J. R. Statist. Soc. B (2004)
66, *Part* 4, *pp.* 941–958

Merging information for semiparametric density estimation

Konstantinos Fokianos

University of Cyprus, Nicosia, Cyprus

[Received December 2002. Final revision March 2004]

Summary. The density ratio model specifies that the likelihood ratio of $m-1$ probability density functions with respect to the mth is of known parametric form without reference to any parametric model. We study the semiparametric inference problem that is related to the density ratio model by appealing to the methodology of empirical likelihood. The combined data from all the samples leads to more efficient kernel density estimators for the unknown distributions. We adopt variants of well-established techniques to choose the smoothing parameter for the density estimators proposed.

Keywords: Bandwidth; Biased sampling; Discrete choice models; Empirical likelihood; Kernel estimator; Retrospective sampling

1. Introduction

Suppose that y denotes a categorical random variable with m categories, and x is a p-dimensional vector of covariates. Then, the so-called multinomial logits model is given by

$$P(y=i|x)=\frac{\exp(\alpha_i^*+x'\beta_i)}{\sum_{k=1}^{m}\exp(\alpha_k^*+x'\beta_k)}, \qquad i=1,\ldots,m. \tag{1}$$

$\alpha_m^*=\beta_m=0$ for identifiability, α_i^* is a scalar parameter, β_i is a $p\times 1$ vector of parameters, for $i=1,\ldots,m-1$, and the marginal distribution of x is left completely unspecified. Model (1) is one of the most popular choices for nominal data analysis and its application is widespread, especially in econometrics and biostatistics. See, for instance, Agresti (1990), chapter 9, and Fahrmeir and Tutz (2001), chapter 3.

Let $\pi_i=P(y=i)$ for $i=1,\ldots,m$ and suppose that there are m independent retrospective samples of sizes $n_1,\ldots,n_m$ acquired from the population with $y=i, i=1,\ldots,m$, respectively. Denote the observed data by x_{ij}, $j=1,\ldots,n_i$, $i=1,\ldots,m$, and assume that the conditional distribution of x given y has density $g_i(x)\equiv g(x|y=i)=\mathrm{d}G_i(x), i=1,\ldots,m$. A straightforward application of Bayes theorem shows that

$$g(x|y=i)=\frac{P(y=i|x)\ f(x)}{\pi_i}.$$

Recalling (1), we obtain the so-called density ratio model

Address for correspondence: Konstantinos Fokianos, Department of Mathematics and Statistics, University of Cyprus, PO Box 20537, 1678 Nicosia, Cyprus.
E-mail: fokianos@ucy.ac.cy

 1369–7412/04/66941

$$\frac{g(x|y=i)}{g(x|y=m)} = \frac{\pi_m}{\pi_i}\exp(\alpha_i^* + x'\beta_i) = \exp(\alpha_i + x'\beta_i), \qquad i = 1, \ldots, m-1, \tag{2}$$

where $\alpha_i = \alpha_i^* + \log(\pi_m/\pi_i)$ for $i = 1, \ldots, m-1$. Clearly when $\beta_i = 0$ then $\alpha_i = 0$. In other words, all the m density functions are assumed unknown but are related, however, through an exponential tilt—or distortion—which determines the difference between them. Model (2) is quite general and includes examples such as the exponential and partial exponential families of distributions. It has been studied in detail by Fokianos *et al.* (2001) who showed how the combined data from all the m samples can be used in the semiparametric large sample problem of estimating each distortion and the reference distribution, and testing the hypothesis that all m distributions are identical. It turns out that this approach generalizes the classical normal-based one-way analysis of variance in the sense that it obviates the need for a completely specified parametric model.

It is worthwhile to mention that expression (2) can be viewed as a biased sampling model with weights depending on parameters. Inference regarding biased sampling models has been discussed by Vardi (1982, 1985), Gill *et al.* (1988) and Bickel *et al.* (1998)—in the case of completely known weight functions—whereas Qin and Zhang (1997), Qin (1998), Gilbert *et al.* (1999) and Gilbert (2000) considered weight functions unknown up to a parameter. Furthermore, model (2) has been suggested by Efron and Tibshirani (1996) for density estimation—a topic that we touch on later—and for logistic discrimination by Anderson (1972, 1979).

A generalization of model (2) is

$$g_i(x) = w(x, \theta_i)\, g_m(x), \qquad i = 1, \ldots, q, \tag{3}$$

where the densities $g_i(x)$, $i = 1, \ldots, m-1$, are not specified and θ_i is a vector of parameters with finite dimension equal to d, for $i = 1, \ldots, m-1$. We assume throughout the paper that w is a *known positive function*—a stipulation which is quite restrictive in applications. Related work on the topic of estimating the weight function $w(\cdot;\cdot)$ includes Sun and Woodroofe (1997), who proposed a nonparametric method assuming that $w(\cdot;\cdot)$ is a monotone density function, and Fokianos (2003) who considered exponential weight functions, such as model (2), whose form is determined by a Box–Cox transformation. The choice of the reference distribution for model (3) is quite arbitrary. It turns out that, for the particular case of model (2), the choice of g_m does not affect the inferential results since the difference of the slopes remains constant—a property that characterizes the multinomial logits models and is usually described as independence of irrelevant alternatives.

Inference for the case $m = 2$ in connection with general binary response regression models subject to retrospective sampling plans under model (3) has been studied by Qin (1998). The aim of this contribution is to extend those results to categorical response regression models under retrospective sampling plans and to apply them to kernel-density-based estimation. In other words, initially we study the large sample behaviour of the finite dimensional parameters in model (3)—see Section 2. The asymptotic theory generalizes the theory of both Qin (1998) and Fokianos *et al.* (2001) in the sense that it covers the more general case of m samples with weights not of the exponential form necessarily. Furthermore, the asymptotic theory of Section 2 parallels the work of Gilbert (2000) who obtained large sample results for biased sampling models with weights depending on an unknown finite dimensional parameter based on a two-step estimation procedure. Section 3, in connection with the theory of Section 2, puts forward kernel density estimators of the unknown probability density functions. It turns out that the estimators which are based on the combined sample have smaller asymptotic mean-square error than the

traditional nonparametric density estimators. A limited simulation study and an application of the methodology to real data are presented in Section 4.

Motivation for this study is initiated by the diverse applications of the density ratio model. To mention a few, Gilbert *et al.* (1999) used model (2) for the analysis of a large scale preventive human immunodeficiency virus vaccine trial to asses differences between vaccine and placebo groups. Hence a density estimate for both groups provides an additional insight into the problem. Likewise, Fokianos *et al.* (1998) considered the problem of combining data from two instruments which measure rainfall rates. Although the problem of density estimation was addressed vaguely by considering a crude histogram, a kernel density estimator will be in general more preferable. For further applications of the density ratio model, see Fokianos *et al.* (2001) or Qin *et al.* (2002).

2. Inference

Consider the m samples with corresponding densities that satisfy equations (3), i.e. consider $q = m - 1$ weight functions $w(x, \theta_i)$, known up to a parameter, let $n = \Sigma_{i=1}^{m} n_i$ be the total sample size and consider the empirical likelihood (Owen, 1988, 1990) based on the pooled data $\{x_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, m\}$,

$$\begin{aligned} L(\theta, G_m) &= \left\{ \prod_{j=1}^{n_1} p_{1j}\, w(x_{1j}, \theta_1) \right\} \left\{ \prod_{j=1}^{n_2} p_{2j}\, w(x_{2j}, \theta_2) \right\} \ldots \prod_{j=1}^{n_m} p_{mj} \\ &= \left(\prod_{i=1}^{m} \prod_{j=1}^{n_i} p_{ij} \right) \prod_{i=1}^{q} \prod_{j=1}^{n_i} w(x_{ij}, \theta_i) \end{aligned} \tag{4}$$

with $p_{ij} = \mathrm{d}G_m(x_{ij})$ and $\theta = (\theta_1', \ldots, \theta_q')'$ a vector of dimension $p = qd$. The log-likelihood is given by

$$l = \log(L) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log(p_{ij}) + \sum_{i=1}^{q} \sum_{j=1}^{n_i} \log\{w(x_{ij}, \theta_i)\}. \tag{5}$$

Maximization of equation (5) is carried out by following a profiling procedure (Qin and Lawless, 1994) whereby first we express each p_{ij} in terms of some finite dimensional parameters and then we substitute them back into the likelihood to produce a parametric function. To be more specific, observe that, when θ is fixed, equation (4) is maximized by maximizing only the product term $\Pi_{i=1}^{m} \Pi_{j=1}^{n_i} p_{ij}$, subject to the constraints

$$p_{ij} \geqslant 0,$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} p_{ij} = 1,$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} p_{ij} \{w(x_{ij}, \theta_k) - 1\} = 0, \qquad k = 1, \ldots, q.$$

The above constraints mimic their population counterpart, i.e. $\int \mathrm{d}G_m(x) = 1$ and $\int w(x, \theta_k)\, \mathrm{d}G_m(x) = 1$, for $k = 1, \ldots, q$.

Following Qin and Lawless (1994), we obtain

$$p_{ij} = \frac{1}{n} \frac{1}{1 + \sum_{k=1}^{q} \mu_k \{w(x_{ij}, \theta_k) - 1\}},$$

with μ_k, $k=1,\ldots,q$, Lagrange multipliers determined by the equations

$$-\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{w(x_{ij},\theta_k)-1}{1+\sum_{k=1}^{q}\mu_k\{w(x_{ij},\theta_k)-1\}}=0, \qquad k=1,\ldots,q.$$

It turns out that the vector of Lagrange multipliers $\mu=(\mu_1,\ldots,\mu_k)'$ is a continuously differentiable function of the parameter θ. Hence equation (5) becomes

$$l\{\theta,\mu(\theta)\}=-\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log\left[1+\sum_{k=1}^{q}\mu_k(\theta)\{w(x_{ij},\theta_k)-1\}\right]+\sum_{i=1}^{q}\sum_{j=1}^{n_i}\log\{w(x_{ij},\theta_i)\}-n\log(n). \tag{6}$$

Denote by $h(x,\theta)$ the vector function $(w(x,\theta_1)-1,\ldots,w(x,\theta_q)-1)'$. Then the following lemma states that there is a maximum in a small neighbourhood of the true parameter with probability approaching 1; see Qin and Lawless (1994), lemma 1. In the following $E_m(\cdot)$ and $\mathrm{var}_m(\cdot)$ denote expectation and variance with respect to G_m.

Lemma 1. Assume that

(a) $E_m\{h(x,\theta_0)\,h'(x,\theta_0)\}$ is positive definite,
(b) $\partial h(x,\theta)/\partial\theta$ is continuous in a neighbourhood of the true value θ_0,
(c) $\|\partial h(x,\theta)/\partial\theta\|$ and $\|h(x,\theta)\|^3$ are bounded by some integrable function $H(x)$ with respect to $G_m(x)$ in this neighbourhood and
(d) the rank of $E\{\partial h(x,\theta)/\partial\theta\}$ is qd.

Then, as $n\to\infty$, the profiled log-likelihood (6) attains its maximum value at some point $\hat\theta$ in the interior of the ball $\|\theta-\theta_0\|\leqslant n^{-1/3}$ and $\hat\theta$ and $\hat\mu=\mu(\hat\theta)$ satisfy the following system of estimating equations:

$$\frac{\partial l(\theta,\mu)}{\partial\theta_l}=-\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{\mu_l\,\partial w(x_{ij},\theta_l)/\partial\theta_l}{1+\sum_{k=1}^{q}\mu_k\{w(x_{ij},\theta_k)-1\}}+\sum_{j=1}^{n_l}\frac{\partial[\log\{w(x_{lj},\theta_l)\}]}{\partial\theta_l}=0, \qquad l=1,\ldots,q,$$

$$\frac{\partial l(\theta,\mu)}{\partial\mu_l}=-\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{w(x_{ij},\theta_l)-1}{1+\sum_{k=1}^{q}\mu_k\{w(x_{ij},\theta_k)-1\}}=0, \qquad l=1,\ldots,q.$$

Lemma 1 implies the existence of the maximum empirical likelihood estimators. In other words, set

$$\hat p_{ij}=\frac{1}{n}\frac{1}{1+\sum_{k=1}^{q}\hat\mu_k\{w(x_{ij};\hat\theta_k)-1\}}$$

to obtain the maximum likelihood estimator of G_m

$$\begin{aligned}\hat G_m(x)&=\sum_{i=1}^{m}\sum_{j=1}^{n_i}\hat p_{ij}\,I(x_{ij}\leqslant x)\\&=\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{1}{1+\sum_{k=1}^{q}\hat\mu_k\{w(x_{ij},\hat\theta_k)-1\}}I(x_{ij}\leqslant x),\end{aligned} \tag{7}$$

where I denotes the indicator function. Similarly, the maximum likelihood estimators of $G_l(x)$, say $\hat{G}_l(x)$, are obtained by

$$\begin{aligned}\hat{G}_l(x) &= \sum_{i=1}^{m}\sum_{j=1}^{n_i} \hat{p}_{ij}\, w(x_{ij},\hat{\theta}_l)\, I(x_{ij}\leqslant x)\\ &= \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{w(x_{ij},\hat{\theta}_l)}{1+\sum_{k=1}^{q}\hat{\mu}_k\{w(x_{ij},\hat{\theta}_k)-1\}} I(x_{ij}\leqslant x),\end{aligned} \tag{8}$$

for $l=1,\ldots,q$. The large sample performance of the finite dimensional estimators is assessed as follows. Set $\zeta_n=(\zeta_{1n},\ldots,\zeta_{qn})$ with $\zeta_{ln}=n_l/n$, for $l=1,\ldots,q$, and assume that $\zeta_{ln}\to\zeta_l$, as $n\to\infty$, for all l. Then $\zeta_n\to\zeta$ and by denoting by θ the true value of the parameter we have the following theorem. Its proof is given in Appendix A.

Theorem 1. In addition to the conditions of lemma 1 assume that

(a) $\partial^2 h(x,\theta)/\partial\theta\,\partial\theta'$ is continuous in a neighbourhood of the true parameter and
(b) there is an integrable function with respect to G_m, say $H(x)$, which bounds $\|\partial^2 h(x,\theta)/\partial\theta\,\partial\theta'\|$.

Then

$$n^{1/2}\begin{pmatrix}\hat{\theta}-\theta\\ \hat{\mu}-\zeta\end{pmatrix}\to N(\mathbf{0},\mathbf{W})$$

in distribution, as $n\to\infty$. The asymptotic covariance matrix $\mathbf{W}$ is defined in Appendix A by equation (20).

Remark 1. In general, the parameter vector θ and the base-line distribution function G_m are not identifiable. However, for model (3), theorem 2 of Gilbert *et al.* (1999) guarantees that the density ratio model is identifiable provided that, for all $\theta,\tilde{\theta}$ with $\theta\neq\tilde{\theta}$, there is an $i\in 1,2,\ldots,m-1$ such that $w(x,\theta_i)$ and $w(x,\tilde{\theta}_i)$ are linearly independent as functions of x.

Remark 2. We point out that the profile log-likelihood (6) for the finite dimensional parameters in a semiparametric setting behaves as a parametric log-likelihood—a fact which has been studied by Murphy and van der Vaart (1997, 2000). Furthermore, when model (2) holds, equation (6) transforms to

$$l_1=-\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log\left\{1+\sum_{k=1}^{q}\rho_k\exp(\alpha_k+x_{ij}'\beta_k)\right\}+\sum_{i=1}^{q}\sum_{j=1}^{n_i}(\alpha_i+x_{ij}'\beta_i)-n\log(n_m), \tag{9}$$

with $\rho_l=n_l/n_m$, $l=1,\ldots,q$. For more, see Fokianos *et al.* (2001).

In this section, the problem of inference for the density ratio model (3) was addressed by following a profiling procedure which enables us to obtain the nonparametric estimators (7) and (8) for the unknown distribution functions. In addition, the score estimating equations—see lemma 1—for the finite dimensional parameters were derived. The question of density estimation is studied next.

3. Combined semiparametric density estimators

The topic of kernel density estimation is the subject of several texts like those of Silverman (1986), Scott (1992) and Wand and Jones (1995). Important contributions on the topic of

semiparametric kernel density estimation include the work by Hjort and Jones (1996), who discussed an estimator based on the concept of local likelihood, and the suggestion of Hjort and Glad (1995) for a semiparametric technique related to kernel density estimation. In addition, Lloyd and Jones (2000), Wu (1996) and Jones (1991) investigated theory for density estimation from biased sampling data. The work by Efron and Tibshirani (1996) developed a new method for estimating an exponential family of probability densities

$$h(x;\theta) = h_0(x)\exp\{\alpha_0 + \alpha_1 s(x)\},$$

based on a random sample, say $x_1, \ldots, x_n$. Here $h_0(x)$ is a carrier density whereas $s(x)$ is a vector of sufficient statistics and $\alpha = (\alpha_0, \alpha_1)'$ denotes a parameter vector. The method is divided into two parts. First, estimation of $h_0(x)$ is achieved by a kernel density estimator, say $\hat{h}_0(x)$. Then, maximization of

$$\prod_{i=1}^{n} \hat{h}_0(x_i)\exp\{\alpha_0 + \alpha_1 s(x_i)\}$$

yields an estimate of α. The above approach is based on Poisson regression methods and can be generalized to m independent samples of observations. It turns out that the exponential family of probability densities coincides with model (2). However, the approach of this work is quite different. First, the enlarged model (3) is considered which is based on m available random samples. In addition, the method is initialized by first estimating the parameters μ and θ—recall lemma 1—and then calculating the maximum likelihood estimators of all the unknown distribution functions (see equations (7) and (8)). Given this inference output, we suggest smoothing the increments of $\hat{G}_i$, $i = 1, \ldots, m$, to obtain new kernel density estimators. In principle, the methodology proposed can be readily generalized to multivariate density estimation. However, for ease of presentation we focus only on univariate measurements.

Recall that $\rho_i = n_i/n_m$, $i = 1, \ldots, m$, and set $w(x, \theta_i) = w_i(x)$ for $i = 1, \ldots, m$. In particular, $\rho_m \equiv 1$ and $w_m(x) \equiv 1$. Smoothing the increments of $\hat{G}_l$, for all l, amounts to the estimators

$$\hat{g}_l(x) = \frac{1}{h_n}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \hat{p}_{ij}\,\hat{w}_l(x_{ij})\,K\left(\frac{x - x_{ij}}{h_n}\right), \qquad l = 1, \ldots, m, \tag{10}$$

where h_n is a sequence of window widths such that $h_n \to 0$ as $n \to \infty$ and K is a kernel function that satisfies the requirements

(a) $\int K(t)\,\mathrm{d}t = 1$ and $\int |K(t)|\,\mathrm{d}t < \infty$,
(b) $\int t\,K(t)\,\mathrm{d}t = 0$ and $\int |t\,K(t)|\,\mathrm{d}t < \infty$ and
(c) $\int t^2\,K(t)\,\mathrm{d}t = k_2$ and $\int t^2|K(t)|\,\mathrm{d}t < \infty$.

By a straightforward calculation

$$\begin{aligned}\int \hat{g}_l(x)\,\mathrm{d}x &= \frac{1}{h_n}\int \sum_{i=1}^{m}\sum_{j=1}^{n_i} \hat{p}_{ij}\,\hat{w}_l(x_{ij})\,K\left(\frac{x - x_{ij}}{h_n}\right)\mathrm{d}x \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n_i} \hat{p}_{ij}\,\hat{w}_l(x_{ij}) = 1,\end{aligned}$$

for all l. In other words, equation (10) introduces a proper probability density function. Furthermore, equation (10) is a semiparametric density estimator since it depends on both the unknown distribution function and the parameters of the model.

Remark 3. It is illuminating to calculate the first and second moment of $\hat{g}_l(x)$ when model (2) holds. With a slight abuse of notation set $x'_{ij} = (x_{ij}, x_{ij}^2)$. Then, it can be shown that

$$\int x\, \hat{g}_l(x)\, \mathrm{d}x = \int x \left\{ \frac{1}{n_m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{\exp(\hat{\alpha}_l + x'_{ij}\hat{\beta}_l)}{1 + \sum_{k=1}^{q} \rho_k \exp(\hat{\alpha}_k + x'_{ij}\hat{\beta}_k)} \right\} \frac{1}{h_n} K\left(\frac{x - x_{ij}}{h_n}\right) \mathrm{d}x$$
$$= \frac{1}{n_l} \sum_{j=1}^{n_l} x_{lj},$$

i.e. the mean of the lth sample, a fact which follows from differentiation of equation (9). Furthermore, we obtain

$$\int x^2\, \hat{g}_l(x)\, \mathrm{d}x = \int x^2 \left\{ \frac{1}{n_m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{\exp(\hat{\alpha}_l + x'_{ij}\hat{\beta}_l)}{1 + \sum_{k=1}^{q} \rho_k \exp(\hat{\alpha}_k + x'_{ij}\hat{\beta}_k)} \right\} \frac{1}{h_n} K\left(\frac{x - x_{ij}}{h_n}\right) \mathrm{d}x$$
$$= \frac{1}{n_l} \sum_{j=1}^{n_l} x_{lj}^2 + h_n^2 k_2,$$

i.e. the second moment of the lth sample plus a small term. Inclusion of higher powers of x corresponds to matching higher order sample moments of the lth sample.

To study the statistical properties of the kernel density estimator (10), it is instructive to consider the random variable

$$\tilde{g}_l(x) = \frac{1}{n_l h_n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{w_l(x_{ij})}{\sum_{k=1}^{m} \rho_k\, w_k(x_{ij})} K\left(\frac{x - x_{ij}}{h_n}\right), \qquad l = 1, \ldots, m. \tag{11}$$

Then, the following lemma holds true—its proof is postponed until Appendix A.

Lemma 2. Assume that conditions (a)–(c) hold and suppose that $g_l(x)$ is continuous at x for every l. Then

$$\hat{g}_l(x) = \tilde{g}_l(x) + O_p(n^{-1/2}),$$

for every l, as $n \to \infty$ and $h_n \to 0$.

Lemma 2 facilitates computation of the asymptotic bias and variance of equation (10), using instead equation (11). Indeed, the bias of expression (11) is given by

$$E\{\tilde{g}_l(x)\} = \frac{1}{h_n} \sum_{i=1}^{m} \rho_i \int \frac{w_l(y)}{\sum_{k=1}^{m} \rho_k\, w_k(y)} K\left(\frac{x - y}{h_n}\right) w_i(y)\, g_m(y)\, \mathrm{d}y$$
$$= \frac{1}{h_n} \int K\left(\frac{x - y}{h_n}\right) g_l(y)\, \mathrm{d}y$$
$$= g_l(x) + \frac{1}{2} h_n^2\, g_l''(x) k_2 + o(h_n^2) \tag{12}$$

as $n \to \infty$ and $h_n \to 0$, and its variance is

$$\begin{aligned}\operatorname{var}\{\tilde{g}_l(x)\} &= \frac{1}{n_l h_n^2} \sum_{i=1}^{m} \rho_i \int \frac{w_l^2(y)}{\left\{ \sum_{k=1}^{m} \rho_k w_k(y) \right\}^2} K^2\left(\frac{x-y}{h_n}\right) g_i(y)\,\mathrm{d}y \\ &\quad - \frac{1}{n_l} \sum_{i=1}^{m} \rho_i \left\{ \int \frac{1}{h_n} \frac{1}{\sum_{k=1}^{m} \rho_k w_k(y)} K\left(\frac{x-y}{h_n}\right) g_i(y)\,\mathrm{d}y \right\}^2 \\ &= \frac{1}{n_l h_n} \frac{\rho_l w_l(x)\, g_l(x)}{\sum_{k=1}^{m} \rho_k w_k(x)} \int K^2(t)\,\mathrm{d}t + o\{(nh_n)^{-1}\}, \end{aligned} \tag{13}$$

with the additional requirement that $nh_n \to \infty$. A careful examination of the above equations reveals the common trade-off problem between random and systematic error. In other words, small values of h_n eliminate bias but introduce substantial variance as opposed to large values of the smoothing parameter which lead to smaller variance but increased bias.

Now fix an index l, and recall the traditional nonparametric density estimator

$$\bar{g}_l(x) = \frac{1}{n_l h_n} \sum_{j=1}^{n_l} K\left(\frac{x - x_{lj}}{h_n}\right), \tag{14}$$

i.e. the kernel estimate based on the lth sample. It is well known—see Silverman (1986), page 36—that, as $n \to \infty$, $h_n \to 0$ and $nh_n \to \infty$,

$$E\{\bar{g}_l(x)\} = g_l(x) + \frac{1}{2} h_n^2\, g_l''(x) k_2 + o(h_n^2),$$

$$\operatorname{var}\{\bar{g}_l(x)\} = \frac{g_l(x)}{n_l h_n} \int K^2(t)\,\mathrm{d}t + o\{(nh_n)^{-1}\}.$$

Equations (12) and (13) when compared with the above results show that the pooled data yield kernel density estimates with the same amount of bias but which are less variable for the same bandwidth selection. Indeed,

$$\frac{\rho_l w_l(x)}{\sum_{k=1}^{m} \rho_k w_k(x)} \leqslant 1,$$

since w_l has been assumed to be positive so that the leading term is less than or equal to 1. In fact, the above factor is strictly less than 1 unless all $n_i = 0$ for $i \neq l$.

The performance of the proposed estimator is based on the asymptotic mean integrated square error (AMISE), which is readily obtained by well-known arguments. The following proposition summarizes the main results and is given without proof.

Proposition 1. Suppose that conditions (a)–(c) hold and assume that $\int K^2(t)\,\mathrm{d}t < \infty$. If $g_l(x)$ is twice continuously differentiable, then, for every l,

(a) as $n \to \infty$, $h_n \to 0$ and $nh_n \to \infty$, we obtain

$$\operatorname{AMISE}\{\hat{g}_l(x)\} = \frac{1}{4} h_n^4 k_2^2 \int g_l''(x)^2\,\mathrm{d}x + \frac{1}{n_l h_n} \int \frac{\rho_l w_l(x)\, g_l(x)}{\sum_{k=1}^{m} \rho_k w_k(x)}\,\mathrm{d}x \int K^2(t)\,\mathrm{d}t,$$

(b) the asymptotically optimal bandwidth, which is found by minimizing AMISE$\{\hat{g}_l(x)\}$, is equal to

$$h_n^* = \left\{ \int \frac{\rho_l \, w_l(x) \, g_l(x)}{\sum_{k=1}^{m} \rho_k \, w_k(x)} \mathrm{d}x \right\}^{1/5} \left\{ \int K^2(t) \, \mathrm{d}t \right\}^{1/5} \left\{ \int g_l''(x)^2 \, \mathrm{d}x \right\}^{-1/5} k_2^{-2/5} \zeta_l^{-1/5} n^{-1/5},$$

(c) assigning h_n^* from (b) to (a), we obtain that the AMISE is equal to

$$\mathrm{AMISE}^*\{\hat{g}_l(x)\} = \frac{5}{4} \left\{ \int \frac{\rho_l \, w_l(x) \, g_l(x)}{\sum_{k=1}^{m} \rho_k \, w_k(x)} \mathrm{d}x \right\}^{4/5} \left\{ \int K^2(t) \, \mathrm{d}t \right\}^{4/5} \times \left\{ \int g_l''(x)^2 \, \mathrm{d}x \right\}^{1/5} k_2^{2/5} \zeta_l^{-4/5} n^{-4/5}.$$

The preceding proposition implies that the optimal bandwidth depends on the weight functions and on the integral $\int g_l''(x)^2 \, \mathrm{d}x$. Furthermore, it follows that the smoothing parameter converges to 0 at a very slow rate—a fact that holds true for the commonly used kernel density estimator (14).

The new semiparametric kernel density estimator reduces the AMISE when it is compared with that of the traditional kernel density estimator (14). This fact can be verified by recalling that

$$\mathrm{AMISE}\{\bar{g}_l(x)\} = \frac{1}{4} h_n^4 k_2^2 \int g_l''(x)^2 \, \mathrm{d}x + \frac{1}{n_l h_n} \int K^2(t) \, \mathrm{d}t,$$

as $n \to \infty$, $h_n \to 0$ and $nh_n \to \infty$. It follows that

$$\mathrm{AMISE}\{\hat{g}_l(x)\} \leqslant \mathrm{AMISE}\{\bar{g}_l(x)\}$$

since

$$\int \frac{\rho_l \, w_l(x) \, g_l(x)}{\sum_{k=1}^{m} \rho_k \, w_k(x)} \mathrm{d}x \leqslant 1,$$

for every l. Furthermore, $\bar{g}_l(x)$ obtains its optimal AMISE value as follows:

$$\mathrm{AMISE}^*\{\bar{g}_l(x)\} = \frac{5}{4} \left\{ \int K^2(t) \, \mathrm{d}t \right\}^{4/5} \left\{ \int g_l''(x)^2 \, \mathrm{d}x \right\}^{1/5} k_2^{2/5} \zeta_l^{-4/5} n^{-4/5},$$

for all l and under the same assumptions as those of proposition 1. Therefore, the asymptotic relative efficiency of $\bar{g}_l(\cdot)$ with respect to $\hat{g}_l(\cdot)$ is given by

$$\mathrm{eff}(\bar{g}_l; \hat{g}_l) \equiv \frac{\mathrm{AMISE}^*\{\hat{g}_l(x)\}}{\mathrm{AMISE}^*\{\bar{g}_l(x)\}} = \left\{ \int \frac{\rho_l \, w_l(x) \, g_l(x)}{\sum_{k=1}^{m} \rho_k \, w_k(x)} \mathrm{d}x \right\}^{4/5} \leqslant 1, \tag{15}$$

for every l. Thus, the density estimator proposed is more efficient than the traditional kernel density estimator unless $n_i = 0$ for $i \neq l$, i.e. only the lth sample is available. The point is demonstrated in Fig. 1 where data have been generated by log-normal distributions which satisfy trivially the density ratio model by inserting the vector function $(\log(x), \log^2(x))'$ in model (2). Fig. 1 shows that the empirically estimated asymptotic relative efficiency (15) for the first and

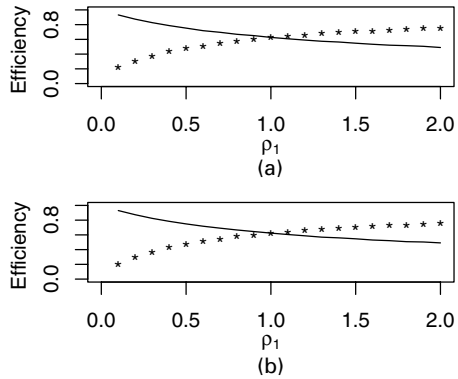

Fig. 1. Empirically estimated asymptotic relative efficiency (15) of the nonparametric kernel density estimator with respect to the combined kernel density estimator for a two-sample problem (the data were generated according to the log-normal distribution LN(μ, σ^2) with corresponding parameters $\mu_1 = 2$, $\mu_2 = 1$, $\sigma_1 = 1.50$ and $\sigma_2 = 2$: ———, asymptotic efficiency of the density estimator for the first sample; $*$, asymptotic efficiency of the density estimator for the second sample) (the results are based on 100 simulations): (a) sample size for the reference distribution $n_2 = 60$; (b) sample size for the reference distribution $n_2 = 100\,(n_1 = \rho_1 n_2)$

second sample decreases and increases respectively as $\rho_1 = n_1/n_2$ increases—a consequence of the fact that when ρ_1 increases there is more available data from the second population. Furthermore, when the sample sizes are both equal, then both semiparametric estimators obtain comparable asymptotic efficiencies.

In summary, we propose semiparametric density estimators (10) which are based on the combination of all available samples under the density ratio model. It was shown that the new estimators have the same amount of bias but they are less variable than the traditional density estimators (14). In addition, estimators which are based on the combination of data are more efficient than the standard kernel density estimators.

4. Applications

4.1. Simulations

A limited simulation study is reported to illustrate empirically the adequacy of the theoretical results with 100 runs and employing a standard Gaussian kernel. The results are based on data from normal distributions with both means and variances unequal. To be more specific, suppose that $X_{11}, \ldots, X_{1n_1}$ is a sample from a normal distribution with mean μ_1 and standard deviation σ_1 and let $X_{21}, \ldots, X_{2n_2}$ be another sample—independent of the first—from a normal distribution with mean μ_2 and standard deviation σ_2. It is a simple exercise to show that model (2) holds with the vector function $(x, x^2)'$ and the appropriate choice of parameters. Recall that $\hat{g}_l(x)$ and $\bar{g}_l(x)$ respectively refer to the semiparametric and nonparametric kernel density estimator respectively as defined by equation (10) and equation (14) respectively. The average AMISEs for both estimators based on 100 simulations and for various sample sizes have been tabulated in Table 1. The results show that in all cases considered the semiparametric kernel density estimator achieves a smaller AMISE than that of equation (14). In this work selection of the smoothing parameter was carried out by empirical estimation of the optimal value in proposition 1, part (b). More specifically, an initial value is chosen, say h_0, and the search for a fixed point value of h is achieved by iterating proposition 1, part (b). The complexity of this method is the estimation of the integral $\int g_l''(x)^2\,\mathrm{d}x$. The first factor of proposition 1, part (b), is consistently estimated by its empirical version. The results of Scott (1992), page 164, who

Table 1. Comparison of AMISEs for the semiparametric and nonparametric kernel density estimators for various sample sizes†

$AMISE\{\hat{g}_1(x)\}$	$AMISE\{\bar{g}_1(x)\}$	$AMISE\{\hat{g}_2(x)\}$	$AMISE\{\bar{g}_2(x)\}$	n_1	n_2
0.0122	0.0194	0.0167	0.0234	30	40
0.0104	0.0151	0.0172	0.0260	40	40
0.0051	0.0074	0.0173	0.0260	50	40
0.0047	0.0066	0.0138	0.0249	60	40
0.0106	0.0160	0.0188	0.0287	30	30
0.0056	0.0079	0.0184	0.0310	40	30
0.0053	0.0071	0.0166	0.0305	50	30
0.0081	0.0102	0.0147	0.0294	60	30

†The observations are drawn from the normal distribution with $\mu_1 = 1$, $\mu_2 = 0$, $\sigma_1 = 1.5$ and $\sigma_2 = 1$. The results are based on 100 simulations.

derived an expression for a normal kernel can be adapted to this situation as follows:

$$\int g_l''(x)^2 \, \mathrm{d}x = \frac{3}{8h^5\sqrt{\pi}} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{k=1}^{m} \sum_{s=1}^{n_k} \left(1 - \Delta_{ijks}^2 + \frac{1}{12}\Delta_{ijks}^4\right) \exp\left(-\frac{1}{4}\Delta_{ijks}^2\right) \times p_{ij}\, \hat{w}_l(x_{ij})\, p_{ks}\, \hat{w}_l(x_{ij}) \tag{16}$$

with $\Delta_{ijks} = (x_{ij} - x_{ks})/h$. It turns out that the fixed point algorithm converges independently of the choice of the starting value. Starting values were obtained by the S-PLUS functions that are provided in Venables and Ripley (1999). Similar results were obtained in situations with three samples, but they are not reported here.

4.2. Data analysis

The new method is illustrated with a real data set which can be accessed from `http://lib.stat.cmu.edu/DASL/Stories/FusionTime.html`. These data consist of results from an experiment in visual perception using random dot stereograms. The subject observes two images which appear to be composed entirely of random dots. However, they are constructed so that a three-dimensional image will be seen, if the images are viewed with a stereo viewer, causing the separate images to fuse.

An experiment was performed to determine whether knowledge of the form of the embedded image affected the time that is required for subjects to fuse the images. One group of subjects (group NV) received either no information or just verbal information about the shape of the embedded object. A second group (group VV) received both verbal information and visual information (e.g. a drawing of the object). The scientific question is whether there are differences between the mean time that is required to fuse the images for the two groups. Previous analyses indicated that there are significant differences after log-transforming the data. After discarding an outlier for the NV group, we obtain box plots of the raw data (Fig. 2) which clearly indicate that time values are positively skewed, and the longer fusion time in the NV group is accompanied by greater variability. Since the subjects in the VV group received more information, we postulate it as the reference sample. The box plots of the data lead us to apply several models which capture the skewed character of the data. For instance log-normal populations or gamma populations serve as reference examples. Table 2 gives negative log-likelihoods for a variety of models. The first row gives the negative log-likelihood value for the model

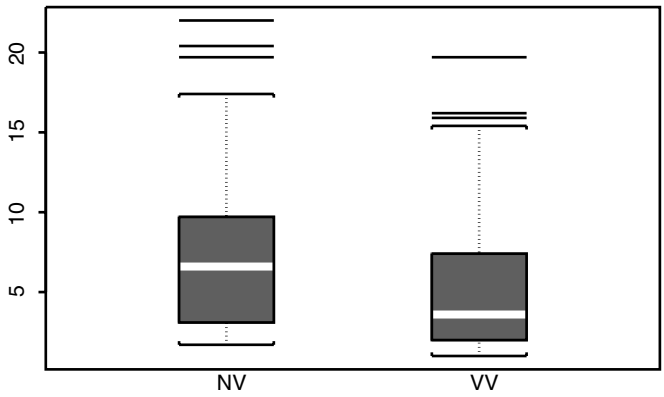

Fig. 2. Box plots of the raw data

Table 2. Fusion times: negative log-likelihood values up to a constant for various models

Model	*Negative log-likelihood*
$(1, \log(x))$	58.482
$(1, x)$	59.108
$(1, \log(x), x)$	58.359
$(1, x, x^2)$	58.690
$(1, \log(x), \log^2(x))$	58.252

$$\log\left\{\frac{g_1(x)}{g_2(x)}\right\} = \theta_1 + \theta_2 \log(x), \tag{17}$$

and so on. Note that $m = 2$, $w_1 = \exp\{\theta_1 + \theta_2 \log(x)\}$, $n_1 = 42$, $n_2 = 35$ and $\rho_1 = 1.2$.

The results illustrate that all the models that were considered are equivalent. Thus, model (17) can be applied to the data. The maximum likelihood estimators of θ_1 and θ_2 turn out to be $\hat{\theta}_1 = -0.985$ and $\hat{\theta}_2 = 0.623$. The standard error of $\hat{\theta}_1$ is 0.478 whereas the standard error of $\hat{\theta}_2$ is 0.303, which leads to the conclusion that there are differences between the two groups.

Turning to the problem of semiparametric density estimation for the VV group, selection of the smoothing parameter is carried out as described in Section 4.1. The proposed iterative method converges to $h = 0.50$ using a standard normal kernel. Both the nonparametric and the semiparametric density estimators are shown in Fig. 3 where both panels have been drawn by setting $h = 0.50$. The new estimator is less variable and smoother than the nonparametric estimator, which is based on the VV group alone. In addition, an estimate of the asymptotic relative efficiency can be obtained by recalling equation (15),

$$\left\{\int \frac{g_2(x)}{\sum_{k=1}^{2} \rho_k w_k(x)}\,\mathrm{d}x\right\}^{4/5} \approx 0.55;$$

i.e. the semiparametric density estimator reduces the AMISE almost by a factor of 2 by using the available information from the NV sample.

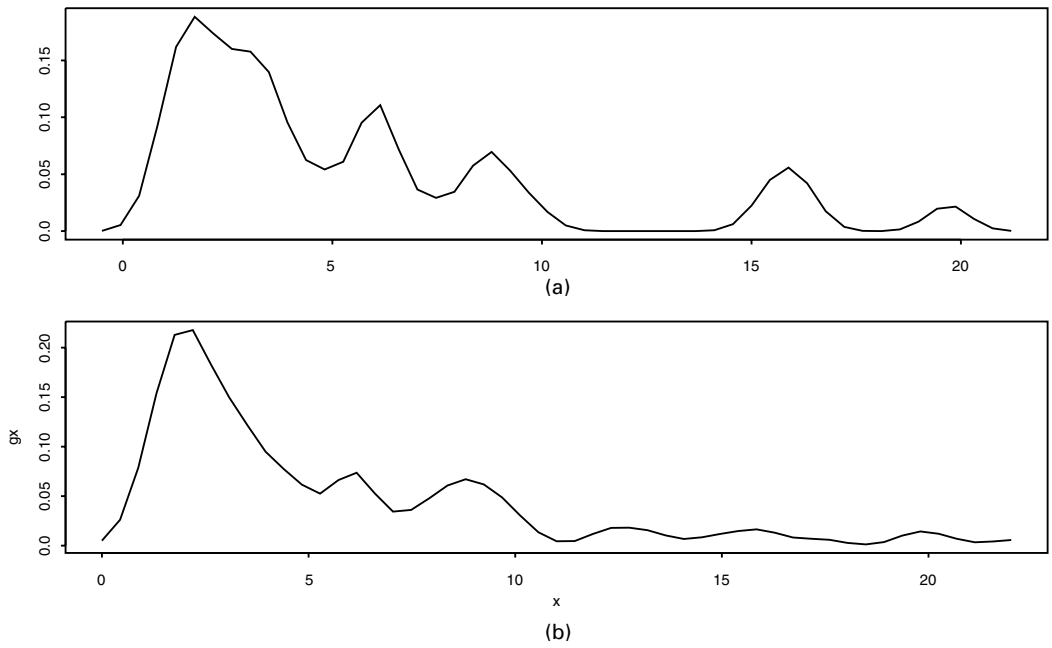

Fig. 3. Density estimators for the VV group with $h = 0.50$: (a) nonparametric density estimator; (b) semiparametric density estimator

Fig. 4 illustrates both of the density estimators for the NV group. Here the optimal bandwidth was calculated to be 0.60. Although there are no strong visible differences between the two estimators, the asymptotic relative efficiency is equal to 0.62 approximately. Furthermore, Figs 3(b) and 4(b) indicate that there are differences between the mean time required by the subjects to fuse the images.

5. Outlook

The subject-matter of this study was inference for the so-called density ratio model which is specified by assuming that the log-ratio of two or more unknown probability density functions is linear in some parameters. The density ratio model has attracted much attention recently as the list of references shows. There are two main explanations for its success. First it relaxes several conventional assumptions in the context of multiple-samples problems and second fitting can be easily implemented in standard software. The contribution of this work was to study the large sample behaviour of parameter estimators and to propose semiparametric density estimators of the unknown probability density functions. The goal was achieved by merging information and using the methodology of empirical likelihood. In particular, it was shown that the kernel density estimators that are based on the combination of information are more efficient than the traditional kernel density estimators.

The problem of density estimation under the two-sample density ratio model is also the subject of a recent contribution by Cheng and Chu (2004). Although our work covers the general case of m samples for model (3), Cheng and Chu (2004) obtain identical results for the case that they considered. Their bandwidth selection criterion is based on a least squares cross-validation scheme whereas the resulting density estimators are employed for a goodness-of-fit test of

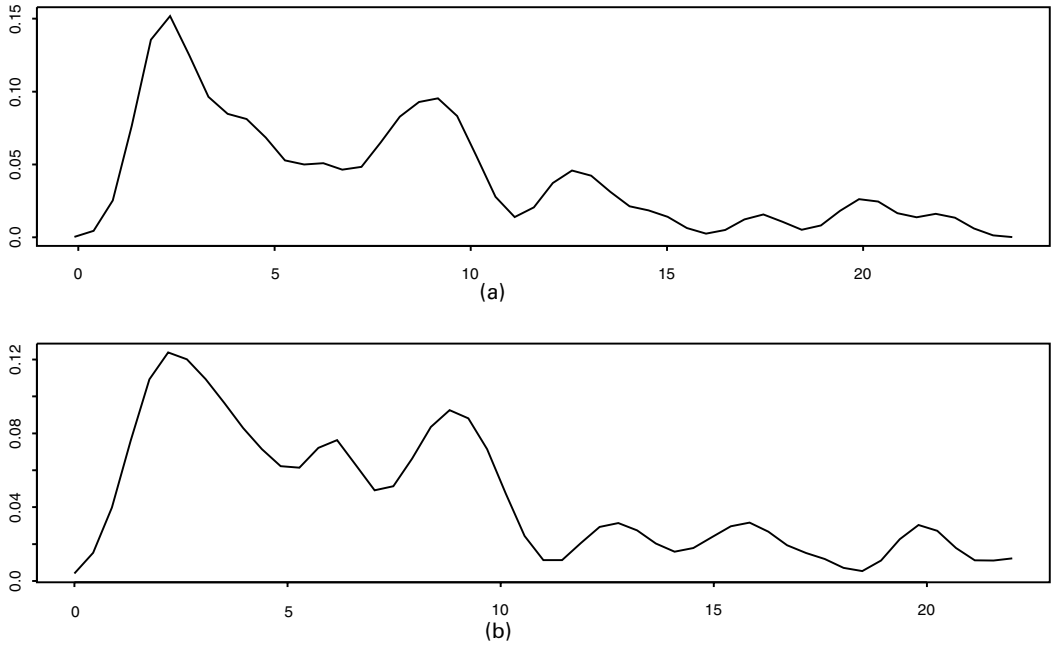

Fig. 4. Density estimators for the NV group with $h = 0.60$: (a) nonparametric density estimator; (b) semiparametric density estimator

model (2) when $m = 2$. To be more specific, they consider the L_2-norm as a measure of distance between $\hat{g}_m(x)$ and $\bar{g}_m(x)$, i.e.

$$L_2 = \int \{\hat{g}_m(x) - \bar{g}_m(x)\}^2 \, \mathrm{d}x,$$

on recalling equations (10) and (14). Apparently the work by Cheng and Chu (2004) is the first to consider a goodness-of-fit test based on the probability density function rather than the cumulative distribution function; see Qin and Zhang (1997). Their test can be readily adapted for the m-sample problem by using the estimators outlined in the theory, i.e. the above display is a test statistic for testing the goodness of fit of model (3). We shall not pursue this point any further.

Several questions need to be addressed for the topic of density estimation when the density ratio model holds true. For instance the problem of bandwidth selection must be investigated thoroughly together with the properties of the resulting estimates. In addition large sample results of the resulting kernel density estimators were not proved—although asymptotic normality should be expected under fairly mild conditions. Another problem of interest is the estimation of the densities derivatives and the quality of approximation. Finally all the results of this paper are proved under the square error loss although it is well known that other norms can be used for density estimation.

Acknowledgements

This work has greatly benefited from the constructive comments of two reviewers. The work of Cheng and Chu (2004) was pointed out by the Joint Editor, whose remarks are also acknowledged.

Appendix A

Throughout the appendix we use the notation E_i, var_i and cov_i to denote expectation, variance and covariance with respect to the ith sample.

A.1. Proof of theorem 1

To prove theorem 1 we note that a Taylor expansion leads to

$$\begin{pmatrix} \hat{\theta}-\theta \\ \hat{\mu}-\zeta \end{pmatrix} = -\begin{pmatrix} \frac{1}{n}\frac{\partial^2 l}{\partial\theta\partial\theta'} & \frac{1}{n}\frac{\partial^2 l}{\partial\theta\partial\mu'} \\ \frac{1}{n}\frac{\partial^2 l}{\partial\mu\partial\theta'} & \frac{1}{n}\frac{\partial^2 l}{\partial\mu\partial\mu'} \end{pmatrix}^{-1}_{(\theta,\zeta)} \begin{pmatrix} \frac{1}{n}\frac{\partial l}{\partial\theta} \\ \frac{1}{n}\frac{\partial l}{\partial\mu} \end{pmatrix}_{(\theta,\zeta)}. \tag{18}$$

Considering the second term on the right-hand side of equation (18) and dropping notation that depends on x, we obtain

$$E\left(\frac{\partial l}{\partial\theta_l}\right)_{(\theta,\zeta)} = -n_l\, E_m\left(\frac{\partial w_l}{\partial\theta_l}\right) + n_l\, E_m\left[\frac{\partial\{\log(w_l)\}}{\partial\theta_l} w_l\right] = 0$$

and

$$E\left(\frac{\partial l}{\partial\mu_l}\right)_{(\theta,\zeta)} = -n\, E_m(w_l - 1) = 0,$$

for $l = 1, \ldots, q$. In addition, we obtain that

$$\mathrm{var}\begin{pmatrix} \frac{1}{\sqrt{n}}\frac{\partial l}{\partial\theta} \\ \frac{1}{\sqrt{n}}\frac{\partial l}{\partial\mu} \end{pmatrix}_{(\theta,\zeta)} = \mathbf{S}.$$

Here $\mathbf{S}$ is a $(d+q)\times(d+q)$ matrix given by

$$\begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{12}' & \mathbf{S}_{22} \end{pmatrix}.$$

The elements of $\mathbf{S}_{11}$ are given by

$$\mathbf{S}_{11}(l,l') = \sum_{i=1}^{m} \zeta_i\zeta_l^2\, \mathrm{var}_i\left(\frac{\partial w_l/\partial\theta_l}{\sum_{k=1}^{m}\rho_k w_k}\right) + \zeta_l\, \mathrm{var}_l\left[\frac{\partial\{\log(w_l)\}}{\partial\theta_l}\right] - 2\zeta_l^2\, \mathrm{cov}_l\left[\frac{\partial w_l/\partial\theta_l}{\sum_{k=1}^{m}\rho_k w_k}; \frac{\partial\{\log(w_l)\}}{\partial\theta_l}\right]$$

for $l = l'$, and

$$\begin{aligned}\mathbf{S}_{11}(l,l') = {} & \sum_{i=1}^{m} \zeta_i\zeta_l\zeta_{l'}\, \mathrm{cov}_i\left(\frac{\partial w_l/\partial\theta_l}{\sum_{k=1}^{m}\rho_k w_k}; \frac{\partial w_{l'}/\partial\theta_{l'}}{\sum_{k=1}^{m}\rho_k w_k}\right) - \zeta_l\zeta_{l'}\, \mathrm{cov}_l\left[\frac{\partial w_{l'}/\partial\theta_{l'}}{\sum_{k=1}^{m}\rho_k w_k}; \frac{\partial\{\log(w_l)\}}{\partial\theta_l}\right] \\ & - \zeta_l\zeta_{l'}\, \mathrm{cov}_l\left[\frac{\partial w_l/\partial\theta_1}{\sum_{k=1}^{m}\rho_k w_k}; \frac{\partial\{\log(w_{l'})\}}{\partial\theta_{l'}}\right]\end{aligned}$$

for $l \neq l'$. The matrix $\mathbf{S}_{12}$ consists of the elements

$$\mathbf{S}_{12}(l,l') = \begin{cases} \sum_{i=1}^{m} \zeta_i\zeta_l\, \mathrm{cov}_i\left(\frac{\partial w_l/\partial\theta_l}{\sum_{k=1}^{m}\rho_k w_k}; \frac{w_l - 1}{\sum_{k=1}^{m}\rho_k w_k}\right) - \zeta_l\, \mathrm{cov}_l\left[\frac{\partial\{\log(w_l)\}}{\partial\theta_l}; \frac{w_l - 1}{\sum_{k=1}^{m}\rho_k w_k}\right] & \text{for } l = l', \\ \sum_{i=1}^{m} \zeta_i\zeta_l\, \mathrm{cov}_i\left(\frac{\partial w_l/\partial\theta_l}{\sum_{k=1}^{m}\rho_k w_k}; \frac{w_{l'} - 1}{\sum_{k=1}^{m}\rho_k w_k}\right) - \zeta_l\, \mathrm{cov}_l\left[\frac{\partial\{\log(w_l)\}}{\partial\theta_l}; \frac{w_{l'} - 1}{\sum_{k=1}^{m}\rho_k w_k}\right] & \text{for } l \neq l'. \end{cases}$$

The elements of $\mathbf{S}_{22}$ are given by

$$\mathbf{S}_{22}(l,l') = \begin{cases} \sum_{i=1}^{m} \zeta_i \operatorname{var}_i\left(\dfrac{w_l - 1}{\sum_{k=1}^{m} \rho_k w_k}\right), & \text{for } l = l', \\ \sum_{i=1}^{m} \zeta_i \operatorname{cov}_i\left(\dfrac{w_l - 1}{\sum_{k=1}^{m} \rho_k w_k}; \dfrac{w_{l'} - 1}{\sum_{k=1}^{m} \rho_k w_k}\right), & \text{for } l \neq l'. \end{cases}$$

Thus, an application of the central limit theorem leads to

$$\begin{pmatrix} \dfrac{1}{\sqrt{n}} \dfrac{\partial l}{\partial \theta} \\ \dfrac{1}{\sqrt{n}} \dfrac{\partial l}{\partial \mu} \end{pmatrix}_{(\theta,\zeta)} \to N(\mathbf{0}, \mathbf{S}) \tag{19}$$

in distribution, as $n \to \infty$. In addition we obtain that

$$-\frac{1}{n} \begin{pmatrix} \dfrac{\partial^2 l}{\partial \theta \partial \theta'} & \dfrac{\partial^2 l}{\partial \theta \partial \mu'} \\ \dfrac{\partial^2 l}{\partial \mu \partial \theta'} & \dfrac{\partial^2 l}{\partial \mu \partial \mu'} \end{pmatrix}_{(\theta,\zeta)} \to \mathbf{D} = \begin{pmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{21} & \mathbf{D}_{22} \end{pmatrix}$$

in probability, by the weak law of large numbers. The elements of $\mathbf{D}_{11}$ are given by

$$\mathbf{D}_{11}(l,l') = \begin{cases} \zeta_l \, E_m\left(\dfrac{\dfrac{\partial w_l}{\partial \theta_l} \dfrac{\partial w_l}{\partial \theta_l'}}{\sum_{k=1}^{m} \rho_k w_k} \sum_{k \neq l}^{m} \rho_k w_k\right), & \text{for } l = l', \\ -\zeta_l \zeta_{l'} \, E_m\left(\dfrac{\dfrac{\partial w_l}{\partial \theta_l} \dfrac{\partial w_{l'}}{\partial \theta_{l'}'}}{\sum_{k=1}^{m} \rho_k w_k}\right), & \text{for } l \neq l'. \end{cases}$$

Similarly, we obtain

$$\mathbf{D}_{12}(l,l') = \begin{cases} \zeta_l \, E_m\left(\dfrac{\partial w_l / \partial \theta_l}{\sum_{k=1}^{m} \rho_k w_k} \sum_{k \neq l}^{m} \rho_k w_k\right), & \text{for } l = l', \\ -\zeta_l \zeta_{l'} \, E_m\left(\dfrac{\dfrac{\partial w_l}{\partial \theta_l} \dfrac{\partial w_{l'}}{\partial \theta_{l'}'}}{\sum_{k=1}^{m} \rho_k w_k}\right), & \text{for } l \neq l'. \end{cases}$$

and finally we end up with

$$\mathbf{D}_{22}(l,l') = \begin{cases} \zeta_l \, E_m\left\{\dfrac{(w_l - 1)^2}{\sum_{k=1}^{m} \rho_k w_k}\right\}, & \text{for } l = l' \\ -E_m\left\{\dfrac{(w_l - 1)(w_{l'} - 1)}{\sum_{k=1}^{m} \rho_k w_k}\right\}, & \text{for } l \neq l'. \end{cases}$$

Thus, we obtain the theorem with

$$\mathbf{W} = \mathbf{D}^{-1} \mathbf{S} \mathbf{D}'^{-1} \tag{20}$$

A.2. Proof of lemma 2

We have that

$$\hat{g}_l(x) - \tilde{g}_l(x) = \frac{1}{h_n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\frac{\hat{w}_l(x_{ij})}{n\left[1 + \sum_{k=1}^{m} \hat{\mu}_k\{\hat{w}_k(x_{ij}) - 1\}\right]} - \frac{w_l(x_{ij})}{n_m\left[1 + \sum_{k=1}^{m} \rho_k\{w_k(x_{ij}) - 1\}\right]} \right),$$

$$K\left(\frac{x - x_{ij}}{h_n}\right) = U'_{1n}(\hat{\theta} - \theta) + U'_{2n}(\hat{\mu} - \mu),$$

by a first-order Taylor series expansion. However, it can be shown that both U_{1n} and U_{2n} are of the order $O_p(1)$. Thus, theorem 1 leads to the result desired.

References

Agresti, A. (1990) *Categorical Data Analysis*. New York: Wiley.

Anderson, J. A. (1972) Separate sample logistic discrimination. *Biometrika*, **59**, 19–35.

Anderson, J. A. (1979) Multivariate logistic compounds. *Biometrika*, **66**, 17–26.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1998) *Efficient and Adaptive Estimation for Semiparametric Models*. New York: Springer.

Cheng, K. F. and Chu, C. K. (2004) Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, to be published.

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

Fahrmeir, L. and Tutz, G. (2001) *Multivariate Statistical Modeling based on Generalized Linear Models*, 2nd edn. New York: Springer.

Fokianos, K. (2003) Box–Cox transformation for semiparametric comparison of two samples. In *Foundations of Statistical Inference* (eds H. R. L. Y. Haitovsky and Y. Ritov), pp. 131–140. Heidelberg: Physica.

Fokianos, K., Kedem, B., Qin, J., Haferman, J. and Short, D. A. (1998) On combining instruments. *J. Appl. Meteorol.*, **37**, 220–226.

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

Gilbert, P. B. (2000) Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. *Ann. Statist.*, **28**, 151–194.

Gilbert, P. B., Lele, S. R. and Vardi, Y. (1999) Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika*, **86**, 27–43.

Gill, R. D., Vardi, Y. and Wellner, J. A. (1988) Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.*, **16**, 1069–1112.

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

Hjort, N. L. and Jones, M. C. (1996) Locally parametric nonparametric density estimation. *Ann. Statist.*, **24**, 1619–1647.

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

Lloyd, C. J. and Jones, M. C. (2000) Nonparametric density estimation from biased data with unknown biasing functions. *J. Am. Statist. Ass.*, **95**, 865–876.

Murphy, S. A. and van der Vaart, A. W. (1997) Semiparametric likelihood ratio inference. *Ann. Statist.*, **25**, 1471–1509.

Murphy, S. A. and van der Vaart, A. W. (2000) On profile likelihood (with discussion). *J. Am. Statist. Ass.*, **95**, 449–485.

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

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

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

Qin, J., Barwick, M., Ashbolt, R. and Dwyer, T. (2002) Quantifying the change of melanoma incidence by Breslow thickness. *Biometrics*, **58**, 665–670.

Qin, J. and Lawless, J. F. (1994) Empirical likelihood and general estimating functions. *Ann. Statist.*, **22**, 300–325.

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

Scott, D. W. (1992) *Multivariate Density Estimation, Theory, Practice, and Visualization*. New York: Wiley.

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

Sun, J. and Woodroofe, M. (1997) Semi-parametric estimates under biased sampling. *Statist. Sin.*, **7**, 545–575.

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

Vardi, Y. (1985) Empirical distribution in selection bias models. *Ann. Statist.*, **13**, 178–203.
Venables, W. N. and Ripley, B. D. (1999) *Modern Applied Statistics with S-Plus*, 3rd edn. New York: Springer.
Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*. London: Chapman and Hall.
Wu, C. O. (1996) Kernel smoothing of the nonparametric maximum likelihood estimates for biased sampling models. *Math. Meth. Statist.*, **5**, 275–298.