

# On asymptotic properties of m-spacing entropy estimators

Yuri Golubev

CNRS, Université Aix-Marseille 1

Bordeaux, 6–7 September 2007

# Outline of the talk

- 1 Introduction
  - The entropy and its statistical applications
  - Delta-method
  - Universal Ibragimov-Khasminskii estimator
- 2 M-spacing entropy estimator
  - The standard Gaussian limit
  - Bias-variance decomposition
  - Main results
- 3 Examples
  - The Gaussian density
  - The Cauchy density

# Entropy estimation

We observe  $\mathbf{X}^n = (X_1, \dots, X_n)$ , where  $X_i$  are i.i.d. with an unknown probability density

$$p(x) = \frac{d}{dx} \mathbf{P}\{X_1 \leq x\}.$$

The goal is to estimate the entropy

$$H(p) = -\mathbf{E} \log[p(X_1)] = - \int_{-\infty}^{\infty} \log[p(x)] p(x) dx.$$

- *Goodness of fit tests*
- *True Random Number Generators testing*
- *Source coding*

# Goodness of fit test

Based on  $\mathbf{X}^n$  we want to test the simple hypothesis

$$H_0 : \quad p(x) = p_0(x)$$

against the alternative

$$H_1 : \quad p(x) \neq p_0(x).$$

The Kolmogorv-Smirnov test:  $H_1$  is accepted when

$$\sup_x \left| F(x, \mathbf{X}^n) - \int_{-\infty}^x p_0(u) du \right| \geq t_\alpha,$$

where

$$F(x, \mathbf{X}^n) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{X_i \leq x\}.$$

# Maximum likelihood test

Suppose we test two simple hypothesis

$$H_0 : p(x) = p_0(x) \text{ against } H_1 : p(x) = p_1(x),$$

where  $p_0(\cdot)$  is known. Then by the ML test, we accept  $H_1$  if

$$\frac{1}{n} \sum_{i=1}^n \log p_1(X_i) - \frac{1}{n} \sum_{i=1}^n \log p_0(X_i) \geq t_\alpha$$

In the goodness of fit,  $p_1(\cdot)$  is assumed to be unknown. However, by the law of large numbers, as  $n \rightarrow \infty$

$$\frac{1}{n} \sum_{i=1}^n \log p_1(X_i) \rightarrow -H(p_1)$$

This remark motivates the entropy goodness of fit test which has the following form

$$-H(\mathbf{X}^n) - \frac{1}{n} \sum_{i=1}^n \log p_0(X_i) \geq t_\alpha,$$

where  $H(\mathbf{X}^n)$  is an entropy estimator.

Example: Darling's (1953) uniformity test.

Suppose we want to estimate the probability  $\mathbf{P}\{X_1 \leq x\}$ , then we use the empirical distribution function

$$F(x, \mathbf{X}^n) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{X_i \leq x\}.$$

This estimate motivates the so-called  $\delta$ -method. If we want to estimate a linear functional

$$L(p) = \int l(x)p(x) dx,$$

we replace  $p(x)$  by

$$p(x, \mathbf{X}^n) = \frac{d}{dx} F(x, \mathbf{X}^n) = \frac{1}{n} \sum_{i=1}^n \delta(X_i - x),$$

where  $\delta(\cdot)$  is the Dirac  $\delta$ -function.

Thus we obtain

$$L(p, \mathbf{X}^n) = \frac{1}{n} \sum_{i=1}^n l(X_i).$$

Statistical analysis of this estimator is simple. For instance,

$$\sqrt{n}[L(p, \mathbf{X}^n) - L(p)] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma(p)),$$

where

$$\sigma^2(p) = \int l^2(p)p(x) dx - L^2(p).$$

Moreover, it is well known that  $L(p, \mathbf{X}^n)$  is an asymptotically optimal estimator in the minimax sense.

Unfortunately, the  $\delta$ -method fails for non-linear functionals  $\Phi(\cdot)$  like entropy.

A natural idea to overcome this difficulty is based on the plug-in principle. Suppose that we have at our disposal a preliminary estimator  $p_0(x, \mathbf{X}^n)$  of the density  $p(x)$ , i.e.

$$p_0(x, \mathbf{X}^n) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right).$$

Assume that  $\Phi(p)$  is a smooth functional such that

$$\Phi(p) \approx \Phi(p_0) + \langle \Phi'_{p_0}, p - p_0 \rangle,$$

where  $\Phi'_{p_0}(x)$  is the Frechet derivative belonging to  $L^2_p$  and

$$\langle q_1, q_2 \rangle = \int q_1(x)q_2(x) dx.$$

Thus we can estimate  $\Phi(p)$  with the help

$$\hat{\Phi}(\mathbf{X}^n) = \Phi(p_0) + \frac{1}{n} \sum_{i=1}^n \Phi'_{p_0}(X_i) - \int \Phi'_{p_0}(x) p_0(x, \mathbf{X}^n) dx.$$

*This method may provide the asymptotically efficient estimator if there exists  $\gamma > 0$  such that*

$$\int [\Phi'_{p_1}(x) - \Phi'_{p_2}(x)]^2 dx \leq C \left[ \int [p_1(x) - p_2(x)]^2 dx \right]^\gamma$$

In the case on entropy this assumption is fulfilled if  $p(x)$  is strictly bounded from below.

# Heuristic motivation

Since

$$H(p) = -\mathbf{E} \log[p(X_1)],$$

we can use the following estimator

$$H(\mathbf{X}^n) = -\frac{1}{n} \sum_{i=1}^n \log[p(X_i, \mathbf{X}^n)] = -\frac{1}{n} \sum_{i=1}^n \log[p(X_{(i)}, \mathbf{X}^n)],$$

where  $X_{(1)} \leq \dots \leq X_{(n)}$  denotes nondecreasing permutation of  $X_1, \dots, X_n$  and  $p(x, \mathbf{X}^n)$  is a density estimator.

The principal idea in the entropy estimation is to plug-in the following density estimator

$$\begin{aligned} p_m(X_{(i)}, \mathbf{X}^n) &= \frac{F(X_{(i+m)}, \mathbf{X}^n) - F(X_{(i)}, \mathbf{X}^n)}{X_{(i+m)} - X_{(i)}} \\ &= \frac{m}{n[X_{(i+m)} - X_{(i)}]}, \end{aligned}$$

where

$$F(x, \mathbf{X}^n) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{X_i \leq x\}, \quad \text{and} \quad F(X_{(i)}, \mathbf{X}^n) = \frac{i}{n}.$$

With this density estimator we arrive at the famous  $m$ -spacing entropy estimator

$$\hat{H}_m(\mathbf{X}^n) = \frac{1}{n} \sum_{i=1}^{n-m} \log \frac{n[X_{(i+m)} - X_{(i)}]}{m}.$$

# Pyke's theorem

In spite of the simplicity of this estimator, its statistical analysis is not banal. In what follows we use the famous Pyke's theorem (1965):

## Theorem

*Let  $U_1, \dots, U_n$  be independent random variables uniformly distributed on  $[0, 1]$  and let  $e_1, \dots, e_{n+1}$  be independent exponentially distributed random variables  $\mathbf{P}\{e_i > x\} = \exp(-x)$ . Then*

$$U_{(k)} \stackrel{\mathcal{D}}{=} \frac{\sum_{i=1}^k e_i}{\sum_{i=1}^{n+1} e_i}.$$

Using Pyke's theorem, we obtain

$$\begin{aligned}\hat{H}_m(\mathbf{X}^n) &= -\frac{1}{n} \sum_{i=1}^{n-m} \log[p(X_{(i)})] + \frac{1}{n} \sum_{i=1}^{n-m} \log \left[ \frac{1}{m} \sum_{k=i}^{i+m-1} e_k \right] \\ &\quad - \left(1 - \frac{m}{n}\right) \log \left[ \frac{1}{n} \sum_{k=1}^{n+1} e_k \right] + \frac{\epsilon_n}{\sqrt{n}},\end{aligned}$$

where

$$\epsilon_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n-m} \log \frac{F[X_{(i+m)}] - F[X_{(i)}]}{p(X_{(i)})[X_{(i+m)} - X_{(i)}]}.$$

Statistical properties of the first term at the right-hand side can be easily analyzed by standard probabilistic methods. Indeed, by the central limit theorem

$$\begin{aligned} & \sqrt{n} \left[ -\frac{1}{n} \sum_{i=1}^{n-m} \log[p(X_{(i)})] - H(p) \right] \\ & \approx -\frac{1}{\sqrt{n}} \sum_{i=1}^n [\log[p(X_i)] + H(p)] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2(p)), \end{aligned}$$

where

$$\sigma^2(p) = \int_{-\infty}^{\infty} \log^2(p(x)) p(x) dx - H^2(p).$$

The second term has also a Gaussian limit

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n-m} \left\{ \log \left[ \sum_{k=i}^{i+m-1} e_k \right] - \Psi(m) \right\} - \sqrt{n} \log \left[ \frac{1}{n} \sum_{k=1}^{n+1} e_k \right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma^2(m)),$$

Cressie (1976) proved

$$\Sigma^2(m) = (2m^2 - 2m + 1)\Psi'(m) - 2m + 1.$$

where  $\Psi(m)$  is digamma function

$$\Psi(m) = \frac{1}{\Gamma(m)} \int_0^{\infty} \log(x) x^{m-1} \exp(-x) dx = \frac{\Gamma'(m)}{\Gamma(m)}.$$

Therefore one can hope that as  $n \rightarrow \infty$

$$\sqrt{n}[\widehat{H}_m(\mathbf{X}^n) - H(p) - \Psi(m) + \log(m)] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2(p) + \Sigma^2(m))$$

To prove this, we have to check that

$$\epsilon_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n-m} \log \frac{F[X_{(i+m)}] - F[X_{(i)}]}{p(X_{(i)})[X_{(i+m)} - X_{(i)}]}$$

is small, i.e.

$$\lim_{n \rightarrow \infty} \mathbf{E} \epsilon_n^2 = 0.$$

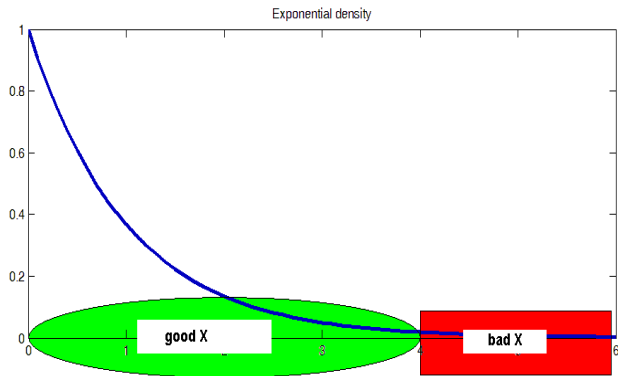
This fact can be easily proved if we suppose the density  $p(\cdot)$  has a compact support and strictly bounded from zero over his support, and the derivative  $p'(x)$  is bounded over this support (Beirlant and van Zuijlen (1985) Beirlant (1986)).

When  $p(x)$  has an unbounded support or vanishes, the problem is non-banal.

To overcome this difficulty, there were several attempts based on the idea to modify the  $m$ -spacing entropy estimator (e.g. Kozachenko and Leonenko (1987) Tsybakov and Van der Meulen (1992)).

The main goal in my talk it to demonstrate that the standard entropy estimator works for vanishing and unbounded probability densities as well.

# The main idea



For  $x \in \mathbb{R}^1$  and a given sequence  $r(n) \geq 1$  define the family of balls in  $\mathbb{R}^1$  by

$$\mathbb{B}^n(x) \stackrel{\text{def}}{=} \left\{ y : |F(x) - F(y)| \leq \frac{2 \log(n)}{n} \right\}.$$

Denote also

$$D^n(x) \stackrel{\text{def}}{=} \sup_{y \in \mathbb{B}^n(x)} \left\{ \frac{p'^2(y)}{p^2(y)} + \frac{|p''(y)|}{p(y)} \right\}.$$

Let

$$\mathbb{Q}_R^n \stackrel{\text{def}}{=} \left\{ x : p(x) \geq \sqrt{D^n(x)} \frac{R(n)}{n} \right\},$$

where  $R(n)$  is a given sequence.

# Controlling the remainder term

## Lemma

Let  $R(n) \geq 14 \log(n)$ . Assume that for some  $\varepsilon > 0$

$$\mathbf{E}|X_1 - X_2|^{\pm\varepsilon} < \infty, \quad \mathbf{E}p^{\pm\varepsilon}(X_1) < \infty,$$

then

$$\begin{aligned} [\mathbf{E}\varepsilon_n^2]^{1/2} &\leq C \log(n) \sqrt{n} \int_{x \notin \mathbb{Q}_R^n} p(x) dx \\ &\quad + \frac{C \log(n)}{n^{3/2}} \int_{x \in \mathbb{Q}_R^n} \frac{D^n(x)}{p(x)} dx + \frac{CmN_n \log(n)}{\sqrt{n}}, \end{aligned}$$

where  $N_n$  is the number of connected components of  $\mathbb{Q}_{r,R}^n$ .

# Conditions A

Suppose  $R(n) \geq 14 \log(n)$  and

- for some  $\varepsilon > 0$   $\mathbf{E}|X_1 - X_2|^{\pm\varepsilon} < \infty$ ,  $\mathbf{E}p^{\pm\varepsilon}(X_1) < \infty$ ,
- 

$$\sqrt{n} \int_{x \notin \mathbb{Q}_R^n} p(x) dx + \frac{1}{n^{3/2}} \int_{x \in \mathbb{Q}_R^n} \frac{D^n(x)}{p(x)} dx = o\left(\frac{1}{\log(n)}\right),$$

- $\lim_{n \rightarrow \infty} mN_n \log(n) / \sqrt{n} = 0$ ,

## Theorem

*Under the above conditions*

$$\limsup_{n \rightarrow \infty} n \mathbf{E} \left[ \widehat{H}_m(\mathbf{X}^n) - H(p) - \Psi(m) + \log(m) \right]^2 \leq C.$$

The following theorem summarizes the principal facts of the talk.

## Theorem

*Under the conditions A*

$$\lim_{n \rightarrow \infty} \sqrt{n} [\hat{H}_m(\mathbf{X}^n) - H(p) - \Psi(m) + \log(m)] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma^2(m) + \sigma^2(p)),$$

where

$$\Sigma^2(m) = (2m^2 - 2m + 1)\Psi'(m) - 2m + 1$$

$$\sigma^2(p) = \int \log^2[p(x)]p(x) dx - H^2(p)$$

and

$$\lim_{n \rightarrow \infty} n \mathbf{E} [\hat{H}_m(\mathbf{X}^n) - H(p) - \Psi(m) + \log(m)]^2 = \sigma^2(p) + \Sigma^2(m).$$

Let

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, x \in \mathbb{R}^1.$$

In this case for large  $n$

$$D^n(x) \approx \frac{|p'(x)|}{p^2(x)} + \frac{|p''(x)|}{p(x)} \approx 2x^2 + 1$$

and it can be shown that

$$\mathbb{Q}_R^n \approx \left\{ x : |x| \leq \sqrt{2 \log \frac{n}{R(n)}} \right\}.$$

Therefore

$$\int_{x \notin \mathbb{Q}_R^n} p(x) dx \leq \frac{CR(n)\sqrt{\log(n)}}{n},$$
$$\int_{x \in \mathbb{Q}_R^n} \frac{\sqrt{D^n(x)}}{p(x)} dx \leq \frac{Cn}{R(n)}.$$

Thus, in order to check Condition A, we need

$$\lim_{n \rightarrow \infty} \log(n) \left[ \frac{R(n)\sqrt{\log(n)}}{\sqrt{n}} + \frac{1}{R(n)\sqrt{n}} \right] = 0.$$

In other words,  $R(n) \leq \sqrt{n} \log^{-3/2-\epsilon}(n)$  for some  $\epsilon > 0$ , but the optimal choice is  $R(n) = 14 \log(n)$ .

Let

$$p(x) = \frac{1}{\pi(1+x^2)}$$

In this case,

$$D^n(x) \leq \frac{C}{x^2}, \quad |x| \geq 1,$$

and for some  $C_1 < C_2$

$$\left\{ x : |x| \leq \frac{C_1 n}{R(n)} \right\} \subseteq \mathbb{Q}_R^n \subseteq \left\{ x : |x| \leq \frac{C_2 n}{R(n)} \right\}.$$

# Cauchy density

Thus we easily obtain

$$\int_{x \in \mathbb{Q}_R^n} \frac{\sqrt{D^n(x)}}{p(x)} dx \leq \left( \frac{Cn}{R(n)} \right)^2,$$
$$\int_{x \notin \mathbb{Q}_R^n} p(x) dx \leq \frac{CR(n)}{n}.$$

Therefore Condition A is fulfilled if

$$\lim_{n \rightarrow \infty} \log(n) \left[ \frac{R(n)}{\sqrt{n}} + \frac{\sqrt{n}}{R^2(n)} \right] = 0,$$

and  $R(n) = n^{1/3}$  provides the optimal choice.