Entropy **2011**, *13*, 650-667; doi:10.3390/e13030650

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

k-Nearest Neighbor Based Consistent Entropy Estimation for Hyperspherical Distributions

Shengqiao Li [1,⋆], Robert M. Mnatsakanov [1,2,⋆] and Michael E. Andrew [1]

[1] Health Effects Laboratory Division, National Institute for Occupational Safety and Health,
Morgantown, WV 26505, USA; E-Mail: mta6@cdc.gov

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

⋆ Authors to whom correspondence should be addressed; E-Mails: Shengqiao.Li@cdc.hhs.gov (S.L.);
rmnatsak@stat.wvu.edu (R.M.).

Received: 22 December 2010; in revised form: 27 January 2011 / Accepted: 28 February 2011 / Published: 8 March 2011

Abstract: A consistent entropy estimator for hyperspherical data is proposed based on the k-nearest neighbor (knn) approach. The asymptotic unbiasedness and consistency of the estimator are proved. Moreover, cross entropy and Kullback-Leibler (KL) divergence estimators are also discussed. Simulation studies are conducted to assess the performance of the estimators for models including uniform and von Mises-Fisher distributions. The proposed knn entropy estimator is compared with the moment based counterpart via simulations. The results show that these two methods are comparable.

Keywords: hyperspherical distribution; directional data; differential entropy; cross entropy; Kullback-Leibler divergence; k-nearest neighbor

1. Introduction

The Shannon (or differential) entropy of a continuously distributed random variable (r.v.) $\boldsymbol{X}$ with probability density function (*pdf*) f is widely used in probability theory and information theory as a measure of uncertainty. It is defined as the negative mean of the logarithm of the density function, *i.e.*,

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

k-Nearest neighbor (knn) density estimators were proposed by Mack and Rosenblatt [1]. Penrose and Yukich [2] studied the laws of large numbers for k-nearest neighbor distances. The nearest neighbor entropy estimators when $\boldsymbol{X} \in \mathbb{R}^p$ were studied by Kozachenko and Leonenko [3]. Singh *et al.* [4] and Leonenko *et al.* [5] extended these estimators using k-nearest neighbors. Mnatsakanov *et al.* [6] studied knn entropy estimators for variable rather than fixed k. Eggermontet *et al.* [7] studied the kernel entropy estimator for univariate smooth distributions. Li *et al.* [8] studied parametric and nonparametric entropy estimators for univariate multimodal circular distributions. Neeraj *et al.* [9] studied knn estimators of circular distributions for the data from the Cartesian product, that is, $[0, 2\pi)^p$. Recently, Mnatsakanov *et al.* [10] proposed an entropy estimator for hyperspherical data based on the moment-recovery (MR) approach (see also Section 4.3).

In this paper, we propose k-nearest neighbor entropy, cross-entropy and KL-divergence estimators for hyperspherical random vectors defined on a unit p-hypersphere $\mathbb{S}^{p-1}$ centered at the origin in p-dimensional Euclidean space. Formally,

$$\mathbb{S}^{p-1} = \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\| = 1\} \tag{2}$$

The surface area S_p of the hypersphere is well known: $S_p = \frac{2\pi^{p/2}}{\Gamma(\frac{p}{2})}$, where Γ is the gamma function. For a part of the hypersphere, the area of a cap with solid angle ϕ relative to its pole is given by Li [11] (*cf.* Gray [12]):

$$S(\phi) = \tfrac{1}{2} S_p \left[1 - \mathrm{sgn}(\cos\phi) I_{\cos^2\phi}\left(\tfrac{1}{2}, \tfrac{p-1}{2}\right)\right] \tag{3}$$

where sgn is the sign function, and $I_x(\alpha, \beta)$ is the regularized incomplete beta function.

For a random vector from the unit circle $\mathbb{S}^1$, the von Mises distribution $\mathrm{vM}(\boldsymbol{\mu}, \kappa)$ is the most widely used model:

$$f_{\mathrm{vM}}(\boldsymbol{x}; \boldsymbol{\mu}, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \boldsymbol{\mu}^T \boldsymbol{x}}$$

where T is the transpose operator, $||\boldsymbol{\mu}|| = 1$ and $\kappa \geq 0$ are the mean direction vector and concentration parameters, and I_0 is the zero-order modified Bessel function of the first kind. Note that the von Mises distribution has a single mode. The multimodal extension to the von Mises distribution is the so-called generalized von Mises model. Its properties are studied by Yfantis and Borgman [13] and Gatto and Jammalamadaka [14].

The generalization of von Mises distribution onto $\mathbb{S}^{p-1}$ is the von Mises-Fisher distribution (also known as Langevin distribution) $\mathrm{vMF}_p(\boldsymbol{\mu}, \kappa)$ with *pdf*,

$$f_p(\boldsymbol{x}; \boldsymbol{\mu}, \kappa) = c_p(\kappa) e^{\kappa \boldsymbol{\mu}^T \boldsymbol{x}} \tag{4}$$

where the normalization constant is

$$c_p(\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa)}$$

and $I_\nu(x)$ is the ν-order modified Bessel function of the first kind. See Mardia and Jupp [15] (p. 167) for details.

Since von Mises-Fisher distributions are members of the exponential family, by differentiating the cumulant generating function, one can obtain the mean and variance of $\boldsymbol{\mu}^T \boldsymbol{X}$:

$$E_f[\boldsymbol{\mu}^T \boldsymbol{X}] = A_p(\kappa)$$

and

$$V_f[\boldsymbol{\mu}^T\boldsymbol{X}] = A_p'(\kappa)$$

where $A_p(\kappa) = I_{p/2}(\kappa)/I_{p/2-1}(\kappa)$, and $A_p'(\kappa) = \frac{d}{d\kappa}A_p(\kappa) = 1 - A_p(\kappa)^2 - (p-1)/\kappa A_p(\kappa)$. See Watamori [16] for details. Thus the entropy of f_p is:

$$H(f_p) = -E_f[\ln f_p(\boldsymbol{X})] = -\ln c_p(\kappa) - \kappa E_f[\boldsymbol{\mu}^T\boldsymbol{X}] = -\ln c_p(\kappa) - \kappa A_p(\kappa) \tag{5}$$

and

$$V_f[\ln f_p(\boldsymbol{X})] = \kappa^2 V_f[\boldsymbol{\mu}^T\boldsymbol{X}] = \kappa^2 A_p'(\kappa) \tag{6}$$

Spherical distributions have been used to model the orientation distribution functions (ODF) in HARDI (High Angular Resolution Diffusion Imaging). Knutsson [17] proposed a mapping from ($p = 3$) orientation to a continuous and distance preserving vector space ($p = 5$). Rieger and Vilet [18] generalized the orientation in any p-dimensional spaces. McGraw *et al.* [19] used vMF$_3$ mixture to model the 3-D ODF and Bhalerao and Westin [20] applied vMF$_5$ mixture to 5-D ODF in the mapped space. Entropy of the ODF is proposed as a measure of anisotropy (Özarslan *et al.* [21], Leow *et al.* [22]). McGraw *et al.* [19] used Rényi entropy for the vMF$_3$ mixture since it has a closed form. Leow *et al.* [22] proposed an exponential isotropy measure based on the Shannon entropy. In addition, KL-divergence can be used to measure the closeness of two ODF's. A nonparametric entropy estimator based on knn approach for hyperspherical data provides an easy way to compute the entropy related quantities.

In Section 2, we will propose the knn based entropy estimator for hyperspherical data. The unbiasedness and consistency are proved in this section. In Section 3, the knn estimator is extended to estimate cross entropy and KL-divergence. In Section 4, we present simulation studies using uniform hyperspherical distributions and aforementioned vMF probability models. In addition, the knn entropy estimator is compared with the MR approach proposed in Mnatsakanov *et al.* [10]. We conclude this study in Section 5.

2. Construction of knn Entropy Estimators

Let $\boldsymbol{X} \in \mathbb{S}^{p-1}$ be a random vector having *pdf* f and $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a set of *i.i.d.* random vectors drawn from f. To measure the nearness of two vectors $\boldsymbol{x}$ and $\mathbf{y}$, we define a distance measure as the angle between them: $\phi = \arccos(\boldsymbol{x}^T\mathbf{y})$ and denote the distance between $\boldsymbol{X}_i$ and its k-th nearest neighbor in the set of n random vectors by $\phi_i := \phi_{n,k,i}$.

With the distance measure defined above and without loss of generality, the naïve k-nearest neighbor density estimate at $\boldsymbol{X}_i$ is thus,

$$f_n(\boldsymbol{X}_i) = \frac{k/n}{S(\phi_i)} \tag{7}$$

where $S(\phi_i)$ is the cap area as expressed by (3).

Let $L_{n,i}$ be the natural logarithm of the density estimate at $\boldsymbol{X}_i$,

$$L_{n,i} = \ln f_n(\boldsymbol{X}_i) = \ln\frac{k/n}{S(\phi_i)} \tag{8}$$

and thus we construct a similar k-nearest neighbor entropy estimator (*cf.* Singh *et al.* [4]):

$$H_n(f) = -\frac{1}{n}\sum_{i=1}^{n}[L_{n,i} - \ln k + \psi(k)] = \frac{1}{n}\sum_{i=1}^{n}\ln[nS(\phi_i)] - \psi(k) \tag{9}$$

where $\psi(k) = \frac{\Gamma'(k)}{\Gamma(k)}$ is the digamma function.

In the sequel, we shall prove the asymptotic unbiasedness and consistency of $H_n(f)$.

2.1. *Unbiasedness of* H_n

To prove the asymptotic unbiasedness, we first introduce the following lemma:

Lemma 2.1. *For a fixed integer $k < n$, the asymptotic conditional mean of $L_{n,i}$ given $\boldsymbol{X}_i = \boldsymbol{x}$, is*

$$E[\lim_{n\to\infty} L_{n,i}|\boldsymbol{X}_i = \boldsymbol{x}] = \ln f(\boldsymbol{x}) + \ln k - \psi(k) \tag{10}$$

Proof. $\forall \ell \in \mathbb{R}$, consider the conditional probability

$$\begin{aligned} P\{L_{n,i} < \ell|\boldsymbol{X}_i = \boldsymbol{x}\} &= P\{f_n(\boldsymbol{X}_i) < e^{\ell}|\boldsymbol{X}_i = \boldsymbol{x}\} \\ &= P\{S(\phi_i) > \frac{k}{n}e^{-\ell}\} \end{aligned} \tag{11}$$

Equation (11) implies that there are at most k samples falling within the cap C_i centered at $\boldsymbol{X}_i = \boldsymbol{x}$ with area $S_{c_i} = \frac{k}{n}e^{-\ell}$.

If we let

$$p_{n,i} = \int_{C_i} f(\boldsymbol{y})\, d\boldsymbol{y}$$

and $Y_{n,i}$ be the number of samples falling onto the cap C_i, then $Y_{n,i} \sim BIN(n,\, p_{n,i})$, is a binomial random variable. Therefore,

$$P\{L_{n,i} < \ell|\boldsymbol{X}_i = \boldsymbol{x}\} = P\{Y_{n,i} < k\}$$

If we let $\frac{k}{n} \to 0$ as $n \to \infty$, then $p_{n,i} \to 0$ as $n \to \infty$. It is reasonable to consider the Poisson approximation of $Y_{n,i}$ with mean $\lambda_{n,i} = np_{n,i} = \frac{ke^{-\ell}}{S_{c_i}}p_{n,i}$. Thus, the limiting distribution of $Y_{n,i}$ is a Poisson distribution with mean:

$$\lambda_i = \lim_{n\to\infty} \lambda_{n,i} = ke^{-\ell} \lim_{n\to\infty} \frac{p_{n,i}}{S_{c_i}} = ke^{-\ell} f(\boldsymbol{x}) \tag{12}$$

Define a random variable L_i having the conditional cumulative density function,

$$F_{L_i,\boldsymbol{x}}(\ell) = \lim_{n\to\infty} P\{L_{n,i} < \ell|\boldsymbol{X}_i = \boldsymbol{x}\}$$

then

$$F_{L_i,\boldsymbol{x}}(\ell) = \sum_{j=0}^{k-1} \frac{[kf(\boldsymbol{x})e^{-\ell}]^j}{j!} e^{-kf(\boldsymbol{x})e^{-\ell}}$$

By taking derivative w.r.t. ℓ, we obtain the conditional *pdf* of L_i:

$$f_{L_i,\boldsymbol{x}}(\ell) = \frac{[kf(\boldsymbol{x})e^{-\ell}]^k}{(k-1)!} e^{-kf(\boldsymbol{x})e^{-\ell}} \tag{13}$$

The conditional mean of L_i is

$$E[L_i|\boldsymbol{X}_i = \boldsymbol{x}] = \int_{-\infty}^{\infty} \ell \cdot \frac{[kf(\boldsymbol{x})e^{-\ell}]^k}{(k-1)!} e^{-kf(\boldsymbol{x})e^{-\ell}}\, d\ell$$

By change of variable, $z = kf(\boldsymbol{x})e^{-\ell}$,

$$\begin{aligned} E[L_i|\boldsymbol{X}_i = \boldsymbol{x}] &= \int_0^{\infty} [\ln f(\boldsymbol{x}) + \ln k - \ln z] \frac{z^{k-1}}{(k-1)!} e^{-z}\, dz \\ &= \ln f(\boldsymbol{x}) + \ln k - \int_0^{\infty} \ln z \frac{z^{k-1}}{(k-1)!} e^{-z}\, dz \\ &= \ln f(\boldsymbol{x}) + \ln k - \psi(k) \end{aligned} \tag{14}$$

□

Corollary 2.2. *Given $\boldsymbol{X}_i = \boldsymbol{x}$, let $\eta_{n,k,\boldsymbol{x}} := nS(\phi_i) = ke^{-L_{n,i}}$, then $\ln \eta_{n,k,\boldsymbol{x}} = \ln k - L_{n,i}$ converges in distribution to $\ln \eta_{k,\boldsymbol{x}} = \ln k - L_i$, and*

$$E[\ln \eta_{k,\boldsymbol{x}}] = -\ln f(\boldsymbol{x}) + \psi(k)$$

Moreover, $\eta_{k,\boldsymbol{x}}$ is a gamma r.v. with the shape parameter k and the rate parameter $f(\boldsymbol{x})$.

Theorem 2.3. *If a* pdf *f satisfies the following conditions: for some $\epsilon > 0$,*
$(A_1): \quad \int_{\mathbb{S}^{p-1}} |\ln f(\boldsymbol{x})|^{1+\epsilon} f(\boldsymbol{x})\, d\boldsymbol{x} < \infty,$
$(A_2): \quad \int_{\mathbb{S}^{p-1}} \int_{\mathbb{S}^{p-1}} \left|\ln[1 - I_{(\boldsymbol{x}^T\boldsymbol{y})^2}(\frac{1}{2}, \frac{p-1}{2})]\right|^{1+\epsilon} f(\boldsymbol{x}) f(\boldsymbol{y})\, d\boldsymbol{x} d\boldsymbol{y} < \infty,$
then the estimator proposed in (9) is asymptotically unbiased.

Proof. According to Corollary 2.2 and condition (A_2), we can show (see (16)–(22)) that for almost all values of $\boldsymbol{x} \in \mathbb{S}^{p-1}$, there exists a positive constant C such that
$(i) \quad E[|\ln \eta_{n,k,\boldsymbol{x}}|^{1+\epsilon}] < C$ for all sufficiently large n.
Hence, applying the moment convergence theorem [23] (p. 186), it follows that

$$\lim_{n\to\infty} E[\ln \eta_{n,k,\boldsymbol{x}}] = E[\ln \eta_{k,\boldsymbol{x}}] = -\ln f(\boldsymbol{x}) + \psi(k)$$

for almost all values of $\boldsymbol{x} \in \mathbb{S}^{p-1}$. In addition, using Fatou's lemma and condition (A_1), we have that

$$\begin{aligned} &\limsup_{n\to\infty} \int_{\mathbb{S}^{p-1}} |E\big[\ln \eta_{n,k,\boldsymbol{x}}\big]|^{1+\epsilon} f(\boldsymbol{x})\, d\boldsymbol{x} \\ &\leq \int_{\mathbb{S}^{p-1}} \limsup_{n\to\infty} |E\big[\ln \eta_{n,k,\boldsymbol{x}}\big]|^{1+\epsilon} f(\boldsymbol{x})\, d\boldsymbol{x} \\ &= \int_{\mathbb{S}^{p-1}} |-\ln f(\boldsymbol{x}) + \psi(k)|^{1+\epsilon} f(\boldsymbol{x})\, d\boldsymbol{x} \\ &\leq C_\epsilon \left(\int_{\mathbb{S}^{p-1}} |-\ln f(\boldsymbol{x})|^{1+\epsilon} f(\boldsymbol{x})\, d\boldsymbol{x} + |\psi(k)|^{1+\epsilon} \right) < \infty \end{aligned}$$

where C_ϵ is a constant. Therefore,

$$\begin{aligned}\lim_{n\to\infty} E[H_n(f)] &= \lim_{n\to\infty} E_f[\ln(nS(\phi_i))] - \psi(k)\\ &= \lim_{n\to\infty} \int_{\mathbb{S}^{p-1}} E\big[\ln \eta_{n,k,\boldsymbol{x}}\big] f(\boldsymbol{x})\,d\boldsymbol{x} - \psi(k)\\ &= \int_{\mathbb{S}^{p-1}} \lim_{n\to\infty} E\big[\ln \eta_{n,k,\boldsymbol{x}}\big] f(\boldsymbol{x})\,d\boldsymbol{x} - \psi(k)\\ &= \int_{\mathbb{S}^{p-1}} E\big[\ln \eta_{k,\boldsymbol{x}}\big] f(\boldsymbol{x})\,d\boldsymbol{x} - \psi(k)\\ &= \int_{\mathbb{S}^{p-1}} [-\ln f(\boldsymbol{x}) + \psi(k)] f(\boldsymbol{x})\,d\boldsymbol{x} - \psi(k)\\ &= H(f)\end{aligned}$$

To show (i), one can follow the arguments similar to those used in the proof of Theorem 1 in [24]. Indeed, we can first establish

(ii) $\quad E[|\ln \eta_{2,1,\boldsymbol{x}}|^{1+\epsilon}] < C.$

Namely, we justify that (i) is valid when $n = 2$ and $k = 1$. But the inequality (ii) follows immediately from the condition (A_2) and

$$\begin{aligned}E\Big[\big|\ln \eta_{2,1,\boldsymbol{x}}\big|^{1+\epsilon}\Big] &= E\Big[\big|\ln[2S(\phi_{1,2})]\big|^{1+\epsilon} | \boldsymbol{X}_1 = \boldsymbol{x}\Big]\\ &= E\Big[\big|\ln\big(S_p\big[1 - \mathrm{sgn}(\boldsymbol{x}^T\boldsymbol{X}_2) I_{(\boldsymbol{x}^T\boldsymbol{X}_2)^2}(\tfrac{1}{2}, \tfrac{p-1}{2})\big]\big)\big|^{1+\epsilon}\Big]\\ &\le C_\epsilon |\ln S_p|^{1+\epsilon} + C_\epsilon |\ln 2|^{1+\epsilon} + C_\epsilon E_f\Big[\big|\ln\big[1 - I_{(\boldsymbol{x}^T\boldsymbol{X}_2)^2}(\tfrac{1}{2}, \tfrac{p-1}{2})\big]\big|^{1+\epsilon} \mathbf{1}(\boldsymbol{x}^T\boldsymbol{X}_2 > 0)\Big]\\ &= C_\epsilon\big(|\ln S_p|^{1+\epsilon} + |\ln 2|^{1+\epsilon}\big) + \frac{1}{2} C_\epsilon E_f\Big[\big|\ln\big[1 - I_{(\boldsymbol{x}^T\boldsymbol{X}_2)^2}(\tfrac{1}{2}, \tfrac{p-1}{2})\big]\big|^{1+\epsilon}\Big] \qquad (15)\end{aligned}$$

Here $\phi_{1,2} = \arccos(\boldsymbol{X}_1^T\boldsymbol{X}_2)$ and $\mathbf{1}(\cdot)$ is the indicator function.

Now let us denote the distribution function of $\eta_{n,k,\boldsymbol{x}}$ by

$$\begin{aligned}G_{n,k,\boldsymbol{x}}(u) &= P(\eta_{n,k,\boldsymbol{x}} \le u) = P(nS(\phi_{n,k,1}) \le u | \boldsymbol{X}_1 = \boldsymbol{x})\\ &= 1 - \sum_{j=0}^{k-1} \binom{n-1}{j} \left(\int_{C_{\boldsymbol{x}}(\phi_n(u))} f(\boldsymbol{y})\,d\boldsymbol{y}\right)^j \left(1 - \int_{C_{\boldsymbol{x}}(\phi_n(u))} f(\boldsymbol{y})\,d\boldsymbol{y}\right)^{n-1-j}\end{aligned}$$

where $\phi_n(u) = S^{-1}(\frac{u}{n})$ and $C_x(\phi)$ is a cap $\{\boldsymbol{y} \in \mathbb{S}^{p-1} : \boldsymbol{y}^T\boldsymbol{x} \ge \cos\phi\}$ with the pole $\boldsymbol{x}$ and base radius $\sin\phi$. Note also that the functions $S(\phi)$ (see (3)) and $\phi_n(u) = S^{-1}(\frac{u}{n})$ are both increasing functions.

Now, one can see (*cf.* (66) in [24]):

$$E[|\ln \eta_{n,k,\boldsymbol{x}}|^{1+\epsilon}] \le I_1 + I_2 + I_3 \qquad (16)$$

where

$$\begin{aligned}I_1 &= (1+\epsilon) \int_0^1 (\ln \frac{1}{u})^\epsilon u^{-1} G_{n,k,\boldsymbol{x}}(u)\,du\\ I_2 &= (1+\epsilon) \int_1^{\sqrt{n}} (\ln u)^\epsilon u^{-1} (1 - G_{n,k,\boldsymbol{x}}(u))\,du\\ I_3 &= (1+\epsilon) \int_{\sqrt{n}}^{nS_p} (\ln u)^\epsilon u^{-1} (1 - G_{n,k,\boldsymbol{x}}(u))\,du\end{aligned}$$

It is easy to see that for sufficiently large n and almost all $\boldsymbol{x} \in \mathbb{S}^{p-1}$:

$$I_1 < (1+\epsilon) f(\boldsymbol{x}) \Gamma(1+\epsilon) < \infty \tag{17}$$

and

$$I_2 \leq (1+\epsilon) \sum_{j=0}^{k-1} [\sup_{\boldsymbol{y} \in \mathbb{S}^{p-1}} f(\boldsymbol{y})]^j f(\boldsymbol{x})^{-j-\epsilon} \Gamma(j+\epsilon) < \infty \tag{18}$$

(*cf.* (89) and (85) in [24], respectively).

Finally, let us show that $I_3 \to 0$ as $n \to \infty$. For each $\boldsymbol{x}$ with $f(\boldsymbol{x}) > 0$, if we choose a $\delta \in (0, f(\boldsymbol{x}))$, then for all sufficiently large n, $\sqrt{n} \int_{C_{\boldsymbol{x}}(\phi_n(\sqrt{n}))} f(\boldsymbol{y})\, d\boldsymbol{y} > f(\boldsymbol{x}) - \delta$, since the area of $C_{\boldsymbol{x}}(\phi_n(\sqrt{n}))$ is equal to $\frac{1}{\sqrt{n}}$. Using arguments similar to those used in (69)–(72) from [24], we have

$$\begin{aligned} I_3 \leq & (1+\epsilon) n^{k-1} k e^{-(n-k-1)(f(\boldsymbol{x})-\delta)\frac{1}{\sqrt{n}}} \\ & \times \int_{\sqrt{n}}^{nS_p} (\ln u)^\epsilon u^{-1} \left(1 - \int_{C_{\boldsymbol{x}}(\phi_n(u))} f(\boldsymbol{y})\, d\boldsymbol{y}\right) du \end{aligned} \tag{19}$$

The integral in (19) after changing the variable, $t = \frac{2u}{n}$, takes the form

$$\begin{aligned} & \int_{\frac{2}{\sqrt{n}}}^{2S_p} \left(\ln \frac{nt}{2}\right)^\epsilon t^{-1} (1 - G_{2,1,\boldsymbol{x}}(t))\, dt \\ & = \left(\int_{\frac{2}{\sqrt{n}}}^{1} + \int_{1}^{2S_p}\right) \left(\ln \frac{nt}{2}\right)^\epsilon t^{-1} (1 - G_{2,1,\boldsymbol{x}}(t))\, dt \end{aligned} \tag{20}$$

since $\phi_n(\frac{nt}{2}) = S^{-1}(\frac{t}{2}) = \phi_2(t)$ and $1 - \int_{C_{\boldsymbol{x}}(\phi_2(t))} f(\boldsymbol{y})\, d\boldsymbol{y} = 1 - G_{2,1,\boldsymbol{x}}(t))$. The first integral in the right side of (20) is bounded as follows:

$$\int_{\frac{2}{\sqrt{n}}}^{1} \left(\ln \frac{nt}{2}\right)^\epsilon t^{-1} (1 - G_{2,1,\boldsymbol{x}}(t))\, dt \leq \frac{\sqrt{n}}{2} \left(\ln \frac{n}{2}\right)^\epsilon \tag{21}$$

while for the second one, we have

$$\int_{1}^{2S_p} \left(\ln \frac{nt}{2}\right)^\epsilon t^{-1} (1 - G_{2,1,\boldsymbol{x}}(t))\, dt \leq C_\epsilon \left(\ln \frac{n}{2}\right)^\epsilon E[\eta_{2,1,\boldsymbol{x}}] + C_\epsilon \left(\ln \frac{n}{2}\right)^\epsilon B \tag{22}$$

where

$$B = \int_{1}^{2S_p} (\ln t)^\epsilon t^{-1} (1 - G_{2,1,\boldsymbol{x}}(t))\, dt = \frac{1}{1+\epsilon} E|\ln \eta_{2,1,\boldsymbol{x}}|^{1+\epsilon}$$

Combination of (15)–(22) and (ii) yields (i).

□

Remark. Note that

$$\begin{aligned} 1 - I_{t^2}(\tfrac{1}{2}, \tfrac{p-1}{2}) & \approx \tfrac{1}{B(\frac{1}{2}, \frac{p-1}{2})} (t^2)^{-\frac{1}{2}} (1-t^2)^{\frac{p-1}{2}} \\ & \approx \frac{2^{\frac{p-1}{2}}}{B(\frac{1}{2}, \frac{p-1}{2})} (1-t)^{\frac{p-1}{2}} \quad \text{as} \quad t \uparrow 1 \end{aligned}$$

where $B(\cdot, \cdot)$ is the beta function. Hence, in the conditions (A_j), $j = 2, 4, 6$ and 8, the difference $1 - I_{(\boldsymbol{x}^T\boldsymbol{y})^2}(\frac{1}{2}, \frac{p-1}{2})$ can be replaced by $1 - \boldsymbol{x}^T\boldsymbol{y}$.

2.2. Consistency of H_n

Lemma 2.4. *Under the following conditions: for some $\epsilon > 0$,*

$(A_3): \quad \int_{\mathbb{S}^{p-1}} |\ln f(\boldsymbol{x})|^{2+\epsilon} f(\boldsymbol{x})\, d\boldsymbol{x} < \infty,$

$(A_4): \quad \int_{\mathbb{S}^{p-1}} \int_{\mathbb{S}^{p-1}} \left|\ln[1 - I_{(\boldsymbol{x}^T\boldsymbol{y})^2}(\frac{1}{2}, \frac{p-1}{2})]\right|^{2+\epsilon} f(\boldsymbol{x}) f(\boldsymbol{y})\, d\boldsymbol{x} d\boldsymbol{y} < \infty,$

the asymptotic variance of $L_{n,i}$ is finite and equals $V_f[\ln f(\boldsymbol{X})] + \psi_1(k)$, where $\psi_1(k)$ is the trigamma function.

Proof. The conditions (A_3) and (A_4), and the argument similar to the one used in the proof of Theorem 2.3, yields

$$\lim_{n\to\infty} E[L_{n,i}^2 | \boldsymbol{X}_i = \boldsymbol{x}] = E[L_i^2 | \boldsymbol{X}_i = \boldsymbol{x}]$$

Therefore, it is sufficient to prove that $V_f[L_i] = V_f(\ln f(\boldsymbol{X})) + \psi_1(k)$. Similarly to (14), we have

$$\begin{aligned} E[L_i^2 | \boldsymbol{X}_i = \boldsymbol{x}] &= \int_0^\infty [\ln f(\boldsymbol{x}) + \ln k - \ln z]^2 \frac{z^{k-1}}{(k-1)!} e^{-z}\, dz \\ &= [\ln f(\boldsymbol{x}) + \ln k]^2 - 2[\ln f(\boldsymbol{x}) + \ln k]\psi(k) + \Gamma''(k)/\Gamma(k) \end{aligned} \tag{23}$$

Since $\Gamma''(k)/\Gamma(k) = \psi^2(k) + \psi_1(k)$,

$$E[L_i^2 | \boldsymbol{X}_i = \boldsymbol{x}] = [\ln f(\boldsymbol{x}) + \ln k - \psi(k)]^2 + \psi_1(k) \tag{24}$$

After some algebra, it can be shown that

$$\begin{aligned} V_f[L_i] &= E_f[(\ln f(\boldsymbol{X}))^2] - (E_f[\ln f(\boldsymbol{X})])^2 + \psi_1(k) \\ &= V_f[\ln f(\boldsymbol{X})] + \psi_1(k) \end{aligned} \tag{25}$$

□

Lemma 2.5. *For a fixed integer $k < n$, $L_{n,i}$ are asymptotically pairwise independent.*

Proof. For a pair of random variables $L_{n,i}$ and $L_{n,j}$ with $i \neq j$ and $\boldsymbol{X}_i \neq \boldsymbol{X}_j$, following the similar argument for Lemma 2.1, C_i and C_j shrink as n increases. Thus, it is safe to assume that C_i and C_j are disjoint for large n, and $L_{n,i}$ and $L_{n,j}$ are independent. Hence Lemma 2.5 follows.

□

Theorem 2.6. *Under the conditions (A_1) through (A_4), the variance of $H_n(f)$ decreases with sample size n, that is*

$$\lim_{n\to\infty} V_f[H_n(f)] = 0 \tag{26}$$

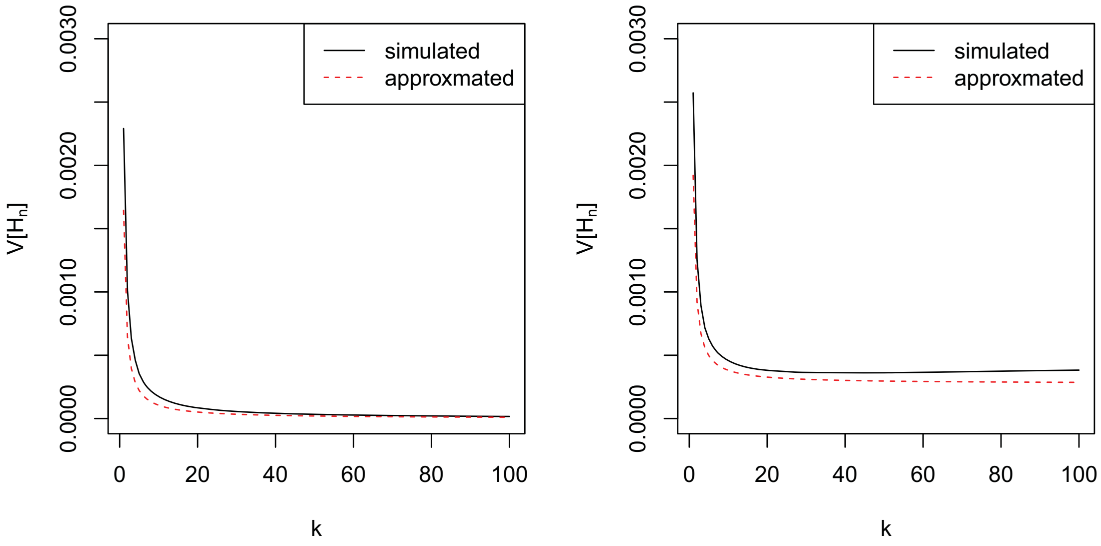
and $H_n(f)$ is a consistent estimator of $H(f)$.

Theorem 2.6 can be established by using Theorem 2.3 and Lemmas 2.4 and 2.5, and

$$\lim_{n\to\infty} V_f[H_n(f)] = \lim_{n\to\infty} \frac{1}{n}\{V_f[\ln f(\boldsymbol{X})] + \psi_1(k)\} = 0$$

For a finite sample, the variance of $H_n(f)$ can be approximated by $\frac{1}{n}\{V_f[\ln f(\boldsymbol{x})] + \psi_1(k)\}$. For instance, for the uniform distribution, $V_f[\ln f(\boldsymbol{x})] = 0$ and $V[H_n(f)] \approx \psi_1(k)/n$ and for a vMF$_p(\boldsymbol{\mu}, \kappa)$, $V[H_n(f)] \approx \frac{1}{n}[\kappa^2 A_p'(\kappa) + \psi_1(k)]$. See the illustration in Figure 1. The simulation was done with sample size $n = 1000$ and the number of simulations was $N = 10,000$. Since $\psi_1(k)$ is a decreasing function, the variance of $H_n(f)$ decreases when k increases.

Figure 1. Variances of $H_n(f)$ by simulation and approximation.



(a) uniform distribution

(b) vMF$_3(\mathrm{e}_3, 1)$

3. Estimation of Cross Entropy and KL-divergence

3.1. Estimation of Cross Entropy

The definition of cross entropy between continuous *pdf*'s f and g is,

$$H(f, g) = -\int f(\boldsymbol{x}) \ln g(\boldsymbol{x})\, d\boldsymbol{x} \tag{27}$$

Given a random sample of size n from f, $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$, and a random sample of size m from g, $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_m\}$, on a hypersphere, denote the knn density estimator of g by g_m. Similarly to (7),

$$g_m(\boldsymbol{X}_i) = \frac{k/m}{S(\varphi_i)} \tag{28}$$

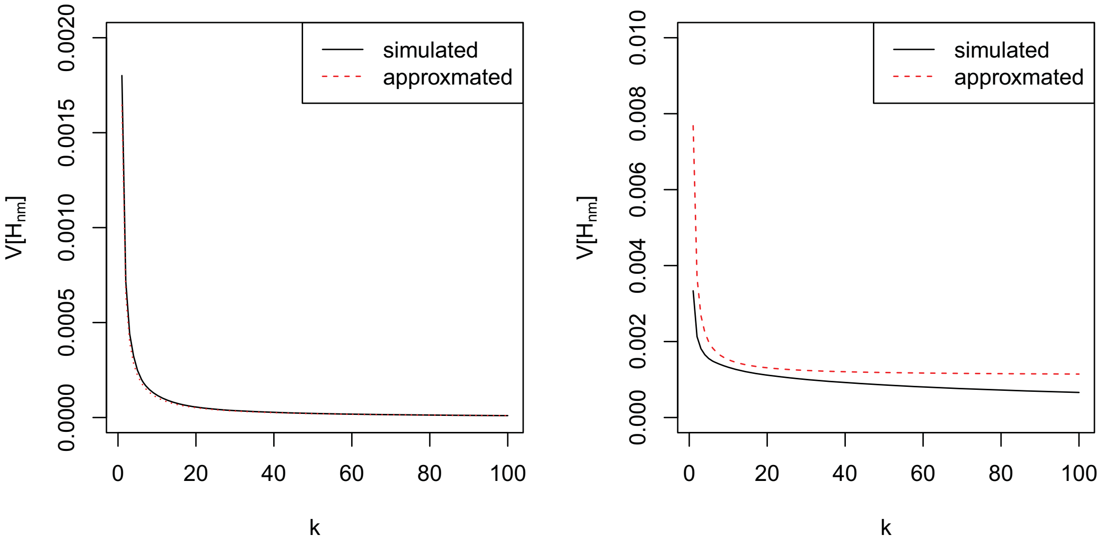
where φ_i is the distance from $\boldsymbol{X}_i$ to its k-th nearest neighbor in $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_m\}$. Analogously to the entropy estimator (9), the cross entropy can be estimated by:

$$H_{n,m}(f, g) = \frac{1}{n}\sum_{i=1}^{n} \ln S(\varphi_i) + \ln m - \psi(k) \tag{29}$$

Under the conditions (A_1)–(A_4), for a fixed integer $k < \min(n, m)$, one can show that $H_{n,m}(f, g)$ is asymptotically unbiased. Moreover, by similar reasoning applied for $H_n(f)$, one can show that $H_{n,m}(f, g)$ is also consistent and $V[H_{n,m}(f, g)] \approx \frac{1}{n}\{V_f[\ln g(\boldsymbol{x})] + \psi_1(k)\}$. For example, when both

f and g are vMF with the same mean direction and different concentration parameters, κ_1 and κ_2, respectively, the approximate variance will be $\frac{1}{n}[\kappa_2^2 A'_p(\kappa_1) + \psi_1(k)]$. Figure 2 shows the approximated and simulated variance of the knn estimators for cross entropy are close to each other and both decrease with k. The simulation is done with sample size $n = m = 1000$ and the number of simulations was $N = 10,000$.

Figure 2. Variances of $H_{n,m}$ by simulation and approximation.



(a) uniform distribution (b) $f = \text{vMF}_3(\text{e}_3, 1)$ and $g = \text{vMF}_3(\text{e}_3, 2)$

3.2. *Estimation of KL-Divergence*

KL-divergence is also known as relative entropy. It is used to measure the similarity of two distributions. Wang *et al.* [24] studied the knn estimator of KL-divergence for distributions defined on $\mathbb{R}^p$. Here we propose the knn estimator of KL-divergence of continuous distribution f from g defined on a hypersphere. The KL-divergence is defined as:

$$KL(f\|g) = E_f[\ln f(\boldsymbol{X})/g(\boldsymbol{X})] = \int f(\boldsymbol{x}) \ln \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}\, d\boldsymbol{x} \tag{30}$$

Equation (30) can also be expressed as $KL(f\|g) = H(f,g) - H(f)$. Then the knn estimator of KL-divergence is constructed as $H_{n,m}(f,g) - H_n(f)$, *i.e.*,

$$KL_{n,m}(f\|g) = \frac{1}{n}\sum_{i=1}^{n} \ln \frac{f_n(\boldsymbol{X}_i)}{g_m(\boldsymbol{X}_i)} = \frac{1}{n}\sum_{i=1}^{n} \ln \frac{S(\varphi_i)}{S(\phi_i)} + \ln \frac{m}{n} \tag{31}$$

where $g_m(\boldsymbol{X}_i)$ is defined as in (28). Besides, for finite samples, the variance of the estimator, $V[KL_{n,m}]$, is approximately $\frac{1}{n}\{V_f[\ln f(\boldsymbol{X})] + V_f[\ln g(\boldsymbol{X})] - 2\text{Cov}_f[\ln f(\boldsymbol{X}), \ln g(\boldsymbol{X})] + 2\psi_1(k)\}$. When f and g are vMF as mentioned above, with concentration parameter κ_1 and κ_2, respectively, we have:

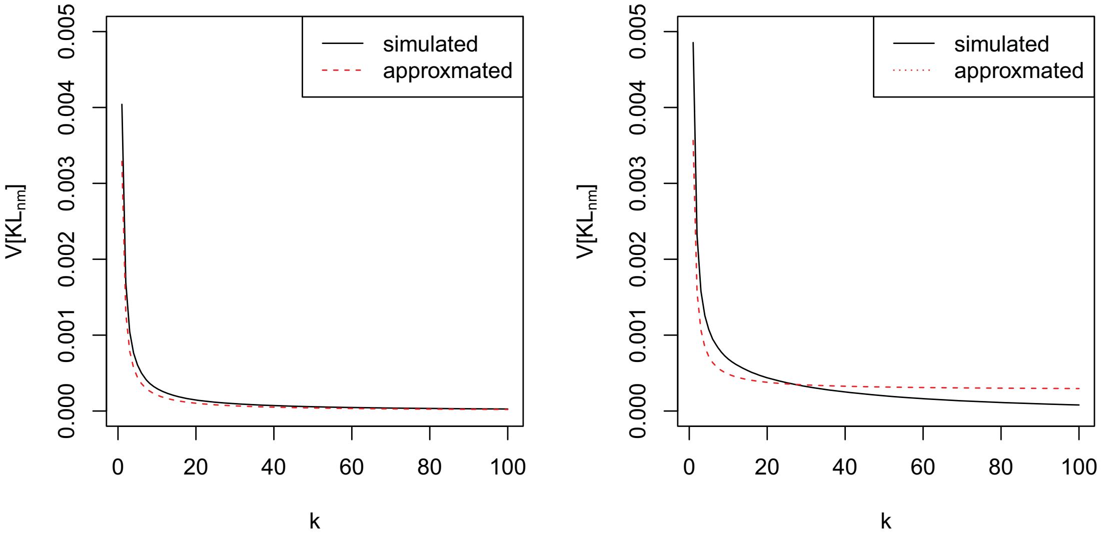
$$V_f[\ln f(\boldsymbol{X})] = \kappa_1^2 A'_p(\kappa_1)$$
$$V_f[\ln g(\boldsymbol{X})] = \kappa_2^2 A'_p(\kappa_1)$$

and

$$\mathrm{Cov}_f[\ln f(\boldsymbol{X}), \ln g(\boldsymbol{X})] = \kappa_1\kappa_2 A_p'(\kappa_1)$$

So the approximate variance is $\frac{1}{n}[(\kappa_1 - \kappa_2)^2 A_p'(\kappa_1) + 2\psi_1(k)]$. Figure 3 shows the approximated and simulated variance of the knn estimators for KL-divergence. The approximation for von Mises-Fisher distribution is not as good as the one for uniform distributions. This could be due to the modality of von Mises-Fisher distributions or the finitude of sample sizes. The larger the sample size, the closer the approximation is to the true value.

Figure 3. Variances of $KL_{n,m}$ by simulation and approximation.



(a) uniform distribution (b) $f = \mathrm{vMF}_3(\mathrm{e}_3, 1)$ and $g = \mathrm{vMF}_3(\mathrm{e}_3, 2)$

In summary, we have

Corollary 3.1. *(1) Under conditions* $(A_1), (A_2)$ *and for some* $\epsilon > 0$,

$(A_5):\quad \int_{\mathbb{S}^{p-1}} |\ln g(\boldsymbol{x})|^{1+\epsilon} f(\boldsymbol{x})\, d\boldsymbol{x} < \infty,$

$(A_6):\quad \int_{\mathbb{S}^{p-1}} \int_{\mathbb{S}^{p-1}} \left|\ln[1 - I_{(\boldsymbol{x}^T\boldsymbol{y})^2}(\frac{1}{2}, \frac{p-1}{2})]\right|^{1+\epsilon} f(\boldsymbol{x})g(\boldsymbol{y})\, d\boldsymbol{x}d\boldsymbol{y} < \infty,$

for a fixed integer $k < \min(n, m)$, *the knn estimator of KL-divergence given in* (31) *is asymptotically unbiased.*

(2) Under condition $(A_3), (A_4)$ *and for some* $\epsilon > 0$,

$(A_7):\quad \int_{\mathbb{S}^{p-1}} |\ln g(\boldsymbol{x})|^{2+\epsilon} f(\boldsymbol{x})\, d\boldsymbol{x} < \infty,$

$(A_8):\quad \int_{\mathbb{S}^{p-1}} \int_{\mathbb{S}^{p-1}} \left|\ln[1 - I_{(\boldsymbol{x}^T\boldsymbol{y})^2}(\frac{1}{2}, \frac{p-1}{2})]\right|^{2+\epsilon} f(\boldsymbol{x})g(\boldsymbol{y})\, d\boldsymbol{x}d\boldsymbol{y} < \infty,$

for a fixed integer $k < \min(n, m)$, *the knn estimator of KL-divergence given in* (31) *is asymptotically consistent.*

To prove the last two corollaries, one can follow the similar steps proposed in Wang *et al.* [24].

4. Simulation Study

To demonstrate the proposed knn entropy estimators and assess their performance for finite samples, we conducted simulations for the uniform distribution and von Mises-Fisher distributions with the

p-coordinate unit vector, $\mathbf{e}_p$, as the common mean direction for $p = 3$ and 10. For each distribution, we drew samples of size $n = 100$, 500 and 1000. All simulations were repeated $N = 10,000$ times. Bias, standard deviation (SD) and root mean squared error (RMSE) were calculated.

4.1. Bias and Standard Deviation

Figures 4–9 show simulated bias and standard deviation of the proposed entropy, cross-entropy and KL-divergence estimators along different k. The pattern for the standard deviation is clear. It decreases sharply then slowly as k increases. This is consistent with the variance approximations described in Sections 2 and 3. The pattern for bias is diverse. For uniform distributions, the bias term is very small. When the underlying distribution has a mode, for example, vMF models used in the current simulations, the relation between bias and k becomes complex and the bias term can be larger for larger k values.

Figure 4. $|\text{Bias}|$ (dashed line) and standard deviation (solid line) of entropy estimate H_n for uniform distributions.

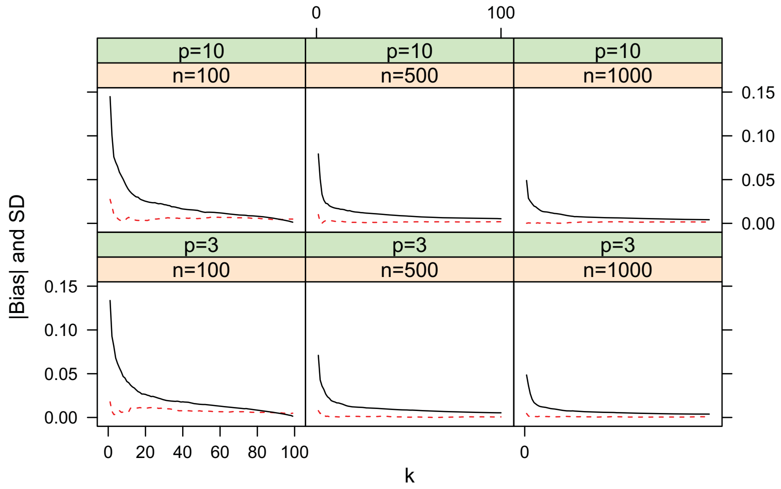


Figure 5. $|\text{Bias}|$ (dashed line) and standard deviation (solid line) of entropy estimate H_n for $\text{vMF}_p(\mathbf{e}_p, 1)$ distributions.

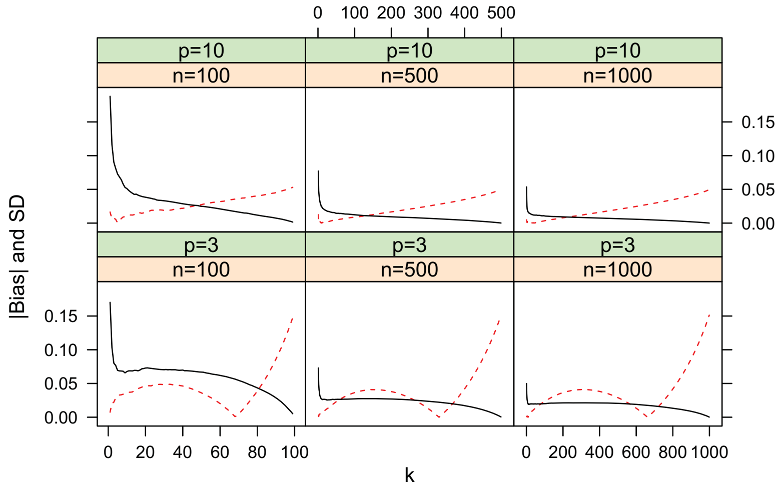


Figure 6. $|\text{Bias}|$ (dashed line) and standard deviation (solid line) of cross entropy estimate $H_{n,m}$ for uniform distributions.

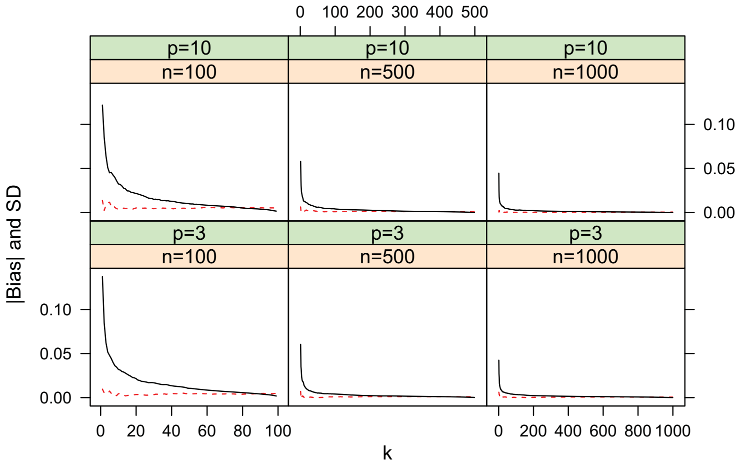


Figure 7. $|\text{Bias}|$ (dashed line) and standard deviation (solid line) of cross entropy estimate $H_{n,m}$ for $f = \text{vMF}_p(\mathbf{e}_p, 1)$ and $g =$ uniform distributions.

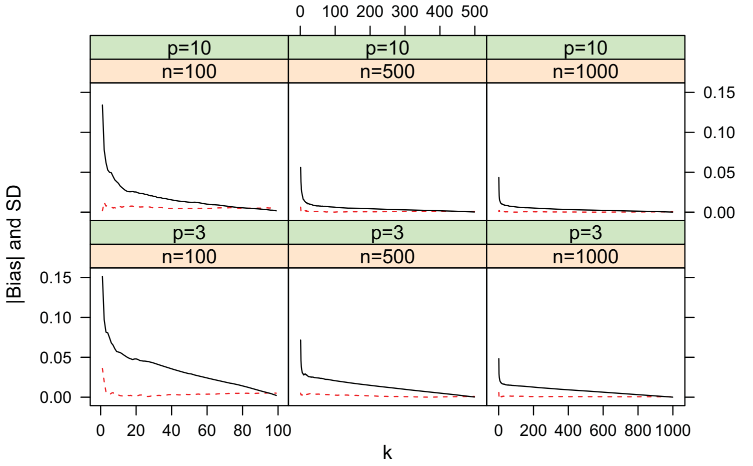


Figure 8. $|\text{Bias}|$ (dashed line) and standard deviation (solid line) of KL-divergence estimate $KL_{n,m}$ for uniform distributions.

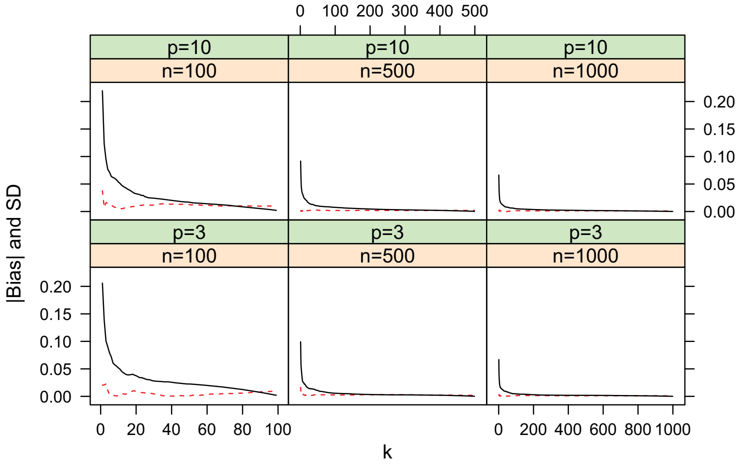
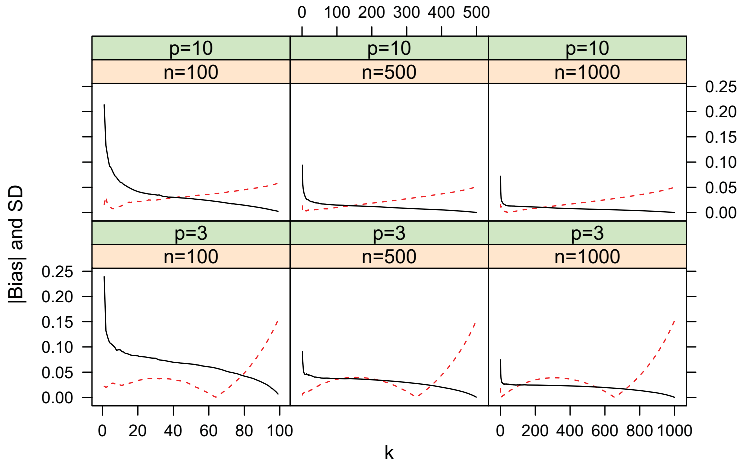


Figure 9. $|\text{Bias}|$ (dashed line) and standard deviation (solid line) of KL-divergence estimate $KL_{n,m}$ for $f = \text{vMF}_p(\mathbf{e}_p, 1)$ and $g =$ uniform distributions.



4.2. Convergence

To validate the consistency, we conducted simulations of different sample size n from 10 to 100,000 for the distribution models used above. Figures 10 and 11 shows the estimates and theoretical values of entropy, cross-entropy and KL-divergence for different sample sizes with $k = 1$ and $k = \lfloor \ln n + 0.5 \rfloor = 2$–12, respectively. The proposed estimators converge to the corresponding theoretical values quickly. Thus the consistency of these estimators are verified. The choice of k is an open problem for knn based estimation approaches. These figures show that using lager k, e.g., the logarithm of n, for lager n, is giving a slightly better preference.

Figure 10. Convergence of estimates with sample size n using the first nearest neighbor. For vMF_p, $\kappa = 1$.

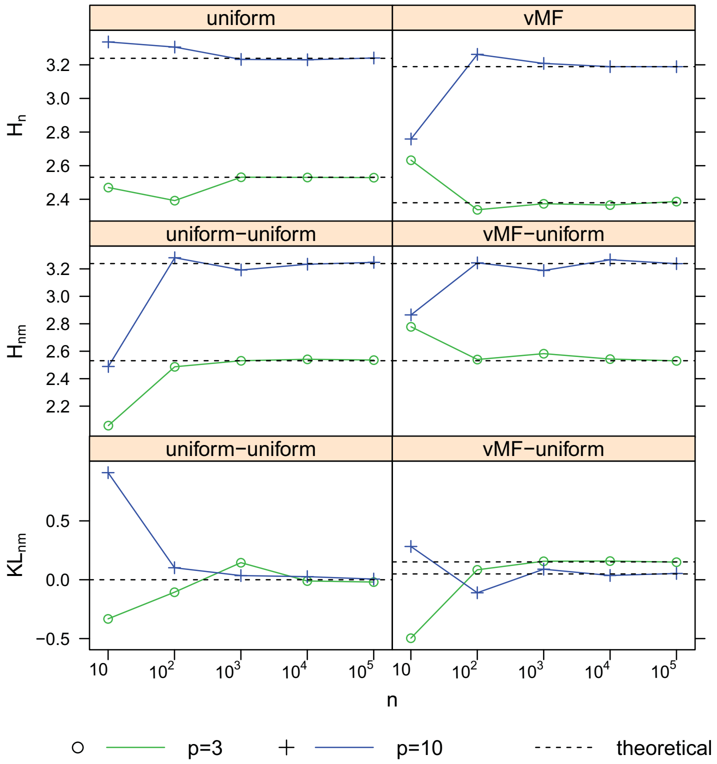
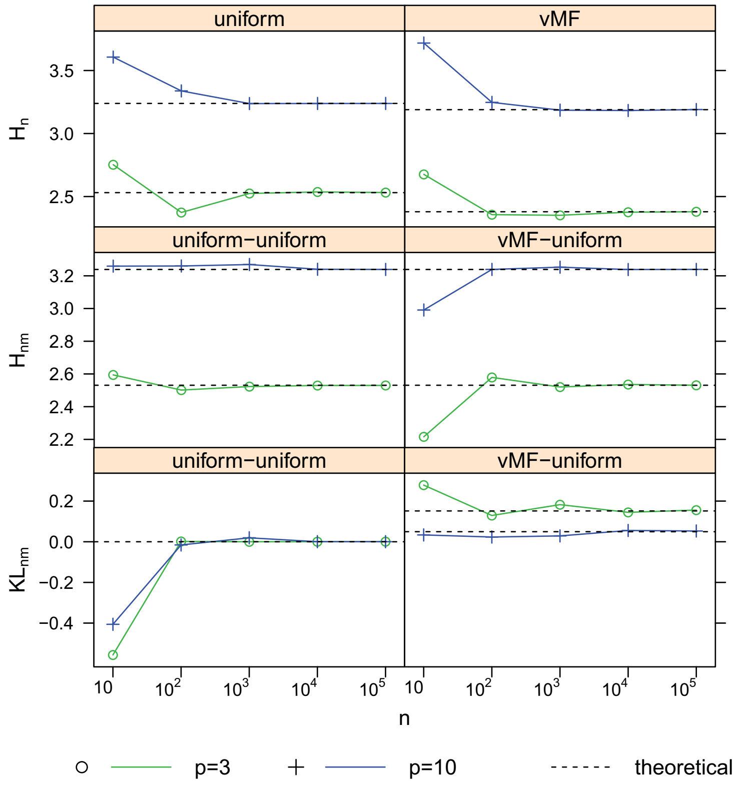


Figure 11. Convergence of estimates with sample size n using $k = \lfloor \ln n + 0.5 \rfloor$ nearest neighbors. For vMF$_p$, $\kappa = 1$.



4.3. Comparison with the Moment-Recovered Construction

Another entropy estimator for hyperspherical data was developed recently by Mnatsakanov *et al.* [10] using MR approach. We call this estimator the MR entropy estimator and denote it by $H_n^{(MR)}(f)$:

$$H_n^{(MR)}(f) = -\frac{1}{n} \sum_{i=1}^{n} \ln P_{n,t}(\boldsymbol{X}_i) + \ln S(\arccos t) \tag{32}$$

where $P_{n,t}(\boldsymbol{X}_i)$ is the estimated probability of the cap $\{\mathbf{y} \in \mathbb{S}^{p-1} : \mathbf{y}^T \boldsymbol{X}_i \geq t\}$ defined by the revolution axis $\boldsymbol{X}_i$ and t is the distance from the cap base to the origin and acts as a tuning parameter. Namely, (see Mnatsakanov *et al.* [10]),

$$P_{n,t}(\boldsymbol{X}_i) = \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} \sum_{k=\lfloor nt \rfloor+1}^{n} \binom{n}{k} (\boldsymbol{X}_j^T \boldsymbol{X}_i)^k (1 - \boldsymbol{X}_j^T \boldsymbol{X}_i)^{n-k} \tag{33}$$

Via simulation study, the empirical comparison between $H_n(f)$ and $H_n^{(MR)}(f)$ was done for the uniform and vMF distributions. The results are presented in Table 1. The values of k and t listed in the table are the optimal ones in the sense of minimizing RMSE. Z-tests and F-tests (at $\alpha = 0.05$) were performed to compare the bias, standard deviation (variance) and RMSE (MSE) between the knn estimators and corresponding MR estimators. In general, for uniform distributions, there are no significant difference for biases. Among other comparisons, the differences are significant. Specifically, knn achieves slightly smaller bias and RMSE values than those of the MR method. The standard deviations of knn method are also smaller for the uniform distribution but larger for vMF distributions than those based on MR approach.

Table 1. Comparison of knn and moment methods by simulations for spherical distributions.

Method			knn				MR			
	p	n	k	bias	SD	RMSE	t	bias	SD	RMSE
Uniform:										
	3	100	99	0.00500	0.00147	0.00521	0.01	0.00523	0.01188	0.01298
	3	500	499	0.00100	0.00013	0.00101	0.01	0.00107	0.00233	0.00257
	3	1000	999	0.00050	0.00005	0.00050	0.01	0.00051	0.00120	0.00130
	10	100	99	0.00503	0.00130	0.00520	0.01	0.00528	0.01331	0.01432
	10	500	499	0.00100	0.00011	0.00101	0.01	0.00102	0.00264	0.00283
	10	1000	999	0.00050	0.00004	0.00050	0.01	0.00052	0.00130	0.00140
vMF$_p(\boldsymbol{e}_p, 1)$:										
	3	100	71	0.01697	0.05142	0.05415	0.30	0.02929	0.04702	0.05540
	3	500	337	0.00310	0.02336	0.02356	0.66	0.00969	0.02318	0.02512
	3	1000	670	0.00145	0.01662	0.01668	0.74	0.00620	0.01658	0.01770
	10	100	46	0.02395	0.02567	0.03511	0.12	0.02895	0.02363	0.03737
	10	500	76	0.00702	0.01361	0.01531	0.40	0.01407	0.01247	0.01881
	10	1000	90	0.00366	0.01026	0.01089	0.47	0.01115	0.00907	0.01437

5. Discussion and Conclusions

In this paper, the knn based estimators for entropy, cross-entropy and Kullback-Leibler divergence are proposed for distributions on hyperspheres. Asymptotic properties such as unbiasedness and consistency are proved and validated by simulation studies using uniform and von Mises-Fisher distribution models. The variances of these estimators decrease with k. For uniform distributions, variance is dominant and bias is negligible. When the underlying distributions are modal, the bias can be large if k is large. In general, we conclude that the behavior of knn and MR entropy estimators have similar performance in terms of root mean square error.

Acknowledgements and Disclaimer

The authors thank the anonymous referees for their helpful comments and suggestions. The research of Robert Mnatsakanov was supported by NSF grant DMS-0906639. The findings and conclusions in this report are those of the author(s) and do not necessarily represent the views of the National Institute for Occupational Safety and Health.

References

1. Mack, Y.; Rosenblatt, M. Multivariate k-nearest neighbor density estimates. *J. Multivar. Anal.* **1979**, *9*, 1–15.

2. Penrose, M.D.; Yukich, J.E. Laws of large numbers and nearest neighbor distances. In *Advances in Directional and Linear Statistics*; Wells, M.T., SenGupta, A., Eds.; Physica-Verlag: Heidelberg, Germany, 2011; pp. 189–199.
3. Kozachenko, L.; Leonenko, N. On statistical estimation of entropy of a random vector. *Probl. Inform. Transm.* **1987**, *23*, 95–101.
4. Singh, H.; Misra, N.; Hnizdo, V.; Fedorowicz, A.; Demchuk, E. Nearest neighbor estimates of entropy. *Am. J. Math. Manag. Sci.* **2003**, *23*, 301–321.
5. Leonenko, N.; Pronzato, L.; Savani, V. A class of Rényi information estimators for multidimensional densities. *Ann. Stat.* **2008**, *36*, 2153–2182; Correction: **2010**, *38*, 3837–3838.
6. Mnatsakanov, R.; Misra, N.; Li, S.; Harner, E. k_n-Nearest neighbor estimators of entropy. *Math. Meth. Stat.* **2008**, *17*, 261–277.
7. Eggermont, P.P.; LaRiccia, V.N. Best asymptotic normality of the kernel density entropy estimator for smooth densities. *IEEE Trans. Inf. Theor.* **1999**, *45*, 1321–1326.
8. Li, S.; Mnatsakanov, R.; Fedorowicz, A.; Andrew, M.E. Entropy estimation of multimodal circular distributions. In Proceedings of Joint Statistical Meetings, Denver, CO, USA, 3–7 August 2008; pp. 1828–1835.
9. Misra, N.; Singh, H.; Hnizdo, V. Nearest neighbor estimates of entropy for multivariate circular distributions. *Entropy* **2010**, *12*, 1125–1144.
10. Mnatsakanov, R.M.; Li, S.; Harner, E.J. Estimation of multivariate Shannon entropies using moments. *Aust. N. Z. J. Stat.* **2011**, in press.
11. Li, S. Concise formulas for the area and volume of a hyperspherical cap. *Asian J. Math. Stat.* **2011**, *4*, 66–70.
12. Gray, A. *Tubes*, 2nd ed.; Birkhäuser-Verlag: Basel, Switzerland, 2004.
13. Yfantis, E.; Borgman, L. An extension of the von Mises distribution. *Comm. Stat. Theor. Meth.* **1982**, *11*, 1695–1076.
14. Gatto, R.; Jammalamadaka, R. The generalized von Mises distribution. *Stat. Methodol.* **2007**, *4*, 341–353.
15. Mardia, K.; Jupp, P. *Directional Statistics*; John Wiley & Sons, Ltd.: New York, NY, USA, 2000.
16. Watamori, Y. Statistical inference of Langevin distribution for directional data. *Hiroshima Math. J.* **1995**, *26*, 25–74.
17. Knutsson, H. Producing a continuous and distance preserving 5-D vector representation of 3-D orientation. In Proceedings of IEEE Computer Society Workshop on Computer Architecture for Pattern Analysis and Image Database Management, Miami Beach, FL, November, 1985; pp. 175–182.
18. Rieger, B.; van Vliet, L. Representing orientation in n-dimensional spaces. *Lect. Notes Comput. Sci.* **2003**, *2756*, 17–24.
19. McGraw, T.; Vemuri, B.; Yezierski, R.; Mareci, T. Segmentation of high angular resolution diffusion MRI modeled as a field of von Mises-Fisher mixtures. *Lect. Notes Comput. Sci.* **2006**, *3953*, 463–475.
20. Bhalerao, A.; Westin, C.F. Hyperspherical von Mises-Fisher mixture (HvMF) modelling of high angular resolution diffusion MRI. *Lect. Notes Comput. Sci.* **2007**, *4791*, 236–243.

21. Özarslan, E.; Vemuri, B.C.; Mareci, T.H. Generalized scalar measures for diffusion MRI using trace, variance, and entropy. *Magn. Reson. Med. Sci.* **2005**, *53*, 866–876.
22. Leow, A.; Zhu, S.; McMahon, K.; de Zubicaray, G.; Wright, M.; Thompson, P. A study of information gain in high angular resolution diffusion imaging (HARDI). In Proceedings of 2008 MICCAI Workshop on Computational Diffusion MRI, New York, NY, USA, 10 September 2008.
23. Loève, M. *Probability Theory I*, 4th ed.; Springer-Verlag: New York, NY, USA, 1977.
24. Wang, Q.; Kulkarni, S.R.; Verdú, S. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Trans. Inf. Theor.* **2009**, *55*, 2392–2405.