PROBABILITY
AND
MATHEMATICAL STATISTICS

Vol. 13, Fasc. 1 (1992), pp. 139–147

GRADE ESTIMATION OF KULLBACK–LEIBLER INFORMATION NUMBER

BY

J. MIELNICZUK (WARSZAWA)

Abstract. An estimator of the Kullback–Leibler information number by using its representation as a functional of the grade density is introduced. Its strong consistency is proved under the mild conditions on the grade density. The same approach is used to study the entropy measure of bivariate dependence (mutual information). Some applications to detection theory are also given.

1. Introduction. Let X and Y be real random variables with distribution functions F_1 and F_2, respectively. We assume that F_2 is absolutely continuous with respect to (w.r.t.) F_1. It can be seen that if F_1 is continuous, then the density g of the random variable $F_1(Y)$ w.r.t. Lebesgue measure exists. We call it the *grade density* (of F_2 w.r.t. F_1). The assumption that F_1 is continuous is imposed throughout the paper. In the paper we deal with estimation of the Kullback–Leibler information number

$$\mathrm{KL}(F_2, F_1) = \int \log\big(dF_2/dF_1(x)\big)dF_2(x), \tag{1.1}$$

where log denotes the natural logarithm. This quantity frequently appears in testing and large deviations theory as the important characteristic of the problem. Our approach to estimate (1.1) is based on the fact that under the above assumptions ensuring the existence of g, $\mathrm{KL}(F_2, F_1)$ is its functional, namely its entropy

$$\mathrm{KL}(F_2, F_1) = \int_0^1 \log g(y)\, dF_2 \circ F_1^{-1}(y) = \int_0^1 \log g(y) g(y)\, dy. \tag{1.2}$$

Thus the problem of estimating the Kullback–Leibler information number can be reduced to that of estimating the entropy of the grade density. Given any estimate of the grade density and any method of estimating the entropy it is possible, straightforwardly or with small modifications, to derive the required estimator. As to the first problem we focus on the histogram estimate of the grade density introduced in [6] and follow the approach of [9] to solve the second one. However, kernel estimates of grade density (cf. [2] and [7]) and various known estimators of entropy (see [4] for an almost exhaustive list of references for this subject) yield a handful of potential competitors to the estimators introduced in this paper. We refer also to [10] where the asymptotic behaviour of MSE for some kernel estimators of entropy is investigated.

We prove in Section 2 strong consistency of the introduced estimator under the mild conditions on the grade density. Estimation of the entropy measure of bivariate dependence (mutual information) can be viewed as the particular form of the considered problem for two dimensions. This is dealt with in Section 3. In the last section, ideas developed in the paper are applied to study the behaviour of a new detector in the detection theory.

It should be noted that it is possible to study in a similar fashion various measures of discrepancy based on the grade density, in particular those having the representation $\int H(dF_2/dF_1)dF_2$ for some measurable function H.

Let $X_1, \ldots, X_n$ be an i.i.d. sequence pertaining to d.f. F_1 and let $Y_1, \ldots, Y_n$ be an i.i.d. sequence pertaining to d.f. F_2. Let $k_n \in N$ and $b_n = (k_n)^{-1}$. Put

$$\hat{g}_n(x) = \frac{\#\{j\colon F_n^1(Y_j) \in A_{ni}\}}{nb_n} \tag{1.3}$$

for $x \in A_{ni}$, $A_{ni} = \big((i-1)b_n, ib_n\big]$ for $i = 1, 2, \ldots, n$ and $A_{n1} = [0, b_n]$; F_n^1 denotes the empirical distribution function of $X_1, \ldots, X_n$. Observe that $\hat{g}_n$ depends only on the ranks of Y's in the combined sample $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$. Denote by $\hat{h}_n(x) = \hat{g}_n(F_n^1(x))$ the corresponding estimate of the Radon–Nikodým derivative $dF_2/dF_1(x)$. We consider the following estimator of $\mathrm{KL}(F_2, F_1)$:

$$\overset{\wedge}{\mathrm{KL}}(F_2, F_1) = \int\limits_{\{z:\hat{g}_n(z) \geqslant a_n\}} \hat{g}_n(z) \log \hat{g}_n(z)\,dz, \tag{1.4}$$

where (a_n) is a sequence of positive numbers tending to 0.

2. Strong consistency of the Kullback–Leibler information number estimator. First we prove an exponential inequality for the difference between $\hat{g}_n$ and the histogram estimate based on an i.i.d. sample pertaining to g. Let

$$\tilde{g}_n(x) = \#\{j\colon F_1(Y_j) \in A_{ni}\}/nb_n$$

for $x \in A_{ni}$ and A_{ni} defined as in Section 1.

LEMMA 1. *Assume that the grade density exists and is bounded. Then*

$$\begin{aligned} &P\big(\sup_{x\in[0,1]} |\hat{g}_n(x) - \tilde{g}_n(x)| > \varepsilon\big) \\ &\qquad \leqslant C\exp(-n\varepsilon^2 b_n^2/32G_0^2) + 4C\exp(-n\varepsilon^2 b_n^2/32), \end{aligned} \tag{2.1}$$

where $G_0 = \sup g$ and C is the constant appearing in the Dvoretzky, Kiefer, Wolfowitz (DKW) *inequality* [5].

Proof. Let

$$A_n = \{\sup_x |F_n^1(x) - F_1(x)| > \delta b_n\},$$

where δ is a positive constant to be chosen later. In view of the DKW inequality we obtain

$$P(A_n) \leqslant C\exp(-2n\delta^2 b_n^2). \tag{2.2}$$

Observe that on the complement of A_n we have

$$\sup_{x\in[0,1]} |\hat{g}_n(x)-\tilde{g}_n(x)| \leqslant \max_{1\leqslant i\leqslant k_n} \#\{j\colon F_1(Y_j)\in C_{i-1,\delta}\cup C_{i\delta}\}/nb_n,$$

where $C_{i\delta} = ((i-\delta)b_n, (i+\delta)b_n]$, $i = 0, \ldots, k_n$. Let $\delta = \varepsilon/8G_0$. Then

$$b_n^{-1}\{P(F_1(Y)\in C_{i-1,\delta})+P(F_1(Y)\in C_{i\delta})\} \leqslant \varepsilon/2$$

and

$$\begin{aligned} &P(\sup|\hat{g}_n(x)-\tilde{g}_n(x)| > \varepsilon) \\ &\qquad \leqslant P\Big(\max_{1\leqslant i\leqslant k_n} \big|\#\{j\colon F_1(Y_j)\in C_{i-1,\delta}\cup C_{i\delta}\}/nb_n \\ &\qquad\qquad -\big(P(F_1(Y)\in C_{i-1,\delta})+P(F_1(Y)\in C_{i\delta})\big)/b_n\big| > \varepsilon/2\Big) \\ &\qquad \leqslant 2P\Big(\max_{0\leqslant i\leqslant k_n} |\#\{j\colon F_1(Y_j)\in C_{i\delta}\}/nb_n - P(F_1(Y)\in C_{i\delta})/b_n| > \varepsilon/4\Big). \end{aligned}$$

Observing that $\#\{j\colon F_1(Y_j)\in C_{i\delta}\}/n = G_n((i+\delta)b_n)-G_n((i-\delta)b_n)$, where G_n is the empirical distribution function pertaining to the sample $F_1(Y_1), \ldots, F_1(Y_n)$, we see that the left-hand side of (2.1) is less than

$$4P(\sup_x |G_n(x)-G(x)| > \varepsilon b_n/8) + P(A_n),$$

where $G(x) = F_2 \circ F_1^{-1}(x)$. Thus using the DKW inequality once again and taking into regard the inequality (2.2) we obtain the result.

Remark 2.1. Observe that $\sup \mathrm{E}|\tilde{g}_n - g| = O(b_n)$ for g boundedly differentiable. Using the DKW inequality for $|\tilde{g}_n - \mathrm{E}\tilde{g}_n|$ it is easy to see that in this case Lemma 1 implies that

$$\sup|\hat{g}_n - g| = O_P((\log n/nb_n^2)^{1/2} + b_n).$$

THEOREM 2.1. *Let F_1 and F_2 be distribution functions such that* $\mathrm{KL}(F_2, F_1)$ *is finite and the grade density is bounded. Assume that there exist positive constants ε_n such that*

(1) $\sum \exp(-cn\varepsilon_n^2 b_n^2) < +\infty$ for $c > 0$;

(2) $\varepsilon_n = o(a_n)$, $0 < a_n$, $a_n \to 0$.

Then $\widehat{\mathrm{KL}}(F_2, F_1) \to \mathrm{KL}(F_2, F_1)$ *a.s.*

Remark 2.2. Note that when b_n is of IMSE-optimal order $n^{-1/3}$ (see [12]) the conditions of the theorem are satisfied for $a_n = An^{-1/6+\varepsilon}$ with $\varepsilon > 0$.

Proof. We put $\hat{g} = \hat{g}_n$ and $\tilde{g} = \tilde{g}_n$ in the proof. Observe that $\tilde{g}_n$ is the histogram estimate of g based on the i.i.d. sample $F_1(Y_1), \ldots, F_1(Y_n)$ pertaining to it. Moreover, conditions (1) and (2) imply the assumptions of Theorem 2 in [9].

Thus

$$\int_{\{\tilde{g}\geqslant a_n\}} \tilde{g}\log\tilde{g} \to I(g) = \mathrm{KL}(F_2, F_1). \tag{2.3}$$

Hence, using the decomposition

$$\begin{aligned}&\int_{\{\hat{g}\geqslant a_n\}} \hat{g}\log\hat{g} - \int_{\{\tilde{g}\geqslant a_n\}} \tilde{g}\log\tilde{g}\\ &\qquad = \int_{\{\hat{g}\geqslant a_n\}} \hat{g}\log\hat{g} - \int_{\{\hat{g}\geqslant a_n\}} \tilde{g}\log\tilde{g} + \int_{\{\hat{g}\geqslant a_n\}} \tilde{g}\log\tilde{g} - \int_{\{\tilde{g}\geqslant a_n\}} \tilde{g}\log\tilde{g},\end{aligned}$$

and the triangle inequality, it is enough to prove

$$\Big|\int_{\{\hat{g}\geqslant a_n\}} \hat{g}\log\hat{g} - \int_{\{\hat{g}\geqslant a_n\}} \tilde{g}\log\tilde{g}\Big| \to 0 \text{ a.s.,} \tag{2.4}$$

$$\Big|\int_{\{\hat{g}\geqslant a_n\}} \tilde{g}\log\tilde{g} - \int_{\{\tilde{g}\geqslant a_n\}} \tilde{g}\log\tilde{g}\Big| \to 0 \text{ a.s.} \tag{2.5}$$

First we prove (2.4). Let us put

$$\begin{aligned}&\int_{\{\hat{g}\geqslant a_n\}} \hat{g}\log\hat{g} - \int_{\{\hat{g}\geqslant a_n\}} \tilde{g}\log\tilde{g}\\ &\qquad = \int_{\{\hat{g}\geqslant a_n\}} \hat{g}\log\hat{g} - \int_{\{\hat{g}\geqslant a_n\}} \hat{g}\log\tilde{g} + \int_{\{\hat{g}\geqslant a_n\}} \hat{g}\log\tilde{g} - \int_{\{\hat{g}\geqslant a_n\}} \tilde{g}\log\tilde{g} = I_1 + I_2.\end{aligned}$$

Using the inequality $|\log x| \leqslant |x-1|+|1/x-1|$ we obtain

$$|I_1| \leqslant \int_{\{x:\hat{g}(x)\geqslant a_n\}} \hat{g}(x)\big((\hat{g}(x))^{-1}+(\tilde{g}(x))^{-1}\big)\,|\hat{g}(x)-\tilde{g}(x)|\,dx. \tag{2.6}$$

In view of (1) and Lemma 1 we get $\sup|\hat{g}(x)-\tilde{g}(x)| < \varepsilon_n$ a.s. Thus, in view of (2), the integration is taken over an x such that $\tilde{g}(x) \geqslant a_n/2$ for sufficiently large n. It follows from (2.6) that

$$|I_1| \leqslant C_1\varepsilon_n/a_n \to 0 \text{ a.s.}$$

Analogously, using the inequality $|\log x| \leqslant x+1/x$ we obtain

$$\begin{aligned}|I_2| &\leqslant \int_{\{x:\hat{g}(x)\geqslant a_n\}} \big(\tilde{g}(x)+(\tilde{g}(x))^{-1}\big)|\hat{g}(x)-\tilde{g}(x)|\,dx\\ &\leqslant C_2 \sup|\hat{g}-\tilde{g}|(1+1/a_n) \to 0 \text{ a.s.}\end{aligned}$$

Observe now that, by Lemma 1 and the assumptions of the theorem,

$$\sup_x |\hat{g}(x)-\tilde{g}(x)| < a_n/2 \text{ a.s.}$$

for sufficiently large n. Thus for such n we infer that the symmetric difference of $\{x\colon \hat{g}(x) \geqslant a_n\}$ and $\{x\colon \tilde{g}(x) \geqslant a_n\}$ is a subset of $\{x\colon 3a_n/2 > \tilde{g}(x) \geqslant a_n/2\}$. Hence

$$\Big|\int_{\{\hat{g}\geqslant a_n\}} \tilde{g}\log\tilde{g} - \int_{\{\tilde{g}\geqslant a_n\}} \tilde{g}\log\tilde{g}\Big| < \Big|\int_{\{a_n/2\leqslant\tilde{g}<3a_n/2\}} \tilde{g}\log\tilde{g}\Big|$$

since the integrand of the last integral does not change sign. However, the last expression is the absolute value of the difference of two estimators of entropy truncated at the levels $3a_n/2$ and $a_n/2$, respectively. Since both the terms tend to $I(g)$ in view of (2.3), the result is proved.

Remark 2.3. Let $\hat{h}_n(x) = \hat{g}_n(F_n^1(x))$ be the corresponding estimate of the Radon–Nikodým derivative, and $Z_1, \ldots, Z_n$ be an i.i.d. sample pertaining to F_2 independent of $X_1, \ldots, X_n, Y_1, \ldots, Y_n$. Note that the alternative estimate of $\mathrm{KL}(F_1, F_2)$ may be considered:

$$\check{\mathrm{KL}}(F_2, F_1) = n^{-1} \sum_{i=1}^{n} \log \hat{h}_n(Z_i) I\{\hat{h}_n(Z_i) \geqslant a_n\}.$$

Assume that g is boundedly differentiable on $[0, 1]$ and

(1) $\sum n \exp(-cna_n^2 b_n^2) < +\infty$ for $c > 0$;

(2) $b_n = o(a_n)$, $0 < a_n$, $a_n \to 0$.

Then it can be proved by using Lemma 1 and Theorem 1 of [9] that $\check{\mathrm{KL}}$ is a strongly consistent estimator of the Kullback–Leibler information number.

3. Estimation of mutual information. Let F_{XY} be a bivariate distribution function with density $f(x, y)$ and $F \otimes G$ be the product of the respective marginal distribution functions. We assume that F_{XY} is absolutely continuous w.r.t. $F \otimes G$. The Kullback–Leibler information number for F_{XY} and $F \otimes G$ can be considered as the dependence index for F_{XY}. This index is called *mutual information* in the literature. Joe [11] introduced its estimator based on plugging the estimates of $f(x, y)$ and the marginal densities into the definition. However, separate estimation of $f(x, y)$ and marginal densities can be avoided by extending the method proposed in Section 1. This consists in estimating the density g of the copula function, i.e. the distribution function of $(F(X), G(Y))$, where (X, Y) is a bivariate random variable distributed according to F_{XY}.

Let $X_1, \ldots, X_n$ be an i.i.d. sample pertaining to F_{XY}, and F_n, G_n be empirical marginal distribution functions based on $X_1, \ldots, X_n$. Define the histogram estimate of the density of the copula function based on $X_1, \ldots, X_n$, $X_i = (X_{1i}, X_{2i})$, by

$$\hat{g}_n(x, y) = \#\{k\colon (F_n(X_{1k}), G_n(X_{2k})) \in A_{nij}\}/nb_n^2, \tag{3.1}$$

where $(x, y) \in A_{nij}$, $A_{nij} = ((i-1)b_n, ib_n] \times ((j-1)b_n, jb_n]$ for $i, j = 2, \ldots, k_n$, when one of i, j is equal to 1, the respective segment of the boundary is included into A_{nij}. Moreover, put $\hat{h}_n(x, y) = \hat{g}_n(F_n(x), G_n(y))$. For the definition and properties of kernel estimates of g see, e.g., [1] and [8].

We define $\widehat{\mathrm{KL}}$ as in (1.4). Theorem 2.1 remains valid with condition (2) replaced by

$$\sum b_n^{-2} \exp(-cn\varepsilon_n^2 b_n^2) < \infty \quad \text{for } c > 0.$$

This can be proved as above by using Theorems 1 and 2 of [9] for $d = 2$ and the lemma stated below. It is easy to see that the above condition implies conditions (2.7)–(2.9) in [9] for two dimensions. Put

$$\tilde{g}_n(x, y) = \#\{k\colon (F(X_{1k}), G(X_{2k}))\in A_{nij}\}/nb_n^2 \quad \text{for } (x, y)\in A_{nij}.$$

Lemma 2. *Assume that $g(x, y)$ exists and is bounded. Then there exist positive constants C_1 and C_2 such that*

$$P\Big(\sup_{x\in[0,1]\times[0,1]} |\hat{g}_n(x)-\tilde{g}_n(x)| > \varepsilon\Big) \leqslant C_1 b_n^{-2}\exp(-C_2 n\varepsilon^2 b_n^2)$$

for $\varepsilon/16G_0 \leqslant 1$, where $G_0 = \sup g$.

Proof. We indicate only the main lines of the proof. Outside the sets

$$A_n = \{\sup_x |F_n(x)-F(x)| > \delta b_n\}, \qquad B_n = \{\sup_x |G_n(x)-G(x)| > \delta b_n\}$$

of probability not exceeding $C\exp(-2n\delta^2 b_n^2)$ we have

$$(3.2)\qquad \sup_x |\hat{g}_n-\tilde{g}_n| \leqslant \max_{1\leqslant i,j\leqslant k_n} \big(\#\{k\colon (F(X_{1k}), G(X_{2k}))\in A_{ij\delta}\} \\ + \#\{k\colon (F(X_{1k}), G(X_{2k}))\in B_{ij\delta}\}\big)/nb_n^2,$$

where

$$A_{ij\delta} = \{((i-1-\delta)b_n, (i-1+\delta)b_n] \\ \cup ((i-\delta)b_n, (i+\delta)b_n]\}\times((j-1-\delta)b_n, (j+\delta)b_n],$$

$$B_{ij\delta} = ((i-1-\delta)b_n, (i+\delta)b_n]\times\{((j-1-\delta)b_n, (j-1+\delta)b_n] \\ \cup ((j-\delta)b_n, (j+\delta)b_n]\}.$$

Observe that

$$P\big((F(X), G(Y))\in A_{ij\delta}\big)b_n^{-2} \leqslant 4G_0\delta(1+2\delta) \leqslant \varepsilon/4$$

for $\delta = \min(1/2, \varepsilon/32G_0)$. Thus

$$P(\{\sup_x |\hat{g}_n-\tilde{g}_n| \geqslant \varepsilon\}\cap A_n^c\cap B_n^c) \\ \leqslant P(\{\max_{1\leqslant i,j\leqslant k_n} |\mu_n(A_{ij\delta})-\mu(A_{ij\delta})| \geqslant \varepsilon b_n^2/4\}) \\ +P(\{\max_{1\leqslant i,j\leqslant k_n} |\mu_n(B_{ij\delta})-\mu(B_{ij\delta})| \geqslant \varepsilon b_n^2/4\}) \\ \leqslant b_n^{-2}\{\max_{1\leqslant i,j\leqslant k_n} P(\{|\mu_n(A_{ij\delta})-\mu(A_{ij\delta})| \geqslant \varepsilon b_n^2/4\}) \\ + \max_{1\leqslant i,j\leqslant k_n} P(\{|\mu_n(A_{ij\delta})-\mu(A_{ij\delta})| \geqslant \varepsilon b_n^2/4\})\},$$

where μ is the measure pertaining to the density g, and μ_n is its empirical counterpart based on $(F(X_{1k}), G(X_{2k}))$, $k = 1, \ldots, n$. In view of the fact that $\mu(A_{ij\delta})$, $\mu(B_{ij\delta}) \leqslant \varepsilon b_n^2/4$ and using Bernstein's inequality (cf. [13], p. 96) we see that the last expression is bounded by $C_1 b_n^{-2}\exp(-C_2 n\varepsilon b_n^2)$, which completes the proof of the lemma.

4. Application in detection. Let us consider the following detection problem: An i.i.d. sample $Z_1, \dots, Z_n$ is known to have density f_1 $(I = 1)$ or f_2 $(I = 2)$. The densities f_1 and f_2 are different on a set of positive Lebesgue measure. The aim is to form the decision $\hat{I}$ classifying the sample to one of those populations on the basis of i.i.d. samples $X_1, \dots, X_k$ and $Y_1, \dots, Y_k$ pertaining to densities f_1 and f_2, respectively. Put $h(x) = f_2(x)/f_1(x)$. Let L_{nk} be the characteristic function of the set for which $I \neq \hat{I}$. The detector $\hat{I}$ is called *strongly consistent* (see, e.g., [3]) if

$$\lim_{k\to\infty} L_k^* = 0 \text{ a.s.} \quad \text{and} \quad L_k^* = \limsup_n L_{nk}.$$

We show that the approach developed in this paper yields a new proposal of a strongly consistent detector. Let g_k and h_k be the grade density and density ratio estimate, respectively, defined in Section 1 and based on the samples $X_1, \dots, X_k$ and $Y_1, \dots, Y_k$. Moreover, let $\bar{g}_k$ be an analogously defined estimate of the density $\bar{g}$ of $F_2(X_1)$ pertaining to these samples. The decision $\hat{I}$ is defined by (cf. [3], p. 280)

$$\hat{I} = \begin{cases} 2 & \text{if } \sum_{i=1}^{n} n^{-1} \log h_k(Z_i) I\{h_k(Z_i) \geqslant a_k\} > c_k, \\ 1 & \text{otherwise,} \end{cases}$$

and

$$c_k = 2^{-1}\Big\{ \int_{\{s: g_k(s) \geqslant a_k\}} g_k(s) \log g_k(s)\,ds - \int_{\{s: \bar{g}_k(s) \geqslant a_k\}} \bar{g}_k(s) \log \bar{g}_k(s)\,ds \Big\}.$$

THEOREM 4.1. *Let g be differentiable on* $[0, 1]$ *with a bounded derivative and let $\bar{g}$ be bounded. Moreover, assume that* $\mathrm{KL}(F_1, F_2)$ *and* $\mathrm{KL}(F_2, F_1)$ *are finite and there exist positive constants ε_n such that*

(1) $\sum \exp(-cn\varepsilon_n^2 b_n^2) < +\infty$ *for* $c > 0$;

(2) $b_n = O(\varepsilon_n)$, $\varepsilon_n = o(a_n)$, $a_n \to 0$.

Then $\hat{I}$ is a strongly consistent detector.

Proof. Assume without loss of generality that $I = 2$. Then conditional on $X_1, \dots, X_k$, $Y_1, \dots, Y_k$ the random variables $\log h_k(Z_i) I\{h_k(Z_i) \geqslant a_k\}$, $i = 1, \dots, n$, are i.i.d. with expected value of their negative part $> -\infty$. As in [3], p. 276, we conclude that

$$n^{-1} \sum_{i=1}^{n} \log h_k(Z_i) I\{h_k(Z_i) \geqslant a_k\} \to \int_{\{z: h_k(z) \geqslant a_k\}} f_2(z) \log h_k(z)\,dz \text{ a.s.} \quad \text{as } n \to \infty.$$

Thus

$$L_k^* = I\Big\{ \int_{\{z: h_k(z) \geqslant a_k\}} f_2(z) \log h_k(z)\,dz \leqslant c_k \Big\} \text{ a.s.}$$

Put $c = 2^{-1}\{\mathrm{KL}(F_2, F_1) - \mathrm{KL}(F_1, F_2)\}$. Since $\mathrm{KL}(F_2, F_1) > c$, it is enough to prove that

$$c_k \to c \text{ a.s.} \tag{4.1}$$

and

$$(4.2)\qquad \int_{\{z:h_k(z)\geq a_k\}} f_2(z)\log h_k(z)dz \to \mathrm{KL}(F_2, F_1).$$

The relation (4.1) is obtained by applying twice Theorem 2.1. It is easy to see that $\sup_x |\mathrm{E}\tilde{g}_k - g| = O(b_k)$ provided that g is boundedly differentiable. Thus, using Lemma 1 and the DKW inequality, for the term $\tilde{g}_k - \mathrm{E}\tilde{g}_k$ we obtain $\sup_x |g_k - g| = O(\varepsilon_k)$ a.s., and the same is true for $\sup_x |h_k - h|$. Thus as in the proof of Theorem 2.1 we show that

$$\int_{\{z:h_k(z)\geq a_k\}} f_2(z)\log h_k(z)dz - \int_{\{z:h_k(z)\geq a_k\}} f_2(z)\log h(z)dz \to 0 \text{ a.s.}$$

Moreover, it is easy to see that

$$\int_{\{z:h_k(z)< a_k\}} f_2(z)\log h(z)dz \to 0 \text{ a.s.}$$

is implied by

$$\int_{\{z:h(z)< 2a_k\}} f_2(z)\log h(z)dz \to 0.$$

However, the last expression is equal to

$$\int_{\{z:g(z)< 2a_k\}} g(z)\log g(z)dz$$

and tends to 0 since $\mathrm{KL}(F_2, F_1)$ exists. This proves (4.2).

REFERENCES

[1] K. Behnen, M. Huskova and G. Neuhaus, *Rank estimators for scores for testing independence*, Statist. Decisions 3 (1985), pp. 239–262.

[2] J. Ćwik and J. Mielniczuk, *Estimating the density ratio with application to discriminant analysis*, Comm. Statist. A 18 (8) (1989), pp. 3057–3071.

[3] L. Devroye and L. Györfi, *Nonparametric Density Estimation. The L^1 View*, Wiley, New York 1984.

[4] E. J. Dudewicz and E. C. van der Meulen, *The empiric entropy, a new approach to nonparametric entropy estimation*, in: M. L. Puri, J. Vilaplana and W. Wertz (Eds.), *New Perspectives in Theoretical and Applied Statistics*, Wiley, New York 1987, pp. 207–227.

[5] A. Dvoretzky, J. Kiefer and J. Wolfowitz, *Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator*, Ann. Math. Statist. 27 (1956), pp. 642–669.

[6] V. Gafrikova and M. Niewiadomska-Bugaj, *Point and interval estimation of discriminant index ar* (in Polish), in: *Proceedings of the Third Meeting on Discriminant Analysis*, Sobótka, Poland, 1989.

[7] I. Gijbels and J. Mielniczuk, *Estimation of Neyman–Pearson curves and related problems in discriminant analysis*, preprint, 1989.

[8] — *Estimating the density of a copula function*, Comm. Statist. A 19 (2) (1990), pp. 445–464.

[9] L. Györfi and E. C. van der Meulen, *Density-free convergence properties of various estimators of entropy*, Comput. Statist. Data Anal. 5 (1987), pp. 425–436.

[10] H. Joe, *Estimation of entropy and other functionals of multivariate density*, Ann. Inst. Statist. Math. 41 (1989), pp. 683–697.

[11] — *Relative entropy measures of multivariate dependence*, J. Amer. Statist. Assoc. 84 (1989), pp. 157–164.

[12] D. Scott, *On optimal and data-based histograms*, Biometrika 66 (1979), pp. 605–610.

[13] R. Serfling, *Approximation Theorems in Mathematical Statistics*, Wiley, New York 1980.

Institute of Computer Science
Polish Academy of Sciences
P.O. Box 22, PKiN
00-901 Warsaw, Poland

Received on 10.9.1990;
revised version on 16.4.1991