Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 52 (2008) 2451 – 2468

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Projection density estimation under a m-sample semiparametric model

Jean-Baptiste Aubin, Samuela Leoni-Aubin*

Univ. de Technologie de Compiègne, Centre de Recherche de Royallieu, Rue Personne de Roberval, BP 20529, 60205 Compiègne, France

Received 5 December 2006; received in revised form 7 August 2007; accepted 8 August 2007
Available online 23 August 2007

Abstract

An m-sample semiparametric model in which the ratio of $m-1$ probability density functions with respect to the mth is of a known parametric form without reference to any parametric model is considered. This model arises naturally from retrospective studies and multinomial logistic regression model. A projection density estimator is constructed by smoothing the increments of the maximum semiparametric empirical likelihood estimator of the underlying distribution function, using the combined data from all the samples. Some asymptotic results on the proposed projection density estimator are established. Connections between our estimator and kernel semiparametric density estimator are pointed out. Some results from simulations and from the analysis of two real data sets are presented.

Keywords: Case-control data; Semiparametric maximum likelihood estimation; Projection density estimation; Truncation index

1. Introduction

Consider m independent random samples $x_{i1}, \ldots, x_{in_i}$, $i=1, \ldots, m$ with probability densities $g_i(x) = \mathrm{d}G_i(x)$, $i=1, \ldots, m$, respectively.

We consider the following semiparametric density ratio model:

$$g_i(x) = w(x, \theta_i) g_m(x), \quad i = 1, \ldots, m-1, \tag{1}$$

where w is a known positive function and θ_k, $k=1, \ldots, m-1$ is a vector of parameters with finite dimension equal to d. The common support of the laws G_i may be known or unknown, discrete or continuous. All the m density functions are assumed unknown but are related, however, through a tilt (or distortion) which determines the difference between them.

The density ratio model has attracted much attention recently, because it relaxes several conventional assumptions in the context of multisamples problems and because fitting can be easily implemented in standard software.

Model (1) can be viewed as a generalisation of multinomial logistic regression (taking $w(x, \theta_i) = \exp\{\theta_{1,i} + x'\theta_{2,i}\}$, where $\theta_{1,i}$ is a scalar parameter and $\theta_{2,i}$ is a $(d-1)$-vector of parameters, for $i=1, \ldots, m-1$, see Fokianos, 2004). This kind of model is one of the most popular choices for nominal data analysis, with several applications especially in

* Corresponding author.
E-mail addresses: jean-baptiste.aubin@utc.fr (J.-B. Aubin), samuela.leoni-aubin@utc.fr (S. Leoni-Aubin).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.08.008

econometrics and biostatistics. This approach also generalizes the classical normal-based one-way analysis of variance in the sense that it obviates the need for a completely specified parametric model, see Fokianos et al. (2001). Moreover, expression (1) can also be seen as a biased sampling model with weights depending on parameters. Vardi (1982), Vardi (1985) and Gill et al. (1988) have discussed inference in biased sampling models with known weight functions. Gilbert (2000) and Gilbert et al. (1999) considered weight functions depending on an unknown finite dimensional parameter, and discussed the identifiability problem of G_m and the maximum likelihood estimation of θ_k, $k = 1, \dots, m-1$ and G_m (see Gilbert et al., 1999, Section 3; Gilbert, 2000, Section 1.1).

For an application of the density ratio model to meteorological data, see Fokianos et al. (1998). For further applications of model (1), see Fokianos et al. (2001), Qin et al. (2002), Fokianos (2004), Cheng and Chu (2004) and Qin and Zhang (2005).

Fokianos et al. (2001) and Keziou and Leoni-Aubin (2005) proposed homogeneity tests in the context of density ratio models, respectively, based on a Wald-type test statistic and on a likelihood ratio test statistic.

Inference for parameters of model (1) in the case $m = 2$ has been studied by Qin (1998), Keziou and Leoni-Aubin (2005, 2007).

The aim of this contribution is to estimate unknown densities in two steps, using the combined data from all the samples, hence taking the information contained in all samples into account. First, applying the empirical likelihood method to model (1), then, using a projection density estimator (see also Aubin and Leoni Aubin, 2007).

A particular case of model (1) has been suggested by Efron and Tibshirani (1996) for density estimation. Combining parametric and nonparametric methods of density estimation, they developed a new method for estimating an exponential family of probability densities

$$g_\theta(x) = g_0(x)\exp\{\theta_0 + s(x)\theta_1\},$$

based on a random sample $x_1, \dots, x_n$. Here g_0 is a carrier density, s is a vector of sufficient statistics and $\theta = (\theta_0, \theta_1)'$ is a parameter vector. Efron and Tibshirani (1996) proposed a two steps estimation procedure of $g_\theta(x)$. First, they estimated $g_0(x)$ by using a kernel density estimator and then they fitted a parametric family. 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.

Recent works (Cheng and Chu, 2004; Fokianos, 2004; Qin and Zhang, 2005) have adopted a quite different approach, using an m-sample (Fokianos, 2004) or a two-sample (Cheng and Chu, 2004; Qin and Zhang, 2005) density ratio model. First, they estimated the parameters' values by maximising a semiparametric likelihood function, and then they obtained the maximum semiparametric likelihood estimator of the unknown distribution function by putting weights on all the observations. Given this inference output, they smoothed the increments of the estimated distribution function to obtain a new kernel density estimator. Specifically, Fokianos (2004) showed that in density ratio model (1), the pooled data leads to more efficient kernel density estimators for the unknown distributions, in the sense that they have the same amount of bias but they are less variable than traditional kernel density estimators. Cheng and Chu (2004) and Qin and Zhang (2005) studied the problem of kernel density estimation under model (1) with $m = 2$. The bandwidth selection criterion of Cheng and Chu (2004) is based on a least square cross validation scheme, whereas resulting density estimators are employed for a goodness-of-fit test of the two-sample density ratio model, using the L_2 norm of the difference between semiparametric and nonparametric kernel density estimators. They also showed that their proposed semiparametric density estimator not only is consistent, but also has the "smallest" asymptotic variance among general nonparametric kernel density estimators. Qin and Zhang (2005) used an iterative approach to select the bandwidth and established asymptotic normality of estimators.

This paper is organised as follows. In Section 2 we recall the estimation method of the finite dimensional parameters in model (1) based on the empirical likelihood approach (see Qin, 1998 and references therein). Section 3, in connection with the theory of Section 2, sets forward projection density estimators of the unknown probability density functions. Section 4 relates to some asymptotic results of the discussed estimators. In particular, we demonstrate that, when the projection basis is chosen in such a way that the Fourier coefficients decay fast enough, the proposed estimator performs better than the semiparametric kernel density estimator. Simulation results are presented in Section 5 to study the finite sample performance of our proposed estimators, and an application of the methodology to real data is also given. Some concluding remarks are provided in Section 6. Finally, proofs of theoretical results are given in the Appendix.

2. Inference in density ratio model

Consider the m samples with corresponding densities that satisfy Eq. (1), let $n := \sum_{i=1}^{m} n_i$ be the total sample size and consider the empirical likelihood (see Owen, 1988, 2001) based on the pooled data $\{x_{ij},\ j=1,\dots,n_i,\ i=1,\dots,m\}$

$$L(\theta, G_m) = \left\{\prod_{j=1}^{n_1} p_{1j} w(x_{1j}, \theta_1)\right\}\left\{\prod_{j=1}^{n_2} p_{2j} w(x_{2j}, \theta_2)\right\}\dots\prod_{j=1}^{n_m} p_{mj},$$

where $p_{ij} := \mathrm{d}G_m(x_{ij})$ and $\theta = (\theta_1^{\mathrm{t}}, \dots, \theta_{m-1}^{\mathrm{t}})^{\mathrm{t}}$ is a $(m-1)d$-vector.

Hence, the log-likelihood is written as follows:

$$l(\theta, p) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \log(p_{ij}) + \sum_{i=1}^{m-1}\sum_{j=1}^{n_i} \log\{w(x_{ij}, \theta_i)\}, \tag{2}$$

where $p := \{p_{ij},\ j=1,\dots,n_i,\ i=1,\dots,m\}$.

Maximisation of Eq. (2) is carried out by employing the two-step profiling approach described in Qin and Lawless (1994). This procedure relies on first maximising the nonparametric part in the full likelihood function with θ fixed, and then maximising the profile log-likelihood function with respect to θ. The profile log-likelihood is then $l(\theta) = \sup_{p\in\mathscr{C}_\theta} l(\theta, p)$, where p is constrained to the set

$$\mathscr{C}_\theta := \left\{p \in \mathbb{R}_+^n : \sum_{i=1}^{m}\sum_{j=1}^{n_i} p_{ij} = 1, \sum_{i=1}^{m}\sum_{j=1}^{n_i} p_{ij}\{w(x_{ij}, \theta_k) - 1\} = 0,\ k=1,\dots,m-1\right\}.$$

The maximisation uses the method of Lagrange multipliers, and it follows that if the set $\mathscr{C}_\theta$ is not empty,

$$p_{ij}(\lambda, \theta) = \frac{1}{n}\frac{1}{1+\sum_{k=1}^{m-1}\lambda_k\{w(x_{ij}, \theta_k) - 1\}}, \tag{3}$$

where $\lambda := \{\lambda_k, k=1,\dots,m-1\}$ is the vector of the Lagrange multipliers determined by the following equations:

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{w(x_{ij}, \theta_k) - 1}{1+\sum_{l=1}^{m-1}\lambda_l\{w(x_{ij}, \theta_l) - 1\}} = 0, \quad k=1,\dots,m-1.$$

It turns out that the vector of the Lagrange multipliers is a continuously differentiable function of the parameter θ, hence Eq. (2) becomes

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

Under some regularity conditions (see Fokianos, 2004), as $n \to \infty$, $\widehat{\theta}$ and $\widehat{\lambda} = \lambda(\widehat{\theta})$ exist, are consistent, satisfy the following system of estimating equations:

$$\frac{\partial l(\theta,\lambda)}{\partial\theta_k} = -\sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{\lambda_k \partial w(x_{ij}, \theta_k)/\partial\theta_k}{1+\sum_{l=1}^{m-1}\lambda_l\{w(x_{ij}, \theta_l) - 1\}} + \sum_{j=1}^{n_k} \frac{\partial[\log\{w(x_{kj}, \theta_k)\}]}{\partial\theta_k} = 0,$$
$$\frac{\partial l(\theta,\lambda)}{\partial\lambda_k} = -\sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{w(x_{ij}, \theta_k) - 1}{1+\sum_{l=1}^{m-1}\lambda_l\{w(x_{ij}, \theta_l) - 1\}} = 0, \quad k=1,\dots,m-1$$

and are asymptotically normal. So we can obtain the following maximum likelihood estimator of G_m:

$$\widehat{G}_m(x) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \widehat{p}_{ij}\mathbb{1}(x_{ij} \leqslant x) = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{1}{1+\sum_{l=1}^{m-1}\widehat{\lambda}_l\{w(x_{ij}, \widehat{\theta}_l) - 1\}}\mathbb{1}(x_{ij} \leqslant x),$$

where $\mathbb{1}$ denotes the indicator function and $\widehat{p}_{ij} = p_{ij}(\widehat{\lambda}, \widehat{\theta})$ is obtained by (3). As a consequence, for $k = 1, \ldots, m-1$, the maximum likelihood estimator of $G_k(x)$, say $\widehat{G}_k(x)$, is obtained by

$$\widehat{G}_k(x) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \widehat{p}_{ij} w(x_{ij}, \widehat{\theta}_k) \mathbb{1}(x_{ij} \leqslant x).$$

To sum up, a profiling procedure gives us the score estimating equations for the finite dimensional parameters and the nonparametric estimators $\widehat{G}_k(x)$, $k = 1, \ldots, m$ for the unknown distribution functions.

We now turn to the question of semiparametric density estimation based on the above inference output.

3. New semiparametric density estimators

Although the problem of density estimation can be addressed by considering a crude histogram (see Fokianos et al., 1998), a smooth density estimator will be in general more preferable.

Cheng and Chu (2004), Fokianos (2004) and Qin and Zhang (2005) propose semiparametric density estimators by modifying classic kernel density estimators (essentially by smoothing the increments of $\widehat{G}_i$, $i = 1, \ldots, m$).

Another smoothing method is obtained by projection. It consists in projecting the density to estimate on a finite dimensional space (for example, the one generated by the first components of a basis of the space of possible densities) and estimate this projection by a moments method (see Cencov, 1962).

In the following, we will design the pooled data $(x_{11}, \ldots, x_{1n_1}, \ldots, x_{mn_m})$ by $(t_1, \ldots, t_n)$. We suppose that for $l = 1, \ldots, m$, the lth sample admits the density g_l with respect to μ such that $g_l \in L^2(\mu)$, where μ is a finite measure. Let $(e_j)_{j\in\mathbb{N}^*}$ be some orthonormal basis in the separable infinite-dimensional Hilbert space $L^2(\mu)$. Since any orthogonal basis with bounded basis vectors may be normalised, the use of orthonormal bases is a matter of convenience rather than restriction. For any $g \in L^2(\mu)$, the sequence of its Fourier coefficients $(a_j)_{j\in\mathbb{N}^*}$ is the unique sequence of parameters defining g, since any function $g \in L^2(\mu)$ is uniquely represented by its series expansion,

$$g = \sum_{j=1}^{\infty} a_j e_j,$$

where $a_j = \langle g, e_j \rangle_{L^2(\mu)} = \int g e_j \, \mathrm{d}\mu$, $j \geqslant 1$. In this work, we will denote by a_j the jth Fourier coefficient of the density function g_m.

The classic projection density estimator (see Cencov, 1962) for $g_m = \sum_{j=1}^{\infty} a_j e_j$ is

$$\overline{g}_{m_{n_m}} = \sum_{j=1}^{k_{n_m}} \overline{a}_{j,n_m} e_j, \tag{4}$$

where (k_{n_m}) is a truncation index sequence such that $k_{n_m} = o(n_m)$ and $(k_{n_m}) \uparrow \infty$ when $n_m \uparrow \infty$. $\overline{a}_{j,n_m} = (1/n_m)\sum_{i=1}^{n_m} e_j(x_{mi})$ is the unbiased estimate of the jth Fourier coefficient a_j. We assume throughout the paper that the basis $(e_j)_{j\in\mathbb{N}^*}$ is uniformly bounded (such that $\exists M < \infty : \sup_j \|e_j\|_\infty < M$).

However, in order to estimate more efficiently g_m, one can merge the information coming from the m samples (instead of only from the last one) whose densities are linked by model (1). Here, every t_i is associated to a p_i, $i = 1, \ldots, n$ (after a rearrangement of (3)), which is estimated by the empirical likelihood method. We recall that the p_i verify $\sum_{i=1}^{n} p_i = 1$. Our modified projection density estimator of g_m is then

$$\widehat{g}_{m_n} = \sum_{j=1}^{k_n} \widehat{a}_{j,n} e_j \quad \text{with } \widehat{a}_{j,n} := \sum_{i=1}^{n} \widehat{p}_i e_j(t_i), \tag{5}$$

where (k_n) is such that $k_n = o(n)$ and $(k_n) \uparrow \infty$ when $n \uparrow \infty$. An empirical choice of (k_n) is proposed at the end of this section. Furthermore, Eq. (5) is a semiparametric density estimator since it depends on both the unknown distribution function and the parameters of model (1).

We show that, if we are able to choose a suitable projection basis $(e_j)_{j\in\mathbb{N}^*}$ (i.e., if $(e_j)_{j\in\mathbb{N}^*}$ is such that the decrease of Fourier coefficients of g_m is strong enough), then the asymptotic mean integrated square error (AMISE) of the estimator defined in (5) can be close to $1/n$ (see Corollary 4.3). This rate (almost a "parametric" one) is better than the one obtained by the kernel estimation method, both in the cases of classic and semiparametric density estimation (see Cheng and Chu, 2004; Fokianos, 2004; Qin and Zhang, 2005).

We deduce projection estimators for the other densities $g_l, l = 1, \dots, m-1$ as follows:

$$\widehat{g}_{l_n} = \sum_{j=1}^{k_n} \widehat{a}_{j,n} e_j \quad \text{with } \widehat{a}_{j,_n} := \sum_{i=1}^{n} \widehat{p}_i w(t_i, \widehat{\theta}_l) e_j(t_i), \quad l = 1, \dots, m-1.$$

Obviously, these estimators enjoy the same asymptotic properties as (5).

We close this section by discussing how to select the truncation index k_n in practice. We want to find an optimal $\widehat{k}_n$ (i.e., which minimizes the mean integrared square error (MISE)). We adapted an approach from Bosq and Lecoure (1987, pp. 272–273). The basic idea is the following: if we state that $MISE(k_n)$ is a strictly convex function (a reasonable condition) which attains its minimum in k_n^{opt}, then, the quantity

$$\Delta(k_n) := MISE(k_n + 1) - MISE(k_n)$$

increases with k_n, and

$$k_n^{\text{opt}} = \min\{k_n : \Delta(k_n) \geqslant 0\}. \tag{6}$$

We have

$$\forall k_n \geqslant 1, \quad MISE(k_n) = \sum_{j=1}^{k_n} \mathbb{E}[(\widehat{a}_{j,_n} - a_j)^2] + \sum_{j > k_n} a_j^2,$$

so

$$\Delta(k_n) = \mathbb{E}[(\widehat{a}_{k_n+1,_n} - a_{k_n+1})^2] - a_{k_n+1}^2.$$

We empirically approximate $\mathbb{E}[(\widehat{a}_{k_n+1,_n} - a_{k_n+1})^2]$ by

$$\widehat{\text{Var}}(\widehat{a}_{j,_n}) = \sum_{l=1}^{m} \left\{ \frac{n_l}{n_l - 1} \sum_{i=s(l-1)+1}^{s(l)} \left[\widehat{p}_i e_{k_n+1}(t_i) - \frac{1}{n_l} \sum_{j=s(l-1)+1}^{s(l)} \widehat{p}_j e_{k_n+1}(t_j) \right]^2 \right\},$$

where $s(l) := \sum_{k=1}^{l} n_k$ and $s(0) := 0$, and $a_{k_n+1}^2$ by $\widehat{a}_{k_n+1,n}^2$. Finally, we obtain

$$\widehat{\Delta}(k_n) := \sum_{l=1}^{m} \left\{ \frac{n_l}{n_l - 1} \sum_{i=s(l-1)+1}^{s(l)} \left[\widehat{p}_i e_{k_n+1}(t_i) - \frac{1}{n_l} \sum_{j=s(l-1)+1}^{s(l)} \widehat{p}_j e_{k_n+1}(t_j) \right]^2 \right\} - \widehat{a}_{k_n+1,n}^2.$$

So, a natural data-based choice of (6) is

$$\widehat{k}_n^{\text{opt}} := \begin{cases} \inf\{k_n \leqslant K : \widehat{\Delta}(k_n) \geqslant 0\}, \\ K \quad \text{if } \widehat{\Delta}(k_n) < 0, \quad \forall k_n \leqslant K, \end{cases} \tag{7}$$

where $K \leqslant n$ is chosen by the user.

4. Asymptotic results

In this section, we consider the AMISE of the semiparametric projection density estimator $\widehat{g}_{m_n}$ (defined by Eq. (5)) as a measure of its global accuracy.

To study statistical properties of $\widehat{g}_{m_n}$, it is useful to consider

$$\tilde{g}_{m_n} = \sum_{j=1}^{k_n} \tilde{a}_{j,n} e_j \quad \text{with } \tilde{a}_{j,n} := \sum_{i=1}^{n} p_i e_j(t_i).$$

Proposition 4.1. *$\tilde{a}_{j,n}$ is an unbiased estimator of a_j.*

Theorem 4.2. *Under classical regularity conditions (see hypotheses in Fokianos, 2004, Theorem 1), we have*

$$\mathbb{E}\|\widehat{g}_{m_n} - g_m\|^2 = \mathcal{O}\left(\frac{k_n}{n}\right) + \sum_{j>k_n} a_j^2.$$

Theorem 4.2 reveals that large values of k_n reduce bias but introduce substantial variance as opposed to small values of the truncation index which lead to smaller variance but increased bias.

In the following corollary we study the AMISE of our estimator in three particular cases.

Corollary 4.3. *Under conditions of Theorem 4.2,*

(i) *if $\forall j \geqslant 1,\ |a_j| < \gamma j^{-\rho}$ where $\gamma > 0$ and $\rho > \frac{1}{2}$, then, for $k_n^* = n^{1/2\rho}$,*

$$\mathbb{E}\|\widehat{g}_{m_n} - g_m\|^2 = \mathcal{O}(n^{(1-2\rho)/2\rho}).$$

(ii) *if $\forall j \geqslant 1,\ |a_j| < \alpha\beta^{-j}$ where $\alpha > 0$ and $\beta > 1$, then, for $k_n^* = \frac{\log n}{2\log\beta}$,*

$$\mathbb{E}\|\widehat{g}_{m_n} - g_m\|^2 = \mathcal{O}\left(\frac{\log n}{n}\right).$$

(iii) *if $\exists J_0 \geqslant 1/\forall j > J_0,\ |a_j| = 0$, then, for any sequence $(v_n) \uparrow \infty$, it exists (k_n^*) such that,*

$$\mathbb{E}\|\widehat{g}_{m_n} - g_m\|^2 = \mathrm{o}\left(\frac{v_n}{n}\right).$$

Corollary 4.3 means that a strong decrease of the Fourier coefficients (in case (i) with $\rho > \frac{5}{2}$ and in cases (ii) and (iii), included in case (i)) implies that a good choice of (k_n) gives us a density estimator which reduces the AMISE when it is compared with that of the semiparametric kernel density estimator (see Cheng and Chu, 2004; Fokianos, 2004; Qin and Zhang, 2005). A strong decrease of the Fourier coefficients means that the user choosed a suitable projection basis $(e_j)_{j\in\mathbb{N}^*}$ with respect to the density to estimate. For example, one can obtain the rate in Corollary 4.3, part (iii) with the trigonometric basis for "periodic" densities, or, more generally, if it exists K_{g_m} such that g_m belongs to the vector space generated by $e_1, \ldots, e_{K_{g_m}}$.

We now give some examples of possible orthonormal bases. Kronmal and Tarter (1968) considered the sine system on the interval $[c; d]$.

$$\sin\left\{j\pi\frac{x-c}{d-c}\right\}, \quad j = 1, 2, \ldots,$$

the cosine system

$$\cos\left\{j\pi\frac{x-c}{d-c}\right\}, \quad j = 0, 1, \ldots$$

and the full trigonometric system

$$\left\{1, \cos\left\{j\pi\frac{x-c}{d-c}\right\}, \sin\left\{j\pi\frac{x-c}{d-c}\right\}\right\}, \quad j = 1, 2, \ldots .$$

The normalised Legendre basis is obtained from the Gram–Schmidt orthonormalisation procedure applied to the power functions $1, x, x^2, \ldots$ on $[-1, 1]$ (see Efromovich, 1999, p. 51).

Other classical choices of bases are possible: Hermite basis, Laguerre basis, Haar basis (see Conway, 1994; Härdle et al., 1998), etc.

One can also use a hybrid basis with the first few vectors from the normalised Legendre basis and consequent basis vectors obtained from the trigonometric basis by the Gram–Schmidt orthonormalisation procedure, or, more generally, "self-made" bases obtained by orthonormalisation of a set of privileged densities and then completed.

Remark 1. Most of these bases have continuous components, so they are expected to yield better results when the density to estimate is continuous; for the same reason, the Haar basis is expected to be appropriate when the density to estimate is not continuous.

We presented some nonuniformly bounded bases; we underline that asymptotic properties shown in this paper are not applicable to projection density estimators on these bases. Nevertheless, some previous works (see Aubin and Massiani, 2003) demonstrated similar properties in a nonparametric context for simply bounded bases. Moreover, in practice, this technical condition is not restrictive, since we consider only the first k_n components $(e_1, \dots, e_{k_n})$ of the projection basis.

In the next proposition, we precise the AMISE of semiparametric and nonparametric projection density estimators for the purpose of comparison.

Proposition 4.4. *Under hypotheses of Theorem* 4.2, *for the same truncation index selection k_n, if $\exists l < m$ such that $n_l \neq 0$, then*

$$AMISE(\widehat{g}_{m_n}) < AMISE(\bar{g}_{m_{n_m}}). \tag{8}$$

The following proposition gives pointwise asymptotic expressions for the bias and variance of the semiparametric density estimator $\widehat{g}_{m_n}(x)$. The corresponding results for the nonparametric projection density estimator $\bar{g}_{m_{n_m}}(x)$ are also provided.

Proposition 4.5. *Under hypotheses of Theorem* 4.2, *if $\sqrt{n}\sum_{j>k_n} a_j e_j(x) \rightarrow \infty$, for the same truncation index selection k_n, if $\exists l < m$ such that $n_l \neq 0$, then*

$$\begin{aligned}
&\mathbb{E}(\widehat{g}_{m_n}(x) - g_m(x)) = -\sum_{j>k_n} a_j e_j(x),\\
&\mathbb{E}(\bar{g}_{m_{n_m}}(x) - g_m(x)) = -\sum_{j>k_n} a_j e_j(x),
\end{aligned} \tag{9}$$

$$\operatorname{Var}(\widehat{g}_{m_n}(x)) = \int p(y)\left(\sum_{j=1}^{k_n} e_j(y)e_j(x)\right)^2 g_m(y)\,\mathrm{d}\mu(y) + \mathrm{o}\left(\frac{k_n}{n}\right) \tag{10}$$

and

$$\operatorname{Var}(\bar{g}_{m_{n_m}}(x)) = \int \frac{1}{n_m}\left(\sum_{j=1}^{k_n} e_j(y)e_j(x)\right)^2 g_m(y)\,\mathrm{d}\mu(y) + \mathrm{o}\left(\frac{k_n}{n}\right),$$

where $\forall y,\ p(y) = \frac{1}{n_m + \sum_{l=1}^{m-1} n_l w(y, \theta_l)} < \frac{1}{n_m}$.

Propositions 4.4 and 4.5 show that the asymptotic bias of $\widehat{g}_{m_n}$ is the same as that of $\bar{g}_{m_{n_m}}$; however, the dominant term of the asymptotic variance of $\widehat{g}_{m_n}$ is smaller than that of $\bar{g}_{m_n}$.

In Corollary 4.3, we showed that if the decrease of Fourier coefficients is strong enough, then $\widehat{g}_{m_n}$ performs better (in terms of AMISE) than the semiparametric kernel density estimator. Similarly, the comparison between pointwise asymptotic expressions for $\widehat{g}_{m_n}$ (Eqs. (9) and (10)) and those of the semiparametric kernel density estimator will depend on the decay of Fourier coefficients.

5. Applications

This section deals with finite sample performance of different density estimators (projection and kernel, semiparametric and classic) and an application to two real data sets.

In practice, for projection density estimators, if the selected basis has a finite support, then the data would usually have to be scaled to that support interval. Using the simplest linear transformation may require estimation of the data-supporting interval. Alternatively, one can use a nonlinear transformation that maps the whole real line to a finite interval, for example, based on the function $\arctan(x)$. Using bases with unlimited support, such as the Hermite system, eliminates the need for a transformation.

We consider orthonormal trigonometric and Legendre bases which are defined on $[-1, 1]$:

- trigonometric basis:

$$e_1(x) = \tfrac{1}{\sqrt{2}} \quad \forall k \geqslant 1, \begin{cases} e_{2k}(x) = \cos((2k-1)\pi x), \\ e_{2k+1}(x) = \sin((2k-1)\pi x). \end{cases}$$

- Legendre basis:

$$e_1(x) = \frac{1}{\sqrt{2}}, \quad e_2(x) = \frac{x}{\sqrt{2/3}}, \quad e_3(x) = \frac{3x^2-1}{\sqrt{8/5}}, \ldots .$$

These choices of bases suppose that we assume that the density to estimate $g_m \in L^2([-1, 1])$. More generally, by a direct transformation, this implies that g_m is assumed to have a compact support. For example, this is not true for a normal or an exponential distribution. Nevertheless, in practice, estimation of g_m on an adapted compact set $\mathscr{S}$ is enough. So, we have to transform our data in such a way that the support of the transformed data is included in $[-1, 1]$. For the sake of simplicity, we take $\mathscr{S} := [\min_i(t_i), \max_i(t_i)]$ in the semiparametric case and $\mathscr{S} := [\min_i(x_i), \max_i(x_i)]$ in the nonparametric case, where $(x_1, \ldots, x_{n_m})$ stands for the mth sample. So, we linearly transform the data:

$$\text{in the nonparametric case} \begin{cases} \min_i(x_i) \longrightarrow -1, \\ \max_i(x_i) \longrightarrow 1, \end{cases}$$
$$\text{in the semiparametric case} \begin{cases} \min_i(t_i) \longrightarrow -1, \\ \max_i(t_i) \longrightarrow 1. \end{cases}$$

We estimate the density over $[-1, 1]$ and make the inverse transformation to obtain the final estimator.

One of the drawbacks of the projection estimators of probability densities with respect to kernel ones is that they can take negative values. To avoid this problem, we consider a slightly modified estimator $\widehat{g}^{(1)}_{m_n}$ such that (see Efromovich, 1999)

$$\forall x \in \mathscr{S}, \quad \widehat{g}^{(1)}_{m_n}(x) := \max(0, \widehat{g}_{m_n}(x)),$$

renormalised in the following way:

$$\forall x \in \mathscr{S}, \quad \widehat{g}^{(2)}_{m_n}(x) := \frac{\widehat{g}^{(1)}_{m_n}(x)}{\int \widehat{g}^{(1)}_{m_n}(t)\,\mathrm{d}t}.$$

$\widehat{g}^{(2)}_{m_n}$ is a better candidate to estimate g_m since it is itself a density.

Moreover, $\int \widehat{g}_{m_n}(y)\,\mathrm{d}y \xrightarrow{\mathrm{P}} 1$ and $\forall x$ such that $g_m(x) > 0$, $\widehat{g}_{m_n}(x) - \widehat{g}^{(1)}_{m_n}(x) \xrightarrow{\mathrm{P}} 0$.

For projection density estimators, we choose the truncation index k_n according to procedure in (7), with $K = 7$. Such a K is large enough to allow a real choice of k_n, but small enough to save computational time.

For kernel density estimators, we consider classic kernel estimators and semiparametric kernel estimators presented in Cheng and Chu (2004), Fokianos (2004) and Qin and Zhang (2005). Selection of the smoothing parameter is carried out by the empirical estimation of the optimal value given in Proposition 1, Part (b) in Fokianos (2004) or in Theorem 2, Part (b) in Qin and Zhang (2005). More specifically, according to Silverman (1986, p. 59), a data-based choice of the bandwidth for the semiparametric kernel estimator (respectively, for the nonparametric density estimator) of g_m is given by the iterative procedure in Qin and Zhang (2005, formula 16) (respectively, Qin and Zhang, 2005, formula 14). We employ a standard Gaussian kernel and we take $\mathscr{S} := [\min_i(t_i), \max_i(t_i)]$ in the semiparametric case and $\mathscr{S} := [\min_i(x_i), \max_i(x_i)]$ in the nonparametric case.

5.1. Simulations

In this section, we report a limited simulation study to illustrate the finite sample performance of the proposed estimators. Our working model is that densities $g_1(t)$ and $g_2(t)$ are related by

$$g_1(t) = g_2(t)\exp\{\theta_1 + \theta_2 t\}. \tag{11}$$

We consider in this study two different cases (see Qin and Zhang, 2005, Section 5). Firstly, we assume that $g_1(t)$ is the density function of a $\mathcal{N}(\mu, 1)$ distribution and $g_2(t)$ is the standard normal density function, so that model (11) holds with $\theta_1 = -\frac{\mu^2}{2}$ and $\theta_2 = \mu$. Secondly, we assume that $g_1(x) = \frac{1}{\mu}\exp -\frac{x}{\mu}$ is the density function of an $\frac{1}{\mu}\mathscr{E}(\frac{1}{\mu})$ distribution and $g_2(x)$ is the density function of an $\mathscr{E}(1)$ distribution, so that model (11) holds with $\theta_1 = -\log\mu$ and $\theta_2 = 1 - \frac{1}{\mu}$. Here we aim to estimate in the first case the standard normal density and in the second case the standard exponential density.

In our simulations, we consider $\mu = 0.25, 0.5, 0.75, 1, 1.25, 1.5$ in the normal case and $\mu = 1.25, 1.5, 1.75, 2, 2.25, 2.5$ in the exponential case. For sample sizes of $n_1 = n_2 = 100$ and each value of μ, we generate 500 independent sets of combined random samples $(x_1, \dots, x_{n_2}, y_1, \dots, y_{n_1})$ from the $\mathcal{N}(0, 1)$ and $\mathcal{N}(\mu, 1)$ distributions in the normal case (respectively, from the $\mathscr{E}(1)$ and $\frac{1}{\mu}\mathscr{E}(\frac{1}{\mu})$ in the exponential case).

Our purpose is to achieve two goals. Firstly, we want to compare $\widehat{g}_{2_n}$ (our semiparametric projection density estimator, defined by Eq. (5) and based on case-control data $y_1, \dots, y_{n_1}, x_1, \dots, x_{n_2}$) to $\overline{g}_{2_{n_2}}$ (the nonparametric projection density estimator defined by Eq. (4) and based on control data $x_1, \dots, x_{n_2}$) by examining their MISE. Secondly, we want to estimate performances of the six following estimators:

- projection estimators on Legendre projection basis,
- projection estimators on trigonometric projection basis and
- kernel estimators

in the semiparametric and nonparametric (classic) cases.

Let $\check{g}_2$ be an estimator of the density g_2. The value of the $MISE(\check{g}_2)$ is empirically approximated by the sample average of the (estimated) Integrated Square Error $\widehat{ISE}(\check{g}_2)$ over the 500 data sets. Here, given each data set, we approximate empirically $ISE(\check{g}_2) = \int(\check{g}_2 - g_2)^2\,\mathrm{d}\mu$ by the quantity

$$\widehat{ISE}(\check{g}_2) = \frac{S_{\max} - S_{\min}}{\rho}\sum_{i=0}^{\rho}(\check{g}_2(\rho_i) - g_2(\rho_i))^2,$$

where $\rho = 200$, $\mathscr{S} = [S_{\min}, S_{\max}]$ is the interval support on which estimate the density and $\rho_i = S_{\min} + \frac{i}{\rho}(S_{\max} - S_{\min})$. This computation procedure is applied to each studied estimator.

Results are quite different if we estimate a normal or an exponential density.

In the "normal case", Table 1 shows that (semiparametric) estimators $\widehat{g}_{2_n}$ are always better than the standard (nonparametric) corresponding ones $\overline{g}_{2_{n_2}}$ in terms of MISE. The best of them seems to be the trigonometric projection

Table 1
Mean (standard deviation) of $ISE(\overline{g}_{2_{n_2}})$ and $ISE(\widehat{g}_{2_{n_2}})$ in the "normal case": each value has been multiplied by 10^3

		Legendre	Trigonometric	Kernel
$\overline{g}_{2_{n_2}}$		24.3 (31.6)	7.0 (6.4)	8.6 (8.0)
	μ			
$\widehat{g}_{2_n}$	0.25	23.1 (37.3)	4.4 (3.7)	6.1 (8.4)
	0.5	21.1 (36.5)	4.6 (3.7)	6.2 (8.5)
	0.75	16.4 (32.1)	4.8 (3.6)	5.9 (7.6)
	1	12.6 (25.9)	5.5 (6.9)	6.6 (7.5)
	1.25	9.2 (17.3)	5.9 (7.5)	6.9 (8.0)
	1.5	8.2 (14.0)	7.0 (11.8)	6.9 (8.0)

Table 2
Mean (standard deviation) of $ISE(\overline{g}_{2_{n_2}})$ and $ISE(\widehat{g}_{2_{n_2}})$ in the "exponential case": each value has been multiplied by 10^3

		Legendre	Trigonometric	Kernel
$\overline{g}_{2_{n_2}}$		12.7 (14.4)	92.8 (22.7)	43.4 (19.0)
	μ			
$\widehat{g}_{2_n}$	1.25	5.8 (6.8)	104.6 (20.9)	34.7 (12.9)
	1.5	5.8 (6.4)	124.0 (27.1)	35.8 (12.8)
	1.75	5.8 (5.6)	136.4 (28.2)	37.6 (14.7)
	2	5.2 (5.5)	151.4 (32.9)	38.1 (15.6)
	2.25	5.7 (5.5)	169.5 (37.5)	39.9 (14.5)
	2.5	6.1 (5.5)	186.0 (41.9)	42.6 (17.1)

estimator, often well adapted when $g_2(S_{\min}) \approx g_2(S_{\max})$. Nevertheless, unreported simulation results indicate that an Hermite projection estimator obtains lower MISE ($\approx 1.5 \times 10^{-3}$). This is a not surprising result since the first component of the Hermite basis *is* the density to estimate (a standard normal).

In the "exponential case", summarised in Table 2, semiparametric estimators obtain smaller ISEs than the classic associated ones. Nevertheless, this result does not hold true when we consider trigonometric projection estimators: an explanation can lie in the fact that, in this particular case, the trigonometric basis is not well adapted since the density to estimate is far from satisfying $g_2(S_{\min}) \approx g_2(S_{\max})$. So, such a projection basis should not be chosen in this case.

Remark 2. These results do not contradict the inequality (8). First of all, this inequality holds when we select the same truncation indexes; moreover, inequality (8) is an asymptotic result.

These results show that usually the semiparametric density estimator achieves a smaller MISE than that of the corresponding nonparametric one. Moreover, if the projection basis is suitable, the semiparametric projection estimator related to this basis obtains a lower MISE than that obtained by the semiparametric kernel estimator.

5.2. Data analysis

The six estimators discussed above are now applied to two familiar data sets.

5.2.1. Example 1—social quotient scores

As reported by Cheng and Chu (2004), Example 5.1, we consider the data consisting of social quotient scores of $n_2 = 21$ control children with learning disabilities and $n_1 = 20$ case children diagnosed as aphasics. Both of them were enrolled in a speech therapy program. Social quotient scores of controls and cases are, respectively, $x_i = 56, 43, 30, 97, 67, 24, 76, 49, 46, 29, 46, 83, 93, 38, 25, 44, 66, 71, 54, 20, 25$ and $y_i = 90, 53, 32, 44, 47, 42, 58, 16, 49, 54, 81, 59, 35, 81, 41, 24, 41, 61, 31, 20$. Qin and Zhang (1997), Zhang (1999) and Cheng and Chu (2004) argue that model

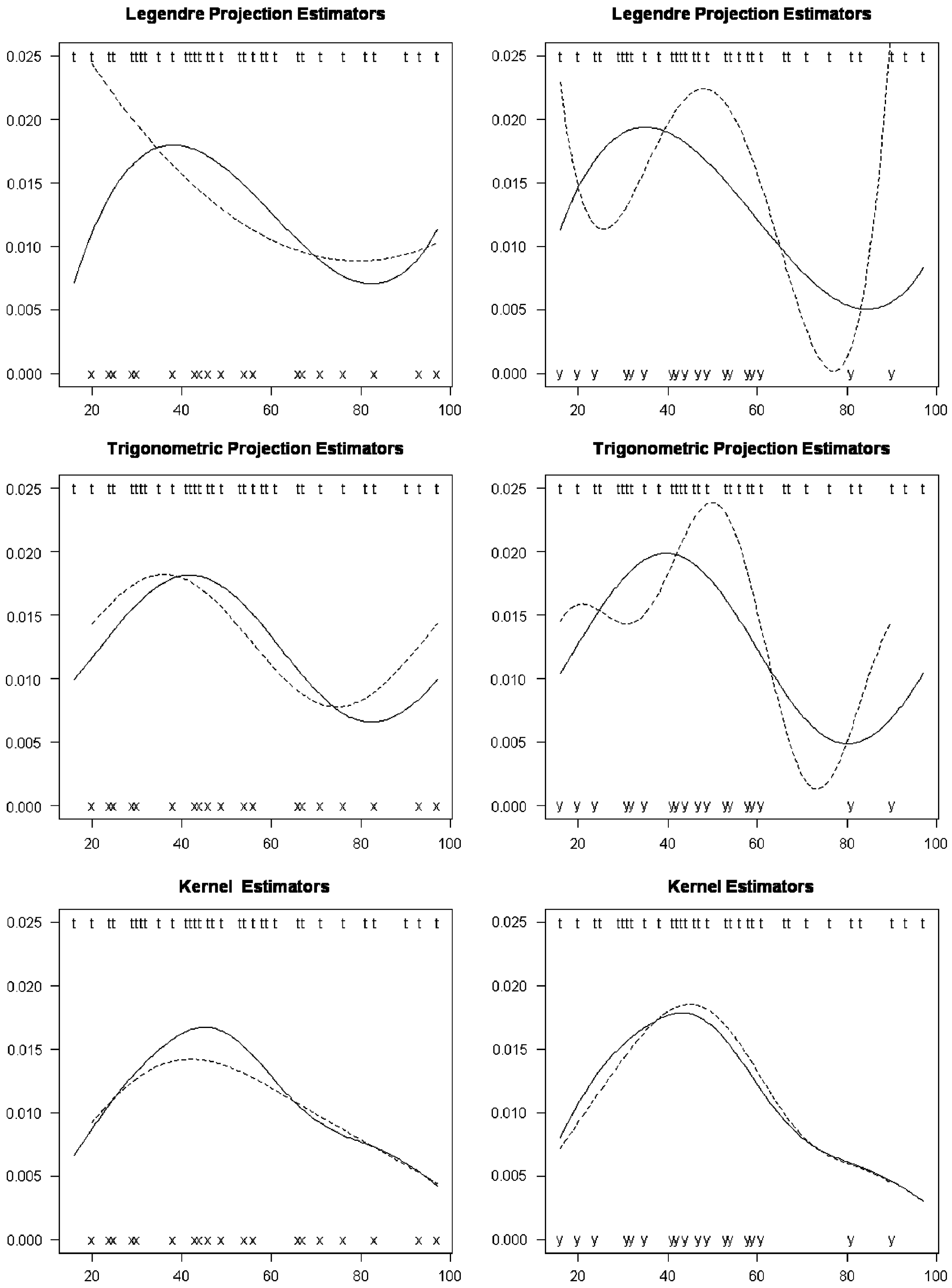
$$g_1(t) = \exp\{\theta_1 + t\theta_2\} g_2(t)$$

can be applied to the data with g_1 and g_2, respectively, standing for the densities of y_i and x_i. The maximum empirical likelihood estimate of (θ_1, θ_2) turns out to be $(\widehat{\theta_1}, \widehat{\theta_2}) = (0.396, -0.008)$.

Selection of the smoothing parameters for the kernel estimators is carried out as described above. Semiparametric and classic kernel estimators of controls' density were computed with bandwidth 10.77 and 15.04, respectively. The iterative procedure for estimation of g_1 in semiparametric and classic cases gives 9.97 and 11.58, respectively.

When we use Legendre basis, we compute the four projection estimators $(\widehat{g}_{2_n}, \widehat{g}_{1_n}, \overline{g}_{2_{n_2}}, \overline{g}_{1_{n_1}})$ with truncation indexes $(k_n, k_n, k_{n_2}, k_{n_1}) = (4, 4, 3, 5)$. These four truncation indexes are chosen according to procedure (7) with $K = 7$ and greater than 2. Similarly, when we use a trigonometric basis, we obtain $(k_n, k_n, k_{n_2}, k_{n_1}) = (3, 3, 3, 4)$.

The curves of these density estimators are shown in Fig. 1. The left panel relates to estimation of controls' density, and the right one to estimation of cases' density. Fig. 1 reveals that semiparametric estimators of g_2 are quite similar. Density for social quotient scores of controls g_2 seems slightly positively skewed and unimodal. Moreover, semiparametric and



Legendre Projection Estimators

Legendre Projection Estimators

Trigonometric Projection Estimators

Trigonometric Projection Estimators

Kernel Estimators

Kernel Estimators

Fig. 1. Example 1—social quotient scores: nonparametric (dashed lines) and semiparametric (solid lines) estimators for controls' density (left side) and cases' density (right side).

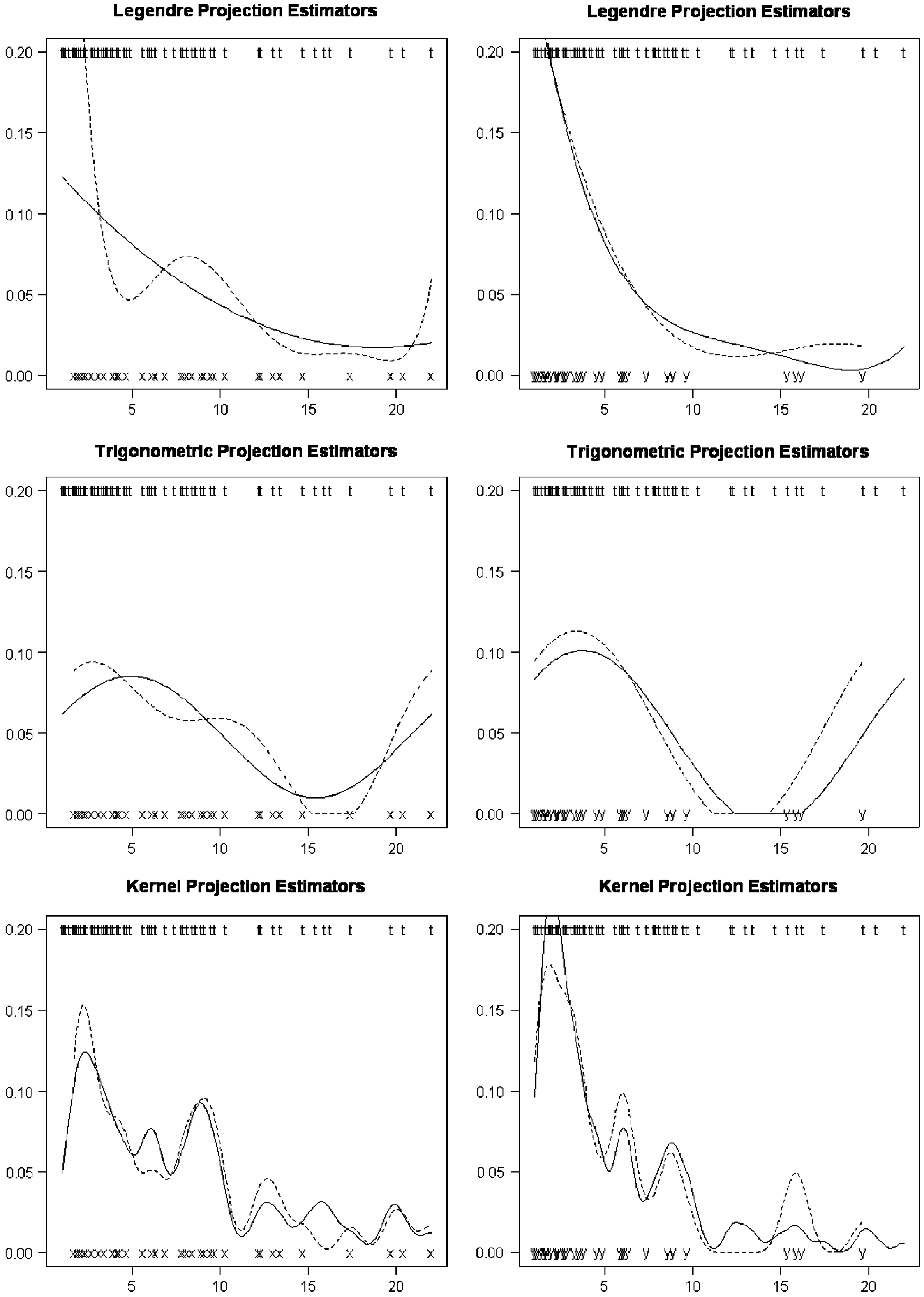


Fig. 2. Example 2—times of image recognition: nonparametric (dashed lines) and semiparametric (solid lines) estimators for NV-group's density (left side) and VV-group's density (right side).

classic estimations present relevant differences. This fact could suggest to us that the model is not appropriate to the data. The same conclusions can be drawn for estimators of cases' density g_1.

5.2.2. Example 2—times of image recognition

Consider the real data set available from http://lib.stat.cmu.edu/DASL/Stories/FusionTime.html. The purpose is to analyse how a visual information given before the experiment reduces time of recognition of the fusion of two images made of random dots stereograms. One group of $n_1 = 42$ subjects (NV group) received no visual information about the shape to recognize. A second group of $n_2 = 35$ subjects received a visual information (VV group). The data consist of the times T_{NV} and T_{VV} required to fuse the two images. After discarding an outlier for the NV group, Fokianos (2004) argues that model

$$g_1(x) = \exp(\theta_1 + \theta_2 \log(x))g_2(x)$$

can be applied to the data, where g_1 and g_2 stand, respectively, for densities of T_{NV} and T_{VV}. Given this model, we obtain $(\widehat{\theta}_1, \widehat{\theta}_2) = (-0.985, 0.623)$.

For kernel estimators of density g_2, the iterative method for the choice of bandwidth converges to 0.47 for the semiparametric estimator and to 0.60 for the nonparametric one. When we estimate the density of the NV group, bandwidths are 0.61 for the semiparametric estimator and 0.60 for the nonparametric one.

When we use Legendre basis, we compute the four projection estimators $(\widehat{g}_{2_n}, \widehat{g}_{1_n}, \overline{g}_{2_{n_2}}, \overline{g}_{1_{n_1}})$ with truncation indexes $(k_n, k_n, k_{n_2}, k_{n_1}) = (5, 3, 4, 7)$. These four truncation indexes are chosen according to the same procedure as in Example 1. For trigonometric projection estimators, we obtain $(k_n, k_n, k_{n_2}, k_{n_1}) = (3, 3, 3, 4)$.

Left and right panels of Fig. 2 illustrate density estimators for NV and VV groups, respectively. First of all, we underline that the choice of the trigonometric basis does not seem to be appropriate to these data. Indeed, observations suggest that $g(S_{\min})$ clearly differs from $g(S_{\max})$. Contrary to the previous example, semiparametric and corresponding nonparametric estimators have a similar shape. The data are positively skewed. In addition, Legendre projection estimators reveal relevant differences between the two groups. Such differences are also visible when we employ kernel estimators, but they are attenuated by irregularities. As observed in Fokianos (2004), semiparametric estimators are smoother than the corresponding nonparametric ones.

6. Conclusions and perspectives

We consider the semiparametric inference problem that is related to the density ratio model by using the methodology of empirical likelihood. This model presents some attractive features: for example, it relaxes several conventional assumptions in the context of multisamples problems.

The contribution of this work is to study the asymptotic behaviour of a new semiparametric projection estimator of the unknown probability density functions. This new estimator is obtained by merging information of all the m samples and using the methodology of empirical likelihood. The proposed semiparametric projection estimator is shown to be more efficient than the semiparametric kernel estimator for suitable projection bases, and it reduces the AMISE when it is compared with that of the traditional projection density estimator, in the sense that, for the same truncation index selection, the combined data provide estimates with the same asymptotic bias but with a smaller asymptotic variance.

The required computation for our approach can be accomplished by using the standard statistical software packages. Nevertheless, we are developing an R package that provides an flexible interface to implement our methodology.

Some authors studied a data-driven version of the projection density estimator in a nonparametric context (see Aubin and Massiani, 2003; Bosq, 2002, 2005; Picard and Tribouley, 2000). This estimator enjoys some local suroptimality properties for the AMISE. An extension of this work can be the use of this data-driven estimator (instead of the classic one) in a semiparametric context. In addition, large sample results of the discussed projection density estimators are not proved, although asymptotic normality should be expected under fairly mild conditions. These developments will be reported in the future.

Acknowledgement

The authors thank a Co-Editor, an Associate Editor, the two referees and I. Steed for their very careful reading and many insightful and valuable comments and suggestions that have greatly improved their original submission.

Appendix

Proof of Proposition 4.1.

$$\mathbb{E}[\tilde{a}_{j,n}] = \mathbb{E}\left[\sum_{i=1}^{n} p_i e_j(t_i)\right] = \sum_{l=1}^{m} n_l \int p(x) e_j(x) w_l(x) g_m(x)\,\mathrm{d}\mu(x),$$

where $p(x) = \frac{1}{\sum_{l=1}^{m} n_l w_l(x)}$, with the conventions $w(\cdot, \theta_l) =: w_l(\cdot)$ and $w_m := 1$. So

$$\mathbb{E}[\tilde{a}_{j,n}] = \int e_j(x) g_m(x)\,\mathrm{d}\mu(x) = a_j. \qquad \square$$

Proof of Theorem 4.2.

$$\begin{aligned}\mathbb{E}\|\widehat{g}_{m_n} - g_m\|^2 &= \mathbb{E}\|\widehat{g}_{m_n} - \tilde{g}_{m_n}\|^2 + \mathbb{E}\|\tilde{g}_{m_n} - g_m\|^2 + 2\mathbb{E}\left[\int (\widehat{g}_{m_n} - \tilde{g}_{m_n})(\tilde{g}_{m_n} - g_m)\,\mathrm{d}\mu\right]\\ &= A + B + C.\end{aligned}$$

We analyse the three terms of this decomposition separately.

Term A:

$$\widehat{g}_{m_n} - \tilde{g}_{m_n} = \sum_{j=1}^{k_n} e_j(\cdot)(\widehat{a}_{j,n} - \tilde{a}_{j,n}) = \sum_{j=1}^{k_n}\sum_{i=1}^{n} e_j(\cdot) e_j(t_i)(\widehat{p_i} - p_i).$$

Under classical regularity conditions (see hypotheses in Fokianos, 2004, Theorem 1), we have the existence and the limit distribution of maximum empirical likelihood estimators,

$$\sqrt{n}\begin{pmatrix}\widehat{\theta} - \theta\\ \widehat{\lambda} - \lambda\end{pmatrix} \longrightarrow \mathcal{N}(0, W),$$

where the asymptotic covariance matrix W is defined in Fokianos (2004), Appendix. By using a first-order Taylor expansion, we have

$$\widehat{p_i} = p_i + \mathcal{O}_{\mathrm{P}}(n^{-3/2}),$$

so,

$$\widehat{g}_{m_n} - \tilde{g}_{m_n}(\cdot) = \mathcal{O}_{\mathrm{P}}(n^{-3/2}) \sum_{j=1}^{k_n}\sum_{i=1}^{n} e_j(\cdot) e_j(t_i) = \mathcal{O}_{\mathrm{P}}(n^{-1/2}) \sum_{j=1}^{k_n}\left(\frac{1}{n}\sum_{i=1}^{n} e_j(t_i)\right) e_j(\cdot).$$

But

$$\sum_{j=1}^{k_n}\left(\frac{1}{n}\sum_{i=1}^{n} e_j(t_i)\right) e_j(\cdot) \longrightarrow \sum_{l=1}^{m} \rho_l g_l(\cdot)$$

as $n \to \infty$, where $\rho_l := \lim_{n\to\infty} \frac{n_l}{n}$, $\forall l \leqslant m$.

So, $\widehat{g}_{m_n} - \tilde{g}_{m_n} = \mathcal{O}_{\mathrm{P}}(n^{-1/2})$, and it follows that term A becomes

$$\mathbb{E}\|\widehat{g}_{m_n} - \tilde{g}_{m_n}\|^2 = \mathcal{O}(n^{-1}).$$

Term B:

$$\mathbb{E}\|\tilde{g}_{m_n} - g_m\|^2 = \mathbb{E}\int \left(\sum_{j=1}^{k_n}(\tilde{a}_{j,n} - a_j)e_j(\cdot) - \sum_{j>k_n} a_j e_j(\cdot)\right)^2 \mathrm{d}\mu(\cdot)$$

$$= \mathbb{E}\int \left(\sum_{j=1}^{k_n}(\tilde{a}_{j,n} - a_j)e_j(\cdot)\right)^2 \mathrm{d}\mu(\cdot) \tag{12}$$

$$- 2\mathbb{E}\int \left(\sum_{j=1}^{k_n}(\tilde{a}_{j,n} - a_j)e_j(\cdot)\sum_{l>k_n} a_l e_l(\cdot)\right) \mathrm{d}\mu(\cdot) \tag{13}$$

$$+ \mathbb{E}\int \left(\sum_{j>k_n} a_j e_j(\cdot)\right)^2 \mathrm{d}\mu(\cdot). \tag{14}$$

Observing that $\int e_j e_k \,\mathrm{d}\mu = \delta_{jk}$, we have

$$\mathbb{E}\int \left(\sum_{j=1}^{k_n}(\tilde{a}_{j,n} - a_j)e_j(\cdot)\right)^2 \mathrm{d}\mu(\cdot) = \sum_{j=1}^{k_n}\mathbb{E}(\tilde{a}_{j,n} - \mathbb{E}(\tilde{a}_{j,n}))^2$$

$$= \sum_{j=1}^{k_n}\mathrm{Var}(\tilde{a}_{j,n}).$$

Since $\mathrm{Var}(e_j(t_i)) < M^2 < \infty$ for all j, i and

$$\sum_{j=1}^{kn}\mathrm{Var}(\tilde{a}_{j,n}) \leqslant \mathcal{O}\left(\frac{1}{n^2}\right)\sum_{j=1}^{k_n}\sum_{i=1}^{n}\mathrm{Var}(e_j(t_i)),$$

terms (12)–(14) become, respectively,

$$\mathbb{E}\int \left(\sum_{j=1}^{k_n}(\tilde{a}_{j,n} - a_j)e_j(\cdot)\right)^2 \mathrm{d}\mu(\cdot) = \mathcal{O}\left(\frac{k_n}{n}\right),$$

$$\mathbb{E}\int \left(\sum_{j=1}^{k_n}(\tilde{a}_{j,n} - a_j)e_j(\cdot)\sum_{l>k_n} a_l e_l(\cdot)\right) \mathrm{d}\mu(\cdot) = 0,$$

and

$$\mathbb{E}\int \left(\sum_{j>k_n} a_j e_j(\cdot)\right)^2 \mathrm{d}\mu(\cdot) = \sum_{j>k_n} a_j^2 \int e_j^2(\cdot)\,\mathrm{d}\mu(\cdot) = \sum_{j>k_n} a_j^2.$$

So, *term B* gives

$$\mathbb{E}\|\tilde{g}_{m_n} - g_m\|^2 = \mathcal{O}\left(\frac{k_n}{n}\right) + \sum_{j>k_n} a_j^2.$$

Finally, since *term A* is negligible with respect to *term B*, the Cauchy–Schwarz inequality implies that the dominant term of $\mathbb{E}\|\widehat{g}_{m_n} - g_m\|^2$ is always $B = \mathcal{O}\left(\frac{k_n}{n}\right) + \sum_{j>k_n} a_j^2$. □

Proof of Corollary 4.3. Under hypotheses of Theorem 4.2, we have

$$\mathbb{E}\|\widehat{g}_{m_n} - g_m\|^2 = \mathcal{O}\left(\frac{k_n}{n}\right) + \sum_{j>k_n} a_j^2.$$

Now, under different conditions on Fourier coefficients a_j, we try to minimize $\mathbb{E}\|\widehat{g}_{m_n} - g_m\|^2$ with respect to k_n.

(i) $\sum_{j>k_n} a_j^2 \leqslant \sum_{j>k_n} \gamma^2 j^{-2\rho}$. Now,

$$\sum_{j>k_n} j^{-2\rho} \leqslant \frac{1}{2\rho - 1} \cdot \frac{1}{k_n^{2\rho-1}},$$

hence

$$\mathbb{E}\|\widehat{g}_{m_n} - g_m\|^2 = \mathcal{O}\left(\frac{k_n}{n}\right) + \mathcal{O}\left(\frac{1}{k_n^{2\rho-1}}\right)$$

and if we take $k_n^* = n^{1/2\rho}$, we have

$$\mathbb{E}\|\widehat{g}_{m_n} - g_m\|^2 = \mathcal{O}(n^{(1-2\rho)/2\rho}).$$

(ii)

$$\sum_{j>k_n} a_j^2 \leqslant \sum_{j\geqslant k_n} \alpha^2 \beta^{-2j} \leqslant \alpha^2 (\beta^{-2})^{k_n} \cdot \frac{1}{1-(1/\beta^2)} \leqslant c_1 \exp\{-c_2 k_n\},$$

where $c_1 = \frac{\alpha^2}{1-(1/\beta^2)}$ and $c_2 = 2\log\beta$, $(c_1, c_2 > 0)$. So $\sum_{j>k_n} a_j^2 = \mathcal{O}(e^{-c_2 k_n})$, and $\mathbb{E}\|\widehat{g}_{m_n} - g_m\|^2 = \mathcal{O}\left(\frac{k_n}{n}\right) + \mathcal{O}(\mathrm{e}^{-c_2 k_n})$.

We are looking for a sequence (k_n^*) such that $\frac{k_n^*}{n} \approx \mathrm{e}^{-c_2 k_n^*}$, that is a sequence such that $\log k_n^* + c_2 k_n^* \approx \log n$. As a first approximation, we can take $\tilde{k}_n^* = \frac{\log n}{c_2}$, hence $\frac{\tilde{k}_n^*}{n} = \frac{\log n}{c_2 n}$, and $\mathrm{e}^{-c_2 \tilde{k}_n^*} = \frac{1}{n}$.

Remark 3. For such a $(\tilde{k}_n^*)$, we obtain the following result:

$$\mathbb{E}\|\widehat{g}_{m_n} - g_m\|^2 = \mathcal{O}\left(\frac{\log n}{n}\right).$$

(iii) $(k_n) \uparrow \infty$, so, for n large enough, $\sum_{j>k_n} a_j^2 = 0$, hence

$$\mathbb{E}\|\widehat{g}_{m_n} - g_m\|^2 = \mathcal{O}\left(\frac{k_n}{n}\right).$$

Therefore, for any sequence $(v_n) \uparrow \infty$, it exists a (k_n^*) (for example, we can take $k_n^* = \sqrt{v_n}$) such that,

$$\mathbb{E}\|\widehat{g}_{m_n} - g_m\|^2 = \mathrm{o}\left(\frac{v_n}{n}\right). \qquad \square$$

Proof of Proposition 4.4. We know by Theorem 4.2 that

$$AMISE(\widehat{g}_{m_n}) = AMISE(\tilde{g}_{m_n}).$$

So, we study the $MISE(\tilde{g}_{m_n})$:

$$MISE(\tilde{g}_{m_n}) = \int \mathrm{Var}\,(\tilde{g}_{m_n}(x)) + [\mathbb{E}(\tilde{g}_{m_n}(x)) - g_m(x)]^2\,\mathrm{d}x.$$

First of all, we consider $\mathbb{E}(\tilde{g}_{m_n}(x))$:

$$\mathbb{E}(\tilde{g}_{m_n}(x)) = \mathbb{E}\left(\sum_{j=1}^{k_n}\sum_{i=1}^{n} p_i e_j(t_i) e_j(x)\right) = \sum_{j=1}^{k_n} \mathbb{E}\left(\sum_{i=1}^{n} p_i e_j(t_i)\right) e_j(x) = \sum_{j=1}^{k_n} a_j e_j(x).$$

So

$$\mathbb{E}(\tilde{g}_{m_n}(x)) - g_m(x) = -\sum_{j>k_n} a_j e_j(x).$$

Similarly, we have:

$$\mathbb{E}(\bar{g}_{m_{n_m}}(x)) - g_m(x) = -\sum_{j>k_n} a_j e_j(x).$$

Then, we consider $\mathrm{Var}(\tilde{g}_{m_n}(x))$:

$$\mathrm{Var}(\tilde{g}_{m_n}(x)) = \sum_{i=1}^{n} \left(\mathbb{E}\left(\sum_{j=1}^{k_n} p_i e_j(t_i) e_j(x) \right)^2 - \mathbb{E}^2 \left(\sum_{j=1}^{k_n} p_i e_j(t_i) e_j(x) \right) \right).$$

We have

$$\begin{aligned} \sum_{i=1}^{n} \mathbb{E}\left(\sum_{j=1}^{k_n} p_i e_j(t_i) e_j(x) \right)^2 &= \int p^2(y) \left(\sum_{j=1}^{k_n} e_j(y) e_j(x) \right)^2 g_m(y) \sum_{l=1}^{m} n_l w_l(y)\, \mathrm{d}\mu(y) \\ &= \int p(y) \left(\sum_{j=1}^{k_n} e_j(y) e_j(x) \right)^2 g_m(y)\, \mathrm{d}\mu(y) \end{aligned} \tag{15}$$

and

$$\sum_{i=1}^{n} \mathbb{E}^2 \left(\sum_{j=1}^{k_n} p_i e_j(t_i) e_j(x) \right) = \sum_{l=1}^{m} n_l \left(\int p(y) \sum_{j=1}^{k_n} e_j(y) e_j(x)\ w_l(y) g_m(y)\, \mathrm{d}\mu(y) \right)^2. \tag{16}$$

Since $\forall i$, $p(t_i) = \mathcal{O}\left(\frac{1}{n}\right)$, then (16) $= \mathcal{O}\left(\frac{1}{n^2}\right) \sum_{l=1}^{m} n_l g_l^2(x)$ as $n \to \infty$, so it is negligible with respect to (15) $= \mathcal{O}\left(\frac{k_n}{n}\right)$ $g_m(x)$. Therefore,

$$\mathrm{Var}(\tilde{g}_{m_n}(x)) = \int p(y) \left(\sum_{j=1}^{k_n} e_j(y) e_j(x) \right)^2 g_m(y)\, \mathrm{d}\mu(y) + \mathrm{o}\left(\frac{k_n}{n} \right). \tag{17}$$

By the same analysis, we get:

$$\mathrm{Var}(\bar{g}_{m_{n_m}}(x)) = \int \frac{1}{n_m} \left(\sum_{j=1}^{k_n} e_j(y) e_j(x) \right)^2 g_m(y)\, \mathrm{d}\mu(y) + \mathrm{o}\left(\frac{k_n}{n_m} \right). \tag{18}$$

We conclude the proof by noting that $\forall y$, if $\exists l < m$ such that $n_l \neq 0$, then

$$p(y) = \frac{1}{n_m + \sum_{l=1}^{m-1} n_l w_l(y)} < \frac{1}{n_m}. \qquad \square$$

Proof of Proposition 4.5. According to the proof of Theorem 4.2, we have that

$$\forall x, \quad \widehat{g}_{m_n}(x) - g_m(x) = \mathcal{O}_{\mathrm{p}}(n^{-1/2}) - \sum_{j>k_n} a_j e_j(x).$$

Since $\sqrt{n} \sum_{j>k_n} a_j e_j(x) \to \infty$, then $\mathbb{E}(\widehat{g}_{m_n}(x) - g_m(x)) = -\sum_{j>k_n} a_j e_j(x)$.

Moreover,

$$\mathrm{Var}(\widehat{g}_{m_n}(x)) = \mathrm{Var}(\widehat{g}_{m_n}(x) - \tilde{g}_{m_n}(x)) + \mathrm{Var}(\tilde{g}_{m_n}(x)) + 2\mathrm{Cov}(\widehat{g}_{m_n}(x) - \tilde{g}_{m_n}(x), \tilde{g}_{m_n}(x)).$$

Cauchy–Schwarz inequality implies

$$\mathrm{Var}(\widehat{g}_{m_n}(x)) \leqslant \mathcal{O}\left(\frac{1}{n}\right) + \mathrm{Var}(\tilde{g}_{m_n}(x)) + 2\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\sqrt{\mathrm{Var}(\tilde{g}_{m_n}(x))}$$

Since $\mathcal{O}\left(\frac{1}{n}\right) = \mathrm{o}(\mathrm{Var}(\tilde{g}_{m_n}(x)))$, we conclude that, for n large enough,

$$\mathrm{Var}(\widehat{g}_{m_n}(x)) \approx \mathrm{Var}(\tilde{g}_{m_n}(x)).$$

Formulae 17 and 18 complete the proof. □

References

Aubin, J., Massiani, A., 2003. Comportement asymptotique d'un estimateur de la densité adaptatif par méthode d'ondelettes. C. R. Acad. Sci. Paris Sér. I Math. 337 (4), 293–296.

Aubin, J.-B., Leoni-Aubin, S., 2007. Merging information for a semiparametric projection density estimation C. R. Acad. Sci. Paris Sér. I Math. 344 (5), 331–335.

Bosq, D., 2002. Estimation localement suroptimale et adaptative de la densité. C. R. Acad. Sci. Paris Sér. I Math. 334 (7), 591–595.

Bosq, D., 2005. Inférence et prévision en grandes dimensions. Economica, Paris.

Bosq, D., Lecoutre, J., 1987. Théorie de l'Estimation Fonctionnelle. Economica, Paris.

Cencov, N., 1962. Estimation of unknown distribution density from observations. Soviet Math. Dokl. 3, 1559–1562.

Cheng, K., Chu, C., 2004. Semiparametric density estimation under a two-sample density ratio model. Bernoulli 10 (4), 583–604.

Conway, J.B., 1994. A Course in Functional Analysis. Springer, Berlin.

Efromovich, S., 1999. Nonparametric Curve Estimation: Methods, Theory and Applications. Springer, New York.

Efron, B., Tibshirani, R., 1996. Using specially designed exponential families for density estimation. Ann. Statist. 24 (6), 2431–2461.

Fokianos, K., 2004. Merging information for semiparametric density estimation. J. R. Statist. Soc. B 66 (4), 941–958.

Fokianos, K., Kedem, B., Qin, J., Haferman, J., Short, D., 1998. On combining instruments. J. Appl. Meteorol. 37.

Fokianos, K., Kedem, B., Qin, J., Short, D., 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 (1), 151–194.

Gilbert, P.B., Lele, S., Vardi, Y., 1999. Maximum likelihood estimation in semiparametric selection bias models with application to aids vaccine trials. Biometrika 86 (1), 27–43.

Gill, R., Vardi, Y., Wellner, J., 1988. Large sample theory of empirical distributions in biased sampling models. Ann. Statist. 16 (3), 1069–1112.

Härdle, W., Kerkyacharian, G., Picard, D., Tsybakov, A., 1998. Wavelets, Approximation, and Statistical Applications. Lecture Notes in Statistics. Springer, New York.

Keziou, A., Leoni-Aubin, S., 2005. Test of homogeneity in semiparametric two-sample density ratio models. C. R. Acad. Sci. Paris Sér. I Math. 340 (12), 905–910.

Keziou, A., Leoni-Aubin, S., 2007. On empirical likelihood for semiparametric two-sample density ratio models. J. Statist. Plann. Inference, in press.

Kronmal, R., Tarter, M., 1968. The estimation of probability densities and cumulatives by fourier series methods. J. Amer. Statist. Assoc. 63, 925–952.

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

Owen, A., 2001. Empirical Likelihood. Chapman & Hall, New York.

Picard, D., Tribouley, K., 2000. Adaptive confidence interval for pointwise curve estimation. Ann. Statist. 28 (1), 298–335.

Qin, J., 1998. Inferences for case-control and semiparametric two-sample density ratio models. Biometrika 85 (3), 619–630.

Qin, J., Lawless, J., 1994. Empirical likelihood and general estimating equations. Ann. Statist. 22 (1), 300–325.

Qin, J., Zhang, B., 1997. A goodness-of-fit test for logistic regression models based on case-control data. Biometrika 84 (3), 609–618.

Qin, J., Zhang, B., 2005. Density estimation under a two-sample semiparametric model. Nonparametric Statist. 17 (6), 665–683.

Qin, J., Berwick, M., Ashbolt, R., Dwyer, T., 2002. Quantifying the change of melanoma incidence by breslow thickness. Biometrics 58 (3), 665–670.

Silverman, B., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall, London.

Vardi, Y., 1982. Nonparametric estimation in the presence of length bias. Ann. Statist. 10 (2), 616–620.

Vardi, Y., 1985. Empirical distributions in selection bias models. Ann. Statist. 13 (1), 178–203.

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