Draft/Working Copy (ISC-1-2011)

The Intersymbol Correlation of a Sequence

edited by

Glenn Takanishi

January 2011

This article studies the correlation of information embedded in a sequence based on Satosi Watanabe's parameter called the intersymbol correlation. Basic concepts of conditional probability, ergodicity, and redundancy are used in formulating the correlation index within a sequence. The occurrence of these symbols of an nth-order Markov chain is governed by an "intersymbol correlation" probability of range n. The conditions of ergodicity and the structure of "ergodic subsets" of sequences of arbitrary length are analyzed. Satosi Watanabe's mathematical method which evaluates the "range" and "strength" of the correlation index of the sequence is studied.

Introduction

This is a study of Satosi Watanabe's article titled "A Study of Ergodicity and Redundancy Based on Intersymbol Correlation of Finite Range" written in 1953. I've edited the text to focus the language specifically on sequences. The content remains unchanged.

This article studies the information embedded in a sequence using concepts of conditional probability, ergodicity, and redundancy leading to the formulation of the correlation index. The intersymbol correlation probability [1] is an indirect measure of the correlation within a sequence. This intersymbol correlation, ISC, probability along with ergodicity is used to define the correlation index, W, [2] of a stationary, ergodic time sequence. The correlation index, W_n, measures the strength of correlation of symbol length n in excess of the correlation of length n-1.

In this study, the data set is a finite sequence, S, of length n composed of symbols or alphabets.

$$S = \{a_1, a_2, ..., a_n\}.$$

A sub-sequence of symbols is called a word. In this language a string or sentence is composed of symbols or alphabets, words and phrase in the sequence S.

In this study, a conditional probability of the form $Q(a_1, a_2, ..., a_{n-1}|\, a_n)$ is called the intersymbol correlation, ISC, probability. It is the probability of the appearance of a symbol a_n following the occurrence of a sub-sequence of $(n-1)$ immediately preceding symbols. n is the range of the ISC in the n-th order Markov chain. The ISC is meaningful only when there exists a "unique, non-vanishing value" [1] of $P\,(a_1, a_2, ..., a_n)$, which fulfills the ergodic property of Markov chains.

Like the ISC, the correlation index, W_n, represents the "range" in the sense that the actual correlation range is the maximum value of n for which $W_n \neq 0$. Theoretically, this determines the applicability of a generalized theory of Markov chains, and practically, this can be use to measure the existing correlation range in a given sub-sequence. $W_n$also represents the "strength" of correlation, in the sense that W_n quantitatively measures the difference of information between the n set of consecutive symbols and the $(n-1)$ set of consecutive symbols in the $n-th$ ordered Markov chain.

The Intersymbol Correlation, ISC

The ISC probability, C_{a_n}, of a symbol a_n is

$$C_{a_n} = Q(a_1, a_2, ..., a_{n-1}|\, a_n).$$

The Correlation Index

The correlation index [1], W_n, of an ergodic sequence, $S = \{a_1, a_2, \ldots, a_n\}$ of length n is

$$\begin{aligned} W_n = & \sum_{a_i} P(a_1, a_2, ..., a_n)\, logP(a_1, a_2, ..., a_n) \\ & - 2 \sum_{a_i} P(a_1, a_2, ..., a_{n-1})\, logP(a_1, a_2, ..., a_{n-1}) \end{aligned}$$

$$+ \sum_{a_i} P(a_1, a_2, ..., a_{n-2}) \, log P(a_1, a_2, ..., a_{n-2}).$$

Only when W_n holds for all $(a_1, a_2, ..., a_n)$, then $W_n = 0$. In other words, for a given value of $m < n$, $W_n = 0$.

Appendix

Derivation of the Correlation Index of a Sequence from a Study of Ergodicity and Redundancy

by Satosi Watanabe

1953

Ergodicity

A finite sequence $S = \{a_1, a_2, ..., a_m, ..., a_n\}$ containing n symbols contains an intersymbol correlation, ISC, probability of:

$$Q(a_1, a_2, ..., a_m \,|\, a_{m+1}, ..., a_{n-1}, a_n), \tag{1}$$

where each one of $a_1, a_2, ..., a_n$can be any one of the n symbols.

Definition 1. *The ISC denoted by* (1) *represents the probability that the last (n-m) symbols of a sequence of n symbols are* $(a_{m+1}, ..., a_n)$ *when it is known that the first m symbols of the sequence are* $(a_1, a_2, ..., a_m)$. *The ISC is written in terms of the conventional Boolean conditional probability P(B/A) as Q(A/B).*

By the very nature of probability, we have

$$\begin{aligned} Q(a_1, a_2, ..., a_m \,|\, a_{m+1}, ..., a_{n-1}, a_n) &\geq 0, \\ \sum_{a_{m+1}} ... \sum_{a_n} Q(a_1, a_2, ..., a_m \,|\, a_{m+1}, ..., a_{n-1}, a_n) &= 1. \end{aligned} \tag{2}$$

If there is no correlation between symbols, the probability of any place in a sequence being occupied by symbol S_i is independent of the preceding symbols. Hence, the only quantity which determines a probability of the type (1) is $Q(S_i)$which represents the probability of symbol S_i appearing at any one place. In this case, we have:

$$Q(a_1, a_2, ..., a_m \,|\, a_{m+1}, ..., a_{n-1}, a_n) = Q(a_{m+1})\, Q(a_{m+2}) \,...\, Q(a_n).$$

If the correlation extends, for instance, over three consecutive symbols, and not more than three, then the probability of a place in a sequence being occupied by symbol S_i will depend on the two symbols directly preceding it, but not on the symbols beyond these two. This means that the quantities $Q(S_i\,, S_j|S_k)$ determines the general probability (1):

$$\begin{aligned} & Q(a_1, a_2, ..., a_m \,|\, a_{m+1}, ..., a_{n-1}, a_n) \\ = \;& Q(a_{m-1}, a_m|a_{m+1})\, Q(a_m, a_{m+1}|a_{m+2}) \,...\, Q(a_{n-2}, a_{n-1} \,|\, a_n). \end{aligned}$$

Theorem 1. *If the ISC does not extend over more than μ consecutive symbols in a sequence, we can factorize (1) as follows:*

$$\begin{aligned} & Q(a_1, a_2, ..., a_m \,|\, a_{m+1}, ..., a_{n-1}, a_n) \\ = \;& Q(a_{m-\mu-2}, ..., a_m | a_{m+1})\, Q(a_{m-\mu+3}, ..., a_{m+1} | a_{m+2}) \,...\, Q(a_{n-\mu+1}, a_{n-1} \,|\, a_n) \end{aligned} \tag{3}$$

This theorem can be used to define the "range-number" of ISC: this number ν is the minimum allowable μ in the decomposition (3).

Assuming the correlation to be of range ν we consider all the possible sequences whose first $(\nu - 1)$ symbols $(a_1, a_2, ..., a_{\nu-1})$. Among these sequences starting with $(a_1, a_2, ..., a_{\nu-1})$, consider the probability of those sequences whose first ν symbols are $(a_1, b_1, b_2, ..., b_{\nu-1})$. If $(a_2, a_3, ..., a_{\nu-1}) = (b_1, b_2, ..., b_{\nu-2})$, and $R(a_1, a_2, ..., a_{\nu-1} |\, b_1, b_2, ..., b_{\nu-1}) = 0$, this probability is given by

$$R(a_1, a_2, ..., a_{\nu-1} |\, b_1, b_2, ..., b_{\nu-1}) \;=\; Q(a_1, a_2, ...,\, a_{\nu-1} \,|\, b_{\nu-1}).$$

In other words, the probability of the sequence $(a_1, b_1, b_2, ..., b_{\nu-1})$ can be written in a matrix form as:

$$\begin{aligned} & (a_1, a_2, ..., a_{\nu-1} | R | b_1, b_2, ..., b_{\nu-1}) \\ = \;& Q(a_1, a_2, ..., a_{\nu-1} |\, b_{\nu-1}) \delta(a_2, b_1) \delta(a_3, b_2) \,...\, \delta(a_{\nu-1}, b_{\nu-2}) \end{aligned} \tag{4}$$

with $\delta(S_i, S_j) = 0$, if i $\neq$ j; and $\delta(S_i, S_j) = 1$, if i = j.

Using this matrix-expression, the probability, in the above set of sequences, of a particular sequence $(b_1, b_2, ..., b_{\nu-1})$ appearing with m symbols between $a_1$and $b_1$is given by

$$T^{(m)}(a_1, a_2, ..., a_{\nu-1} |\, b_1, b_2, ..., b_{\nu-1}) \;=\; (a_1, a_2, ..., a_{\nu-1} | R^{(m)} | b_1, b_2, ..., b_{\nu-1}), \tag{5}$$

where $R^{(m)}$ means the m-th power of R in the sense of matrix-multiplication.

With the help of the quantity (5), we can further calculate the probability of a given sequence of any length $(\mu - 1)$, say $(b_1, b_2, ..., b_{\mu-1})$, appearing at any position after the initial $(a_1, a_2, ..., a_{\mu-1})$. If $\mu > \nu$ this probability will be

$$\begin{aligned} & T^{(m)}(a_1, a_2, ..., a_{\mu-1} |\, b_1, b_2, ..., b_{\mu-1}) \\ = \;& T^{(m)}(a_1, a_2, ..., a_{\nu-1} |\, b_1, b_2, ..., b_{\nu-1})\, Q(b_1, b_2, ..., b_{\nu-1} |\, b_\nu) \,...\, Q(b_{\mu-\nu}, ..., b_{\mu-2} |\, b_{\mu-1}) \end{aligned} \tag{6}$$

where m is the number of symbols between $a_1$and b_1.

If $\mu > \nu$, we have

$$\begin{aligned} & T^{(m)}\,(a_1, a_2, ..., a_{\nu-1} |\, b_1, b_2, ..., b_{\mu-1}) \\ = \;& \sum_{b_\mu} ... \sum T^{(m)}_{b_{\nu-1}}(a_1, a_2, ..., a_{\nu-1} |\, b_1, b_2, ..., b_{\mu-1}, b_\mu, ..., b_{\nu-1}), \end{aligned} \tag{7}$$

where m bears the same meaning.

The average probability of sequence $(b_1, b_2, ..., b_{\mu-1})$ with the "separation-distance" not larger than m will be:

$$\begin{aligned} & U^{(m)}(a_1, a_2, ..., a_{\nu-1}|\, b_1, b_2, ..., b_{\mu-1}) \\ = \; & \frac{1}{m} \sum T_{l=1}^{(l)m}(a_1, a_2, ..., a_{\nu-1}|\, b_1, b_2, ..., b_{\mu-1}). \end{aligned} \tag{8}$$

We now define what we mean by ergodicity with respect to this article. Consider all the possible, infinitely long sequences which start with a given initial sequence $(a_1, a_2, ..., a_{\nu-1})$ and also consider the average probability of the sequence $(\mathrm{b}_1, b_2, ..., b_{\mu-1})$ appearing in any position. This probability has the mathematical expression:

$$\lim_{m\to\infty} \quad U^{(m)}(a_1, a_2, ..., a_{\nu-1}|\, b_1, b_2, ..., b_{\mu-1}). \tag{9}$$

The word average here implies a two-fold averaging: first, averaging over all possible sequences with a fixed position where the sequence $(\mathrm{b}_1, b_2, ..., b_{\mu-1})$ should appear, and second, averaging over all the possible positions of this sequence. The first averaging is mathematically represented by the matrix multiplication in (5), and the second averaging by the summation in (8).

> **Definition II**. If $\lim_{m\to\infty} U^{(m)}(a_1, a_2, ..., a_{\nu-1}|\, b_1, b_2, ..., b_{\mu-1})$ converges to a unique, non-vanishing limit independent of $(a_1, a_2, ..., a_{\nu-1})$, where $(\mathrm{a}_1, a_2, ..., a_{v-1})$ can be taken arbitrarily from a certain family of $(\nu - 1)$ symbol sequences and $(b_1 b_2, ..., b_{\mu-1})$ can be taken arbitrarily from a certain family of $(\mu - 1)$ symbol sequences, then we speak of ergodicity with regard to these families.

We shall presently see that the quantity (9) with a fixed initial sequence $(a_1, a_2, ..., a_{\nu-1})$ and a fixed final sequence $(b_1, b_2, ..., b_{\mu-1})$ indeed converges to a limit, say:

$$U^{(\infty)}(a_1, a_2, ..., a_{\nu-1}|\, b_1, b_2, ..., b_{\mu-1}), \tag{10}$$

but this limit is not necessarily larger than zero, not is it in general necessarily independent of the initial sequence. In order to understand clearly the situation, let us invoke some well-known mathematical theorems regarding the Markov chain.

The ordinary Markov chain formally pertains to a two-symbol correlation probability $(\alpha|R|\beta)$, $(\alpha, \beta = 1, 2, ..., M)$:

$$(\alpha|R|\beta) \; \geq \; 1, \quad \textstyle\sum_\beta (\alpha|R|\beta) \quad = 1. \tag{11}$$

In accordance with the usual rule of matrix multiplication, we further introduce

$$(\alpha|R^m|\beta) \; = \; \sum_\kappa \sum_\lambda ... \sum_\mu (\alpha|R|\kappa)(\kappa|R|\lambda)...(\mu|R|\beta). \tag{12}$$

Then, we have the following theorems:

Theorem II. The quantity defined by

$$U^{(m)}(\alpha|\beta) = \sum_{l=1}^{m} \frac{1}{m}(\alpha|R^l|\beta) \tag{13}$$

for any given pair $(\alpha.\beta)$ converges to a limit as $m \to \infty$:

$$U^{(\infty)}(\alpha|\beta) = lim_{m\to\infty} U^{(m)}(\alpha|\beta). \tag{14}$$

Theorem III. The entire set G of symbols $(\alpha = 1, 2, ..., M)$ can be divided into a "vanishing" subset V and a certain number of "closed" subsets $C_i (i = 1, 2, ...)$ in such a way that

$U^{(\infty)}(\alpha|\beta) = 0$ for α belonging to G, and for β belonging to V,

$U^{(\infty)}(\alpha|\beta) > 0$ *for* α and β belonging to the same C_i,

$U^{(\infty)}(\alpha|\beta) = 0$ *for* α and β belonging to different C's.

Theorem IV. $U^{(\infty)}(\alpha|\beta)$is independent of α, if α and β belong to the same C.

Coming back to our original topic, if the correlation-range is two, and if $\mu = \nu$, these theorems can be directly applied to our problem involved in definition II. If the correlation-range is > 2, we only need to consider a sequence of $(\nu - 1)$ symbols collectively as a symbol α. The R's defined in (4) indeed satisfy (11). The cases: $\mu \neq \nu$ can be handled with the help of (6) and (7).

From Theorem II follows quite generally:

Theorem V. The limit (10) exists.

We will now discuss first the case $\mu = \nu$ in the light of theorems II, III and IV. According to theorem III, the entire set of $(\nu - 1)$ symbol sequences is subdivided into a vanishing subset V and a certain number of closed subsets C_i. If the final sequence of (10) belongs to V, the $U^{(\infty)}$ is zero independently of the initial sequence. For a given final sequence belonging to one of the closed subsets, $U^{(\infty)}$ will be zero if the initial sequence belongs to another closed subset, and will have a constant non-vanishing value insofar as the initial sequence belongs to the same closed subset as the final sequence. Thus,

Theorem VI. When $\mu = \nu$, ergodicity in the sense of definition II holds if and only if the initial family and the final family are the same closed subset.

In the cases where $\mu > \nu$, we construct an "extended" closed subset D_i of $(\mu - 1)$ symbols by taking those $(\mu - 1)$ symbols sequences $(b_1, b_2, ..., b_{\mu-1})$ whose first $(\nu - 1)$ symbols coincide with one of the members of the $(\nu - 1)$ symbols closed subset C_i and which satisfy the condition

$$Q(b_1 b_2, ..., b_{\nu-1}|\, b_v)\, Q(b_2 b_3, ..., b_\nu|\, b_{v+1}) \, ... Q(b_{\mu-\nu}, ..., b_{\mu-2}|\, b_{\mu-1}) \neq 0. \tag{15}$$

The extended vanishing subset will be composed of all those $(\mu - 1)$ symbols sequences whose first $(\nu - 1)$ symbols coincide with one of the members of the $(\nu - 1)$ symbols vanishing subset, or whose first

$(\nu-1)$ symbols coincide with one of the members of some closed subset but whost last $(\mu-\nu)$ symbols violate the condition (15). The entire set of possible $(\mu-1)$ symbol sequences are thus covered by the D's and V, and there is no possible overlapping. If the $(\mu-1)$ symbol final sequence of (10) is a member of this extended vanishing subset, $\mathrm{U}^{(\infty)}$ will certainly vanish whatever the initial sequence may be. If the final sequence belongs to an extended closed subset D_i will vanish for an initial sequence belonging to a C_j different from the one, C_i, which corresponds to D_i, and will have a constant non-vanishing value for any initial sequence belonging to C_i.

Theorem VII. When $\mu > \nu$, ergodicity holds if and only if the initial family is one of the closed subset C_i and the final family is the extended closed subset D_i corresponding to C_i.

In the cases where $\mu > \nu$, we encounter a rather peculiar situation. From a closed subset C_i we construct a retrenched subset E_i of $(\mu-1)$ symbol sequences. E_i is the set of those $(\mu-1)$ symbols of at least one of the members of C_i. The retrenched vanishing subset is defined as the totality of all those $(\mu-1)$ symbol sequences which do not belong to any one of the retrenched closed subsets. In case of the extended closed subsets, a given sequence of $(\mu-1)$ symbols could not belong to more than one D_i, since the division made in theorem III does not allow for any overlapping. However, in the present case of retrenched subsets, a given $(\mu-1)$ symbol sequence may well belong to more than one E. If the $(\mu-1)$-th symbol final sequence of (10) belongs to the retrenched vanishing subset, $\mathrm{U}^{(\infty)}$ will always vanish. If the $(\mu-1)$ symbol final sequence belongs to $\mathrm{E}_i, E_j, ..., E_k$, then $\mathrm{U}^{(\infty)}$ will be zero for an initial sequence belonging to a C different from any one of the corresponding subsets $\mathrm{C}_i, C_j, ..., C_k$. For the same final sequence, $\mathrm{U}^{(\infty)}$ may thus have different non-vanishing values according as to which one of the $\mathrm{C}_i, C_j, ..., C_k$ the initial sequence belongs to.

Theorem VIII. When $\mu < \nu$, ergodicity holds for the initial family identical with one of the closed subset C_i and the final family identical with the corresponding retrenched subset E_i.

In the foregoing considerations, we have systematically omitted the initial sequences belonging to the vanishing subset V. The reason for this is that the $\mathrm{U}^{(\infty)}$ depends in this case on the detailed structure of the intersymbol correlation, and that we cannot draw a conclusion of general validity. Of course, if the final sequence also belongs to V, then $\mathrm{U}^{(\infty)}$ vanishes.

Regarding the closed subsets of $(\nu-1)$ symbols, we should like to mention the following interesting property. We have obviously

$$U^{(\infty)}(a_1, ..., a_{\nu-1}|b_2, ..., b_\nu) = \sum_{b_1} U^{(\infty)}(a_1, ..., a_{\nu-1}|b_1, ..., b_{\nu-1})\, Q(b_1, b_2, ..., b_{\nu-1}|b_\nu), \tag{16}$$

from which we infer the following theorem.

Theorem IX. $(b, b_3, ..., b_\nu)$ is a member of C_i, if there is any symbol b_1 such that (b_1,b_2,...,$b_{\nu-1}$) is a member of C_i and $Q(b_1, b_2, ..., b_{\nu-1}|b_\nu) \neq 0$.

For a given (b_1,b_2,...,$b_{\nu-1}$) there must be at least one b_ν such that $Q(b_1, b_2, ..., b_{\nu-1}|b_\nu) \neq 0$, on account of (3). Hence,

> **Theorem X.** If (b_1,b_2,...,$b_{\nu-1}$) is a member of C_i, then there is always a member of C_i whose first $(\nu - 2)$ symbols are (b_2,b_3,...,$b_{\nu-1}$).

Before closing this section, a simple illustration may be given. Suppose an sequence is composed of three symbols S_1, S_2, and S_3. Suppose that it has an ISC range of 3.

$Q(S_1,S_1|S_1) = 1, \quad Q(S_1,S_2|S_1) = 1 \quad Q(S_1,S_3|S_1) = 1$
$Q(S_2,S_1|S_2) = 1, \quad Q(S_2,S_2|S_2) = 1 \quad Q(S_2,S_3|S_1) = 1$
$Q(S_3,S_1|S_1) = 1, \quad Q(S_3,S_2|S_1) = 1 \quad Q(S_3,S_3|S_1) = 1$

Then the $(\nu - 1)$ symbol subsets are:

C_1: (S_1, S_1)
C_2: (S_1, S_2), (S_2, S_1)
C_3: (S_2, S_2)
V: (S_1, S_3), (S_2, S_1), (S_2, S_3), (S_3, S_2), (S_3, S_3)

The extended 3-symbol subsets are:

D : (S_1, S_1, S_1)
D : (S_1, S_2, S_1), (S_2, S_1, S_2)
D : (S_2, S_2, S_3)
V' : all other 3-symbol sequences

The retrenched 1-symbol subsets are:

E : S_1
E : S_1, S_2
E : S_2
V : S_3

We can see the overlapping we have discussed. As a result, $U^{(\infty)}$ with the final sequence S_1, for instance, becomes 3-valued.

$U^{(\infty)}(S_1, S_1|S_1) = 1$
$U^{(\infty)}(S_1, S_2|S_1) = \frac{1}{2}$
$U^{(\infty)}(S_2, S_1|S_1) = \frac{1}{2}$
$U^{(\infty)}(S_2, S_2|S_1) = 0$
All other $U^{(\infty)}(|S_1) = 1$

Redundancy

In this section, we will constantly use a quantity denoted by

$$P(a_1, a_2, ..., a_n) \geq 1. \tag{17}$$

Definition III. $P(a_1, a_2, ..., a_n)$ represents the probability of observing an ergodic sequence of an arbitrarily symbol-length n with consecutive order $(a_1, a_2, ..., a_n)$.

From this definition follow the normalization conditions

$$\sum_{a_1} ... \sum_{a_n} P(a_1, a_2, ..., a_n) = 1. \tag{18}$$

According to the viewpoint of the last section, the existence of a unique value of such a probability is not unconditionally guaranteed. Only if the initial sequence $(b_1, b_2, ..., b_{\nu-1})$ is limited to within a closed subset, say C_i, then

$$U^{(\infty)}(b_1, ..., b_{\nu-1}|a_1, a_2, ..., a_n)$$

becomes independent of $(b_1, b_2, ..., b_{\nu-1})$, i.e., a function only of $(a_1, a_2, ..., a_n)$. If this is the case, we can write

$$U^{(\infty)}(b_1, ..., b_{\nu-1}|a_1, a_2, ..., a_n) = P(a_1, a_2, ..., a_n). \tag{19}$$

According to the theorems of the last section, if $(a_1, a_2, ..., a_n)$belongs to C_i, or its extended subset D, or its retrenched subset E_i, P will be finite, and otherwise zero. We have therefore to restrict the "infinitely long sequences" of definition III to only those which start with initial sequences belonging to one closed subset. The condition regarding P does not require that all the P's should be non-vanishing, hence the restriction on the final sequences, in the sense of definition II, is not necessary. On account of ergodicity, two sequences starting from two different initial sequences of the same closed subset becomes, in the long run, statistically independent. It is true that we can evade the restriction on the initial sequences by giving a certain "weight" to each of the closed subsets, which would lead to a unique value of each P. However, from the viewpoint that the sequences are determined solely by the correlation probability, this alternative is not acceptable, since it involves an arbitrary "weight" of each closed subset. Our discussion of this section will be based on the assumption that the initial sequences are limited to a single subset. The generalization of the results to the case of "weighted" subsets is very simple.

It should be noted that, as a result of the limitation of the initial sequences to a single subset, it may well happen that some of the generally possible sequences $(a_1, a_2, ..., a_{\nu-1})$ in the correlation probability $Q(q_1, a_2, ..., a_{\nu-1}|a_\nu)$ actually never occur in the possible set of sequences. Thus the actual range of correlation may become smaller than the range defined with regard to the entire possibilities of the a's. For instance, in the illustration of the last section, if we limit ourselves to the initial subset C_2, all 3-symbol Q's except $Q(S_1, S_2, |S_1) = 1$ and $Q(S_2, S_1|S_2) = 1$ will become meaningless. These two 3-symbol correlation probabilities reduce to the following two 2-symbol correlation probabilities $Q(S_1|S_2) = 1$, and $Q(S_2|S_1) = 1$. The range is thus reduced from three to two.

If a population of very long sample sequences is given, we can always evaluate $P(a_1, a_2, ..., a_n)$ by just counting the frequency of each seqment $(a_1, a_2, ..., a_n)$. However, if divide this entire population into, say, two groups, the values of $P(a_1, a_2, ..., a_n)$ may be different in the two groups. This discrepancy may be caused by a difference in correlation probabilities and/or by a difference in the initial sequences. In this section, however, we assume that we have a single population from which the quantities of the type $P(a_1, a_2, ..., a_n)$ are uniquely determined.

The quantity $P(a_1, a_2, ..., a_n)$ has, besides (18), the following property.

$$\begin{aligned} & \sum_a P(a_1, a_2, ..., a_k, b_1, ..., b_m, a_{k+m+1}, ..., a_n) \\ = \; & P(b_1, b_2, ..., b_m). \end{aligned} \tag{20}$$

This is obvious from the statistical viewpoint, but can also be verified from the standpoint of (19).

According to (6), we have for $n \geq \nu$,

$$P(a_1, a_2, ..., a_n) = P(a_1, a_2, ..., a_{\nu-1}) Q(a_1, a_2, ..., a_{\nu-1}|a_\nu) \cdots Q(a_{n-\nu-1}, ..., a_{n-1}|a_n), \tag{21}$$

or more generally,

$$P(a_1, a_2, ..., a_n) = P(a_1, a_2, ..., a_{\mu-1}) Q(a_1, a_2, ..., a_{\mu-1}|a_\mu) \cdots Q(a_{n-\mu-1}, ..., a_{n-1}|a_n), \tag{22}$$

provided $n \geq \mu \geq \nu$. The equivalence of (21) and (22) can readily be seen with the help of (3) and (6). In particular, for $n \geq \mu \geq \nu$, we get from (22)

$$Q(a_1, a_2, ..., a_{\mu-1}|a_\mu) = \frac{P(a_1, a_2, ..., a_\mu)}{P(a_1, a_2, ..., a_{\mu-1})}. \tag{23}$$

This is just what should be according to definitions I and III. (23) may be considered as the definition of $Q(a_1, a_2, ..., a_{\mu-1}|a_\mu)$ even for $\mu < \nu$. However, with such Q's with $\mu < \nu$, (22) will be true, since the Q's with $\mu < \nu$ cannot describe fully the existing correlation. Substituting (23) into (22), we get

$$P(a_1, a_2, ..., a_n) = \frac{P(a_1, a_2, ..., a_\mu) P(a_2, a_3, ..., a_{\mu+1}) \cdots P(a_{n-\mu+1}, ..., a_n)}{P(a_1, a_2, ..., a_\mu) \cdots P(a_{n-\mu+1}, ..., a_{n-1})}, \tag{24}$$

provided $n \geq \mu \geq \nu$. The actual range ν is thus the minimum value of μ for which the decomposition (24) is allowed.

For an allowed value of μ, if a further decomposition of range $\mu - 1$ is still allowed, i.e., if $\mu - 1 \geq \nu$, then we get from (24)

$$P(a_1, a_2, ..., a_\mu) = \frac{P(a_1, a_2, ..., a_{\mu-1}) P(a_2, a_3, ..., a_\mu)}{P(a_1, a_2, ..., a_{\mu-1})}, \tag{25}$$

for all $(a_1, a_2, ..., a_\mu)$. But if $\mu - 1 < \nu$, the left side of (25) will not be equal to its right side for at least one sequence $(a_1, a_2, ..., a_\mu)$. Thus we are led to use (25) as a criterion to determine whether $\mu > \nu$ or not. If (25) holds for all $(a_1, a_2, ..., a_\mu)$, then $\mu > \nu$; if not, $\mu \leq \nu$. Indeed, if (25) is possible, we have by virtue of (23)

$$\begin{aligned} Q(a_1, a_2, ..., a_{\mu-1}|a_\mu) &= \frac{P(a_1, a_2, ..., a_\mu)}{P(a_1, a_2, ..., a_{\mu-1})}, \\ &= \frac{P(a_2, a_3, ..., a_\mu)}{P(a_2, a_3, ..., a_{\mu-1})} \\ &= Q(a_2, a_3, ..., a_{\mu-1}|a_\mu), \end{aligned} \tag{26}$$

i.e., Q of range μ is reducible to Q of range $\mu - 1$. In light of theorem I, this means that the actual range is $\mu - 1$ or less. If (25) breaks down for at least one sequence $(a_1, a_2, ..., a_\mu)$, then (26) does not hold in general, meaning that the actual range is larger than $\mu - 1$.

Theorem XI. If and only if (25) holds for all $(a_1, a_2, ..., a_\mu)$, the actual correlation range ν is $\mu - 1$ or less.

This criterion is interesting, for here the P's instead of the Q's are the quantities which are primarily given. The criterion of theorem XI can be brought to a more concise form by the help of the well-known theorem attriuted to W. Gibbs.

Theorem XII. If

$$f_i \geq 0,\ g_i \geq 0,\ and \sum_i f_i = \sum_i g_i,\ (i = 1, 2, ..., f), \tag{27}$$

then

$$W = \sum_i f_i \, log \, f_i - \sum_i f_i \, log \, g_i \geq 0, \tag{28}$$

where the equality holds only when $f_i = g_i$for all i.

Now let's call the left-hand side and the right-hand side of (25) respectively.

$$f(a_1, a_2, ..., a_\mu) = P(a_1, a_2, ..., a_\mu) \tag{29}$$

$$g(a_1, a_2, ..., a_\mu) = \frac{P(a_1, a_2, ..., a_{\mu-1})P(a_2, a_3, ..., a_\mu)}{P(a_2, a_3, ..., a_{\mu-1})} \tag{30}$$

Consider the index i of theorem III as a collective index for the various possible sequences of symbol-length μ. On account of (18) and (20), the conditions (27) are satisfied, and we obtain

$$\begin{aligned} W_\mu = & \sum P(a_1, a_2, ..., a_\mu)\, log P(a_1, a_2, ..., a_\mu) \\ & - 2 \sum P(a_1, a_2, ..., a_{\mu-1})\, log P(a_1, a_2, ..., a_{\mu-1}) \\ & + \sum P(a_1, a_2, ..., a_{\mu-2})\, log P(a_1, a_2, ..., a_{\mu-2}) \\ & \geq 0 \end{aligned} \tag{31}$$

Only when (25) holds for all $(a_1, a_2, ..., a_\mu)$, then $W_\mu = 0$. In other words, for a given value of ν, $W_\mu = 0$ for $\mu > \nu$. This leads to a convenient way to determine the actual range.

Theorem XIII. The actual range ν is the maximum value of μ for which $W_\mu = 0$.

The W's defined by (31) are called the "correlation indices".

For $\mu = 2$, the definition of W_μ in (31) should be understood as meaning

$$W_2 = \sum P(a_1, a_2)\, log P(a_1, a_2) - 2 \sum P(a_1)\, log P(a_1), \tag{32}$$

for we have here $g(a_1, a_2) = P(a_1)P(a_2)$.

We now proceed to find out the average amount of information carried by a sub-sequence of length n in a language in which the P's exist. A specific sub-sequence $(a_1, a_2, ..., a_n)$ has probability $P(a_1, a_2, ..., a_n)$. Thus the information per symbol carried by this sub-sequence is

$$-\frac{1}{n} log\, P(a_1, a_2, ..., a_n).$$

The probability of occurrence of such a message being $P(a_1, a_2, ..., a_n)$, the average information per symbol for various possible sub-sequences of length n is given by

$$I_n = -\frac{1}{n} \sum P(a_1, a_2, ..., a_n) \log P(a_1, a_2, ..., a_n)\,. \tag{33}$$

If the existing correlation is of range ν, the P can be decomposed as in (24) with $\mu = \nu$. A straight forward calculation with the help of (18) and (20) gives

$$\begin{aligned} I_n &= I_{n,\nu} = -\frac{1}{n}(n - \nu + 1) \sum P(a_1, a_2, ..., a_\nu) \log P(a_1, a_2, ..., a_\nu) \\ &= +\frac{1}{n}(n - \nu) \sum P(a_1, a_2, ..., a_{\nu-1}) \log P(a_1, a_2, ..., a_{\nu-1}) \end{aligned} \tag{34}$$

For an obvious reason this ν can be the actual minimum range or any ν that is larger than this. Supposing ν in (34) to be the actual minimum range, let us find the error in the calculation based on the assumption that the actual range were ν-1. This is easily found to be

$$I_{n,\nu} - I_{n,\nu-1} = -\frac{(n - \nu - 1)}{n} W_\nu. \tag{35}$$

Repeating this process, we obtain

$$I_n - I^0 = I_{n,\nu} - I^0 = -\sum_{\mu=2}^{\nu} \frac{n-\mu-1}{n} W_\mu, \tag{36}$$

where

$$I^0 \quad = I_{n,1} = \quad -\sum P(a_1) log\, P(a_1). \tag{37}$$

Since W_μ vanishes anyway for $\mu > \nu$, we can state that

Theorem XIV. The average information per symbol carried by a sub-sequence of length n is

$$I_n = I^0 - \sum_{\mu=2}^{\infty} \frac{n-\mu-1}{n} W_\mu, \tag{38}$$

insofar as n is larger than the actual correlation range.

Since the W's are zero or positive, the intersymbol correlation, ISC, tends to decrease the amount of information. Thus, W_μ can be thought to represent the "strength" of correlation. By defintion, I_n cannot be negative, hence, there is an upper limit to the total "strength" of the correlation which is

$$\sum_{\mu=2}^{\infty} \frac{n-\mu-1}{n} W_\mu \leq \sum_{\mu=2}^{\infty} W_\mu \leq I^0. \tag{39}$$

For $n \gg \nu$, we obtain from (38),

$$I_n \approx I_\infty = I^0 - \sum_{\mu=2}^{\infty} W_\mu \;\; (n \gg \nu) \tag{40}$$

showing that if take a sufficiently long segment as a unit, the information per symbol becomes independent of the length of the segment. This indirectly justifies the usual procedure according to which an infinitely long message is cut into segments of sufficient length and the segments are treated as if they did not have any correlation among them.

The quantity called "redundancy" is defined by [3]

$$R = \frac{I^0 - I^\infty}{I^0}. \tag{41}$$

Theorem XV. The redundancy of a language which is characterized by the correlation indices W_μ is given by

$$R = \frac{1}{I^0} \sum_\mu W_\mu, \quad 0 \leq R \leq 1. \tag{42}$$

In the illustration of the last section, if we limit the initial sequences to C_2, we get

$$W_2 = log\, 2, \quad W_3 = W_4 = ... = 0$$

$$I^0 = log\, 2, \quad I_\infty = 0, \quad R = 100\%.$$

This last result is not surprising, because the possible infinite sequences are limited to

$$... S_1\, S_2\, S_1\, S_2 ...,$$

which certainly cannot convey any information.

References

1. Satosai Watanabe, A Study of Ergodicity and Redundancy Based on Intersymbol Correlation of Finite Range, 1953, Monterey US Naval Postgraduate School, Research Paper No. 4

2. Satosi Watanabe, Information Theoretical Analysis of Multivariate Correlation, 1960, IBM Journal of Research and Development, 4-1:66

3. Stanford Goldman, Information Theory, Prentice-Hall, New York, 1953, p. 45