AMERICAN JOURNAL OF MATHEMATICAL AND MANAGEMENT SCIENCES

NEAREST NEIGHBOR ESTIMATES OF ENTROPY

Harshinder Singh[1,2], Neeraj Misra[1,3], Vladimir Hnizdo[2], Adam Fedorowicz[2], Eugene Demchuk[2]

[1] Department of Statistics, West Virginia University, Morgantown, WV 26506-6330, USA
[2] Health Effects Laboratory Division, National Institute for Occupational Safety and Health, Morgantown, 26505-2888, USA
[3] Department of Mathematics, Indian Institute of Technology Kanpur, Kanpur 208 016, India

SYNOPTIC ABSTRACT

Motivated by the problems in molecular sciences, we introduce new non-parametric estimators of entropy which are based on the k^{th} nearest neighbor distances between the n sample points, where k ($\leq n-1$) is a fixed positive integer. These provide competing estimators to an estimator proposed by Kozachenko and Leonenko (1987), which is based on the first nearest neighbor distances of the sample points. These estimators are helpful in the evaluation of entropies of random vectors. We establish the asymptotic unbiasedness and consistency of the proposed estimators. For some standard distributions, we also investigate their performance for finite sample sizes using Monte Carlo simulations. The proposed estimators are applied to estimate the entropy of internal rotation in the methanol molecule, which can be characterized by a one-dimensional random vector, and of diethyl ether, which is described by a four-dimensional random vector.

Key Words and Phrases: Bias, diethyl ether, enthalpy, entropy, free energy, internal rotation, k^{th} nearest neighbor, methanol, molecular dynamics simulations, root mean squared error, torsional angles.

2003, VOL. 23, NOS. 3 & 4, 301-321
0196-6324/03/030301-21 $26.00

1. INTRODUCTION

There are random fluctuations in the atomic coordinates of a molecule. The extent of these fluctuations determines the thermodynamic functions and shapes of molecules. The evaluation of thermodynamic functions, including entropy, is an important problem in molecular biology, chemistry and molecular physics (Landau and Lifshitz, 1980). The entropy of a molecule depends on fluctuations in its internal coordinates; i.e., bond lengths, valence angles and torsional angles. Since variations in bond lengths and valence angles are relatively small, the internal entropy is mainly determined by the random fluctuations of torsional angles.

Researchers have developed probabilistic models of the torsional angles of molecules for the evaluation of entropy. Karplus and Kushik (1981) and Levy, Karplus, Kushik, and Perahia (1984) proposed modeling the p torsional angles of a macromolecule by a multivariate normal distribution. The entropy of the multivariate normal distribution, having the probability density function $g(.)$ and the variance-covariance matrix Σ, is given by

$$E\left[-\ln g(X)\right] = \frac{p}{2} + \frac{1}{2}\ln\left[(2\pi)^p \mid \Sigma \mid\right].$$

An entropy estimator can be calculated by using the logarithm of the determinant of the estimated variance-covariance matrix S of independent observations of the torsional angles. These observations could be obtained from molecular dynamics simulations.

The assumption of a multivariate normal distribution of torsional angles limits the applicability of this approximation to molecules with small fluctuations around rotatable bonds. In general, the fluctuations in molecules are often large, and moreover, multiple modes are commonly observed in the marginal distributions of torsional angles. Taking this into account, Demchuk and Singh (2001) proposed a circular probabilistic approach to the modeling of torsional angels in molecules. Assuming a trimodal von Mises distribution for the torsional angle of the methanol molecule, they established a bathtub-shaped distribution for the torsional potential energy of the molecule. The trimodal von Mises distribution provided an excellent fit to the histogram of observations of the torsional angle.

Singh, Hnizdo, and Demchuk (2002) proposed a bivariate circular probability model on the torus, which is a natural torus version of the bivariate normal distribution. This distribution has a natural extension to more than two variables. However, the marginal distributions of the proposed bivariate distribution are symmetric, and thus this model is not applicable when distributions of torsional angles are skewed. Such skewness of angular variables is

not uncommon in large molecules. These complexities suggest that the evaluation of molecular entropy may substantially benefit from the flexibility offered by a non-parametric approach.

Let $X_1, X_2, \ldots, X_n$ be n independent copies of a p-dimensional random variable X, having an absolutely continuous distribution function $F(x)$ and probability density function (pdf) $f(x)$. The entropy of the pdf $f(x)$ (or the random variable X) is defined by

$$H(f) = E\left[-\ln f(X)\right] = -\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} I_S(x) f(x)\ln f(x)dx, \qquad (1)$$

where $dx = dx_1 \cdots dx_p$, S is the region where the pdf $f(.)$ is positive and $I_S(.)$ is the indicator function of the set S.

Several non-parametric estimators of entropy are discussed in the literature. Some of them were obtained by replacing the probability density $f(x)$ in the definition of entropy by its non-parametric kernel or histogram estimator $f_n(x)$ (Beirlant, Dudewicz, Györfi, and Van der Meulen, 1997, Scott, 1992). Another one is based on spacings (Dudewicz and van der Meulen, 1981, Vasicek, 1976), but it is applicable in only one dimension ($p = 1$). For $p = 1$, the estimator based on spacings is given by

$$\hat{V}_m^{(n)} = \frac{1}{n-m}\sum_{i=1}^{n-m}\ln\left(\frac{n}{m}(X_{i+m:n} - X_{i:n})\right) - \Psi(m) + \ln m, \qquad (2)$$

where m ($\leq n-1$) is a fixed positive integer, $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ are the order statistics of $X_1, X_2, \ldots, X_n$ and $\Psi(m) = \frac{\Gamma'(m)}{\Gamma(m)}$ is the digamma function. Note that the estimator $\hat{V}_m^{(n)}$ is based on m-spacings $X_{i+m:n} - X_{i:n}, i = 1, \ldots, n-m$. We call $\hat{V}_m^{(n)}$ as the m-spacing estimator. Under suitable conditions, Vasicek (1976) proved that $\hat{V}_m^{(n)}$ is a consistent estimator of the entropy $H(f)$, defined by (1). For $p = 1$ and under appropriate conditions, the m-spacing estimator $\hat{V}_m^{(n)}$ is also known to be asymptotically normal (Cressie, 1976, Dudewicz and van der Meulen, 1981), i.e.,

$$\sqrt{n}\left(\hat{V}_m^{(n)} - H(f)\right) \xrightarrow{d}$$

$$N\left(0, (2m^2 - 2m + 1)\Psi'(m) - 2m + 1 + \mathrm{Var}(\ln f(X))\right), \qquad (3)$$

as $n \to \infty$.

For the general p, Kozachenko and Leonenko (1987) proposed a non-parametric estimate of entropy based on the nearest neighbor distances between the sample points; the idea is traced back to Dobrushin (1958). Let $S_{r,z}$ be the sphere

of radius $r > 0$, centered at $z \in R^p$, the p-dimensional Euclidean space. The volume of $S_{r;z}$ is given by

$$V_r = \frac{\pi^{p/2} r^p}{\Gamma(p/2+1)}.$$

Let $\rho_i = \min\{||X_i - X_j||, j \in \{1,2,\ldots,n\} - \{i\}\}$ and let $\gamma = 0.5772\cdots$ denote Euler's constant. Based on the first nearest neighbor distances, the ρ_i values, Kozachenko and Leonenko (1987) proposed the following estimator H_n of $H(f)$, given by (1),

$$H_n = \frac{p}{n}\sum_{i=1}^{n} \ln \rho_i + \ln\left[\frac{\pi^{p/2}}{\Gamma(\frac{p}{2}+1)}\right] + \gamma + \ln(n-1). \tag{4}$$

They proved the asymptotic unbiasedness and the consistency of the estimator H_n. For $p = 1$, Tsybakov and van der Meulen (1996) established the mean square root n consistency of a truncated version of H_n.

Since the random variable X is absolutely continuous, the nearest neighbor distances (the ρ_i values) are expected to be small positive numbers. Due to the presence of n ln factors in the expression of H_n, given by (4), small fluctuations in the small ρ_i values will result in relatively higher fluctuations in the values of H_n. Therefore, a valid objection to using estimator H_n is that it can be used in practice only if the small ρ_i values are recorded to high accuracy, which is often not the case. This difficulty can be avoided if one constructs an estimator based on higher (second or higher) nearest neighbor distances. In this paper, we extend the idea of Kozachenko and Leonenko (1987) by constructing estimates of entropy, which are based on the second and higher nearest neighbor distances.

In Section 2 of this paper, we define an estimator of the entropy based on the k^{th} nearest neighbor distances, where k ($\leq n-1$) is a fixed positive integer. We call this estimator as the k-nearest neighbor estimator. We prove the asymptotic unbiasedness and consistency of k-nearest neighbor estimators. For $p = 1$, we compare the orders of the asymptotic variances of the k-nearest neighbor estimators and the m-spacing estimators of Vasicek (1976), as defined by (2). For $p = 1$ and for some standard distributions, using Monte Carlo simulations, small sample comparisons of the performances of the k-nearest neighbor estimators and the m-spacing estimators are made under the bias and root mean squared error criteria. For finite sample sizes and for some standard distributions, using Monte Carlo simulations, we also study the performances of the k-nearest neighbor estimators in terms of their biases and root mean squared errors. We observe that, under the criterion of root mean squared error, the estimators based on the fourth ($k = 4$) order nearest neighbor distances perform reasonably well.

In section 3, we give an example to illustrate the computation of nearest neighbor estimators of entropy. In section 4, we use these estimators to estimate entropy of the methanol molecule, having one torsional angle, and of diethyl ether, which has four torsional angles.

2. k^{th} NEAREST NEIGHBOR ESTIMATOR OF ENTROPY

Based on a random sample $X_1, X_2, \ldots, X_n$ from the distribution having pdf $f(x)$, our aim is to estimate the entropy $H(f)$, given by (1).

A reasonable estimator of entropy $H(f)$ is of the form

$$\hat{H}(f) = -\frac{1}{n}\sum_{i=1}^{n} \ln\left[\hat{f}(X_i)\right],$$

where $\hat{f}(.)$ is a suitable estimator of the pdf $f(.)$.

Let $1 \leq k \leq n$ be a given positive integer, and, for $i = 1, \ldots, n$, let

$$R_{i,k,n} = \text{Euclidean distance from } X_i \text{ to its } k^{th} \text{ closest neighbor.} \tag{5}$$

Then, a reasonable estimate (say $\hat{f}(X_i)$) of $f(X_i)$ is given by

$$\hat{f}(X_i)\frac{\pi^{p/2} R_{i,k,n}^p}{\Gamma(p/2+1)} = \frac{k}{n}.$$

Note that $\frac{R_{i,k,n}^p \pi^{p/2}}{\Gamma(p/2+1)}$ is the volume of sphere having radius $R_{i,k,n}$, defined by (5). The above equation gives

$$\hat{f}(X_i) = \frac{k\Gamma(p/2+1)}{n\pi^{p/2} R_{i,k,n}^p}, \; i = 1, 2, \ldots, n,$$

and, therefore, a reasonable estimate of $H(f)$ is given by

$$\hat{G}_k^{(n)}(f) = -\frac{1}{n}\sum_{i=1}^{n} \ln\left[\hat{f}(X_i)\right] = \frac{1}{n}\sum_{i=1}^{n} T_i^{(n)}, \tag{6}$$

where

$$T_i^{(n)} = \ln\left[\frac{n\pi^{p/2} R_{i,k,n}^p}{k\Gamma(p/2+1)}\right], \; i = 1, 2, \ldots, n. \tag{7}$$

The following theorem gives the asymptotic mean of the estimator $\hat{G}_k^{(n)}(f)$.

Theorem 8. *The asymptotic mean of the estimator $\hat{G}_k^{(n)}(f)$ is given by*

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

where $L_0 = 0$, $L_j = \sum_{i=1}^{j} \frac{1}{i}$, $j \geq 1$ and $\gamma = 0.5772\cdots$ is Euler's constant.

Proof. We note that the random variables $T_1^{(n)}, T_2^{(n)}, \ldots, T_n^{(n)}$, as given by (7), are identically distributed and thus

$$E\left[\hat{G}_k^{(n)}(f)\right] = E\left[T_1^{(n)}\right].$$

For a real number r, we have

$$P\left[T_1^{(n)} > r | X_1 = x\right] = P\left[R_{1,k,n} > \rho_{r,n} | X_1 = x\right],$$

where,

$$\rho_{r,n} = \left[\frac{k\Gamma(p/2+1)e^r}{n\pi^{p/2}}\right]^{\frac{1}{p}}.$$

Thus,

$$P\left[T_1^{(n)} > r | X_1 = x\right] = \sum_{i=0}^{k-1} \binom{n-1}{i} \left[P(S_{\rho_{r,n};x})\right]^i \left[1 - P(S_{\rho_{r,n};x})\right]^{n-1-i},$$

where

$$P\left(S_{\rho_{r,n};x}\right) = \int_{S_{\rho_{r,n},x}} f(t)dt.$$

Since $\rho_{r,n} \to 0$, as $n \to \infty$, we have

$$\lim_{n\to\infty} \left[nP\left(S_{\rho_{r,n};x}\right)\right] = ke^r \lim_{n\to\infty} \frac{P\left(S_{\rho_{r,n};x}\right)}{V_{\rho_{r,n}}} = ke^r f(x).$$

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

$$\begin{aligned}\lim_{n\to\infty} P\left[T_1^{(n)} > r | X_1 = x\right] &= \sum_{i=0}^{k-1} \frac{\left[kf(x)e^r\right]^i}{i!} e^{-kf(x)e^r} \\ &= P\left[T_x > r\right],\end{aligned}$$

where, for given x, the random variable T_x has the pdf

$$h_{T_x}(y) = \frac{[kf(x)e^y]^k}{(k-1)!} e^{-kf(x)e^y}, -\infty < y < \infty.$$

Thus,

$$\begin{aligned}\lim_{n\to\infty} E\left[T_1^{(n)} | X_1 = x\right] &= \int_{-\infty}^{\infty} y \frac{[kf(x)e^y]^k}{(k-1)!} e^{-kf(x)e^y} dy \\ &= \int_0^{\infty} [\ln z - \ln k - \ln f(x)] \frac{z^{k-1}}{(k-1)!} e^{-z} dz \\ &= \frac{1}{\Gamma(k)} \int_0^{\infty} \left[\ln(z) z^{k-1} e^{-z}\right] dz - \ln k - \ln f(x) \\ &= \Psi(k) - \ln k - \ln f(x), \end{aligned} \tag{9}$$

where $\Psi(k) = \frac{\Gamma'(k)}{\Gamma(k)}$ is the digamma function. From Abramowitz and Stegun (1965, pp. 258), we have $\Psi(k) = L_{k-1} - \gamma$. Therefore, (9) yields

$$\lim_{n\to\infty} E\left[T_1^{(n)}\right] = L_{k-1} - \gamma - \ln k + H(f).$$

Thus, the estimator $\hat{G}_k^{(n)}(f)$ is asymptotically biased. Therefore, we consider the modified estimator

$$\begin{aligned}\hat{H}_k^{(n)}(f) &= \hat{G}_k^{(n)}(f) - L_{k-1} + \gamma + \ln k \\ &= \frac{p}{n} \sum_{i=1}^{n} \ln R_{i,k,n} + \ln\left[\frac{\pi^{p/2}}{\Gamma(p/2+1)}\right] + \gamma - L_{k-1} + \ln n, \end{aligned} \tag{10}$$

which is asymptotically unbiased. We call the estimator $\hat{H}_k^{(n)}(f)$ as the k-nearest neighbor estimator. Note that, for $k = 1$, this estimator is almost the same as the estimator given by (4), which was proposed by Kozachenko and Leonenko (1987); the only difference in the two estimators is the use of $\ln(n-1)$ in place of $\ln n$ in the estimator given by (4).

The following theorem proves the consistency of the estimator $\hat{H}_k^{(n)}(f)$, given by (10).

Theorem 11. $\lim_{n\to\infty} Var\left[\hat{H}_k^{(n)}(f)\right] = 0.$

Proof. Since the distribution of the random vector $\left(T_1^{(n)}, T_2^{(n)}, \ldots, T_n^{(n)}\right)$ is same as any permutation of it, we have

$$\mathrm{Var}\left[\hat{G}_k^{(n)}(f)\right] = \frac{\mathrm{Var}\left[T_1^{(n)}\right]}{n} + \frac{n(n-1)}{n^2} \mathrm{Cov}\left(T_1^{(n)}, T_2^{(n)}\right). \tag{12}$$

For a positive integer k, define $Q_k = \sum_{j=k}^{\infty} \frac{1}{j^2}$. Then, for fixed x, we have

$$\begin{aligned}\lim_{n\to\infty} E\left[(T_1^{(n)})^2 | X_1 = x\right] &= \int_0^{\infty} [\ln z - \ln k - \ln f(x)]^2 \frac{e^{-z} z^{k-1}}{\Gamma(k)} dz \\ &= \frac{\Gamma''(k)}{\Gamma(k)} + (\ln k)^2 + [\ln f(x)]^2 - 2\ln(k)\frac{\Gamma'(k)}{\Gamma(k)} \\ &\quad - 2\ln\left(f(x)\right)\frac{\Gamma'(k)}{\Gamma(k)} + 2(\ln k)(\ln f(x)).\end{aligned}$$

On using the results of digamma and polygamma functions (Abramowitz and Stegun (1965, pp. 260)), we have $\Psi'(k) = (\Gamma'(k)/\Gamma(k))' = Q_k$ and therefore

$$\frac{\Gamma''(k)}{\Gamma(k)} = Q_k + [L_{k-1} - \gamma]^2.$$

Therefore,

$$\lim_{n\to\infty} E\left[(T_1^{(n)})^2\right] = Q_k + (L_{k-1}-\gamma)^2 + (\ln k)^2 + E\left[(\ln f(X))^2\right] - 2\left[L_{k-1}-\gamma\right]\ln(k) + 2\left[L_{k-1}-\gamma\right]H(f) - 2H(f)\ln(k) = Q_k + \mathrm{Var}\left[\ln f(X)\right] + \left[L_{k-1}-\gamma-\ln k + H(f)\right]^2,$$

implying that

$$\lim_{n\to\infty} \mathrm{Var}\left[T_1^{(n)}\right] = \lim_{n\to\infty} E\left[(T_1^{(n)})^2\right] - \lim_{n\to\infty}\left(E\left[T_1^{(n)}\right]\right)^2 = Q_k + \mathrm{Var}\left[\ln f(X)\right]. \quad (13)$$

For finding the limiting covariance between $T_1^{(n)}$ and $T_2^{(n)}$, we consider, for $-\infty < r < \infty$ and $-\infty < s < \infty$,

$$P\left[T_1^{(n)} > r, T_2^{(n)} > s | X_1 = x, X_2 = y\right] = P\left[R_{1,k,n} > \rho_{r,n}, R_{2,k,n} > \rho_{s,n} | X_1 = x, X_2 = y\right].$$

For $x \neq y$, since $\rho_{r,n}$ and $\rho_{s,n}$ tend to 0 as $n \to \infty$, we may assume that, for large n, $S_{\rho_{r,n};x} \cap S_{\rho_{s,n};y} = \phi$, the empty set. Thus, for large n,

$$\begin{aligned} &P\left[\, T_1^{(n)} > r, T_2^{(n)} > s | X_1 = x, X_2 = y\right] \\ &= P\left[\text{at most } (k-1) \text{ of } X_3, \ldots, X_n \in S_{\rho_{r,n};x} \right. \\ &\qquad \left. \text{and at most } (k-1) \text{ of } X_3, \ldots, X_n \in S_{\rho_{r,n};y}\right] \\ &= \sum_{0\le i,j\le k-1, i+j\le n-2} \frac{(n-2)!}{i!j!(n-2-i-j)!}\left[P\left(S_{\rho_{r,n};x}\right)\right]^i \left[P\left(S_{\rho_{r,n};y}\right)\right]^j \\ &\qquad \left[1 - P\left(S_{\rho_{r,n};x}\right) - P\left(S_{\rho_{r,n};y}\right)\right]^{n-2-i-j} \\ &= \sum_{0\le i,j\le k-1, i+j\le n-2} \frac{(n-2)!}{i!j!(n-2-i-j)!}\frac{k^i e^{ir}}{n^i}\frac{k^j e^{js}}{n^j}\left[\frac{P\left(S_{\rho_{r,n};x}\right)}{V_{\rho_{r,n}}}\right]^i \\ &\left[\frac{P\left(S_{\rho_{r,n};y}\right)}{V_{\rho_{s,n}}}\right]^j \left[1 - \frac{1}{n}\left\{ke^r\frac{P\left(S_{\rho_{r,n};x}\right)}{V_{\rho_{r,n}}} + ke^s\frac{P\left(S_{\rho_{r,n};y}\right)}{V_{\rho_{s,n}}}\right\}\right]^{n-2-i-j} \end{aligned}$$

Since,

$$\lim_{n\to\infty}\frac{(n-2)!}{(n-2-i-j)!n^i n^j} = 1,$$

we have,

$$\begin{aligned} &\lim_{n\to\infty} P\left[\, T_1^{(n)} > r, T_2^{(n)} > s | X_1 = x, X_2 = y\right] \\ &= \sum_{i=0}^{k-1}\sum_{j=0}^{k-1} \frac{\left[kf(x)e^r\right]^i}{i!}\frac{\left[kf(y)e^s\right]^j}{j!} e^{-\left[kf(x)e^r + kf(y)e^s\right]} \\ &= P\left[T_x > r\right] P\left[T_y > s\right], \end{aligned}$$

where, for given z, the random variable T_z has the pdf

$$h_{T_z}(y) = \frac{\left[kf(z)e^y\right]^k}{(k-1)!} e^{-kf(z)e^y}, \quad -\infty < y < \infty.$$

Therefore,

$$\lim_{n\to\infty} E\left[T_1^{(n)} T_2^{(n)}\right] = \lim_{n\to\infty}\left[E[T_1^{(n)}]E[T_2^{(n)}]\right],$$

implying that

$$\lim_{n\to\infty} \mathrm{Cov}\left[T_1^{(n)}, T_2^{(n)}\right] = 0. \quad (14)$$

Hence, from (12), (13) and (14), we have

$$\lim_{n\to\infty} \mathrm{Var}\left[\hat{G}_k^{(n)}(f)\right] = 0,$$

implying that

$$\lim_{n\to\infty} \mathrm{Var}\left[\hat{H}_k^{(n)}(f)\right] = 0.$$

From expressions (12), (13) and (14), it follows that the asymptotic variance of $\hat{H}_k^{(n)}(f)$ is of the order

$$\mathrm{Var}\left[\hat{H}_k^{(n)}(f)\right] = \mathrm{Var}\left[\hat{G}_k^{(n)}(f)\right] \approx \frac{Q_k + \mathrm{Var}\left[\ln f(X)\right]}{n} = E_k^{(n)}, \text{ say,} \quad (15)$$

where $Q_k = \sum_{j=k}^{\infty}\frac{1}{j^2}$. Also, since $\hat{H}_k^{(n)}(f)$ is asymptotically unbiased, the asymptotic mean squared error of $\hat{H}_k^{(n)}(f)$ is also of the order given by (15). For $p = 1$, it follows from (3) that the asymptotic variance of the Vasicek's (1976) m-spacing estimator is of the order

$$\mathrm{Var}\left[\hat{V}_m^{(n)}(f)\right] \approx \frac{(2m^2-2m+1)Q_m - 2m + 1 + \mathrm{Var}\left[\ln f(X)\right]}{n} = F_m^{(n)}, \text{ say.} \quad (16)$$

For $p = 1$, it will be interesting to compare the orders of the asymptotic variances of the m-spacing estimators and the k-nearest neighbor estimators. The following lemma will be useful in making such comparisons.

Lemma 17. *For all $t > 0$*
(a) $\frac{t}{1-e^{-t}} < 1 + \frac{t}{2} + \frac{t^2}{12}$.
(b) $\frac{t}{1-e^{-t}} > 1 + \frac{t}{2} + \frac{t^2}{12} - \frac{t^4}{720}$.

Proof. (a) Consider

$$g(t) = 1 - \frac{t}{2} + \frac{t^2}{12} - e^{-t}\left[1 + \frac{t}{2} + \frac{t^2}{12}\right], \; t \geq 0.$$

Then

$$\begin{aligned} g'(t) &= -\frac{1}{2} + \frac{t}{6} + e^{-t}\left[\frac{1}{2} + \frac{t}{3} + \frac{t^2}{12}\right], \\ g''(t) &= \frac{1}{6} - e^{-t}\left[\frac{1}{6} + \frac{t}{6} + \frac{t^2}{12}\right] \end{aligned}$$

and

$$g'''(t) = \frac{t^2}{12}e^{-t} > 0, \; \forall\, t > 0.$$

Thus,

$$\begin{aligned} g''(t) &> g''(0) = 0, \; \forall\, t > 0 \\ \Rightarrow g'(t) &> g'(0) = 0, \; \forall\, t > 0 \\ \Rightarrow g(t) &> g(0) = 0, \; \forall\, t > 0. \end{aligned}$$

This proves the assertion (a).
(b) Consider

$$m(t) = -1 + \frac{t}{2} - \frac{t^2}{12} + \frac{t^4}{720} + e^{-t}\left[1 + \frac{t}{2} + \frac{t^2}{12} - \frac{t^4}{720}\right], \; t \geq 0.$$

Then

$$\begin{aligned} m'(t) &= \frac{1}{2} - \frac{t}{6} + \frac{t^3}{180} - e^{-t}\left[\frac{1}{2} + \frac{t}{3} + \frac{t^2}{12} + \frac{t^3}{180} - \frac{t^4}{720}\right], \\ m''(t) &= -\frac{1}{6} + \frac{t^2}{60} + e^{-t}\left[\frac{1}{6} + \frac{t}{6} + \frac{t^2}{15} + \frac{t^3}{90} - \frac{t^4}{720}\right], \\ k(t) \equiv \frac{m'''(t)}{t} &= \frac{1}{30} - e^{-t}\left[\frac{1}{30} + \frac{t}{30} + \frac{t^2}{60} - \frac{t^3}{720}\right] \end{aligned}$$

and

$$k'(t) = \frac{t^2}{48}e^{-t}\left[1 - \frac{t}{15}\right], \; t \geq 0.$$

Thus $k(t)$ is strictly increasing if $t \in [0, 15)$ and it is strictly decreasing if $t \in (15, \infty)$. Therefore, for all $t > 0$,

$$\begin{aligned} k(t) &> \min\left[k(0), \lim_{t\to\infty} k(t)\right] \\ &= \min\left[0, \frac{1}{30}\right] = 0 \\ \Rightarrow m''(t) &> m''(0) = 0, \; \forall\, t > 0 \\ \Rightarrow m'(t) &> m'(0) = 0, \; \forall\, t > 0 \\ \Rightarrow m(t) &> m(0) = 0, \; \forall\, t > 0. \end{aligned}$$

Hence the result follows.

Theorem 18. *Let $E_k^{(n)}$ and $F_m^{(n)}$ be defined by (15) and (16), respectively. Then, for $p = 1$, $n = 1, 2, \ldots$, and $m = 1, 2, \ldots, n-1$,*
(a) $F_m^{(n)} < E_m^{(n)}$.
(b) $E_{3m}^{(n)} < F_m^{(n)}$.

Proof. Using the representation (6.4.1) of Abramowitz and Stegun (1965, pp. 260), we have, for $p = 1, 2, \ldots$,

$$Q_p = \Psi'(p) = \int_0^\infty \frac{t}{1-e^{-t}} e^{-pt} dt.$$

For fixed $n \in \{1, 2, \ldots\}$ and $m \in \{1, \ldots, n-1\}$, consider

$$\begin{aligned} &\Delta_1(m) \\ &= n\left[E_m^{(n)} - F_m^{(n)}\right] \\ &= Q_m - (2m^2 - 2m + 1)Q_m + 2m - 1 \\ &= 2m - 1 - 2m(m-1)Q_m \\ &= 2m - 1 - 2m(m-1)\int_0^\infty \frac{t}{1-e^{-t}} e^{-mt} dt \\ &> 2m - 1 - 2m(m-1)\int_0^\infty \left[1 + \frac{t}{2} + \frac{t^2}{12}\right] e^{-mt} dt \\ &= 2m - 1 - 2m(m-1)\frac{6m^2 + 3m + 1}{6m^3} \\ &= \frac{2m+1}{3m^2} > 0, \end{aligned}$$

where the inequality above follows using Lemma 17 (a). This proves the assertion (a).

(b) For fixed $n \in \{1,2,\ldots\}$ and $m \in \{1,\ldots,n-1\}$, consider

$$\begin{aligned}
\Delta_2(m) &= n\left[F_m^{(n)} - E_{3m}^{(n)}\right] \\
&= (2m^2-2m+1)Q_m - 2m + 1 - Q_{3m} \\
&= 2m(m-1)Q_m + Q_m - Q_{3m} - (2m-1) \\
&= 2m(m-1)\int_0^\infty \frac{t}{1-e^{-t}} e^{-mt} dt \\
&\quad + \int_0^\infty \frac{t}{1-e^{-t}} e^{-mt}(1-e^{-2mt})dt - (2m-1) \\
&> 2m(m-1)\int_0^\infty \left[1 + \frac{t}{2} + \frac{t^2}{12} - \frac{t^4}{720}\right] e^{-mt} dt \\
&\quad + \int_0^\infty \left[1 + \frac{t}{2} + \frac{t^2}{12} - \frac{t^4}{720}\right] e^{-mt}(1-e^{-2mt})dt - (2m-1) \\
&= (2m^2-2m+1)\frac{30m^4+15m^3+5m^2-1}{30m^5} \\
&\quad - \frac{2430m^4+405m^3+45m^2-1}{7290m^5} - (2m-1) \\
&= \frac{810m^3+684m^2+486m-242}{7290m^5} > 0,
\end{aligned}$$

where the inequality above follows using Lemma 17 (b). This proves the assertion (b).

The above theorem suggests that, for a suitably large n and for a fixed $m \leq n-1$, the m-spacing estimator $\hat{V}_m^{(n)}$ has smaller variance than the m-nearest neighbor estimator $\hat{H}_m^{(n)}$, whereas the $3m$-nearest neighbor estimator $\hat{H}_{3m}^{(n)}$ has smaller variance than the m-spacing estimator $\hat{V}_m^{(n)}$. The m-spacing estimators $\hat{V}_m^{(n)}$ have limited applicability as they can be used only in one dimension ($p=1$).

Clearly, the expression in (15) is a decreasing function of k. Thus, the asymptotics suggests that, under the criterion of mean squared error, larger values of k are better. However, for practical purposes (fixed and small sample sizes), the choice of k is an important issue. For practical applications a rule of thumb about the order of k may be obtained from Monte Carlo simulations.

For finite sample sizes ($n = 10, 25, 50$) and for some standard distributions, using Monte Carlo simulations, we studied the performances of estimators $\hat{H}_k^{(n)}(f)$s, for various k values, under the criterion of root mean squared error (RMSE). We also report the bias of these estimators. We considered the following distributions: (i) the uniform distribution on [0,1], (ii) Student's t distribution with $\nu = 4, 10$ degrees of freedom, (iii) the standard normal distribution having mean 0 and variance 1, (iv) standard (having scale parameter 1) gamma distributions with shape parameters $\alpha = 1, 3, 6$, and (v)

standard (having means 0 and variances 1) bivariate normal distributions with correlation coefficients $\rho = 0, 0.5, 0.9$. The value of p is 1 for the distributions mentioned in (i) - (iv), where as $p = 2$ for the distributions mentioned in (v). It is worth emphasizing here that, in any Monte Carlo study, use of a quality random number generator is very important. Many commonly used random number generators fail to pass the available extensive tests, called TESTRAND (Karian and Dudewicz, 1999). It has been observed that using a random number generator which fails to pass the TESTRAND tests may vitiate simulation study (Chen, McCoskey, and Kao, 1999). For our Monte Carlo study, we generated U(0, 1) random numbers using the URN13 random generator given in Karian and Dudewicz (1999). This random number generator is known to pass the TESTRAND tests (Karian and Dudewicz, 1999). Then we used suitable transformations to generate random numbers from the distributions mentioned in (i)-(v). Ten thousand random samples of size n, with varying n, were generated from each of these distributions. The values of estimator $\hat{H}_k^{(n)}(f)$ were calculated from each of these samples for $k = 1, \ldots, n-1$. These were used to estimate the bias and RMSE of these estimators for each distribution. Some of these values are listed in Table 1 in the columns 4-7. For each studied sample size and for all distributions under the study, our Monte Carlo simulations, which were conducted for each $k \in \{1, \ldots, n-1\}$, suggested that the estimator $\hat{H}_4^{(n)}(f)$ performs reasonably well. Figure 1 gives the plot of RMSE of $H_k^{(n)}$ as a function of k for $n = 10, 25$, and 50 for uniform distribution on $[0, 1]$. To save space, we are not reporting results of all studied values of k for all the distributions, rather results for only important values of k ($k = 1, 4$ and for the k (denoted by k_b) which yields the smallest RMSE/absolute bias) are reported here.

The following conclusions are evident from the Monte Carlo study:
(i) In many cases, estimates based on higher values of k seem preferable under the root mean squared error criterion. However, in most of such cases, gain in using higher k, in comparison with $k = 4$, is not substantial and these small gains are overshadowed by the complexities involved in the computation of n k^{th} nearest neighbor distances in the expression of $\hat{H}_k^{(n)}(f)$.
(ii) In most cases, the RMSE of the estimator $\hat{H}_4^{(n)}(f)$ is close to the RMSE of the estimator $\hat{H}_{k_b}^{(n)}(f)$; here k_b is the best choice of k under the criterion of RMSE. Moreover, in most cases, the estimator $\hat{H}_4^{(n)}(f)$ has significantly smaller RMSE in comparison with the RMSE of $\hat{H}_1^{(n)}(f)$.
(iii) In view of the above observations, the estimator $\hat{H}_4^{(n)}(f)$ seems to be a reasonable choice.

Similar Monte Carlo study was carried out to study the performances of the m-spacing estimators $\hat{V}_m^{(n)}(f)$ and to compare these with the nearest neighbor estimators for finite sample sizes. The results for $m = 1, 4$ and for the m

(denoted by m_b) which yields the smallest RMSE/absolute bias are reported in the last four columns of the Table 1. We observe that the spacing estimator with $m = 1$ has consistently smaller root mean squared error than the nearest neighbor estimator with $k = 1$. Figure 2 gives the plot of RMSE of $V_m^{(n)}$, as a function of m, for $n = 10, 25$, and 50 for the uniform distribution on $[0, 1]$.

Remark 19. *We note that because of the possibilities of ties in bootstrap samples and the presence of* $\ln(R_{i,k,n}^p)$ *factors in the expression of nearest neighbor estimates, the application of bootstrap methods to the nearest neighbor estimates can be problematic. However, chances of such a problem occurring are less if one considers estimates based on higher order nearest neighbor distances.*

3. EXAMPLE

We illustrate the computation of nearest neighbor estimates $\hat{H}_k^{(n)}$, for $k = 1, 2$, based on a random sample of five observations taken from a continuous distribution on $[0, 2\pi]$ with unknown entropy. Since the distribution is one dimensional, we have $p = 1$. The random observations are 1.99, 2.31, 4.76, 2.64 and 4.65. We note that the nearest observation to the first observation is the second observation and thus the first nearest neighbor distance for the first observation is given by $R_{1,1,5} = |1.99 - 2.31| = 0.32$. Similarly, we have $R_{2,1,5} = |2.31 - 1.99| = 0.32$, $R_{3,1,5} = |4.76 - 4.65| = 0.11$, $R_{4,1,5} = |2.64 - 2.31| = 0.33$, $R_{5,1,5} = |4.65 - 4.76| = 0.11$. Using (10), an estimate of entropy of the population distribution based on the first nearest neighbor distances is given by

$$\hat{H}_1^{(5)}(f) = \frac{1}{5}\sum_{i=1}^{5} \ln R_{i,1,5} + \ln\left(\frac{\pi^{1/2}}{\Gamma(3/2)}\right) + 0.5772 - 0 + \ln 5 = 1.32.$$

The second nearest neighbor of the first observation is the fourth observation and thus the second nearest neighbor distance for the first observation is given by $R_{1,2,5} = |1.99 - 2.64| = 0.65$. Similarly, we have $R_{2,2,5} = |2.31 - 2.64| = 0.33$, $R_{3,2,5} = |4.76 - 2.64| = 2.12$, $R_{4,2,5} = |2.64 - 1.99| = 0.65$, $R_{5,2,5} = |4.65 - 2.64| = 2.01$. Again, using (10), an estimate of entropy based on the second nearest neighbor distances is given by

$$\hat{H}_2^{(5)}(f) = \frac{1}{5}\sum_{i=1}^{5} \ln R_{i,2,5} + \ln\left(\frac{\pi^{1/2}}{\Gamma(3/2)}\right) + 0.5772 - 1 + \ln 5 = 1.77.$$

Table 1: Biases and RMSEs of $\hat{H}_k^{(n)}(f)$ ($\hat{V}_m^{(n)}(f)$) for Five (Four) Common Distributions

Distribution	n	Criterion	k_b	$k=1$	$k=4$	$k=k_b$	m_b	$m=1$	$m=4$	$m=m_b$
Uniform	10	Bias	1	.095	.204	.095	6	-.050	-.048	-.043
on $[0,1]$		RMSE	3	.469	.325	.316	9	.302	.229	.156
	25	Bias	1	.032	.082	.032	15	-.020	-.020	-.017
		RMSE	5	.298	.166	.165	24	.171	.103	.060
	50	Bias	1	.018	.041	.018	19	-.010	-.009	-.008
		RMSE	6	.209	.108	.101	49	.118	.059	.029
Student's t	10	Bias	1	.001	-.105	.001	9	-.217	-.377	-.129
Distribution		RMSE	7	.533	.363	.331	9	.461	.494	.407
$\nu = 4$	25	Bias	1	-.011	-.081	-.011	23	-.118	-.266	.001
		RMSE	20	.347	.240	.196	22	.279	.332	.248
	50	Bias	1	.004	-.048	.004	45	-.076	-.184	-.020
		RMSE	41	.245	.170	.137	45	.195	.231	.176
$\nu = 10$	10	Bias	1	.005	-.055	.005	9	-.196	-.332	-.158
		RMSE	7	.513	.328	.296	9	.423	.444	.343
	25	Bias	1	-.003	-.051	-.003	24	-.105	-.232	.038
		RMSE	20	.332	.209	.172	23	.254	.291	.205
	50	Bias	1	-.003	-.033	-.003	47	-.065	-.159	.021
		RMSE	39	.234	.149	.120	46	.171	.200	.138
Standard	10	Bias	6	.024	-.018	.002	9	-.182	-.301	-.170
Normal		RMSE	7	.502	.308	.282	9	.404	.413	.317
	25	Bias	1	.003	-.036	.003	24	-.097	-.209	-.006
		RMSE	19	.321	.197	.167	24	.238	.266	.180
	50	Bias	1	.004	-.027	.004	47	-.057	-.141	-.016
		RMSE	38	.230	.141	.115	47	.163	.180	.117
Gamma	10	Bias	3	.041	.046	.035	8	-.119	-.160	-.002
Distribution		RMSE	4	.558	.402	.402	7	.447	.411	.376
$\alpha = 1$	25	Bias	6	.014	.002	.000	19	-.067	-.125	.011
		RMSE	8	.349	.238	.232	18	.269	.252	.229
	50	Bias	20	.009	-.001	-.001	36	-.040	-.083	-.005
		RMSE	15	.251	.175	.156	36	.187	.173	.159
$\alpha = 3$	10	Bias	4	.023	.002	.002	9	-.160	.267	-.100
		RMSE	7	.514	.331	.322	9	.405	.408	.344
	25	Bias	16	.004	-.026	.001	23	-.092	-.187	.000
		RMSE	11	.331	.205	.188	22	.244	.258	.199
	50	Bias	34	.003	-.018	.000	45	-.053	-.127	-.004
		RMSE	7	.232	.143	.132	45	.168	.177	.138
$\alpha = 6$	10	Bias	2	.019	-.016	.000	9	-.174	-.286	-.131
		RMSE	7	.504	.316	.305	9	.406	.407	.324
	25	Bias	17	.009	-.027	-.001	24	-.090	-.200	.035
		RMSE	18	.326	.199	.180	23	.238	.263	.188
	50	Bias	1	.002	-.022	.002	46	-.056	-.135	-.014
		RMSE	36	.228	.141	.123	46	.165	.178	.128
Standard	10	Bias	3	-.022	.038	.001				
Bivariate		RMSE	4	.553	.397	.397				
Normal	25	Bias	10	-.033	-.064	-.005				
$\rho = 0.0$		RMSE	10	.359	.253	.228				
	50	Bias	23	-.032	-.068	.005				
		RMSE	23	.251	.184	.158				
$\rho = 0.5$	10	Bias	2	-.010	.083	.001				
		RMSE	4	.557	.410	.410				
	25	Bias	8	-.034	-.047	-.002				
		RMSE	7	.354	.246	.230				
	50	Bias	5	-.025	-.012	.002				
		RMSE	5	.255	.176	.168				
$\rho = 0.9$	10	Bias	1	.053	.435	.053				
		RMSE	2	.576	.636	.500				
	25	Bias	2	-.018	.076	.006				
		RMSE	4	.356	.264	.264				
	50	Bias	5	-.027	-.012	.005				
		RMSE	6	.251	.173	.168				

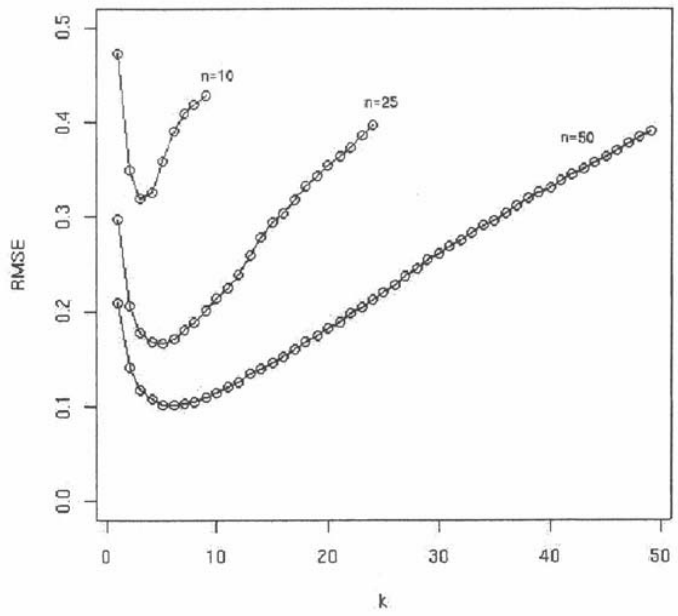

Figure 1: Plot of RMSE of $H_k^{(n)}$, as a function of k, for $n = 10, 25$, and 50 (Uniform distribution on $[0, 1]$)

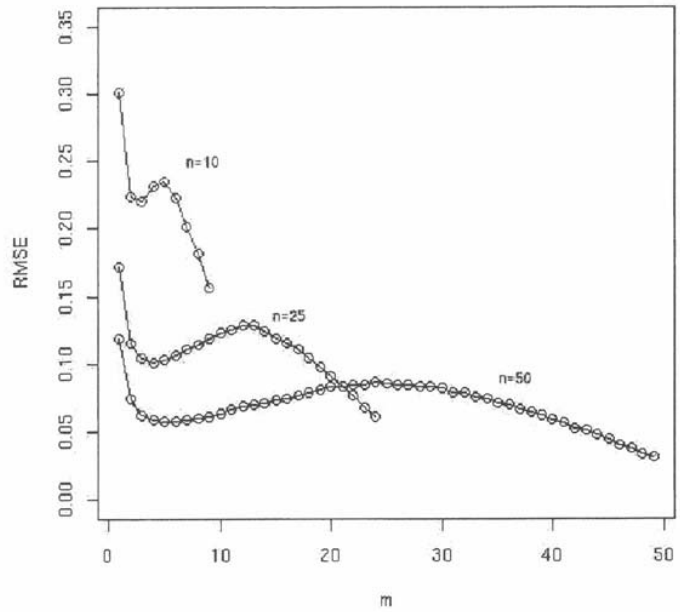

Figure 2: Plot of RMSE of $V_m^{(n)}$, as a function of m, for $n = 10, 25$, and 50 (Uniform distribution on $[0, 1]$)

4. APPLICATIONS

Four different estimates $\hat{H}_k^{(n)}(f)$, $k = 1, 2, 3, 4$, were applied to calculate the internal entropy of a simple molecule (methanol) with the only one internal-rotational degree of freedom, i.e., $p = 1$. A random sample of 15,000 observations of the torsional angle was generated using molecular dynamics (MD) simulations.

MD is a computational method which relies on Newton's equations of motion to describe the evolution of a system of atoms which interact via the potential energy function. The equations of motion are solved iteratively, and thus the coordinates of atoms are propagated in time generating an approximation to the statistical mechanical general population called the *ensemble* of the system. A collection of specific averages over the ensemble, known as thermodynamic potentials, describes the macroscopic behavior of the system, i.e., these functions express such properties of the system which can be experimentally observed. Therefore, methods of MD are widely used in modern chemistry, biology and material science to predict evolution and transformations of the system, e.g., binding of drugs to receptors. Nowadays MD-derived technologies have become an important supplement in decision making and strategic planning in many areas of industry and natural science.

Usually MD is started from an experimentally determined (X-ray or NMR) static three-dimensional structure of a molecule, or its quantum-mechanically calculated set of co-ordinates. Thus at first, each atom is assigned a predetermined position in space. Then, the atoms are assigned initial velocities that are randomly sampled from the Maxwellian distribution at a specific temperature and pressure. Atoms interact with each other according to the laws of physics, and therefore the whole system is able to evolve in time. Evolution is achieved by numerical propagation of Newtonian equations of motion at sufficiently small time intervals, typically each 10^{-15} s. Characteristics of the evolving system are recorded. The time interval of recording is chosen so that during it the system "forgets" about the previous state; i.e., the snapshots are independent of each other. A result of this procedure is a randomly sampled Boltzmann-Gibbs ensemble of the given molecular system simulated under specific well-controlled conditions; i.e., it is the full statistical-mechanical description of the process that is required by statistical-mechanical theory. In this theory, the observed macroscopic quantities are expressed as statistical averages over the Boltzmann-Gibbs ensemble of microscopic realizations. To this end, the theory provides a connection between simulated microscopic and experimentally measured macroscopic observations (Landau and Lifshitz, 1980, for more details).

Gibbs free energy G is a thermodynamic potential that describes the fu-

ture of the system (e.g., binding, separation, fusion/decay, etc.) at constant temperature and pressure. G is given by the expression

$$G = H - TS,$$

where H is the enthalpy (the average internal potential energy of the system), S is the entropy of the system and T is the temperature. The change in free energy between the two states is given by

$$\delta(G) = \delta(H) - T\delta(S),$$

where $\delta(H)$ is the change in the enthalpy and δS is the change in the entropy. The change from one state to another is spontaneous if the corresponding change in the free energy $\delta(G) < 0$. The system always naturally tends to evolve in the direction of a state with minimum free energy. Enthalpy is easily calculated as the first moment of the ensemble of energy states. Entropy is a fundamental thermodynamic function which describes the "volume" of phase space from which the system samples its microscopic realizations. Entropy depends on the full ensemble of microstates and can not be trivially obtained from MD simulations like enthalpy (refer Landau and Lifshitz, 1980). Development of new methods for accurate estimation of entropy is an important goal in improving the predictive power of statistical-mechanical methods.

Based on a random sample of 15,000 observations on the torsional angle of methanol generated using MD simulations, the estimates of entropy, given by $\hat{H}_k^{(n)}(f)$, $k = 1, 2, 3, 4$, for the methanol molecule at room temperature are 1.840, 1.777, 1.770 and 1.756, respectively. The estimate of entropy obtained by Demchuk and Singh (2001), using a probabilistic modeling approach, is 1.744. The numeric estimate $\hat{H}_4^{(n)}(f) = 1.756$ obtained using the fourth nearest neighbor distances is within 0.7% of this number, which is based on a precise parametric fit. The estimates of entropy, given by $\hat{V}_m^{(n)}$, for $m = 1, 2, 3, 4$, are 1.776, 1.736, 1.740 and 1.758 respectively.

A similar calculation was done using $\hat{H}_k^{(n)}(f)$, for $k = 1, 2, 3, 4$, to estimate the full internal-rotational entropy of diethyl ether at room temperature. This molecule has four torsional degrees of freedom and thus $p = 4$. Based on a random sample of 15,000 observations on the four torsional angles generated using MD simulations, these estimates of entropy are 3.216, 3.236, 3.196 and 3.199, respectively.

ACKNOWLEDGMENTS

The authors are grateful to Dan S. Sharp, E. James Harner and Sidney C. Soderholm for helpful discussions and to Lingyi Zheng and J. Tan for their help with the preparation of the data. A.F. was supported by a National

Research Associateship of the National Academy of Science. The authors are also thankful to an anonymous reviewer for insightful comments that led to an improved presentation.

REFERENCES

Abramowitz, M., & Stegun, A.(1965). Handbook of Mathematical Functions. Dover, New York.

Beirlant, J., Dudewicz, E. J., Györfi, L., & van der Meulen, E. C. (1997). Nonparametric estimation of entropy: An overview. International Journal of Mathematical and Statistical Sciences, 6, 17-39.

Chen, B., McCoskey, S.K. & Kao, C. (1999). Estimation and inference of a cointegrated regression in panel data: a Monte Carlo study. American Journal of Mathematical and Management Sciences, 19, 75-114.

Cressie, N. (1976). On the logarithms of high-order spacings. Biometrika, 63, 345-355.

Demchuk, E. & Singh, H. (2001). Statistical thermodynamics of hindered rotation from computer simulations. Molecular Physics, 99, 627-636.

Dobrushin, R. L. (1958). A simple method of empirical estimation of the entropy of a stationary sequence. Teor. Veroyatn. Ee Primen, 3, 462-464.

Dudewicz, E.J. & van der Meulen, E.C. (1981). Entropy-based tests of uniformity. Journal of the American Statistical Association, 76, 967-974.

Karian, Z.A. & Dudewicz, E.J. (1999). Modern Statistical, Systems, and GPSS Simulation, 2nd edition. CRC Press, Boca Raton, FL.

Karplus, M. & Kushik, J.N. (1981). Method for estimating the configurational entropy of macromolecules. Macromolecules, 14, 325-332.

Kozachenko, L. F. & Leonenko, N. N. (1987). Sample estimates of entropy of a random vector. Problems of Information Transmission, 23, 95-101.

Landau L. D. & Lifshitz, E. M. (1980). Statistical Physics, Part 1, 3rd edition. Oxford: Pergamon.

Levy, R.M., Karplus, M., Kushik, J. & Perahia, D. (1984). Evaluation of the configurational entropy for proteins: Application to molecular dynamics simulations of an α-helix. Macromolecules, 17, 1370-1374.

Scott, D. (1992). Multivariate Density Estimation: Theory, Practice and Visualization. John Wiley and Sons Inc., New York.

Singh, H., Hnizdo, V. & Demchuk, E. (2002). Probabilistic modeling of two

dependent circular variables. Biometrika, 89, 719-723.

Tsybakov, A. B. & van der Meulen, E. C. (1996). Root-n consistent estimators of entropy for densities with unbounded support. Scandinavian Journal of Statistics, 23, 75-83.

Vasicek, O. (1976). On a test for normality based on sample entropy. Journal of Royal Statistical Society, Series B, 38, 54-59.