2010

Nearest Neighbor Estimates of Entropy for Multivariate Circular Distributions

Neeraj Misra
West Virginia University

Harshinder Singh
West Virginia University

Vladimir Hnizdo
National Institute for Occupational Safety and Health

Entropy **2010**, *12*, 1125-1144; doi:10.3390/e12051125

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Nearest Neighbor Estimates of Entropy for Multivariate Circular Distributions

Neeraj Misra [1,3,⋆], **Harshinder Singh** [1,2,†] **and Vladimir Hnizdo** [2]

[1] Department of Statistics, West Virginia University, Morgantown, West Virginia 26506, USA

[2] National Institute for Occupational Safety and Health, Morgantown, West Virginia 26505, USA

[3] Department of Mathematics and Statistics, Indian Institute of Technology Kanpur, Kanpur 208 016, India

[†] Deceased.

[⋆] Author to whom correspondence should be addressed; E-Mail: neeraj@iitk.ac.in.

Received: 26 February 2010 / Accepted: 29 April 2010 / Published: 6 May 2010

Abstract: In molecular sciences, the estimation of entropies of molecules is important for the understanding of many chemical and biological processes. Motivated by these applications, we consider the problem of estimating the entropies of circular random vectors and introduce non-parametric estimators based on circular distances between n sample points and their k th nearest neighbors (NN), where k ($\leq n-1$) is a fixed positive integer. The proposed NN estimators are based on two different circular distances, and are proven to be asymptotically unbiased and consistent. The performance of one of the circular-distance estimators is investigated and compared with that of the already established Euclidean-distance NN estimator using Monte Carlo samples from an analytic distribution of six circular variables of an exactly known entropy and a large sample of seven internal-rotation angles in the molecule of tartaric acid, obtained by a realistic molecular-dynamics simulation.

Keywords: circular random variables; differential entropy; non-parametric estimation; nearest neighbor; circular distance; molecular simulation

---

1. Introduction

Estimation of entropies of molecules is an important problem in molecular sciences. Internal configurational entropy of a molecule is the entropy of the joint distribution of the internal molecular coordinates (bond lengths, bond angles, and dihedral angles), and as such it is a measure of random fluctuations in these coordinates. Most significant contribution to the internal configurational entropy of a molecule comes from the fluctuations in dihedral angles (also called internal-rotation angles). Many important properties of complex molecules, such as their stability and adopted conformation, depend on random fluctuations in their internal coordinates. Estimation of the internal configurational entropy of molecules is therefore important for understanding many chemical and biological processes, such as the spontaneity of a chemical reaction, protein folding, intermolecular protein-protein interactions, and protein-ligand interactions. It is also a key in the design of drugs that can stabilize the normally folded molecular structure or correct a misfolded structure, since protein misfolding is a cause of several diseases such as Alzheimer disease, mad cow disease, cystic fibrosis, and some types of cancer.

Estimation of the internal entropy of macromolecules, such as proteins, is a challenging problem because of the large number of correlated internal molecular coordinates. A commonly used method of estimating the internal entropy of a molecule, known as the quasi-harmonic approach, is based on the assumption of a multivariate normal distribution for the internal molecular coordinates [1]. Misra *et al.* [2] discussed the decision theoretic estimation of the entropy of a multivariate normal distribution and obtained improvements over the best affine equivariant estimator under the squared error loss function. However, the assumption of a multivariate normal distribution for the internal coordinates of a molecule is appropriate only at low temperatures, when the fluctuations in its internal coordinates are small. At higher temperatures, the dihedral angles of a complex molecule exhibit multimodes and skewness in their distributions, and the multivariate normal distribution becomes inadequate.

Demchuk and Singh [3] discussed a circular probability approach for modeling the dihedral angles of a molecule in the estimation of internal rotational entropy. As an illustration, they modeled the torsional angle of the methanol molecule by a trimodal von Mises distribution and derived a bath-tub-shaped distribution for the torsional potential energy of the molecule. Singh *et al.* [4] introduced a torus version of a bivariate normal distribution for modeling two dihedral angles. The marginal distributions of the model are symmetric unimodal or symmetric bimodal depending on the configurations of the parameters. A multivariate generalization of this bivariate model has been proposed by Mardia *et al.* [5]. Hnizdo *et al.* [6] and Darian *et al.* [7] used a Fourier series expansion approach for modeling univariate and bivariate distributions of molecular dihedral angles. Complex molecules, however, have many significantly correlated dihedral angles, whose joint distribution can take an arbitrary form. For this reason, a non-parametric approach for estimating the entropy of a circular random vector of arbitrary dimensions is desirable.

Several non-parametric estimators of the entropy of an m-dimensional random variable X have been discussed in the literature. A common approach is to replace the probability density function (pdf) $f(\cdot)$ in the definition of the differential entropy,

$$H(f) = E_f(-\ln f(X)) \tag{1}$$

by its non-parametric kernel or histogram density estimator [8,9]. However, in most practical situations, implementation of such estimates in higher dimensions becomes difficult. In one dimension ($m = 1$), several authors have proposed estimates of entropy in the context of testing goodness of fit [10,11]. Singh *et al.* [12] proposed the following asymptotically unbiased and consistent nearest-neighbor (NN) estimator of the entropy $H(f)$:

$$\hat{H}_{k,n} = \frac{m}{n}\sum_{i=1}^{n} \ln R_{i,k,n} + \ln \frac{\pi^{m/2}}{\Gamma(m/2+1)} + \gamma - L_{k-1} + \ln n \tag{2}$$

Here, $R_{i,k,n}$ is the Euclidean distance of a point X_i to its k^{th}, $k \leq n-1$ nearest (in the Euclidean-distance sense) neighbor in a random sample $X_1, X_2, ..., X_n$ from the distribution $f(\cdot)$; $\gamma = 0.5772\cdots$ is Euler's constant, $L_0 = 0$, $L_j = \sum_{i=1}^{j} 1/i$, $j = 1, 2, \cdots$, and $\Gamma(\cdot)$ is the usual gamma function. For $k = 1$, the estimator $\hat{H}_{k,n}$ reduces to the NN estimator proposed by Kozachenko and Leonenko [13]. Results similar to those of [12] have been also reported by Goria *et al.* [14]. For the purpose of the estimation of the information-theoretic quantity of mutual information, Kraskov *et al.* [15] generalized the first-nearest-neighbor estimator of [13] in terms of k^{th} nearest-neighbor distances in a general metric, giving, however, explicit expressions only for the maximum and Euclidean metrics and without providing formal proofs of asymptotic unbiasedness and consistency. For $m = k = 1$, Tsybakov and van der Meulen [16] established the mean-square-root-n consistency of a truncated version of $\hat{H}_{k,n}$. Earlier, Loftsgaarden and Quesenberry [17] had used NN distances to construct non-parametric estimates of a multivariate pdf. Recently, Mnatsakanov *et al.* [18] studied k-NN estimators of entropy in which the parameter k is assumed to be a function of the sample size.

The NN entropy estimator (2) uses Euclidean distances between the sample points. However, when the random variable X is circular, it is natural to base an NN estimate of entropy on a circular distance rather than the Euclidean distance. A circular observation can be regarded as a point on a circle of unit radius. Once an initial direction and an orientation of the circle have been chosen, each circular observation can be specified by the angle from the initial direction to the point on the circle corresponding to the observation. In this paper, we construct estimates of the entropy of an m-dimensional circular random vector $\Theta \in (0, 2\pi]^m$ based on two different definitions of circular distances. Let $\phi = (\phi_1, \ldots, \phi_m) \in [0, 2\pi)^m$ and $\psi = (\psi_1, \ldots, \psi_m) \in [0, 2\pi)^m$ be two observations on an m-dimensional circular random vector Θ. We define two circular distance functions $d_1(\cdot, \cdot)$ and $d_2(\cdot, \cdot)$ as follows:

$$d_1(\phi, \psi) = \sqrt{\sum_{i=1}^{m} (\pi - |\pi - |\phi_i - \psi_i||)^2} \tag{3}$$

and

$$d_2(\phi, \psi) = \sqrt{2\sum_{i=1}^{m} [1 - \cos(\phi_i - \psi_i)]} \tag{4}$$

Note that $\pi - |\pi - |\phi_i - \psi_i||$, $i = 1, \ldots, m$, is the arc length between the points $(\cos\phi_i, \sin\phi_i)$ and $(\cos\psi_i, \sin\psi_i)$ on the unit circle S_1. On the other hand, $[2(1-\cos(\phi_i - \psi_i))]^{1/2}$ is the Euclidean distance between the points $(\cos\phi_i, \sin\phi_i)$ and $(\cos\psi_i, \sin\psi_i)$ on the unit circle S_1.

In Section 2 and Section 3, we propose explicit expressions for NN estimators of entropy based on the circular distance functions $d_1(\cdot, \cdot)$ and $d_2(\cdot, \cdot)$, respectively, and prove their asymptotic

unbiasedness and consistency (with some mathematical details given in an Appendix). In Section 4, we compare the performance of the estimator based on the circular distance d_1 with that of the Euclidean-distance estimator (2) using Monte Carlo simulations from an analytic 6-dimensional circular distribution, where the exact joint entropy is known. We there also apply the d_1-distance estimator to the problem of estimating the entropy of a 7-dimensional joint distribution of the internal-rotation angles in the molecule of tartaric acid, using a large sample of these angles obtained by a realistic molecular-dynamics simulation.

2. Nearest Neighbor Estimates of Entropy Based on Circular Distance d_1

For constructing nearest neighbor estimates of the entropy of a circular random vector based on the distance function (3), we first derive the expression for the volume of a ball

$$N_r(\psi) = \{\theta \in [0, 2\pi)^m : d_1(\theta, \psi) < r\} \tag{5}$$

centered at $\psi \in [0, 2\pi)^m$ and having a radius $r \in [0, \sqrt{m}\pi]$, $\sqrt{m}\pi$ being the maximum value of $d_1(\cdot, \cdot)$.

Lemma 2.1. Let $r \in [0, \sqrt{m}\pi]$, $\psi \in [0, 2\pi)^m$ and let V_r be the volume of the ball $N_r(\psi)$, defined by (5). Then

$$V_r = (2\pi)^m A_m \left(\frac{r^2}{\pi^2}\right) \tag{6}$$

where $A_m(\cdot)$ denotes the cumulative distribution function of the sum of m independent and identically distributed random variables each having a beta distribution with parameters $\alpha = \frac{1}{2}$, $\beta = 1$.

Proof. Without loss of generality, we may take $\psi = (0, \ldots, 0)$. Then

$$V_r = \int_R d\theta = (2\pi)^m \Pr\left(\sum_{i=1}^{m} (\pi - \mid \pi - U_i \mid)^2 < r^2\right)$$

where $R = \{\theta = (\theta_1, \ldots, \theta_m) : \theta \in [0, 2\pi)^m, \sum_{i=1}^m (\pi - |\pi - \theta_i|)^2 < r^2\}$, $d\theta = d\theta_1 \ldots d\theta_m$ and $U_1, \ldots, U_m$ are independent and identically distributed uniform random variables over the interval $(0, 2\pi)$. Define $C_i = (\pi - \mid \pi - U_i \mid)^2/\pi^2$, $i = 1, \cdots, m$. Then $C_1, \cdots, C_m$ are independent and identically distributed beta random variables, having parameters $\alpha = \frac{1}{2}$, $\beta = 1$. Hence the result follows.

Remark 2.1. (i) For $m = 1$ and $x \in [0, 1]$, we have $A_1(x) = \sqrt{x}$. For $m = 2$ and $x \in [0, 2]$, it can be verified that

$$A_2(x) = \begin{cases} \frac{\pi x}{4}, & \text{if } 0 \le x \le 1 \\ \sqrt{x-1} + \frac{x}{2}(2 \arcsin \sqrt{\frac{1}{x}} - \frac{\pi}{2}), & \text{if } 1 < x \le 2 \end{cases}$$

(ii) For $m = 3$ and $x \in [0, 3]$, it can be verified that

$$A_3(x) = \begin{cases} \frac{\pi x^{3/2}}{6}, & \text{if } 0 \le x \le 1 \\ \frac{\pi}{12}(-3 + 9x - 4x^{3/2}), & \text{if } 1 \le x \le 2 \\ \sqrt{x-2} + \frac{1}{12}\left\{(1-3x)(\pi - 4 \arcsin \frac{1}{\sqrt{x-1}}) + (4-12x) \arctan \frac{x-3}{2\sqrt{x-2}}\right\} & \\ +\frac{1}{3}x^{3/2}\left\{\arctan \sqrt{x(x-2)} + \arctan \frac{x-\sqrt{x}-1}{\sqrt{x-2}} - \arctan \frac{x+\sqrt{x}-1}{\sqrt{x-2}}\right\}, & \text{if } 2 < x \le 3 \end{cases}$$

(iii) For a general m (≥ 1) and $x \in [0,1]$, it can be verified that

$$A_m(x) = \frac{(\pi x)^{\frac{m}{2}}}{2^m \Gamma(1+\frac{m}{2})}$$

(iv) For any $m \geq 2$ and for $1 \leq x \leq m$, $A_m(x)$ satisfies the following recursion relation

$$A_m(x) = \begin{cases} \int_0^1 A_{m-1}(x-s) f_C(s) ds, & \text{if } 1 \leq x \leq m-1 \\ A_1(x-m+1) + \int_{x-m+1}^1 A_{m-1}(x-s) f_C(s) ds, & \text{if } m-1 \leq x \leq m \end{cases}$$

where $f_C(\cdot)$ is the pdf of a beta random variable with parameters $\alpha = 1/2$ and $\beta = 1$.

The circular distance (3) becomes the Euclidean distance $d_E(\phi,\psi) = [\sum_{i=1}^m (\phi_i - \psi_i)^2]^{1/2}$ when $|\phi_i - \psi_i| \leq \pi$, $i = 1,2,\ldots,m$. Circular-distance balls $N_{r_q}(\theta)$, where $\theta \in (0,2\pi)^m$ and $\lim_{q\to\infty} r_q = 0$, thus tend to the corresponding Euclidean-distance balls as $q \to \infty$. We can therefore apply the Lebesgue differentiation theorem [19] to the probability density function $f(\theta)$ of a circular random variable $\Theta \in [0,2\pi)^m$ in the form

$$\lim_{q\to\infty} \frac{1}{V_{r_q}} \int_{N_{r_q}(\theta)} f(\mu)\, d\mu = f(\theta), \quad \text{at almost all } \theta \in [0,2\pi)^m \tag{7}$$

where V_{r_q} is given by (6). Equation (7) suggests that, given a sufficiently large random sample $\Theta_1,\ldots,\Theta_n$ from the distribution of Θ, the probability density function $f(\theta)$ can be approximated at almost all $\theta \in [0,2\pi)^m$ by

$$\hat{f}_n^{(1)}(\theta) = \frac{|\, N_r(\theta)\,|}{nV_r} = \frac{|\, N_r(\theta)\,|}{n(2\pi)^m A_m(\frac{r^2}{\pi^2})} \tag{8}$$

where $|\, N_r(\theta)\,|$ denotes the cardinality of the set $\{i : \Theta_i \in N_r(\theta)\}$ and r is sufficiently small.

Guided by this insight, we will now construct nearest neighbor estimates of entropy for an m-variate circular random vector Θ, having a probability density function $f(\cdot)$. Let $\Theta_1,\ldots,\Theta_n$ be a random sample from the distribution of Θ and let $k \in \{1,\ldots,n-1\}$ be a given positive integer. For $i \in \{1,\ldots,n\}$, let $d_1(i,k,n)$ denote the circular distance of Θ_i from its k^{th} closest neighbor, with respect to the circular distance $d_1(\cdot,\cdot)$, *i.e.*,

$$d_1(i,k,n) = k^{\text{th}} \text{ smallest of } \{d_1(\Theta_j,\Theta_i)\ j = 1,\ldots,n,\ j \neq i\},\ i = 1,\ldots,n$$

Assume that the sample size n is sufficiently large, so that the distances $d_1(i,k,n)$ are small, on the average. Then, based on approximation (8), a reasonable estimator of $f(\Theta_i)$ is

$$\hat{f}_{k,n}^{(1)}(\Theta_i) = \frac{k}{n(2\pi)^m A_m(\frac{d_1^2(i,k,n)}{\pi^2})}$$

and thus a reasonable estimator of the entropy $H(f) = E(-\ln f(\Theta))$ is

$$\hat{G}_{k,n}^{(1)} = -\frac{1}{n}\sum_{i=1}^n \ln \hat{f}_{k,n}^{(1)}(\Theta_i) = \frac{1}{n}\sum_{i=1}^n \ln\left(\frac{n}{k}(2\pi)^m A_m\Big(\frac{d_1^2(i,k,n)}{\pi^2}\Big)\right) \tag{9}$$

In the following theorem, we derive the expression for the asymptotic mean of the estimator $\hat{G}_{k,n}^{(1)}$. Apart from the arguments for the interchange of the limit and the integral signs, the proof is similar to that of Theorem 8 of Singh *et al.* [12], who in their proof interchange the limit and the integral signs without mentioning the conditions under which it is allowed.

Theorem 2.1. Suppose that there exists an $\epsilon > 0$, such that

$$\int_{[0,2\pi)^m} |\ln f(\theta)|^{1+\epsilon} f(\theta) d\theta < \infty \tag{10}$$

and

$$\int_{[0,2\pi)^m} \int_{[0,2\pi)^m} |\ln d_1(\theta,\mu)|^{1+\epsilon} f(\theta) f(\mu) d\theta d\mu < \infty \tag{11}$$

Then, for a fixed $k \in \{1, 2, \ldots, \}$ (not depending on n)

$$\lim_{n\to\infty} E_f\left(\hat{G}_{k,n}^{(1)}\right) = L_{k-1} - \gamma - \ln k + H(f)$$

where $\hat{G}_{k,n}^{(1)}$ is defined by (9), $L_0 = 0$, $L_j = \sum_{i=1}^{j} \frac{1}{i}$, $j = 1, 2, \ldots$ and $\gamma = -\int_0^\infty (\ln t) e^{-t} dt = 0.5772\cdots$ is Euler's constant.

Proof. Let

$$T_{i,k,n} = \ln\left(\frac{n}{k}(2\pi)^m A_m\Big(\frac{d_1^2(i,k,n)}{\pi^2}\Big)\right),\ i = 1, \ldots, n$$

Then $T_{1,k,n}, \ldots, T_{n,k,n}$ are identically distributed random variables. Therefore,

$$E_f\left(\hat{G}_{k,n}^{(1)}\right) = E_f\left(T_{1,k,n}\right) = \int_{[0,2\pi)^m} E_f\left(S_{\theta,k,n}\right) f(\theta) d\theta \tag{12}$$

where, for a given $\theta \in [0, 2\pi)^m$, $S_{\theta,k,n}$ is a random variable having the same distribution as that of the conditional distribution of $T_{1,k,n}$ given $\Theta_1 = \theta$.

$$\rho_{k,n}(u) = \pi\sqrt{A_m^{-1}\left(\frac{ke^u}{n(2\pi)^m}\right)}$$

where $A_m^{-1}(\cdot)$ denotes the inverse function of $A_m(\cdot)$. Using standard arguments, we get

$$\begin{aligned} P_f\left(S_{\theta,k,n} \le u\right) &= 1 - P_f\left(d_1(1,k,n) > \rho_{k,n}(u) \mid \Theta_1 = \theta\right) \\ &= 1 - \sum_{j=0}^{k-1} \binom{n-1}{j} \left(P_f\left(N_{\rho_{k,n}(u)}(\theta)\right)\right)^j \left(1 - P_f\left(N_{\rho_{k,n}(u)}(\theta)\right)\right)^{n-1-j} \end{aligned}$$

where $N_r(\cdot)$ is defined by (5). For a fixed $u \in (-\infty, \infty)$, $k \in \{1, 2, \ldots\}$ and for almost all values of $\theta \in [0, 2\pi)^m$, using Lemma 2.1, we have

$$\lim_{n\to\infty}\left(nP_f\left(N_{\rho_{k,n}(u)}(\theta)\right)\right) = ke^u \lim_{n\to\infty}\left(\frac{1}{V_{\rho_{k,n}(u)}} \int_{N_{\rho_{k,n}(u)}(\theta)} f(\mu) d\mu\right) = ke^u f(\theta)$$

Therefore, using the Poisson approximation to the binomial distribution, we get

$$\lim_{n\to\infty} P_f\left(S_{\theta,k,n} \le u\right) = 1 - \sum_{j=0}^{k-1} e^{-kf(\theta)e^u} \frac{\left(kf(\theta)e^u\right)^j}{j!} = \frac{1}{\Gamma(k)} \int_0^{kf(\theta)e^u} e^{-t} t^{k-1} dt \tag{13}$$

for almost all values of $\theta \in [0, 2\pi)^m$. For a fixed $\theta \in [0, 2\pi)^m$, let $S_{\theta,k}$ be a random variable having the pdf

$$g_{\theta,k}(u) = \frac{e^{-kf(\theta)e^u}\left(kf(\theta)e^u\right)^k}{\Gamma(k)},\ -\infty < u < \infty$$

Then, in view of (13),

$$S_{\theta,k,n} \xrightarrow{d} S_{\theta,k}, \text{ as } n \to \infty \tag{14}$$

where $\xrightarrow{d}$ stands for the convergence in distribution. For each fixed $\theta \in [0, 2\pi)^m$, it can be verified that

$$E_f\left(S_{\theta,k}\right) = \frac{1}{\Gamma(k)} \int_0^\infty (\ln t) e^{-t} t^{k-1} dt - \ln k - \ln f(\theta) = L_{k-1} - \gamma - \ln k - \ln f(\theta)$$

Under the condition (10), it can be shown (for details, see the Appendix) that, for almost all values of $\theta \in [0, 2\pi)^m$, there exists a constant C (not depending on n) such that for all sufficiently large values of n

$$E_f\left(\mid S_{\theta,k,n} \mid^{1+\epsilon}\right) < C \tag{15}$$

Then, in view of (14) and the moment convergence theorem, it follows that

$$\lim_{n\to\infty} E_f\left(S_{\theta,k,n}\right) = E_f\left(S_{\theta,k}\right) = L_{k-1} - \gamma - \ln k - \ln f(\theta) \tag{16}$$

for almost all values of $\theta \in [0, 2\pi)^m$ Using Fatou's lemma, we get

$$\begin{aligned}
&\limsup_{n\to\infty} \int_{[0,2\pi)^m} \mid E_f\left(S_{\theta,k,n}\right) \mid^{1+\epsilon} f(\theta) d\theta \\
&\quad \leq \int_{[0,2\pi)^m} \limsup_{n\to\infty} \mid E_f\left(S_{\theta,k,n}\right) \mid^{1+\epsilon} f(\theta) d\theta \\
&\quad = \int_{[0,2\pi)^m} \mid L_{k-1} - \gamma - \ln k - \ln f(\theta) \mid^{1+\epsilon} f(\theta) d\theta \\
&\quad \leq 2^{\epsilon-1} \left(\mid L_{k-1} - \gamma - \ln k \mid^{1+\epsilon} + \int_{[0,2\pi)^m} \mid \ln f(\theta) \mid^{1+\epsilon} f(\theta) d\theta \right) < \infty
\end{aligned}$$

by (10). Therefore,

$$\lim_{n\to\infty} \int_{[0,2\pi)^m} E_f\left(S_{\theta,k,n}\right) f(\theta) d\theta = \int_{[0,2\pi)^m} \lim_{n\to\infty} E_f\left(S_{\theta,k,n}\right) f(\theta) d\theta = \int_{[0,2\pi)^m} E_f\left(S_{\theta,k}\right) f(\theta) d\theta$$

Now the result follows from (12) and (16).

Since the estimator $\hat{G}_{k,n}^{(1)}$ is not asymptotically unbiased, we propose the following (asymptotic) bias corrected estimator for estimating the entropy $H(f)$:

$$\hat{H}_{k,n}^{(1)} = \frac{1}{n} \sum_{i=1}^{n} \ln\left(A_m\left(\frac{d_1^2(i,k,n)}{\pi^2} \right) \right) + \ln\left(n(2\pi)^m\right) - L_{k-1} + \gamma \tag{17}$$

Thus, we have the following corollary to Theorem 2.1.

Corollary 2.1. Under the assumptions of Theorem 2.1, the estimator $\hat{H}_{k,n}^{(1)}$ is asymptotically unbiased for estimating the entropy $H(f)$.

The following theorem provides conditions under which the estimator $\hat{H}_{k,n}^{(1)}$ is consistent for estimating the entropy $H(f)$.

Theorem 2.2. Suppose that there exists an $\epsilon > 0$, such that

$$\int_{[0,2\pi)^m} \mid \ln f(\theta) \mid^{2+\epsilon} f(\theta) d\theta < \infty \tag{18}$$

and

$$\int_{[0,2\pi)^m} \int_{[0,2\pi)^m} |\ln d_1(\theta,\mu)|^{2+\epsilon} f(\theta) f(\mu) d\theta d\mu < \infty \tag{19}$$

Then, for a fixed $k \in \{1, 2, \ldots\}$ (not depending on n),

$$\lim_{n\to\infty} \mathrm{Var}_f \left(\hat{H}_{k,n}^{(1)} \right) = 0$$

and thus $\hat{H}_{k,n}^{(1)}$ is a consistent estimator of the entropy $H(f)$.

Under conditions (18) and (19), in the proof of Theorem 2.2, the steps involved in justifying the interchange of the limit and the integral sign are tedious but virtually identical to the arguments used in the proof of Theorem 2.1. The remaining part of the proof is identical to the proof of Theorem 11 of Singh *et al.* [12]. We therefore omit the proof of Theorem 2.2.

Remark 2.2. For small values of $m \geq 4$, the function $A_m(\cdot)$ involved in the evaluation of estimate $\hat{H}_{k,n}^{(1)}$ can be computed using numerical integration. For moderate and large values of m, which is the case with many molecules encountered in molecular sciences, using the central limit theorem one can get a reasonable approximation of $A_m(\cdot)$ by the cumulative distribution function of a normal distribution having mean $m/3$ and variance $4m/45$.

3. Nearest Neighbor Estimates of Entropy Based on Circular Distance d_2

Let $\psi \in [0, 2\pi)^m$. In order to construct nearest neighbor estimates of entropy based on the distance function $d_2(\cdot,\cdot)$, defined by (4), we require the volume of the ball to be

$$S_r(\psi) = \{\theta \in [0, 2\pi)^m : d_2(\theta, \psi) < r\} \tag{20}$$

centered at $\psi \in [0, 2\pi)^m$ and having radius $r \in [0, 2\sqrt{m}]$.

Lemma 3.1. Let $r \in [0, 2\sqrt{m}]$, $\psi \in [0, 2\pi)^m$ and let W_r be the volume of the ball $S_r(\psi)$, defined by (20). Then

$$W_r = (2\pi)^m B_m \left(\frac{r^2}{4} \right)$$

where $B_m(\cdot)$ denotes the cumulative distribution function of the sum of m independent and identically distributed random variables each having a beta distribution with parameters $\alpha = \frac{1}{2}$ and $\beta = \frac{1}{2}$

Proof. We have

$$W_r = \int_{\{\theta\in[0,2\pi)^m,\ \sum_{i=1}^m (1-\cos(\theta_i)) < \frac{r^2}{2}\}} d\theta = (2\pi)^m \mathrm{Pr} \left(\sum_{i=1}^m \frac{1-\cos U_i}{2} < \frac{r^2}{4} \right)$$

where $U_1, \ldots, U_m$ are independent and identically distributed uniform random variables over the interval $[0, 2\pi)$.

Define $D_i = (1 - \cos U_i)/2,\ i = 1, \ldots, m$. Then, $D_1, \ldots, D_m$ are independent and identically distributed beta random variables, having parameters $\alpha = \frac{1}{2}$ $\beta = \frac{1}{2}$. Hence the result follows.

Remark 3.1. (i) For $m = 1$ and $x \in (0, 1)$,

$$B_1(x) = \frac{2}{\pi} \arcsin(\sqrt{x}) = \frac{1}{\pi} \arccos(1 - 2x)$$

(ii) $B_m(x)$ satisfies a similar recursion relation as that satisfied by $A_m(x)$ and given in Remark 2.1 (iv).

For $i \in \{1, \ldots, n\}$, let $d_2(i, k, n)$ denote the circular distance of Θ_i from its k^{th} closest neighbor with respect to the circular distance $d_2(\cdot, \cdot)$, defined by (4). Assume that the sample size n is sufficiently large, so that on average the distances $d_2(i, k, n)$ are small. Then, based on approximation (7), a reasonable estimator of $f(\Theta_i)$ is

$$\hat{f}^{(2)}_{k,n}(\Theta_i) = \frac{k}{n(2\pi)^m B_m\left(\frac{1}{4}\, d_2^2(i,k,n)\right)}$$

and thus a reasonable estimator of the entropy $H(f) = E(-\ln f(\Theta))$ is

$$\hat{G}^{(2)}_{k,n} = -\frac{1}{n}\sum_{i=1}^{n} \ln \hat{f}^{(2)}_{k,n}(\Theta_i) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(\frac{n(2\pi)^m}{k} B_m\left(\tfrac{1}{4}\, d_2^2(i,k,n)\right)\right) \tag{21}$$

The proof of the following theorem is identical to the proof of Theorem 2.1 and therefore it is omitted.

Theorem 3.1. Suppose that there exists an $\epsilon > 0$ such that (10) holds and

$$\int_{[0,2\pi)^m} \int_{[0,2\pi)^m} |\ln d_2(\theta, \mu)\ |^{1+\epsilon}\ f(\theta) f(\mu) d\theta d\mu < \infty \tag{22}$$

Then, for a fixed $k \in \{1, 2, \ldots, \}$ (not depending on n),

$$\lim_{n\to\infty} E_f\left(\hat{G}^{(2)}_{k,n}\right) = L_{k-1} - \gamma - \ln k + H(f)$$

where $\hat{G}^{(2)}_{k,n}$ is defined by (21).

Since the estimator $\hat{G}^{(2)}_{k,n}$ is not asymptotically unbiased, we propose the following (asymptotic) bias corrected estimator for estimating the entropy $H(f)$:

$$\hat{H}^{(2)}_{k,n} = \frac{1}{n}\sum_{i=1}^{n} \ln\left(B_m\left(\tfrac{1}{4}\, d_2^2(i,k,n)\right)\right) + \ln\left(n(2\pi)^m\right) - L_{k-1} + \gamma \tag{23}$$

Thus, we have the following corollary to Theorem 3.1.

Corollary 3.1. Under the assumptions of Theorem 3.1, the estimator $\hat{H}^{(2)}_{k,n}$ is asymptotically unbiased for estimating the entropy $H(f)$.

The following theorem provides conditions under which the estimator $\hat{H}^{(2)}_{k,n}$ is consistent for estimating the entropy $H(f)$. The proof of the theorem follows using the arguments similar to the one given for the proof of Theorem 2.2.

Theorem 3.2. Suppose that there exists an $\epsilon > 0$ such that (18) holds and

$$\int_{[0,2\pi)^m} \int_{[0,2\pi)^m} |\ln d_2(\theta, \mu)\ |^{2+\epsilon}\ f(\theta) f(\mu) d\theta d\mu < \infty \tag{24}$$

Then, for a fixed $k \in \{1, 2, \ldots\}$ (not depending on n),

$$\lim_{n\to\infty} \text{Var}_f\left(\hat{H}^{(2)}_{k,n}\right) = 0$$

and therefore, under conditions (18) and (24), $\hat{H}^{(2)}_{k,n}$ is a consistent estimator of the entropy $H(f)$.

Remark 3.2. (i) Using Remark 3.1 (i) and the fact that $\arccos(x) \in [0, \pi]$ is a decreasing function of $x \in [-1, 1]$, for $m = 1$ and $i \in \{1, \ldots, n\}$, we have

$$\begin{aligned} B_1\left(\tfrac{1}{4}\, d_2^2(i,k,n)\right) &= \frac{1}{\pi} \arccos\left(1 - \frac{d_2^2(i,k,n)}{2}\right) \\ &= \frac{1}{\pi} \times k^{\text{th}} \text{ smallest of } \left\{\arccos\left(\cos\left(\Theta_j - \Theta_i\right)\right),\ j = 1, \ldots, n,\ j \neq i\right\} \\ &= \frac{1}{\pi} \times k^{\text{th}} \text{ smallest of } \left\{\pi - \mid \pi - \mid \Theta_j - \Theta_i \mid\mid,\ j = 1, \ldots, n,\ j \neq i\right\} \end{aligned}$$

since $\arccos(\cos x) = \pi - \mid \pi - \mid x \mid\mid,\ x \in [-2\pi, 2\pi]$
Therefore, for $m = 1$ and $i \in \{1, \ldots, n\}$, we have

$$B_1\left(\tfrac{1}{4}\, d_2^2(i,k,n)\right) = \frac{d_1(i,k,n)}{\pi} \ \Rightarrow\ \hat{G}_{k,n}^{(1)} = \hat{G}_{k,n}^{(2)} \ \Rightarrow\ \hat{H}_{k,n}^{(1)} = \hat{H}_{k,n}^{(2)}$$

Thus for $m = 1$, estimators $\hat{H}_{k,n}^{(1)}$ and $\hat{H}_{k,n}^{(2)}$, based on circular distance functions $d_1(\cdot,\cdot)$ and $d_2(\cdot,\cdot)$ respectively, are identical.

(ii) For small values of $m \geq 2$, the function $B_m(\cdot)$ involved in the evaluation of estimate $\hat{H}_{k,n}^{(2)}$ can be computed using numerical integration. For moderate and large values of m, which is the case with many molecules encountered in molecular sciences, using the central limit theorem one can get a reasonable approximation of $B_m(\cdot)$ by the cumulative distribution function of a normal distribution having mean $m/2$ and variance $m/8$.

With $k \in \{1, 2, \ldots\}$, (17) and (23) define two classes of estimators for the entropy $H(f)$. The biases and variances of these estimators depend on k, the sample size n, and the pdf $f(\cdot)$ and its dimensions m. It would be useful to have the knowledge of the biases and variances as functions of k, n, m, and some characteristic of $f(\cdot)$, such as $\mu = E(h(f(\Theta)))$, where $h(\cdot)$ is a function such that a reliable estimate of μ can be obtained using the available data on Θ. We have not been able to derive any meaningful expressions for the biases and variances of the proposed estimators, and this problem is under further investigation.

4. Monte Carlo Results and a Molecular Entropy Example

The performance of an entropy estimator can be investigated rigorously by using Monte Carlo samples from a distribution for which the entropy is known exactly. While analytic distributions of more than two correlated circular variables with exactly calculable entropic attributes do not seem available, one may construct a distribution of higher dimensionality as a product of a suitable number of bivariate distributions. To test the performance of the circular-distance estimator (17), we used an analytic 6-dimensional circular distribution given as the product of three bivariate circular distributions, each of the form [4]

$$f(\theta_1, \theta_2) = C e^{\kappa_1 \cos(\theta_1 - \mu_1) + \kappa_2 \cos(\theta_2 - \mu_2) + \lambda \sin(\theta_1 - \mu_1) \sin(\theta_2 - \mu_2)}$$

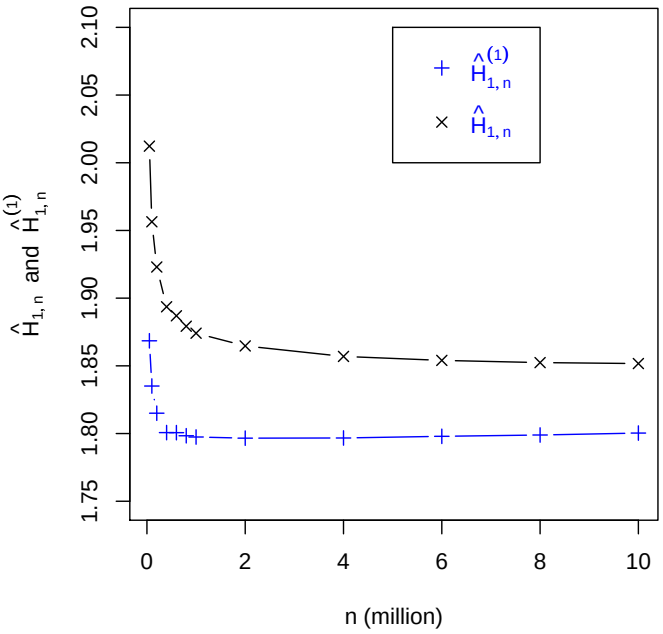
which, as a circular analogue of the bivariate normal distribution, can be called the bivariate von Mises distribution. Details pertaining to the 6-dimensional distribution used and the Monte Carlo sampling are given in [20], where the same circular distribution was used in an investigation of the combined mutual-information-expansion and Euclidean-distance-NN method of entropy estimation.

Table 1. Circular- and Euclidean-distance estimates $\hat{H}^{(1)}_{k,n}$ and $\hat{H}_{k,n}$, respectively, from samples of size n of the analytic distribution of 6 circular variables; the exact entropy value is (to 4 decimals) $H = 1.8334$.

$n \times 10^{-6}$	$\hat{H}^{(1)}_{1,n}/\hat{H}_{1,n}$	$\hat{H}^{(1)}_{2,n}/\hat{H}_{2,n}$	$\hat{H}^{(1)}_{3,n}/\hat{H}_{3,n}$	$\hat{H}^{(1)}_{4,n}/\hat{H}_{4,n}$	$\hat{H}^{(1)}_{5,n}/\hat{H}_{5,n}$
0.05	1.86851	1.92019	1.96379	2.00395	2.03653
	2.01225	2.09929	2.16540	2.22380	2.27390
0.10	1.81503	1.86968	1.90481	1.92899	1.95164
	1.95640	2.01952	2.07339	2.11255	2.14865
0.20	1.80070	1.83713	1.85672	1.87498	1.89122
	1.92305	1.96691	2.00110	2.03086	2.05650
0.40	1.8007	1.81582	1.82840	1.83904	1.84862
	1.89363	1.92652	1.95001	1.96939	1.98668
0.60	1.80066	1.80693	1.81496	1.82296	1.83097
	1.88696	1.90952	1.92764	1.94358	1.95783
0.80	1.79837	1.80297	1.81063	1.81660	1.82222
	1.87936	1.89971	1.91649	1.92946	1.94125
1.00	1.79539	1.80017	1.80566	1.81030	1.81566
	1.87317	1.89238	1.90694	1.91850	1.92915
2.00	1.79660	1.79533	1.79736	1.79970	1.80266
	1.86471	1.87555	1.88493	1.89298	1.90073
4.00	1.79673	1.79383	1.79404	1.79480	1.79613
	1.85696	1.86477	1.87136	1.87702	1.88211
6.00	1.79795	1.79491	1.79385	1.79419	1.79458
	1.85403	1.86071	1.86552	1.87029	1.87414
8.00	1.79893	1.79484	1.79373	1.79322	1.79337
	1.85240	1.85745	1.86197	1.86545	1.86891
10.00	1.80036	1.79562	1.79426	1.79350	1.79329
	1.85170	1.85578	1.85969	1.86287	1.86583

Table 1 presents the circular-distance estimates $\hat{H}^{(1)}_{k,n}$, $k = 1, \ldots, 5$ obtained from samples of sizes in the range $n = 5 \times 10^4$–1×10^7, together with the corresponding Euclidean-distance estimates $\hat{H}_{k,n}$. Figure 1 displays the estimates $\hat{H}^{(1)}_{k=1,n}$ and $\hat{H}_{k=1,n}$ as functions of the sample size n. Noting that the exact entropy value here is, to 4 decimal places, $H = 1.8334$, we observe that as n increases the circular-distance estimates initially "undershoot" the exact value and then start to approach it slowly from below. In contrast, the Euclidean-distance estimates approach the exact value monotonically from above. Interestingly, the biases of the two kinds of estimates at sample sizes $n \gtrsim 1$ million are approximately equal in absolute value. The behavior of the circular-distance estimates at $k = 2, \ldots, 5$ is similar to that at $k = 1$, and the estimate values at different k's become very close at $n \gtrsim 1$ million.

Figure 1. Plots of the circular- and Euclidean-distance estimates $\hat{H}^{(1)}_{1,n}$ and $\hat{H}_{1,n}$, respectively, as functions of the sample size n, for the analytic distribution of 6 circular variables; the exact entropy value is (to 4 decimals) $H = 1.8334$.



To investigate the usefulness of circular-distance NN estimators in the problem of evaluating the configurational entropy of internal rotations in molecules, we used the circular-distance estimator $\hat{H}^{(1)}_{k,n}$ to estimate the entropy of the joint distribution of internal-rotation angles in the molecule of tartaric acid, where the number of variables is $m = 7$. Samples of size n up to 14.4 million of the internal-rotation angles were obtained from a molecular dynamics simulation of the (R,S) stereoisomer of this molecule [21]. Figure 2 shows marginal histograms, smoothed using a Gaussian kernel, of the seven internal-rotation angles of tartaric acid; note that these marginals display markedly non-Gaussian features. The code ANN [22] (with our modification for the circular distance d_1), which utilizes a k-d tree algorithm [23], was used for finding the k-th NN distances between sample points. Figure 3 presents the estimates $\hat{H}^{(1)}_{k,n}$, $k = 1, ..., 5$ as functions of the sample size n. The values of $\hat{H}^{(1)}_{k,n}$ decrease as n increases, while, at a fixed value of $n \lesssim 7$ million, they increase as k increases; at greater values of n, the estimates at different k's become quite close in value. Figure 4 compares the circular-distance estimates $\hat{H}^{(1)}_{1,n}$ and $\hat{H}^{(1)}_{5,n}$ with the corresponding Euclidean-distance estimates $\hat{H}_{k,n}$. We note that an $n \to \infty$ extrapolated Euclidean-distance estimate $\hat{H} = 5.04 \pm 0.01$ was obtained for this entropy in [21]. Again, as in the case of the analytic circular distribution, this value approximately equals the arithmetic mean of the circular and Euclidean distance estimates at sample sizes $n \gtrsim 1$ million.

Figure 2. Smoothed marginal histograms of the internal-rotation angles ϕ_i, $i = 1, \ldots, 7$ of the (R,S) isomer of tartaric acid obtained by molecular dynamics simulations.

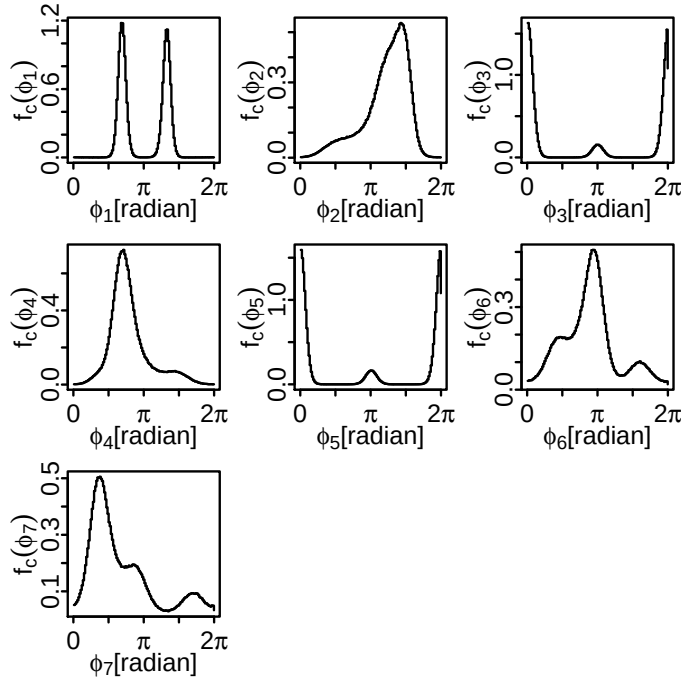


Figure 3. Circular-distance nearest-neighbor estimates $\hat{H}^{(1)}_{k,n}$, $k = 1, \ldots, 5$ of the entropy of the 7-dimensional joint distribution of internal-rotation angles in the (R,S) isomer of tartaric acid as functions of the sample size n. The estimates $\hat{H}^{(1)}_{k,n}$ at a fixed $n \lesssim 7$ million increase in value as k increases.

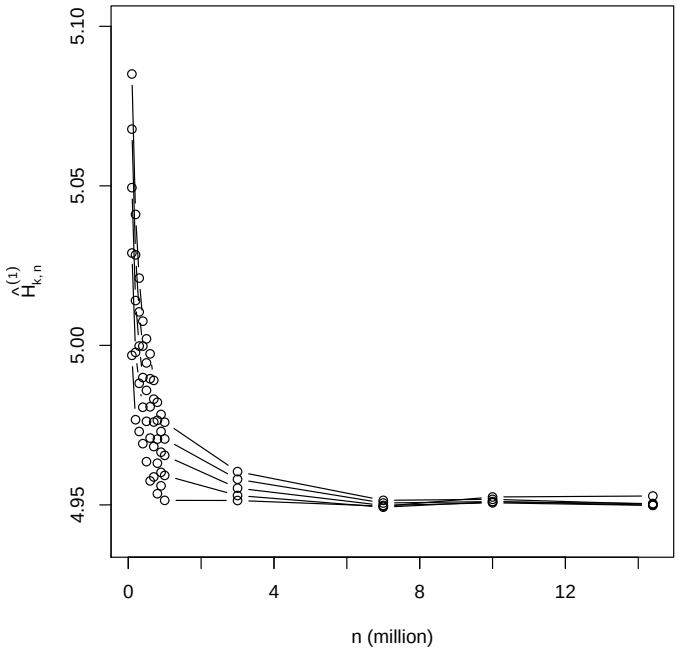
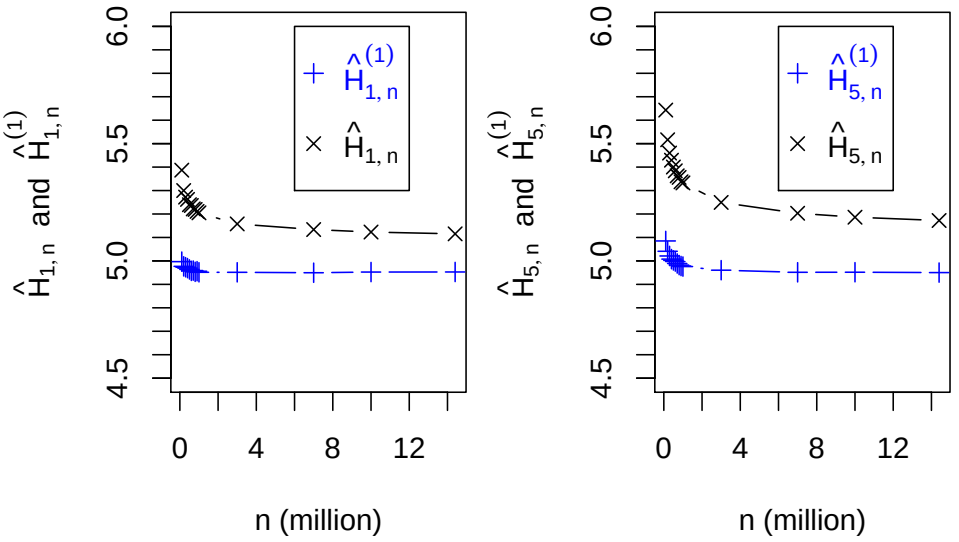


Figure 4. Circular-distance nearest-neighbor estimates $\hat{H}^{(1)}_{1,n}$ and $\hat{H}^{(1)}_{5,n}$ of the internal-rotation entropy of tartaric acid as functions of the sample size n compared with the Euclidean-distance nearest-neighbor estimates $\hat{H}_{1,n}$ and $\hat{H}_{5,n}$. An $n \to \infty$ extrapolated estimate is $\hat{H} = 5.04 \pm 0.01$ [21].



Perhaps surprisingly, the results of both the analytic-distribution and molecular-simulation studies undertaken here indicate that only when relatively small data samples are available, the use of a circular-distance estimator has some advantage over the Euclidean-distance estimator. On samples of large size, needed for sufficient convergence of an NN estimate of the entropy of a multivariate distribution, the circular-distance estimates obtained did not have a significantly smaller bias than the Euclidean-distance estimates. In view of such findings, one may question whether the additional computational complexity of a circular-distance estimate is worth the effort. However, we observed that as the sample size increased, the circular NN distances in the sample became quickly so small that the circular-distance estimator $\hat{H}^{(1)}_{k,n}$ coincided in value with the simpler Euclidean-distance estimator $\hat{H}_{k,n}$ in which the same NN-distance values were used. This is explained by the fact that when the circular NN distances $d_1(i,k,n) \leq \pi$, the estimator $\hat{H}^{(1)}_{k,n}$ can be replaced with the estimator $\hat{H}_{k,n}$ in which the NN distances $d_1(i,k,n)$ are substituted for the Euclidean NN distances $R(i,k,n)$; this fact follows directly from Remark 2.1 (iii). The only extra computational effort is then expended in finding the circular, instead of Euclidean, NN distances in a given sample.

Acknowledgments

The authors are thankful to Jun Tan for carrying out the circular-distance NN calculations, and to E. James Harner, Cecil Burchfiel, Robert Mnatsakanov and Dan S. Sharp for helpful discussions. 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.

Appendix

Here we provide the proof of (15). For a fixed $\theta \in [0, 2\pi)^m$, let $F_{\theta,k,n}(\cdot)$ be the distribution function of the random variable $S_{\theta,k,n}$. To establish (15), we will first show that, for $n = 2$ and $k = 1$,

$$E_f\left(\mid S_{\theta,1,2} \mid^{1+\epsilon}\right) = \int_{-\infty}^{\infty} \mid u \mid^{1+\epsilon} dF_{\theta,1,2}(u) < \infty \tag{25}$$

for almost all values of $\theta \in [0, 2\pi)^m$. In order to establish (25), consider

$$\begin{aligned} E_f\left(\mid S_{\theta,1,2} \mid^{1+\epsilon}\right) &= E_f\left(\left|\ln\left[2(2\pi)^m A_m\left(d_1^2(1,1,2)/\pi^2\right)\right]\right|^{1+\epsilon} \Big| \Theta_1 = \theta\right) \\ &= E_f\left(\left|\ln\left[2(2\pi)^m\right] + \ln\left[A_m\left(d_1^2(\theta,\Theta_2)/\pi^2\right)\right]\right|^{1+\epsilon}\right) \\ &\leq 2^\epsilon\left(\mid \ln\left[2(2\pi)^m\right] \mid^{1+\epsilon} + E_f\left(\left|\ln\left[A_m\left(d_1^2(\theta,\Theta_2)/\pi^2\right)\right]\right|^{1+\epsilon}\right)\right) \end{aligned}$$

Thus, to establish (25), it is enough to show that the second term in the above expression is finite. Note that, for $x \in [0, m]$,

$$A_m(x) = \Pr\left(\sum_{i=1}^m C_i \leq x\right) \geq \prod_{i=1}^m \Pr\left(C_i \leq \frac{x}{m}\right) = \left(\frac{x}{m}\right)^{m/2}$$

and thus

$$\begin{aligned} E_f\left(\left|\ln\left[A_m\left(d_1^2(\theta,\Theta_2)/\pi^2\right)\right]\right|^{1+\epsilon}\right) &\leq m^{1+\epsilon} E_f\left(\left|\ln[d_1(\theta,\Theta_2)/\pi\sqrt{m}]\right|^{1+\epsilon}\right) \\ &\leq m^{1+\epsilon}2^\epsilon\left(E_f\left(|\ln[d_1(\theta,\Theta_2)]|^{1+\epsilon}\right) + \left|\ln(\pi\sqrt{m})\right|^{1+\epsilon}\right) \\ &= m^{1+\epsilon}2^\epsilon\left(\int_{[0,2\pi)^m} |\ln[d_1(\theta,\mu)]|^{1+\epsilon} f(\mu)d\mu + \left|\ln(\pi\sqrt{m})\right|^{1+\epsilon}\right) \end{aligned}$$

In view of assumption (11), it follows that

$$\int_{[0,2\pi)^m} \mid \ln d_1(\theta,\mu) \mid^{1+\epsilon} f(\mu)d\mu < \infty$$

for almost all values of $\theta \in [0, 2\pi)^m$. Therefore, (25) is established.

Now we will establish (15). Consider

$$E_f\left(\mid S_{\theta,k,n} \mid^{1+\epsilon}\right) = \int_{-\infty}^{0} \mid u \mid^{1+\epsilon} dF_{\theta,k,n}(u) + \int_0^{\infty} \mid u \mid^{1+\epsilon} dF_{\theta,k,n}(u) \tag{26}$$

We can write

$$\begin{aligned} \int_0^{\infty} \mid u \mid^{1+\epsilon} dF_{\theta,k,n}(u) &= (1+\epsilon)\left(\int_0^{\ln\sqrt{n}} u^\epsilon\left(1 - F_{\theta,k,n}(u)\right)du + \int_{\ln\sqrt{n}}^{\infty} u^\epsilon\left(1 - F_{\theta,k,n}(u)\right)du\right) \\ &= (1+\epsilon)\left(I_1(n) + I_2(n)\right), \text{ say.} \end{aligned} \tag{27}$$

We have,

$$I_2(n) = \int_{\ln\sqrt{n}}^{\infty} u^\epsilon\left(\sum_{j=0}^{k-1}\binom{n-1}{j}\left(P_f\left(N_{\rho_{k,n}(u)}(\theta)\right)\right)^j\left(1 - P_f\left(N_{\rho_{k,n}(u)}(\theta)\right)\right)^{n-1-j}\right)du$$

For $j \in \{0, 1, \ldots, k-1\}$, we have $\binom{n-1}{j} \leq \frac{n-1}{n-k}\binom{n-2}{j}$ Therefore,

$$\begin{aligned} I_2(n) &\leq \frac{n-1}{n-k}\int_{\ln\sqrt{n}}^{\infty} u^{\epsilon}\left(1 - P_f\left(N_{\rho_{k,n}(u)}(\theta)\right)\right) \\ &\quad \times \left(\sum_{j=0}^{k-1}\binom{n-2}{j}\left(P_f\left(N_{\rho_{k,n}(u)}(\theta)\right)\right)^j\left(1 - P_f\left(N_{\rho_{k,n}(u)}(\theta)\right)\right)^{n-2-j}\right) du \\ &= \frac{n-1}{n-k}\int_{\ln\sqrt{n}}^{\infty} u^{\epsilon}\left(1 - P_f\left(N_{\rho_{k,n}(u)}(\theta)\right)\right) P\left(B_{k,n-k-1} \geq P_f\left(N_{\rho_{k,n}(u)}(\theta)\right)\right) du \end{aligned} \quad (28)$$

where $B_{a,b}$ denotes the beta random variable with parameter (a, b), $a > 0, b > 0$. For $u > \ln\sqrt{n}$, we have $P_f\left(N_{\rho_{k,n}(u)}(\theta)\right) \geq P_f\left(N_{\rho_{k,n}(\ln\sqrt{n})}(\theta)\right)$ Therefore, (28) yields

$$I_2(n) \leq \frac{n-1}{n-k} P\left(B_{k,n-k-1} \geq P_f\left(N_{\rho_{k,n}(\ln\sqrt{n})}(\theta)\right)\right)\int_{\ln\sqrt{n}}^{\infty} u^{\epsilon}\left(1 - P_f\left(N_{\rho_{k,n}(u)}(\theta)\right)\right) du \quad (29)$$

Note that $\lim_{n\to\infty}\rho_{k,n}(\ln\sqrt{n}) = 0$. Thus, on using (7), we get

$$\lim_{n\to\infty}\left(\frac{\sqrt{n}}{k} P_f\left(N_{\rho_{k,n}(\ln\sqrt{n})}(\theta)\right)\right) = f(\theta)$$

for almost all values of $\theta \in [0, 2\pi)^m$.

For a $\theta \in [0, 2\pi)^m$, for which $f(\theta) > 0$, choose $\delta \in (0, f(\theta))$. Then, for sufficiently large values of n,

$$P_f\left(N_{\rho_{k,n}(\ln\sqrt{n})}(\theta)\right) > \frac{k}{\sqrt{n}}(f(\theta) - \delta)$$

and therefore, for sufficiently large values of n,

$$\begin{aligned} &P\left(B_{k,n-k-1} \geq P_f\left(N_{\rho_{k,n}(\ln\sqrt{n})}(\theta)\right)\right) \leq P\left(B_{k,n-k-1} \geq \frac{k}{\sqrt{n}}(f(\theta) - \delta)\right) \leq \frac{E\left(B_{k,n-k-1}^2\right)}{\left(\frac{k}{\sqrt{n}}(f(\theta) - \delta)\right)^2} \\ &= \frac{k+1}{(n-1)k(f(\theta) - \delta)^2} \end{aligned}$$

Therefore, for sufficiently large values of n and for almost all values of $\theta \in [0, 2\pi)^m$, (29) yields

$$I_2(n) \leq \frac{k+1}{k(n-k)(f(\theta) - \delta)^2}\int_{\ln\sqrt{n}}^{\infty} u^{\epsilon}\left(1 - P_f\left(N_{\rho_{k,n}(u)}(\theta)\right)\right) du \quad (30)$$

On making the change of variable $z = \ln(2k/n) + u$ in the integral in (30), we get

$$\begin{aligned} &\int_{\ln\sqrt{n}}^{\infty} u^{\epsilon}\left(1 - P_f\left(N_{\rho_{k,n}(u)}(\theta)\right)\right) du = \int_{\ln\frac{2k}{\sqrt{n}}}^{\infty}\left(u + \ln\frac{n}{2k}\right)^{\epsilon}\left(1 - P_f\left(N_{\rho_{k,n}(u+\ln(\frac{n}{2k}))}(\theta)\right)\right) du \\ &= \int_{\ln\frac{2k}{\sqrt{n}}}^{\infty}\left(u + \ln\frac{n}{2k}\right)^{\epsilon}\left(1 - P_f\left(N_{\rho_{1,2}(u)}(\theta)\right)\right) du = \int_{\ln\frac{2k}{\sqrt{n}}}^{0}\left(u + \ln\frac{n}{2k}\right)^{\epsilon}\left(1 - P_f\left(N_{\rho_{1,2}(u)}(\theta)\right)\right) du \\ &\quad + \int_0^{\infty}\left(u + \ln\frac{n}{2k}\right)^{\epsilon}\left(1 - P_f\left(N_{\rho_{1,2}(u)}(\theta)\right)\right) du \end{aligned} \quad (31)$$

We also have

$$\int_{\ln\frac{2k}{\sqrt{n}}}^{0}\left(u+\ln\frac{n}{2k}\right)^{\epsilon}\left(1-P_f\left(N_{\rho_{1,2}(u)}(\theta)\right)\right)du \le \left(\ln\frac{n}{2k}\right)^{\epsilon}\int_{\ln\frac{2k}{\sqrt{n}}}^{0}\left(1-P_f\left(N_{\rho_{1,2}(u)}(\theta)\right)\right)du$$
$$=\left(\ln\frac{n}{2k}\right)^{\epsilon}\int_{\frac{2k}{\sqrt{n}}}^{1}\frac{1}{u}\left(1-P_f\left(N_{\rho_{1,2}(\ln u)}(\theta)\right)\right)du \le \frac{\sqrt{n}}{2k}\left(\ln\frac{n}{2k}\right)^{\epsilon}\int_{\frac{2k}{\sqrt{n}}}^{1}\left(1-P_f\left(N_{\rho_{1,2}(\ln u)}(\theta)\right)\right)du$$
$$\le \frac{\sqrt{n}}{2k}\left(\ln\frac{n}{2k}\right)^{\epsilon} \tag{32}$$

and

$$\int_{0}^{\infty}\left(u+\ln\frac{n}{2k}\right)^{\epsilon}\left(1-P_f\left(N_{\rho_{1,2}(u)}(\theta)\right)\right)du \;\le\; C_\epsilon\left(\left(\ln\frac{n}{2k}\right)^{\epsilon}\int_{0}^{\infty}\left(1-P_f\left(N_{\rho_{1,2}(u)}(\theta)\right)\right)du + \int_{0}^{\infty}u^{\epsilon}\left(1-P_f\left(N_{\rho_{1,2}(u)}(\theta)\right)\right)du\right) \tag{33}$$

where $C_\epsilon = \max(1, 2^{\epsilon-1})$.

Note that, for $u \in (-\infty, \infty)$,

$$1-P_f\left(N_{\rho_{1,2}(u)}(\theta)\right) = 1-P_f\left(S_{\theta,1,2}\le u\right) = 1-F_{\theta,1,2}(u)$$

Therefore (33) yields

$$\int_{0}^{\infty}\left(u+\ln\frac{n}{2k}\right)^{\epsilon}\left(1-P_f\left(N_{\rho_{1,2}(u)}(\theta)\right)\right)du \;\le\; C_\epsilon\left(\left(\ln\frac{n}{2k}\right)^{\epsilon}\int_{0}^{\infty}\left(1-F_{\theta,1,2}(u)\right)du + \int_{0}^{\infty}u^{\epsilon}\left(1-F_{\theta,1,2}(u)\right)du\right) \tag{34}$$

In view of (25), we have

$$\int_{0}^{\infty}\left(1-F_{\theta,1,2}(u)\right)du<\infty \;\text{ and }\; \int_{0}^{\infty}u^{\epsilon}\left(1-F_{\theta,1,2}(u)\right)du<\infty \tag{35}$$

Therefore, using (30)-(35), we conclude that

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

for almost all values of θ.

Now consider

$$I_1(n) = \int_{0}^{\ln\sqrt{n}}u^{\epsilon}\left(1-F_{\theta,k,n}(u)\right)du \tag{37}$$
$$= \int_{0}^{\ln\sqrt{n}}u^{\epsilon}\left(\sum_{j=0}^{k-1}\binom{n-1}{j}\left(P_f\left(N_{\rho_{k,n}(u)}(\theta)\right)\right)^{j}\left(1-P_f\left(N_{\rho_{k,n}(u)}(\theta)\right)\right)^{n-1-j}\right)du$$
$$= \int_{0}^{\ln\sqrt{n}}u^{\epsilon}P\left(B_{k,n-k}\ge P_f\left(N_{\rho_{k,n}(u)}(\theta)\right)\right)du \tag{38}$$

For $u<\ln\sqrt{n}$, we have

$$0\le\rho_{k,n}(u)\le\pi\sqrt{A_m^{-1}\left(\frac{k}{\sqrt{n}(2\pi)^m}\right)}\longrightarrow 0, \;\text{ as }\; n\to\infty$$

Therefore, for $u \in (-\infty, \infty)$ and for almost all values of $\theta \in [0, 2\pi)^m$, we have

$$\lim_{n\to\infty} \left(\frac{n}{ke^u} P_f \left(N_{\rho_{k,n}(u)}(\theta) \right) \right) = f(\theta)$$

uniformly in u.

For $f(\theta) > 0$, let $\delta \in (0, f(\theta))$. Then, for sufficiently large n, $u < \ln \sqrt{n}$ and for almost all values of $\theta \in [0, 2\pi)^m$, we have

$$P_f \left(N_{\rho_{k,n}(u)}(\theta) \right) > \frac{k}{n} \left(f(\theta) - \delta \right) e^u$$

and therefore

$$\begin{aligned} P\left(B_{k,n-k} \geq P_f \left(N_{\rho_{k,n}(u)}(\theta) \right) \right) &\leq P\left(B_{k,n-k} \geq \frac{k}{n}(f(\theta) - \delta)e^u \right) \\ &\leq \frac{E\left(B_{k,n-k} \right)}{\frac{k}{n}(f(\theta) - \delta)e^u} = \frac{e^{-u}}{f(\theta) - \delta} \end{aligned} \tag{39}$$

Using (39) in (37) we conclude that, for sufficiently large values of n, and for almost all values of θ

$$I_1(n) \leq \frac{1}{f(\theta) - \delta} \int_0^{\ln \sqrt{n}} u^\epsilon e^{-u} du \leq \frac{1}{f(\theta) - \delta} \int_0^\infty u^\epsilon e^{-u} du \ < \ \infty \tag{40}$$

Using (36) and (40) in (27), we conclude further that there exists a constant D_1 such that, for sufficiently large values of n,

$$\int_0^\infty \mid u \mid^{1+\epsilon} dF_{\theta,k,n}(u) du < D_1 \tag{41}$$

for almost all values of $\theta \in [0, 2\pi)^m$.

Now consider

$$\int_{-\infty}^0 \mid u \mid^{1+\epsilon} dF_{\theta,k,n}(u) = \int_{-\infty}^0 (-u)^{1+\epsilon} dF_{\theta,k,n}(u) = (1+\epsilon) \int_{-\infty}^0 (-u)^\epsilon F_{\theta,k,n}(u) du \tag{42}$$

Note that, for each $u \in (-\infty, 0)$ and for almost all values of $\theta \in [0, 2\pi)^m$,

$$\lim_{n\to\infty} \left(\frac{n}{ke^u} P_f \left(N_{\rho_{k,n}(u)}(\theta) \right) \right) = f(\theta)$$

uniformly in $u < 0$, *i.e.*, for almost all values of $\theta \in [0, 2\pi)^m$, $u \in (-\infty, 0)$ and every $\delta > 0$

$$P_f \left(N_{\rho_{k,n}(u)}(\theta) \right) < \frac{k}{n}(f(\theta) + \delta)e^u$$

for sufficiently large values of n. Therefore,

$$\begin{aligned} F_{\theta,k,n}(u) &= P\left(B_{k,n-k} \leq P_f \left(N_{\rho_{k,n}(u)}(\theta) \right) \right) \\ &\leq P\left(B_{k,n-k} \leq \frac{k}{n}(f(\theta) + \delta)e^u \right) \\ &= \sum_{j=k}^{n-1} \binom{n-1}{j} \left(\frac{k}{n} \left(f(\theta) + \delta \right) e^u \right)^j \left(1 - \frac{k}{n} \left(f(\theta) + \delta \right) e^u \right)^{n-1-j} \\ &\leq \frac{n-1}{n} \left(f(\theta) + \delta \right) e^u \leq \left(f(\theta) + \delta \right) e^u \end{aligned} \tag{43}$$

Using (43) in (42), we conclude that, for sufficiently large values of n and for almost all values of $\theta \in [0, 2\pi)^m$,

$$\begin{aligned} \int_{-\infty}^{0} \mid u \mid^{1+\epsilon} dF_{\theta,k,n}(u) &= (1+\epsilon) \int_{-\infty}^{0} (-u)^{\epsilon} F_{\theta,k,n}(u) du \\ &\leq (1+\epsilon)\,(f(\theta)+\delta) \int_{-\infty}^{0} (-u)^{\epsilon} e^{u} du \; < \; \infty \end{aligned} \tag{44}$$

Finally, on using (41) and (44) in (26), we conclude (15).

References

1. Karplus, M.; Kushick, J.N. Method for estimating the configurational entropy of macromolecules. *Macromolecules* **1981**, *14*, 325–332.
2. Misra, N.; Singh, H.; Demchuk, E. Estimation of the entropy of a multivariate normal distribution. *J. Multivar. Anal.* **2005**, *92*, 324–342.
3. Demchuk, E.; Singh, H. Statistical thermodynamics of hindered rotation from computer simulations. *Mol. Phys.* **2001**, *99*, 627–636.
4. Singh, H.; Hnizdo, V.; Demchuk, E. Probabilistic modeling of two dependent circular variables. *Biometrika* **2002**, *89*, 719–723.
5. Mardia, K.V.; Hughes, G.; Taylor, C.C.; Singh, H. A multivariate von Mises distribution with applications to bioinformatics. *Can. J. Stat.* **2008**, *36*, 99–109.
6. Hnizdo, V.; Fedorowicz, A.; Singh, H.; Demchuk, E. Statistical thermodynamics of internal rotation in a hindering potential of mean force obtained from computer simulations. *J. Comput. Chem.* **2003**, *24*, 1172–1183.
7. Darian, E.; Hnizdo, V.; Fedorowicz, A.; Singh, H.; Demchuk, E. Estimation of the absolute internal-rotation entropy of molecules with two torsional degrees of freedom from stochastic simulations. *J. Comput. Chem.* **2005**, *26*, 651–660.
8. Beirlant, J.; Dudewicz, E.J.; Gyorfi, L.; van der Meulen, E.C. Nonparametric estimation of entropy: An overview. *Internat. J. Math. Stat.* **1997**, *6*, 17–39.
9. Scott, D. *Multivariate Density Estimation: Theory, Practice and Visualization*; Wiley: New York, NY, USA, 1992.
10. Vasicek, O. On a test for normality based on sample entropy. *J. R. Stat. Soc.* Series B **1976**, *38*, 54–59.
11. Dudewicz, E.J.; van der Meulen, E.C. Entropy-based tests of uniformity. *J. Am. Stat. Assoc.* **1981**, *76*, 967–974.
12. Singh, H.; Misra, N.; Hnizdo, V.; Fedorowicz, E.; Demchuk, E. Nearest neighbor estimates of entropy. *Am. J. Math. Manag. Sci.* **2003**, *23*, 301–321.
13. Kozachenko, L.F.; Leonenko, N.N. Sample estimates of entropy of a random vector. *Prob. Inf. Trans.* **1987**, *23*, 95–101.
14. Goria, M.N.; Leonenko, N.N.; Novi Inveradi, P.L. A new class of random vector entropy estimators and its applications. *Nonparam. Stat.* **2005**, *17*, 277–297.
15. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138-1–066138-16.

16. Tsybakov, A.B.; van der Meulen, E.C. Root-n consistent estimators of entropy for densities with unbounded support. *Scan. J. Stat.* **1996**, *23*, 75–83.
17. Loftsgaarden, D.O.; Quesenberry, C.P. A non-parametric estimate of a multivariate density function. *Ann. Math. Stat.* **1965**, *36*, 1049–1051.
18. Mnatsakanov, R.M.; Misra, N.; Li, Sh.; Harner, E.J. k_n-Nearest neighbor estimators of entropy. *Math. Meth. Stat.* **2008**, *17*, 261–277.
19. Lebesgue, H. Sur l'intégration des fonctions discontinues. *Ann. Ecole Norm.* **1910**, *27*, 361–450.
20. Hnizdo, V.; Tan, J.; Killian, B.J.; Gilson, M.K. Efficient calculation of configurational entropy from molecular simulations by combining the mutual-information expansion and nearest-neighbor methods. *J. Comput. Chem.* **2008**, *29*, 1605–1614.
21. Hnizdo, V.; Darian, E.; Fedorowicz, A.; Demchuk, E.; Li, S.; Singh, H. Nearest-neighbor nonparametric method for estimating the configurational entropy of complex molecules. *J. Comput. Chem.* **2006**, *28*, 655–668.
22. Arya, S.; Mount, D.M. Approximate nearest neighbor searching. In the Proceedings of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, 25–27 January 1993; p. 271; Available online: http:// www.cs.umd.edu/∼mount/ANN/ (accessed on 5 May 2010).
23. Friedman, J.H.; Bentley, J.L.; Finkel, R.A. An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Software* **1977**, *3*, 209–226.