NOTICE: The copyright law of the United States (Title 17, U.S. Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research."

The CDC library absorbs the cost of copyright fees charged by publishers when applicable and the cost of articles and books obtained from other libraries. Copyright fees average $35.00 and fees charged by the lending libraries are between $10 and $15 per request

Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. **53**(3), 2011, 271–288 doi: 10.1111/j.1467-842X.2011.00633.x

ESTIMATION OF MULTIVARIATE SHANNON ENTROPY USING MOMENTS

ROBERT M. MNATSAKANOV[1,2,*], SHENGQIAO LI[2] AND E. JAMES HARNER[1]

West Virginia University and National Institute for Occupational Safety and Health

Summary

Three new entropy estimators of multivariate distributions are introduced. The two cases considered here concern when the distribution is supported by a unit sphere and by a unit cube. In the former case, the consistency and the upper bound of the absolute error for the proposed entropy estimator are established. In the latter one, under the assumption that only the moments of the underlying distribution are available, a non-traditional estimator of the entropy is suggested. We also study the practical performances of the constructed estimators through simulation studies and compare the estimators based on the moment-recovered approaches with their counterparts derived by using the histogram and kth nearest neighbour constructions. In addition, one worked example is briefly discussed.

Key words: moment-recovered estimates; mean squared error; spherical data.

1. Introduction

In the field of molecular physics, when investigating the thermodynamic properties of a complex molecular system, it is important to know the differential Shannon entropy of a probability density function (pdf) f:

$$H(f) = -\int f(\boldsymbol{x}) \ln f(\boldsymbol{x})\, d\lambda(\boldsymbol{x}), \tag{1}$$

which represents the measure of random fluctuations of a molecule's bond (torsional) angles, the dihedral angles, etc., distributed according to f. The term λ in (1) denotes the uniform distribution defined on the range of corresponding angles. In most cases f is multimodal and skewed, which is why the use of only parametric models for evaluation of $H(f)$ is not appropriate.

Below, in Figure 1, the structure of the bromo (chloromethyoxy) molecule with two torsional angles ϕ_1 and ϕ_2 is illustrated, while Figure 4 (see Section 5) demonstrates the estimated joint density function of $f(\phi_1, \phi_2)$ based on a molecular dynamic (MD) simulation conducted by researchers at the National Institute for Occupational Safety and Health (NIOSH). In this calculation we applied the moment-recovered pdf $f_{\mathbf{a},\widehat{\nu}}$ introduced in Section 3.

*Author to whom correspondence should be addressed.
[1]Department of Statistics, P.O. Box 6330, West Virginia University, Morgantown, WV 26506, USA. e-mail: rmnatsak@stat.wvu.edu
[2]National Institute for Occupational Safety and Health, 1095 Willowdale Road, Morgantown, WV 26505, USA

Acknowledgments. The first author would like to thank Estate V. Khmaladze and Frits H. Ruymgaart for helpful discussions. The authors also thank the referees for their comments and suggestions, which led to improvements of the original version. The research of Robert M. Mnatsakanov is supported by NSF grant DMS-0906639. 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.

© 2011 Australian Statistical Publishing Association Inc. Published by Blackwell Publishing Asia Pty Ltd.

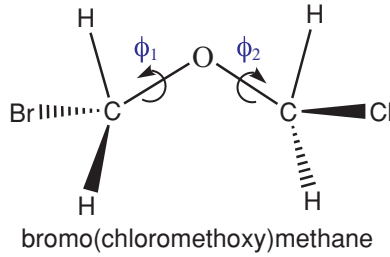


Figure 1. The structure of a bromo(chloromethyoxy) methane molecule with two torsional angles ϕ_1 and ϕ_2.

In the following we will assume that a sequence of i.i.d. random vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots,$ defined on the unit sphere $\mathbf{S}^{p-1} \subset \mathbb{R}^p$ or on the unit cube $[0, 1]^p$, is given. Denote the common distribution of $\boldsymbol{X}_i$s by P and the density of P by f (with respect to the uniform distribution λ); that is,

$$f(\boldsymbol{x}) = \frac{dP}{d\lambda}(\boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \mathbf{S}^{p-1} \text{ or } \boldsymbol{x} \in [0, 1]^p.$$

Let $\hat{f}_n$ denote a nonparametric estimate of f constructed from $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$. There are two methods for estimating the entropy $H(f)$: one is based on replacing the pdf f in definition (1) by its estimate $\hat{f}_n$, say a kernel or histogram density estimator; the other one is based on using the formula

$$\hat{H}_n = -\frac{1}{n}\sum_{i=1}^{n} \ln(\hat{f}_n(\boldsymbol{X}_i))$$

(see, for example, Joe 1989; Hall & Morton 1993; and Beirlant *et al.* 1997, among others). Note that the implementation of a method of estimating a functional, such as $H(f)$ by $H(\hat{f}_n)$, when $\hat{f}_n$ is a kernel density estimator becomes inaccurate owing to evaluations of the multiple integral in $H(\hat{f}_n)$ (see, for example, Joe 1989). In Gyöfri & van der Meulen (1987) the strong consistency of the truncated version of the entropy estimator $H(\hat{f}_n)$ when $\hat{f}_n$ is a histogram estimator of f was established. The authors replaced the integral in $H(\hat{f}_n)$ by the one taken over the set $A_n = \{\boldsymbol{x} : \hat{f}_n(\boldsymbol{x}) \geq a_n\}$ with $0 < a_n \to 0$, and showed that $H(\hat{f}_n)$ is easier to evaluate.

From a computational point of view, the second method is preferable, especially when the dimensionality p is large (Misra, Singh & Hnizdo 2010; Li, Mnatsakanov & Andrew 2011). Note that in applying the second method, Hall & Morton (1993) and Joe (1989) proved that for kernel-density entropy estimators the rate of convergence is $n^{-1/2}$ when the dimensionality is $p \leqslant 3$, while for histogram-type constructions the rate of convergence of $\hat{H}_n$ is $n^{-1/2}$ only if $p = 1, 2$. From a computational viewpoint, in high-dimensional models, one of the most efficient nonparametric estimates of the Shannon entropy has been shown to be the one based on the kth nearest neighbour (k-NN) approach. These types of estimates have been studied extensively during the last three decades. See, for example, Kozhachenko & Leonenko (1987), Singh *et al.* (2003), Mnatsakanov *et al.* (2008), and Misra *et al.* (2010), among others. The rate of convergence of the k-NN entropy estimator was derived for $p = 1$

© 2011 Australian Statistical Publishing Association Inc.

and $k = 1$ in Tsybakov & van der Meulen (1996). The authors used the truncation technique and assumed that the density f has an exponentially decreasing tail.

Recently, a moment-type estimation technique, namely a procedure that recovers the underlying distribution from its assigned moments (estimated moments), has been developed and applied both in univariate and in multivariate cases: see Mnatsakanov (2008a; 2008b) and Gzyl & Tagliani (2010) for the univariate case; and Mnatsakanov & Li (2010) and Mnatsakanov (2011) for the multivariate case. Hence, to estimate the entropy $H(f)$ we propose to use the moment-recovered (MR) approximations (Mnatsakanov 2008a; 2011) of corresponding distributions supported by a unit sphere and by a unit cube.

The main aim of the present article is to investigate the L_1-rate of convergence of the entropy estimate based on the histogram type density estimate $\hat{f}_n^{(1)}$ and compare it via simulation with its MR counterpart based on the moment-empirical estimate $\hat{f}_n^{(2)}$ (see Section 2). Although we used the simulated spherical data when $p = 2$ and 3, our approach is useful for large values of $p > 3$ as well. See, for example, Li *et al.* (2011), where we calculated the MR entropy estimate for spherical data and compared its biases and mean squared errors (MSEs) with those based on the k-NN constructions when $p = 3$ and $p = 10$. In addition, we applied the two-dimensional MR approximation of the unknown pdf f and studied the asymptotic behaviour of the corresponding MR entropy estimator via simulations on a unit cube.

The paper is organized as follows. In Section 2 we describe the constructions of the estimates $\hat{f}_n^{(k)}$, $k = 1, 2$. We also explain how the form of $\hat{f}_n^{(2)}$ is connected to the solution of the Hausdorff moment problem. In addition, we outline the closeness of these two constructions and demonstrate it graphically. In this section we pay more attention to deriving the asymptotic properties of the entropy estimate based on $\hat{f}_n^{(1)}$, as it is easier to determine. Namely, in Theorems 1 and 2 the consistency and the L_1-rate of convergence are established for the corresponding entropy estimator based on $\hat{f}_n^{(1)}$. The proofs of these statements are provided in the Appendix. In Section 3 we apply the two-dimensional version of the MR construction for estimation of the entropy $H(f)$ when the only available information about f represents the sequence of its moments. In Section 4, simulation studies on a sphere $\mathbf{S}^{p-1}$ and on $[0, 1]^p$ when $p = 2, 3$ are conducted. In Tables 1 and 2 we record the L_1-errors and MSEs of the proposed entropy estimates when the target distributions are von Mises and uniform, and in Tables 3 and 4 the biases and MSEs of a bivariate von Mises distribution and a Dirichlet distribution are presented. In Figures 2 and 3 we plot the curves of target and MR pdfs for von Mises and Dirichlet distributions. Finally, in Section 5 we present an example studied by researchers in NIOSH. Figure 4 provides the estimated pdf of a worked example derived using MD simulation. In Section 6 we outline the main advantages of the estimates proposed in this work.

2. Estimation on the unit sphere $\mathbf{S}^{p-1}$

In the case of spherical data, two types of estimates of the pdf f are used when estimating the unknown entropy $H(f)$ from (1). The first one, $\hat{f}_n^{(1)}$, is based on the histogram type of construction in Ruymgaart (1989), and the second one, $\hat{f}_n^{(2)}$ in (10), is based on the moment-empirical cumulative distribution function (cdf) introduced and studied in Mnatsakanov & Ruymgaart (2003).

© 2011 Australian Statistical Publishing Association Inc.

First, let us recall the estimate $\hat{f}_n^{(1)}$. Consider the subsets of $\mathbf{S}^{p-1}$ parametrized as follows: for any unit vector $\boldsymbol{x} \in \mathbf{S}^{p-1}$ and a number $t \in (0, 1)$, consider a hyperplane perpendicular to the vector $\boldsymbol{x}$ and crossing the line passing through the points $\mathbf{0}$ and $\boldsymbol{x}$ at $t\boldsymbol{x}$. This hyperplane separates the sphere into two parts. The closure of the smaller of these two parts is denoted by $C_{\boldsymbol{x}}(t) = \{\boldsymbol{y} \in \mathbf{S}^{p-1} : \boldsymbol{y}^\top \boldsymbol{x} > t\}$. This is a cap with pole $\boldsymbol{x}$ and radius $(1 - t^2)^{1/2}$. Now for any Borel set B in $\mathbb{R}^p$, define the empirical measure

$$\hat{P}_n(B) = \frac{1}{n} \sum_{i=1}^{n} I_B(\boldsymbol{X}_i).$$

Here $I_B(\cdot)$ denotes an indicator function of the set B. Since at each $\boldsymbol{x} \in \mathbf{S}^{p-1}$

$$f(\boldsymbol{x}) = \lim_{t \uparrow 1} \frac{P(C_{\boldsymbol{x}}(t))}{\lambda(C_{\boldsymbol{x}}(t))},$$

a natural estimator of the density f at $\boldsymbol{x}$ is

$$\hat{f}_n^{(1)}(\boldsymbol{x}) = \frac{\hat{P}_n(C_{\boldsymbol{x}}(t_n))}{\lambda(C_{\boldsymbol{x}}(t_n))}, \tag{2}$$

where $t_n \in (0, 1)$ and $t_n \uparrow 1$ as $n \to \infty$. The uniform rate of convergence of $\hat{f}_n^{(1)}$ and the speed of convergence of t_n to 1 are specified in theorem 3.1 of Ruymgaart (1989).

In definition (2), it is important to have the direct expression of $\lambda(C_{\boldsymbol{x}}(t_n))$. Consider a sphere $S^{p-1}(r) \subset \mathbb{R}^p$ of radius $r > 0$. Its total area $\lambda(S^{p-1}(r)) = 2\pi^{p/2} r^{p-1}/\Gamma(p/2)$. One can easily derive the area of the corresponding cap $C_{\boldsymbol{x}}(t, r) \subset S^{p-1}(r)$ as well:

Lemma 1. *If* $\mathrm{Beta}(\cdot, a, b)$ *is a cdf of a beta distribution with the parameters* $a, b > 0$, *then for* $t \in (0, r)$

$$\lambda(C_{\boldsymbol{x}}(t, r)) = \frac{1}{2} \lambda(S^{p-1}(r)) \, \mathrm{Beta}\left(\frac{r^2 - t^2}{r^2}, \frac{p-1}{2}, \frac{1}{2} \right).$$

For the proof of Lemma 1 we refer to Li (2011). This formula, with $r = 1$ and $C_{\boldsymbol{x}}(t) := C_{\boldsymbol{x}}(t, 1)$, that is,

$$\lambda(C_{\boldsymbol{x}}(t)) = \frac{\pi^{(p-1)/2}}{\Gamma((p-1)/2)} \int_0^{1-t^2} u^{(p-3)/2} (1 - u)^{-1/2} \, du, \tag{3}$$

will be used in numerical calculations in Section 4. Based on (2) and (3) we derive the first estimator of $H(f)$:

$$\widehat{H}_n^{(1)} = -\frac{1}{n} \sum_{i=1}^{n} \ln \hat{f}_{n,-i}^{(1)}(\boldsymbol{X}_i). \tag{4}$$

Here $\hat{f}_{n,-i}^{(1)}$ denotes the so-called leave-one-out density estimate: it has the same form as $\hat{f}_n^{(1)}$ in (2) but is computed by means of $\{\boldsymbol{X}_j, j \neq i, j = 1, \ldots, n\}$. The leave-one-out kernel and histogram density estimates of the entropy have been studied by Hall & Morton (1993), while the kernel density estimate was examined by Joe (1989). See also the references in Beirlant *et al.* (1997).

© 2011 Australian Statistical Publishing Association Inc.

Our second construction is based on the solution of the Hausdorff moment problem. To be more specific, consider the moment-determinate cdf f defined on the interval $[0, T)$, $T \leqslant \infty$, and denote its moments by

$$\mu_{k,F} = \int_0^T t^k d\, F(t) = (\mathcal{K}F)(k), \quad k = 0, 1, \ldots. \tag{5}$$

According to (5) the operator $\mathcal{K}$ can be considered as an operator mapping F into the vector of moments $\mu_F = (\mu_{0,F}, \mu_{1,F}, \ldots)$, where $\mu_{0,F} = 1$. The inverse of the operator $\mathcal{K}$ can be approximated by the sequence of operators defined for any $T > 0$. When $T = 1$, this sequence is defined as follows:

$$\left(\mathcal{K}_m^{-1}\, \mu_F\right)(t) = \sum_{k=0}^{[mt]} \sum_{j=k}^{m} \binom{m}{j} \binom{j}{k} (-1)^{j-k}\, \mu_{j,F}, \quad 0 \leqslant t < 1, \tag{6}$$

where $[mt]$ denotes the integer part of mt, $m = 0, 1, \ldots$. More precisely, the following statement is true (see Mnatsakanov & Ruymgaart 2003):

$$\mathcal{K}_m^{-1}\mathcal{K}F \to_{\mathrm{w}} F, \quad \text{as } m \to \infty. \tag{7}$$

By $\to_{\mathrm{w}}$, we mean the weak convergence of the cdfs (i.e. convergence at each continuity point of the limiting cdf). Under some extra conditions on F, one can prove that (7) is valid in the sense of the uniform and L_1-norms (see Mnatsakanov 2008a).

Let $\widehat{\mu}_n = (\widehat{\mu}_{0,n}, \widehat{\mu}_{1,n}, \ldots, \widehat{\mu}_{m,n})$, with

$$\widehat{\mu}_{k,n} = \frac{1}{n} \sum_{i=1}^{n} Y_i^k, \quad k = 0, 1, \ldots, m, \tag{8}$$

be the estimated vector of moments up to order m. Here $Y_1, \ldots, Y_n$ are n i.i.d. copies of Y distributed according to F. Our approach yields the construction of the so-called moment-empirical cdf based on the data $Y_1, \ldots, Y_n$:

$$F_n^* = \mathcal{K}_m^{-1}\, \widehat{\mu}_n, \tag{9}$$

with $m = m(n) \to \infty$ as $n \to \infty$.

Mnatsakanov & Ruymgaart (2003) proved, for $m = n$, that $n^{1/2}(F_n^* - F)$ converges in distribution to the zero-mean Gaussian process U with the same covariance function $\mathrm{E}(U(t)\, U(s)) = F(t \wedge s) - F(t)\, F(s)$ as we have for the ordinary empirical process

$$U_n = n^{1/2}(\hat{F}_n - F) \quad \text{with} \quad \hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} I_{[0,t]}(Y_i), \quad 0 \leqslant t < 1.$$

Our second construction $\hat{f}_n^{(2)}$ is based on replacing the empirical measure $\hat{P}_n$ in (2) by the moment-empirical counterpart denoted here by P_n^*. Namely, in order to estimate the density $f(\boldsymbol{x})$ at any fixed point $\boldsymbol{x} \in \mathbf{S}^{p-1}$, let us consider the projection of $\boldsymbol{X}_i$ on $\boldsymbol{x}$ and take $Y_i = Y_{i,\boldsymbol{x}} = \boldsymbol{X}_i^{\top} \boldsymbol{x} \geq 0$ in (8). Now, one can easily derive from (6) and (9) that the smoothed

© 2011 Australian Statistical Publishing Association Inc.

moment-empirical version estimating $P(C_{\boldsymbol{x}}(t)) = P(Y_{1,\boldsymbol{x}} > t) = 1 - F_{\boldsymbol{x}}(t)$ has the form

$$P_n^*(C_{\boldsymbol{x}}(t)) = 1 - F_n^*(t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=[mt]+1}^{m}\binom{m}{k}\left(\boldsymbol{X}_i^\top \boldsymbol{x}\right)^k\left(1-\boldsymbol{X}_i^\top \boldsymbol{x}\right)^{m-k}, \quad 0 \leqslant t < 1.$$

The moment-empirical density estimator of $f(\boldsymbol{x})$ is constructed as

$$\hat{f}_n^{(2)}(\boldsymbol{x}) = \frac{P_n^*(C_{\boldsymbol{x}}(t))}{\lambda(C_{\boldsymbol{x}}(t))}, \quad \text{as } t \uparrow 1, \quad \boldsymbol{x} \in \mathbf{S}^{p-1}. \tag{10}$$

One can easily see that, given $\boldsymbol{X}_i$, $i = 1, \ldots, n$, we have

$$\sum_{k=[mt]+1}^{m}\binom{m}{k}\left(\boldsymbol{X}_i^\top \boldsymbol{x}\right)^k\left(1-\boldsymbol{X}_i^\top \boldsymbol{x}\right)^{m-k} \sim I\left\{\boldsymbol{X}_i^\top \boldsymbol{x} > t\right\} \quad \text{for each } \boldsymbol{x} \in \mathbf{S}^{p-1}, \quad 0 \leqslant t < 1,$$

for large m; that is, $P_n^*(C_{\boldsymbol{x}}(t)) \sim \hat{P}_n(C_{\boldsymbol{x}}(t))$. That is why in the rest of this section and in Section 4 we will focus our interest on the asymptotic properties of $\hat{H}_n^{(1)}$ based on $\hat{f}_n^{(1)}$, and will compare the performance of $\widehat{H}_n^{(1)}$ with

$$\widehat{H}_n^{(2)} = -\frac{1}{n}\sum_{i=1}^{n}\ln \hat{f}_{n,-i}^{(2)}(\boldsymbol{X}_i). \tag{11}$$

The estimate $\hat{f}_{n,-i}^{(2)}$ in (11) is defined in a similar way to $\hat{f}_{n,-i}^{(1)}$ in (4). The rate of convergence of $t = t_n$ to 1, as $n \to \infty$, can be specified in a similar way to in Ruymgaart (1989). Furthermore, in Section 4 we chose the optimal parameter $t = t^*$ by optimizing the L_1-error or MSE of $\widehat{H}_n^{(k)}$, $k = 1, 2$, with respect to t.

Remark 1. From a computational point of view, the estimate $\widehat{H}_n^{(2)}$ compared to $\widehat{H}_n^{(1)}$ is more convenient to use because the values of $\ln \hat{f}_{n,-i}^{(2)}$ at some of the $\boldsymbol{X}_i$s in (11) will not degenerate to $-\infty$.

Now let us consider the following conditions:

(A_1): $n(1-t_n)^{(p-1)/2} \to \infty$ as $t_n \uparrow 1, n \to \infty$;
(A_2): $B_1 = \inf_{\boldsymbol{x} \in s^{p-1}} f(\boldsymbol{x}) > 0$;
(A_3): for some $k_n \to \infty$, as $n \to \infty$, we have

$$M_n = \left(\frac{n\,\lambda(C_{\boldsymbol{x}}(t_n))}{K_n\,\ln(1/\lambda(C_{\boldsymbol{x}}(t_n)))}\right)^{1/2} \to \infty.$$

Theorem 1. *Suppose that f is continuous and conditions (A_j), $j = 1, 2$, are satisfied. Then*

$$\mathrm{E}\left(\widehat{H}_n^{(1)}\right) \to H(f) \text{ and } \operatorname{var}\left\{\widehat{H}_n^{(1)}\right\} \to 0 \text{ as } t_n \uparrow 1, \quad n \to \infty.$$

Theorem 2. *If f is continuous and conditions (A_j), $j = 1, 2, 3$, hold, then*

$$\rho_n = E(|\widehat{H}_n^{(1)} - H(f)|) \leq \left(\frac{B_2}{n}\right)^{1/2} + \frac{1}{\theta}\,\Delta(f, t_n) + o(1/M_n) \quad \text{as } t_n \uparrow 1,\ n \to \infty,$$

for some $0 < \theta < 1$ and $\Delta(f,t) = \sup_{(\boldsymbol{x},\boldsymbol{y}) \in \mathbf{S}^{p-1} \times C_{\boldsymbol{x}}(t) \times C_{\boldsymbol{x}}(t)} |f(\boldsymbol{x}) - f(\boldsymbol{y})|$.

© 2011 Australian Statistical Publishing Association Inc.

For the proofs of Theorems 1 and 2, see the Appendix.

Remark 2. Hall & Morton (1993) noted that the estimates of entropy based on the leave-one-out procedures are very sensitive to outliers in the case of $\boldsymbol{X}_i \in \mathbb{R}^p$. Under the condition (A_2), the estimates $\widehat{H}_n^{(k)}$, $k = 1, 2$, defined by (4) and (11), are free from this shortcoming.

3. Estimation on the unit cube $[0, 1]^p$

Suppose now that the moment-determinate cdf F is absolutely continuous with respect to the Lebesgue measure λ and has support $[0, 1]^p$, $p \geqslant 2$. For simplicity of notation, we consider only the case when $p = 2$. Denote the corresponding density function by f and assume that f is observed indirectly and only its moments are available (or can be estimated from the data). To approximate $H(f)$ one can first recover f via the constructions studied in Mnatsakanov (2008b, 2011), where the univariate and multivariate models are considered, respectively.

To describe this construction in the two-dimensional case let us consider the ordinary moments of f:

$$\mu_{m,j} = \int\int t^m u^j f(t, u)\, dt\, du, \quad m, j \in \mathbb{N} = \{0, 1, \ldots\}.$$

In this case, the moment sequence $\boldsymbol{\nu} = \{\mu_{m,j}, m, j \in \mathbb{N}\}$ determines f uniquely; that is, the pdf f is Hausdorff moment-determinate (see conditions in Shohat & Tamarkin 1943). Our MR approximation of f has the form

$$f_{\mathbf{a},\boldsymbol{\nu}}(x, y) = \frac{\Gamma(\alpha + 2)\,\Gamma(\alpha' + 2)}{\Gamma([\alpha x] + 1)\Gamma([\alpha' y] + 1)} \times \sum_{m=0}^{\alpha - [\alpha x]} \sum_{j=0}^{\alpha' - [\alpha' y]} \frac{(-1)^{m+j}\, \mu_{m+[\alpha x], j+[\alpha' y]}}{m!\, j!\, (\alpha - [\alpha x] - m)!\, (\alpha' - [\alpha' y] - j)!}. \tag{12}$$

Here $\mathbf{a} = (\alpha, \alpha')$ with α and $\alpha' \to \infty$ at an appropriate rate.

Remark 3. If a density f has a compact support $[0, T]^2$, $0 < T < \infty$, then we can use the following MR approximation of f:

$$f_{\mathbf{a},\boldsymbol{\nu}}(x, y) = \frac{1}{T^{[\alpha x/T] + [\alpha' y/T] + 2}} \times \frac{\Gamma(\alpha + 2)\,\Gamma(\alpha' + 2)}{\Gamma([\alpha x/T] + 1)\,\Gamma([\alpha' y/T] + 1)} \times \sum_{m=0}^{\alpha - [\alpha x/T]} \sum_{j=0}^{\alpha' - [\alpha' y/T]} \frac{\left(-\frac{1}{T}\right)^{m+j} \hat{\mu}_{m+[\alpha x/T], j+[\alpha' y/T]}}{m!\, j!\, (\alpha - [\alpha x/T] - m)!\, (\alpha' - [\alpha' y/T] - j)!}.$$

Mnatsakanov & Li (2010) prove the following theorem.

© 2011 Australian Statistical Publishing Association Inc.

Theorem. *Let* $\boldsymbol{\nu} = \{\mu_{m,j}, (m, j) \in \mathbb{N}_{\mathbf{a}}\}$ *and assume that* f *is continuous on* $[0, 1]^2$. *Then* $f_{\mathbf{a},\boldsymbol{\nu}} \to f$ *uniformly as* $\alpha, \alpha' \to \infty$, *and for some* $0 < \delta < 1/2$,

$$||f_{\mathbf{a},\boldsymbol{\nu}} - f||_\infty \leq \bar{\Delta}(f, \delta) + \frac{2\,||f||_\infty}{\delta^4\,(\alpha + 2)\,(\alpha' + 2)}.$$

Here $||f||_\infty$ denotes the sup-norm of f on $[0, 1]^2$, while $\bar{\Delta}(f, \delta) = \sup_{(x,y)\in[0,1]^2} \sup_{(t,s)\in S(x,y;\delta)} |f(t, s) - f(x, y)|$ represents the modulus of continuity of f, and $S(x, y; \delta) = \{(t, s) = |t - x| \leqslant \delta; |s - y| \leqslant \delta\}$ with $0 < \delta < 1/2$, and $\mathbb{N}_{\mathbf{a}} = \{0, 1, \ldots, \alpha\} \times \{0, 1, \ldots, \alpha'\}$.

As special cases, consider two families of functions $f : [0, 1]^2 \to \mathbb{R}$:

Case 1: f is a polynomial of order up to $p + q$:

$$f(t, s) = \sum_{m=0}^{p} \sum_{j=0}^{q} a_{mj}\, t^m s^j\,.$$

The class of all such functions with $a_{00} = 0$ and all a_{mj} finite will be denoted by $\mathcal{P}_{p+q}$.

Case 2: consider the class of Lipschitz-continuous functions on $[0, 1]^2$:

$$\mathcal{L}_{\beta,L} = \{f : |f(x, y) - f(t, s)| \leq L\,\{|x - t|^\beta + |y - s|^\beta\}\}, \quad 0 < \beta \leq 1 \quad \text{and} \quad L > 0.$$

For these two cases it was proved in Mnatsakanov & Li (2010) that:

(i) *if* $f \in \mathcal{P}_{p+q}$ *and* $\alpha = \alpha' \to \infty$, *then* $||f_{\mathbf{a},\boldsymbol{\nu}} - f||_\infty \sim \frac{C_1}{\alpha}$ *, for some constant* C_1;

(ii) *if* $f \in \mathcal{L}_{\beta,L}$, *and* $C_2 = (\beta + 2)\,L^{\frac{2}{\beta+2}}\,(\frac{2}{\beta})^{\frac{\beta}{\beta+2}}$, *then*

$$||f_{\mathbf{a},\boldsymbol{\nu}} - f||_\infty \leq \frac{C_2}{(\alpha + 2)^{\frac{\beta}{\beta+2}}} + o\left(\frac{1}{\alpha}\right), \quad \text{as} \quad \alpha = \alpha' \to \infty.$$

Hence, one can use the estimated sequence of moments $\widehat{\boldsymbol{\nu}}$ (instead of the theoretical moment sequence $\boldsymbol{\nu}$ in (12)) and estimate the entropy $H(f)$ by means of a construction similar to (4). For example, given the empirical moment sequence $\widehat{\boldsymbol{\nu}} = \{\widehat{\mu}\,(j, m), (j, m) \in \mathbb{N}_{\mathbf{a}}\}$ with the components

$$\widehat{\mu}(j, m) = \frac{1}{n} \sum_{i=1}^{n} X_i^j\, Y_i^m\,,$$

we can estimate the entropy $H(f)$ using

$$\widehat{H}_{\mathbf{a}} = -\frac{1}{n} \sum_{i=1}^{n} \ln(f_{\mathbf{a},\widehat{\boldsymbol{\nu}}}(\boldsymbol{X}_i)), \quad \boldsymbol{X}_i = (X_i, Y_i). \tag{13}$$

Here $\{\boldsymbol{X}_i\}_{i=1}^n$ is a sample from f and $f_{\mathbf{a},\widehat{\boldsymbol{\nu}}}$ is the MR pdf (12) with α and $\alpha' \to \infty$. When applying (13) it is important to investigate the behaviour of the estimate $\widehat{H}_{\mathbf{a}}$ as a function of α, α' and the sample size n. In Section 4 we simulate the samples from the von Mises and Dirichlet distributions with different sample sizes. The optimal α^* (when $\alpha = \alpha'$) are defined by minimizing the simulated MSE of $\widehat{H}_{\mathbf{a}}$. In particular, in Table 4 we record the biases and

© 2011 Australian Statistical Publishing Association Inc.

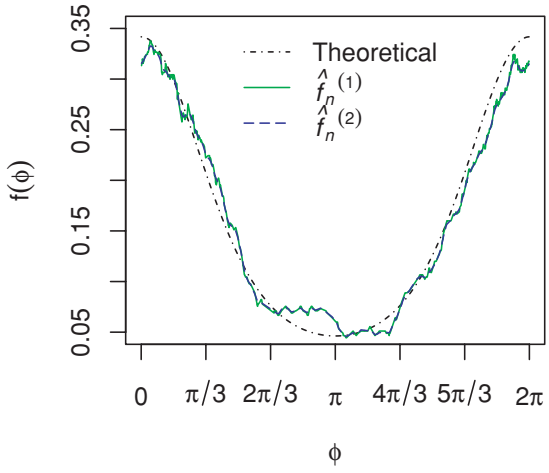


Figure 2. The graphs of the histogram density estimate $\hat{f}_n^{(1)}$ and the moment-empirical density estimate $\hat{f}_n^{(2)}$ of the von Mises distribution with parameters $\kappa = 1$ and $\phi_0 = 0$ when $t = 0.9$. In black and white plot the estimates $\hat{f}_n^{(k)}$, $k = 1, 2$ have almost overlapping lines.

MSEs of $\widehat{H}_{\mathbf{a}}$ for the three-dimensional Dirichlet distribution. Here we chose some of the values of $\alpha = \alpha' \in \{10 + 5j$, with $j = 0, 1, \ldots, 15\}$.

4. Simulation study

In this section we simulate samples on the unit sphere and unit cube. For each distribution, we drew samples of sizes $n = 50,\ 100,\ 500,$ and 1000, and for each sample size n we repeated our calculations $N = 100$ times.

von Mises Distribution on $\mathbf{S}^1$: Consider first the case of a two-dimensional unit sphere $\mathbf{S}^1$. We conducted the simulations for estimating the pdf f_{vM} of a univariate von Mises distribution vM(κ, ϕ_0) with concentration κ and mean direction ϕ_0 given by

$$f_{\mathrm{vM}}(\phi) = \frac{1}{2\pi I_0(\kappa)} \mathrm{e}^{\kappa \cos(\phi - \phi_0)}, \quad 0 \le \phi < 2\pi,$$

where $I_0(\kappa)$ denotes the modified Bessel function of order 0. Figure 2 displays the performance of the estimates $\hat{f}_n^{(k)}$, $k = 1, 2$, respectively, with $n = 500$, and $t = 0.9$ for the vM(1, 0) distribution. From Figure 2 we see that the moment-empirical density estimate $\hat{f}_n^{(2)}$ is somewhat smoother when compared to the histogram density estimate $\hat{f}_n^{(1)}$.

The exact expression for the entropy of the vM(κ, ϕ_0) distribution is

$$H(f_{\mathrm{vM}}) = \ln(2\pi I_0(\kappa)) - \kappa I_1(\kappa)/I_0(\kappa),$$

where $I_1(\kappa)$ is the modified Bessel function of order 1. Its value $H(f_{\mathrm{vM}}) = 1.6274$, when $\kappa = 1$ and $\phi_0 = 0$. In Table 1, the values of entropies $\widehat{H}_n^{(k)}$, $k = 1, 2$, as well as the L_1-errors and MSEs are recorded for different sample sizes. From this table one can see that the behaviours of $\widehat{H}_n^{(k)}$, $k = 1, 2$ are similar to each other. We also conclude that the optimal $t = t^*$ is an increasing function of the sample size.

© 2011 Australian Statistical Publishing Association Inc.

TABLE 1

Comparison of $\widehat{H}_n^{(1)}$ and $\widehat{H}_n^{(2)}$: the von Mises distribution on $\mathbf{S}^1$.

	$\widehat{H}_n^{(1)}$				$\widehat{H}_n^{(2)}$			
n	t	L_1-error	MSE	$\widehat{H}_n^{(1)}$	t	L_1-error	MSE	$\widehat{H}_n^{(2)}$
50	0.4	0.0694	0.0072	1.6473	0.2	0.0678	0.0065	1.6497
100	0.3	0.0417	0.0029	1.6375	0.4	0.0421	0.0026	1.6416
500	0.9	0.0201	0.0007	1.6281	0.5	0.0219	0.0007	1.6411
1000	0.7	0.0138	0.0003	1.6305	0.8	0.0142	0.0003	1.6311

TABLE 2

Comparison of $\widehat{H}_n^{(1)}$ and $\widehat{H}_n^{(2)}$: the uniform distribution on $\mathbf{S}^2$.

	$\widehat{H}_n^{(1)}$				$\widehat{H}_n^{(2)}$			
n	t	L_1-error	MSE	$\widehat{H}_n^{(1)}$	t	L_1-error	MSE	$\widehat{H}_n^{(2)}$
50	0	0.0203	0.0007	2.5392	0.1	0.0212	0.0007	2.5443
100	0	0.0108	0.0002	2.5374	0.1	0.0107	0.0002	2.5383
500	0	0.0022	0.0000	2.5320	0.1	0.0022	0.0000	2.5322
1000	0	0.0011	0.0000	2.5316	0.1	0.0011	0.0000	2.5318

Uniform Distribution on $\mathbf{S}^2$: Consider a uniform distribution on a unit sphere $\mathbf{S}^2$ with $f(\boldsymbol{x}) = (4\pi)^{-1}$ for $\boldsymbol{x} \in \mathbf{S}^2$. The theoretical entropy in this case is $H(f) = \ln(4\pi) = 2.531$. We applied Lemma 1 for calculations of $\widehat{H}_n^{(k)}$, $k = 1, 2$. In Table 2 the values of the entropy estimates $\widehat{H}_n^{(k)}$, $k = 1, 2$, as well as the L_1-errors and MSEs are recorded for different sample sizes. From this table we conclude that the performances of $\widehat{H}_n^{(1)}$ and $\widehat{H}_n^{(2)}$ are similar to each other. In addition, we see that the optimal $t = t^*$ is not an increasing function of the sample size. This can be explained by the fact that f is a constant function on $\mathbf{S}^2$.

Bivariate von Mises Distribution on $[0, 2\pi)^2$: Let us define on $[0, 2\pi)^2$ a bivariate circular distribution (see Singh, Hnizdo & Demchuk 2002):

$$f_{\text{cir}}(\phi_1, \phi_2; \mu_1, \mu_2, \kappa_1, \kappa_2, \lambda) = \frac{1}{4\pi^2 C}\exp\{\kappa_1 \cos(\phi_1 - \mu_1) + \kappa_2 \cos(\phi_2 - \mu_2) + 2\lambda \sin(\phi_1 - \mu_1)\sin(\phi_2 - \mu_2)\}.$$

Here

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

The entropy of f has the following form: $H(f) = 2\ln 2\pi + \ln C - D/C$, where

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

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

© 2011 Australian Statistical Publishing Association Inc.

TABLE 3

The bias and MSE of $\widehat{H}_{\mathbf{a}}$: the bivariate von Mises distribution on $[0, 2\pi]^2$.

n	α	Bias	MSE	$\widehat{H}_{\mathbf{a}}$
50	10	0.0280	0.0224	2.5628
100	15	0.0175	0.0110	2.5523
500	25	0.0035	0.0019	2.5383
1000	30	−0.0014	0.0013	2.5334

We simulated the samples from f_{cir} with the parameters $\kappa_1 = \kappa_2 = 2$, $\mu_1 = \mu_2 = 3\pi/2$, and $\lambda = 0.5$. To estimate the entropy $H(f_{\text{cir}}) = 2.5348$ we applied the formula presented in Remark 2. In all our calculations below we assume $\alpha = \alpha'$. Table 3 displays the various values of α along with the corresponding bias terms and MSEs of $\widehat{H}_{\mathbf{a}}$ for each sample size. Moreover, one can compare the optimal biases and MSEs of $\widehat{H}_{\mathbf{a}}$ with the corresponding terms of the k-NN entropy estimator proposed in Mnatsakanov *et al.* (2008) for the same distribution f_{cir}. Table 3 justifies that, for small samples, $\widehat{H}_{\mathbf{a}}$ has a better performance than the k-NN entropy estimator. Our simulations confirm that the convergence rate of $\widehat{H}_{\mathbf{a}}$ is of the order $n^{-1/2}$, as the values of the product $(n \text{ MSE})^{1/2}$ are almost constant: 1.058, 1.051, 0.962, 1.154, when the sample size $n = 50, 100, 500, 1000$, and the optimal $\alpha^* = 10, 15, 25, 30$, respectively. Note that for the k-NN entropy estimator of $H(f_{\text{cir}})$, the values of the product $(n \text{ MSE})^{1/2}$ were a little larger: 1.50, 1.30, 1.28, 1.49 when $n = 50, 500, 1000, 10000$, respectively (see also Mnatsakanov *et al.* 2008).

Dirichlet Distribution on $[0, 1]^2$: For distributions on a unit cube $[0, 1]^{k-1}$, we choose the kth-order Dirichlet distribution with shape parameter vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{k-1}, \gamma_k)^\top$, denoted here by Dir $(\boldsymbol{\gamma})$, and the corresponding pdf is given by

$$f_{\text{Dir}}(x_1, \ldots, x_{k-1}; \boldsymbol{\gamma}) = \frac{1}{B(\boldsymbol{\gamma})} \prod_{i=1}^{k} x_i^{\gamma_i - 1}, \quad \text{with} \quad \sum_{j=1}^{k-1} x_j < 1 \quad \text{and} \quad x_k = 1 - \sum_{j=1}^{k-1} x_j.$$

Here $B(\boldsymbol{\gamma}) = \prod_{j=1}^{k} \Gamma(\gamma_j) / \Gamma(\gamma_0)$ with $\gamma_0 = \sum_{j=1}^{k} \gamma_j$. It is straightforward to show that the $\mathbf{r}$th order geometric moment of f_{Dir} is equal to

$$\mu_{\mathbf{r}} = \frac{B(\boldsymbol{\gamma} + \mathbf{r})}{B(\boldsymbol{\gamma})}$$

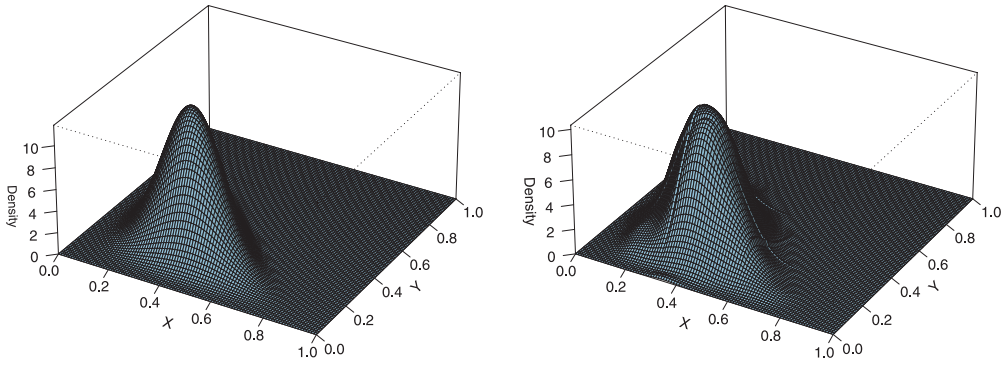
for any $\mathbf{r} = (r_1, \ldots, r_{k-1}, 0)^\top$. The theoretical entropy of f_{Dir} is given by

$$H(f_{\text{Dir}}) = \ln B(\boldsymbol{\gamma}) + (\gamma_0 - k)\psi(\gamma_0) - \sum_{j=1}^{k} (\gamma_j - 1)\psi(\gamma_j),$$

where $\psi(\cdot)$ is the digamma function.

Consider the pdf f_{Dir} when $k = 3$ and $\boldsymbol{\gamma} = (5, 5, 5)$. The graphs of f_{Dir} and its MR approximant $f_{\mathbf{a},\hat{\nu}}$ defined according to (12) and (13) (when $\alpha = \alpha' = 100$ and the sample size $n = 1000$) are shown in Figure 3(a) and (b) , respectively. From Figure 3 we see that the MR construction (12) works quite well in this model. The value of the theoretical Shannon entropy $H(f_{\text{Dir}}) = -1.638$. Table 4 below lists the values of biases and MSEs of $\widehat{H}_{\mathbf{a}}$ for different sample sizes. The values of biases and MSEs of $\widehat{H}_{\mathbf{a}}$ become smaller when

© 2011 Australian Statistical Publishing Association Inc.



(a) (b)

Figure 3. (a) Graph of the pdf f_{Dir} of the Dir (5, 5, 5) distribution and (b) graph of the moment-recovered pdf $f_{\mathbf{a},\widehat{\boldsymbol{\nu}}}$ based on empirical moments $\widehat{\boldsymbol{\nu}}$ with the sample size $n = 1000$ and $\alpha = \alpha' = 100$.

TABLE 4

The bias and MSE of $\widehat{H}_{\mathbf{a}}$: the Dirichlet Dir (5, 5, 5) distribution on $[0, 1]^2$.

n	α	Bias	MSE	$\widehat{H}_{\mathbf{a}}$
50	30	0.0152	0.0086	−1.6231
100	35	0.0191	0.0042	−1.6192
100	40	0.0127	0.0053	−1.6256
500	70	−0.0022	0.0012	−1.6405
500	75	−0.0045	0.0011	−1.6428
1000	85	−0.0009	0.0005	−1.6392

n increases. Furthermore, we see from Table 4 that by increasing the sample size from $n = 50$ to 1000, the corresponding values of optimal α^* (when $\alpha = \alpha'$) increase and can be specified as follows: $\alpha^* = 30, 40, 70, 85$, respectively. In addition, the values of the product $(n \text{ MSE})^{1/2}$ are: 0.66, 0.73, 0.78, 0.71 for $n = 50$, 100, 500, 1000, respectively. Hence, in this example the simulations confirm that the rate of convergence of MSE is of the order of $n^{-1/2}$.

5. Worked example

Researchers at NIOSH studied the bromo(chloromethyoxy) methane molecule structure, which has two torsional angles (ϕ_1 and ϕ_2), as shown in Figure 1. The torsional angle data of 1000 samples were obtained using MD simulation. The estimated pdf $f_{\mathbf{a},\widehat{\boldsymbol{\nu}}}$ of (ϕ_1, ϕ_2) is shown in Figure 4 , where we took $\alpha = \alpha' = 30$. The corresponding MR entropy estimate of $H(f)$ is calculated as well, and $\widehat{H}_{\mathbf{a}} = 2.54$.

6. Discussion and conclusion

In this paper, the asymptotic behaviours of three multivariate entropy estimators, $\widehat{H}_n^{(1)}$, $\widehat{H}_n^{(2)}$, and $\widehat{H}_{\mathbf{a}}$, introduced in (4), (11) and (13), respectively, are investigated. We assumed that

© 2011 Australian Statistical Publishing Association Inc.

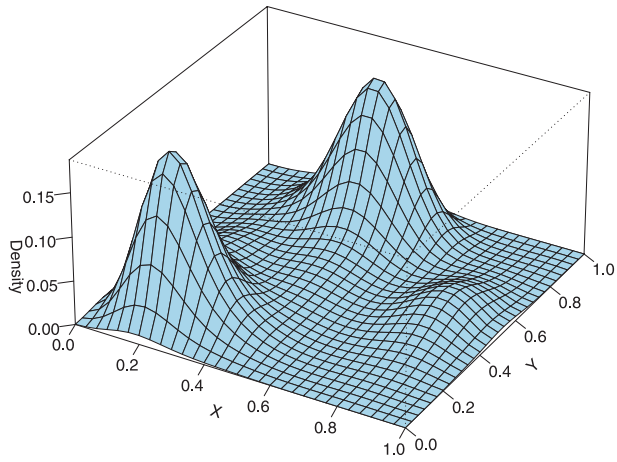


Figure 4. The estimated torsional angles distribution of bromo(chloromethyoxy) methane by $f_{\mathbf{a},\widehat{\mathbf{v}}}$ with $\alpha = \alpha' = 30$.

the support of the underlying distribution is a unit sphere or a unit cube in $\mathbb{R}^p$, $p \geq 2$. These constructions have several advantages, as follows.

(i) The performances of the estimators $\widehat{H}_n^{(1)}$ and $\widehat{H}_n^{(2)}$ are based on the density estimates $\hat{f}_{n,-i}^{(1)}$ and $\hat{f}_{n,-i}^{(2)}$, respectively, which depend on the choice of only one parameter, $t = t_n$. They provide very accurate precision compared with the entropy estimates based on the kernel density construction (it is well known that the kernel density approach fails to perform well in high-dimensional Euclidean space, e.g. Scott 1992).

(ii) Both estimates $\widehat{H}_n^{(k)}$, $k = 1, 2$, use the circular distance between the data points on the unit sphere (instead of the Euclidean one).

(iii) The entropy estimator $\widehat{H}_{\mathbf{a}}$ in (13) can be used (while other methods such as the kernel or k-NN entropy estimates can not) when the only available data represent the sequence of estimated moments of f.

(iv) The statement of Theorem 2 provides the upper bound for the L_1-error for $\widehat{H}_n^{(1)}$, when $p > 1$. The form of the bound depends on the smoothness of the density function f and on the constant M_n defined under the condition (A_3).

(v) $\widehat{H}_{\mathbf{a}}$ is preferable to the k-NN entropy estimator when f is observed directly for small sample sizes, but, for large sample sizes, these two approaches have similar performances.

The disadvantage of $\widehat{H}_{\mathbf{a}}$ is that the calculation of its values is time-consuming if the dimensionality p is very large. In addition to this, to have a 'good' performance for $\widehat{H}_{\mathbf{a}}$, the calculations should be conducted with a high accuracy when both components of $\mathbf{a} = (\alpha, \alpha')$ are large (>30), as shown in the simulation study.

Finally, we showed via simulation studies in several models that the bias and L_1- and L_2-errors of MR entropy estimators for the spherical data are similar to those based on the histogram, $\hat{f}_n^{(1)}$, and k-NN density estimators when the sample sizes are large enough. In addition, $\widehat{H}_n^{(1)}$ and $\widehat{H}_n^{(2)}$ can be used efficiently for sufficiently large values of p (see section 4.3 in Li *et al.* 2011).

© 2011 Australian Statistical Publishing Association Inc.

APPENDIX: The Proofs of Theorems

To prove Theorem 1, let us denote $\nu_{n-1,\boldsymbol{x}} = (n-1)\hat{P}_{n-1}(C_{\boldsymbol{x}}(t_n))$ and

$$\hat{f}^{(1)}_{n-1}(\boldsymbol{X}_i) = \hat{f}^{(1)}_{n,-i}(\boldsymbol{X}_i) = \frac{\nu_{n-1,X_i}}{(n-1)\lambda\big(C_{X_i}(t_n)\big)}.$$

Proof of Theorem 1. Assume

$$T_{i,n} = \ln\big(\hat{f}^{(1)}_{n,-i}(\boldsymbol{X}_i)\big) = \ln\left(\frac{\nu_{n-1,X_i}}{(n-1)\,\lambda\big(C_{X_i}(t_n)\big)}\right), \qquad i = 1, 2, \ldots, n,$$

so that

$$\widehat{H}^{(1)}_n = -\frac{1}{n}\sum_{i=1}^{n} T_{i,n}.$$

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

$$\mathrm{E}_f\big(\widehat{H}^{(1)}_n\big) = -\mathrm{E}_f(T_{1,n}).$$

For a fixed $\boldsymbol{x} \in \mathbf{S}^{p-1}$, let $S_{\boldsymbol{x},n}$ be a random variable having the same distribution as the conditional distribution of $T_{1,n}$ given $\mathbf{X}_1 = \boldsymbol{x}$. Let $F_{\boldsymbol{x},n}(\cdot)$ be the corresponding cdf of $S_{\boldsymbol{x},n}$. Then

$$\mathbf{E}_f(T_{1,n}) = \int_{\mathbf{S}^{p-1}} \mathbf{E}_f(S_{\boldsymbol{x},n}) f(\boldsymbol{x})\, d\lambda(\boldsymbol{x}).$$

Note that, for a fixed $\boldsymbol{x} \in \mathbf{S}^{p-1}$ and $u \in (-\infty, \infty)$,

$$F_{\boldsymbol{x},n}(u) = \mathbf{P}\big(\nu_{n-1,X_1} \le \mathrm{e}^u\ (n-1)\,\lambda(C_{\boldsymbol{x}_1}(t_n))\big|\boldsymbol{X}_1 = \boldsymbol{x}\big).$$

Therefore,

$$\begin{aligned} F_{\boldsymbol{x},n}(u) &= \mathbf{P}(\nu_{n-1,\boldsymbol{x}} \le \mathrm{e}^u\,(n-1)\,\lambda(C_{\boldsymbol{x}}(t_n))) \\ &= \mathbf{P}\left(Z_n \le \frac{(n-1)\,\lambda(C_{\boldsymbol{x}}(t_n))\left(\mathrm{e}^u - \dfrac{p_n(\boldsymbol{x})}{\lambda(C_{\boldsymbol{x}}(t_n))}\right)}{((n-1)\,p_n(\boldsymbol{x})(1-p_n(\boldsymbol{x})))^{1/2}}\right), \end{aligned}$$

where $\nu_{n-1,\boldsymbol{x}} \sim \mathrm{Bi}\,(n-1,\, p_n(\boldsymbol{x}))$, and

$$Z_n = \frac{\nu_{n-1,\boldsymbol{x}} - (n-1)\,p_n(\boldsymbol{x})}{\{(n-1)\,p_n(\boldsymbol{x})(1-p_n(\boldsymbol{x}))\}^{1/2}} \quad \text{with} \quad p_n(\boldsymbol{x}) = \int_{C_{\boldsymbol{x}}(t_n)} f(\mathbf{y})\, d\lambda(\mathbf{y}).$$

Note that, for almost all $\boldsymbol{x} \in \mathbf{S}^{p-1}$, $\lim_{n\to\infty} p_n(\boldsymbol{x}) = 0$. Applying Lemma 1 and condition (A_1), we have

$$\lim_{n\to\infty}((n-1)\lambda(C_{\boldsymbol{x}}(t_n))) = \infty,$$

© 2011 Australian Statistical Publishing Association Inc.

for almost all values of $\boldsymbol{x}$. Consequently, when n is sufficiently large and $u < \ln f(\boldsymbol{x})$, we have

$$e^u - \frac{p_n(\boldsymbol{x})}{\lambda(C_{\boldsymbol{x}}(t_n))} < 0,$$

for almost all values of $\boldsymbol{x}$. Now, on using lemma 2.1 (i) from Mnatsakanov *et al.* (2008) it follows that, for almost all values of $\boldsymbol{x}$, $F_{\boldsymbol{x},n}(u) \to 0$, as $n \to \infty$. Similarly, for each fixed $u > \ln f(\boldsymbol{x})$, we have $F_{\boldsymbol{x},n}(u) \to 1$, as $n \to \infty$, for almost all values of $\boldsymbol{x}$. Hence, for almost all values of $\boldsymbol{x}$, the limiting ($n \to \infty$) distribution of $S_{\boldsymbol{x},n}$ is degenerate at $\ln f(\boldsymbol{x})$. Furthermore, application of Lebesgue's dominant convergence theorem (see Billigsley 1995, p. 209) on the right-hand side of

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

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

$$F_T(u) = \int I\{u > \ln f(\boldsymbol{x})\}\, f(\boldsymbol{x})\, d\lambda(\boldsymbol{x}) = 1 - Q_f(e^u).$$

The function $Q_f(\cdot) = \int I\{f(\boldsymbol{x}) > \cdot\} f(\boldsymbol{x})\, d\lambda(\boldsymbol{x})$ on the right-hand side of the previous equation represents the so-called Q-structural function of f. Using the properties of Q_f (see Khmaladze 1988), we derive

$$\mathrm{E}(T) = \int u\, dF_T(u) = -\int \ln \tau\, dQ_f(\tau) = \int f(\boldsymbol{x}) \ln f(\boldsymbol{x})\, d\lambda(\boldsymbol{x}) = -H(f).$$

To prove Theorem 1, it remains to show that

$$-\lim_{n\to\infty} \mathrm{E}(T_{1,n}) = -\mathrm{E}(T) = H(f). \tag{14}$$

Under the conditions of Theorem 1, it follows that there exist an $\varepsilon > 0$ and a constant C_1 (not depending on n) such that, for all sufficiently large values of n,

$$\mathrm{E}\left(|T_{1,n}|^{1+\varepsilon}\right) < C_1.$$

See also the proofs of theorems 2.1 and 2.2, and the appendix in Mnatsakanov *et al.* (2008). Consequently, applying the corollary of theorem 25.12 in Billingsley (1995, p. 338), we derive (14). By a similar argument to the one applied in Mnatsakanov *et al.* (2008) we can show as well that $\mathrm{var}\{H_n^{(1)}\} \to 0$.

To prove Theorem 2, we find the following notations helpful:

$$\begin{aligned}
I_{1n} &= \mathrm{E}\left(\left|\int_{\mathbf{S}^{p-1}} \ln\left(\hat{f}_{n-1}^{(1)}(\boldsymbol{x})\right) d\hat{P}_n(\boldsymbol{x}) - \int_{\mathbf{S}^{p-1}} \ln\left(\mathrm{E}\hat{f}_{n-1}^{(1)}(\boldsymbol{x})\right) d\hat{P}_n(\boldsymbol{x})\right|\right)\\
I_{2n} &= \mathrm{E}\left(\left|\int_{\mathbf{S}^{p-1}} \ln\left(\mathrm{E}\hat{f}_{n-1}^{(1)}(\boldsymbol{x})\right) d\hat{P}_n(\boldsymbol{x}) - \int_{\mathbf{S}^{p-1}} \ln f(\boldsymbol{x})\, d\hat{P}_n(\boldsymbol{x})\right|\right)\\
I_{3n} &= \mathrm{E}\left(\left|\int_{\mathbf{S}^{p-1}} \ln f(\boldsymbol{x})\, d(\hat{P}_n(\boldsymbol{x}) - P(\boldsymbol{x}))\right|\right).
\end{aligned} \tag{15}$$

© 2011 Australian Statistical Publishing Association Inc.

Proof of Theorem 2. Since

$$\widehat{H}_n^{(1)} = -\frac{1}{n}\sum_{i=1}^{n} \ln\left(\hat{f}_{n,-i}^{(1)}(\boldsymbol{X}_i)\right) = -\int_{\mathbf{S}^{p-1}} \ln\left(\hat{f}_{n-1}^{(1)}(\boldsymbol{x})\right) d\hat{P}_n(\boldsymbol{x}),$$

one can see that

$$\rho_n \le I_{1n} + I_{2n} + I_{3n}. \tag{16}$$

Condition (A_2) yields

$$\begin{aligned} I_{3n} &= \mathrm{E}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(\ln f(\boldsymbol{X}_i) - \mathrm{E}\ln f(\boldsymbol{X}_i))\right|\right) \le \left(\mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n}\ln f(\boldsymbol{X}_i)\right)\right)^{1/2} \\ &= \left(\frac{1}{n}\mathrm{var}\left(\ln f(\boldsymbol{X}_1)\right)\right)^{1/2} \le \left(\frac{B_2}{n}\right)^{1/2}. \end{aligned} \tag{17}$$

Now let us consider I_{2n}. Application of the Taylor expansion formula for $\ln u$, that is,

$$\ln u = \ln c + (u-c)\ \ln'(\theta\, c + (1-\theta)\, u))\ \text{ for some }\ 0 < \theta < 1,$$

with $u = \mathrm{E}(\hat{f}_{n-1}^{(1)}(\boldsymbol{x}))$ and $c = f(\boldsymbol{x})$ leads to

$$\left|\ln\left(\mathrm{E}\left(\hat{f}_{n-1}^{(1)}(\boldsymbol{x})\right)\right) - \ln(f(\boldsymbol{x}))\right| \le \frac{\left|\mathrm{E}\left(\hat{f}_{n-1}^{(1)}(\boldsymbol{x})\right) - f(\boldsymbol{x})\right|}{\theta f(\boldsymbol{x}) + (1-\theta)\, E\left(\hat{f}_{n-1}^{(1)}(\boldsymbol{x})\right)}$$

which is valid for any $\boldsymbol{x} \in \mathbf{S}^{p-1}$. Hence, we derive

$$\begin{aligned} I_{2n} &\le \mathrm{E}\left(\int_{\mathbf{S}^{p-1}} \frac{\left|\mathrm{E}\left(\hat{f}_{n-1}^{(1)}(\boldsymbol{x})\right) - f(\boldsymbol{x})\right|}{\theta\, f(\boldsymbol{x})}\, d\,\hat{P}_n(\boldsymbol{x})\right) \\ &= \frac{1}{\theta}\,\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\left|\mathrm{E}\left(\hat{f}_{n-1}^{(1)}(\boldsymbol{X}_i)\right) - f(\boldsymbol{X}_i)\right|}{f(\boldsymbol{X}_i)}\right) = \frac{1}{\theta}E\left(\frac{\left|\mathrm{E}\left(\hat{f}_{n-1}^{(1)}(\boldsymbol{X}_1)\right) - f(\boldsymbol{X}_1)\right|}{f(\boldsymbol{X}_1)}\right) \\ &= \frac{1}{\theta}\int_{\mathbf{S}^{p-1}} \mathrm{E}\left(\frac{\left|\mathrm{E}\left(\hat{f}_{n-1}^{(1)}(\boldsymbol{X}_1)\right) - f(\boldsymbol{X}_1)\right|}{f(\boldsymbol{X}_1)}\,\middle|\, \boldsymbol{X}_1 = \boldsymbol{x}\right) f(\boldsymbol{x})\, d\lambda(\boldsymbol{x}) \\ &= \frac{1}{\theta}\int_{\mathbf{S}^{p-1}} \left|\mathrm{E}\left(\hat{f}_{n-1}^{(1)}(\boldsymbol{x})\right) - f(\boldsymbol{x})\right| d\lambda(\boldsymbol{x}). \end{aligned} \tag{18}$$

Now from (2) we have

$$\left|E\left(\hat{f}_{n-1}^{(1)}(\boldsymbol{x})\right) - f(\boldsymbol{x})\right| = \left|\frac{P(C_x(t_n))}{\lambda(C_x(t_n))} - f(\boldsymbol{x})\right| = \left|\frac{1}{\lambda(C_x(t_n))}\int_{C_x(t_n)} [f(\boldsymbol{y}) - f(\boldsymbol{x})] d\lambda(\boldsymbol{y})\right|. \tag{19}$$

Combining (18) and (19) we derive

$$I_{2n} \le \frac{1}{\theta}\,\Delta(f, t_n). \tag{20}$$

© 2011 Australian Statistical Publishing Association Inc.

Using an argument similar to the one used in (18), one can easily find

$$I_{1n} \le \mathrm{E}\int_{\mathbf{S}^{p-1}} \frac{\left|\hat{f}_{n-1}^{(1)}(\boldsymbol{x}) - \mathrm{E}\big(\hat{f}_{n-1}^{(1)}(\boldsymbol{x})\big)\right|}{\theta\,\mathrm{E}\big(\hat{f}_{n-1}^{(1)}(\boldsymbol{X})\big)}\, d\,\hat{P}_n(\boldsymbol{X}) = \frac{1}{\theta}\mathrm{E}\left\{\frac{\left|\hat{f}_{n-1}^{(1)}(\boldsymbol{X}_1) - \mathrm{E}\big(\hat{f}_{n-1}^{(1)}(\boldsymbol{X}_1)\big)\right|}{\mathrm{E}\hat{f}_{n-1}^{(1)}(\boldsymbol{X}_1)}\right\}$$

$$= \frac{1}{\theta}\int_{\mathbf{S}^{p-1}} \mathrm{E}\left\{\frac{\left|\hat{f}_{n-1}^{(1)}(\boldsymbol{X}_1) - \mathrm{E}\big(\hat{f}_{n-1}^{(1)}(\boldsymbol{X}_1)\big)\right|}{\mathrm{E}\big(\hat{f}_{n-1}^{(1)}(\boldsymbol{X}_1)\big)}\,\middle|\, \boldsymbol{X}_1 = \boldsymbol{x}\right\} f(\boldsymbol{x})\, d\lambda(\boldsymbol{x}). \tag{21}$$

Note also that the integrand in (21) is an integrable function because

$$\mathrm{E}\left\{\frac{\left|\hat{f}_{n-1}^{(1)}(\boldsymbol{X}_1) - \mathrm{E}\big(\hat{f}_{n-1}^{(1)}(\boldsymbol{X}_1)\big)\right|}{\mathrm{E}\big(\hat{f}_{n-1}^{(1)}(\boldsymbol{X}_1)\big)}\,\middle|\, \boldsymbol{X}_1 = \boldsymbol{x}\right\} \le \frac{\mathrm{E}\big(\hat{f}_{n-1}^{(1)}(\boldsymbol{x})\big) + \mathrm{E}\big(\hat{f}_{n-1}^{(1)}(\boldsymbol{x})\big)}{\mathrm{E}\big(\hat{f}_{n-1}^{(1)}(\boldsymbol{x})\big)} = 2,$$

On the other hand, using the inequality

$$\mathrm{E}\big(\hat{f}_{n-1}^{(1)}(\boldsymbol{x})\big) = \frac{P(C_{\boldsymbol{x}}(t_n))}{\lambda(C_{\boldsymbol{x}}(t_n))} = \frac{1}{\lambda(C_{\boldsymbol{x}}(t_n))}\int_{C_{\boldsymbol{x}}(t_n)} f(\mathbf{y})\,\lambda(\mathbf{y}) \ge B_1 > 0$$

and the result from theorem 3.1 in Ruymgaart (1989):

$$\sup_{\boldsymbol{x}\in\mathbf{S}^{p-1}} M_n \left|\hat{f}_{n-1}^{(1)}(\boldsymbol{x}) - \mathrm{E}\big(\hat{f}_{n-1}^{(1)}(\boldsymbol{x})\big)\right| \to 0 \ \text{ a.s.},$$

we conclude that for almost all $\boldsymbol{x} \in \mathbf{S}^{p-1}$ and sufficiently large M_n defined according to (A_3):

$$\mathrm{E}\left\{\frac{\left|\hat{f}_{n-1}^{(1)}(\boldsymbol{X}_1) - \mathrm{E}\big(\hat{f}_{n-1}^{(1)}(\boldsymbol{X}_1)\big)\right|}{\mathrm{E}\big(\hat{f}_{n-1}^{(1)}(\boldsymbol{X}_1)\big)}\,\middle|\, \boldsymbol{X}_1 = \boldsymbol{x}\right\} \le o(1/M_n)$$

as $t_n \to 1$, $n \to \infty$. Hence,

$$I_{1n} \le o(1/M_n) \quad \text{as} \quad t_n \to 1, n \to \infty. \tag{22}$$

The combination of (15)–(17), (20) and (22) leads to the proof of Theorem 2.

References

BEIRLANT, J., DUDEWICZ, E.J., GYÖRFI, L. & VAN DER MEULEN, E.C. (1997). Nonparametric entropy estimation: an overview. *Inter. J. Math. Statist. Sci.* **6**, 17–39.

BILLINGSLEY, P. (1995). *Probability and Measure*. New York: John Wiley and Sons, Inc.

GYÖFRI, L. & VAN DER MEULEN, E.C. (1987). Density-free convergence properties of various estimators of entropy. *Comput. Statist. Data Anal.* **5**, 425–436.

GZYL, H. & TAGLIANI, A. (2010). Hausdorff moment problem and fractional moments. *Appl. Math. Comput.* **216**, 3319–3328.

HALL, P. & MORTON, S. C. (1993). On the estimation of entropy. *Ann. Inst. Statist. Math.* **45**, 69–88.

JOE, H. (1989). Estimation of entropy and other functionals of a multivariate density. *Ann. Inst. Statist. Math.* **41**, 683–697.

KHMALADZE, E.V. (1988). *The Statistical Analysis of a Large Number of Rare Events*. Technical Report MS-R8804, Center for Mathematics and Computer Science, Amsterdam.

KOZACHENKO, L.F. & LEONENKO, N.N. (1987). Sample estimates of entropy of a random vector. *Problems Inform. Transmission* **23**, 95–101.

LI, S. (2011). Concise formulas for the area and volume of a hyperspherical cap. *Asian J. Math. Statist.* **4**, 66–70.

© 2011 Australian Statistical Publishing Association Inc.

LI, S., MNATSAKNOV, R.M. & ANDREW, M.E. (2011). k-Nearest neighbor based consistent entropy estimation for hyperspherical distributions. *Entropy* **13**, 650–667.

MISRA, N., SINGH, H. & HNIZDO, V. (2010). Nearest neighbor estimates of entropy for multivariate circular distributions. *Entropy* **12**, 1125–1144.

MNATSAKANOV, R.M. (2008a). Hausdorff moment problem: Reconstruction of distributions. *Statist. Probab. Lett.* **78**, 1612–1218.

MNATSAKANOV, R.M. (2008b). Hausdorff moment problem: Reconstruction of probability density functions. *Statist. Probab. Lett.* **78**, 1869–1877.

MNATSAKANOV, R.M. (2011). Moment-recovered approximations of multivariate distributions: The Laplace transform inversion. *Statist. Probab. Lett.* **81**, 1–7.

MNATSAKANOV, R.M. & LI, S. (2010). The Radon transform inversion using moments. (Manuscript)

MNATSAKANOV, R. & RUYMGAART, F.H. (2003). Some properties of moment-empirical cdf's with application to some inverse estimation problems. *Math. Methods Statist.* **12**, 478–495.

MNATSAKANOV, R., MISRA, N., LI, SH. & HARNER, E.J. (2008). k_n-Nearest neighbor estimators of entropy. *Math. Methods Statist.* **17**, 261–277.

RUYMGAART, F.H. (1989). Strong uniform convergence of density estimators on spheres. *J. Statist. Plann. Inference.* **23**, 45–52.

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

SHOHAT, J.A. & TAMARKIN, J.D. (1943). *The Problem of Moments*. New York: American Mathematical Society.

SINGH, H., HNIZDO, V. & DEMCHUK, E. (2002). Probabilistic model for two dependent circular variables. *Biometrika* **89**, 719–723.

SINGH, H., MISRA, N., HNIZDO, V., FEDOROWICZ, A. & DEMCHUK, E. (2003). Nearest neighbor estimates of entropy. *Amer. J. Math. Management Sci.* **23**, 301–321.

TSYBAKOV, A.B. & VAN DER MEULEN, E.C. (1996). Root-n consistent estimators of entropy for densities with unbounded support. *Scand. J. Statist.* **23**, 75–83.

© 2011 Australian Statistical Publishing Association Inc.