NONLINEAR TIME SERIES ANALYSIS, WITH APPLICATIONS TO MEDICINE

José María Amigó

Centro de Investigación Operativa, Universidad Miguel Hernández, Elche (Spain)

OUTLINE OF THE COURSE

LECTURE 1 INFORMATION THEORY

LECTURE 2: DYNAMICAL SYSTEMS

LECTURE 3: SYMBOLIC DYNAMICS

LECTURE 4: NONLINEAR METHODS IN MEDICINE I

LECTURE 5: NONLINEAR METHODS IN MEDICINE II

LECTURE 1

INFORMATION THEORY

OUTLINE

Entropy is a measure of the uncertainty of a random variable.

Notation:

- X a random variable.
- $p(x) = \Pr\{X = x\}$, the *probability function* of X,

$$0 \leq p(x) \leq 1, \quad \sum_{x \in \mathcal{X}} p(x) = 1.$$

- The *alphabet* $\mathcal{X}$ is the set of all possible outcomes of X.
- The outcomes of X are also called *letters*, *symbols* or *states*.
- If $\#\mathcal{X} < \infty$, X is called a *finite-alphabet*, or *finite-state* rv

1. Information and Shannon entropy

Definition. The *entropy* $H(X)$ of a finite-alphabet rv X is

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x).$$

Remarks.

- Units: $\log_2 \to$ bits, $\log_e \to$ nats, $\log_{10} \to$ dits.
- If $p(x) = 0$, then $p(x) \log p(x) = 0 \log 0 = 0$ by convention.
- $H(X)$ depends on X only through $p(x)$: $H(X) = H(p)$.

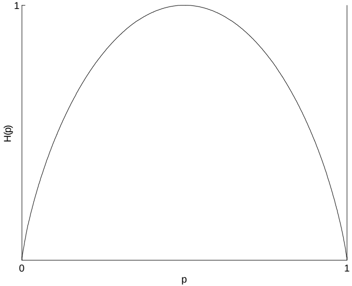
Example. Let

$$X=\begin{cases}1 & \text{with probability } p,\\ 0 & \text{with probability } 1-p.\end{cases}$$

Then

$$H(X)=-p\log p-(1-p)\log(1-p)=:H(p).$$

Example (cont'd)



- $H(p) = 0$ when $p = 0$ or $p = 1$,
- Maximum at $p = 1/2$: $H(1/2) = \log 2 = 1$ bit.
- In general,

$$H_{\max}(p) = H(\tfrac{1}{\#\mathcal{X}}, ..., \tfrac{1}{\#\mathcal{X}}) = -\sum_{x\in\mathcal{X}} \tfrac{1}{\#\mathcal{X}} \log \tfrac{1}{\#\mathcal{X}} = \log \#\mathcal{X}.$$

1. Information and Shannon entropy

Definition. The *Rényi entropy of order* q, where $q \geq 0$ and $q \neq 1$, is

$$H_q(X) = \frac{1}{1-q} \log \sum_{x \in \mathcal{X}} p(x)^q.$$

- $H_0(X) = \log |\mathcal{X}|$ (*Hartley entropy*)
- $\lim_{q \to 1} H_q(X) \equiv H_1(X) =$ *Shannon entropy*
- $H_2(X) = -\log \sum p(x)^2 = -\log \Pr\{X = Y\}$ where X, Y are $i.i.d.$ (*collision entropy*)
- $\lim_{q \to \infty} H_q(X) \equiv H_\infty(X) = \min\{-\log p(x)\} = -\log \max\{p(x)\}$ (*min-entropy*)

Property.

$$\boxed{H_0(X) \geq H_1(X) \geq ... \geq H_\infty(X).}$$

BRIEF CHRONOLOGY OF ENTROPY

- In **physics** (as a measure of *disorder*):

  Boltzmann (1877), Gibbs (1902), von Neumann (1927),...

- In **Information theory** (as a measure of *uncertainty*):

  Shannon (1948), Kullback-Leibler (1951), Rényi (1961),...

- In **metric dynamical systems** (as a measure of *randomness*):

  Kolmogorov (1958), Sinai (1959),...

- In **continuous dynamical systems** (as a measure of *complexity*):

  Adler-Konheim-McAndrew (1965), Bowen (1971), ...

Boltzmann's tomb at the *Zentralfriedhof* in Vienna

Claude E. Shannon (1916-2001)

1. Information and Shannon entropy

According to K. Denbigh[1]:

When Shannon had invented his quantity and consulted von Neumann on what to call it, von Neumann replied: "Call it entropy. It is already in use under that name and besides, it will give you a great edge in debates because nobody knows what entropy is anyway".

[1] K. Denbigh. In *Maxwell's Demon, Entropy, Information, Computing* (ed. H.S. Leff and A.F. Rex), pp. 109-115. Princeton University Press,1990.

2. Joint entropy and conditional entropy

- The *joint entropy* of two rv X and Y is

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y),$$

where

$$p(x,y) = \Pr\{X = x, Y = y\}.$$

- The entropy of Y conditioned on X, or *conditional entropy* $H(Y\,|X)$ is

$$H(Y\,|X) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y\,|x)\,,$$

where

$$p\left(y|\,x\right) = \frac{p(x,y)}{p(x)}.$$

2. Joint entropy and conditional entropy

Properties

- $H(X,Y) = H(Y,X)$
- $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$ (*Chain rule*)
- $H(X,Y) \leq H(X) + H(Y)$
- $H(X,Y) = H(X) + H(Y)$ iff X and Y are independent (i.e., $p(x,y) = p(x)p(y)$)
- $H(X|Y) \leq H(X)$
- $H(X|Y) = H(X)$ iff X and Y are independent

Example. Let (X, Y) have the following probability function $p(x, y)$:

$Y \backslash X$	1	2	3	4	$p(y)$
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{4}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{4}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
4	$\frac{1}{4}$	0	0	0	$\frac{1}{4}$
$p(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	

Then

$$\begin{aligned} p(y|\,1) &= \left(\tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{8}, \tfrac{1}{2}\right), \; p(y|\,2) = \left(\tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4}, 0\right), \\ p(y|\,3) &= \left(\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{2}, 0\right), \; p(y|\,4) = \left(\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{2}, 0\right). \end{aligned}$$

2. Joint entropy and conditional entropy

Example (cont'd). It follows (in bits):

$$
\begin{aligned}
H(X,Y) &= \sum\sum p(x,y)\log_2 p(x,y) = \tfrac{27}{8} \\
H(X) &= \sum p(x)\log_2 p(x) = \tfrac{7}{4} \\
H(Y) &= \sum p(y)\log_2 p(y) = 2 \\
H(X|Y) &= \sum\sum p(x,y)\log_2 p(x|y) = \tfrac{11}{8} \\
H(Y|X) &= \sum\sum p(x,y)\log_2 p(y|x) = \tfrac{13}{8}
\end{aligned}
$$

3. Mutual information

Definition. Let X and Y be two rv with a joint probability function $p(x,y)$ and marginal probability functions $p(x)$ and $p(y)$, respectively. The *mutual information* $I(X;Y)$ is

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}.$$

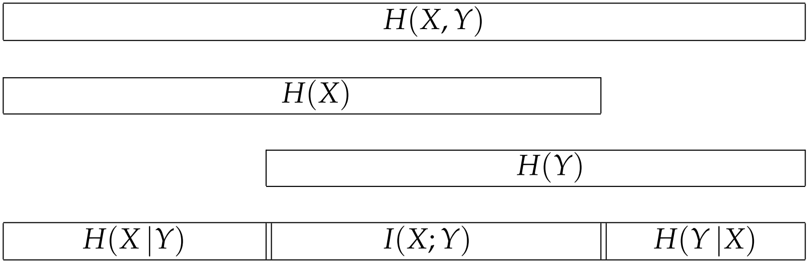
Interpretation: $I(X;Y)$ is the information on X due to the knowledge of Y, as well as the information on Y due to the knowledge of X.

3. Mutual information

Properties.

- $I(X;Y) \geq 0$
- $I(X;Y) = 0$ iff X and Y are independent
- $I(X;Y) = I(Y;X)$
- $I(X;Y) = H(X) - H(X|\,Y) = H(Y) - H(Y|\,X)$
- $I(X;Y) = H(X) + H(Y) - H(X,Y)$
- $I(X;X) = H(X)$
- $I(X;Y) \geq I(X;\varphi(Y))$ for any map φ (*data processing inequality*)
- $I(X;Y) = I(X;\varphi(Y))$ if φ is one-to-one

Graphical summary:



4. The multivariate case

To go from the previous univariate and bivariate cases to the n-variate case, just consider $X_1, ..., X_n$ a vector-valued rv. For example,

$$\begin{aligned} I(X_1, ..., X_n; Y_1, ..., Y_m) &= H(X_1, ..., X_n) + H(Y_1, ..., Y_m) \\ &\quad - H(X_1, ..., X_n, Y_1, ..., Y_m). \end{aligned}$$

Theorem (*Chain rule for entropy*). If $X_1, X_2, ..., X_n$ are rv with joint probability function $p(x_1, ..., x_n)$, then

$$\begin{aligned} H(X_1, X_2, ..., X_n) &= H(X_1) + H(X_2 \,|X_1) + H(X_3 \,|X_2, X_1) + ... \\ &\quad + H(X_n \,|X_{n-1}, ..., X_1) \\ &= \sum_{i=1}^{n} H(X_i \,|X_{i-1}, ..., X_1)\,. \end{aligned}$$

4. The multivariate case

There is a similar *chain rule for the mutual information*:

$$\begin{aligned} I(X_1, X_2, ..., X_n; Y) &= I(X_1; Y) + I(X_2; Y\,|X_1) + I(X_3; Y\,|X_2, X_1) + ... \\ &\quad + I(X_n; Y\,|X_{n-1}, ..., X_1) \\ &= \sum_{i=1}^{n} I(X_i; Y\,|X_{i-1}, ..., X_1)\,. \end{aligned}$$

5. Random processes

Random processes model the repetition of a random experiment in time.

Definition. A (discrete-time) *random process* $\mathbf{X}$ is a one-sided sequence

$$\{X_n\}_{n\in\mathbb{N}} := X_1, X_2, ..., X_n, ... \text{ (or } \{X_n\}_{n\in\mathbb{N}_0} := X_0, X_1, ..., X_n, ...)$$

or a two-sided sequence

$$\{X_n\}_{n\in\mathbb{Z}} := ..., X_{-n}, ..., X_{-1}, X_0, X_1, ..., X_n, ...$$

of rv with the same alphabet $\mathcal{X}$ (but not necessarily with the same probability functions).

Remark. In *Information Theory*, random processes are supposed to be one-sided.

X is characterized by the joint probability functions

$$\Pr\{X_{n_1} = x_1, ..., X_{n_k} = x_k\}$$

for all $k \geq 1$ and $n_1, ..., n_k$.

Definition. A random process is said to be *stationary* if

$$\Pr\{X_{n_1} = x_1, ..., X_{n_k} = x_k\} = \Pr\{X_{n_1+h} = x_1, ..., X_{n_k+h} = x_k\}$$

for every $k, h \geq 0$, and every $x_1, ..., x_k \in \mathcal{X}$.

Interpretation: The statistical properties do not depend on 'time'.

5. Random processes

Finite-alphabet stationary random processes model *information sources*.

$$\boxed{\quad\mathbf{X}\quad} ======> x_1x_2...x_n...$$

- $(x_n)_{n\geq 1} = x_1, x_2, ...$ is a *message* output by the source.
- Each block $x_k^{k+L-1} = x_k, x_{k+1}, ..., x_{k+L-1}$ is a *word*.

Example. $\mathbf{X} = X_1, X_2, ...$ is said to be a *Markov process* if for $n = 1, 2, ...$

$$p\left(x_{n+1} \middle| x_n, x_{n-1}, ..., x_1\right) = p\left(x_{n+1} \middle| x_n\right)$$

for all $x_1, ..., x_n, x_{n+1} \in \mathcal{X}$. It follows

$$p(x_1, x_2, ..., x_n) = p(x_1) p\left(x_2 \middle| x_1\right) p\left(x_3 \middle| x_2\right) \cdots p\left(x_n \middle| x_{n-1}\right).$$

If $X_1, X_2, ..$ are independent rv, then

$$p\left(x_{n+1} \middle| x_n, x_{n-1}, ..., x_1\right) = p(x_{n+1})$$

for any $n \geq 1$. Such processes are also called *memoryless*.

Example.

1. *Coin tossing*: $p(1) = p, p(0) = 1 - p$. Then

$$p\left(x_6 = 1 \middle| x_5 = 0, x_4 = 1, x_3 = 1, x_2 = 0, x_1 = 0\right) = p(1) = p.$$

2. *English language*:

$$p(x_6 = P \,|x_5 = O, x_4 = R, x_3 = T, x_2 = N, x_1 = E) \geq \tfrac{5}{7}$$

(entrochal, entrochite, entropic, entropically, entropion, entropium, entropy).

5. Random processes

Definition. The *entropy* (rate) of a random process $\mathbf{X} = \{X_n\}_{n\geq 1}$ is

$$\begin{aligned} h(\mathbf{X}) &= \lim_{n\to\infty} \frac{1}{n} H(X_1, ..., X_n) \\ &= -\lim_{n\to\infty} \frac{1}{n} \sum_{x_1,...,x_n \in \mathcal{X}} p(x_1, ..., x_n) \log p(x_1, ..., x_n), \end{aligned}$$

provided the limit exists.

Remarks.

- The units of $h(\mathbf{X})$ are bits/symbol, nats/symbol, dits/symbol, etc.
- The expression

$$h(X_1, ..., X_n) = \frac{1}{n} H(X_1, ..., X_n)$$

is called the *entropy of order* n.

If $\mathbf{X}$ is stationary, then $h(\mathbf{X})$ always exists and $h(\mathbf{X}) \leq \log|\mathcal{X}|$.

Theorem. If $\mathbf{X} = \{X_n\}_{n\geq 1}$ is a *stationary* random process, then

$$\lim_{n\to\infty} H\left(X_n \middle|\, X_{n-1}, ..., X_1\right) \searrow h(\mathbf{X}).$$

Consequences.

- $h(X_1, ..., X_n)$ and $H\left(X_n \middle|\, X_{n-1}, ..., X_1\right)$ overestimate $h(\mathbf{X})$.
- Independent processes are the *least predictable*, hence the most random ones.

Example.

1. If $\mathbf{X}$ is *i.i.d.*, then

$$h(\mathbf{X}) = \lim_{n\to\infty} \frac{H(X_1, ..., X_n)}{n} = \lim_{n\to\infty} \frac{nH(X_1)}{n} = H(X_1).$$

2. If $\mathbf{X}$ is an m-state stationary Markov process with *probability transition matrix*

$$P = (P_{ij})_{1\le i,j\le m}, \text{ where } P_{ij} := \Pr\{X_{n+1} = j | \, X_n = i\}$$

and *stationary probability distribution*

$$\mathbf{p} = (p_1, ..., p_m), \text{ where } \mathbf{p}P = \mathbf{p},$$

then

$$h(\mathbf{X}) = -\sum_{i=1}^{m}\sum_{j=1}^{m} p_i P_{ij} \log P_{ij}.$$

Other information-theoretical quantities can be also extended from random variables to random processes.

Definition. The *mutual information* between two stationary random processes $\mathbf{X} = \{X_i\}$ and $\mathbf{Y} = \{Y_j\}$ is given by

$$i(X;Y) = \lim_{n\to\infty} \frac{1}{n} I(X_1, ..., X_n; Y_1, ..., Y_n).$$

6. Estimation of the entropy rate

The estimation of $h(\mathbf{X})$ in practice faces two basic obstacles:

- Real life data sets are finite, while the $h(\mathbf{X})$ involves an infinite limit.
- The convergence of $h(X_1, ..., X_n) \rightarrow h(\mathbf{X})$ is slow.

We consider two methods:

1. *Maximum likelihood*, *naive* or *plug-in* estimation (MLE)
2. *Lempel-Ziv complexity* (LZC).

6.1. Estimation of the entropy rate: MLE

Task: Estimate $h(\mathbf{X})$ from a word $x_1^N = x_1, ..., x_N$ output by $\mathbf{X}$.

Naive solution:

$$h(\mathbf{X}) = \lim_{n\to\infty} h(X_1, ..., X_n) \simeq \hat{h}(X_1, ..., X_n) \text{ with } n \gg 1,$$

where $\hat{h}(X_1, ..., X_n)$ is the so-called *maximum likelihood estimator*

$$\hat{h}(X_1, ..., X_n) = -\frac{1}{n} \sum \hat{p}(x_1, ..., x_n) \log \hat{p}(x_1, ..., x_n),$$

where $\hat{p}(x_1, ..., x_n)$ is the *nth order empirical distribution*, i.e.,

$$\hat{p}(x_1, ..., x_n) = \frac{1}{N-n-1} \sum_{i=1}^{N-n-1} \mathbf{1}(X_i = x_1, ..., X_{i+n-1} = x_n)$$

where $\mathbf{1}(\cdot)$ is the *indicator function*.

6.1. Estimation of the entropy rate: MLE

Problem: As the window size n grows, we run into trouble.

1. The number of windows (i.e. samples) decreases as $N - n + 1$.
2. The number of length-n blocks $x_1, ..., x_n$ grows as $(\#\mathcal{X})^n$.

This situation is called *undersampling*.

Example. Illustration of undersampling with a 2-state Markov process.

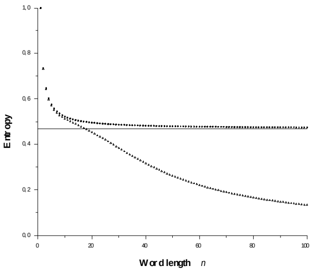


Figure: Entropy estimation of a 2-state Markov chain with transition probabilities $p_{01} = p_{10} = 0.1$ ($h(\mathbf{X}) = 0.469$ bits/symbol).

Remedies.

- *Algebraic*: algebraic correction terms[2].
- *Graphical*: extrapolation of the scaling region[3].

[2] P. Grassberger, Phys. Lett. A 128 113 (1985) 369. L. Paninski, Neural Comp. 15 (2003) 1191.

[3] Strong et al., Phys. Rev. Lett. 80 (1998) 197.

Example. Extrapolating the scaling region over the undersampling region

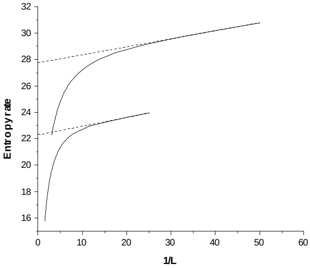


Figure: Extrapolating the linear part of $h(X_1, ..., X_L)$ vs $1/L$, over the undersampling region.

6.2. Estimation of the entropy rate: LZC

Lempel-Ziv complexity is based on *pattern matching*.

Applications:

- Data compression (WinZip, etc.)
- Estimation of the entropy

Versions: LZ76, LZ78,...

6.2 Estimation of the entropy rate: LZC

Given a finite message $x_1^N = x_1, x_2, ..., x_N$, LZ76 decomposes it in *minimal blocks*.

Example. Decomposition of $x_1^{19} = 0101101000110111001 0$.

$$
\begin{array}{lcl}
\mathbf{0}101101000110111 0010 & \rightarrow & 0|\,101101000110111 0010 \\
0|\,\mathbf{1}01101000110111 0010 & \rightarrow & 0|\,1|\,01101000110111 0010 \\
0|\,1|\,\mathbf{011}01000110111 0010 & \rightarrow & 0|\,1|\,\mathbf{011}|\,01000110111 0010
\end{array}
$$

etc. At the end:

$$x_1^{19} \rightarrow 0|\,1|\,011|\,0100|\,011011|\,1001|\,0$$

Thus, x_1^{19} has been decomposed into 7 minimal blocks.

6.2. Estimation of the entropy rate: LZC

Definition. Given a word $x_1^N = x_1, x_2, ..., x_N$ with $\#\mathcal{X} = k$,

- the *complexity* of x_1^N, $C(x_1^N)$, is the number of its minimal blocks,
- the *normalized complexity* of x_1^N is

$$c(x_1^N) = \frac{C(x_1^N)}{N/\log_k N} = \frac{C(x_1^N)}{N}\log_k N.$$

In the preceding example: $C(x_1^{19}) = 7$, hence

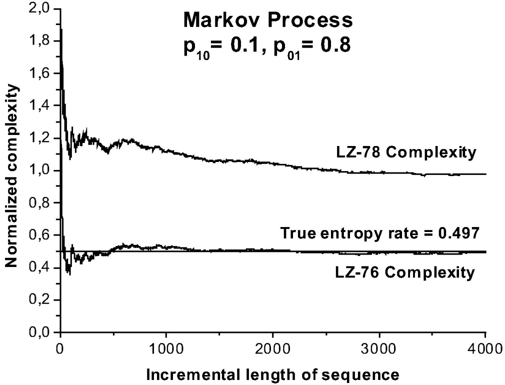
$$c(x_1^{19}) = \tfrac{7}{19}\log_2 19 = 1.565 \text{ bits/symbol}$$

- A finite-alphabet process is *ergodic* if it is memoryless on sufficiently long time scales.
- An ergodic process is the most general dependent process for which the *Strong Law of Large Numbers* holds.

Theorem. If $\mathbf{X}$ is an *ergodic* process, then

$$\lim_{N\to\infty} c(x_1^N) = h(\mathbf{X}) \text{ with probability 1.}$$

Numerical simulation[4].



[4] J.M. Amigó et al, Neural Comp. 16 (2004) 717.

References

1. R.B. Ash, *Information Theory*. Dover Publications, New York, 1990.
2. T.M. Cover and J.A. Thomas, *Elements of Information Theory*, 2nd edition. New York, John Wiley & Sons, 2006.
3. D. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003.
4. L. Paninski, Estimation of entropy and mutual information, *Neural Computation* **15** (2003) 1191.