

# Statistical estimation of the Shannon entropy

Denis Dimitrov

(Lomonosov Moscow State University)

*den.dimitrov@gmail.com*

December 8, 2020

# Overview

## 1 Introduction

- History
- The concept of discrete entropy
- Entropy of absolutely continuous variables
- Relation between two concepts
- Another types of entropy

## 2 Statistical estimation of the Shannon entropy

- Problem statement
- Discrete variables case
- Differential entropy estimate
- Case of mixture of absolutely continuous distributions

## 3 Some applications

- Problem statement
- Detection of defects of porous media

# History

The notion of entropy belongs to the principle ones in Physics and Mathematics.

R.Clausius is considered as the father of the entropy concept. The important contributions to the development of this concept were made by L.Boltzmann, J.Gibbs and M.Planck.

Mathematicians also were preoccupied with the entropy. The works by C.Shannon, A.N.Kolmogorov, Ya.G.Sinai, A.Rényi, A.S.Holevo, T.Tsallis are worth mentioning in this regard.

# The concept of discrete entropy

Let  $\xi : \Omega \rightarrow S$ , where  $S$  is some finite or numerable set. For such a random variables C.Shannon introduced a concept of entropy

Definition (Discrete case)

$$H(\xi) := - \sum_{x \in S} p_x \log p_x,$$

where  $p_x = \mathbf{P}(\xi = x)$ ,  $x \in S$  and  $0 \log 0 := 0$  by continuity.

# Properties

- $H(\xi) \geq 0$  for all  $\xi$  with finite or numerable set of values.
- $H(\xi) = 0$  iff  $\xi = c = \text{const}$  a.s.,  $c \in S$ .
- For  $S = \{x_1, \dots, x_n\}$ , by the Jensen inequality for concave function  $g(t) = \log t$ ,  $t > 0$ ,

$$H(\xi) = \sum_{i=1}^n p_{x_i} \log \frac{1}{p_{x_i}} \leq \log \left( \sum_{i=1}^n p_{x_i} \frac{1}{p_{x_i}} \right) = \log n.$$

Thus  $H(\xi)$  takes maximal value when  $\xi$  has uniform distribution on  $S$ .

- One can view the Shannon entropy  $H(\xi)$  as a "measure" of uncertainty of a random variable  $\xi$ .

# Differential entropy

Let  $\xi : \Omega \rightarrow \mathbb{R}^d$ ,  $f(x)$  be a probability density function w.r.t. Lebesgue measure  $\mu$  on  $\mathbb{R}^d$ ,  
 $S = \text{supp}(f) := \{x \in \mathbb{R}^d : f(x) > 0\}$ .

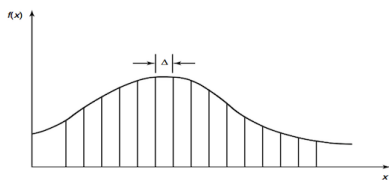
## Definition (Differential entropy)

$$h(\xi) := - \int_S f(x) \log f(x) dx.$$

Note that  $h(\xi)$  can be less than 0. For instance, if  $\xi \sim \text{U}[0, a]$ ,  $a < 1$ , then  $h(\xi) = \log a < 0$ .

## Relation between two concepts

Consider a random variable  $X : \Omega \rightarrow \mathbb{R}$  having density  $f(x)$ .



For all  $\Delta > 0$  and all  $i \in \mathbb{Z}$ , choose

$x_i(\Delta) \in [i\Delta, (i+1)\Delta)$  such that  $f(x_i(\Delta))\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x) dx$ .

Let  $X^\Delta(\omega) := x_i(\Delta)$  if  $i\Delta \leq X(\omega) < (i+1)\Delta$ .

Then  $X^\Delta \rightarrow X$  as  $\Delta \rightarrow 0$ . However

$H(X^\Delta) + \log \Delta \rightarrow h(X)$  as  $\Delta \rightarrow 0$ . Thus the differential entropy is not an extension of the discrete entropy concept.

## Another types of entropy

### ■ Rényi entropy

$$h_q^* := \frac{1}{1-q} \log \int_{\mathbb{R}^d} f^q(x) dx, \quad q \neq 1,$$

### ■ Tsallis entropy

$$h_q := \frac{1}{q-1} \left( 1 - \int_{\mathbb{R}^d} f^q(x) dx \right), \quad q \neq 1.$$

One can see that  $h_q^* \rightarrow h$  and  $h_q \rightarrow h$  as  $q \rightarrow 1$ .



## Problem statement

Let  $\xi$  be a discrete random vector with values in finite set  $S$ .

We have a sample of i.i.d. random vectors  $X_1, X_2, \dots, X_n$  with values in a finite set  $S$  and having the same law as a vector  $\xi$ .

The goal is to provide a statistical estimate of discrete  $H(\xi)$ .

# Ideas of estimation of the discrete entropy

When  $n$  is large enough (a set  $S$  is fixed), one can apply the "plug-in" method. For all  $x \in S$  define

$$\hat{p}_{n,x} := \frac{\sum_{i=1}^n \mathbb{I}\{X_i = x\}}{n}.$$

Introduce

$$\hat{H}_n := - \sum_{x \in S} \hat{p}_{n,x} \log \hat{p}_{n,x}.$$

It is clear (by SLLN) that  $\hat{p}_{n,x} \rightarrow p_x$  a.s.,  $n \rightarrow \infty$ .  
Thus  $\hat{H}_n \rightarrow H(\xi)$  a.s when  $n \rightarrow \infty$ .

# Ideas of estimation of the discrete entropy

In 1958 R. Dobrushin proposed a new method to estimate discrete entropy when  $|S|$  is large with respect to  $n$ . He has defined

$\eta := \min \{1 < k \leq n : X_k = X_1\}$ . It turns out that  $E \log \eta \approx H(\xi) - \gamma$ , where  $\gamma$  is the Euler's constant,  $\gamma \approx 0.577$ .

So, the idea of approximation scheme is to generate  $M$  (large enough) independent observations  $y_1, \dots, y_M$  of the random variable  $\eta$ . Then one can define  $\hat{H}_n := \frac{\sum_{m=1}^M \log y_m}{M} + \gamma$  and by the SLLN  $\hat{H}_n \approx H(\xi)$  when  $n$  is large enough.

## Kozachenko-Leonenko estimate

**Notation**

Let us now consider a random vector  $\xi : \Omega \rightarrow \mathbb{R}^d$ ,  $d \geq 1$ , with a density  $f(x)$ .

For each  $i = 1, \dots, n$ , set  $\rho_i := \min_{j \neq i} \|X_i - X_j\|$ .  
 Let  $V_d = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$  be a volume of the unit ball in  $\mathbb{R}^d$ .

Then, for  $n \in \mathbb{N}$ , the Kozachenko-Leonenko estimate of a differential entropy  $h(\xi)$  is provided by the formula

$$\hat{h}_n := -\psi(1) + \log(n-1) + \frac{1}{n} \sum_{i=1}^n \log(\rho_i^d V_d).$$

# Kozachenko-Leonenko estimate

In this section we discuss the result of our recent joint paper with A. Bulinski

For  $x \in \mathbb{R}^d$ ,  $r > 0$ ,  $R > 0$  and a density  $f$  introduce the following functionals.

- $I_f(x, r) := \frac{\int_{B(x,r)} f(y) dy}{r^d V_d},$
- $M_f(x, R) := \sup_{r \in (0, R]} I_f(x, r),$
- $m_f(x, R) := \inf_{r \in (0, R]} I_f(x, r).$

Note that  $M_f(x, R)$  is an analog of the Hardy-Littlewood maximal function.

## Kozachenko-Leonenko estimate

## Theorem (B&amp;D, 2017. Asymptotic unbiasedness)

*Assume that, for some  $p > 1$  and some positive  $\varepsilon_i, R_i$ , where  $i = 1, 2$ , the following conditions are satisfied*

- $K_f(p) := \int_{\mathbb{R}^d} |\log \|x - y\||^p f(x) f(y) dx dy < \infty$ ,
- $Q_f(\varepsilon_1, R_1) := \int_{\mathbb{R}^d} M_f^{\varepsilon_1}(x, R_1) f(x) dx < \infty$ ,
- $T_f(\varepsilon_2, R_2) := \int_{\mathbb{R}^d} m_f^{-\varepsilon_2}(x, R_2) f(x) dx < \infty$ .

*Then the estimates  $\hat{h}_n$  are asymptotically unbiased, i.e.*

$$\lim_{n \rightarrow \infty} \mathbb{E} \hat{h}_n = h.$$

## Kozachenko-Leonenko estimate

Theorem (B&D, 2017.  $L^2$ -consistency)

*Assume that, for some  $p > 2$  and some positive  $\varepsilon_i, R_i$ , where  $i = 1, 2$ , the following conditions are satisfied*

- $K_f(p) := \int_{\mathbb{R}^d} |\log \|x - y\||^p f(x) f(y) dx dy < \infty$ ,
- $Q_f(\varepsilon_1, R_1) := \int_{\mathbb{R}^d} M_f^{\varepsilon_1}(x, R_1) f(x) dx < \infty$ ,
- $T_f(\varepsilon_2, R_2) := \int_{\mathbb{R}^d} m_f^{-\varepsilon_2}(x, R_2) f(x) dx < \infty$ .

*Then the estimates  $\hat{h}_n$  are  $L^2$ -consistent, i.e.*

$$\mathbb{E}(\hat{h}_n - h)^2 \rightarrow 0, \quad n \rightarrow \infty.$$

## Gaussian random vector

## Corollary (B&amp;D, 2017)

*Let  $\xi$  be a Gaussian random vector in  $\mathbb{R}^d$  with  $E\xi = \nu$  and a nondegenerate covariance matrix  $\Sigma$  (i.e.  $\xi$  has a density). Then  $E(\hat{h}_n - h)^2 \rightarrow 0$  as  $n \rightarrow \infty$ , where  $h = \frac{1}{2} \log \det(2\pi e \Sigma)$ .*

The proof involves the following inequalities  $m_f(x, R) \geq C f(x)$  and  $\int_{\mathbb{R}^d} f(x)^{1-\varepsilon} dx < \infty$  permitting to claim that  $T_f(\varepsilon, R) < \infty$ .



# Case of mixture of absolutely continuous distributions

## Notation

Assume that  $f_1, \dots, f_M$  are the probability density functions w.r.t. the Lebesgue measure  $\mu$  in  $\mathbb{R}^d$ .

Let us define, for all  $x \in \mathbb{R}^d$  and  $\alpha_1, \dots, \alpha_M$  such that  $0 < \alpha_m < 1$ ,  $m \in \{1, \dots, M\}$ ,  $\sum_{m=1}^M \alpha_m = 1$  a mixture

$$f(x) := \sum_{m=1}^M \alpha_m f_m(x)$$

## Case of mixture of absolutely continuous distributions

## Theorem (D, 2018. Asymptotic unbiasedness)

*Assume that, for some  $p > 1$  and some positive  $\varepsilon, c, C, R$  the following conditions are satisfied*

- $\int_{\mathbb{R}^d} |\log \|x - y\||^p f_m(x) f_m(y) dx dy < \infty$  for all  $m = 1, \dots, M$ .
- $cM_{f_m}(x, R) \leq f_m(x) \leq Cm_{f_m}(x, R)$  for all  $m = 1, \dots, M$ .
- $f_m \in L^{1-\varepsilon} \cap L^{1+\varepsilon}$  for all  $m = 1, \dots, M$ .

*Then the estimates  $\hat{h}_n$  of differential entropy of random variable  $\xi$  with mixture density  $f(x)$  are asymptotically unbiased, i.e.  $\lim_{n \rightarrow \infty} \mathbb{E} \hat{h}_n = h$ .*

## Case of mixture of absolutely continuous distributions

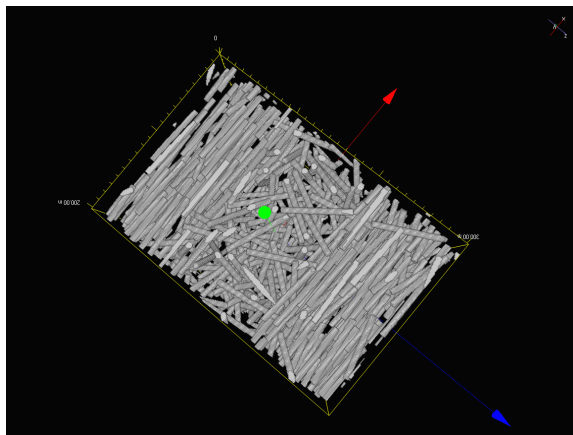
Theorem (D, 2018.  $L^2$ -consistency)

*Assume that, for some  $p > 2$  and some positive  $\varepsilon, c, C, R$  the following conditions are satisfied*

- $\int_{\mathbb{R}^d} |\log \|x - y\||^p f_m(x) f_m(y) dx dy < \infty$  for all  $m = 1, \dots, M$ .
- $cM_{f_m}(x, R) \leq f_m(x) \leq Cm_{f_m}(x, R)$  for all  $m = 1, \dots, M$ .
- $f_m \in L^{1-\varepsilon} \cap L^{1+\varepsilon}$  for all  $m = 1, \dots, M$ .

*Then the estimates  $\hat{h}_n$  of differential entropy of random variable  $\xi$  with mixture density  $f(x)$  are  $L^2$ -consistent, i.e.  $E(\hat{h}_n - h)^2 \rightarrow 0, n \rightarrow \infty$ .*

# Problem statement

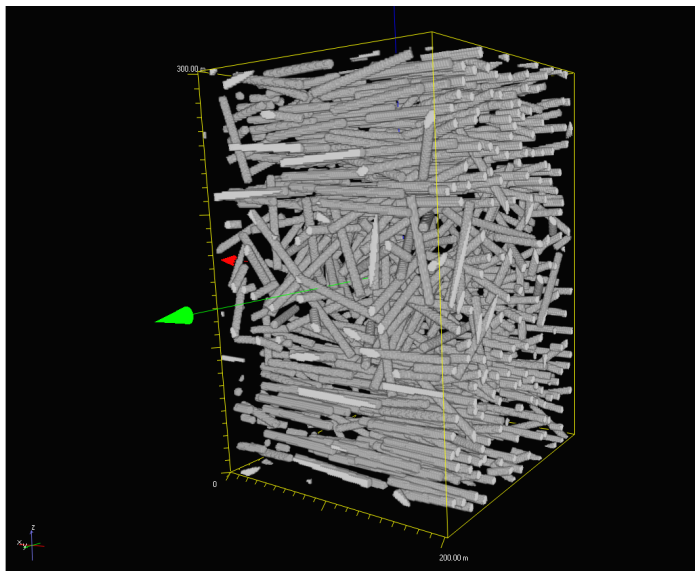


Axis  $X$  is red,  $Y$  is green,  $Z$  is blue.

$$\mathbb{P} = [0, 200] \times [0, 200] \times [0, 300],$$

$$\mathbb{P}_{inhom} = [0, 200] \times [0, 200] \times [100, 200].$$

# Problem statement



# Detection of defects of porous media. Algorithm

- Make an uniform three-dimensional grid  $\mathbb{S}$  that covers our material.
- At each point of the grid one can take a small cube with center at this point. Then calculate an estimate of differential entropy of fibres' directions distribution in the cube.

After that procedure one will have a set of the local estimates of a differential entropy at each point:

$$\mathcal{H}^{\mathbb{S}} = \{\hat{h}(M)\}_{M \in \mathbb{S}}.$$

# Detection of defects of porous media. Algorithm

Let  $\hat{\mu} = \mu(\mathcal{H}^{\mathbb{S}})$  be a sample median,  
 $\hat{m} = \frac{1}{|\mathbb{S}|} \sum_{M \in \mathbb{S}} \hat{H}(M)$  be a sample mean,  
 $\hat{\sigma}^2 = \frac{1}{|\mathbb{S}|-1} \sum_{M \in \mathbb{S}} (\hat{H}(M) - \hat{m})^2$  be a sample  
 variance.

The estimation of inhomogeneity area now can be  
 written as

$$\hat{R}_N = \{M \in \mathbb{P} : |\hat{H}(M) - \hat{\mu}| > k \cdot \hat{\sigma}\}$$

# Detection of defects of porous media. Results

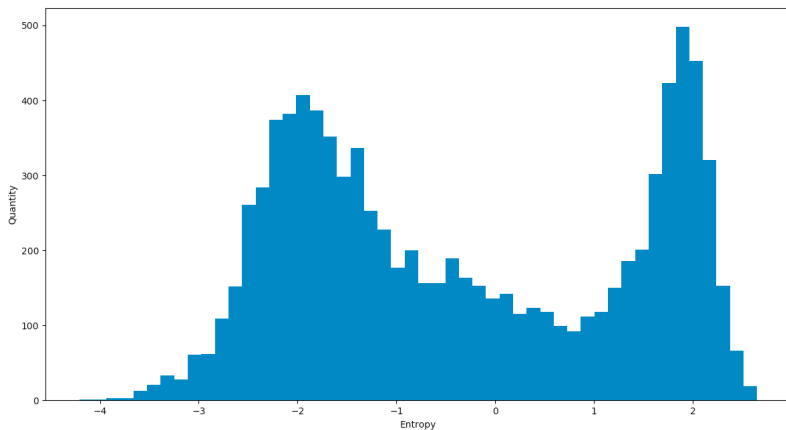


Figure:  $\hat{\mu} = -0.68$ ,  $\hat{\sigma} = 1.70$ ,  $k = 1.5$



# Detection of defects of porous media. Results

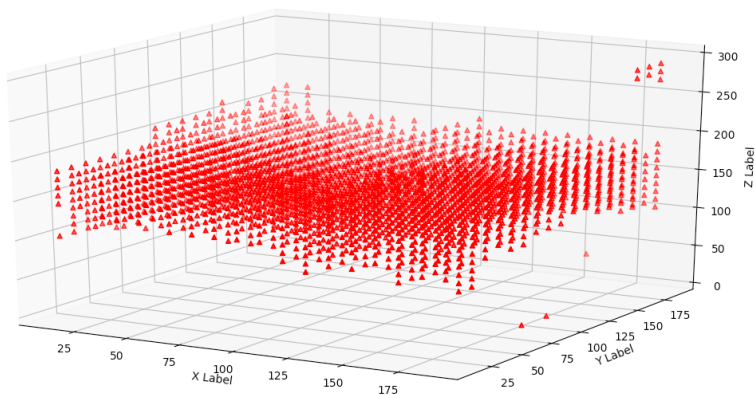






Figure:  $r_{true}^{inhom} \approx 0.85$ ,  $r_{true}^{hom} \approx 0.99$





## Possible improvements

- Adaptive median
- Adaptive  $k$  in  $k$ -sigma rule

# References

-  Alonso-Ruiz, P., Spodarev, E.: Entropy-based inhomogeneity detection in porous media. *arXiv preprint*, arXiv:1611.02241
-  Bulinski, A. V., Dimitrov, D.V.: Statistical estimation of the Shannon entropy (to appear).
-  Delattre, S., Fournier, N.: On the Kozachenko-Leonenko Entropy Estimator. *Journal of Statistical Planning and Inference*, (2017), DOI:  
<http://dx.doi.org/10.1016/j.jspi.2017.01.004> (accepted manuscript)
-  Dobrushin, R.L.: A simplified method of experimentally evaluating the entropy of a stationary sequence. *Theory of Probability and its Applications*, (1958), 3:4, 428–430

# References

-  Kozachenko, L.F., Leonenko, N.N.: Sample estimate of the entropy of a random vector. *Problems of Information Transmission*, **23**, Issue 2, 9-16 (1987)
-  Leonenko, N.N., Pronzato, L., Savani V.: A class of Rényi information estimations for multidimensional densities. *The Annals of Statistics*, **36**, 2153–2182 (2008). Correction: *The Annals of Statistics*, **38**, 3837-3838 (2010)
-  Shannon, C.E.: A Mathematical Theory of Communication. *Bell Systems Technical Journal*, **27**, July and October, 379–423 and 623–656 (1948)
-  Singh, S., Pószoc, B.: Analysis of  $k$ -nearest neighbor distances with application to entropy estimation, *arXiv preprint*, arXiv: 1603.08578v2