Efficient Nonparametric Entropy Estimation

Tom Berrett

University of Cambridge, England

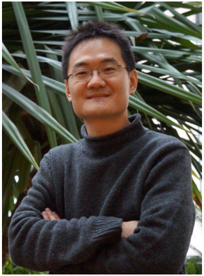
tbb26@cam.ac.uk

June 20, 2016

Collaborators

Material in talk based on joint work with

Richard Samworth



Ming Yuan

Definition of Entropy

The *differential entropy* of a random vector X with density function f is defined as

$$H(f) = -\int_{\mathcal{X}} f(x) \log f(x)\, dx = -\mathbb{E}[\log f(X)]$$

where $\mathcal{X} = \{x : f(x) > 0\}$. It is usually thought of as a measure of the unpredictability of X.

- Maximised among random variables on a compact set A of positive Lebesgue measure by the uniform distribution on A.
- For $\mathrm{N}(\mu, \sigma^2)$ random variables $H = (1/2) \log(2\pi e \sigma^2)$, the maximum possible for a distribution on $\mathbb{R}$ with fixed variance.

The above are examples of *maximum entropy distributions*.

Nearest Neighbour Distances

- Given $X_1, \ldots, X_n$ define the nearest neighbours $X_{(1),i}, \ldots, X_{(n-1),i}$ such that $\|X_{(1),i} - X_i\| \leq \ldots \leq \|X_{(n-1),i} - X_i\|$ and

$$\rho_{(k),i} = \|X_{(k),i} - X_i\|$$

- We use

$$k \approx (n-1)\rho_{(k),i}^d V_d f(X_i)$$

for $X_1, \ldots, X_n \overset{iid}{\sim} f$ to estimate $f(X_i)$.

- Kozachenko and Leonenko (1987) used this idea for the estimator

$$\hat{H}_n = \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{\rho_{(k),i}^d V_d (n-1)}{e^{\Psi(k)}}\right)$$

in the special case $k = 1$.

Bias Results

Let α be the number of moments of f. Then, under regularity conditions (including twice continuous differentiability):

- For $d \geq 3$ and $\alpha > 2d/(d-2)$ we have

$$\mathbb{E}(\hat{H}_n) - H = -\frac{\Gamma(k+2/d)}{2V_d^{2/d}(d+2)\Gamma(k)n^{2/d}} \int_{\mathbb{R}^d} \frac{\Delta f(x)}{f(x)^{2/d}}\, dx + o\left(\frac{k^{2/d}}{n^{2/d}}\right)$$

  as $n \to \infty$.

- If $d \leq 2$ or $d \geq 3$ and $\alpha \leq 2d/(d-2)$ we have

$$\mathbb{E}(\hat{H}_n) - H = o\left(\left(\frac{k}{n}\right)^{\alpha/(\alpha+d)-\tau}\right)$$

  as $n \to \infty$ for every $\tau > 0$.

Asymptotic Normality and Efficiency

Assume $d \leq 3$, k diverges with n (with restrictions) and regularity conditions on f. Then

$$n^{1/2}(\hat{H}_n - H) \xrightarrow{d} N\big(0, \operatorname{Var} \log f(X_1)\big),$$

and

$$n\mathbb{E}\{(\hat{H}_n - H)^2\} \to \operatorname{Var} \log f(X_1)$$

as $n \to \infty$.

- For $d \leq 3$ and appropriate k and regularity conditions the Kozachenko–Leonenko estimator can be efficient.
- For $d \geq 4$ a non-negligible bias prevents efficiency.

Efficiency in Higher Dimensions

For $w \in \mathbb{R}^k$ such that $\sum_{j=1}^k w_j = 1$ define the estimator

$$\hat{H}_n^w = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k w_j \log \xi_{(j),i},$$

where $\xi_{(j),i} = e^{-\Psi(j)} V_d (n-1) \rho_{(j),i}^d$, a weighted sum of previous estimators for different values of the tuning parameter. For $d \geq 3$, $\alpha > 2d/(d-2)$ and with restrictions on w we have

$$\mathbb{E}\hat{H}_n^w - H = -\frac{n^{-2/d}}{2V_d^{2/d}(d+2)} \left\{ \int_{\mathcal{X}} \frac{\Delta f(x)}{f(x)^{2/d}} \, dx \right\} \sum_{j=1}^k w_j \frac{\Gamma(j+2/d)}{\Gamma(j)} + o\left(\frac{k^{2/d}}{n^{2/d}} \right).$$

Looks like we can choose w to cancel out the leading bias.

Efficiency in Higher Dimensions

More generally, consider

$$\mathcal{W}_n := \Big\{ w \in \mathbb{R}^k : \sum_{j=1}^{k} w_j \frac{\Gamma(j + 2\ell/d)}{\Gamma(j)} = 0 \quad \text{for } \ell = 1, \ldots, \lfloor d/4 \rfloor$$

$$\sum_{j=1}^{k} w_j = 1 \text{ and } w_j = 0 \text{ if } j \notin \{\lfloor k/d \rfloor, \lfloor 2k/d \rfloor, \ldots, k\} \Big\}.$$

Then, for $w_n \in \mathcal{W}_n$ with $\sup_n \|w_n\| < \infty$ and regularity conditions on k and f (including $d/2$ times continuous differentiability)

$$n^{1/2}(\hat{H}_n^{w_n} - H) \xrightarrow{d} N\big(0, \mathrm{Var} \log f(X_1)\big),$$

and

$$n\mathbb{E}\{(\hat{H}_n^{w_n} - H)^2\} \to \mathrm{Var} \log f(X_1).$$

References

Berrett, T. B., Samworth, R. J. and Yuan, M. (2016). Efficient multivariate entropy estimation via k-nearest neighbour distances. http://arxiv.org/abs/1606.00304.

Biau, G. and Devroye, L. (2015) *Lectures on the Nearest Neighbor Method*. Springer, New York.

Delattre, S. and Fournier, N. (2016) On the Kozachenko–Leonenko entropy estimator. Available at `arXiv:1602.07440`.

Gao, W., Oh, S. and Viswanath, P. (2016) Demystifying fixed k-nearest neighbor information estimators. Available at `arXiv:1604.03006`.

Kozachenko, L. F. and Leonenko, N. N. (1987). Sample estimate of the entropy of a random vector. *Probl. Inform. Transm.*, **23**, 95–101.

Tsybakov, A. B. and Van der Meulen, E. C. (1996). Root-n consistent estimators of entropy for densities with unbounded support. *Scand. J. Stat.*, **23**, 75–83.