Entropy and Semantic: a mathematical approach to Authorship Attribution, plagiarism detection and key words extraction

Workshop on "Web Information and Quality Evaluation"
Universidad Politécnica de Valencia

M. Degli Esposti

desposti@dm.unibo.it

Department of Mathematics
University of Bologna

13-15 September 2010

Main objective of the talk

1. present a (narrow) point of view from mathematical-physics on Automatic Text categorization and information retrieval in general

Main objective of the talk

1. present a (narrow) point of view from mathematical-physics on Automatic Text categorization and information retrieval in general
2. bring to your attention some recent results that appeared in the community of mathematics and physics

Main objective of the talk

1. present a (narrow) point of view from mathematical-physics on Automatic Text categorization and information retrieval in general
2. bring to your attention some recent results that appeared in the community of mathematics and physics
3. discuss a "simple" question: how far can we go just with "entropy" (or related) , without linguistics and computational linguistics ?

Simple, but important, observations...

Although the information to be encoded by language is usually highly complex, it can be readily projected onto a string of words.

Simple, but important, observations...

Although the information to be encoded by language is usually highly complex, it can be readily projected onto a string of words.

In recent years the use of tools drawn from statistical physics and dynamical systems has quantitatively revealed rich linguistic structures at many scales, ranging from the domain of syntax to the organization of whole lexicons and literary corpora.

Simple, but important, observations...

Although the information to be encoded by language is usually highly complex, it can be readily projected onto a string of words.

In recent years the use of tools drawn from statistical physics and dynamical systems has quantitatively revealed rich linguistic structures at many scales, ranging from the domain of syntax to the organization of whole lexicons and literary corpora.

However, a fundamental question that has not been directly addressed so far is how statistical structures relate to the function of encoding complex information

Simple, but important, observations...

Although the information to be encoded by language is usually highly complex, it can be readily projected onto a string of words.

In recent years the use of tools drawn from statistical physics and dynamical systems has quantitatively revealed rich linguistic structures at many scales, ranging from the domain of syntax to the organization of whole lexicons and literary corpora.

However, a fundamental question that has not been directly addressed so far is how statistical structures relate to the function of encoding complex information

two recent papers....

In the following two papers quantitative measures have been introduce to captures the relationship between the statistical structure of word sequences and their semantic content.

two recent papers....

In the following two papers quantitative measures have been introduce to captures the relationship between the statistical structure of word sequences and their semantic content.

- E Alvarez-Lacalle, B Dorow, JP Eckmann and E Moses: "Hierarchical structures induce long-range dynamical correlations in written texts", *Proceedings of the National Academy of Sciences*, **103** (21), pp. 7956-7961 (2006)

two recent papers....

In the following two papers quantitative measures have been introduce to captures the relationship between the statistical structure of word sequences and their semantic content.

- E Alvarez-Lacalle, B Dorow, JP Eckmann and E Moses: "Hierarchical structures induce long-range dynamical correlations in written texts", *Proceedings of the National Academy of Sciences*, **103** (21), pp. 7956-7961 (2006)

- M. A. Montemurro and D. Zanette: "Towards the quantification of the semantic information encoded in written language", *arxiv.org/abs/0907.1558v2* (2009)

a (very) big vector space and few notations

W_{all} is a vector space in which each word of the English language represents a base vector.

a (very) big vector space and few notations

W_{all} is a vector space in which each word of the English language represents a base vector.

Given a text $x \in \mathcal{A}^*$, we restrict the analysis to the subspace W_{text} of the words appearing at least once in x.

a (very) big vector space and few notations

W_{all} is a vector space in which each word of the English language represents a base vector.

Given a text $x \in \mathcal{A}^*$, we restrict the analysis to the subspace W_{text} of the words appearing at least once in x.

$\mathcal{D} = \mathcal{D}(x)$ is the dictionary of x, i.e. the set of distinct words, ordered using for example the rank.

a (very) big vector space and few notations

W_{all} is a vector space in which each word of the English language represents a base vector.

Given a text $x \in \mathcal{A}^*$, we restrict the analysis to the subspace W_{text} of the words appearing at least once in x.

$\mathcal{D} = \mathcal{D}(x)$ is the dictionary of x, i.e. the set of distinct words, ordered using for example the rank.

At each word ω_j is associated a *canonical vector* $\mathbf{e}_j$.

a (very) big vector space and few notations

W_{all} is a vector space in which each word of the English language represents a base vector.

Given a text $x \in \mathcal{A}^*$, we restrict the analysis to the subspace W_{text} of the words appearing at least once in x.

$\mathcal{D} = \mathcal{D}(x)$ is the dictionary of x, i.e. the set of distinct words, ordered using for example the rank.

At each word ω_j is associated a *canonical vector* $\mathbf{e}_j$.

Arbitrary directions in this vector space are therefore combinations of words.

a (very) big vector space and few notations

W_{all} is a vector space in which each word of the English language represents a base vector.

Given a text $x \in \mathcal{A}^*$, we restrict the analysis to the subspace W_{text} of the words appearing at least once in x.

$\mathcal{D} = \mathcal{D}(x)$ is the dictionary of x, i.e. the set of distinct words, ordered using for example the rank.

At each word ω_j is associated a *canonical vector* $\mathbf{e}_j$.

Arbitrary directions in this vector space are therefore combinations of words.

Among these combinations one is interested in those that represents certain *topics*, or *concepts* that are discussed in the text.

the *window of attention*....

These word groups are looked for within a *window of attention* of words-size a, e.g. $a = 200$ words .

the *window of attention*....

These word groups are looked for within a *window of attention* of words-size a, e.g. $a = 200$ words .

This window represents the words that *have just been read*, and these comprise at each point of the text a momentary alert*vector of attention*.....

the *window of attention*....

These word groups are looked for within a *window of attention* of words-size a, e.g. $a = 200$ words .

This window represents the words that *have just been read*, and these comprise at each point of the text a momentary alert*vector of attention*.....

but first, the *corpus*...(and the stemming)

the Corpus

In Eckmann's paper, the authors used 12 books in their English version.

the Corpus

In Eckmann's paper, the authors used 12 books in their English version.

Nine of them were novels :

- *War and Peace* (WP) by Tolstoi,
- *Don Quixote* (QJ) by Cervantes,
- *The Iliad* (IL) by Homer,
- *Moby-Dick* or *The Whale* (MD) by Melville,
- *David Crockett* (DC) by Abbott,
- *The adventure of Tom Sawyer* (TS) by Twain,
- *Naked Lunch* (NK) by Burroughs,
- *Hamlet* (HM) by Shakespeare,
- *The Metamorphosis* (MT) by Kafka.

the Corpus

In Eckmann's paper, the authors used 12 books in their English version.

Nine of them were novels :

- *War and Peace* (WP) by Tolstoi,
- *Don Quixote* (QJ) by Cervantes,
- *The Iliad* (IL) by Homer,
- *Moby-Dick* or *The Whale* (MD) by Melville,
- *David Crockett* (DC) by Abbott,
- *The adventure of Tom Sawyer* (TS) by Twain,
- *Naked Lunch* (NK) by Burroughs,
- *Hamlet* (HM) by Shakespeare,
- *The Metamorphosis* (MT) by Kafka.

In addition :

- *Relativity: The Special and the General Theory* (EI) by Einstein
- *Critique of Pure Reason* (KT) by Kant
- *The Republic* (RP) by Plato.

the Corpus

Book	Length	m_{thr}	d_{thr}	P	d_{conv}	Exponent
MT	22,375	4	377	17.6	25	0.45 (0.05)
HM	32,564	5	446	16.4	30	0.95 (0.07)
NK	62,190	8	762	20.6	60	0.80 (0.05)
TS	73,291	8	669	17.5	40	0.47 (0.04)
DC	77,728	8	816	20.5	80	0.45 (0.08)
IL	152,400	12	830	22.7	70	0.38 (0.04)
MD	213,682	14	1,177	20.2	70	0.44 (0.05)
QJ	402,870	20	1,293	19.6	75	0.36 (0.03)
WP	529,547	23	1,576	24.3	200	0.45 (0.05)
EI	30,715	5	474	26.4	50	0.85 (0.10)
RP	118,661	11	628	15.6	70	0.57 (0.05)
KT	197,802	14	704	27.9	50	0.30 (0.03)

Figure: **Corpus parameters and results:** m_{thr} is the threshold for the number of occurrences and d_{thr} is the number of words kept after thresholding. P is the percentage of the words in the book that passes the threshold, $P = \sum_{j=1}^{d_{\text{thr}}} m_j / L$. d_{conv} is the dimension at which a power law is bring fit. The absolute values of the negative exponents of the fit are given in the last column, together with their error in parentheses.

Cleaning and Stemming

Each of the book was processed by eliminating punctuation and extracting the words.

Cleaning and Stemming

Each of the book was processed by eliminating punctuation and extracting the words.

Each word has been *stemmed* by querying `WORDNET 2.0`.

Cleaning and Stemming

Each of the book was processed by eliminating punctuation and extracting the words.

Each word has been *stemmed* by querying `WORDNET 2.0`.

The *leading* word for this query was retained, keeping the information on whether it was originally a noun, a verb, or an adjective.

Cleaning and Stemming

Each of the book was processed by eliminating punctuation and extracting the words.

Each word has been *stemmed* by querying `WORDNET 2.0`.

The *leading* word for this query was retained, keeping the information on whether it was originally a noun, a verb, or an adjective.

A list of *stop words* that carry no significant meaning has been defined and at each of them were assigned a value of zero:determiners, pronouns, and the like

Cleaning and Stemming

Each of the book was processed by eliminating punctuation and extracting the words.

Each word has been *stemmed* by querying `WORDNET 2.0`.

The *leading* word for this query was retained, keeping the information on whether it was originally a noun, a verb, or an adjective.

A list of *stop words* that carry no significant meaning has been defined and at each of them were assigned a value of zero:determiners, pronouns, and the like

Moreover were rejected those words that occur significantly in at least 11 of the 12 texts in the corpus.

Cleaning and Stemming

Each of the book was processed by eliminating punctuation and extracting the words.

Each word has been *stemmed* by querying `WORDNET 2.0`.

The *leading* word for this query was retained, keeping the information on whether it was originally a noun, a verb, or an adjective.

A list of *stop words* that carry no significant meaning has been defined and at each of them were assigned a value of zero:determiners, pronouns, and the like

Moreover were rejected those words that occur significantly in at least 11 of the 12 texts in the corpus.

Books were thus transformed into a list of *stemmed words*, and used for constructing the mathematical objects we will now discuss. ....

the vector of attention

Fix a window size a (e.g. $a = 200$ words).

the vector of attention

Fix a window size a (e.g. $a = 200$ words).

We define its (normalized) *vector of attention* $\mathbf{V}$ as:

$$\mathbf{V} = \left[\sum_j m_j^2(a)\right]^{-\frac{1}{2}} \sum_j m_j(a)\mathbf{e}_j,$$

where the sum can be thought over all dictionary $\mathcal{D}(x)$.

the vector of attention

Fix a window size a (e.g. $a = 200$ words).

We define its (normalized) *vector of attention* **V** as:

$$\mathbf{V} = \left[\sum_j m_j^2(a)\right]^{-\frac{1}{2}} \sum_j m_j(a)\mathbf{e}_j,$$

where the sum can be thought over all dictionary $\mathcal{D}(x)$.

Now we would like to *project* the vector **V** onto a *smaller subspace related with different concepts or themes that appear in the text*.

Symmetric Connectivity Matrix

The starting point is the construction of a *symmetric connectivity matrix* M.

Symmetric Connectivity Matrix

The starting point is the construction of a *symmetric connectivity matrix* M.

Definition (The symmetric connectivity matrix M)

Given a text x, the matrix M has rows and columns indexed by words, and the entry M_{ij} counts how often word ω_i occurs within a distance $a/2$ on either side of word ω_j.

the Normalized Symmetric Connectivity Matrix

The connectivity matrix R of an equivalent *random/shuffled book* :

$$R_{ij} = \frac{a}{L} m_i m_j,$$

the Normalized Symmetric Connectivity Matrix

The connectivity matrix R of an equivalent *random/shuffled book* :

$$R_{ij} = \frac{a}{L} m_i m_j,$$

Definition

Given a text x and a context length a, the normalized connectivity matrix N is defined as:

$$N_{ij} \; = \; R_{ij}^{-\frac{1}{2}} \left(M_{ij} - R_{ij} \right).$$

the Normalized Symmetric Connectivity Matrix

The connectivity matrix R of an equivalent *random/shuffled book* :

$$R_{ij} = \frac{a}{L} m_i m_j ,$$

Definition

Given a text x and a context length a, the normalized connectivity matrix N is defined as:

$$N_{ij} \; = \; R_{ij}^{-\frac{1}{2}} \left(M_{ij} - R_{ij} \right) .$$

This normalization quantifies the extent to which the analyzed text deviates from a random book (with the same words distribution) measured in units of its standard deviation.

Projecting down: SVD

We now project onto a smaller subspace by keeping only those d basis vector with *highest singular values*.

Projecting down: SVD

We now project onto a smaller subspace by keeping only those d basis vector with *highest singular values*.

The idea behind this choice of *principal directions* is that the most important vectors in this decomposition describe *concepts*.

Projecting down: SVD

We now project onto a smaller subspace by keeping only those d basis vector with *highest singular values*.

The idea behind this choice of *principal directions* is that the most important vectors in this decomposition describe *concepts*.

Given d vectors fro the SVD basis, every word can be projected onto a unique superposition of those basic vectors, i.e.:

$$\mathbf{e}_k \to \sum_{j=1}^{d} S_{kj}\mathbf{v}_j,$$

where $\mathbf{e}_k$ is the *canonical* vector representing word ω_k.

few experiments......

Table 2. Examples of the highest singular components for three books

MD(1)	MD(5)	EI(1)	EI(2)	TS(1)	TS(2)
whale	bed	surface	*planet*	spunk	*ticket*
ahab	room	euclidean	*sun*	wart	*bible*
starbuck	queequeg	rod	*ellipse*	nigger	*verse*
sperm	*dat*	continuum	*mercury*	huck	*blue*
boat	*aye*	geometry	*orbital*	tell	*pupil*
cry	door	universe	*orbit*	stump	*yellow*
aye	*moby*	curve	*star*	johnny	*ten*
stubb	*dick*	numbers	*angle*	reckon	*spunk*
sir	landlord	slab	*arc*	bet	*thousand*
leviathan	*ahab*	plane	*newton*	water	*red*

Given are component one and five of *Moby-Dick* (MD), one and two of Einstein (EI) and of *Tom Sawyer* (TS). The coefficients of the words in the singular component may be positive (plain text) or negative (italic), and their absolute values range from 0.1 to 0.37.

a dynamic analysis

The idea is now to slide the *window of attention* of fixed size $a = 200$ along the text and observe how the corresponding vectors **V** moves in the vector space spanned by the SVD.

a dynamic analysis

The idea is now to slide the *window of attention* of fixed size $a = 200$ along the text and observe how the corresponding vectors $\mathbf{V}$ moves in the vector space spanned by the SVD.

If this vector space were irrelevant to the text, then the trajectory defined in this space would perform a *random walk*.

a dynamic analysis

The idea is now to slide the *window of attention* of fixed size $a = 200$ along the text and observe how the corresponding vectors $\mathbf{V}$ moves in the vector space spanned by the SVD.

If this vector space were irrelevant to the text, then the trajectory defined in this space would perform a *random walk*.

If, on the contrary, the evolution of the text is reflected in this vector space, then the trajectory should *trace out the concepts in a systematic way*, and some evidence of this will be observed (and hopefully measured)

Trajectories and time

Trajectories in this vector space can be connected to the process of reading of the text by replacing the notion of distance along the text with the time it takes to read it

$$t = \ell \times \delta t,$$

Trajectories and time

Trajectories in this vector space can be connected to the process of reading of the text by replacing the notion of distance along the text with the time it takes to read it

$$t = \ell \times \delta t,$$

with ℓ the distance into the text and δt the average time it takes a hypothetical reader to read a word.

a dynamic analysis

At each time t we define in this way a vector of attention, $\mathbf{V}(t)$ corresponding to the window $[t/\delta t - a/2, t/\delta t + a/2]$.

a dynamic analysis

At each time t we define in this way a vector of attention, $\mathbf{V}(t)$ corresponding to the window $[t/\delta t - a/2, t/\delta t + a/2]$.

We project the vector $\mathbf{V}(t)$ onto the first d vectors $\mathbf{v}_j$:

$$\mathbf{V}(t) \leftarrow \sum_{j=1}^{d} S_j(t)\mathbf{v}_j,$$

a dynamic analysis

At each time t we define in this way a vector of attention, $\mathbf{V}(t)$ corresponding to the window $[t/\delta t - a/2, t/\delta t + a/2]$.

We project the vector $\mathbf{V}(t)$ onto the first d vectors $\mathbf{v}_j$:

$$\mathbf{V}(t) \leftarrow \sum_{j=1}^{d} S_j(t)\mathbf{v}_j,$$

The *moving unit vector* $\mathbf{V}(t) \in R^d$ is a dynamical system and it is natural to study its *autocorrelation function* in time:

$$C(\tau) = \langle \mathbf{V}(t) \cdot \mathbf{V}(t+\tau) \rangle_t ,$$

where $\langle \cdot \rangle_t$ is the *time average*.

autocorrelation function in *Tom Sawyer*

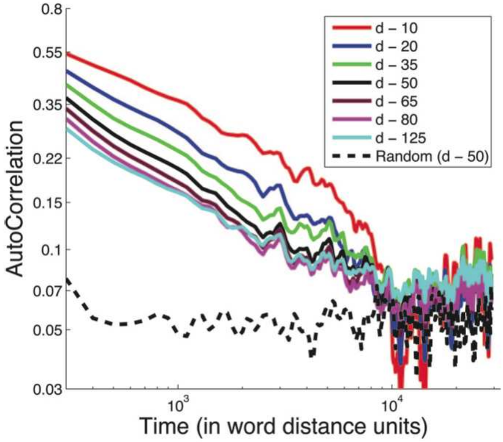


Figure: Log-log plot of the autocorrelation function for the *Adventures of Tom Sawyer* using different numbers of singular components for building the dynamics. For comparison, the autocorrelation of a randomized version of the book is also shown.

autocorrelation function in the other books...

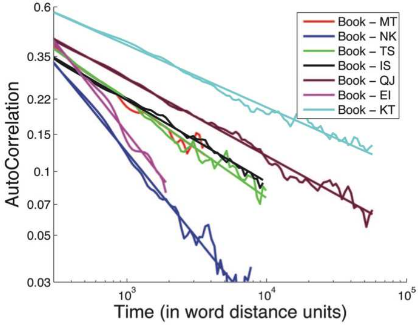


Figure: Autocorrelation functions and fits fro seven of the book listed.

autocorrelation function in the other books...

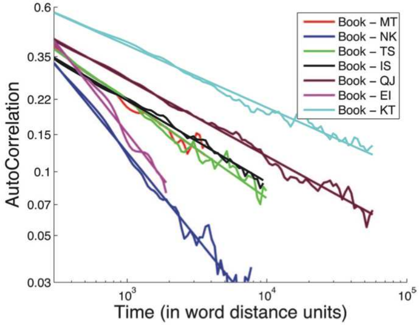


Figure: Autocorrelation functions and fits fro seven of the book listed.

autocorrelation function in the other books...

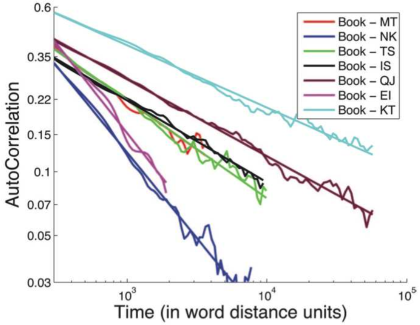
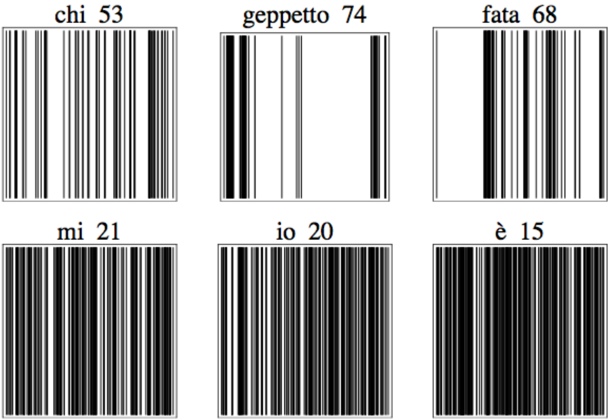


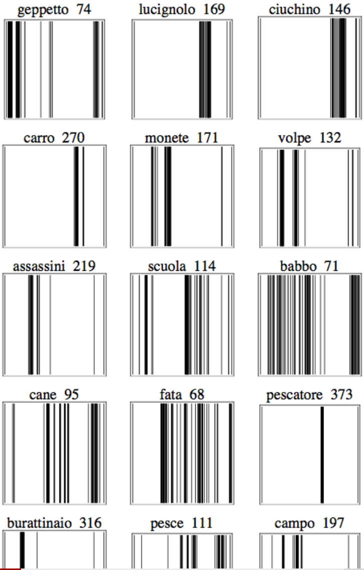
Figure: Autocorrelation functions and fits fro seven of the book listed.

authors claim that this range is much longer than what we found when measuring correlations among sentences, *without* using the concept vectors.

Spectrum of words



Spectrum of words: Pinocchio



Few notations

A given text x of N words is divided in P parts, each of word-length N_j, $j = 1, 2, \ldots, P$.

Few notations

A given text x of N words is divided in P parts, each of word-length N_j, $j = 1, 2, \ldots, P$.

Assume ω is a word that appears n_j times in part j, with $j = 1, \ldots, P$: $\mu(\omega|j) := n_j/N_j$ can be considerate as the conditional probability of finding word ω in part j.

Few notations

A given text x of N words is divided in P parts, each of word-length N_j, $j = 1, 2, \dots, P$.

Assume ω is a word that appears n_j times in part j, with $j = 1, \dots, P$: $\mu(\omega|j) := n_j/N_j$ can be considerate as the conditional probability of finding word ω in part j.

We also denote by $\mu(j) = N_j/N$ the *a priori* probability that the word ω appears in part j, then

$$\sum_{j=1}^{P} \mu(\omega|j)\mu(j) = \mu(\omega),$$

where $\mu(\omega) = n/N$ stands for the overall probability of occurrences of a word in the whole text.

Bayes's rule

We look for the *inverted* probability $\mu(j|\omega)$, which tell us how likely is that we are looking into part j given that we saw an instance of word ω in the text.

$$\mu(j|\omega) = \frac{\mu(\omega|j)\mu(j)}{\sum_{k=1}^{P} \mu(\omega|k)\mu(k)} = n_j/n.$$

Now we can write Shannon mutual

$$I(x,\mathcal{D}) = \sum_{\omega\in\mathcal{D}} \mu(\omega) \sum_{j=1}^{P} \mu(j|\omega) \log\left(\frac{\mu(j|\omega)}{\mu(j)}\right).$$

Entropy of a word (in a given text x)

$$h(x|\omega) := -\sum_{j=1}^{P} \mu(j|\omega) \log \mu(j|\omega), \qquad \mu(j|\omega) = n_j/n$$

Entropy of a word (in a given text x)

$$h(x|\omega) := -\sum_{j=1}^{P} \mu(j|\omega) \log \mu(j|\omega), \qquad \mu(j|\omega) = n_j/n$$

moreover, we also average over shuffling

$$\left\langle \hat{h}(x|\omega) \right\rangle := -\sum_{j=1}^{P} \langle \hat{\mu}(j|\omega) \log \hat{\mu}(j|\omega) \rangle .$$

Definition

Relevant words are ranked w.r.t.

$$h(x|\omega) - \left\langle \hat{h}(x|\omega) \right\rangle$$

Shuffling and Averaging...

We can use elementary methods to compute an analytic expression of the entropy $< \hat{h}(x|\omega) >$.

Shuffling and Averaging...

We can use elementary methods to compute an analytic expression of the entropy $< \hat{h}(x|\omega) >$.

For a word ω that appears m_j times in part j with a frequency n over the text x, this entropy takes the following form:

$$\hat{h}(x|\omega) := -\sum_{j=1}^{P} \frac{m_j}{n} \log \frac{m_j}{n}.$$

Shuffling and Averaging...

We can use elementary methods to compute an analytic expression of the entropy $< \hat{h}(x|\omega) >$.

For a word ω that appears m_j times in part j with a frequency n over the text x, this entropy takes the following form:

$$\hat{h}(x|\omega) := -\sum_{j=1}^{P} \frac{m_j}{n} \log \frac{m_j}{n}.$$

We now compute the average over all possible realizations of the random text:

$$\left\langle \hat{h}(x|\omega) \right\rangle = -\sum_{\substack{m_1+\cdots+m_P=n \\ m_j \leq N/P}} \mu(m_1, \dots, m_P) \sum_{j=1}^{P} \frac{m_j}{n} \log \frac{m_j}{n},$$

where $\mu(m_1, \dots, m_P)$ is the probability of finding m_j words ω in part j, with $j = 1, \dots, P$.

Shuffling and Averaging: we can use symmetry

$$\left\langle \hat{h}(x|\omega) \right\rangle = -P \sum_{m=1}^{\min(n,N/P)} \mu(m) \frac{m}{n} \log \frac{m}{n},$$

where the margin probability $\mu(n)$ is given by the probability of finding m instances of word ω in one part, together with $(N/P - m)$ words different from ω, and reads:

Shuffling and Averaging: we can use symmetry

$$\left\langle \hat{h}(x|\omega) \right\rangle = -P \sum_{m=1}^{\min(n,N/P)} \mu(m) \frac{m}{n} \log \frac{m}{n},$$

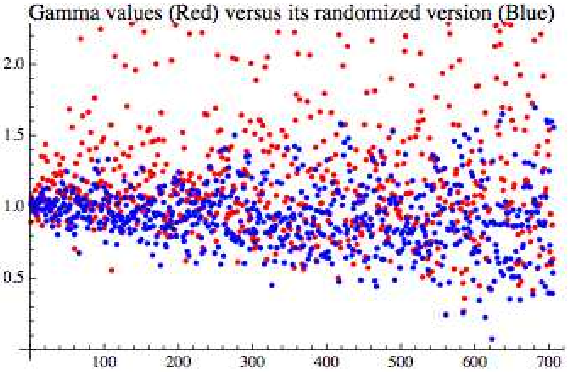
where the margin probability $\mu(n)$ is given by the probability of finding m instances of word ω in one part, together with $(N/P - m)$ words different from ω, and reads:

$$\mu(m) = \frac{\binom{n}{m}\binom{N-n}{N/P-m}}{\binom{N}{N/P}}.$$

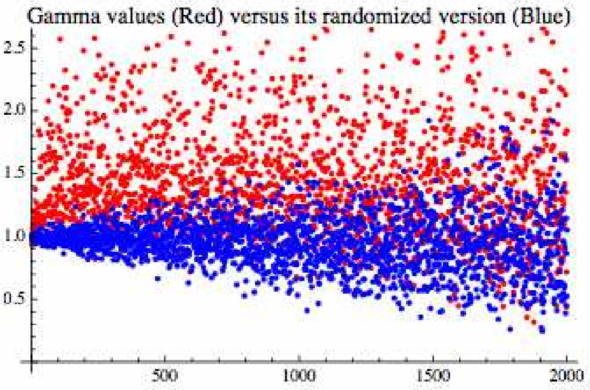
and use Gaussian approximation, to get:

$$\left\langle \hat{h}(x|\omega) \right\rangle \approx 1 - \frac{P-1}{2n \log P}$$

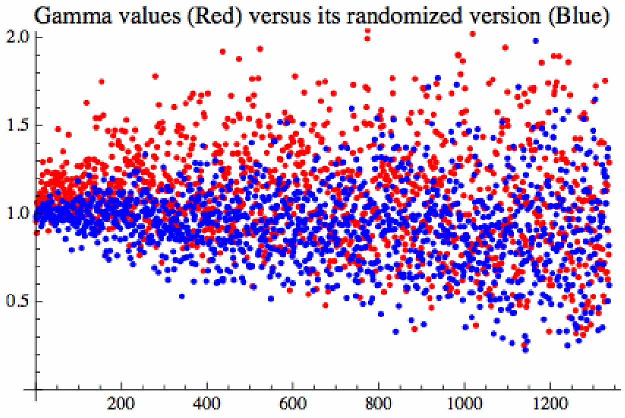
Pinocchio's words Entropy distribution



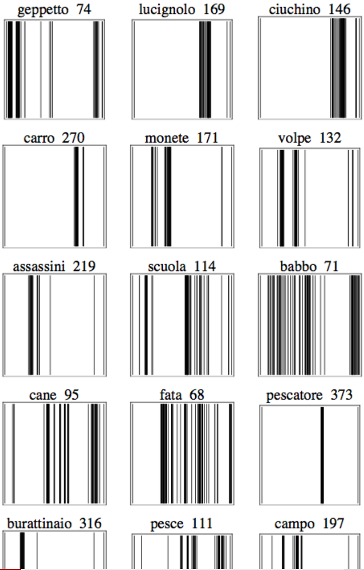
Kant's words Entropy distribution



Dante's words Entropy distribution

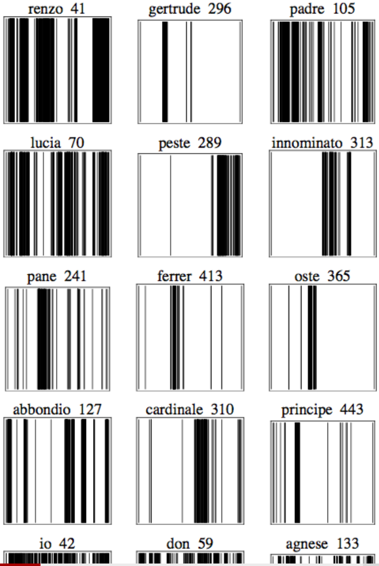


Spectrum of words: Pinocchio



Anna Karerina

Promessi Sposi



A.A. with K-L

Authorship Attribution algorithms based on Relative Entropy (K-L Divergence).

A.A. with K-L

Authorship Attribution algorithms based on Relative Entropy (K-L Divergence).

...we start with wrong assumptions (i.e. the author is a stochastic source) and we end up with interesting results.......

A.A. with K-L

Authorship Attribution algorithms based on Relative Entropy (K-L Divergence).

...we start with wrong assumptions (i.e. the author is a stochastic source) and we end up with interesting results.......

a mathematical problem: given two unknown stochastic (stationary and ergodic) sources μ and ν, compute/approximate the relative entropy

$$d(\mu \| \nu)$$

just by using two finite realizations $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ of μ and ν respectively......

μ ergodic, stationary stochastic source

Just to recall the main defintions..........

μ ergodic, stationary stochastic source

Just to recall the main defintions..........

n-block entropy

$$H_n(\mu) := -\sum_{|\omega|=n} \mu(\omega) \log \mu(\omega).$$

μ ergodic, stationary stochastic source

Just to recall the main defintions..........

n-block entropy

$$H_n(\mu) := -\sum_{|\omega|=n} \mu(\omega)\log\mu(\omega).$$

entropy rate and n-conditional entropy

$$\begin{aligned} h_n(\mu) \quad &\underset{\text{entropy rate}}{:=} \quad H_{n+1}(\mu) - H_n(\mu) \\ &\underset{\text{conditional entropy}}{=} \quad \sum_{\omega_1^n \in \mathcal{A}^n, a\in\mathcal{A}} \mu(\omega_1^n a)\log\mu(a|w_1^n) \\ &:= \quad \mathbb{E}_{\mu_{n+1}}\left(\log\mu(a|\omega_1^n)\right), \end{aligned}$$

μ ergodic, stationary stochastic source

Just to recall the main defintions..........

n-block entropy

$$H_n(\mu) := - \sum_{|\omega|=n} \mu(\omega) \log \mu(\omega).$$

entropy rate and n-conditional entropy

$$\begin{aligned} h_n(\mu) \quad & \underset{\text{entropy rate}}{:=} \quad H_{n+1}(\mu) - H_n(\mu) \\ & \underset{\text{conditional entropy}}{=} \quad \sum_{\omega_1^n \in \mathcal{A}^n, a \in \mathcal{A}} \mu(\omega_1^n a) \log \mu(a|w_1^n) \\ & := \quad \mathbb{E}_{\mu_{n+1}} \left(\log \mu(a|\omega_1^n) \right), \end{aligned}$$

Entropy of μ

$$h(\mu) = \lim_{n\to\infty} \frac{H_n(\mu)}{n} = \lim_{n\to\infty} h_n(\mu) = \mathbb{E}_\mu \left(\log \mu(a|\omega_1^\infty) \right)$$

Cross and Relative entropy: $h(\mu||\nu) = h(\mu) + d(\mu||\nu)$

Cross and Relative entropy: $h(\mu||\nu) = h(\mu) + d(\mu||\nu)$

n-conditional cross entropy:

Cross and Relative entropy: $h(\mu||\nu) = h(\mu) + d(\mu||\nu)$

n-conditional cross entropy:

$$h_n(\mu||\nu) = -\sum_{\omega\in A^n,\, a\in A} \mu(\omega a) \log \nu(a|\omega),$$

Cross and Relative entropy: $h(\mu||\nu) = h(\mu) + d(\mu||\nu)$

n-conditional cross entropy:

$$h_n(\mu||\nu) = -\sum_{\omega\in A^n,\, a\in A} \mu(\omega a) \log \nu(a|\omega),$$

cross entropy

Cross and Relative entropy: $h(\mu||\nu) = h(\mu) + d(\mu||\nu)$

n-conditional cross entropy:

$$h_n(\mu||\nu) = -\sum_{\omega \in A^n,\, a \in A} \mu(\omega a) \log \nu(a|\omega),$$

cross entropy

$$h(\mu||\nu) = \lim_{k \to +\infty} \frac{1}{n} H_k(\mu||\nu) = \lim_{n \to +\infty} h_n(\mu||\nu),$$

Cross and Relative entropy: $h(\mu||\nu) = h(\mu) + d(\mu||\nu)$

n-conditional cross entropy:

$$h_n(\mu||\nu) = -\sum_{\omega\in A^n,\, a\in A} \mu(\omega a)\log\nu(a|\omega),$$

cross entropy

$$h(\mu||\nu) = \lim_{k\to+\infty}\frac{1}{n}H_k(\mu||\nu) = \lim_{n\to+\infty} h_n(\mu||\nu),$$

relative entropy (Kullback-Leibler divergence)

Cross and Relative entropy: $h(\mu||\nu) = h(\mu) + d(\mu||\nu)$

n-conditional cross entropy:

$$h_n(\mu||\nu) = -\sum_{\omega\in A^n,\, a\in A} \mu(\omega a) \log \nu(a|\omega),$$

cross entropy

$$h(\mu||\nu) = \lim_{k\to+\infty} \frac{1}{n} H_k(\mu||\nu) = \lim_{n\to+\infty} h_n(\mu||\nu),$$

relative entropy (Kullback-Leibler divergence)

$$d(\mu||\nu) \qquad =$$

Cross and Relative entropy: $h(\mu||\nu) = h(\mu) + d(\mu||\nu)$

n-conditional cross entropy:

$$h_n(\mu||\nu) = -\sum_{\omega\in A^n,\, a\in A} \mu(\omega a)\log \nu(a|\omega),$$

cross entropy

$$h(\mu||\nu) = \lim_{k\to+\infty} \frac{1}{n} H_k(\mu||\nu) = \lim_{n\to+\infty} h_n(\mu||\nu),$$

relative entropy (Kullback-Leibler divergence)

$$d(\mu||\nu) \qquad = \lim_{n\to\infty} E_\mu \left(\log \frac{\mu(\omega_n|\omega_1^{n-1})}{\nu(\omega_n|\omega_1^{n-1})}\right)$$

Cross and Relative entropy: $h(\mu||\nu) = h(\mu) + d(\mu||\nu)$

n-conditional cross entropy:

$$h_n(\mu||\nu) = - \sum_{\omega \in A^n,\, a \in A} \mu(\omega a) \log \nu(a|\omega),$$

cross entropy

$$h(\mu||\nu) = \lim_{k \to +\infty} \frac{1}{n} H_k(\mu||\nu) = \lim_{n \to +\infty} h_n(\mu||\nu),$$

relative entropy (Kullback-Leibler divergence)

$$\begin{aligned} d(\mu||\nu) \quad &= \lim_{n\to\infty} E_\mu \left(\log \tfrac{\mu(\omega_n|\omega_1^{n-1})}{\nu(\omega_n|\omega_1^{n-1})} \right) \\ &= \lim_{n\to\infty} \textstyle\sum_{\omega_1^n A^n} \mu(\omega_1^n) \log \tfrac{\mu(\omega_n|\omega_1^{n-1})}{\nu(\omega_n|\omega_1^{n-1})}. \end{aligned}$$

Three methods for computing K-L divergence

1. *Zippers*: *cross-parsing* and Merhav-Ziv Theorem

Three methods for computing K-L divergence

1. *Zippers*: *cross-parsing* and Merhav-Ziv Theorem
2. NSRPS: *Non Sequential Recursive Pair Substitution*

Three methods for computing K-L divergence

1. *Zippers*: *cross-parsing* and Merhav-Ziv Theorem
2. NSRPS: *Non Sequential Recursive Pair Substitution*
3. BWT: *The Burrows-Wheeler Transform*

LZ78

In LZ78 a parsing into blocks (often referred to as *words*) of variable length is performed according to the following rule:

LZ78

In LZ78 a parsing into blocks (often referred to as *words*) of variable length is performed according to the following rule:

> *the next word is the shortest word that hasn't been previously seen in the parse*

LZ78

In LZ78 a parsing into blocks (often referred to as *words*) of variable length is performed according to the following rule:

> *the next word is the shortest word that hasn't been previously seen in the parse*

Every new parsed word is added to a *dictionary*, which can then be used for reference to proceed in the parsing.

LZ78

In LZ78 a parsing into blocks (often referred to as *words*) of variable length is performed according to the following rule:

> *the next word is the shortest word that hasn't been previously seen in the parse*

Every new parsed word is added to a *dictionary*, which can then be used for reference to proceed in the parsing.

an example of *LZ78*-parsing

$$a_1^n = \texttt{accbbabcbcbbabbcbcabbb}$$

an example of *LZ78*-parsing

$$a_1^n = \texttt{accbbabcbcbbabbcbcabbb}$$

The final result of the parse is:

$$\texttt{a|c|cb|b|ab|cbc|bb|abb|cbca|bbb}$$

Ziv's Theorem

Theorem

If μ is a stationary ergodic process,

$$\frac{c(n)\log c(n)}{n} \xrightarrow[n\to\infty]{} h_\mu \text{ almost surely}$$

Ziv's Theorem

Theorem

If μ is a stationary ergodic process,

$$\frac{c(n)\log c(n)}{n} \xrightarrow[n\to\infty]{} h_\mu \text{ almost surely}$$

Theorem

(Ziv, Merhav) If X is stationary and ergodic with positive entropy and Y is a Markov chain $P_n \ll Q_n$ asymptotically, then

$$\lim_{n\to\infty} \frac{c_n(x|y)\log n}{n} = h(P) + d(P\|Q) \quad (P \times Q) - a.s.$$

Returning and Waiting times

Entropy and cross entropy can be related to the asymptotic behavior of properly defined returning times and waiting times, respectively.

Returning and Waiting times

Entropy and cross entropy can be related to the asymptotic behavior of properly defined returning times and waiting times, respectively.

returning time

$$R(w_1^n) = \min\{k > 1 : w_k^{k+n-1} = w_1^n\}$$

Returning and Waiting times

Entropy and cross entropy can be related to the asymptotic behavior of properly defined returning times and waiting times, respectively.

returning time

$$R(w_1^n) = \min\{k > 1 : \ w_k^{k+n-1} = w_1^n\}$$

waiting time

$$W(w_1^n, z) = \min\{k > 1 : \ z_k^{k+n-1} = w_1^n\}$$

Returning and Waiting times

Entropy and cross entropy can be related to the asymptotic behavior of properly defined returning times and waiting times, respectively.

returning time

$$R(w_1^n) = \min\{k > 1 : w_k^{k+n-1} = w_1^n\}$$

waiting time

$$W(w_1^n, z) = \min\{k > 1 : z_k^{k+n-1} = w_1^n\}$$

Returning and Waiting times

Entropy and cross entropy can be related to the asymptotic behavior of properly defined returning times and waiting times, respectively.

returning time

$$R(w_1^n) = \min\{k > 1 : w_k^{k+n-1} = w_1^n\}$$

waiting time

$$W(w_1^n, z) = \min\{k > 1 : z_k^{k+n-1} = w_1^n\}$$

Note that $W(w_1^n, w) = R(w_1^n)$.

Two important results

Theorem (Entropy and returning time)

If μ is a stationary, ergodic process, then

$$\lim_{n\to\infty} \frac{1}{n} \log R(w_1^n) = h(\mu) \qquad \mu\text{--}a.s.$$

Two important results

Theorem (Entropy and returning time)

If μ is a stationary, ergodic process, then

$$\lim_{n\to\infty} \frac{1}{n} \log R(w_1^n) = h(\mu) \qquad \mu\text{−}a.s.$$

Theorem (Relative entropy and waiting time)

If μ is stationary and ergodic, ν is k-Markov and $\mu_n << \nu_n$, then

Two important results

Theorem (Entropy and returning time)

If μ is a stationary, ergodic process, then

$$\lim_{n\to\infty} \frac{1}{n} \log R(w_1^n) = h(\mu) \qquad \mu\text{−}a.s.$$

Theorem (Relative entropy and waiting time)

If μ is stationary and ergodic, ν is k-Markov and $\mu_n << \nu_n$, then

$$\lim_{n\to\infty} \frac{1}{n} \log W(w_1^n, z) = h(\mu) + d(\mu||\nu) = h(\mu||\nu), \qquad (\mu \times \nu)\text{−}a.s.$$

A *real* scenario: Gramsci's articles

A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

A *real* scenario: Gramsci's articles


A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

- During the period 1914-1928, Gramsci produced an enormous number of articles on different national newspaper.

A *real* scenario: Gramsci's articles


A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

- During the period 1914-1928, Gramsci produced an enormous number of articles on different national newspaper.
- Most of these article (hundreds, if not thousands) are NOT signed

A *real* scenario: Gramsci's articles


A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

- During the period 1914-1928, Gramsci produced an enormous number of articles on different national newspaper.
- Most of these article (hundreds, if not thousands) are NOT signed
- Other possible authors : Bordiga, Serrati, Tasca, Togliatti...

A *real* scenario: Gramsci's articles


A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

- During the period 1914-1928, Gramsci produced an enormous number of articles on different national newspaper.
- Most of these article (hundreds, if not thousands) are NOT signed
- Other possible authors : Bordiga, Serrati, Tasca, Togliatti...

A *real* scenario: Gramsci's articles


A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

- During the period 1914-1928, Gramsci produced an enormous number of articles on different national newspaper.
- Most of these article (hundreds, if not thousands) are NOT signed
- Other possible authors : Bordiga, Serrati, Tasca, Togliatti...the aim is to recognize the articles really written by A. Gramsci...

A *real* scenario: Gramsci's articles

A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

- During the period 1914-1928, Gramsci produced an enormous number of articles on different national newspaper.
- Most of these article (hundreds, if not thousands) are NOT signed
- Other possible authors : Bordiga, Serrati, Tasca, Togliatti...the aim is to recognize the articles really written by A. Gramsci...
- Quite positive results for the period 1915-1917

A *real* scenario: Gramsci's articles


A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

- During the period 1914-1928, Gramsci produced an enormous number of articles on different national newspaper.
- Most of these article (hundreds, if not thousands) are NOT signed
- Other possible authors : Bordiga, Serrati, Tasca, Togliatti...the aim is to recognize the articles really written by A. Gramsci...
- Quite positive results for the period 1915-1917

A *real* scenario: Gramsci's articles


A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

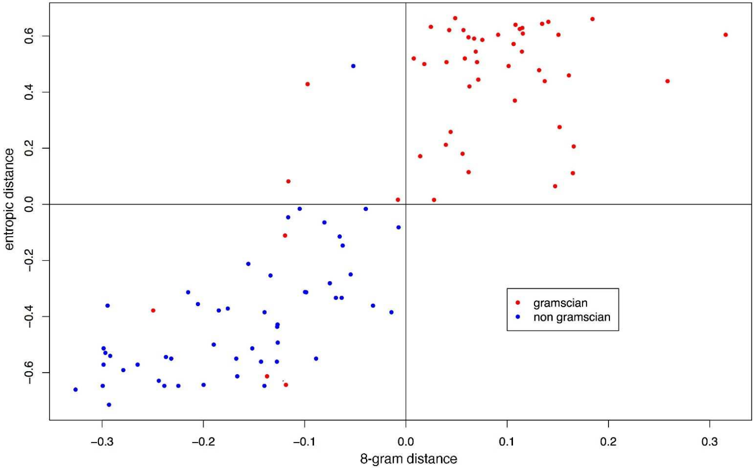
- During the period 1914-1928, Gramsci produced an enormous number of articles on different national newspaper.
- Most of these article (hundreds, if not thousands) are NOT signed
- Other possible authors : Bordiga, Serrati, Tasca, Togliatti...the aim is to recognize the articles really written by A. Gramsci...
- Quite positive results for the period 1915-1917 (!?!?)

A *real* scenario: Gramsci's articles

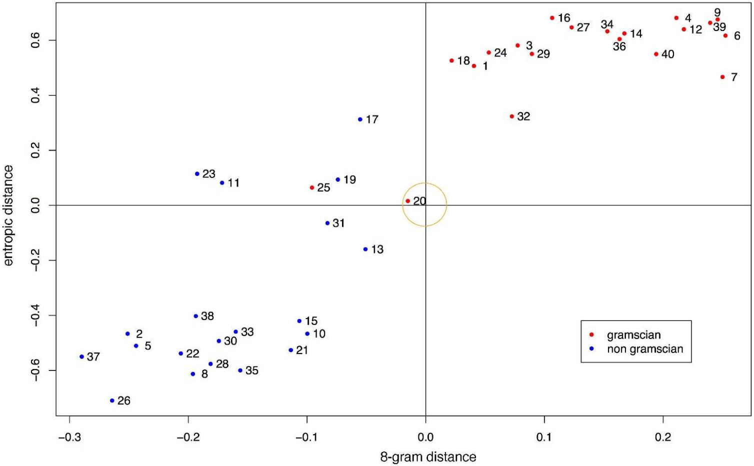

A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

- During the period 1914-1928, Gramsci produced an enormous number of articles on different national newspaper.
- Most of these article (hundreds, if not thousands) are NOT signed
- Other possible authors : Bordiga, Serrati, Tasca, Togliatti...the aim is to recognize the articles really written by A. Gramsci...
- Quite positive results for the period 1915-1917 (!?!?)
- Joint collaboration with D. Benedetto, E. Caglioti e M. Lana, for the new *Edizione Nazionale delle Opere di Gramsci (2007-2008)*

A *real* scenario: Gramsci's articles



A *real* scenario: Gramsci's articles



Reference

- D. Benedetto, E. Caglioti, G. Cristadoro and —-: "Relative entropy via non-sequential recursive pair substitution", *Journal of Statistical Mechanics: Theory and Experiments* , in press (2010)

a family of transformations on sequences and the corresponding operators on distributions:

given $a, b \in \mathcal{A}$, $\alpha \notin \mathcal{A}$ and $\mathcal{A}' = \mathcal{A} \cup \{\alpha\}$, a *pair substitution* is a map

$$G_{ab}^{\alpha} : \mathcal{A}^* \to \mathcal{A}'^*$$

a family of transformations on sequences and the corresponding operators on distributions:

given $a, b \in \mathcal{A}$, $\alpha \notin \mathcal{A}$ and $\mathcal{A}' = \mathcal{A} \cup \{\alpha\}$, a *pair substitution* is a map

$$G_{ab}^{\alpha} : \mathcal{A}^* \to \mathcal{A}'^*$$

which substitutes sequentially, from left to right, the occurrences of ab with α.

a family of transformations on sequences and the corresponding operators on distributions:

given $a, b \in \mathcal{A}$, $\alpha \notin \mathcal{A}$ and $\mathcal{A}' = \mathcal{A} \cup \{\alpha\}$, a *pair substitution* is a map

$$G_{ab}^{\alpha} : \mathcal{A}^* \to \mathcal{A}'^*$$

which substitutes sequentially, from left to right, the occurrences of *ab* with α.

For example

$$G_{01}^{2}\,(0010001011100100) = 020022110200.$$

a family of transformations on sequences and the corresponding operators on distributions:

given $a, b \in \mathcal{A}$, $\alpha \notin \mathcal{A}$ and $\mathcal{A}' = \mathcal{A} \cup \{\alpha\}$, a *pair substitution* is a map

$$G_{ab}^{\alpha} : \mathcal{A}^* \to \mathcal{A}'^*$$

which substitutes sequentially, from left to right, the occurrences of ab with α.

For example

$$G_{01}^{2}\,(0010001011100100) = 020022110200.$$

or:

$$G_{00}^{2}(0001000011) = 2012211.$$

a family of transformations on sequences and the corresponding operators on distributions:

given $a, b \in \mathcal{A}$, $\alpha \notin \mathcal{A}$ and $\mathcal{A}' = \mathcal{A} \cup \{\alpha\}$, a *pair substitution* is a map

$$G_{ab}^{\alpha} : \mathcal{A}^* \to \mathcal{A}'^*$$

which substitutes sequentially, from left to right, the occurrences of ab with α.

For example

$$G_{01}^{2}\,(0010001011100100) = 020022110200.$$

or:

$$G_{00}^{2}(0001000011) = 2012211.$$

$G = G_{ab}^{\alpha}$ is always an injective but not surjective map that can be immediately extended also to infinite sequences $w \in \mathcal{A}^{\mathbb{N}}$.

the action of G

G shorten the original sequence:

$$\frac{1}{Z_{ab}(\omega_1^n)} := \frac{|G_{ab}^{\alpha}(\omega_1^n)|}{|\omega_1^n|}$$

the action of G

G shorten the original sequence:

$$\frac{1}{Z_{ab}(\omega_1^n)} := \frac{|G_{ab}^{\alpha}(\omega_1^n)|}{|\omega_1^n|} = 1 - \frac{\sharp\{ab \subseteq \omega_1^n\}}{n},$$

the action of G

G shorten the original sequence:

$$\frac{1}{Z_{ab}(\omega_1^n)} := \frac{|G_{ab}^{\alpha}(\omega_1^n)|}{|\omega_1^n|} = 1 - \frac{\sharp\{ab \subseteq \omega_1^n\}}{n},$$

For *μ-typical* sequences we can pass to the limit and define:

the action of G

G shorten the original sequence:

$$\frac{1}{Z_{ab}(\omega_1^n)} := \frac{|G_{ab}^{\alpha}(\omega_1^n)|}{|\omega_1^n|} = 1 - \frac{\sharp\{ab \subseteq \omega_1^n\}}{n},$$

For *μ-typical* sequences we can pass to the limit and define:

$$\frac{1}{Z^{\mu}} := \lim_{n\to\infty} \frac{|G(\omega_1^n)|}{|\omega_1^n|} = \begin{cases} 1 - \mu(ab) & \text{if } a \neq b \\ 1 - \mu(aa) + \mu(aaa) - \mu(aaaa) + \cdots & \text{if } a = b \end{cases}$$

invariance of the entropy

- *Invariance of entropy*

$$h(\mathcal{G}\mu) = Z\, h(\mu).$$

invariance of the entropy

- *Invariance of entropy*

$$h(\mathcal{G}\mu) = Z\, h(\mu).$$

- *Decreasing of the 1-conditional entropy*

$$h_1(\mathcal{G}\mu) \leq Z h_1(\mu).$$

invariance of the entropy

- *Invariance of entropy*

$$h(\mathcal{G}\mu) = Z\,h(\mu).$$

- *Decreasing of the 1-conditional entropy*

$$h_1(\mathcal{G}\mu) \leq Z h_1(\mu).$$

- $\mathcal{G}$ maps 1-Markov measures in 1-Markov measures:

$$h(\mathcal{G}\mu) \leq h_1(\mathcal{G}\mu) \leq Z h_1(\mu) = Z h(\mu) = h(\mathcal{G}\mu)$$

invariance of the entropy

- *Invariance of entropy*

$$h(\mathcal{G}\mu) = Z\, h(\mu).$$

- *Decreasing of the 1-conditional entropy*

$$h_1(\mathcal{G}\mu) \leq Zh_1(\mu).$$

- $\mathcal{G}$ maps 1-Markov measures in 1-Markov measures:

$$h(\mathcal{G}\mu) \leq h_1(\mathcal{G}\mu) \leq Zh_1(\mu) = Zh(\mu) = h(\mathcal{G}\mu)$$

- *Decreasing of the k-conditional entropy*

$$h_k(\mathcal{G}\mu) \leq Zh_k(\mu).$$

invariance of the entropy

- *Invariance of entropy*

$$h(\mathcal{G}\mu) = Z\, h(\mu).$$

- *Decreasing of the 1-conditional entropy*

$$h_1(\mathcal{G}\mu) \leq Z h_1(\mu).$$

- $\mathcal{G}$ maps 1-Markov measures in 1-Markov measures:

$$h(\mathcal{G}\mu) \leq h_1(\mathcal{G}\mu) \leq Z h_1(\mu) = Z h(\mu) = h(\mathcal{G}\mu)$$

- *Decreasing of the k-conditional entropy*

$$h_k(\mathcal{G}\mu) \leq Z h_k(\mu).$$

- $\mathcal{G}$ maps k-Markov measures in k-Markov measures.

invariance of the entropy

- *Invariance of entropy*

$$h(\mathcal{G}\mu) = Z\, h(\mu).$$

- *Decreasing of the 1-conditional entropy*

$$h_1(\mathcal{G}\mu) \leq Z h_1(\mu).$$

- $\mathcal{G}$ maps 1-Markov measures in 1-Markov measures:

$$h(\mathcal{G}\mu) \leq h_1(\mathcal{G}\mu) \leq Z h_1(\mu) = Z h(\mu) = h(\mathcal{G}\mu)$$

- *Decreasing of the k-conditional entropy*

$$h_k(\mathcal{G}\mu) \leq Z h_k(\mu).$$

- $\mathcal{G}$ maps k-Markov measures in k-Markov measures.

invariance of the entropy

- *Invariance of entropy*

$$h(\mathcal{G}\mu) = Z\, h(\mu).$$

- *Decreasing of the 1-conditional entropy*

$$h_1(\mathcal{G}\mu) \leq Z h_1(\mu).$$

- $\mathcal{G}$ maps 1-Markov measures in 1-Markov measures:

$$h(\mathcal{G}\mu) \leq h_1(\mathcal{G}\mu) \leq Z h_1(\mu) = Z h(\mu) = h(\mathcal{G}\mu)$$

- *Decreasing of the k-conditional entropy*

$$h_k(\mathcal{G}\mu) \leq Z h_k(\mu).$$

- $\mathcal{G}$ maps k-Markov measures in k-Markov measures.

These properties, *roughly speaking*, reflect the fact that:

the amount of information of $G(\omega)$, which is equal to that of ω, is more concentrated on the pairs of consecutive symbols.

iterating G...

$\mathcal{A}_1$, A_2, $\ldots \mathcal{A}_N, \ldots$ will be an increasing alphabet sequence

iterating G...

$\mathcal{A}_1$, A_2, $\ldots \mathcal{A}_N, \ldots$ will be an increasing alphabet sequence

Given N and chosen $a_{\boldsymbol{N}}$, $b_{\boldsymbol{N}} \in \mathcal{A}_{N-1}$:

iterating G...

$\mathcal{A}_1$, A_2, $\ldots \mathcal{A}_N, \ldots$ will be an increasing alphabet sequence

Given N and chosen $a_{\boldsymbol{N}}$, $b_{\boldsymbol{N}} \in \mathcal{A}_{N-1}$:

$\alpha_{\boldsymbol{N}} \notin A_{N-1}$ is a new symbol and define the new alphabet as $\mathcal{A}_N = \mathcal{A}_{N-1} \cup \{\alpha_{\boldsymbol{N}}\}$;

iterating G...

$\mathcal{A}_1$, A_2, $\ldots \mathcal{A}_N, \ldots$ will be an increasing alphabet sequence

Given N and chosen $a_{\boldsymbol{N}}$, $b_{\boldsymbol{N}} \in \mathcal{A}_{N-1}$:

$\alpha_{\boldsymbol{N}} \notin A_{N-1}$ is a new symbol and define the new alphabet as $\mathcal{A}_N = \mathcal{A}_{N-1} \cup \{\alpha_{\boldsymbol{N}}\}$;

G_N is the substitution map $G_N = G^{\alpha_{\boldsymbol{N}}}_{a_{\boldsymbol{N}} b_{\boldsymbol{N}}} : \mathcal{A}^*_{N-1} \to \mathcal{A}^*_N$ which substitutes whit $\alpha_{\boldsymbol{N}}$ the occurrences of the pair $a_{\boldsymbol{N}} b_{\boldsymbol{N}}$;

iterating G...

$\mathcal{A}_1$, A_2, $\ldots \mathcal{A}_N, \ldots$ will be an increasing alphabet sequence

Given N and chosen $a_{\boldsymbol{N}}$, $b_{\boldsymbol{N}} \in \mathcal{A}_{N-1}$:

$\alpha_{\boldsymbol{N}} \notin A_{N-1}$ is a new symbol and define the new alphabet as $\mathcal{A}_N = \mathcal{A}_{N-1} \cup \{\alpha_{\boldsymbol{N}}\}$;

G_N is the substitution map $G_N = G^{\alpha_{\boldsymbol{N}}}_{a_{\boldsymbol{N}} b_{\boldsymbol{N}}} : \mathcal{A}^*_{N-1} \to \mathcal{A}^*_N$ which substitutes whit $\alpha_{\boldsymbol{N}}$ the occurrences of the pair $a_{\boldsymbol{N}} b_{\boldsymbol{N}}$;

$\mathcal{G}_N$ the corresponding map from the measures on $A^{\mathbb{Z}}_{N-1}$ to the measures on $A^{\mathbb{Z}}_N$;

iterating G...

$\mathcal{A}_1$, A_2, $\dots \mathcal{A}_N, \dots$ will be an increasing alphabet sequence

Given N and chosen $a_{\boldsymbol{N}}$, $b_{\boldsymbol{N}} \in \mathcal{A}_{N-1}$:

$\alpha_{\boldsymbol{N}} \notin A_{N-1}$ is a new symbol and define the new alphabet as $\mathcal{A}_N = \mathcal{A}_{N-1} \cup \{\alpha_{\boldsymbol{N}}\}$;

G_N is the substitution map $G_N = G^{\alpha_{\boldsymbol{N}}}_{a_{\boldsymbol{N}} b_{\boldsymbol{N}}} : \mathcal{A}^*_{N-1} \to \mathcal{A}^*_N$ which substitutes whit $\alpha_{\boldsymbol{N}}$ the occurrences of the pair $a_{\boldsymbol{N}} b_{\boldsymbol{N}}$;

$\mathcal{G}_N$ the corresponding map from the measures on $A^{\mathbb{Z}}_{N-1}$ to the measures on $A^{\mathbb{Z}}_N$;

we define by Z_N the corresponding normalization factor $Z_N = Z^{\alpha_{\boldsymbol{N}}}_{a_{\boldsymbol{N}} b_{\boldsymbol{N}}}$.

over-line to denote iterated quantities

$$\bar{G}_N \quad := \quad G_N \circ G_{N-1} \circ \cdots \circ G_1,$$

over-line to denote iterated quantities

$$\bar{G}_N := G_N \circ G_{N-1} \circ \cdots \circ G_1, \qquad \bar{\mathcal{G}}_N := \mathcal{G}_N \circ \mathcal{G}_{N-1} \circ \cdots \circ \mathcal{G}_1$$

over-line to denote iterated quantities

$$\bar{G}_N := G_N \circ G_{N-1} \circ \cdots \circ G_1, \qquad \bar{\mathcal{G}}_N := \mathcal{G}_N \circ \mathcal{G}_{N-1} \circ \cdots \circ \mathcal{G}_1$$

and also

$$\bar{Z}_N = Z_N Z_{N-1} \cdots Z_1.$$

asymptotic of $\bar{Z}_N$

The asymptotic properties of $\bar{Z}_N$ clearly depend on the pairs chosen in the substitutions.

asymptotic of $\bar{Z}_N$

The asymptotic properties of $\bar{Z}_N$ clearly depend on the pairs chosen in the substitutions.

In particular, if at any step N the chosen pair $a_N b_N$ is the pair of maximum of frequency of $\mathcal{A}_{N-1}$ then (Theorem 4.1 in BCG):

$$\lim_{N\to\infty} \bar{Z}_N = +\infty$$

asymptotic properties of the entropy

Theorem (Entropy via NSRPS)

If

$$\lim_{N\to\infty} \bar{Z}_N = +\infty$$

asymptotic properties of the entropy

Theorem (Entropy via NSRPS)

If

$$\lim_{N\to\infty} \bar{Z}_N = +\infty$$

then

$$h(\mu) = \lim_{N\to\infty} \frac{1}{\bar{\bar{Z}}_N} h_1(\mu_N)$$

asymptotic properties of the entropy

Theorem (Entropy via NSRPS)

If

$$\lim_{N\to\infty} \bar{Z}_N = +\infty$$

then

$$h(\mu) = \lim_{N\to\infty} \frac{1}{\bar{Z}_N} h_1(\mu_N)$$

i.e. $\mu_N := \bar{G}_N \mu$ *becomes asymptotically 1-Markov*.

generalization to the cross and relative entropy

Theorem (Invariance of relative entropy for pair substitution)

If μ is ergodic, ν is a Markov chain and $\mu_n << \nu_n$, then if G is a pair substitution

$$d(\mathcal{G}\mu||\mathcal{G}\nu) = Z^{\mu} d(\mu||\nu)$$

generalization to the cross and relative entropy

Theorem (Invariance of relative entropy for pair substitution)

If μ is ergodic, ν is a Markov chain and $\mu_n << \nu_n$, then if G is a pair substitution

$$d(\mathcal{G}\mu||\mathcal{G}\nu) = Z^{\mu} d(\mu||\nu)$$

Theorem (KL divergence via NSRPS)

If $\overline{Z}_N^{\nu} \to +\infty$ as $N \to +\infty$,

$$h(\mu||\nu) = \lim_{N\to+\infty} \frac{h_1(\mathcal{G}_N\mu||\mathcal{G}_N\nu)}{\overline{Z}_N^{\mu}}$$

BWT in few words

$\omega = \omega_1\omega_2 \cdots \omega_n \in \mathcal{A}^n$ is a finite string on some ordered, finite alphabet.

BWT in few words

$\omega = \omega_1\omega_2\cdots\omega_n \in \mathcal{A}^n$ is a finite string on some ordered, finite alphabet.

Generate all the n cyclic rotations:

$$\omega_1\omega_2\cdots\omega_n, \quad \omega_2\omega_3\cdots\omega_n\omega_1, \qquad \ldots\ldots \qquad \omega_n\omega_1\omega_2\cdots\omega_{n-1}.$$

BWT in few words

$\omega = \omega_1\omega_2\cdots\omega_n \in \mathcal{A}^n$ is a finite string on some ordered, finite alphabet.

Generate all the n cyclic rotations:

$$\omega_1\omega_2\cdots\omega_n, \quad \omega_2\omega_3\cdots\omega_n\omega_1, \qquad \ldots\ldots \qquad \omega_n\omega_1\omega_2\cdots\omega_{n-1}.$$

Sort them *from right-to-left* in lexicographic order.

BWT in few words

$\omega = \omega_1\omega_2\cdots\omega_n \in \mathcal{A}^n$ is a finite string on some ordered, finite alphabet.

Generate all the n cyclic rotations:

$$\omega_1\omega_2\cdots\omega_n, \quad \omega_2\omega_3\cdots\omega_n\omega_1, \qquad \ldots\ldots \qquad \omega_n\omega_1\omega_2\cdots\omega_{n-1}.$$

Sort them *from right-to-left* in lexicographic order.

Form a matrix $\mathcal{M}$ whose rows are the sorted cyclic permutations.

BWT in few words

$\omega = \omega_1\omega_2\cdots\omega_n \in \mathcal{A}^n$ is a finite string on some ordered, finite alphabet.

Generate all the n cyclic rotations:

$$\omega_1\omega_2\cdots\omega_n, \quad \omega_2\omega_3\cdots\omega_n\omega_1, \qquad \ldots\ldots \qquad \omega_n\omega_1\omega_2\cdots\omega_{n-1}.$$

Sort them *from right-to-left* in lexicographic order.

Form a matrix $\mathcal{M}$ whose rows are the sorted cyclic permutations.

$\texttt{bwt}(\omega)$ is defined as the *first* column of $\mathcal{M}$.

an example of BWT

mississippi#
ississippi#m
ssissippi#mi
sissippi#mis
issippi#miss
ssippi#missi
sippi#missis
ippi#mississ
ppi#mississi
pi#mississip
i#mississipp
#mississippi

$\Longrightarrow$

F		L
m	ississippi	#
s	sissippi#m	i
#	mississipp	i
s	sippi#miss	i
p	pi#mississ	i
i	ssissippi#	m
p	i#mississi	p
i	#mississip	p
s	issippi#mi	s
s	ippi#missi	s
i	ssippi#mis	s
i	ppi#missis	s

FIG. 1. Example of Burrows–Wheeler transform for the string s = mississippi. The matrix on the right has the rows sorted in right-to-left lexicographic order. The string bwt(s) is the first column F with the symbol # removed; in this example, bwt(s) = msspipissii.

why the BWT can be important in constructing efficient entropy indicators

why the BWT can be important in constructing efficient entropy indicators

given a fixed finite string $s \in \mathcal{A}^N$, for each substring ω of s, all characters in s following ω are grouped together inside $\texttt{bwt}(s)$.

why the BWT can be important in constructing efficient entropy indicators

given a fixed finite string $s \in \mathcal{A}^N$*, for each substring* ω *of* s*, all characters in* s *following* ω *are grouped together inside* $\texttt{bwt}(s)$.

Think now at s as an asymptotically larger string coming from a stochastic sources, we might conclude that:

why the BWT can be important in constructing efficient entropy indicators

given a fixed finite string $s \in \mathcal{A}^N$, for each substring ω of s, all characters in s following ω are grouped together inside $\texttt{bwt}(s)$.

Think now at s as an asymptotically larger string coming from a stochastic sources, we might conclude that:

$\texttt{bwt}(s)$ *looks like a piecewise i.i.d. process*.

just a remark

The *context sorting* properties of the BWT, suggest a method to estimate conditional empirical distribution based on segmentation of the BWT output.

The algorithm in four steps

The algorithm in four steps

1. Run the BWT on a realization of the source.

The algorithm in four steps

1. Run the BWT on a realization of the source.
2. Partition the BWT output sequence x into T_x segments. For example using a *uniform segmentation strategy*.

The algorithm in four steps

1. Run the BWT on a realization of the source.
2. Partition the BWT output sequence x into T_x segments. For example using a *uniform segmentation strategy*.
3. Estimate the *first-order distribution* within each segment. We denote by $n_j(a)$ the number of occurrences of the symbol $a \in \mathcal{A}$ in the jth segment, and by $\hat{\mu}(a, j)$ the probability estimate of symbol a again in the jth segment:

$$\hat{\mu}(a, j) = \frac{n_j(a)}{\sum_{b \in \mathcal{A}} n_j(b)}.$$

The contribution to the entropy estimate of the empirical distribution in the jth segment is given by

$$\log \hat{\mu}(j) = \sum_{a \in \mathcal{A}} n_j(a) \log \hat{\mu}(a, j).$$

The algorithm in four steps

1. Run the BWT on a realization of the source.
2. Partition the BWT output sequence x into T_x segments. For example using a *uniform segmentation strategy*.
3. Estimate the *first-order distribution* within each segment. We denote by $n_j(a)$ the number of occurrences of the symbol $a \in \mathcal{A}$ in the jth segment, and by $\hat{\mu}(a,j)$ the probability estimate of symbol a again in the jth segment:

$$\hat{\mu}(a,j) = \frac{n_j(a)}{\sum_{b \in \mathcal{A}} n_j(b)}.$$

The contribution to the entropy estimate of the empirical distribution in the jth segment is given by

$$\log \hat{\mu}(j) = \sum_{a \in \mathcal{A}} n_j(a) \log \hat{\mu}(a,j).$$

4. Average the individual estimates over the segments to get the estimate:

$$\hat{h}(x_1^n) := -\frac{1}{n} \sum_{k=1}^{T_x} \log \hat{\mu}(j)$$

the Main Theorem

Theorem

Let $x \in \mathcal{A}^n$ be a sequence of length n generated from a stationary ergodic sources μ, with entropy h_μ.

the Main Theorem

Theorem

Let $x \in \mathcal{A}^n$ be a sequence of length n generated from a stationary ergodic sources μ, with entropy h_μ.

The entropy estimator using uniform segmentation with segment length $c(n) = \alpha \cdot n^\gamma$ converges to the entropy rate almost surely:

the Main Theorem

Theorem

Let $x \in \mathcal{A}^n$ be a sequence of length n generated from a stationary ergodic sources μ, with entropy h_μ.

The entropy estimator using uniform segmentation with segment length $c(n) = \alpha \cdot n^\gamma$ converges to the entropy rate almost surely:

$$\lim_{|x|=n\to\infty} \hat{h}(x) = h_\mu, \qquad a.s.$$

BWT for K-L estimates

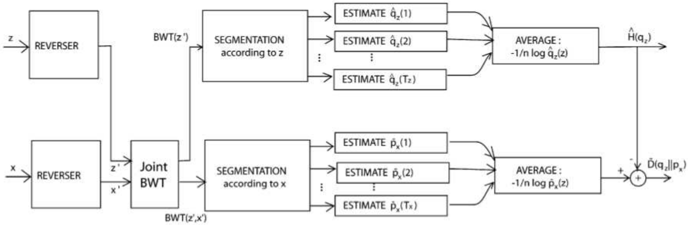


Fig. 2. Block diagram of the divergence estimator via the BWT.

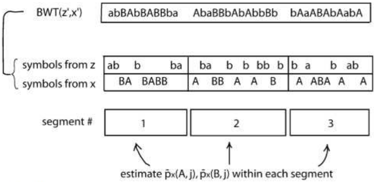


Fig. 3. The joint BWT segmentation and estimation.