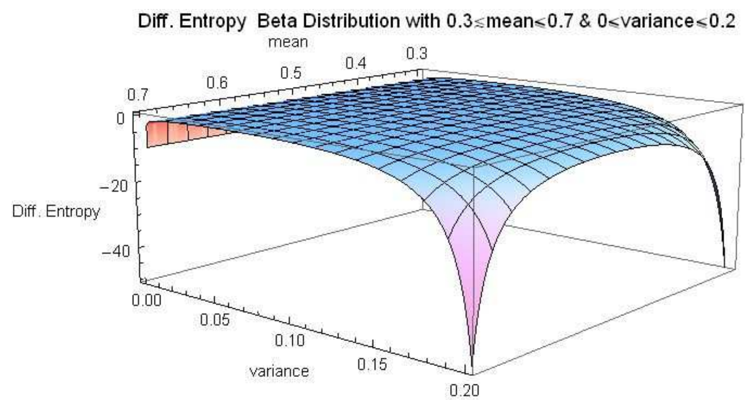
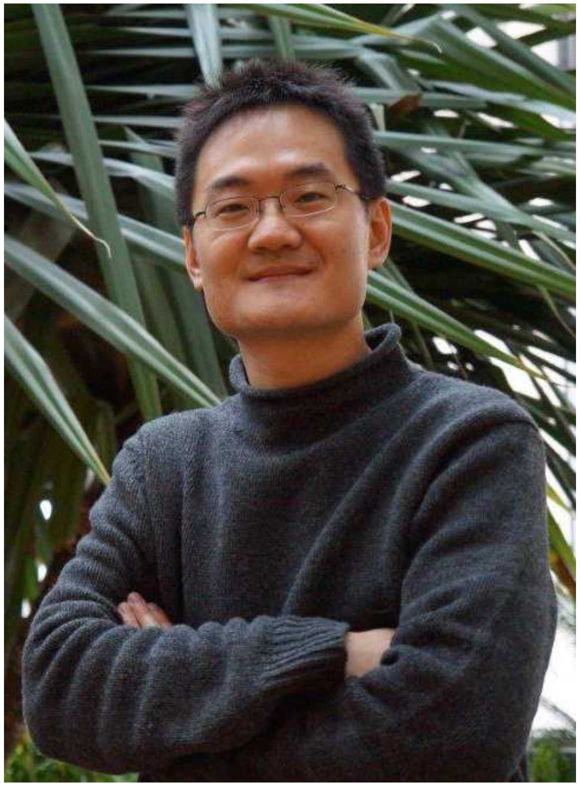
EFFICIENT MULTIVARIATE ENTROPY ESTIMATION VIA k-NEAREST NEIGHBOUR DISTANCES



Richard Samworth, University of Cambridge

Joint work with Thomas B. Berrett and Ming Yuan

Collaborators



Differential entropy

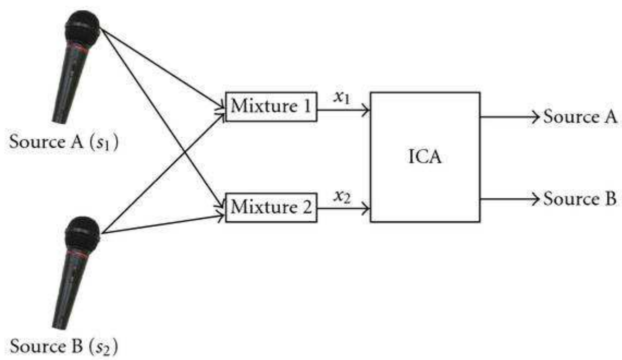
The *(differential) entropy* of a random vector X with density function f is defined as

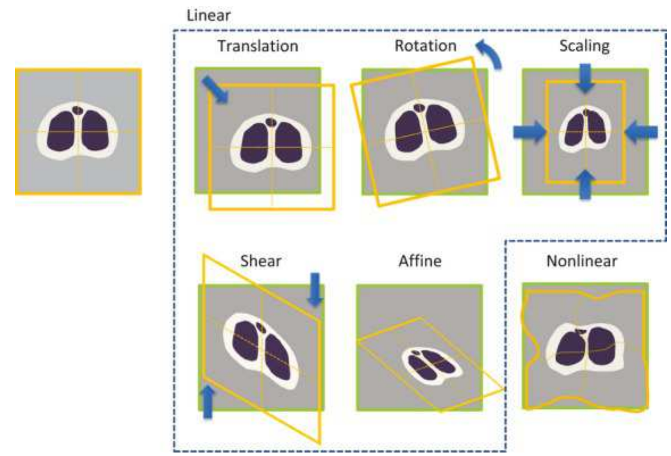
$$H = H(f) := -\mathbb{E}\{\log f(X)\} = -\int_{\mathcal{X}} f \log f$$

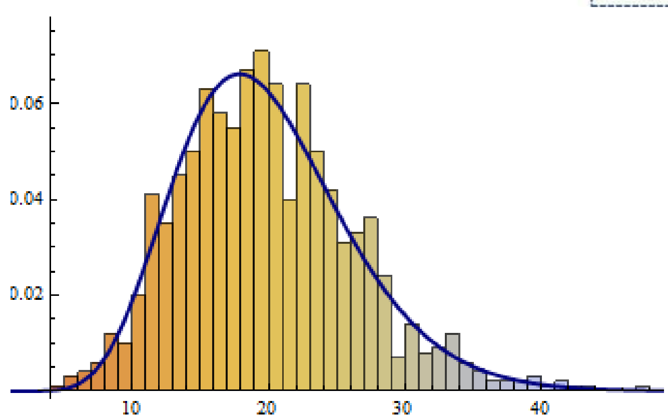
where $\mathcal{X} := \{x : f(x) > 0\}$.

The quantity $-\log f(X)$ is often thought of as a measure of information content, so H measures unpredictability.

Why estimate entropy?







Kozachenko–Leonenko estimators

Let $X_1, \ldots, X_n \overset{\mathrm{iid}}{\sim} f$ on $\mathbb{R}^d$. Let $X_{(k),i}$ denote the kth nearest neighbour of X_i, and let

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

The Kozachenko–Leonenko estimator of the entropy H is

$$\hat{H}_n = \hat{H}_n(X_1, \ldots, X_n) := \frac{1}{n}\sum_{i=1}^{n} \log\biggl(\frac{\rho_{(k),i}^d V_d (n-1)}{e^{\Psi(k)}}\biggr),$$

where $V_d := \pi^{d/2}/\Gamma(1+d/2)$ denotes the volume of the unit d-dimensional Euclidean ball and where Ψ denotes the digamma function.

Intuition

KL estimators attempt to mimic 'oracle' estimator $H_n^* := -n^{-1}\sum_{i=1}^n \log f(X_i)$ based on a k-nearest neighbour density estimate approximation

$$\frac{k}{n-1} \approx V_d \rho_{(k),1}^d f(X_1).$$

Previous work focuses on $k = 1$ or (recently) k fixed, and often assumes f is compactly supported (Kozachenko and Leonenko, 1987; Tsybakov and Van der Meulen, 1996; Singh et al., 2003; Mnatsakanov et al., 2008; Biau and Devroye, 2015; Delattre and Fournier, 2017; Singh and Póczos, 2016; Gao et al., 2016).

The trouble with full support

A Taylor expansion of $H(f)$ around a density estimator $\hat{f}$ yields

$$H(f) \approx -\int_{\mathbb{R}^d} f(x) \log \hat{f}(x)\, dx - \frac{1}{2}\biggl(\int_{\mathbb{R}^d} \frac{f^2(x)}{\hat{f}(x)}\, dx - 1\biggr).$$

When f is bounded away from zero on its support, one can estimate the (smaller order) second term to obtain efficient estimators in higher dimensions (Laurent, 1996).

However, when f is not bounded away from zero on its support such procedures are no longer effective.

Intuition regarding bias

Let $\xi_i := \frac{\rho^d_{(k),i} V_d(n-1)}{e^{\Psi(k)}}$**, and for** $u \in [0, \infty)$**, define**

$$F_{n,x}(u) := \mathbb{P}(\xi_i \leq u | X_i = x) = \sum_{j=k}^{n-1} \binom{n-1}{j} p^j_{n,x,u}(1-p_{n,x,u})^{n-1-j},$$

where $p_{n,x,u} := \int_{B_x(r_{n,u})} f(y)\,dy$ **and** $r_{n,u} := \bigl\{\frac{e^{\Psi(k)}u}{V_d(n-1)}\bigr\}^{1/d}$**.**
Also define a limiting distribution function

$$F_x(u) := e^{-\lambda_{x,u}} \sum_{j=k}^{\infty} \frac{\lambda^j_{x,u}}{j!},$$

where $\lambda_{x,u} := uf(x)e^{\Psi(k)}$**.**

More intuition regarding bias

We expect that

$$\mathbb{E}(\hat{H}_n) = \int_{\mathcal{X}} f(x) \int_0^\infty \log u \, dF_{n,x}(u) \, dx$$
$$\approx \int_{\mathcal{X}} f(x) \int_0^\infty \log u \, dF_x(u) \, dx$$
$$= \int_{\mathcal{X}} f(x) \int_0^\infty \log\Big(\frac{te^{-\Psi(k)}}{f(x)}\Big) e^{-t} \frac{t^{k-1}}{(k-1)!} \, dt \, dx = H,$$

where we have substituted $t = \lambda_{x,u}$.

Definition of parameter space

Let $\mathcal{F}_d$ **denote all density functions on** $\mathbb{R}^d$**, and let** $\mu_\alpha(f) := \int_{\mathcal{X}} \|x\|^\alpha f(x)\,dx$.

Let $\mathcal{A}$ **consist of all decreasing** $a : (0,\infty) \rightarrow [1,\infty)$ **with** $a(\delta) = o(\delta^{-\epsilon})$ **as** $\delta \searrow 0, \forall \epsilon > 0$. **For an** $m := \lceil \beta \rceil - 1$**-times differentiable** $f \in \mathcal{F}_d$ **and** $a \in \mathcal{A}$**, let** $r_a(x) := \{8d^{\frac{1}{2}} a(f(x))\}^{\frac{-1}{\beta \wedge 1}}$ **and**

$$M_{f,a,\beta}(x) := \max_{t=1,\ldots,m} \frac{\|f^{(t)}(x)\|}{f(x)} \vee \sup_{y \in B_x^\circ(r_a(x))} \frac{\|f^{(m)}(y) - f^{(m)}(x)\|}{f(x)\|y-x\|^{\beta-m}}.$$

For $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta := (0,\infty)^4 \times \mathcal{A}$**, set**

$$\mathcal{F}_{d,\theta} := \left\{ f \in \mathcal{F}_d : \mu_\alpha(f) \leq \nu, \|f\|_\infty \leq \gamma, \sup_{x : f(x) \geq \delta} M_{f,a,\beta}(x) \leq a(\delta) \right\}.$$

The bias of the KL estimator

Fix $d \in \mathbb{N}$ **and** $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta$**. Let** $k^* = k_n^* = O(n^{1-\epsilon})$ **as** $n \to \infty$ **for some** $\epsilon > 0$**.**

There exist $\lambda_1, \ldots, \lambda_{\lceil \beta/2 \rceil - 1} \in \mathbb{R}$**, depending only on** f **and** d**, such that for every** $\epsilon > 0$

$$\sup_{f \in \mathcal{F}_{d,\theta}} \biggl| \mathbb{E}\hat{H}_n - H - \sum_{l=1}^{\lceil \beta/2 \rceil - 1} \frac{\Gamma\bigl(k + \frac{2l}{d}\bigr)\Gamma(n)}{\Gamma(k)\Gamma\bigl(n + \frac{2l}{d}\bigr)} \lambda_l \biggr| = O\biggl(\frac{k^{\frac{\alpha}{\alpha+d} - \epsilon}}{n^{\frac{\alpha}{\alpha+d} - \epsilon}} \vee \frac{k^{\frac{\beta}{d}}}{n^{\frac{\beta}{d}}} \biggr),$$

uniformly for $k \in \{1, \ldots, k^*\}$**, with** $\lambda_l = 0$ **if** $2l \geq d\alpha/(\alpha + d)$**.**

The limitation of KL estimators

From our bias result, if $d \geq 3$, $\alpha > \frac{2d}{d-2}$, $\beta > 2$, then uniformly for $k \in \{1, \ldots, k^*\}$,

$$\sup_{f \in \mathcal{F}_{d,\theta}} \biggl| \mathbb{E}\hat{H}_n - H + \frac{V_d^{-2/d}\Gamma(k+2/d)}{2(d+2)\Gamma(k)n^{2/d}} \int_{\mathcal{X}} \frac{\Delta f(x)}{f(x)^{2/d}}\, dx \biggr| = o\biggl(\frac{k^{2/d}}{n^{2/d}}\biggr).$$

In particular, when $d \geq 4$ and $\int_{\mathcal{X}} \frac{\Delta f(x)}{f(x)^{2/d}}\, dx \neq 0$, the bias precludes the efficiency of $\hat{H}_n$.

Weighted KL estimators

For weights $w_1, \ldots, w_k$ **with** $\sum_{j=1}^{k} w_j = 1$**, define**

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

where $\xi_{(j),i} := e^{-\Psi(j)} V_d (n-1) \rho_{(j),i}^d$ **(e.g. Moon et al., 2016). If**

$$\sum_{j=1}^{k} w_j \frac{\Gamma(j+2/d)}{\Gamma(j)} = 0,$$

then when $d = 4$**,** $\alpha > d$ **and** $\beta > 2$**, we can make** $\sup_{f \in \mathcal{F}_{d,\theta}} |\mathbb{E}\hat{H}_n^w - H| = o(n^{-1/2})$**. If** $d = 5$ **then the same conclusion holds when** $\beta > 5/2$**.**

Choosing weights in the general case

Let

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

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

Then there exists $k_d \in \mathbb{N}$ **such that for** $k \geq k_d$**, we can find** $w = w^{(k)} \in \mathcal{W}^{(k)}$ **with** $\sup_{k \geq k_d} \|w^{(k)}\| < \infty$**. For such** w**,**

$$\sup_{f \in \mathcal{F}_{d,\theta}} \bigl| \mathbb{E}\hat{H}_n^w - H \bigr| = O\biggl(\max\biggl\{ \frac{k^{\frac{\alpha}{\alpha+d} - \epsilon}}{n^{\frac{\alpha}{\alpha+d} - \epsilon}}, \frac{k^{\frac{2(\lfloor d/4 \rfloor + 1)}{d}}}{n^{\frac{2(\lfloor d/4 \rfloor + 1)}{d}}}, \frac{k^{\frac{\beta}{d}}}{n^{\frac{\beta}{d}}} \biggr\} \biggr),$$

for each $\epsilon > 0$**, uniformly for** $k \in \{1, \ldots, k^*\}$**.**

Asymptotic variance

Let $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta$ **with** $\alpha > d$ **and** $\beta > 0$**. Let** k_0^* **and** k_1^* **satisfy** $k_0^* \leq k_1^*$**,** $k_0^*/\log^5 n \to \infty$ **and** $k_1^* = O(n^{\tau_1})$**, where**

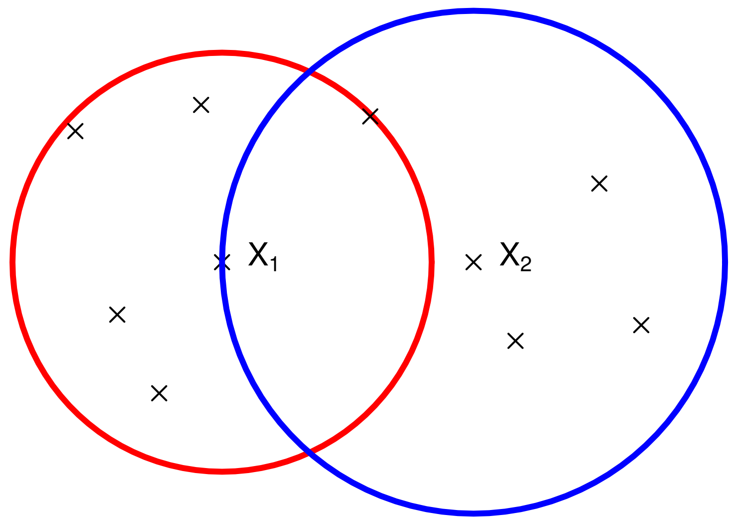
$$\tau_1 < \min\left\{ \frac{2\alpha}{5\alpha + 3d}, \frac{\alpha - d}{2\alpha}, \frac{4(\beta \wedge 1)}{4(\beta \wedge 1) + 3d} \right\}.$$

Write $V(f) := \mathrm{Var} \log f(X_1) = \int_{\mathcal{X}} f \log^2 f - H(f)^2$**. Then for any** $w = w^{(k)} \in \mathcal{W}^{(k)}$ **with** $\sup_{k \geq k_d} \|w^{(k)}\| < \infty$**,**

$$\sup_{k \in \{k_0^*, \ldots, k_1^*\}} \sup_{f \in \mathcal{F}_{d,\theta}} \left| n \mathrm{Var}\, \hat{H}_n^w - V(f) \right| \to 0$$

as $n \to \infty$**.**

Variance challenges



Here, X_1 is one of the five nearest neighbours of X_2, but not vice-versa.

Efficiency in arbitrary dimensions

Let $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta$ **with** $\alpha > d$ **and** $\beta > d/2$**. Let** k_0^* **and** k_1^* **satisfy** $k_0^* \leq k_1^*$, $k_0^*/\log^5 n \to \infty$, $k_1^* = O(n^{\tau_1})$ **and** $k_1^* = o(n^{\tau_2})$**, where**

$$\tau_2 := \min\biggl(1 - \frac{d/4}{1 + \lfloor d/4 \rfloor}, 1 - \frac{d}{2\beta}\biggr).$$

Then for any $w = w^{(k)} \in \mathcal{W}^{(k)}$ **with** $\sup_{k \geq k_d} \|w^{(k)}\| < \infty$**,**

$$\sup_{k \in \{k_0^*, \ldots, k_1^*\}} \sup_{f \in \mathcal{F}_{d,\theta}} n\mathbb{E}\bigl\{(\hat{H}_n^w - H_n^*)^2\bigr\} \to 0$$

as $n \to \infty$**. In particular,**

$$\sup_{k \in \{k_0^*, \ldots, k_1^*\}} \sup_{f \in \mathcal{F}_{d,\theta}} \bigl|n\mathbb{E}\{(\hat{H}_n^w - H(f))^2\} - V(f)\bigr| \to 0.$$

A confidence interval

The asymptotic variance $V(f)$ can be estimated by $\hat{V}_n^w := \max(\tilde{V}_n^w, 0)$, where

$$\tilde{V}_n^w := \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} w_j \log^2 \xi_{(j),i} - (\hat{H}_n^w)^2.$$

Fixing $q \in (0,1)$, this suggests that an asymptotic $(1-q)$-level confidence interval for $H(f)$ is given by

$$I_{n,q} := \big[\hat{H}_n^w - n^{-1/2} z_{q/2} (\hat{V}_n^w)^{1/2}, \hat{H}_n^w + n^{-1/2} z_{q/2} (\hat{V}_n^w)^{1/2}\big],$$

where z_q is the $(1-q)$th quantile of the standard normal distribution (Delattre and Fournier, 2017).

Asymptotic normality

Under the previous conditions,

$$\sup_{k \in \{k_0^*,\ldots,k_1^*\}} \sup_{f \in \mathcal{F}_{d,\theta}} d_{\mathrm{BL}}\Bigl(\mathcal{L}\bigl(n^{1/2}(\hat{H}_n^w - H(f))\bigr), N\bigl(0, V(f)\bigr)\Bigr) \to 0$$

as $n \to \infty$. Consequently,

$$\sup_{q \in (0,1)} \sup_{k \in \{k_0^*,\ldots,k_1^*\}} \sup_{f \in \mathcal{F}_{d,\theta}} \Bigl|\mathbb{P}\bigl(I_{n,q} \ni H(f)\bigr) - (1-q)\Bigr| \to 0.$$

Local asymptotic minimax lower bound

Fix $d \in \mathbb{N}$, $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta$ **and** $f \in \mathcal{F}_{d,\theta}$. **For** $t > 0$ **and a measurable** $g : \mathbb{R}^d \to \mathbb{R}$, **let**

$$f_{t,g}(x) := \frac{2c(t)}{1 + e^{-2tg(x)}} f(x),$$

where $c(t)$ **is a normalisation constant. For** $\lambda \in \mathbb{R}$, **let** $g_\lambda := -\lambda\{\log f + H(f)\}$. **If** $\mathcal{I}$ **denotes the set of finite subsets of** $\mathbb{R}$, **then for any estimator sequence** $(\tilde{H}_n)$,

$$\sup_{I \in \mathcal{I}} \liminf_{n \to \infty} \max_{\lambda \in I} n\mathbb{E}_{f_{n^{-1/2},g_\lambda}} \bigl[\bigl\{\tilde{H}_n - H(f_{n^{-1/2},g_\lambda})\bigr\}^2\bigr] \geq V(f).$$

Moreover, if $t|\lambda| \leq 1 \wedge \{144V(f)\}^{-1/2}$, **then** $f_{t,g_\lambda} \in \mathcal{F}_{d,\theta'}$, **where** $\theta' = (\alpha, \beta, 4\gamma, 4\nu, \tilde{a})$ **and** $\tilde{a}(\delta) := C_{\beta,d}\, a(\delta/4)^{\lfloor\beta\rfloor^2 + \lfloor\beta\rfloor + 1}$.

Summary

- **Kozachenko–Leonenko entropy estimators can be efficient for $d \leq 3$, but are typically not when $d \geq 4$**
- **By incorporating weights to kill main bias terms, we obtain efficient estimators in arbitrary dimensions, subject to sufficient moments and smoothness**
- **Future applications: testing log-concavity, independence...**

`http://arxiv.org/abs/1606.00304v3.`

References

- Berrett, T. B., Samworth, R. J. and Yuan, M. (2016) Efficient multivariate entropy estimation via k-nearest neighbour distances. http://arxiv.org/abs/1606.00304v3.
- Biau, G. and Devroye, L. (2015) Lectures on the Nearest Neighbor Method. Springer, New York.
- Delattre, S. and Fournier, N. (2017) On the Kozachenko–Leonenko entropy estimator. *J. Statist. Plann. Inf.*, 185, 69–93.
- Gao, W., Oh, S. and Viswanath, P. (2016) Demystifying fixed k-nearest neighbor information estimators. http://arxiv.1604.03006.
- Goria, M. N., Leonenko, N. N., Mergel, V. V. and Novi Inverardi, P. L. (2005). A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *J. Nonparam. Statist.*, 17, 277–297.

- **Kozachenko, L. F. and Leonenko, N. N. (1987). Sample estimate of the entropy of a random vector. Problemy Peredachi Informatsii, 23, 9–16.**
- **Laurent, B. (1996) Efficient estimation of integral functionals of a density. *Ann. Statist.*, 24, 659–681.**
- **Mnatsakanov, R. M., Misra, N., Li, S. and Harner, E. J. (2008). Kn-nearest neighbor estimators of entropy. *Math. Meth. Statist.*, 17, 261–277.**
- **Moon, K. R., Sricharan, K., Greenewald, K., Hero, A. O. (2016). Improving convergence of divergence functional ensemble estimators.** `https://arxiv.org/pdf/1601.06884v1.pdf`**.**
- **Singh, H., Misra, N., Hnizdo, V., Fedorowicz, A. and Demchuk, E. (2003). Nearest neighbor estimates of entropy. *Amer. J. Math. Man. Sci.*, 23, 301–321.**
- **Singh, S. and Póczos, B. (2016) Analysis of k nearest neighbor distances with application to entropy estimation. http://arxiv.1603.08578.**
- **Tsybakov, A. B. and Van der Meulen, E. C. (1996). Root-n consistent estimators of entropy for densities with unbounded support. Scand. J. Statist., 23, 75–83.**