Asymptotic normality of relative entropy in multivariate density estimation

Alain Berlinet, Edward C. van Der Meulen

Pub. Inst. Stat. Paris
XXXXI, fasc. 1 et 2, 1997, 3 à 27

Asymptotic normality of relative entropy in multivariate density estimation

Alain Berlinet, Université Montpellier II
László Györfi, Technical University of Budapest
Edward C. van der Meulen, Katholieke Universiteit Leuven

January 13, 1997

Abstract

We show that the centered and normalized relative entropy error of a consistent histogram based density estimate is asymptotically normal with asymptotic variance less than or equal to 1 for all multivariate densities f which have a finite relative entropy with respect to a given reference density g and which satisfy a mild condition on the boundary of their support.

I Introduction

Density estimation is typically an intermediate tool for making statistical inferences on the actual probability law, so a meaningful error criterion for estimation of an unknown probability density function should be related to an error criterion for distribution estimation. Such error criteria can be derived from dissimilarity measures of probability measures, like the f-divergences introduced by Csiszár (1963, 1967). The two most important f-divergences in mathematical statistics and information theory are the total variation and the relative entropy.

If μ and ν are probability measures on $\mathbb{R}^d$ $(d \geq 1)$ then the total variation

of μ and ν is defined by

$$V(\mu,\nu) = \sup_A |\mu(A) - \nu(A)|,$$

where the supremum is taken over all Borel sets A.

If μ and ν are absolutely continuous with respect to a σ-finite measure λ with densities f and g respectively, then

$$\|f - g\| = \int |f(x) - g(x)|\lambda(dx) = 2\,V(\mu,\nu),$$

so that an L_1-consistent density estimate implies a distribution estimate consistent in total variation. The L_1-theory can appear more natural in the sense that it does not require any additional assumption on the density. However the relative entropy has some specific statistical properties not shared by other dissimilarity measures. For instance its additivity property is used as a basic tool in projection pursuit density estimation.

If μ is absolutely continuous with respect to the Lebesgue measure then there are L_1-consistent density estimators (histogram, kernel, etc.) for all densities, and there are distribution estimators consistent in total variation for all absolutely continuous distributions (Devroye and Györfi (1985)). However, given any sequence of density estimators $\{f_n\}$ the rate of convergence of the expected L_1 error

$$E\|f - f_n\|$$

can be arbitrary slow (Devroye (1983)).

According to these facts one can have an L_1-consistent density estimator, but its rate of convergence can be slow unless we have some conditions on the unknown f. This motivates for using a stronger mode of convergence like consistency in relative entropy. If the density estimator is the histogram then for $d = 1$ the best rate of convergence of $E\|f - f_n\|$ is of order $n^{-1/3}$ and this order can be achieved under some smoothness and tail conditions. Under the same conditions, Berlinet, Devroye and Györfi (1995) proved the asymptotic normality of the random part of the L_1 error

$$\|f - f_n\| - E\|f - f_n\|$$

with order $n^{-1/2}$, so the rate of convergence of the random part of the L_1 error is much faster than that of the expected L_1 error.

The main aim of this paper is to show the asymptotic normality of relative entropy with order $(n\sqrt{h_n})^{-1}$ and uniformly bounded asymptotic variance. The estimate under consideration was introduced by Barron (1988). It involves a density from which the true one is supposed to be not too far (in a mild sense). Such a reference density is at our disposal when dealing with parametric models or in nonparametric frameworks when a pilot estimator is available (for instance in plug-in methods or Rao-Blackwellization procedures) or when suitable information is provided on the tail of the unknown distribution. The limit distribution is independent of the dimension and of the true density when it has full support with respect to the reference density.

II Relative entropy

If μ and ν are probability measures on $\mathbb{R}^d$ then the relative entropy (information divergence, I-divergence, Kullback-Leibler information number) of μ with respect to ν is defined by

$$I(\mu,\nu) = \sup_{\{A_j\}} \sum_j \mu(A_j) \ln \frac{\mu(A_j)}{\nu(A_j)},$$

where the supremum is taken over all finite Borel measurable partitions $\{A_j\}$ of $\mathbb{R}^d$. Throughout the paper we will use the standard convention $0 \ln 0 = 0 \ln(0/0) = 0$. The following inequality, also called Pinsker's inequality, upperbounds the total variation in terms of relative entropy (cf. Csiszár (1967), Kemperman (1969) and Kullback (1967)):

$$2\{V(\mu,\nu)\}^2 \leq I(\mu,\nu),$$

which means that the relative entropy dominates the total variation.

If μ and ν are absolutely continuous with respect to a σ-finite measure λ, with densities f and g, respectively, then the relative entropy of μ with

respect to ν becomes the relative entropy $D(f,g)$ of f with respect to g, i.e.

$$I(\mu,\nu) = \int_{\mathbb{R}^d} f(x) \ln \frac{f(x)}{g(x)} \lambda(dx) = D(f,g).$$

Recently, there has been quite an interest in studying the relative entropy $I(\mu,\nu)$ and the relative entropy $D(f,g)$ when ν or g is empirical or data-based. In this regard, suppose we observe i.i.d. random variables $X_1, \ldots, X_n$ from an unknown probability distribution μ. If $\mu_n^* = \mu_n^*(.; X_1, \ldots, X_n)$ is a distribution estimate of μ, then $\{\mu_n^*\}$ is said to be consistent in relative entropy if

$$\lim_{n \longrightarrow \infty} I(\mu, \mu_n^*) = 0 \ a.s.$$

Analogously, if $X_1, \ldots, X_n$ are i.i.d. according to a probability density function f and $\{f_n^*\}$ is a sequence of density estimators, then $\{f_n^*\}$ is said to be consistent in relative entropy if $\lim_{n \longrightarrow \infty} D(f, f_n^*) = 0$ a.s. (Alternatively, one may consider convergence of $D(f, f_n^*)$ to 0 in probability or consistency in expected relative entropy which means that $\lim_{n \longrightarrow \infty} E(D(f, f_n^*)) = 0$.)

We first summarize some important results concerning the consistent estimation of a distribution or density in relative entropy, which were recently established and are relevant to this paper.

Barron, Györfi and van der Meulen (1992) showed that if one imposes a certain condition on the class of distributions from which we are estimating the unknown one, namely that there exists a known probability measure ν such that $I(\mu,\nu) < \infty$, then one can construct a distribution estimator which is a.s. consistent in relative entropy for all distributions in the class.

As is well-known, the condition $I(\mu,\nu) < \infty$ implies that μ is absolutely continuous with respect to ν. The distribution estimate proposed in Barron, Györfi and van der Meulen (1992) implies a consistent density estimate as follows: define a sequence of integers $\{m_n\}$, $0 < m_n < n$, $n \geq 2$, and let $h_n = 1/m_n$. Choose a reference density g and let ν denote the probability measure with density g. Next, introduce partitions $P_n = \{A_{n,1}, A_{n,2}, ..., A_{n,m_n}\}$, $n \geq 2$, of $\mathbb{R}^d$ such that the $A_{n,i}$'s are rectangles with $\nu(A_{n,i}) = h_n$. For

$$a_n = \frac{1}{nh_n + 1}$$

consider the following density estimate introduced by Barron (1988):

$$\begin{aligned} f_n(x) &= ((1-a_n)\mu_n(A_n(x))/h_n + a_n)g(x) \\ &= (n\mu_n(A_n(x)) + 1)a_n g(x), \end{aligned} \tag{1}$$

where μ_n stands for the empirical measure for the sample $X_1, ..., X_n$ and $A_n(x) = A_{n,i}$ if $x \in A_{n,i}$. For $d = 1$ we can get this estimate if we transform first the data into $[0,1]$ by the distribution function of g, then construct a histogram on $[0,1]$ by a uniform partition into m_n intervals, take the mixture of this histogram and the uniform density with weights $1 - a_n$ and a_n, resp., and finally transform back this mixture to the real line. The advantage of the histogram based density estimate (1) is that it avoids the problem of empty rectangles, i.e., if f is the underlying density then $D(f, f_n)$ is well defined and is finite. If we were to use the standard histogram density estimate $\hat{f}_n = \mu_n(A_n(x))/h_n$ instead of f_n then $D(f, \hat{f}_n)$ would be infinite with positive probability. We refer to f_n also as the modified histogram density estimator.

Now assume that

$$D(f, g) < \infty.$$

Then the Corollary in Barron, Györfi, van der Meulen (1992) states that if

$$\lim_{n \longrightarrow \infty} h_n = 0, \quad \lim_{n \longrightarrow \infty} nh_n = \infty, \tag{2}$$

then

$$\lim_{n \longrightarrow \infty} D(f, f_n) = 0 \ a.s.$$

and

$$\lim_{n \longrightarrow \infty} E(D(f, f_n)) = 0.$$

By using Mc Diarmid's methodology (Devroye, 1991) it is possible to get upper bounds for $Var(D(f, f_n))$ and the probability of deviation of $D(f, f_n)$ as stated in the following Lemma.

Lemma 1 *Let f_n be defined by (1) with $D(f, g) < \infty$. Then*

$$Var(D(f, f_n)) \leq \frac{n}{4} \left[\max_{1 \leq j \leq m_n} \mu(A_{nj}) \right]^2 (\ln 2)^2$$

and, for any $\epsilon > 0$,

$$P\left(|D(f,f_n) - ED(f,f_n)| \geq \epsilon\right) \leq 2\exp\left(\frac{-2\epsilon^2}{n\left[\max_{1\leq j\leq m_n}\mu(A_{nj})\right]^2(\ln 2)^2}\right).$$

The upper bounds given here are obtained through a very short and simple proof. Although they are rather crude, it is worth noticing that they are valid for any n, independent of g and (a_n) and that in the proof of Lemma 1 no hypothesis on the partition P_n is needed. As we will see Theorem 1 will provide much better asymptotic upper bounds than Lemma 1. For any given sequence of partitions and any suitable sequence $\eta = (\eta_n)$ let $\mathcal{F}_\eta$ be the set of densities f for which

$$\forall n \in \mathbb{N}^*, \qquad \max_{1\leq j\leq m_n}\mu(A_{nj}) \leq \eta_n.$$

Then we have by Lemma 1, for any $n \geq 2$, the uniform bounds

$$\sup_{f\in\mathcal{F}_\eta} Var(D(f,f_n)) \leq \frac{n}{4}\eta_n^2(\ln 2)^2$$

and

$$\sup_{f\in\mathcal{F}_\eta} P\left(|D(f,f_n) - ED(f,f_n)| \geq \epsilon\right) \leq 2\exp\left(\frac{-2\epsilon^2}{n\eta_n^2(\ln 2)^2}\right).$$

If η is chosen in such a way that

$$\eta_n = O\left(\frac{1}{\sqrt{n}}\right)$$

then we can get upper bounds for

$$Var(D(f,f_n)) \text{ and } P\left(|D(f,f_n) - ED(f,f_n)| \geq \epsilon\right)$$

which are independent of n and uniform over $\mathcal{F}_\eta$.

Györfi and van der Meulen (1994) showed how to get a density g for which $D(f,g) < \infty$. For example, for $d = 1$, the density g which is constant if $|x| < 1$ and behaves like $constant/x^2$ if $|x| \geq 1$ is good for this purpose if the differential entropy of f is finite and $E(\ln|X|)^+ < \infty$. On the other hand Györfi and van der Meulen (1994) showed also that, given any sequence of

density estimators $\{f_n^*\}$ there always exists an absolutely continuous probability measure μ with density f where f has finite differential entropy and arbitrarily many derivatives such that the sequence $\{f_n^*\}$ is not consistent in relative entropy (see also Györfi, Páli and van der Meulen (1994)).

Remark 1. Hall (1990) examined the consistency properties of the relative entropy (Kullback-Leibler loss) $(D(f, \hat{f}_n))$ between a probability density f and a standard (unmodified) histogram density estimator $\hat{f}_n$, when f is supported on the interval $[0, 1]$ and $f(x)$ and $f(1-x)$ behave like x^a and x^b, respectively, as $x \longrightarrow 0$, where $a, b \geq 0$. Hall (1990) uses $D(f, \hat{f}_n)$ as a criterion for selecting the number of bins m_n in $\hat{f}_n$. Barron and Sheu (1991) considered the estimation of a density f, also defined on a bounded interval, by an exponential family estimator $\hat{p}_n$, where $\hat{p}_n$ is found by application of the principle of minimum relative entropy subject to empirical constraints. They showed that if f satisfies a certain smoothness condition, then $D(f, \hat{p}_n)$ tends to 0 in probability at a certain rate, which depends on the smoothness condition and the type of expansion used to approximate $\ln f$.

While the result of Barron, Györfi, van der Meulen (1992) provides sufficient conditions for the consistency of a histogram based density estimator f_n in relative entropy, and in expected relative entropy, we prove in this paper a limit law for the centered relative entropy $D(f, f_n) - ED(f, f_n)$. The relative entropy $D(f, f_n)$ can be written as the sum of two terms, $D(f, f_n) - ED(f, f_n)$ and $ED(f, f_n)$, respectively. The first term $D(f, f_n) - ED(f, f_n)$ is the *random part* and represents the global error minus the expected global error. The second term, $ED(f, f_n)$, is the *nonrandom part*. Similarly to the expected L_1 error, the nonrandom part $ED(f, f_n)$ may have arbitrarily slow rate of convergence, unless some additional conditions on the smoothness and tail of f are imposed.

For the particular histogram density estimator f_n considered here, it is shown in Barron, Györfi, van der Meulen (1992) that

$$E(D(f, f_n)) \leq \frac{m_n}{n} + D(f, Ef_n).$$

For $d = 1$, Barron and Sheu (1991) proved under some strict smoothness conditions on f (f defined on a bounded interval and having finite Fisher

information) that

$$D(f, Ef_n) \leq O(1/m_n^2),$$

so it follows in this case that

$$E(D(f, f_n)) \leq \frac{m_n}{n} + O(1/m_n^2).$$

Therefore the good choice for m_n is in this case $n^{1/3}$ and then

$$E(D(f, f_n)) \leq O(n^{-2/3}).$$

Unfortunately Lemma 1 is not informative when nh_n^2 tends to infinity.
In this paper we show that without any additional condition on f, the centered relative entropy $D(f, f_n) - ED(f, f_n)$ is asymptotically normal with guaranteed rate. For the particular choice $m_n = n^{1/3}$ the rate of convergence of the random part is of order $n^{-5/6}$, which is much faster than the rate of the non-random part. This result shows that, since $D(f, f_n) - ED(f, f_n)$ is small with respect to $ED(f, f_n)$ and $D(f, f_n) - ED(f, f_n)$ has nice asymptotic behavior, all the information about the asymptotic behavior of $D(f, f_n)$ is contained in the expected global error $ED(f, f_n)$.

In proving the asymptotic normality of $D(f, f_n) - ED(f, f_n)$, obviously new aspects come in, as compared to proving the classical central limit theorem or pointwise asymptotic normality, since the global error is not a sum of independent random variables. For proving the asymptotic normality of $D(f, f_n) - ED(f, f_n)$ we use the technique of Poissonization (which stems from the fact that a multinomial distribution can be written as the conditional distribution of a set of independent Poisson random variables given their sum) and an inversion technique for obtaining characteristic functions of conditional distributions.

III Asymptotic normality of the relative entropy

In the sequel we show the asymptotic normality of the random part of the relative entropy so that the asymptotic variance is less than or equal to 1 for

all densities f for which the consistency in relative entropy of the estimate f_n is guaranteed. The rate of convergence is shown to be of order $(n\sqrt{h_n})^{-1}$.

Theorem 1 *Let μ and ν be probability measures on $\mathbb{R}^d$ with densities f and g with respect to the Lebesgue measure. Let $\bar{S}_\mu$ be the closure of the set $S_\mu = \{x : f(x) \neq 0\}$ and let f_n be given by*

$$f_n(x) = (n\mu_n(A_n(x)) + 1)a_n g(x)$$

with (h_n) satisfying

$$\lim_{n\longrightarrow\infty} h_n = 0, \quad \lim_{n\longrightarrow\infty} nh_n = \infty.$$

If $D(f,g) < \infty$ and $\nu(\bar{S}_\mu - S_\mu) = 0$ then

$$n\sqrt{2h_n}[D(f,f_n) - E(D(f,f_n))] \xrightarrow{\mathcal{D}} \mathcal{N}(0,\sigma^2)$$

as $n \longrightarrow \infty$, where $\sigma^2 = \nu(S_\mu) > 0$.

Remark 2. If Φ stands for the standard normal distribution function then by Theorem 1 for all f with $D(f,g) < \infty$ and $\epsilon > 0$

$$\begin{aligned}\lim_{n\longrightarrow\infty} P\{n\sqrt{2h_n}|D(f,f_n) - E(D(f,f_n))| > \epsilon\} &= 2\Phi\left(-\frac{\epsilon}{\sqrt{\nu(S_\mu)}}\right) \\ &\leq 2\Phi(-\epsilon),\end{aligned}$$

where the last bound is density-free. For fixed f, the asymptotic upper bounds for $Var(D(f,f_n))$ and $P\{|D(f,f_n) - E(D(f,f_n))| > \epsilon\}$ are respectively

$$\frac{1}{2h_n n^2} \text{ and } \sqrt{\frac{2}{\pi}} \exp\left(-\epsilon^2 n^2 h_n\right),$$

much better than those deduced from Lemma 1.

Remark 3: Application to goodness-of-fit. To show the usefulness of Theorem 1 in statistical inference let us consider the problem of testing the goodness-of-fit of a continuous distribution μ to a set of n observations grouped into m_n equal probability sets A_{nj}. This classical problem has

raised many questions such as:
(i) how to choose the number m_n of classes,
(ii) on which statistic should a testing procedure be based,
(iii) which criterion should be used to compare different tests?
Hoeffding (1965) (cf. Barron (1989)) showed that the likelihood ratio tests for the hypothesis testing problem $\mu = \mu_0$ versus $\mu \neq \mu_0$ based on the relative frequencies $\mu_n(A_{n,j})$ accept if $D_n(\mu_n, \mu_0) < r$ for some $r > 0$, where

$$D_n(\mu_n, \mu_0) = \sum_{j=1}^{m_n} \mu_n(A_{n,j}) \ln \frac{\mu_n(A_{n,j})}{\mu_0(A_{n,j})}.$$

If $d = 1$ and μ_0 is uniform on $[0, 1]$ then

$$D_n(\mu_n, \mu_0) = T_n - \ln h_n,$$

where

$$T_n = \sum_{j=1}^{m_n} \mu_n(A_{n,j}) \ln \mu_n(A_{n,j}).$$

Similarly, a classical goodness-of-fit test for the same hypothesis testing problem is based on (reversed order) χ^2-divergence:

$$\chi^2(\mu_n, \mu_0) = \sum_{j=1}^{m_n} \frac{(\mu_n(A_{nj}) - \mu_0(A_{nj}))^2}{\mu_0(A_{nj})}.$$

Again, if $d = 1$ and μ_0 is uniform on $[0, 1]$ then

$$\chi_n^2(\mu_n, \mu_0) = S_n/h_n - 1,$$

where

$$S_n = \sum_{j=1}^{m_n} \mu_n(A_{n,j})^2.$$

From the point of view of Pitman efficiency, Quine and Robinson (1985) proved, among other things, that the chi-square test statistic S_n and the likelihood ratio test statistic T_n are equivalent for testing the null hypothesis

$$H_0 : f(x) = 1_{(0,1)}(x)$$

versus the sequence of neighboring alternatives

$$H_{1n} : \bar{f}_n(x) = (1 + l_n h_n(x))1_{(0,1)}(x)$$

under the conditions $m_n \longrightarrow \infty$, $n/m_n \longrightarrow \infty$ and suitable conditions on $l_n \longrightarrow 0$. Now, the reversed order χ^2-divergence can be written as the square of the L_2-norm of

$$\bar{f}_n - 1_{(0,1)}$$

which was shown by Beirlant and Györfi (1994) to be the best among p^{th} powers of L_p-norms in terms of Pitman efficiency. All these statistics can be interpreted as reversed order divergences, and under the null-hypothesis these statistics are of order $n^{-1/2}$. $D_n(\mu_n, \mu_0)$ is the reversed order relative entropy restricted to the partition, and is, because of its connection to the likelihood ratio test, perhaps more common in statistics than the relative entropy $D_n(\mu_0, \mu_n)$ considered here. Because of the possible empty cells the divergence $D_n(\mu_0, \mu_n)$ can be infinite with positive probability. We can substitute μ_n by the distribution estimate μ_n^* derived from the density estimate f_n, thus

$$D_n(\mu_0, \mu_n^*) = \sum_{j=1}^{m_n} \mu_0(A_{n,j}) \ln \frac{\mu_0(A_{n,j})}{\mu_n(A_{n,j})(1 - a_n) + h_n a_n},$$

from which the test statistic can be derived

$$T_n^* = -\sum_{j=1}^{m_n} \mu_0(A_{n,j}) \ln(\mu_n(A_{n,j})(1 - a_n) + h_n a_n).$$

T_n^* is called cross-entropy by some authors (cf. Parzen (1991)). This motivates us to propose another goodness-of-fit test for testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$, as an alternative to the classical likelihood ratio test mentioned above. This test is based on $D_n(\mu_0, \mu_n^*)$ and accepts if $D_n(\mu_0, \mu_n^*) < r$ for some $r > 0$. If $d = 1$ and μ_0 is uniform on $[0, 1]$ then

$$D_n(\mu_0, \mu_n^*) = T_n^* + \ln h_n.$$

It is easy to see that

$$D(f, f_n) - ED(f, f_n) = T_n^* - ET_n^*,$$

so because of Theorem 1 the random part of the test statistic T_n^* is of order $n^{-1}h_n^{-1/2}$. These results show the need for further research on tests based on relative entropy. Also the validity of Theorem 1 in any dimension makes it possible to consider as null hypothesis any density f with finite relative entropy with respect to some density g. This is important since the reduction of the null hypothesis to the uniform density is not possible in higher dimensions.

Remark 4. From the proof of Theorem 1 (cf. (3) and (4)) it will appear that $D(f, f_n) - ED(f, f_n)$ has the same asymptotic distribution as $I_n^* - E(I_n^*)$ where

$$I_n^* = \frac{1}{2} \sum_{j=1}^{m_n} \frac{\mu(A_{n,j})(\mu_n(A_{n,j}) - \mu(A_{n,j}))^2}{(\mu(A_{n,j}) + 1/n)^2}.$$

Notice that I_n^* is again of the type of a reversed order χ^2-divergence. Heuristically, the result of Theorem 1 can now be explained as follows. Invoking the limit theory of chi-square statistics, it can be argued that under suitable technical conditions, as $n \longrightarrow \infty$ and m_n remains fixed, the statistic

$$2nI_n^* = n \sum_{j=1}^{m_n} \frac{\mu(A_{n,j})(\mu_n(A_{n,j}) - \mu(A_{n,j}))^2}{(\mu(A_{n,j}) + 1/n)^2}.$$

will have as asymptotic distribution the distribution of a chi-square random variable Y with $m_n - 1$ degrees of freedom. Now, letting $m_n = 1/h_n$ tend to infinity, we have that

$$\frac{Y - m_n}{\sqrt{2m_n}} = \sqrt{\frac{h_n}{2}}(Y - 1/h_n) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1).$$

So intuitively it should be clear that, after it has been properly centered and normalized, the random variable

$$\sqrt{\frac{h_n}{2}} 2nI_n^* = n\sqrt{2h_n} I_n^* = I_n$$

will asymptotically have a standard normal distribution as $n \longrightarrow \infty$ and $m_n \longrightarrow \infty$ under suitable technical conditions. Theorem 1 makes these conditions and this statement precise.

Proof of Theorem 1. From the finiteness of $D(f,g)$ and the definition of f_n it follows that $D(f,f_n)$ and $ED(f,f_n)$ are finite and that, denoting by λ the Lebesgue measure on $\mathbb{R}^d$, we have

$$\begin{aligned} D(f,f_n) &= \int_{\mathbb{R}^d} f(x) \ln\left(\frac{f(x)}{(n\mu_n(A_n(x))+1)a_n g(x)}\right) \lambda(dx) \\ &= D(f,g) - \ln(a_n) - \sum_{j=1}^{m_n} \mu(A_{nj}) \ln(1+n\mu_n(A_{nj})), \end{aligned}$$

and

$$\begin{aligned} D(f,f_n) - ED(f,f_n) &= -\sum_{j=1}^{m_n} \mu(A_{nj}) \left(\ln\left(\frac{1+n\mu_n(A_{nj})}{c_{nj}}\right) \right. \\ &\quad \left. -E\ln\left(\frac{1+n\mu_n(A_{nj})}{c_{nj}}\right)\right) \end{aligned}$$

where $\{c_{nj}\}$ is any sequence of positive numbers. With the intention of using a Taylor expansion of the logarithms at the point 1, we put

$$\begin{aligned} c_{nj} &= 1+n\mu(A_{nj}) \\ \text{and } R_{nj} &= \frac{n(\mu_n(A_{nj})-\mu(A_{nj}))}{c_{nj}}, \ 1 \le j \le m_n. \end{aligned}$$

Then the arguments of the logarithms in the above expression of $D(f,f_n) - ED(f,f_n)$ are equal to $(1+R_{nj})$, and we can approximate the difference between $\ln(1+R_{nj})$ and its expectation by a linear combination of R_{nj} and R_{nj}^2.
More precisely, let

$$I_n = n\sqrt{2h_n} \sum_{j=1}^{m_n} \mu(A_{nj}) \frac{R_{nj}^2}{2}, \tag{3}$$

$$\begin{aligned} J_n &= -n\sqrt{2h_n} \sum_{j=1}^{m_n} \mu(A_{nj}) \ln(1+R_{nj}) \\ \text{and } K_n &= n\sqrt{2h_n} \sum_{j=1}^{m_n} \mu(A_{nj}) R_{nj}. \end{aligned}$$

We can write

$$\begin{aligned} n\sqrt{2h_n}(D(f,f_n) - ED(f,f_n)) &= J_n - EJ_n \\ &= J_n + K_n - I_n - E(J_n + K_n - I_n) \\ &\quad -K_n + EK_n \\ &\quad +I_n - EI_n. \end{aligned}$$

For this decomposition it is good to have the following two lemmas, the proofs of which are deferred to the next section:

Lemma 2 *Under the conditions of Theorem 1 we have*

$$\lim_{n\longrightarrow\infty} E(K_n)^2 = 0.$$

Lemma 3 *Under the conditions of Theorem 1 we have*

$$\lim_{n\longrightarrow\infty} E(J_n + K_n - I_n - E(J_n + K_n - I_n))^2 = 0.$$

According to these lemmas the random variables $K_n - EK_n$ and $J_n + K_n - I_n - E(J_n + K_n - I_n)$ tend to zero in L_2. In order to prove Theorem 1 it remains to show that

$$I_n - E(I_n) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\sigma^2). \tag{4}$$

The basic technique applied is Poissonization: let N_n be a Poisson (n) random variable independent of $\{X_i\}$. The empirical measure for Poisson sample size N_n is defined as follows:

$$\mu_{N_n}(A) = \frac{\#\{i : X_i \in A, 1 \le i \le N_n\}}{n}.$$

The Poisson approximation technique is formulated by a result of Beirlant, Györfi, Lugosi (1994, Proposition): Let

$$\tilde{I}_n = \sum_{j=1}^{m_n} g_{nj}(\mu_{N_n}(A_{nj})),$$

where g_{nj} $(n, j \geq 1)$ are real measurable functions with

$$E\left(g_{nj}(\mu_{N_n}(A_{nj}))\right) = 0 \quad (n, j \geq 1).$$

Assume that, as n tends to infinity, we have for any $(t, v) \in \mathbb{R}^2$,

$$E\left(\exp\left(it\tilde{I}_n + iv\frac{N_n - n}{\sqrt{n}}\right)\right) \longrightarrow e^{-(t^2\sigma^2+v^2)/2}.$$

Then

$$\sum_{j=1}^{m_n} g_{nj}(\mu_n(A_{nj})) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2).$$

Looking at the expression of I_n it appears that the functions g_{nj} could be chosen as quadratic functions to get the asymptotic distribution of I_n from the above Proposition. More precisely, putting for $j \geq 1$

$$g_{nj}(x) = n\sqrt{2h_n}\frac{n^2\mu(A_{nj})}{2(n\mu(A_{nj})+1)^2}\left((x - \mu(A_{nj}))^2 - E(\mu_{N_n}(A_{nj}) - \mu(A_{nj}))^2\right),$$

we have

$$\sum_{j=1}^{m_n} g_{nj}(\mu_n(A_{nj})) = I_n - E(\overline{I_n}),$$

where

$$\overline{I_n} = n\sqrt{2h_n}\sum_{j=1}^{m_n}\frac{n^2\mu(A_{nj})}{2(n\mu(A_{nj})+1)^2}(\mu_{N_n}(A_{nj}) - \mu(A_{nj}))^2.$$

Now we use the following lemma the proof of which is also given in the next section.

Lemma 4 *Under the conditions of Theorem 1 we have*

$$\lim_{n\longrightarrow\infty}(E(\overline{I_n}) - E(I_n)) = 0.$$

According to the Proposition in Beirlant, Györfi and Lugosi (1994) and Lemma 3, it now suffices to show that

$$S_n = t\tilde{I}_n + v\frac{N_n - n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, t^2\sigma^2 + v^2), \tag{5}$$

$(t, v) \in \mathbb{R}^2$, in order that $I_n - E(I_n)$ has an asymptotic $\mathcal{N}(0, \sigma^2)$ distribution. We prove this fact, using Lyapunov's central limit theorem. Note that

$$Var(S_n) = \sum_{j=1}^{m_n} t^2 Var(g_{nj}(\mu_{N_n}(A_{nj})))$$

$$+2tv\sqrt{n}\sum_{j=1}^{m_n} E\left(g_{nj}(\mu_{N_n}(A_{nj}))(\mu_{N_n}(A_{nj}) - \mu(A_{nj}))\right) + v^2.$$

We wish to show that $Var(S_n) \longrightarrow t^2\sigma^2 + v^2$. First we show that

$$Var(\tilde{I}_n) = \sum_{j=1}^{m_n} Var(g_{nj}(\mu_{N_n}(A_{nj}))) \longrightarrow \sigma^2. \tag{6}$$

We have

$$Var\left(g_{nj}(\mu_{N_n}(A_{nj}))\right) = n^2 2h_n \frac{(n^2\mu(A_{nj}))^2}{4(n\mu(A_{nj})+1)^4}$$

$$\times[E(\mu_{N_n}(A_{nj}) - \mu(A_{nj}))^4 - \left(E(\mu_{N_n}(A_{nj}) - \mu(A_{nj}))^2\right)^2]$$

$$= h_n\left(\frac{n\mu(A_{nj})}{n\mu(A_{nj})+1}\right)^4 + h_n\frac{(n\mu(A_{nj}))^3}{2\left(n\mu(A_{nj})+1\right)^4}.$$

Therefore

$$Var(\tilde{I}_n) = \int_{\mathbb{R}^d}\left(\frac{nh_n z_n(x)}{nh_n z_n(x)+1}\right)^4 \nu(dx) + h_n\sum_{j=1}^{m_n}\frac{(n\mu(A_{nj}))^3}{2\left(n\mu(A_{nj})+1\right)^4},$$

where

$$z_n(x) = \frac{\mu(A_n(x))}{\nu(A_n(x))}.$$

Let

$$z(x) = \frac{d\mu}{d\nu}(x).$$

The above integral is lower bounded by

$$\int_{\mathbb{R}^d} 1_{\{z(x)>0\}}\left(\frac{nh_n z_n(x)}{nh_n z_n(x)+1}\right)^4 \nu(dx)$$

which tends to $\nu(S_\mu)$ by the dominated convergence theorem. Since

$$\sum_{j=1}^{m_n} h_n\frac{(n\mu(A_{nj}))^3}{2\left(n\mu(A_{nj})+1\right)^4} \leq \frac{h_n}{2}\sum_{j=1}^{m_n}\frac{n\mu(A_{nj})}{2(n\mu(A_{nj})+1)^2} = \frac{\overline{C}_n}{2} \longrightarrow 0,$$

it follows that

$$\liminf_{n\longrightarrow\infty} Var(\tilde{I}_n) \geq \nu(S_\mu).$$

Now

$$\begin{aligned} h_n \sum_{j=1}^{m_n} \left(\frac{n\mu(A_{nj})}{n\mu(A_{nj})+1} \right)^4 &\leq h_n \sum_{j=1}^{m_n} \frac{n\mu(A_{nj})}{n\mu(A_{nj})+1} \\ &= \int_{\mathbb{R}^d} 1_{\{z_n(x)>0\}} \frac{nh_n z(x)}{nh_n z_n(x)+1} \nu(dx). \end{aligned}$$

If one denotes the above integral by $\overline{D}_n$ one can prove in the same way as in the proof of Lemma 2 that

$$\overline{D}_n - D_n \longrightarrow 0,$$

where

$$\begin{aligned} D_n &= \int_{\mathbb{R}^d} 1_{\{z_n(x)>0\}} \frac{nh_n z(x)}{nh_n z(x)+1} \nu(dx) \\ &\leq \int_{\mathbb{R}^d} \frac{nh_n z(x)}{nh_n z(x)+1} \nu(dx) \longrightarrow \nu(S_\mu). \end{aligned}$$

Thus

$$\limsup_{n\longrightarrow\infty} Var(\tilde{I}_n) \leq \nu(S_\mu),$$

which proves (6). To complete the asymptotics for $Var(S_n)$ it remains to show that

$$\sqrt{n} \sum_{j=1}^{m_n} E\left(g_{nj}(\mu_{N_n}(A_{nj}))(\mu_{N_n}(A_{nj}) - \mu(A_{nj}))\right) \longrightarrow 0.$$

This follows immediately from the fact that

$$E\left(\mu_{N_n}(A_{nj}) - \mu(A_{nj})\right)^3 = \frac{\mu(A_{nj})}{n^2}.$$

Hence $Var(S_n) \longrightarrow t^2\sigma^2 + v^2$. To finish the proof of (5), by Lyapunov's central limit theorem it suffices to show that

$$\sum_{j=1}^{m_n} E\left|g_{nj}(\mu_{N_n}(A_{nj}))\right|^3 + n^{-3/2} \sum_{j=1}^{m_n} E\left|n\mu_{N_n}(A_{nj}) - n\mu(A_{nj})\right|^3 \longrightarrow 0.$$

Using moment properties of Poisson variables one gets

$$n^{-3/2}\sum_{j=1}^{m_n} E\,|n\mu_{N_n}(A_{nj}) - n\mu(A_{nj})|^3 \le \sqrt{3}\max_{1\le j\le m_n}\sqrt{\mu(A_{nj})} + \frac{1}{\sqrt{n}}$$

and

$$\begin{aligned}\sum_{j=1}^{m_n} E\,|g_{nj}(\mu_{N_n}(A_{nj}))|^3 \le \quad & (2h_n)^{3/2}\sum_{j=1}^{m_n}\left(\frac{n^3\mu(A_{nj})}{2(n\mu(A_{nj})+1)^2}\right)^3 \\ & \times\frac{150}{n^6}((n\mu(A_{nj}))^3 + (n\mu(A_{nj}))^2 + n\mu(A_{nj})) \\ < \quad & 160\sqrt{h_n}\,\overline{D}_n.\end{aligned}$$

It is easy to see that both terms tend to zero. So S_n is asymptotically $\mathcal{N}(0, t^2\sigma^2 + v^2)$ and the conclusion follows.

IV Proofs of the lemmas

Proof of Lemma 1. As shown in the proof of Theorem 1, we have

$$D(f, f_n) = D(f, g) - \ln(a_n) - \xi(X_1, \cdots, X_n)$$

with ξ defined on $\mathbb{R}^n$ by

$$\xi(x_1, \cdots, x_n) = \sum_{j=1}^{m_n}\mu(A_{nj})\ln(1 + N_j)$$

where

$$N_j = \sum_{i=1}^{n} 1_{A_{nj}}(x_i).$$

Suppose that x_i belongs to A_{nk} and that x_i' belongs to $A_{nk'}$ with $k \neq k'$. We have

$$\begin{aligned}\xi(x_1, \cdots, x_i', \cdots, x_n) - \xi(x_1, \cdots, x_i, \cdots, x_n) \quad = \quad & \mu(A_{nk})\ln\left(\frac{N_k}{N_k + 1}\right) \\ & +\mu(A_{nk'})\ln\left(\frac{N_{k'} + 2}{N_{k'} + 1}\right)\end{aligned}$$

with $1 \leq N_k \leq n$ and $0 \leq N_{k'} \leq n-1$.
Therefore

$$|\xi(x_1, \cdots, x_i', \cdots, x_n) - \xi(x_1, \cdots, x_i, \cdots, x_n)| \leq \max_{1 \leq j \leq m_n} \mu(A_{nj}) \ln 2.$$

We are now in a position to apply Mc Diarmid's methodology to the function ξ (see Devroye (1991), Theorem 2 and 3). The two inequalities follow.

Proof of Lemma 2. Applying the fact that

$$\bar{K}_n = n\sqrt{2h_n} \sum_{j=1}^{m_n} (\mu_n(A_{nj}) - \mu(A_{nj})) = 0$$

it is enough to prove that

$$\lim_{n \longrightarrow \infty} E(K_n - \bar{K}_n)^2 = 0.$$

For $j \neq k$, we have

$$E((\mu_n(A_{nj}) - \mu(A_{nj}))(\mu_n(A_{nk}) - \mu(A_{nk}))) = -\frac{1}{n}\mu(A_{nj})\mu(A_{nk})$$

thus

$$\begin{aligned} E(K_n - \bar{K}_n)^2 &= E\left(n\sqrt{2h_n} \sum_{j=1}^{m_n} \frac{1}{n\mu(A_{nj}) + 1} (\mu_n(A_{nj}) - \mu(A_{nj})) \right)^2 \\ &\leq n^2 2h_n \sum_{j=1}^{m_n} \frac{1}{(n\mu(A_{nj}) + 1)^2} E\left(\mu_n(A_{nj}) - \mu(A_{nj})\right)^2 \\ &= n^2 2h_n \sum_{j=1}^{m_n} \frac{1}{(n\mu(A_{nj}) + 1)^2} \frac{\mu(A_{nj})}{n} (1 - \mu(A_{nj})) \\ &\leq 2\, C_n, \end{aligned}$$

with

$$C_n = h_n \sum_{j=1}^{m_n} \frac{n\mu(A_{nj})}{(n\mu(A_{nj}) + 1)^2} = \int_{\mathbb{R}^d} \phi(nh_n z_n(x))\ \nu(dx).$$

and

$$\phi(t) = \frac{t}{(t+1)^2} \leq \frac{1}{4} \quad \text{if} \quad t \geq 0.$$

If $x \in S_\mu$ then, almost surely, $z_n(x)$ tends to $z(x) > 0$ and $\phi(nh_n z_n(x))$ tends to 0.
If $x \in \mathbb{R}^d - \bar{S}_\mu$ then x is in the interior of $\{x : f(x) = 0\}$. Thus, for n large enough, $z_n(x) = 0$ and $\phi(nh_n z_n(x)) = 0$. As $\nu(\bar{S}_\mu - S_\mu) = 0$, the dominated convergence theorem implies that C_n tends to 0. This ends the proof of Lemma 2.

Proof of Lemma 3. First note that

$$R_{nj} = n(\mu_n(A_{nj}) - \mu(A_{nj}))/(n\mu(A_{nj}) + 1)$$

is a random variable taking its values in $[\delta_{nj}^-, \delta_{nj}^+]$ with

$$-1 < \frac{-n}{n+1} \leq \delta_{nj}^- = \frac{-n\mu(A_{nj})}{n\mu(A_{nj}) + 1} \leq 0$$

and

$$0 \leq \delta_{nj}^+ \leq \frac{n}{n\mu(A_{nj}) + 1} \leq n.$$

Thus one can write

$$J_n + K_n - I_n - E(J_n + K_n - I_n)$$

$$= n\sqrt{2h_n} \sum_{j=1}^{m_n} \mu(A_{nj})[\varphi(R_{nj}) - E(\varphi(R_{nj})]$$

where $\varphi : (-1, +\infty) \longrightarrow \mathbb{R}$ is defined by

$$\varphi(x) = -log(1 + x) + x - \frac{x^2}{2}.$$

The function φ is strictly decreasing and satisfies $\varphi(0) = 0$. Let $\alpha > 1/3$ and $\beta = -1 + 1/3\alpha$ such that $-n/(n+1) < \beta$. We have

$$|\varphi(R_{nj})| \leq \alpha |R_{nj}|^3 \qquad if \quad R_{nj} \geq \beta$$

$$\text{and} \qquad |\varphi(R_{nj})| < log(n\mu(A_{nj}) + 1) \qquad if \quad R_{nj} < 0.$$

For $j \in \{1, \cdots, m_n\}$, R_{nj} is a strictly increasing function of $\mu_n(A_{nj})$. Denote its inverse by R_{nj}^{-1}. Using Hoeffding's formula (Dharmadhikari and Joag-dev,

1987, p 148) we get

$$\begin{aligned} Cov(\varphi(R_{nj}),\varphi(R_{nk})) &= \iint_{\mathbb{R}^2} \{P(\varphi(R_{nj}) \le x, \varphi(R_{nk}) \le y) \\ &- P(\varphi(R_{nj}) \le x) P(\varphi(R_{nk}) \le y)\}\, dx\, dy \end{aligned}$$

$$\begin{aligned} &= \iint_{\mathbb{R}^2} \left\{ P\left(\mu_n(A_{nj}) \ge R_{nj}^{-1}(\varphi^{-1}(x)), \mu_n(A_{nk}) \ge R_{nk}^{-1}(\varphi^{-1}(y))\right)\right. \\ &- \left. P\left(\mu_n(A_{nj}) \ge R_{nj}^{-1}(\varphi^{-1}(x))\right) P\left(\ \mu_n(A_{nk}) \ge R_{nk}^{-1}(\varphi^{-1}(y))\ \right)\right\} dx\, dy. \end{aligned}$$

This last quantity is non-positive for $j \neq k$ by a result of Mallows (1968) on multinomial probabilities. Therefore for $j \neq k$

$$E\left([\varphi(R_{nj}) - E(\varphi(R_{nj}))][\varphi(R_{nk}) - E(\varphi(R_{nk}))]\right) \le 0.$$

Thus

$$E(J_n + K_n - I_n - E(J_n + K_n - I_n))^2$$

$$\begin{aligned} &\le n^2 2h_n \sum_{j=1}^{m_n} \mu(A_{nj})^2 E[\varphi(R_{nj}) - E(\varphi(R_{nj})]^2 \\ &\le n^2 2h_n \sum_{j=1}^{m_n} \mu(A_{nj})^2 E[\varphi(R_{nj})^2] \\ &\le n^2 2h_n \sum_{j=1}^{m_n} \mu(A_{nj})^2 \alpha^2 E[R_{nj}^6] \\ &\quad + n^2 2h_n \sum_{j=1}^{m_n} \mu(A_{nj})^2 \left(log(n\mu(A_{nj}) + 1)\right)^2 P(R_{nj} < \beta) \\ &\le n^2 2h_n \alpha^2 \sum_{j=1}^{m_n} \mu(A_{nj})^2 \frac{n\mu(A_{nj}) + 25(n\mu(A_{nj}))^2 + 15(n\mu(A_{nj}))^3}{(n\mu(A_{nj}) + 1)^6} \\ &\quad + n^2 2h_n \sum_{j=1}^{m_n} \mu(A_{nj})^2 \left(log(n\mu(A_{nj}) + 1)\right)^2 exp\left\{-\frac{n}{4}\beta^2 \mu(A_{nj})\right\} \\ &\le 82h_n \alpha^2 \sum_{j=1}^{m_n} \frac{n\mu(A_{nj})}{(n\mu(A_{nj}) + 1)^2} \\ &\quad + 2\int_{z(x)>0} [nh_n z_n(x) log(nh_n z_n(x) + 1)]^2 exp\left\{-nh_n z_n(x)\frac{\beta^2}{4}\right\} \nu(dx). \end{aligned}$$

In the next to last inequality Bennett's inequality (1962) was applied to the zero mean random variables

$$\Big(1_{A_{nj}}(X_i) - \mu(A_{nj})\Big)_{1 \le i \le n}$$

as in Lemma 3 of Györfi and van der Meulen (1987). The first term of the right hand side of the last inequality tends to 0 according to the proof of Lemma 2. For the second term, the integrand tends to 0 in a dominated way, so it tends to 0, too.

Proof of Lemma 4.

$$E(\overline{I_n}) - E(I_n) = n\sqrt{2h_n} \sum_{j=1}^{m_n} \frac{n^2 \mu(A_{nj})}{2(n\mu(A_{nj}) + 1)^2} \qquad [E((\mu_{N_n}(A_{nj}) - \mu(A_{nj}))^2)$$

$$- \quad E((\mu_n(A_{nj}) - \mu(A_{nj}))^2)]$$

hence

$$E(\overline{I_n}) - E(I_n) = \sqrt{2h_n} \sum_{j=1}^{m_n} \frac{(n\mu(A_{nj}))^2}{2(n\mu(A_{nj}) + 1)^2} \mu(A_{nj}) \le \frac{\sqrt{2h_n}}{2} \longrightarrow 0.$$

Acknowledgement. We would like to thank Igor Vajda who pointed out the missing condition $\nu(\bar{S}_\mu - S_\mu) = 0$ in an earlier version of the paper.

References

[1] Barron, A. R. (1988). The convergence in information of probability density estimators. Presented at *IEEE ISIT, Kobe, Japan, June 19-24,1988.*

[2] Barron, A. R. (1989) Uniformly powerful goodness of fit tests. *Annals of Statistics*, 17, pp. 107-124.

[3] Barron, A. R., Györfi, L. and van der Meulen, E. C. (1992). Distribution estimates consistent in total variation and in two types of information divergence. *IEEE Trans. on Information Theory*, 38, pp. 1437-1454.

[4] Barron, A. R. and Sheu, C. (1991). Approximation of density functions by sequences of exponential families. *Annals of Statistics*, 19, pp. 1347-1369.

[5] Beirlant, J. and Györfi, L. (1994) Pitman efficiencies of L_p-goodness-of-fit tests *Kybernetika*, 30, pp. 223-232.

[6] Beirlant, J., Györfi, L. and Lugosi, G. (1994). On the asymptotic normality of the L_1- and L_2-errors in histogram density estimation. *Canadian J. Statistics*, 3, pp. 309-318.

[7] Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *J. Amer. Statist. Assoc.*, 57, pp. 33-45.

[8] Berlinet, A., Devroye, L. and Györfi, L. (1994). Asymptotic normality of L_1 error in density estimation. *Statistics*, 26, pp. 329-343.

[9] Csiszár, I. (1963). Eine informationstheoretische ungleichung und ihre anwendung auf den Beweis der ergodizität von Markoffschen Ketten. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 8, pp. 85-107.

[10] Csiszár, I. (1967). Information-type measures of divergence of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2, pp. 299-318.

[11] Devroye, L. (1983). On arbitrary slow rates of global convergence in density estimation. *Zeitschrift für Wahscheinlichkeitstheorie und verwandte Gebiete*, 62, pp. 475-483.

[12] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: the L_1 View.* Wiley.

[13] Devroye, L. (1991) Exponential inequalities in nonparametric estimation, in *Nonparametric Functional Estimation and Related Topics* (G. Roussas, ed), Kluwer Academic Publishers, pp. 31-44.

[14] Dharmadhikari, S. and Joag-dev, K. (1987). *Unimodality, convexity and applications.* Academic Press.

[15] Györfi, L., Páli, I. and van der Meulen, E. C. (1994). There is no universal source code for infinite alphabet. *IEEE Trans. on Information Theory*, 40, pp. 267-271.

[16] Györfi, L. and van der Meulen, E. C. (1987). Density-free convergence properties of various estimators of entropy. *Computational Statistics and Data Analysis*, 5, pp. 425-436.

[17] Györfi, L. and van der Meulen, E. C. (1991) A consistent goodness-of-fit test based on the total variation distance, in *Nonparametric Functional Estimation and Related Topics* (G. Roussas, ed), Kluwer Academic Publishers, pp. 631-645.

[18] Györfi, L. and van der Meulen, E. C. (1994). There is no density estimate consistent in information divergence for all densities. *Transactions of the Twelfth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, pp. 88-90.

[19] Hall, P. (1990). Akaike's information criterion and Kullback-Leibler loss for histogram density estimation. *Probability Theory and Related Fields*, 85, pp. 449-467.

[20] Hoeffding, W. (1965). Asymptotically optimal tests for multinomial distributions. *Ann. Math. Statist.*, 36, pp. 369-408.

[21] Kemperman, J. H. B. (1969). An optimum rate of transmitting information. *Ann. Math. Statist.*, 40, pp. 2156-2177.

[22] Kullback, S. (1967). A lower bound for discrimination in terms of variation. *IEEE Trans. Information Theory*, 13, pp. 126-127.

[23] Mallows, C. L. (1968). An inequality involving multinomial probabilities. *Biometrika*, 55, pp. 422-424.

[24] Parzen, E. (1991). Goodness of fit tests and entropy. *Journal of Combinatorics, Information and System sciences*, 16, pp. 129-136.

[25] Quine, M. P. and Robinson J. (1985) Efficiences of chi-square and likelihood ratio goodness-of-fit tests. *Annals of Statistics*, 13, pp. 727-742.