ISSN 1066-5307, Mathematical Methods of Statistics, 2008, Vol. 17, No. 3, pp. 261–277. © Allerton Press, Inc., 2008.

k_n-Nearest Neighbor Estimators of Entropy

R. M. Mnatsakanov[1*], N. Misra[2], Sh. Li[1], and E. J. Harner[3]

[1] West Virginia Univ. and National Inst. for Occupational Safety and Health, Morgantown, USA

[2] West Virginia Univ., Morgantown, USA, and Indian Inst. of Technology, Kanpur, India

[3] West Virginia Univ., Morgantown, USA

Received October 21, 2006; in final form, August 7, 2008

Abstract—For estimating the entropy of an absolutely continuous multivariate distribution, we propose nonparametric estimators based on the Euclidean distances between the n sample points and their k_n-nearest neighbors, where $\{k_n : n = 1, 2, \dots\}$ is a sequence of positive integers varying with n. The proposed estimators are shown to be asymptotically unbiased and consistent.

Key words: entropy estimator, k_n-nearest neighbor.

2000 Mathematics Subject Classification: primary 62G05; secondary 94A17.

DOI: 10.3103/S106653070803006X

1. INTRODUCTION AND PRELIMINARIES

Problems in molecular science, fluid mechanics, chemistry, physics, etc., often require the estimation of the unknown differential entropy $H(f)$ (see (1) below) of the probability density function (p.d.f.) f, also known as the Shannon entropy. Researchers at the National Institute for Occupational Safety and Health are modeling the random fluctuations of molecules by means of molecular dynamics simulation in order to study their properties and functions. We focus on estimating the configurational entropy, which measures the freedom of the molecular system to explore its available configuration space. Specifically, entropy evaluation is necessary to understand the stability of a conformation and the changes that take place in going from one conformation to another. The entropy of a molecular conformation depends on the following variables: bond lengths, bond angles, and torsional (or dihedral) angles. Entropy is mainly determined by the fluctuations in torsional angles since bond lengths and bond angles are rather stable variables. The difficulty is that the number of torsional angles can be quite large, i.e., the dimension p of the conformation space can be high.

An attempt to construct models describing the fluctuations of torsional angles in molecules (different from the normal distribution), was presented in Demchuk and Singh [1]. It was shown that the von Mises distribution can be used when modeling the torsional angles in certain cases.

The form of f is complex and unknown, except for certain simple molecules such as methanol. Since the marginal distributions of the fluctuations of molecular dihedral angles generally exhibit multimodes and skewness, a nonparametric approach to entropy estimation seems appropriate. The construction of entropy estimates based on k-nearest neighbor (kNN) techniques will help us to overcome the dimensionality problem.

kNN estimation techniques have been studied extensively in both the computer science and statistical literature; the estimators have a simple construction and can be easily implemented computationally. However, algorithms implementing smoothing kernel density estimation techniques generally have fallen short in high-dimensional cases with massive data sets (see, for example, Scott [2]).

To describe the construction of the k-nearest neighbor estimator of entropy, let $X_1, \dots, X_n$ be n independent copies of a p-dimensional random variable (r.v.) X with a common distribution having a

[*] E-mail: rmnatsak@stat.wvu.edu

p.d.f. f (with respect to the Lebesgue measure defined on a support of f). The entropy of f (or of the r.v. X) is defined by

$$H(f) = \mathbf{E}_f(-\log f(X)) = -\int_{\mathbb{R}^p} f(x)\log f(x)\,dx, \tag{1}$$

where $\mathbb{R}^p$ denotes the p-dimensional space. In the sequel, to avoid the uncertainty (when $f(x) = 0$), we consider the integration in (1) over the set $D = \{x \in \mathbb{R}^p\colon f(x) > 0\}$. Based on a random sample $X_1, \ldots, X_n$, our goal is to estimate the entropy $H(f)$, as defined in (1), and establish the L_2-consistency of corresponding k-nearest neighbor estimators as $k = k_n \to \infty$ and $n \to \infty$.

At first, let us introduce the construction based on the k-nearest neighbor density estimators studied by Kozachenko and Leonenko [3] for $k = 1$, and by Singh *et al.* [4] and Goria *et al.* [5], for $k \geq 1$. Namely, the following asymptotically unbiased and consistent estimator has been proposed:

$$\hat{H}_{k,n} = \frac{p}{n}\sum_{i=1}^{n}\log d_{i,k,n} + \log\left(\frac{\pi^{p/2}}{\Gamma(p/2+1)}\right) + \gamma - L_{k-1} + \log n, \tag{2}$$

where, for each $k \leq n-1$, $d_{i,k,n}$ is the Euclidean distance between X_i and its kth nearest neighbor among $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$, and $\gamma = 0.5772\ldots$ is the Euler constant, $L_0 = 0$, $L_j = \sum_{i=1}^{j} i^{-1}$, $j = 1, 2, \ldots$

Note that the rate of convergence of $\hat{H}_{k,n}$ is still not derived when $p \geq 2$. Another very important question related to this construction is the derivation of the optimal (for instance, in the sense of Root Mean Squared Error (RMSE)) values of k as a function of the sample size n and dimension p. Currently, we are working on these questions, which are addressed partially via the simulation studies in Section 3.

In the one-dimensional case ($p = 1$), several authors have proposed estimators of entropy in the context of goodness-of-fit testing (Vasicek [6], Dudewicz and van der Meulen [7]):

$$\hat{V}_{m,n} = \frac{1}{n-m}\sum_{i=1}^{n-m}\log(X_{(i+m)} - X_{(i)}) - \psi(m) + \log n,$$

where $m\,(\leq n-1)$ is a fixed positive integer, $X_{(1)}, \ldots, X_{(n)}$ are the order statistics, and $\psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$ is the digamma function (see Abramowitz and Stegun [8] for the properties of the gamma $\Gamma(\cdot)$ and digamma functions). Under certain conditions the asymptotic normality of $\hat{V}_{m,n}$ is proved. Namely, as $n \to \infty$,

$$\sqrt{n}\left(\hat{V}_{m,n} - H(f)\right) \xrightarrow{d} N\big(0, (2m^2 - 2m + 1)\,\psi'(m) - 2m + 1 + \mathrm{Var}\{\log f(X_1)\}\big). \tag{3}$$

Let us note that the asymptotic variance of the estimator $\hat{H}_{k,n}$, when k is fixed and $n \to \infty$, has the following form:

$$\mathrm{Var}\{\hat{H}_{k,n}\} \sim \frac{\psi'(k) + \mathrm{Var}\{\log f(X_1)\}}{n} \tag{4}$$

(see, Singh *et al.* [4], formula (15)). Here the function $\psi'(k) = \sum_{j=k}^{\infty} j^{-2}$ is a decreasing function of k, so that the larger the value of k, the smaller the variance of $\hat{H}_{k,n}$. That is why it is reasonable to assume that k should be an increasing function of n. Monte Carlo simulations can be done to obtain practical rules of thumb for the order of k. For example, in simulation studies (see Goria *et al.* [5]), a heuristic formula $k = [\sqrt{n} + 0.5]$ was suggested for small sample sizes in the range of 10–10^2. They choose the beta, gamma, Laplace, Cauchy, and Student-t distributions for simulating the data. In Section 3, the simulations from the uniform, beta, gamma, normal, von Mises, bivariate circular, and log-normal distributions with sample sizes n in the range 50–10^4 are conducted. We approximated the optimal rates of k_n (via the minimization of RMSE) for different distributions and derived $k_n = [c\,\log^2 n]$, where, for example, in the univariate uniform model, $c = 0.44$. For the choices of c in other models see the lines before Table 1 in Section 3.

Now compare the asymptotic variances of $\hat{V}_{m,n}$ and $\hat{H}_{k,n}$ defined in (3) and (4), respectively, as $n \to \infty$, and $k = m$, $p = 1$. The comparison shows that an increasing k yields a better behavior for $\hat{H}_{k,n}$ (compared with $\hat{V}_{m,n}$). Indeed, in Singh *et al.* [4], it was proved that $\mathrm{Var}\{\hat{V}_{m,n}\} < \mathrm{Var}\{\hat{H}_{m,n}\}$, while $\mathrm{Var}\{\hat{H}_{3m,n}\} < \mathrm{Var}\{\hat{V}_{m,n}\}$ for sufficiently large n, and for each fixed $m \le n-1$. Hence, the estimate $\hat{V}_{m,n}$ based on the m-spacings $X_{(i+m)} - X_{(i)}$ is better than the m-nearest neighbor in one-dimensional models. Unfortunately the construction of $\hat{V}_{m,n}$ cannot be extended to the multidimensional case. Note also that the class of entropy estimates based on m-spacings of increasing order $m = m(n)$ was proposed in van Es [9], where their almost sure consistency and asymptotic normality were derived under the condition that f is bounded. In Goria *et al.* [5] it is also assumed that f is bounded while we do not require this condition (see Theorems 2.1 and 2.2) in Section 2.

Now let $R_{i,n}$ be the distance between X_i and its k_nth nearest neighbor among $X_1, \ldots, X_{i-1}$, $X_{i+1}, \ldots, X_n$, $i = 1, \ldots, n$, i.e., $R_{i,n} := d_{i,k_n,n}$. For a positive real number r and $x \in D$, let $N_x(r)$ denote the ball of radius r centered at x (with respect to the Euclidean distance denoted as $\|x - y\|$):

$$N_x(r) = \{y \in \mathbb{R}^p \colon \|x - y\| \le r\}, \tag{5}$$

and let $V(r)$ be the volume of $N_x(r)$, i.e.,

$$V(r) = \int\limits_{N_x(r)} dy = c_p r^p. \tag{6}$$

Here $c_p = \pi^{p/2}/\Gamma(p/2+1)$ is the volume of the unit ball in $\mathbb{R}^p$. Then,

$$\lim_{r \downarrow 0} \frac{1}{V(r)} \int\limits_{N_x(r)} f(y)\, dy = f(x), \tag{7}$$

for almost all x (with respect to the probability measure induced by the p.d.f. f), where $N_x(r)$ and $V(r)$ are defined in (5) and (6), respectively.

Equation (7) suggests (for small values of r) a histogram type approximation of the p.d.f. $f(x)$ at point $x \in D$:

$$\hat{f}_n^{(1)}(x) = \frac{|N_x(r)|}{nV(r)}.$$

Here $|N_x(r)|$ denotes the cardinality of the set $\{i \colon X_i \in N_x(r)\}$. Namely, to construct a nonparametric estimate of $H(f) = \mathbf{E}_f\big(-\log f(X)\big)$ one can consider the following estimator:

$$\hat{H}_n = -\frac{1}{n} \sum_{i=1}^{n} \log\big[\hat{f}_n^{(1)}(X_i)\big].$$

For the case, where the distance between the points x and y in $\mathbb{R}^p$ is defined as

$$\|x - y\| = \max_{1 \le j \le p} |x_j - y_j|,$$

the histogram type estimate $\hat{H}_n$ with $V(r) = 2^p\, r^p$ was studied in Hall and Morton [10].

A common approach to estimating the entropy of a multivariate distribution is to replace the p.d.f. f in definition (1) by its nonparametric kernel or histogram density estimator (Beirlant *et al.* [11], Scott [2]). However, in many practical situations, the implementation of such estimates in high dimensions (which is the case with the dimension of dihedral angles of macromolecules) becomes difficult. The asymptotic properties of kernel-type entropy estimators of $H(f)$ have been studied in Hall and Morton [10], Joe [12], Ivanov and Rozhkova [13], among others. It was mentioned (see, for example, Hall and Morton [10]) that the rate of convergence is $n^{-1/2}$ when the dimension is $p \le 3$ for kernel-density entropy estimators, while for histogram-type constructions $\hat{H}_n$, the rate of convergence is $n^{-1/2}$ only if $p = 1, 2$. In particular, $\sqrt{n}$-consistency of the histogram entropy estimator was proved under certain tail conditions and smoothness properties of f when $p = 1$.

It was mentioned above that the asymptotic behavior of k-nearest neighbor entropy estimators is related to the construction based on spacings. That is why one may expect to have $\sqrt{n}$-consistency for our estimator $\hat{G}_n$ in (9) for univariate non-uniform models, and $(nk_n)^{-1/2}$-consistency when f is a uniform density (cf. van Es [9]) as $n \to \infty$ and $k = k_n \to \infty$. Table 2 in Section 3 demonstrates this conjecture. For distributions with unbounded support, Tsybakov and van der Meulen [14] studied the $\sqrt{n}$-consistency of nearest neighbor ($k = 1$) entropy estimators based on a truncation technique, when $p = 1$ and the density function f has exponentially decreasing tails. The $\sqrt{n}$-consistency of k-nearest neighbor entropy estimators with unbounded support in $\mathbb{R}^p$ is still an open question when $p \geq 2$. The solution of this problem is connected with derivation of the rate of convergence for $\mathrm{Cov}(T_{1,n}, T_{2,n}) \to 0$ (see (21)), which is not easy to show. The simulation studies below show that, when $p = 2$, it is still possible to have $\sqrt{n}$-consistency for k-nearest neighbor entropy estimators (see Table 3 in Section 3).

Finally, let us introduce the entropy estimators based on the k_n-nearest neighbor density estimator of $f(X_i)$ proposed by Loftsgaarden and Quesenberry [15]:

$$\hat{f}_n^{(2)}(X_i) = \frac{k_n}{nD_{i,n}(\mathbf{X})}, \qquad i = 1, \ldots, n,$$

where $\mathbf{X} = (X_1, \ldots, X_n)'$ and

$$D_{i,n}(\mathbf{X}) = \int\limits_{N_{X_i}(R_{i,n})} dy = \frac{\pi^{p/2} R_{i,n}^p}{\Gamma(p/2+1)} = c_p\, R_{i,n}^p.$$

Denote the corresponding estimator of $H(f)$ by

$$\tilde{G}_n = -\frac{1}{n}\sum_{i=1}^{n} \log(\hat{f}_n^{(2)}(X_i)) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{nD_{i,n}(\mathbf{X})}{k_n}\right) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{nc_pR_{i,n}^p}{k_n}\right). \tag{8}$$

Combining (21), (23), and (25) (see Section 2), one can compare the asymptotic variance of $\hat{G}_n$ with the asymptotic variance of $\hat{H}_{k,n}$ given in (4). We can see that asymptotically, $\mathrm{Var}\{\hat{G}_n\}$ is smaller than $\mathrm{Var}\{\hat{H}_{k,n}\}$ as $n \to \infty$ and k is fixed.

Note that, when $k_n \equiv k$ is a fixed integer, the asymptotic bias of the estimator $\tilde{G}_n$ is equal to $\psi(k) - \log k$ (see Singh *et al.* [4], Goria *et al.* [5], and Leonenko *et al.* [16]), where $\psi(k) = L_{k-1} - \gamma$. The form of the bias term can be explained by observing that the limiting distribution of the summands in (8), given $X_i = x$, represents the gamma distribution with mean $-\log f(x) + \psi(k) - \log k$.

Note also that the estimators studied in Goria *et al.* [5] and $\hat{H}_{k,n}$ defined in (2) can be rewritten in the form (8) with $b_n := \exp(\psi(k_n))$ instead of k_n in the denominator:

$$\hat{G}_n = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{nc_p\, R_{i,n}^p}{b_n}\right). \tag{9}$$

Remark 1. The proofs of Theorems 2.1 and 2.2 in Section 2 are valid for constructions (8) and (9) as $k_n \to \infty$, but we will focus only on the L_2-consistency of the estimator (9). Observe also that $\psi(k) \sim \log k$, when k is large, and hence $b_n \sim k_n$ as $n \to \infty$. Our proof is based on the property that the limiting conditional distributions of the summands in (9), given $X_i = x$, are degenerate at $-\log f(x)$ as $k_n \to \infty$.

After submitting the present paper, we discovered the work of Leonenko *et al.* [16], where the consistency of the kNN estimators of the Tsallis, Rényi, and Shannon entropies have been derived when k is a fixed integer and f is bounded (in the cases of estimating the Shannon entropy $H(f)$ and the functional $I_q = \int_{\mathbb{R}^p} f^q(x)\,dx$, $q > 1$). The authors also made a conjecture concerning the almost sure consistency of $\hat{G}_n$ when f belongs to the class of uniformly continuous functions $\mathcal{F}$ such that $0 < c_1 \leq f(x) \leq c_2 < \infty$ for some c_1, c_2, and $k_n \to \infty$, $k_n/n \to 0$, $k_n/\log n \to \infty$ as $n \to \infty$. In Theorems 2.1 and 2.2 below we do not require f to be a bounded and uniformly continuous function when estimating the Shannon entropy $H(f)$.

Remark 2. Using our approach from Section 2, the weak and L_2-consistency of the k_n-nearest neighbor estimators of the Rényi and Tsallis entropies given by

$$H_q^* = \frac{1}{1-q} \log I_q \qquad \text{and} \qquad H_q = \frac{1}{q-1}(1 - I_q), \quad q \neq 1,$$

respectively, can be derived as $k = k_n \to \infty$ without requiring that the function f is bounded. See, for example, Corollary 2.1, where the asymptotic unbiasedness of the Tsallis estimator $\hat{H}_{n,q}$ is proved. The proof of L_2-consistency of $\hat{H}_{n,q}$ is long and will be presented in a forthcoming paper. Note also that the limits of both entropies, H_q and H_q^*, as $q \to 1$ are equal to $H(f)$. So, to estimate the Shannon entropy $H(f)$, one can apply the kNN estimators of the Tsallis or Rényi entropies with properly chosen $k = k_n \to \infty$, $q = q_n \to 1$ as $n \to \infty$. This question is currently under investigation. When k a fixed integer, the estimation of $H(f)$ by means of $\lim_{q\to 1} \hat{H}_{n,k,q}$ (where $\hat{H}_{n,k,q}$ is the Tsallis entropy estimator from Leonenko *et al.* [16]) is reduced to (9), where $k_n = k$.

2. ASYMPTOTIC UNBIASEDNESS AND CONSISTENCY

Standard results stated in the following lemma will be useful in the proofs of Theorems 2.1 and 2.2, which study the asymptotic properties of the estimators.

Lemma 2.1 (see Billingsley [17], p. 105). (i) *Let $Y_m \sim \mathrm{Bin}(m, p_m)$, $m = 1, 2, \dots$, where $\{p_m, m = 1, 2, \dots\}$ is a sequence of real numbers satisfying*

$$0 < p_m < 1, \qquad m = 1, 2, \dots, \qquad p_m \to 0 \quad \textit{and} \quad m p_m \to \infty \quad \textit{as} \quad m \to \infty. \tag{10}$$

Then the sequence of r.v.'s

$$Z_m = \frac{Y_m - m p_m}{\sqrt{m p_m (1 - p_m)}}, \qquad m = 1, 2, \dots,$$

has a limiting ($m \to \infty$) normal distribution with mean 0 *and variance* 1.

(ii) *Let $\mathbf{Y}_m = (Y_{1,m}, Y_{2,m}) \sim \mathrm{Mult}(m, p_{1,m}, p_{2,m}, p_{3,m})$, $m = 1, 2, \dots$, where $p_{1,m}$ and $p_{2,m}$, $m = 1, 2, \dots$, satisfy the conditions in* (10) *and $p_{3,m} = 1 - p_{1,m} - p_{2,m} > 0$. Define,*

$$Z_{i,m} = \frac{Y_{i,m} - m p_{i,m}}{\sqrt{m p_{i,m}(1 - p_{i,m})}}, \qquad i = 1, 2, \quad m = 1, 2, \dots$$

Then, the sequence $\mathbf{Z}_m = (Z_{1,m}, Z_{2,m})'$, $m = 1, 2, \dots$, of r.v.'s has a limiting bivariate normal distribution with mean vector $(0, 0)'$ and covariance matrix $((\sigma_{i,j}))$, where $\sigma_{1,1} = \sigma_{2,2} = 1$ and $\sigma_{1,2} = \sigma_{2,1} = 0$.

In the following theorem, we establish that the estimator $\hat{G}_n$ defined in (9) is asymptotically unbiased for estimating the entropy $H(f)$.

Theorem 2.1. *Suppose that there exists an $\varepsilon > 0$ such that*

$$\int\limits_{\mathbb{R}^p} |\log f(x)|^{1+\varepsilon} f(x)\, dx < \infty \tag{11}$$

and

$$\int\limits_{\mathbb{R}^p} \int\limits_{\mathbb{R}^p} \big|\log(\|x - y\|)\big|^{1+\varepsilon} f(x) f(y)\, dx\, dy < \infty. \tag{12}$$

Then

$$\lim_{n\to\infty} \mathbf{E}_f(\hat{G}_n) = H(f) \qquad \textit{as} \quad k_n \to \infty, \quad k_n/\sqrt{n} \to 0, \quad n \to \infty.$$

Proof. For $b_n = \exp(\psi(k_n))$ define,

$$T_{i,n} = \log\left(\frac{nD_{i,n}(\mathbf{X})}{b_n}\right) = \log\left(\frac{nc_pR_{i,n}^p}{b_n}\right), \qquad i = 1, 2, \ldots, n,$$

so that

$$\hat{G}_n = \frac{1}{n}\sum_{i=1}^{n} T_{i,n}.$$

Clearly, $T_{1,n}, T_{2,n}, \ldots, T_{n,n}$ are identically distributed. Therefore,

$$\mathbf{E}_f(\hat{G}_n) = \mathbf{E}_f(T_{1,n}).$$

For a fixed $x \in D$, let $S_{x,n}$ be a r.v. whose distribution is the conditional distribution of $T_{1,n}$ given $X_1 = x$. Let $F_{x,n}(\cdot)$ be the corresponding d.f. Then

$$\mathbf{E}_f(\hat{G}_n) = \mathbf{E}_f(T_{1,n}) = \int\limits_{\mathbb{R}^p} \mathbf{E}_f(S_{x,n}) f(x)\, dx.$$

Note that, for a fixed $x \in D$ and $u \in (-\infty, \infty)$,

$$F_{x,n}(u) = \mathbf{P}_f\big\{R_{1,n} \le \rho_n(u) \mid X_1 = x\big\},$$

where

$$\rho_n(u) = \left(\frac{b_n\, \Gamma(p/2+1)\, e^u}{n\pi^{p/2}}\right)^{1/p}. \tag{13}$$

Therefore,

$$\begin{aligned} F_{x,n}(u) &= \mathbf{P}_f\{R_{1,n} \le \rho_n(u) \mid X_1 = x\} \\ &= \mathbf{P}_f\{\text{at least } k_n \text{ of } X_2, \ldots, X_n \in N_x(\rho_n(u))\} = \mathbf{P}\{B_{n-1}(x,u) \ge k_n\} \\ &= \mathbf{P}\left\{Z_n \ge \sqrt{k_n}\, \frac{1 - (n-1)p_n(x,u)/k_n}{\sqrt{(n-1)p_n(x,u)(1-p_n(x,u))/k_n}}\right\}, \end{aligned} \tag{14}$$

where $B_{n-1}(x,u) \sim \mathrm{Bin}(n-1, p_n(x,u))$ and

$$Z_n = \frac{B_{n-1}(x,u) - (n-1)p_n(x,u)}{\sqrt{(n-1)p_n(x,u)(1-p_n(x,u))}} \qquad \text{with} \quad p_n(x,u) = \int\limits_{N_x(\rho_n(u))} f(y)\, dy. \tag{15}$$

Note that, for every fixed $u \in (-\infty, \infty)$, $\lim_{n\to\infty} \rho_n(u) = 0$. On combining (7) and (13), we get

$$\lim_{n\to\infty}\left[\frac{n}{k_n}\, p_n(x,u)\right] = \lim_{n\to\infty}\left[\frac{b_n}{k_n}\, e^u\, \frac{1}{V(\rho_n(u))} \int\limits_{N_x(\rho_n(u))} f(y)\, dy\right] = e^u\, f(x), \tag{16}$$

for almost all x. Consequently, since $k_n \to \infty$, it follows from (16) that, for a fixed $u \in (-\infty, \infty)$

$$p_n(x,u) \to 0 \qquad \text{and} \qquad n\, p_n(x,u) \to \infty \quad \text{as} \quad n \to \infty,$$

and, when n is sufficiently large and $u < -\log f(x)$, we have

$$1 - \frac{(n-1)p_n(x,u)}{k_n} > 0, \tag{17}$$

for almost all x. Now on using Lemma 2.1 (i) in combination with (17) and (14), where $k_n \to \infty$, it follows that, for almost all x, $F_{x,n}(u) \to 0$ as $n \to \infty$. Similarly, for each fixed $u > -\log f(x)$, we have $F_{x,n}(u) \to 1$ as $n \to \infty$ for almost all x. Hence, for almost all x, the limiting $(n \to \infty)$ distribution of

$S_{x,n}$ is degenerate at $-\log f(x)$. Furthermore, an application of the Lebesgue's dominated convergence theorem (see Billigsley [18], p. 209) in the right-hand side of

$$F_n(u) = P(T_{1,n} \le u) = \int F_{x,n}(u) f(x)\, dx,$$

yields $T_{1,n} \xrightarrow{d} T$. Here T denotes a r.v. distributed according to

$$F_T(u) = \int I\{u > -\log f(x)\} f(x)\, dx := Q_f(e^{-u}).$$

The function $Q_f(\cdot) = \int I\{f(x) > \cdot\} f(x)\, dx$ on the right-hand side of the previous equation represents the so-called Q-structural function of f. Using the properties of Q_f (see Khmaladze [19]), we derive

$$\mathbf{E}_f(T) = \int u\, dF_T(u) = -\int \log\lambda\, dQ_f(\lambda) = -\int f(x)\log f(x)\, dx = H(f).$$

To prove Theorem 2.1, it is sufficient to show $\mathbf{E}_f(T_{1,n}) \to \mathbf{E}_f(T)$ as $n \to \infty$. Under the conditions (11) and (12) it can be shown (for details see the Appendix) that, for almost all x, there exists a constant C_1 (not depending on n) such that for all sufficiently large n

$$\mathbf{E}_f\big(|S_{x,n}|^{1+\varepsilon}\big) < C_1 \qquad \text{and} \qquad \mathbf{E}_f(|T_{1,n}|^{1+\varepsilon}) < C_1. \tag{18}$$

Consequently, it follows that (see, for example, Corollary of Theorem 25.12 in Billingsley [18], p. 338)):

$$\lim_{n\to\infty} \mathbf{E}_f(T_{1,n}) = \mathbf{E}_f(T) = H(f).$$

□

The following theorem establishes the consistency of the estimator $\hat{G}_n$.

Theorem 2.2. *Suppose that there exists an $\varepsilon > 0$ such that*

$$\int_{\mathbb{R}^p} |\log f(x)|^{2+\varepsilon} f(x)\, dx < \infty \tag{19}$$

and

$$\int_{\mathbb{R}^p}\int_{\mathbb{R}^p} |\log(\|x-y\|)|^{2+\varepsilon} f(x) f(y)\, dx\, dy < \infty. \tag{20}$$

Then $\hat{G}_n \xrightarrow{L_2} H(f)$, i.e., $\hat{G}_n$ is a consistent estimator of $H(f)$ as $n \to \infty$ and $k_n \to \infty$, $k_n/\sqrt{n} \to 0$.

Proof. Using the notation in the proof of Theorem 2.1, we have $\hat{G}_n = \frac{1}{n}\sum_{i=1}^n T_{i,n}$. Since the distribution of the random vector $(T_{1,n}, T_{2,n}, \ldots, T_{n,n})'$ is the same for any its permutation, we have

$$\mathrm{Var}_f(\hat{G}_n) = \frac{1}{n}\mathrm{Var}_f(T_{1,n}) + \frac{n(n-1)}{n^2}\mathrm{Cov}_f(T_{1,n}, T_{2,n}). \tag{21}$$

Also,

$$\mathbf{E}_f(T_{1,n}^2) = \int_{\mathbb{R}^p} \mathbf{E}_f(S_{x,n}^2) f(x)\, dx,$$

where $S_{x,n}$ has the same distribution as the conditional distribution of $T_{1,n}$ given $X_1 = x$. On using arguments similar to those used in the Appendix, it can be shown that, for almost all x, there exists a constant C_2 (not depending on n) such that

$$\mathbf{E}_f(|S_{x,n}|^{2+\varepsilon}) < C_2 \qquad \text{and} \qquad \mathbf{E}_f(|T_{1,n}|^{2+\varepsilon}) < C_2 \tag{22}$$

for all sufficiently large n.

In a way similar to that in the proof of Theorem 2.1, using condition (19), we get

$$\lim_{n\to\infty} \mathbf{E}_f(T_{1,n}^2) = \mathbf{E}_f(T^2) = \int u^2\, dQ_f(e^{-u}) = \int (\log\lambda)^2\, dQ_f(\lambda) = \int (\log f(x))^2 f(x)\, dx$$

and

$$\begin{aligned}\lim_{n\to\infty} \mathrm{Var}_f(T_{1,n}) &= \lim_{n\to\infty} \mathbf{E}_f(T_{1,n}^2) - \lim_{n\to\infty} \big(\mathbf{E}_f(T_{1,n})\big)^2 \\ &= \int\limits_{\mathbb{R}^p} (\log f(x))^2 f(x)\, dx - (H(f))^2 = \mathrm{Var}_f(\log f(X_1)). \qquad (23)\end{aligned}$$

Also, we can write

$$\mathbf{E}_f(T_{1,n}\, T_{2,n}) = \int\limits_{\mathbb{R}^p}\int\limits_{\mathbb{R}^p} \mathbf{E}_f(S_{x,y,n}^{(1)} S_{x,y,n}^{(2)}) f(x) f(y)\, dx\, dy,$$

where $\mathbf{S}_{x,y,n} = \big(S_{x,y,n}^{(1)}, S_{x,y,n}^{(2)}\big)'$ has the same distribution as the conditional distribution of $(T_{1,n}, T_{2,n})'$ given $(X_1, X_2) = (x, y)$. For fixed $x, y \in D$ and $u_1, u_2 \in (-\infty, \infty)$

$$\mathbf{P}_f\big\{S_{x,y,n}^{(1)} > u_1, S_{x,y,n}^{(2)} > u_2\big\} = \mathbf{P}_f\big\{R_{1,n} > \rho_n(u_1), R_{2,n} > \rho_n(u_2) \mid X_1 = x,\, X_2 = y\big\}$$

with $\rho_n(\cdot)$ defined in (13). Since $\rho_n(u_1)$ and $\rho_n(u_2)$ tend to zero as $n \to \infty$, for $x \neq y$ we may assume that $N_x(\rho_n(u_1)) \cap N_x(\rho_n(u_2)) = \emptyset$, the empty set, for large n. Thus, for large n,

$$\begin{aligned}\mathbf{P}_f\big\{S_{x,y,n}^{(1)} > u_1, S_{x,y,n}^{(2)} > u_2\big\} &= \mathbf{P}_f\big\{\text{at most } \ k_n - 1 \ \text{ of } X_3, \ldots, X_n \in N_x(\rho_n(u_1)) \\ &\qquad \text{and at most } \ k_n - 1 \ \text{ of } \ X_3, \ldots, X_n \in N_x(\rho_n(u_2))\big\} \\ &= \mathbf{P}_f\big\{M_{1,n-2} \le k_n - 1,\, M_{2,n-2} \le k_n - 1\big\},\end{aligned}$$

where $(M_{1,n-2}, M_{2,n-2}) \sim \mathrm{Mult}\,\big(n-2,\, p_n(x, u_1), p_n(y, u_2), 1 - p_n(x, u_1) - p_n(y, u_2)\big)$ with $p_n(\cdot,\cdot)$ defined in (15). Note that, for fixed $u_1, u_2 \in (-\infty, \infty)$, according to (7) and (13) $p_n(x, u_1)$ and $p_n(y, u_2)$ satisfy condition (16) with $u = u_1$ and $u = u_2$, for almost all x and y. Consequently, for fixed $u_1, u_2 \in (-\infty, \infty)$, it follows that for almost all x and y

$$p_n(x, u_1) \to 0, \quad p_n(y, u_2) \to 0 \qquad \text{and} \qquad np_n(x, u_1) \to \infty, \quad np_n(y, u_2) \to \infty$$

as $n \to \infty$. Now on using Lemma 2.1 (ii) it follows that, for almost all x and y, the limiting distribution of $\big(S_{x,y,n}^{(1)}, S_{x,y,n}^{(2)}\big)$ is degenerate at $\big(-\log f(x), -\log f(y)\big)$.

Under the assumptions of the theorem, using arguments similar to those used in the Appendix, it can be shown that for almost all x and y there exists a constant C_3 (not depending on n) such that

$$\mathbf{E}_f\big(|S_{x,y,n}^{(1)}\, S_{x,y,n}^{(2)}|^{1+\varepsilon}\big) < C_3 \qquad (24)$$

for all sufficiently large n. Then using Corollary of Theorem 25.12 in Billingsley [17], it follows that

$$\lim_{n\to\infty} \mathbf{E}_f\big(S_{x,y,n}^{(1)}\, S_{x,y,n}^{(2)}\big) = \log f(x)\, \log f(y).$$

Again, combining Fatou's Lemma (see Billingsley [18], p. 338) and condition (19), we obtain

$$\begin{aligned}\lim_{n\to\infty} \mathbf{E}_f\big(T_{1,n}\, T_{2,n}\big) &= \lim_{n\to\infty} \int\limits_{\mathbb{R}^p}\int\limits_{\mathbb{R}^p} \mathbf{E}_f\big(S_{x,y,n}^{(1)} S_{x,y,n}^{(2)}\big) f(x) f(y)\, dx\, dy \\ &= \int\limits_{\mathbb{R}^p}\int\limits_{\mathbb{R}^p} \lim_{n\to\infty} \mathbf{E}_f\big(S_{x,y,n}^{(1)} S_{x,y,n}^{(2)}\big) f(x) f(y)\, dx\, dy \\ &= \int\limits_{\mathbb{R}^p}\int\limits_{\mathbb{R}^p} \log f(x) \log f(y) f(x) f(y)\, dx\, dy = \big(H(f)\big)^2\end{aligned}$$

and

$$\lim_{n\to\infty} \mathrm{Cov}_f\,\big(T_{1,n},\, T_{2,n}\big) = \lim_{n\to\infty} \mathbf{E}_f\big(T_{1,n}\, T_{2,n}\big) - \lim_{n\to\infty} \big(\mathbf{E}_f(T_{1,n})\,\mathbf{E}_f(T_{2,n})\big)$$
$$= \big(H(f)\big)^2 - \big(H(f)\big)^2 = 0. \tag{25}$$

Finally, since $\mathbf{E}_f(\hat{G}_n - H(f))^2 = \mathrm{Var}_f(\hat{G}_n) + (\mathbf{E}_f(\hat{G}_n) - H(f))^2$, the statement of Theorem 2.2 follows from Theorem 2.1, (21), (23), and (25). □

Let $b_n^* = \{\Gamma(k_n + 1 - q)/\Gamma(k_n)\}^{1/(1-q)}$. In a similar way as in Leonenko *et al.* [16], define the estimator of the Tsallis entropy as follows:

$$\hat{H}_{n,q} = \frac{1}{q-1}\{1 - \hat{I}_{n,q}\} \qquad \text{with} \quad \hat{I}_{n,q} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{nc_pR_{i,n}^p}{b_n^*}\right)^{1-q}, \quad q \neq 1.$$

Corollary 2.1. (i) *If $q < 1$ and $I_q < \infty$, then*

$$\lim_{n\to\infty} \mathbf{E}_f(\hat{I}_{n,q}) = I_q \quad \textit{and} \quad \lim_{n\to\infty} \mathbf{E}_f(\hat{H}_{n,q}) = H_q \quad \textit{as} \quad k_n \to \infty, \quad k_n/n \to 0, \quad n \to \infty. \tag{26}$$

(ii) *If $I_s < \infty$ for $1 < s \le 2$, then* (26) *holds for* $1 < q < 2$.

Proof. To show that $\hat{I}_{n,q}$ is asymptotically unbiased, let

$$T_{i,n} = nc_pR_{i,n}^p/b_n^* \quad \text{and} \quad \rho_n^*(u) = \{(ub_n^*)/(nc_p)\}^{1/p}.$$

Using an argument similar to the one in the proof of Theorem 2.1, we see that the conditional limiting distribution of $T_{1,n}$, given $X_1 = x$, is degenerate at $1/f(x)$ and the limiting distribution of $T_{1,n}$ is defined as follows: $F_T(u) = Q_f(u^{-1})$. To justify this let us notice that $b_n^*/k_n \to 1$, and instead of (16) we have:

$$\lim_{n\to\infty}\left[\frac{n}{k_n}\, p_n(x,u)\right] = \lim_{n\to\infty}\left[\frac{b_n^*\, u}{k_n}\, \frac{1}{V(\rho_n^*(u))} \int\limits_{N_x(\rho_n^*(u))} f(y)\, dy\right] = uf(x).$$

Case (i): Since the function $u \to u^{1-q}$ is bounded on any bounded interval from $(0,\infty)$, an application of the generalized Herry–Bray Lemma, see Loève [20], p. 187, gives

$$\lim_{n\to\infty} \mathbf{E}_f(T_{1,n}^{1-q}) = \mathbf{E}_f(T^{1-q}) = \int u^{1-q}\, dF_T(u) = -\int \lambda^{q-1}\, dQ_f(\lambda) = -\int f^{q-1}(x) f(x)\, dx = I_q.$$

Case (ii): Since for any $0 < u < 1$, we have $\rho_n^*(u) \le \rho_n^*(1) \to 0$ as $n \to \infty$, we conclude:

$$\lim_{n\to\infty}\left[\frac{n}{u\,k_n}\, p_n(x,u)\right] = f(x)$$

for almost all x, uniformly in $0 < u < 1$. Hence, for any positive δ, we can take sufficiently large n such that $p_n(x,u) \le n\,u\,(f(x)+\delta)/k_n$ for any $0 < u < 1$, and obtain

$$F_{x,n}(u) = P\{B_{k_n,n-k_n} \le p_n(x,u)\} \le P\Big\{B_{k_n,n-k_n} \le \frac{k_n}{n}\,(f(x)+\delta)u\Big\}$$
$$= P\big(\bar{Q}_n \ge k_n\big) \le \frac{E(\bar{Q}_n)}{k_n} = \frac{n-1}{n}(f(x)+\delta)\,u \le (f(x)+\delta)\,u,$$

for all sufficiently large n. Here $\bar{Q}_n \sim \mathrm{Bin}\,\big(n-1, k_n\big(f(x)+\delta\big)u/n\big)$ (cf. (40) in the Appendix). Now let $S_{x,n}$ be a r.v. whose distribution is the same as the conditional distribution of $T_{1,n}$ given $X_1 = x$. Using the previous inequality and integration by parts, as in Leonenko et al. [16], we derive

$$\mathbf{E}_f(S_{x,n}^{\beta}) = \int\limits_0^\infty u^\beta\, dF_{n,x}(u) \le 1 - \frac{\beta(f(x)+\delta)}{(\beta+1)} = K_\varepsilon(x),$$

for almost all x and for all sufficiently large n. Here $\beta = (1-q)(1+\varepsilon) < 0$ is such that $\beta + 1 > 0$ if $0 \le \varepsilon < (2-q)/(q-1)$, and $\int_{\mathbb{R}^p} K_\varepsilon(x) f(x)\, dx < \infty$. Hence the conditions similar to (18) are satisfied. An application of Corollary from Theorem 25.12 in Billingsley [18], p. 338, yields $\lim_{n\to\infty} \mathbf{E}_f(T_{1,n}^{1-q}) = I_q$. $\square$

Remark 3. Denote the unit torus by S_1^p, where $S_1 = \{x \in \mathbb{R}^2 \colon \|x\| = 1\}$ is a unit circle. Each circular observation $X = (X_1, \ldots, X_p)$ on S_1^p can be specified by an angular observation $\theta = (\theta_1, \ldots, \theta_p) \in [0, 2\pi)^p$ and vice versa, since $S_1 = \{(\cos\theta_j, \sin\theta_j) \colon \theta_j \in [0, 2\pi)\}$ for $j = 1, \ldots, p$. For measuring distances between angular random variables, say $\theta = (\theta_1, \ldots, \theta_p) \in [0, 2\pi)^p$ and $\mu = (\mu_1, \ldots, \mu_p) \in [0, 2\pi)^p$ two distance functions of interest are

$$d_1(\theta, \mu) = \sqrt{\sum_{j=1}^{p} \big(\pi - \big|\pi - |\theta_j - \mu_j|\big|\big)^2}$$

and

$$d_2(\theta, \mu) = \sqrt{2 \sum_{j=1}^{p} \big(1 - \cos(\theta_j - \mu_j)\big)}$$

(see also the formulas (2.3.13) and (2.3.5) in Mardia and Jupp [21], when $p = 1$). For estimating the entropy of circular distributions, Misra *et al.* [22] used distance functions $d_1(\cdot,\cdot)$ and $d_2(\cdot,\cdot)$ to construct estimators based on distances between the n sample points and their kth nearest neighbors, where k ($\le n-1$) is a fixed positive integer (not depending on n). The results of Section 2 can be extended, in a straightforward manner, to circular distances $d_1(\cdot,\cdot)$ and $d_2(\cdot,\cdot)$ as well.

3. SIMULATION STUDY

In this section we study the RMSE of $\hat{G}_n$ via simulations from different distributions, i.e., the uniform, beta, gamma, normal, von Mises, bivariate circular, and the univariate and bivariate log-normal distributions.

First, recall that the density function $f(\theta; \mu, \kappa)$ (with respect to the Lebesgue measure on $[0, 2\pi)$) of von Mises circular distribution is

$$f(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa\cos(\theta - \mu)}, \qquad 0 \le \theta < 2\pi,$$

where μ and κ are the mean direction and concentration parameters (see, for example, Singh *at al.* [23]). The value of the entropy of $f(\theta; \mu, \kappa)$ is given by $H(f) = \log(2\pi I_0(\kappa)) - \frac{\kappa I_1(\kappa)}{I_0(\kappa)}$, where the $I_p(\cdot)$ represent the Bessel functions of order $p = 0, 1$. In a similar way let us define on $[0, 2\pi)^2$ a bivariate circular distribution as introduced in Singh *at al.* [23]:

$$f(\theta_1, \theta_2; \mu_1, \mu_2, \kappa_1, \kappa_2, \lambda) = \frac{1}{4\pi^2 C} e^{\kappa_1\cos(\theta_1 - \mu_1) + \kappa_2\cos(\theta_2 - \mu_2) + 2\lambda\sin(\theta_1 - \mu_1)\sin(\theta_2 - \mu_2)}.$$

Here

$$C = \sum_{p=0}^{\infty} \binom{2p}{p} \left(\frac{\lambda^2}{\kappa_1\kappa_2}\right)^p I_p(\kappa_1) I_p(\kappa_2).$$

Its theoretical entropy is $H(f) = 2\log 2\pi + \log C - D/C$, where

$$D = \sum_{p=0}^{\infty} \binom{2p}{p} \left(\frac{\lambda^2}{\kappa_1\kappa_2}\right)^p \big(\kappa_1 I_{p+1}(\kappa_1) I_p(\kappa_2) + \kappa_2 I_p(\kappa_1) I_{p+1}(\kappa_2) + 2p I_p(\kappa_1) I_p(\kappa_2)\big)$$

and the $I_p(\cdot)$ represent the Bessel functions of order $p = 0, 1, \ldots$

The Monte Carlo simulations justify, for distributions like the circular and log-normal, that the optimal k (in the sense of RMSE) is a slowly increasing function of n for the dimension $p = 1, 2$. See Figs. 1 and 2 and Table 1, where the optimal k's, corresponding RMSE's, and bias terms are recorded. The sample sizes for the simulations were: $n = 50$, $n = 500$, and $n = 1000$, and the number of repetitions was $N = 10,000$. Actually, when the sample size is $n = 50$, there are two minima of RMSE for the bivariate circular distribution model with values of 0.223 and 0.212 at $k = 3$ and at $k = 14$, respectively (see the first curve in Fig. 1). In Table 1 we recorded only the first minimum value.

Table 1. Distribution models used for simulations

Model	Optimal k (RMSE, BIAS)		
	$n = 50$	$n = 500$	$n = 1000$
von Mises with $\kappa = 2$, $\mu = \frac{3\pi}{2}$	7 (0.149, −0.0575)	11 (0.0426, −0.00764)	18 (0.0282, −0.00542)
Bivariate circular with $\kappa = (2, 2)$, $\mu = (\frac{3\pi}{2}, \frac{3\pi}{2})$, $\lambda = 0.5$	3 (0.223, 0.0542)	10 (0.0580, 0.0164)	15 (0.0404, 0.0111)
Log-Normal with $\mu = 10$, $\sigma^2 = 1$	7 (0.194, −0.0524)	13 (0.0585, −0.0130)	13 (0.0408, −0.00702)
Bivariate Log-Normal with $\mu_1 = \mu_2 = 10$, $\sigma_1 = \sigma_2 = 1$, $\rho = 0.6$	6 (0.296, 0.0274)	12 (0.0899, 0.00453)	15 (0.0629, 0.00167)

During the simulation studies we found that, for example in one-dimensional cases, the rate of convergence of k_n-nearest neighbor entropy estimators is $n^{-1/2}$ for non-uniform distributions, and $(nk_n)^{-1/2}$ for the uniform one (cf. van Es [9]). In addition, in Table 2 we calculated the values of $\sqrt{n}$ RMSE for all models except the Uniform $[0, 1]$, and the values of $\sqrt{nk_n}$ RMSE for the Uniform $[0, 1]$. We can see from Table 2 that these products are almost constants for each model. Also, we approximated the optimal k_n as follows: $k_{n,opt} = [c \log^2 n]$, where $c = 0.44$ for the Uniform $[0, 1]$, $c = 0.67$ for the von Mises, $c = 0.03$ for the Normal $(0, 1)$ and Log-Normal $(10, 1)$, and $c = 0.31$ for the Gamma $(6, 1)$ distributions, respectively.

Table 2. $\sqrt{n\,k_n} \cdot$ RMSE for the uniform distribution and $\sqrt{n} \cdot$ RMSE for the other univariate distributions

n	50	100	200	500	700	1000	2000	5000	7000
Uniform	1.78	1.78	1.73	1.73	1.61	1.71	1.68	1.67	1.63
Normal	0.91	0.89	0.88	0.84	0.82	0.83	0.80	0.78	0.77
Log-Normal	1.38	1.35	1.34	1.31	1.30	1.29	1.29	1.29	1.26
Gamma $(6, 1)$	0.92	0.90	0.89	0.85	0.84	0.83	0.83	0.81	0.79
Beta $(3, 2)$	0.65	0.66	0.62	0.50	0.58	0.48	0.56	0.54	0.53
von Mises	1.06	1.03	1.00	0.95	0.92	0.89	0.88	0.85	0.84

Finally, in Table 3 the values of $\sqrt{n}$ RMSE for two-dimensional uniform, circular, normal and log-normal distributions are presented. We conclude that, for these distributions, one can expect to have $\sqrt{n}$-consistency as well.

Table 3. $\sqrt{n} \cdot \mathrm{RMSE}$ for bivariate distributions

n	50	500	1000	10000
Uniform	1.46	1.36	1.34	1.40
Circular	1.50	1.30	1.28	1.49
Log Normal	2.09	2.01	1.99	2.07
Normal	1.12	1.35	1.32	1.27

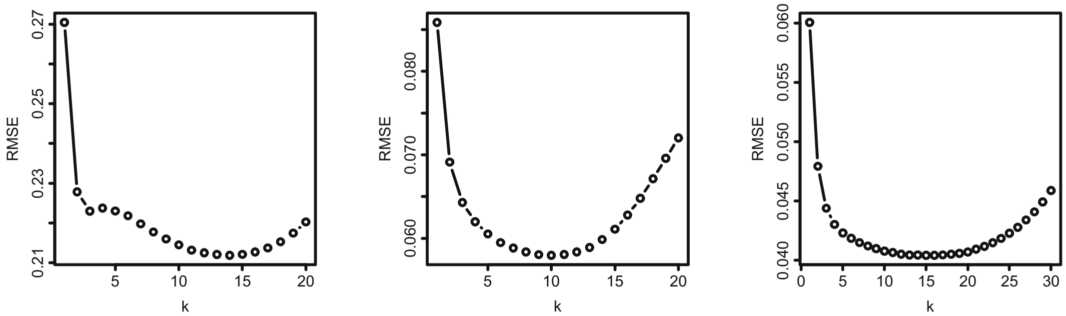

Fig. 1. Bivariate circular distribution, $n = 50, 500$ and 1000.

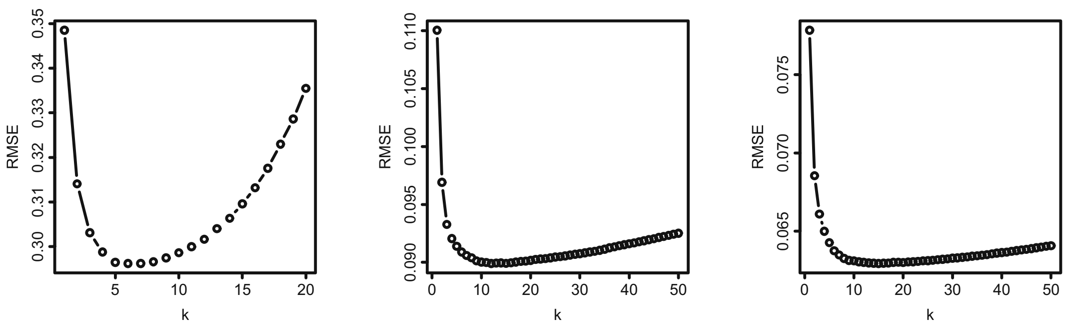

Fig. 2. Bivariate Log-Normal distribution, $n = 50, 500$ and 1000.

APPENDIX

Here we provide the proof of the first inequality in (18). The proofs of the first inequalities in (22) and (24), being cumbersome and virtually identical to the proof of (18), are omitted.

Let $R_1 = \|X_1 - X_2\|$,

$$V(R_1) = \int\limits_{N_x(R_1)} dy = \frac{\pi^{p/2} R_1^p}{\Gamma(p/2+1)},$$

and

$$T_1 = \log\big(2V(R_1)\big) = \log\left(\frac{2\,\pi^{p/2} R_1^p}{\Gamma(p/2+1)}\right).$$

Further, for a fixed $x \in D$, let S_x be a r.v. having the same distribution as the conditional distribution of T_1 given $X_1 = x$. To prove the first inequality in (18), we will first establish that

$$\mathbf{E}_f\big(|S_x|^{1+\varepsilon}\big) < \infty \tag{27}$$

for almost all x. We have

$$\mathbf{E}_f(|S_x|^{1+\varepsilon}) = \mathbf{E}_f\big(|\log(2V(R_1))|^{1+\varepsilon} \mid X_1 = x\big)$$

$$\leq 2^{\varepsilon}\left[\left|\log\left(\frac{2\,\pi^{p/2}}{\Gamma(p/2+1)}\right)\right|^{1+\varepsilon}+p^{1+\varepsilon}\,\mathbf{E}_f\big(|\log R_1|^{1+\varepsilon}\mid X_1=x\big)\right]$$
$$=2^{\varepsilon}\left[\left|\log\left(\frac{2\,\pi^{p/2}}{\Gamma(p/2+1)}\right)\right|^{1+\varepsilon}+p^{\varepsilon}\int\limits_{\mathbb{R}^p}\big|\log(\|x-y\|)\big|^{1+\varepsilon}f(y)\,dy\right].$$

In view of (12), it follows that

$$\int\limits_{\mathbb{R}^p}|\log(\|x-y\|)|^{1+\varepsilon}f(y)\,dy<\infty$$

for almost all x. Therefore (27) is established.

Now we will establish the first inequality in (18). We have

$$\mathbf{E}_f\big(|S_{x,n}|^{1+\varepsilon}\big)=\int\limits_{-\infty}^{0}|u|^{1+\varepsilon}\,dF_{x,n}(u)+\int\limits_{0}^{\infty}|u|^{1+\varepsilon}\,dF_{x,n}(u). \tag{28}$$

We can write

$$\begin{aligned}\int\limits_{0}^{\infty}|u|^{1+\varepsilon}\,dF_{x,n}(u)&=(1+\varepsilon)\int\limits_{0}^{\infty}u^{\varepsilon}(1-F_{x,n}(u))\,du\\&=(1+\varepsilon)\Bigg[\int\limits_{0}^{\log\sqrt{n}}u^{\varepsilon}(1-F_{x,n}(u))\,du+\int\limits_{\log\sqrt{n}}^{\infty}u^{\varepsilon}(1-F_{x,n}(u))\,du\Bigg]\\&=(1+\varepsilon)\big[I_1(x,n)+I_2(x,n)\big],\qquad\text{say.}\end{aligned} \tag{29}$$

We have

$$\begin{aligned}I_2(x,n)&=\int\limits_{\log\sqrt{n}}^{\infty}u^{\varepsilon}(1-F_{x,n}(u))\,du=\int\limits_{\log\sqrt{n}}^{\infty}u^{\varepsilon}P\{B_{n-1}(x,u)\leq k_n-1\}\,du\\&=\int\limits_{\log\sqrt{n}}^{\infty}u^{\varepsilon}\Bigg[\sum_{j=0}^{k_n-1}\binom{n-1}{j}(p_n(x,u))^j(1-p_n(x,u))^{n-1-j}\Bigg]\,du.\end{aligned}$$

For $j\in\{0,1,\ldots,k_n-1\}$, we have

$$\binom{n-1}{j}\leq\frac{n-1}{n-k_n}\binom{n-2}{j}.$$

Therefore,

$$\begin{aligned}I_2(x,n)&\leq\frac{n-1}{n-k_n}\int\limits_{\log\sqrt{n}}^{\infty}u^{\varepsilon}(1-p_n(x,u))\Bigg[\sum_{j=0}^{k_n-1}\binom{n-2}{j}(p_n(x,u))^j(1-p_n(x,u))^{n-2-j}\Bigg]\,du\\&=\frac{n-1}{n-k_n}\int\limits_{\log\sqrt{n}}^{\infty}u^{\varepsilon}(1-p_n(x,u))P\{B_{k_n,n-k_n-1}\geq p_n(x,u)\}\,du,\end{aligned}$$

where $B_{a,b}$ denotes the beta r.v. with parameters (a,b), $a>0$, $b>0$. For $u>\log\sqrt{n}$, we have $p_n(x,u)>p_n(x,\log\sqrt{n})$. Therefore,

$$I_2(x,n)\leq\frac{n-1}{n-k_n}P\{B_{k_n,n-k_n-1}\geq p_n(x,\log\sqrt{n})\}\int\limits_{\log\sqrt{n}}^{\infty}u^{\varepsilon}(1-p_n(x,u))\,du.$$

Note that

$$\rho_n(\log\sqrt{n}) = \left(\frac{b_n\,\Gamma(p/2+1)}{\sqrt{n}\;\pi^{p/2}}\right)^{1/p} \to 0 \qquad \text{as} \quad \frac{b_n}{\sqrt{n}} \to 0, \quad n\to\infty.$$

Therefore,

$$\lim_{n\to\infty}\left[\frac{\sqrt{n}}{k_n}\,p_n(x,\log\sqrt{n})\right] = \lim_{n\to\infty}\left[\frac{b_n}{k_n}\,\frac{1}{V(\rho_n(\log\sqrt{n}))}\int\limits_{N_x(\rho_n(\log\sqrt{n}))} f(y)\,dy\right] = f(x),$$

for almost all x. Let us now, for $x\in D$, choose a $\delta\in(0,f(x))$. Then, for all sufficiently large n,

$$p_n(x,\log\sqrt{n}\,) > \frac{k_n}{\sqrt{n}}\,(f(x)-\delta)$$

and therefore

$$\begin{aligned} P\big\{B_{k_n,n-k_n-1}\ge p_n(x,\log\sqrt{n})\big\} &\le P\Big\{B_{k_n,n-k_n-1}\ge \frac{k_n}{\sqrt{n}}\,(f(x)-\delta)\Big\}\\ &\le \frac{E(B^2_{k_n,n-k_n-1})}{\big(\frac{k_n}{\sqrt{n}}\,(f(x)-\delta)\big)^2} = \frac{k_n+1}{(n-1)\,k_n\,(f(x)-\delta)^2}. \end{aligned}$$

Thus, for almost all x,

$$I_2(x,n) \le \frac{k_n+1}{(n-k_n)\,k_n\,(f(x)-\delta)^2}\int\limits_{\log\sqrt{n}}^{\infty} u^{\varepsilon}(1-p_n(x,u))\,du, \tag{30}$$

for all sufficiently large n. On making the change of variable $z=\log\big(2b_n/n\big)+u$ in the integral in (30), we get

$$\int\limits_{\log\sqrt{n}}^{\infty} u^{\varepsilon}(1-p_n(x,u))\,du = \int\limits_{\log(2b_n/\sqrt{n})}^{\infty}\left(u+\log\left(\frac{n}{2b_n}\right)\right)^{\varepsilon}\left(1-p_n\left(x,u+\log\left(\frac{n}{2b_n}\right)\right)\right)du.$$

Note that

$$\rho_n\left(u+\log\left(\frac{n}{2b_n}\right)\right) = \left(\frac{\Gamma(p/2+1)e^{u}}{2\,\pi^{p/2}}\right)^{1/p}$$

and thus

$$p_n\left(x,u+\log\left(\frac{n}{2b_n}\right)\right) = \int\limits_{N_x(\rho_n(u+\log(n/(2b_n))))} f(y)\,dy = q(x,u), \quad \text{say},$$

does not depend on n. Therefore,

$$\begin{aligned} \int\limits_{\log\sqrt{n}}^{\infty} u^{\varepsilon}(1-p_n(x,u))\,du &= \int\limits_{\log(2b_n/\sqrt{n})}^{\infty}\left(u+\log\left(\frac{n}{2b_n}\right)\right)^{\varepsilon}\big(1-q(x,u)\big)\,du\\ &= \int\limits_{\log(2b_n/\sqrt{n})}^{0}\left(u+\log\left(\frac{n}{2b_n}\right)\right)^{\varepsilon}\big(1-q(x,u)\big)\,du\\ &\quad + \int\limits_{0}^{\infty}\left(u+\log\left(\frac{n}{2b_n}\right)\right)^{\varepsilon}\big(1-q(x,u)\big)\,du. \end{aligned} \tag{31}$$

For the first term in the last equation we have from (31)

$$\int\limits_{\log(2b_n/\sqrt{n})}^{0} \left(u+\log\left(\frac{n}{2b_n}\right)\right)^{\varepsilon}\bigl(1-q(x,u)\bigr)\,du \le \left(\log\left(\frac{n}{2b_n}\right)\right)^{\varepsilon} \int\limits_{\log(2b_n/\sqrt{n})}^{0} \bigl(1-q(x,u)\bigr)\,du$$

$$= \left(\log\left(\frac{n}{2b_n}\right)\right)^{\varepsilon} \int\limits_{2b_n/\sqrt{n}}^{1} \bigl(1-q(x,\log u)\bigr)\frac{du}{u} \le \frac{\sqrt{n}}{2b_n}\left(\log\left(\frac{n}{2b_n}\right)\right)^{\varepsilon} \int\limits_{2b_n/\sqrt{n}}^{1} \bigl(1-q(x,\log u)\bigr)\,du$$

$$\le \frac{\sqrt{n}}{2b_n}\left(\log\left(\frac{n}{2b_n}\right)\right)^{\varepsilon}. \tag{32}$$

To estimate the second term note that

$$\int\limits_{0}^{\infty} \left(u+\log\left(\frac{n}{2b_n}\right)\right)^{\varepsilon}\bigl(1-q(x,u)\bigr)\,du$$

$$\le F_\varepsilon\left[\left(\log\left(\frac{n}{2b_n}\right)\right)^{\varepsilon}\int\limits_0^\infty \bigl(1-q(x,u)\bigr)\,du + \int\limits_0^\infty u^{\varepsilon}\bigl(1-q(x,u)\bigr)\,du\right], \tag{33}$$

where $F_\varepsilon = \max(1, 2^{\varepsilon-1})$. Note now that, for $u \in (-\infty,\infty)$ and for R_1 and S_x defined in the beginning of this Appendix,

$$F_x(u) = P_f(S_x \le u) = \int\limits_{N_x(\rho_n(u+\log(n/2b_n)))} f(y)\,dy = q(x,u).$$

Therefore (33) yields

$$\int\limits_{0}^{\infty} \left(u+\log\left(\frac{n}{2b_n}\right)\right)^{\varepsilon}\bigl(1-F_x(u)\bigr)\,du$$

$$\le F_\varepsilon\left[\left(\log\left(\frac{n}{2b_n}\right)\right)^{\varepsilon}\int\limits_0^\infty \bigl(1-F_x(u)\bigr)\,du + \int\limits_0^\infty u^{\varepsilon}\bigl(1-F_x(u)\bigr)\,du\right]. \tag{34}$$

In view of (27),

$$\int\limits_0^\infty \bigl(1-F_x(u)\bigr)\,du < \infty \qquad \text{and} \qquad \int\limits_0^\infty u^{\varepsilon}\bigl(1-F_x(u)\bigr)\,du < \infty. \tag{35}$$

On using (30)–(35), we conclude that

$$\lim_{n\to\infty} I_2(x,n) = 0 \tag{36}$$

for almost all x. Now let us consider

$$I_1(x,n) = \int\limits_0^{\log\sqrt{n}} u^{\varepsilon}\bigl(1-F_{x,n}(u)\bigr)\,du$$

$$= \int\limits_0^{\log\sqrt{n}} u^{\varepsilon}\left[\sum_{j=0}^{k_n-1}\binom{n-1}{j}(p_n(x,u))^j(1-p_n(x,u))^{n-1-j}\right]du$$

$$= \int\limits_0^{\log\sqrt{n}} u^{\varepsilon} P\bigl(B_{k_n,n-k_n} \ge p_n(x,u)\bigr)\,du. \tag{37}$$

For $u < \log\sqrt{n}$,

$$0 \le \rho_n(u) \le \rho_n(\log\sqrt{n}) = \left(\frac{b_n\Gamma(p/2+1)}{\sqrt{n}\,\pi^{p/2}}\right)^{1/p} \to 0 \qquad \text{as} \quad \frac{b_n}{\sqrt{n}} \to 0, \quad n \to \infty.$$

Therefore, for almost all x,

$$\lim_{n\to\infty}\left(\frac{n}{k_n e^u} p_n(x,u)\right) = \lim_{n\to\infty}\left[\frac{b_n}{k_n}\frac{1}{V(\rho_n(u))}\int\limits_{N_x(\rho_n(u))} f(y)\,dy\right] = f(x),$$

uniformly in u.

For $x \in D$, let $\delta \in (0, f(x))$. Then, for all sufficiently large n, $u < \log\sqrt{n}$, and for almost all x, we have $p_n(x,u) > \frac{k_n}{n}(f(x)-\delta)e^u$ and therefore

$$P\{B_{k_n,n-k_n} \ge p_n(x,u)\} \le P\left\{B_{k_n,n-k_n} \ge \frac{k_n}{n}(f(x)-\delta)e^u\right\} \le \frac{E(B_{k_n,n-k_n})}{\frac{k_n}{n}(f(x)-\delta)e^u} = \frac{e^{-u}}{f(x)-\delta}.$$

On using the above inequality in (37), for almost all x and for all sufficiently large n, we get

$$I_1(x,n) \le \frac{1}{f(x)-\delta}\int\limits_0^{\log\sqrt{n}} u^\varepsilon e^{-u}\,du \le \frac{1}{f(x)-\delta}\int\limits_0^\infty u^\varepsilon e^{-u}\,du < \infty. \tag{38}$$

From (29), (36), and (38) we conclude that there exists a constant D_1 such that for almost all x and for all sufficiently large n

$$\int\limits_0^\infty |u|^{1+\varepsilon}\,dF_{x,n}(u) < D_1. \tag{39}$$

Now consider

$$\int\limits_{-\infty}^0 |u|^{1+\varepsilon}\,dF_{x,n}(u) = \int\limits_{-\infty}^0 (-u)^{1+\varepsilon}\,dF_{x,n}(u) = (1+\varepsilon)\int\limits_{-\infty}^0 (-u)^\varepsilon F_{x,n}(u)\,du.$$

For $u < 0$,

$$0 \le \rho_n(u) \le \left(\frac{b_n\,\Gamma(p/2+1)}{n\,\pi^{p/2}}\right)^{1/p} \to 0 \qquad \text{as} \quad n \to \infty.$$

Thus, for $u < 0$, the convergence $\rho_n(u) \to 0$ as $n \to \infty$ is uniform. Therefore, for almost all x,

$$\lim_{n\to\infty}\left(\frac{n}{k_n e^u}\,p_n(x,u)\right) = f(x)$$

uniformly in $u < 0$, i.e., for almost all x, $u \in (-\infty, 0)$, and every $\delta > 0$ we have $p_n(x,u) < \frac{k_n}{n}(f(x)+\delta)e^u$ for all sufficiently large n. Let $Q_n \sim \text{Bin}\big(n-1, k_n\big(f(x)+\delta\big)e^u/n\big)$. Then

$$\begin{aligned} F_{x,n}(u) &= P\{B_{k_n,n-k_n} \le p_n(x,u)\} \le P\left\{B_{k_n,n-k_n} \le \frac{k_n}{n}(f(x)+\delta)e^u\right\} \\ &= P\big(Q_n \ge k_n\big) \le \frac{E(Q_n)}{k_n} = \frac{n-1}{n}(f(x)+\delta)\,e^u \le (f(x)+\delta)e^u. \end{aligned} \tag{40}$$

Thus, for almost all x,

$$\int\limits_{-\infty}^0 |u|^{1+\varepsilon}\,dF_{x,n}(u) = (1+\varepsilon)\int\limits_{-\infty}^0 (-u)^\varepsilon F_{x,n}(u)\,du \le (1+\varepsilon)(f(x)+\delta)\int\limits_{-\infty}^0 (-u)^\varepsilon e^u\,du < \infty \tag{41}$$

for all sufficiently large n. Now on using (39) and (41) in (28), we conclude the first inequality in (18).

ACKNOWLEDGMENTS

The authors thank the referee for his helpful comments and suggestions. The authors are also thankful to Vladimir Hnizdo, Dan Sharp, and Cecil Burchfiel for helpful discussions. The work was supported by the National Institute for Occupational Safety and Health, contract no. 212-2005-M-12857. The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Institute for Occupational Safety and Health.

REFERENCES

1. E. Demchuk and H. Singh, "Statistical Thermodynamics of Hindered Rotation from Computer Simulations", Molecular Physics **99**, 627–636 (2001).
2. D. Scott, *Multivariate Density Estimation: Theory, Practice and Visualization* (Wiley, New York, 1992).
3. L. F. Kozachenko and N. N. Leonenko, "Sample Estimates of Entropy of a Random Vector", Probl. Inform. Transmission **23**, 95–101 (1987).
4. H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and F. Demchuk, "Nearest Neighbor Estimates of Entropy", Amer. J. Math. and Management Sci. **23**, 301–321 (2003).
5. M. N. Goria, N. N. Leonenko, V. V. Mergel, and P. L. Novi Inverardi, "A New Class of Random Vector Entropy Estimators and Its Applications in Testing Statistical Hypotheses", J. Nonparam. Statist. **17**, 277–297 (2005).
6. O. Vasicek, "A Test of Normality Based on Sample Entropy", J. Roy. Statist. Soc. Ser. B **38**, 254–256 (1976).
7. E. J. Dudewicz and E. C. van der Meulen, (1981). "Entropy-Based Tests of Uniformity", J. Amer. Statist. Assos. **76**, 967–974 (1981).
8. M. Abramowitz and A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1965).
9. B. van Es, "Estimating Functionals Related to a Density by Class of Statistics Based on Spacing", Scand. J. Statist. **19**, 61–72 (1992).
10. P. Hall and S. C. Morton, "On the Estimation of Entropy", Ann. Inst. Statist. Math. **45**, 69–88 (1993).
11. J. Beirlant, E. J. Dudewicz, L. Gyorfi, and E. C. van der Meulen, "Nonparametric Entropy Estimation: an Overview", Internat. J. Math. Statist. Sci. **6**, 17–39 (1997).
12. H. Joe, "Estimation of Entropy and Other Functionals of a Multivariate Density", Ann. Inst. Statist. Math. **41**, 683–697 (1989).
13. A. V. Ivanov and M. N. Rozhkova, "Properties of the Statistical Estimate of the Entropy of a Random Vector with a Probability Density", Probl. Inform. Transmission **17**, 171–178 (1981).
14. A. B. Tsybakov and E. C. van der Meulen, "Root-n Consistent Estimators of Entropy for Densities with Unbounded Support", Scand. J. Statist. **23**, 75–83 (1996).
15. D. O. Loftsgaarden and C. P. Quesenberry, "A Non-Parametric Estimate of a Multivariate Density Function", Ann. Math. Statist. **36**, 1049–1051 (1965).
16. N. Leonenko, L. Pronzato, and V. Savani, "A Class of Rényi Information Estimators for Multidimensional Densities", Ann. Statist. (2007) (in press).
17. P. Billingsley, *Convergence of Probability Measures* (Wiley, New York, 1968).
18. P. Billingsley, *Probability and Measure* (Wiley, New York, 1995).
19. E. V. Khmaladze, *The Statistical Analysis of a Large Number of Rare Events*, Technical Report MS-R8804 (Centre for Mathematics and Computer Science, Amsterdam, 1988).
20. M. Loève, *Probability Theory* (Springer, New York, 1977), Vol. I.
21. K. V. Mardia and P. E. Jupp, *Directional Statistics* (Wiley, New York, 2000).
22. N. Misra, H. Singh, and V. Hnizdo, *Nearest Neighbor Estimates of Entropy for Multivariate Circular Distributions*, Preprint (2006).
23. H. Singh, V. Hnizdo, and E. Demchuk, "Probabilistic Model for Two Dependent Circular Variables", Biometrika **89**, 719–723 (2002).