use the L_2 norm as a measure of the 'distance' between $\hat{g}(t)$ and $\tilde{g}(t)$. Define

$$L_{2,n} = \int\{\hat{g}(t) - \tilde{g}(t)\}^2\,\mathrm{d}t.$$

In the following we shall develop a general asymptotic theorem for $L_{2,n}$ for the situation where model (1.1) may not be satisfied. The proof of Theorem 4.2 will be given in the Appendix. Define

$$m_1 = \kappa_S(1-\zeta)\zeta^{-1}\int g(t)w(t)/D(t)\mathrm{d}t,$$
$$m_2 = \int\{N(t)/D^*(t) - g(t)\}^2\mathrm{d}t,$$
$$m_3 = \kappa_2\int\{N(t)/D^*(t) - g(t)\}\{N(t)/D^*(t) - g(t)\}^{(2)}\,\mathrm{d}t,$$
$$v_1 = 2\kappa^*(1-\zeta)^2\zeta^{-2}\int\{g(t)w(t)/D(t)\}^2\,\mathrm{d}t,$$
$$v_2 = 4(1-\zeta)\zeta^{-1}\left[\int\{N(t)/D^*(t) - g(t)\}^2\{g(t)w^*(t)/D^*(t)\}\mathrm{d}t + CA^{-1}BA^{-1}C^{\mathrm{T}}\right],$$

where

$$D(t) = \zeta + (1-\zeta)w(t), \qquad \kappa^* = \int_{-2}^{2} K*K(u)^2\,\mathrm{d}u,$$

$$A = \begin{pmatrix} A_0 & A_1 \\ A_1^{\mathrm{T}} & A_2 \end{pmatrix}, \qquad B = \begin{pmatrix} A_0(1-A_0) & A_1(1-A_0) \\ A_1^{\mathrm{T}}(1-A_0) & A_2 - A_1^{\mathrm{T}}A_1, \end{pmatrix},$$

$$C = \int \{N(t)/D^*(t) - g(t)\} N(t) w^*(t) D^*(t)^{-2} \{1,\, r(t)\} \mathrm{d}t,$$

$$A_0 = \int N(t) w^*(t) D^*(t)^{-1}\, \mathrm{d}t, \qquad A_1 = \int N(t) w^*(t) D^*(t)^{-1} r(t) \mathrm{d}t,$$

$$A_2 = \int N(t) w^*(t) D^*(t)^{-1} r(t)^{\mathrm{T}} r(t) \mathrm{d}t,$$

and the symbol $*$ denotes convolution.

Theorem 4.2. *Suppose assumptions* (B1)–(B7) *hold and b satisfies* $nb^6 \to 0$ *as* $n \to \infty$. *If model* (1.1) *is valid, then the limiting distribution of* $L_{2,n}$ *can be expressed as*

$$nb^{1/2}(L_{2,n} - n^{-1}b^{-1}m_1) \Rightarrow N(0,\, v_1), \qquad \text{as } n \to \infty. \tag{4.1}$$

On the other hand, if $v_2 \neq 0$, *then the limiting distribution of* $L_{2,n}$ *can be expressed as*

$$n^{1/2}(L_{2,n} - m_2 - b^2 m_3) \Rightarrow N(0,\, v_2), \qquad \text{as } n \to \infty. \tag{4.2}$$

Note that under model (1.1), the quantities m_1 and v_1 may be estimated by

$$\tilde{m}_1 = \kappa_S (1-\zeta_n)\zeta_n^{-1} \sum_{i=1}^{n} \tilde{p}(T_i)\tilde{w}(T_i)\{\zeta_n + (1-\zeta_n)\tilde{w}(T_i)\}^{-1},$$

$$\tilde{v}_1 = 2\kappa^*(1-\zeta_n)^2\zeta_n^{-2} \sum_{i=1}^{n} \tilde{p}(T_i)\tilde{w}(T_i)^2\tilde{g}(T_i)\{\zeta_n + (1-\zeta_n)\tilde{w}(T_i)\}^{-2},$$

where $\zeta_n = n_0/n$ and $\tilde{w}(t) = \exp\{\tilde{\alpha} + r(t)\tilde{\beta}\}$. Employing the properties of $\tilde{\alpha}$ and $\tilde{\beta}$, we can easily prove that, as $n \to \infty$,

$$\tilde{m}_1 - m_1 = O_p(n^{-1/2}) \quad \text{and} \quad \tilde{v}_1 - v_1 = o_p(1).$$

Then we can conclude that

$$nb^{1/2}(L_{2,n} - n^{-1}b^{-1}\tilde{m}_1)\tilde{v}_1^{-1/2} \Rightarrow N(0,\, 1), \qquad \text{as } n \to \infty,$$

without the need for conditions beyond those stated in Theorem 4.2. Based on this result, suppose we select α^* as the significance level and let z_{α^*} denote the upper $100\alpha^*$ percentile point of $N(0,\, 1)$. Then the test procedure is to reject model (1.1) if

$$L_{2,n} \geqslant n^{-1}b^{-1/2}(\tilde{v}_1^{1/2} z_{\alpha^*} + b^{-1/2}\tilde{m}_1) \equiv L_{2,n,\alpha^*}.$$

Note also that, in general, if model (1.1) is not satisfied, then $v_2 \neq 0$ and $m_2 > 0$. In this case, the corresponding limiting power function will be

$$\lim_{n\to\infty} \Pr\{n^{1/2}(L_{2,n} - m_2 - b^2 m_3)v_2^{-1/2} > n^{1/2}(L_{2,n,\alpha^*} - m_2 - b^2 m_3)v_2^{-1/2}\}$$

$$= \lim_{n\to\infty} [1 - \Phi\{n^{1/2}(L_{2,n,\alpha^*} - m_2 - b^2 m_3)v_2^{-1/2}\}] = 1,$$

where $\Phi(t)$ is the standard normal distribution function. Hence the test is consistent under such situation.

Example 4.1. Suppose we let $g(t)$ be the $N(0, 1)$ density and $h(t) = g(t)\exp(a_0 - r_0 t^{2/3})$, where a_0 is the normalizing constant determined by the choice of r_0. Suppose we further assume $r(t) = t$ in model (1.1). Then

$$m_2 = (1-\zeta)^2 \int \phi(t)^2 \{\exp(a_0 - r_0 t^{2/3}) - \exp(\alpha^* + \beta^* t)\}^2 \{\zeta + (1-\zeta)\exp(\alpha^* + \beta^* t)\}^{-2}\, \mathrm{d}t,$$

where ϕ is the standard normal density and (α^*, β^*) is the unique solution of

$$\int (1,\ t)^{\mathrm{T}} \phi(t) \{\exp(a_0 - r_0 t^{2/3}) - \exp(\alpha + \beta t)\} \{\zeta + (1-\zeta)\exp(\alpha + \beta t)\}^{-1}\, \mathrm{d}t = 0.$$

Figure 2 gives some numerical results for values of m_2. It is clear that $m_2 > 0$ for $r_0 \neq 0$, the value of m_2 increases as ζ decreases, and, for each fixed ζ value, m_2 is a monotonic increasing function of $|r_0|$. In this situation, our test for the validity of model (1.1) is consistent.

Remark 4.1. A Kolmogorov–Smirnov type test statistic based on the semiparametric maximum likelihood estimate $\tilde{G}(t)$ of the distribution function $G(t)$ has been proposed by Qin and Zhang (1997) for testing the validity of model (1.1). The formulation of $\tilde{G}(t)$ was given in Section 2. However, to apply their procedures, one needs to use a bootstrap method to find critical values of their test. Another method of applying $\tilde{G}$ is a chi-squared

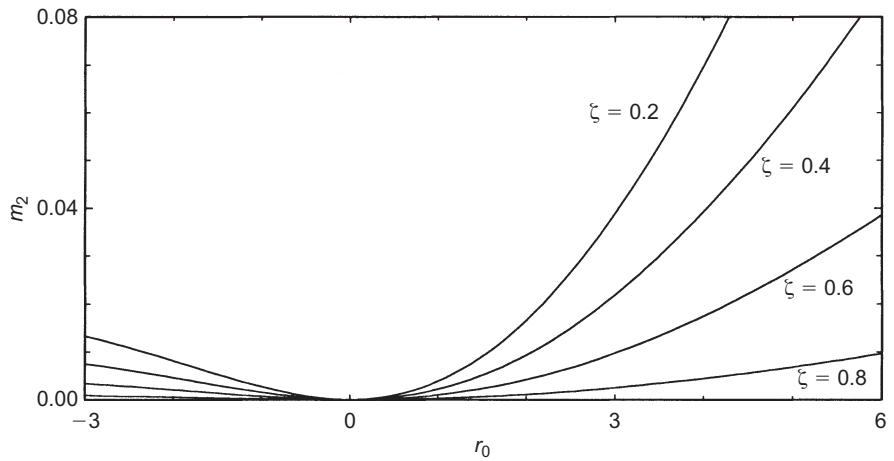

Figure 2. Plot of the value of m_2 in Example 4.1.

goodness-of-fit test, suggested by Zhang (1999). His test depends on the partition of the space of X. But examples exist showing that different partitions may lead to completely different conclusions; see Cheng and Chen (2003). More recently, Zhang (2001) has suggested an information matrix test. This test was derived by extending White's (1982) approach to the semiparametric profile likelihood setting. The test statistic requires a high-dimensional matrix inversion and chi-squared approximation for its distribution. Usually, such an approximation becomes unacceptable when the dimension of X is greater than or equal to 2 (the corresponding degrees of freedom of the chi-squared distribution are greater than or equal to 6). In contrast, our approach for testing model (1.1) is quite different in nature. Our test statistic is based on the difference between the semiparametric and nonparametric density estimates of g. Under conditions more general than those given by Zhang (1999; 2001), our test is shown to be asymptotically consistent.

5. Simulations and examples

To evaluate the performance of the semiparametric and the nonparametric density estimators, empirical studies were carried out. The kernel function used by each discussed estimator was the Epanechnikov kernel. Simulation studies are presented in Section 5.1, and our methods were applied to two real data examples in Section 5.2.

5.1. Simulations

A simulation study was performed to compare the performance of $\tilde{g}(t)$ and $\hat{g}(t)$. Our working model is that densities $g(t)$ and $h(t)$ are related by $h(t) = g(t)\exp(\alpha + \beta_1 t)$. However, the true densities $g(t)$ and $h(t)$ were taken respectively as the density functions of $N(0, \tau^2)$ and $N(\mu, \sigma^2)$. Hence $g(t)$ and $h(t)$ are related by $h(t) = g(t)\exp(\alpha + t\beta_1 + t^2\beta_2)$, where $\alpha = \ln(\tau/\sigma) - \mu^2/(2\sigma^2)$, $\beta_1 = \mu/\sigma^2$, $\beta_2 = \frac{1}{2}(\tau^{-2} - \sigma^{-2})$. The values $\tau = \frac{1}{2}$, $\beta_1 = -1, -\frac{1}{2}, 0, \frac{1}{2}, 1$ and $\beta_2 = -1, -\frac{1}{2}, 0, \frac{1}{2}, 1$ were considered. As a consequence, the working model is misspecified when $\beta_2 \neq 0$. Given the value of τ, for each pair (β_1, β_2), the values of σ^2 and μ were determined by $\sigma^2 = 1/(\tau^{-2} - 2\beta_2)$ and $\mu = \sigma^2\beta_1$, and four sample sizes $n_0 = n_1 = 50$, $n_0 = n_1/4 = 50$, $n_0 = n_1 = 100$, and $n_0 = n_1/4 = 100$ were employed in our study. For each setting, 1000 independent sets of data $(X_1, \ldots, X_{n_0}, Z_1, \ldots, Z_{n_1})$ were generated.

Given each data set, the value of $\tilde{g}(t)$ and that of the corresponding $CV(b; \tilde{g})$ were calculated on an equally spaced logarithmic grid of 101 values of b in $[\frac{1}{10}, 1]$. Given the value of b, the value of the mean integrated square error MISE$(b; \tilde{g})$ for $\tilde{g}(t)$ was empirically approximated by the sample average of the integrated square error ISE$(b; \tilde{g})$ over the 1000 data sets. Here ISE$(b; \tilde{g}) = \int\{\tilde{g}(t) - g(t)\}^2\,\mathrm{d}t$ was empirically approximated by the quantity $(1/\rho)\sum_{i=0}^{4\rho}\{\tilde{g}(\rho_i) - g(\rho_i)\}^2$ with $\rho = 200$ and $\rho_i = -2 + i/\rho$. After evaluation on the grid, $\tilde{b}_M$ and $\tilde{b}_{CV}$ for $\tilde{g}(t)$ were taken respectively as global minimizers of MISE$(b; \tilde{g})$ and $CV(b; \tilde{g})$ on the grid of b. See Marron and Wand (1992) on the point that an equally spaced grid of parameters is typically not a very efficient design for this type of

grid search. When $\tilde{b}_M$ and $\tilde{b}_{CV}$ were obtained, the values of ISE($\tilde{b}_M$; $\tilde{g}$) and ISE($\tilde{b}_{CV}$; $\tilde{g}$) were calculated, and their sample averages over the 1000 data sets were used to measure respectively the best and the practical performance of $\tilde{g}(t)$.

The same computation procedures were applied to $\hat{g}(t)$. Let $\hat{b}_M$, $\hat{b}_{CV}$, and ISE(b; $\hat{g}$) be similarly defined. The simulation results are summarized in the following figures and tables.

Given the working model $(\beta_1, \beta_2) = (\frac{1}{2}, 0)$ with the sample sizes $n_0 = n_1 = 50$, the performance of the two estimators is presented in Figure 3. Figure 3 gives one realization, and shows that density estimates obtained practically and optimally by our proposed estimator $\tilde{g}(t)$ perform better than $\hat{g}(t)$, in the sense of having smaller sample mean square errors for most $t \in [-2, 2]$.

Table 1 shows that, for each model setting ($\beta_2 = 0$), the best performance of $\tilde{g}(t)$ is better than that of $\hat{g}(t)$, since the minimum sample MISE over the grid derived by $\tilde{g}(t)$ is less than that obtained by $\hat{g}(t)$. Table 2 contains the sample mean and standard deviation of ISE($\tilde{b}_{CV}$; $\tilde{g}$) for our proposed estimator $\tilde{g}(t)$, and those of ISE($\hat{b}_{CV}$; $\hat{g}$) for the ordinary kernel density estimator $\hat{g}(t)$. Considering the values of the sample mean and standard deviation, for each model setting, the practical performance of $\tilde{g}(t)$ is still better than that of $\hat{g}(t)$.

When $\beta_2 \neq 0$, we are in the non-model settings. For each fixed n_0, n_1 and β_1, the sample averages and standard deviations of ISE($\tilde{b}_M$; $\tilde{g}$) are increasing in β_2. However, except for $\beta_2 = \frac{1}{2}$ and 1, the averages and standard deviations of ISE($\tilde{b}_M$; $\tilde{g}$) are smaller than the corresponding values of ISE($\hat{b}_M$; $\hat{g}$). For the $\beta_2 = \frac{1}{2}$ case, the differences are not very significant, particularly when sample sizes n_0 and n_1 are small. On the other hand, except for the $\beta_2 = 1$ case, the averages and standard deviations of ISE($\tilde{b}_{CV}$; $\tilde{g}$) are smaller than the corresponding values of ISE($\hat{b}_{CV}$; $\hat{g}$). Our unreported simulation results also indicate that when $\beta_2 = 0.1, 0.2$ and 0.3, $\tilde{g}(t)$ has better optimal and practical performance than $\hat{g}(t)$. Thus, under minor misspecification of the model (such as $|\beta_2| \leqslant 0.3$), the semiparametric density estimate $\tilde{g}(t)$ still seems to be better than the usual density estimate $\hat{g}$.

5.2. Examples

The estimators discussed were applied to two familiar data sets. Density estimates obtained by $\tilde{g}$ and $\tilde{h}$ are given so that distributions of data sets under study can be compared when model (1.1) is satisfied.

Example 5.1. Consider the data set reported by Glovsky and Rigrodsky (1964). The purpose is to analyse and compare the developmental histories of children with learning disabilities. Forty-one children were enrolled in a speech therapy program at the training school at Vineland, New Jersey. Among them, 20 children had been diagnosed as aphasics (cases) and the remaining 21 children were a random sample of children with learning disabilities (controls). The social quotient scores of these children on the Vineland Social Maturity Scale were $X_i = 56, 43, 30, 97, 67, 24, 76, 49, 46, 29, 46, 83, 93, 38, 25, 44, 66, 71, 54, 20, 25$ for

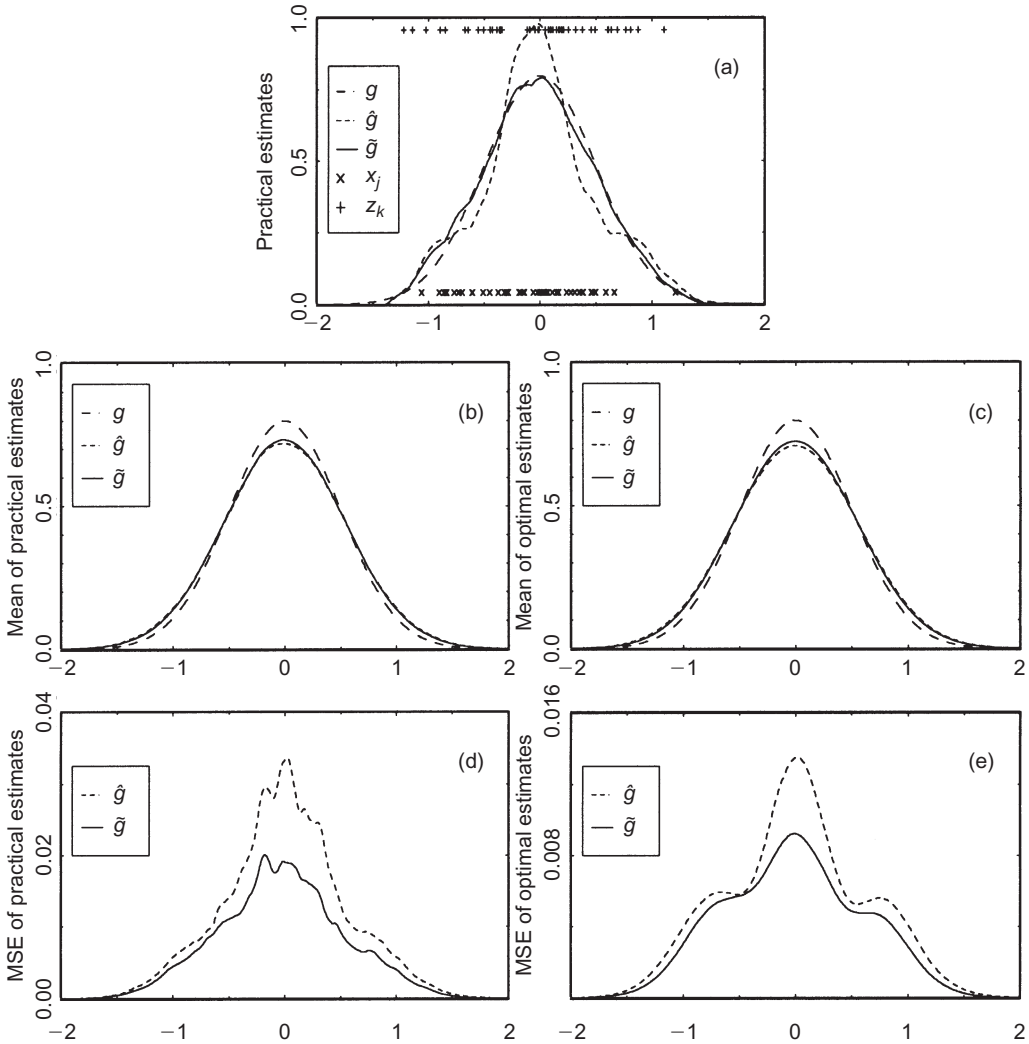

Figure 3. An artificial example. (a) One simulated data set and density estimates produced by $\hat{g}(t)$ and $\tilde{g}(t)$ using their least-squares cross-validated bandwidths. (b) Sample averages of practical density estimates generated by discussed estimators employing their least-squares cross-validated bandwidths. (c) Sample averages of optimal density estimates obtained by discussed estimators using bandwidths as minimizers of their sample mean integrated square errors. (d) Sample mean square error of practical density estimates. (e) Sample mean square error of optimal density estimates.

controls, and $Z_k = 90, 53, 32, 44, 47, 42, 58, 16, 49, 54, 81, 59, 35, 81, 41, 24, 41, 61, 31, 20$ for cases.

Qin and Zhang (1997) and Zhang (1999) argue that the densities $g(t)$ of X_i and $h(t)$ of Z_k can be related by model (1.1) with $r(t) = t$. Using this model, our semiparametric

Table 1. Values of the sample average and standard deviation (in parentheses) of ISE($\tilde{b}_M$; $\tilde{g}$) for $\tilde{g}$, and of ISE($\hat{b}_M$; $\hat{g}$) for $\hat{g}$; each value has been multiplied by 10^3

		$n_0 = n_1 = 50$	$n_0 = n_1/4 = 50$	$n_0 = n_1 = 100$	$n_0 = n_1/4 = 100$
$\hat{g}$		16.8 (13.1)	16.8 (13.1)	10.6 (7.78)	10.6 (7.78)
$\tilde{g}$	β_1		$\beta_2 = -1$		
	-1	9.29 (8.96)	6.54 (8.09)	5.05 (4.44)	3.40 (3.98)
	$-\frac{1}{2}$	8.82 (8.78)	6.31 (8.10)	4.69 (4.28)	3.23 (3.96)
	0	8.66 (8.73)	6.25 (8.13)	4.57 (4.24)	3.19 (4.00)
	$\frac{1}{2}$	8.82 (8.80)	6.33 (8.14)	4.71 (4.36)	3.24 (4.04)
	1	9.28 (9.00)	6.57 (8.14)	5.09 (4.57)	3.42 (4.13)
			$\beta_2 = -\frac{1}{2}$		
	-1	10.8 (9.70)	7.55 (8.56)	6.06 (4.93)	4.00 (4.28)
	$-\frac{1}{2}$	10.4 (9.56)	7.27 (8.56)	5.74 (4.80)	3.81 (4.28)
	0	10.3 (9.51)	7.21 (8.58)	5.65 (4.79)	3.76 (4.33)
	$\frac{1}{2}$	10.4 (9.53)	7.32 (8.49)	5.78 (4.87)	3.82 (4.34)
	1	10.8 (9.65)	7.60 (8.45)	6.12 (5.03)	4.01 (4.38)
			$\beta_2 = 0$		
	-1	13.3 (10.8)	10.1 (9.14)	7.99 (5.82)	5.90 (4.88)
	$-\frac{1}{2}$	13.1 (10.8)	9.80 (9.08)	7.85 (5.75)	5.66 (4.85)
	0	13.0 (10.8)	9.72 (9.12)	7.81 (5.78)	5.57 (4.84)
	$\frac{1}{2}$	13.1 (10.6)	9.87 (9.08)	7.89 (5.83)	5.63 (4.83)
	1	13.3 (10.6)	10.2 (9.04)	8.08 (5.92)	5.87 (4.85)
			$\beta_2 = \frac{1}{2}$		
	-1	17.5 (12.6)	17.5 (10.6)	11.7 (7.40)	12.6 (6.57)
	$-\frac{1}{2}$	18.0 (12.8)	17.6 (10.4)	12.2 (7.46)	12.9 (6.32)
	0	18.3 (12.7)	17.7 (10.3)	12.5 (7.55)	13.0 (6.16)
	$\frac{1}{2}$	18.1 (12.6)	17.7 (10.4)	12.3 (7.59)	12.8 (6.16)
	1	17.6 (12.3)	17.6 (10.5)	11.8 (7.61)	12.5 (6.44)
			$\beta_2 = 1$		
	-1	24.0 (15.3)	34.6 (14.3)	17.9 (9.74)	29.5 (10.1)
	$-\frac{1}{2}$	27.2 (16.0)	38.6 (13.7)	21.2 (10.2)	33.8 (9.56)
	0	28.7 (15.9)	39.9 (13.5)	22.6 (10.4)	35.1 (9.20)
	$\frac{1}{2}$	27.4 (15.7)	38.7 (13.9)	21.3 (10.4)	33.6 (9.35)
	1	24.2 (15.2)	34.8 (14.4)	18.1 (10.2)	29.5 (10.1)

maximum likelihood estimate of (α, β) is $(\tilde{\alpha}, \tilde{\beta}) = (0.3961, -0.0080)$. To compute the goodness-of-fit test statistic $L_{2,n}$, we first note that using the control data and the least-squares cross-validation criterion discussed in Remark 3.3, the minimizer of $CV(b; \hat{g})$ is $\hat{b}_{CV} = 34.5593$. The minimization was performed on a grid of 5001 equally spaced logarithmic values in the interval [1, 100], and the minimizer was taken on the grid. The same bandwidth value was used for computing both $\hat{g}(t)$ and $\tilde{g}(t)$. Next, $L_{2,n} = \int_{-19}^{132} \{\hat{g}(t) - \tilde{g}(t)\}^2 \, \mathrm{d}t \equiv L_{2,n}$ (observed) was empirically approximated by the quantity $(1/\rho)\sum_{i=0}^{151\times\rho} \{\hat{g}(\rho_i) - \tilde{g}(\rho_i)\}^2$ with $\rho = 200$ and $\rho_i = -19 + i/\rho$. Here the lower limit of

Table 2. Values of the sample average and standard deviation (in parentheses) of ISE($\tilde{b}_{CV}$; $\tilde{g}$) for $\tilde{g}$, and of ISE($\hat{b}_{CV}$; $\hat{g}$) for $\hat{g}$; each value has been multiplied by 10^3

		$n_0 = n_1 = 50$	$n_0 = n_1/4 = 50$	$n_0 = n_1 = 100$	$n_0 = n_1/4 = 100$
$\hat{g}$		35.7 (40.5)	35.7 (40.5)	18.9 (17.4)	18.9 (17.4)
$\tilde{g}$	β_1		$\beta_2 = -1$		
	-1	23.2 (25.1)	17.0 (16.7)	11.7 (12.5)	9.05 (9.00)
	$-\frac{1}{2}$	23.0 (25.8)	16.4 (15.9)	11.3 (12.0)	8.81 (8.75)
	0	22.4 (25.5)	16.6 (16.9)	11.5 (12.6)	8.76 (8.81)
	$\frac{1}{2}$	21.5 (23.7)	16.1 (15.9)	11.4 (11.3)	8.65 (8.83)
	1	21.9 (23.8)	16.6 (16.3)	11.6 (11.5)	8.76 (8.79)
			$\beta_2 = -\frac{1}{2}$		
	-1	23.7 (24.0)	16.4 (14.6)	12.6 (12.2)	8.82 (7.53)
	$-\frac{1}{2}$	23.8 (25.1)	16.0 (14.7)	11.9 (11.5)	8.25 (7.20)
	0	23.0 (23.7)	16.0 (14.7)	12.1 (11.9)	8.41 (7.32)
	$\frac{1}{2}$	22.3 (22.6)	15.5 (13.7)	11.9 (10.4)	8.51 (7.37)
	1	22.2 (21.8)	16.0 (14.4)	12.2 (10.9)	8.54 (7.59)
			$\beta_2 = 0$		
	-1	25.6 (24.1)	17.7 (13.5)	13.9 (11.4)	9.58 (6.77)
	$-\frac{1}{2}$	25.7 (24.6)	16.9 (13.5)	13.5 (11.0)	9.01 (6.42)
	0	25.4 (22.8)	16.5 (12.8)	13.7 (11.4)	9.10 (6.51)
	$\frac{1}{2}$	24.2 (21.8)	16.8 (12.7)	13.5 (10.3)	9.08 (6.54)
	1	24.7 (22.3)	17.4 (13.5)	13.8 (11.0)	9.35 (7.02)
			$\beta_2 = \frac{1}{2}$		
	-1	29.8 (25.7)	23.8 (13.8)	17.5 (12.2)	15.5 (7.57)
	$-\frac{1}{2}$	29.3 (23.4)	22.9 (13.0)	17.4 (11.2)	15.3 (7.11)
	0	29.5 (22.2)	23.2 (12.7)	17.8 (11.5)	15.5 (6.95)
	$\frac{1}{2}$	28.2 (21.0)	23.1 (12.7)	17.2 (10.7)	15.3 (6.96)
	1	28.7 (23.9)	23.5 (14.0)	17.4 (11.9)	15.4 (7.63)
			$\beta_2 = 1$		
	-1	36.2 (26.6)	40.6 (17.1)	23.6 (13.7)	32.1 (10.9)
	$-\frac{1}{2}$	37.7 (24.5)	43.4 (15.3)	26.0 (12.5)	35.9 (9.99)
	0	38.7 (23.0)	44.3 (14.7)	27.4 (12.6)	37.1 (9.55)
	$\frac{1}{2}$	36.8 (21.9)	43.4 (15.2)	26.3 (13.1)	35.7 (9.68)
	1	35.9 (25.6)	40.6 (16.6)	23.7 (13.4)	32.1 (10.7)

the integral is $-19 \approx -\hat{b}_{CV} + \min_{1\leqslant i\leqslant n}\{T_i\}$ and the upper limit is $132 \approx \hat{b}_{CV} + \max_{1\leqslant i\leqslant n}\{T_i\}$. Working through the computations, one obtains $L_{2,n}$ (observed) $= 2.2941 \times 10^{-5}$. Further, using $\kappa_S = \int_{-1}^{1} K(u)^2 du = \frac{3}{5}$, $\kappa^* = \int_{-2}^{2} K * K(u)^2 du = 0.4338$, and $\zeta_n = n_0/n = 21/41$, one can compute values of $\tilde{m}_1$ and $\tilde{v}_1$. Finally, by (4.1), the p-value of the goodness-of-fit test can be approximated by $1 - \Phi\ [n\hat{b}_{CV}^{1/2}\{L_{2,n}$ (observed) $- n^{-1}\hat{b}_{CV}^{-1}\tilde{m}_1\}\tilde{v}_1^{-1/2}] = 0.8296$. This shows that model (1.1) with $r(t) = t$ cannot be rejected. In this situation, the densities $g(t)$ and $h(t)$ can be respectively estimated by the semi-parametric density estimates $\tilde{g}(t)$ and $\tilde{h}(t)$. Here $\tilde{b}_{CV}(g)$ and $\tilde{b}_{CV}(h)$ for $\tilde{g}(t)$ and $\tilde{h}(t)$,

respectively, were determined similarly to $\hat{b}_{CV}$ on the same grid. Thus one obtains $\tilde{b}_{CV}(g) = 32.5862$ and $\tilde{b}_{CV}(h) = 20.5133$. The curves of $\tilde{g}(t)$ and $\tilde{h}(t)$ are plotted in Figure 4. The result shows that the social quotient scores of ordinary children with learning disabilities cluster around 39, while those of children with aphasia cluster around 45. Both $\tilde{g}(t)$ and $\tilde{h}(t)$ appear to be unimodal and symmetric. Further, their spreads seem not to differ substantially.

Example 5.2. Consider the data set given by Hosmer and Lemeshow (1989, p. 3). The purpose is to analyse the relationship between age and coronary heart disease. One hundred subjects participated in the study. Qin and Zhang (1997) concluded that model (1.1) with $r(t) = (t, t^2)$ was strongly supported by the data. Given this model, our semiparametric maximum likelihood estimate of $(\alpha, \beta_1, \beta_2)$ is $(\tilde{\alpha}, \tilde{\beta}_1, \tilde{\beta}_2) = (-3.9589, 0.0613, 0.00055)$. Using the same computation steps as in Example 5.1, one has $\hat{b}_{CV} = 11.5085$, $L_{2,n}$ (observed) $= \int_8^{80} \{\hat{g}(t) - \tilde{g}(t)\}^2 \, dt = 3.3449 \times 10^{-6}$, and the approximate p-value is 0.8731. Thus model (1.1) with $r(t) = (t, t^2)$ cannot be rejected. Further, $\tilde{b}_{CV}(g) = 11.4706$ and $\tilde{b}_{CV}(h) = 6.4759$ can be applied to compute $\tilde{g}(t)$ and $\tilde{h}(t)$. The curves of $\tilde{g}(t)$ and $\tilde{h}(t)$ are plotted in Figure 5. For the coronary heart disease population, the estimated density $\tilde{h}(t)$ appears to be bimodal, clustering around ages 49 and 58. On the other hand, the estimated density $\tilde{g}(t)$ for the non-diseased population appears to be unimodal and symmetric, clustering around age 36. This shows the tendency for individuals with no evidence of coronary heart disease to be younger than those with evidence of coronary heart disease.

6. Final remarks

In this paper we have proposed a new semiparametric density estimate $\tilde{g}(t)$ {or $\tilde{h}(t)$} based on a two-sample density ratio model. Our estimate is motivated by consideration of a class

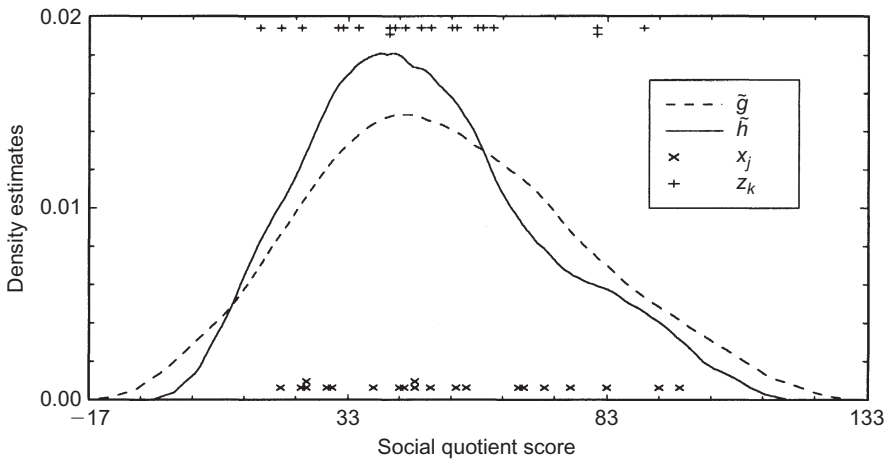

Figure 4. Child speech therapy example.

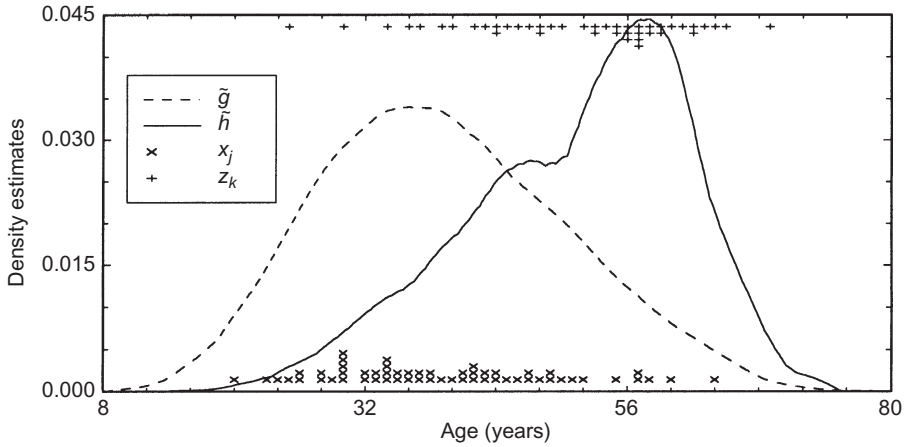

Figure 5. Coronary heart disease example.

of general density estimators. The proposed estimator is shown to be asymptotically unbiased for $g(t)$, like the ordinary kernel density estimator, but has the smallest asymptotic variance. From our formulation, it is seen that to define $\tilde{g}(t)$ we have used weights $\tilde{p}(T_i) = \{n_0 + n_1\tilde{w}(T_i)\}^{-1}$, with $\tilde{w}(T_i) = \exp\{\tilde{\alpha} + r(t)\tilde{\beta}\}$, for all $i = 1, \ldots, n$. A referee suggests using a natural alternative estimate $g^*(t)$ with weights $p^*(T_i) = 1$, for $i = 1, \ldots, n_0$, and weights $p^*(T_i) = \tilde{w}(T_i)^{-1}$, for $i = n_0 + 1, \ldots, n_0 + n_1$. Tedious computations show that under model (1.1) both $\tilde{g}(t)$ and $g^*(t)$ have the same asymptotic bias, but the asymptotic variance of $g^*(t)$ is $n^{-1}b^{-1}\{\zeta + (1-\zeta)/w(t)\}g(t)\kappa_S + o(n^{-1}b^{-1})$. By the Cauchy–Schwarz inequality, this asymptotic variance is larger than that of $\tilde{g}(t)$ unless $w(t) = 1$ or $\zeta = 1$. However, our unreported simulation results show that under minor misspecification of the model ($|\beta_2| \leqslant 0.3$), as discussed in Section 5.1, the differences between the averages and standard deviations of ISE($\tilde{b}_M$; $\tilde{g}$) and ISE(b_M^*; g^*) are not significant. The same conclusion also holds for ISE($\tilde{b}_{CV}$; $\tilde{g}$) and ISE(b_{CV}^*; g^*). The definitions of b_M^* and b_{CV}^* are the same as those given in Section 5.1.

Acknowledgement

The authors thank the editor, an associate editor, and two referees for their kind suggestions which greatly improved the presentation of this paper. The research of the first author was supported by the Ministry of Education program for promoting academic excellence of universities under grant no. 91-H-FA07-1-4. The research of the second author was supported by the Ministry of Education program for improving fundamental science education.

Appendix

Proof of Theorem 3.1. The proofs of (3.2) and (3.4) can be found in Silverman (1986, Section 3.3), and hence are omitted. To verify (3.1) and (3.3), using the consistency of $(\tilde{\alpha}, \tilde{\beta})$ with (α, β) and applying the second-order Taylor expansion theorem to $\tilde{w}(T_i)$, we have

$$\tilde{p}(T_i) = p(T_i) - c_i\{\tilde{\alpha} - \alpha + r(T_i)(\tilde{\beta} - \beta)\} + O[p(T_i)\{\tilde{\alpha} - \alpha + r(T_i)(\tilde{\beta} - \beta)\}^2],$$

where $c_i = p(T_i)^2 n_1 w(T_i)$ and the coefficient corresponding to the O term is asymptotically uniformly bounded over the subindex i. Using this result, $\tilde{g}(t)$ can be decomposed into

$$\tilde{g}(t) = S_1(t) + \mu_2(t)(\tilde{\alpha} - \alpha) + \mu_3(t)(\tilde{\beta} - \beta) + S^{\#}(t), \tag{A.1}$$

where

$$S_1(t) = \sum_{i=1}^{n} p(T_i) K_b(t - T_i),$$

$$S^{\#}(t) = \{S_2(t) - \mu_2(t)\}(\tilde{\alpha} - \alpha) + \{S_3(t) - \mu_3(t)\}(\tilde{\beta} - \beta) + S_4(t),$$

$$S_2(t) = \sum_{i=1}^{n} c_i K_b(t - T_i), \qquad \mu_2(t) = \mathrm{E}\{S_2(t)\},$$

$$S_3(t) = \sum_{i=1}^{n} c_i K_b(t - T_i) r(T_i), \qquad \mu_3(t) = \mathrm{E}\{S_3(t)\},$$

$$S_4(t) = O\left[\sum_{i=1}^{n} p(T_i) K_b(t - T_i)\{\tilde{\alpha} - \alpha + r(T_i)(\tilde{\beta} - \beta)\}^2\right].$$

Using (B1)–(B4), a straightforward calculation gives the following asymptotic results:

$$\mathrm{E}\{S_1(t)\} = g(t) + \tfrac{1}{2} b^2 \kappa_2 g^{(2)}(t) + O(b^3), \tag{A.2}$$

$$\mathrm{var}\{S_1(t)\} = n^{-1} b^{-1} D(t)^{-1} g(t) \kappa_S + o(n^{-1} b^{-1}), \tag{A.3}$$

$$\mathrm{E}\{S_2(t)\} = (1 - \zeta) g(t) w(t) D(t)^{-1} + O(b^2), \tag{A.4}$$

$$\mathrm{var}\{S_2(t)\} = n^{-1} b^{-1} \kappa_S g(t)\{(1 - \zeta) w(t)\}^2 D(t)^{-3} + o(n^{-1} b^{-1}), \tag{A.5}$$

$$\mathrm{E}\{S_3(t)\} = (1 - \zeta) g(t) w(t) D(t)^{-1} r(t) + O(b^2), \tag{A.6}$$

$$\mathrm{var}\{S_3(t)\} = n^{-1} b^{-1} \kappa_S g(t)\{(1 - \zeta) w(t)\}^2 D(t)^{-3} r(t)^{\mathrm{T}} r(t) + o(n^{-1} b^{-1}), \tag{A.7}$$

$$\mathrm{E}\{S_4(t)\} = O(n^{-1}), \qquad \mathrm{E}\{S_4(t)^2\} = O(n^{-2}), \tag{A.8}$$

for $t \in \mathbb{R}$. Using (A.1)–(A.8), $\mathrm{E}(\tilde{\alpha} - \alpha) = o(n^{-1/2})$, $\mathrm{var}(\tilde{\alpha}) = O(n^{-1})$, $\mathrm{E}(\tilde{\beta} - \beta) = o(n^{-1/2})$, $\mathrm{var}(\tilde{\beta}) = O(n^{-1})$, the Cauchy–Schwarz inequality, and (B.4), equations (3.1) and (3.3) follow. The proof of Theorem 3.1 is complete. □

Proof of Theorem 4.1. Following essentially the same proof of Theorem 3.1 and replacing

(α, β) with (α^*, β^*), by means of a straightforward calculation, the asymptotic bias and variance of $\tilde{g}(t)$ in Theorem 4.1 follow. The proof of Theorem 4.1 is complete. □

Proof of Theorem 4.2. We begin by proving (4.1). Using (A.1), $L_{2,n}$ can be decomposed into

$$L_{2,n} = \int \{\mu_0(t) - \mu_1(t) + \psi_0(t) - \psi_1(t) - S^*(t)\}^2 \, \mathrm{d}t = I_1 + I_2 + I_3 + I_4 + I_5 + I_6.$$

Here

$$\mu_0(t) = \mathrm{E}\{\hat{g}(t)\}, \qquad \mu_1(t) = \mathrm{E}\{S_1(t)\},$$
$$S^*(t) = \mu_2(t)(\tilde{\alpha} - \alpha) + \mu_3(t)(\tilde{\beta} - \beta) + S^{\#}(t),$$
$$\psi_0(t) = \sum_{j=1}^{n_0} K_{0j}(t), \qquad \psi_1(t) = \sum_{k=1}^{n_1} K_{1k}(t),$$
$$K_{0j}(t) = \{n_0^{-1} - p(X_j)\}K_b(t - X_j) - \mathrm{E}[\{n_0^{-1} - p(X_j)\}K_b(t - X_j)],$$
$$K_{1k}(t) = p(Z_k)K_b(t - Z_k) - \mathrm{E}\{p(Z_k)K_b(t - Z_k)\},$$
$$I_1 = \int \{\mu_0(t) - \mu_1(t)\}^2 \, \mathrm{d}t, \qquad I_2 = \int \{\psi_0(t) - \psi_1(t)\}^2 \, \mathrm{d}t,$$
$$I_3 = 2\int \{\mu_0(t) - \mu_1(t)\}\{\psi_0(t) - \psi_1(t)\} \, \mathrm{d}t, \qquad I_4 = \int S^*(t)^2 \, \mathrm{d}t,$$
$$I_5 = -2\int \{\mu_0(t) - \mu_1(t)\}S^*(t) \, \mathrm{d}t, \qquad I_6 = -2\int \{\psi_0(t) - \psi_1(t)\}S^*(t) \, \mathrm{d}t.$$

Using (B1)–(B7), (A.2)–(A.8), $n^{-1}b^{-6} = o(1)$, and the central limit theorem for quadratic forms in de Jong (1987), by means of a straightforward calculation, (4.1) follows by showing the following asymptotic results:

$$I_1 = O(b^6), \qquad nb^{1/2}(I_2 - n^{-1}b^{-1}m_1) \Rightarrow N(0, v_1),$$
$$\mathrm{E}(I_3) = 0, \qquad \mathrm{E}(I_4) = O(n^{-1}), \qquad \mathrm{E}(I_5) = o(n^{-1/2}b^3), \qquad \mathrm{E}(I_6) = O(n^{-1}),$$
$$\mathrm{E}(I_3^2) = O(n^{-1}b^6), \qquad \mathrm{E}(I_4^2) = O(n^{-2}), \qquad \mathrm{E}(I_5^2) = O(n^{-1}b^6), \qquad \mathrm{E}(I_6^2) = O(n^{-2}).$$

We now give the proof of (4.2). Using (A.1), $L_{2,n}$ can be decomposed into

$$L_{2,n} = \int \{\mu_0(t) - \mu_1(t) + S^{\&}(t) - S^{\#}(t)\}^2 \, \mathrm{d}t = J_1 + J_2 + J_3 + J_4 + J_5 + J_6.$$

Here

$$S^{\&}(t) = \psi_0(t) - \psi_1(t) - \mu_2(t)(\tilde{\alpha} - \alpha) - \mu_3(t)(\tilde{\beta} - \beta),$$
$$J_1 = \int \{\mu_0(t) - \mu_1(t)\}^2 \, \mathrm{d}t, \qquad J_2 = \int S^{\&}(t)^2 \, \mathrm{d}t,$$
$$J_3 = 2\int \{\mu_0(t) - \mu_1(t)\}S^{\&}(t) \, \mathrm{d}t, \qquad J_4 = \int S^{\#}(t)^2 \, \mathrm{d}t,$$

$$J_5 = -2\int\{\mu_0(t) - \mu_1(t)\}S^{\#}(t)\,\mathrm{d}t, \qquad J_6 = -2\int S^{\&}(t)S^{\#}(t)\,\mathrm{d}t.$$

Following essentially the proof of (4.1), replacing (α, β) with (α^*, β^*), and using (B1)–(B7), $n^{-1}b^{-6} = o(1)$, the central limit theorem, and Lemma 1 of Qin and Zhang (1997), by means of a straightforward calculation, (4.2) follows by showing the following asymptotic results:

$$J_1 = m_2 + b^2 m_3 + O(b^3), \qquad n^{1/2}J_3 \Rightarrow N(0, v_2),$$

$$\mathrm{E}(J_2) = O(n^{-1}b^{-1}), \quad \mathrm{E}(J_4) = O(n^{-1}), \quad \mathrm{E}(J_5) = O(n^{-1}b^{-1}), \quad \mathrm{E}(J_6) = O(n^{-1}),$$

$$\mathrm{E}(J_2^2) = O(n^{-2}b^{-2}), \quad \mathrm{E}(J_4^2) = O(n^{-2}), \quad \mathrm{E}(J_5^2) = O(n^{-2}b^{-2}), \quad \mathrm{E}(J_6^2) = O(n^{-2}).$$

The proof of Theorem 4.2 is complete. □

References

Cheng, K.F. and Chen, L.C. (2003) Testing goodness-of-fit of a logistic regression model with case–control data. *J. Statist. Plann. Inference*. To appear.

de Jong, P. (1987) A central limit theorem for generalized quadratic forms. *Probab. Theory Related Fields*, **75**, 261–277.

Efron, B. and Tibshirani, R. (1996) Using specially designed exponential families for density estimation. *Ann. Statist*., **24**, 2431–2461.

Epanechnikov, V.A. (1969) Nonparametric estimation of a multidimensional probability density. *Theory Probab. Appl*., **14**, 153–158.

Fokianos, K. (2002) Merging information for semiparametric density estimation. Technical report, Department of Mathematics and Statistics, University of Cyprus.

Fokianos, K., Kedem, B., Qin, J. and Short, D.A. (2001) A semiparametric approach to the one-way layout. *Technometrics*, **43**, 56–64.

Glovsky, L. and Rigrodsky, S. (1964) A developmental analysis of mentally deficient children with early histories of aphasis. *Training School Bull*., **61**, 76–96.

Hjort, N.L. and Glad, I.K. (1995) Nonparametric density estimation with a parametric start. *Ann. Statist*., **23**, 882–904.

Hosmer, D.J. and Lemeshow, S. (1989) *Applied Logistic Regression*. New York: Wiley.

Jones, M.C. (1991) Kernel density estimation for length biased data. *Biometrika*, **78**, 511–519.

Marron, J.S. and Wand, M.P. (1992) Exact mean integrated square error. *Ann. Statist*., **20**, 712–736.

Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case–control studies. *Biometrika*, **66**, 403–411.

Qin, J. (1998) Inferences for case–control and semiparametric two-sample density ratio model. *Biometrika*, **85**, 619–630.

Qin, J. and Zhang, B. (1997) A goodness-of-fit test for logistic regression models based on case–control data. *Biometrika*, **84**, 609–618.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. New York: Chapman & Hall.

Vardi, Y. (1982) Nonparametric estimation in the presence of length bias. *Ann. Statist*., **10**, 616–620.

Vardi, Y. (1985) Empirical distributions in selection bias models. *Ann. Statist*., **13**, 178–203.

White, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–16.

Zhang, B. (1999) A chi-squared goodness-of-fit test for logistic regression models based on case–control data. *Biometrika*, **86**, 531–539.

Zhang, B. (2000) M-estimation under a two-sample semiparametric model. *Scand. J. Statist*., **27**, 263–280.

Zhang, B. (2001) An information matrix test for logistic regression models based on case–control data. *Biometrika*, **88**, 921–932.

Zhao, L.P., Kristal, A.R. and White, E. (1996) Estimating relative risk functions in case–control studies using a nonparametric logistic regression. *Amer. J. Epidemiology*, **144**, 598–609.

Received November 2002 and revised January 2004