Metadata of the chapter that will be visualized in SpringerLink

Book Title	Fuzzy Sets and Operations Research	
Series Title		
Chapter Title	The Semantic Information Method Compatible with Shannon, Popper, Fisher, and Zadeh's Thoughts	
Copyright Year	2019	
Copyright HolderName	Springer Nature Switzerland AG	
Corresponding Author	Family Name	**Lu**
	Particle	
	Given Name	**Chenguang**
	Prefix	
	Suffix	
	Role	
	Division	College of Intelligence Engineering and Mathematics
	Organization	Liaoning Engineering and Technology University
	Address	Fuxin, Liaoning, 123000, China
	Email	survival99@gmail.com

Abstract	Popper and Fisher's hypothesis testing thoughts are very important. However, Shannon's information theory does not consider hypothesis testing. The combination of information theory and likelihood method is attracting more and more researchers' attention, especially when they solve Maximum Mutual Information (MMI) and Maximum Likelihood (ML). This paper introduces how we combine Shannon's information theory, likelihood method, and fuzzy sets theory to obtain the Semantic Information Method (SIM) for optimizing hypothesis testing better. First, we use the membership functions of fuzzy sets proposed by Zadeh as the truth functions of hypotheses; then, we use the truth functions to produce likelihood functions, and bring such likelihood functions into Kullback-Leibler and Shannon's information formulas to obtain the semantic information formulas. Conversely, the semantic information measure may be used to optimize the membership functions. The maximum semantic information criterion is equivalent to the ML criterion; however, it is compatible with Bayesian prediction, and hence can be used in cases where the prior probability distribution is changed. Letting the semantic channel and the Shannon channel mutually match and iterate, we can achieve MMI and ML for tests, estimations, and mixture models. This iterative algorithm is called Channels' Matching (CM) algorithm. Theoretical analyses and several examples show that the CM algorithm has fast speed, clear convergence reason, and wild potential applications. The further studies of the SIM related to the factor space and information value are discussed.

Keywords (separated by '-')	Hypothesis testing - Shannon's theory - Maximum likelihood - Maximum mutual information - Membership function - Semantic information - Factor space - Information value

The Semantic Information Method Compatible with Shannon, Popper, Fisher, and Zadeh's Thoughts

Chenguang Lu(✉)

College of Intelligence Engineering and Mathematics,
Liaoning Engineering and Technology University,
Fuxin, Liaoning 123000, China
survival99@gmail.com

Abstract. Popper and Fisher's hypothesis testing thoughts are very important. However, Shannon's information theory does not consider hypothesis testing. The combination of information theory and likelihood method is attracting more and more researchers' attention, especially when they solve Maximum Mutual Information (MMI) and Maximum Likelihood (ML). This paper introduces how we combine Shannon's information theory, likelihood method, and fuzzy sets theory to obtain the Semantic Information Method (SIM) for optimizing hypothesis testing better. First, we use the membership functions of fuzzy sets proposed by Zadeh as the truth functions of hypotheses; then, we use the truth functions to produce likelihood functions, and bring such likelihood functions into Kullback-Leibler and Shannon's information formulas to obtain the semantic information formulas. Conversely, the semantic information measure may be used to optimize the membership functions. The maximum semantic information criterion is equivalent to the ML criterion; however, it is compatible with Bayesian prediction, and hence can be used in cases where the prior probability distribution is changed. Letting the semantic channel and the Shannon channel mutually match and iterate, we can achieve MMI and ML for tests, estimations, and mixture models. This iterative algorithm is called Channels' Matching (CM) algorithm. Theoretical analyses and several examples show that the CM algorithm has fast speed, clear convergence reason, and wild potential applications. The further studies of the SIM related to the factor space and information value are discussed.

Keywords: Hypothesis testing · Shannon's theory · Maximum likelihood
Maximum mutual information · Membership function · Semantic information
Factor space · Information value

1 Introduction

Although Shannon's information theory [1] has achieved great successes, his information concept does not accord with daily usages of "information". For example, Shannon's (amount of) information is irrelevant to the truth and falsity of statements or predictions. Another problem is that the Shannon theory does not contain hypothesis

B.-Y. Cao and Y.-B. Zhong (Eds.): ICFIE 2017, AISC 872, pp. 1–16, 2019.
https://doi.org/10.1007/978-3-030-02777-3_19

testing thought, and therefore, we cannot use Shannon's mutual information as criterion to optimize tests, estimations, and predictions. Popper's theory of scientific advances [2] and Fisher's Likelihood Method (LM) [3] contain hypothesis testing thoughts. Popper initiates to use (semantic) information criterion to evaluate and optimize scientific hypotheses. His information concept is more accordant with daily usages. Yet, Popper did not provide proper semantic information formula. So far, what is used to resolve hypothesis testing problems is Fisher's likelihood method, which plays an important role in statistical inference and machine learning. Yet, it is unclear how the LM is related to semantic meaning and semantic information. Still, the LM is not compatible with Bayesian prediction well.

According Davidson's truth-conditional semantics [4], we can use the truth function of a hypothesis to represent its semantic meaning. According to Zadeh's fuzzy sets theory [5], a membership function is also a (fuzzy) truth function of a hypothesis.

There have been some iterative methods for Maximum Mutual Information (MMI) and Maximum Likelihood (ML), including the Newton method [6], EM algorithm [7], and minimax method [8]. Still, we want a better method.

Recently, we found that Lu's semantic information formulas [9–12] could combine information measures, likelihood functions, and membership functions better for hypothesis testing. Although Lu did not mention "likelihood" in his earlier studies, in fact, his "predicted probability distribution" is likelihood function. Using the concepts of likelihood and semantic channel, we can state Lu's Semantic Information Method (SIM) better. We also found that letting the semantic channel and the Shannon channel mutually match and iterate, we could achieve MMI and ML for tests, estimations, and mixture models conveniently.

In this paper, we first restate Lu's SIM in terms of likelihood and semantic channel. That is to use the fuzzy truth function to produce the likelihood function, and put the likelihood function into the Kullback-Laibler (KL) information formula and the Shannon mutual information formula to obtain sematic information formulas. Such a semantic information measure may contain Popper and Fisher's hypothesis testing thoughts. We shall show that new semantic information measure can be used to evaluate and optimize semantic communication, to improve the LM for variable sources, and to optimize the membership functions according to sampling distributions. Then, we simply introduce new iterative algorithm: Channels' Matching algorithm or the CM algorithm. For further studies, the information value related to portfolio and factor space are also simply discussed.

2 Semantic Channel, Semantic Communication Model, and Semantic Bayesian Prediction

2.1 Shannon Channel and Transition Probability Function

The semantic channel and the Shannon channel may mutually affect. First, we simply introduce the Shannon channel [1].

Let X be a discrete random variable representing a fact with alphabet $A = \{x_1, x_2, \ldots, x_m\}$, let Y be a discrete random variable representing a message with alphabet

$B = \{y_1, y_2, \ldots, y_n\}$, and let Z be a discrete random variable representing a observed condition with alphabet $C = \{z_1, z_2, \ldots, z_w\}$. A message sender chooses Y to predict X according to Z. For example, in weather forecasts, X is a rainfall, Y is a forecast such as "There will be light to moderate rain tomorrow", and Z is a set of meteorological data. In medical tests, X is an infected or uninfected person, Y is positive or negative (testing result), and Z is a laboratory datum or a set of laboratory data.

We use $P(X)$ to denote the probability distribution of X and call $P(X)$ the source, and we use $P(Y)$ to denote the probability distribution of Y and call $P(Y)$ the destination. We call $P(y_j|X)$ with certain y_j and variable X the transition probability function from X to y_j. Then a Shannon's channel is composed of a group of transition probability functions [1]:

$$P(Y|X) \Leftrightarrow \begin{bmatrix} P(y_1|x_1) & P(y_1|x_2) & \cdots & P(y_1|x_m) \\ P(y_2|x_1) & P(y_2|x_2) & \cdots & P(y_2|x_m) \\ \cdots & \cdots & \cdots & \cdots \\ P(y_n|x_1) & P(y_n|x_2) & \cdots & P(y_n|x_m) \end{bmatrix} \Leftrightarrow \begin{bmatrix} P(y_j|X) \\ P(y_j|X) \\ \cdots \\ P(y_n|X) \end{bmatrix} \tag{1}$$

The transition probability function has two properties:

(1) $P(y_j|X)$ is different from the conditional probability function $P(Y|x_i)$ or $P(X|y_j)$ in that whereas the latter is normalized, the former is not. In general, $\sum_i P(y_j|e_i) \neq 1$.
(2) $P(y_j|X)$ can be used to make Bayesian prediction to get the posterior probability distribution $P(X|y_j)$ of X. To use it by a coefficient k, the two predictions are equivalent, i.e.

$$\frac{P(X)kP(y_j|X)}{\sum_i P(x_i)kP(y_j|x_i)} = \frac{P(X)P(y_j|X)}{\sum_i P(x_i)P(y_j|x_i)} = P(X|y_j) \tag{2}$$

2.2 Semantic Channel and Semantic Communication Model

In terms of hypothesis testing, X is a sample point or a piece of evidence and Y is a hypothesis or a prediction. We need a sample sequence or a sampling distribution $P(X|.)$ to test a hypothesis to see how accurate the hypothesis is.

Let Θ be a random variable for a fuzzy set (defined by Zadeh [5]) and let θ_j be a value taken by Θ when $Y = y_j$. We also treat θ_j as a predictive model (or sub-model). A predicate $y_j(X)$ means "X is in θ_j" whose truth function is $T(\theta_j|X) \epsilon [0,1]$. Because $T(\theta_j|X)$ is constructed with some parameters, we may also treat θ_j as a set of model parameters.

In contrast to the popular likelihood method, we use sub-models $\theta_1, \theta_2, \ldots, \theta_n$ instead of one model θ or Θ, where a sub-model θ_j is defined by a truth function $T(\theta_j|X)$. The likelihood function $P(X|\theta_j)$ here is equivalent to $P(X|y_j, \theta)$ in popular likelihood method. A sample used to test y_j is a sub-sample or conditional sample. We use the sampling distribution $P(X)$ or $P(X|y_j)$ instead of the sample sequence $x(1), x(2), \ldots$. to test a hypothesis. These changes will make the new method more flexible and more compatible with the Shannon information theory.

When $X = x_i$, $y_j(X)$ become $y_j(x_i)$, which is a proposition with truth value $T(\theta_j|x_i)$. We have the semantic channel:

$$T(\Theta|X) \Leftrightarrow \begin{bmatrix} T(\theta_1|x_1) & T(\theta_1|x_2) & \cdots & T(\theta_1|x_m) \\ T(\theta_2|x_1) & T(\theta_2|x_2) & \cdots & T(\theta_2|x_m) \\ \cdots & \cdots & \cdots & \cdots \\ T(\theta_n|x_1) & T(\theta_n|x_2) & \cdots & T(\theta_n|x_m) \end{bmatrix} \Leftrightarrow \begin{bmatrix} T(\theta_1|X) \\ T(\theta_2|X) \\ \cdots \\ T(\theta_n|X) \end{bmatrix} \tag{3}$$

This semantic channel can also be used for Bayesian prediction, i.e., semantic Bayesian prediction, to produce likelihood function:

$$P(X|\theta_j) = P(X)T(\theta_j|X)/T(\theta_j), \quad T(\theta_j) = \sum_i P(x_i)T(\theta_j|x_i) \tag{4}$$

where $T(\theta_j)$ may be called the logical probability of y_j. If $T(\theta_j|X) \propto P(y_j|X)$, then the semantic Bayesian prediction is equivalent to Bayesian prediction according to Eq. (2). Lu called this formula the set-Bayesian formula in 1991 [9] and put it into a semantic information measure. According to Dubois and Prade' paper [13], Thomas (1981) and Natvig (1983) proposed this formula earlier.

AQ1

We can also consider that $T(\theta_j|X)$ is defined with normalized likelihood (function), i.e., $T(\theta_j|X) = kP(\theta_j|X)/P(\theta_j) = kP(X|\theta_j)/P(X)$, where k is a coefficient that makes the maximum of $T(\theta_j|X)$ be 1. With $P(X)$, $T(\theta_j|X)$ and $P(X|\boldsymbol{\theta}_j)$ can ascertain each other.

Note that $T(\theta_j)$ is the logical probability of y_j, whereas $P(y_j)$ is the probability of choosing y_j. They are very different. $T(\Theta)$ is also not normalized, and generally there is $T(\boldsymbol{\theta}_1) + T(\boldsymbol{\theta}_2)\ldots + T(\theta_n) > 1$. Consider hypotheses y_1 = "There will be light rain", y_2 = "There will be moderate rain", and y_3 = "There will be light to moderate rain". According to their semantic meanings, $T(\theta_3) \approx T(\theta_1) + T(\theta_2)$; however, there may be $P(y_3) < P(y_1)$. Particularly, when y_j is a tautology, $T(\theta_j) = 1$ whereas $P(y_j)$ is almost 0. The $P(X|\theta_j)$ is a likelihood function and is also different from $P(X|y_j)$ which is a sampling distribution.

The semantic communication model is shown in Fig. 1.

A semantic channel is supported by a Shannon channel. For weather forecasts, the transition probability function $P(y_j|X)$ indicates the rule of choosing a forecast y_j. The rules used by different forecasters may be different and have more or fewer mistakes. Whereas, $T(\theta_j|X)$ indicates the semantic meaning of y_j that is understood by the audience. The semantic meaning is generally publicly defined and may also come from (or be affected by) the past rule of choosing y_j. To different people, the semantic meaning should be similar.

2.3 Is Likelihood Function or Truth Function Provided by the GPS's Positioning?

Consider the semantic meaning of the small circle (or the arrow) in the map on a GPS device. The circle tells where the position of the device is. A clock, a balance, or a

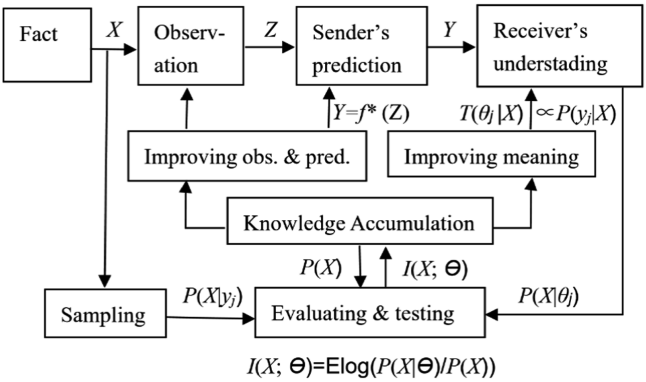

Fig. 1. The semantic communication model. Information comes from testing the semantic likelihood function $P(X|_j)$ by the sampling distribution $P(X|y_j)$.

thermometer is similar to a GPS device in that their actions may be abstracted as y_j = "$X \approx x_j$", $j = 1, 2, \ldots, n$. The Y with such a meaning may be called an unbiased estimate, and its transition probability functions $P(y_j|X)$ constitute a Shannon channel. This semantic channel may be expressed by

$$T(\theta_j|X) = \exp\left[-|X - x_j|^2/(2d^2)\right], \ \mathrm{j} = 1, 2, \ldots, n \tag{5}$$

where d is the standard deviation.

Consider a particular environment (in Fig. 2) where a GPS device is used in a car. The positioning circle is on a building. The left side of the building is a highway and the right side is a road. We must determine the most possible position of the car. If we think that the circle provides a likelihood function, we should infer "The car is most possibly on the building". However, common sense would indicate that this conclusion is wrong. Alternatively, we can understand the semantic meaning of the circle by a transition probability function. However, the transition probability function is difficult to obtain, especially when the GPS has a systematical deviation. One may posit that we can use a guessed transition probability function and neglect its coefficient. This idea is a good one. In fact, the truth function in Eq. (5) is just such a function. With the truth function, we can obtain the likelihood function by the semantic Bayesian prediction:

$$P(X|\theta_j) = \frac{P(X)\exp[-(X - x_j)^2/(2d^2)]}{\sum_i P(X)\exp[-(X - x_j)^2/(2d^2)]} \tag{6}$$

This likelihood function accords with common sense and avoids conclusion "The car is most likely on the building". This example shows that a semantic channel is simpler and more understandable than the corresponding Shannon channel.

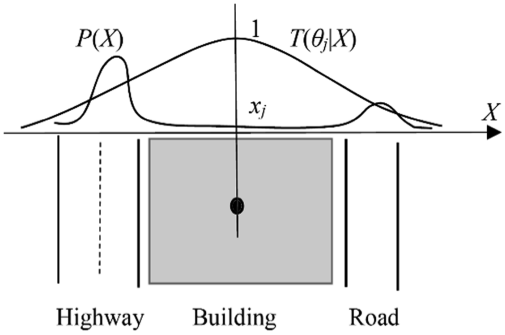

Fig. 2. The illustration of a GPS's positioning. When the prior distribution $P(X)$ is uneven and variable, using a truth function to make a semantic Bayesian prediction will be better than using a likelihood function to predict directly

3 Semantic Information Measure and the Optimization of the Semantic Channel

3.1 Semantic Information Measure Defined with Log Normalized Likelihood

In the Shannon information theory, there is only the statistical probability without the logical probability and likelihood (predicted probability). However, Lu defined semantic information measure by these three types of probabilities at the same time.

The (amount of) semantic information conveyed by y_j about x_i is defined as [10]:

$$I(x_i;\theta_j) = \log\frac{P(x_i|\theta_j)}{P(x_i)} = \log\frac{T(\theta_j|x_i)}{T(\theta_j)} \tag{7}$$

where semantic Bayesian prediction is used; it is assumed that the prior likelihood is equal to the prior probability distribution. For an unbiased estimation, its truth function and semantic information are illustrated in Fig. 3.

This formula contains Popper's thought [2] that the less the logical probability is, the more information there is if the hypothesis can survive tests; a tautology cannot be falsified and hence contains no information.

Bringing $T(\theta_j|X)$ in Eq. (5) into Eq. (7), we have

$$I(x_i;\theta_j) = \log(1/T(\theta_j)) - \left|X - x_j\right|^2/\left(2d^2\right) \tag{8}$$

where $\log(1/T(\theta_j))$ is the semantic information measure defined by Bar-Hillel and Carnap [14]. So, semantic information increases with either logical probability or deviation decreasing. The smaller deviation means that the hypothesis survives tests better.

Averaging $I(x_i;\theta_j)$, we obtain semantic (or generalized) Kullback-Leibler (KL) information (see [15] for the KL information or divergence):

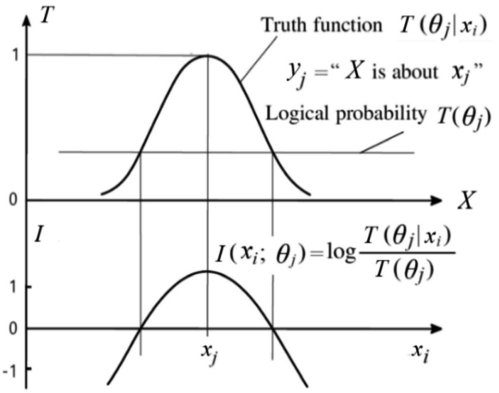

Fig. 3. Semantic information is defined with normalized likelihood. The less the logical probability is, the more information there is; the larger the deviation is, the less information there is; lastly, a wrong estimation may convey negative information.

$$I(X;\theta_j)=\sum_i P(x_i|y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=\sum_i P(x_i|y_j)\log\frac{T(\theta_j|x_i)}{T(\theta_j)} \tag{9}$$

The statistical probability or frequency $P(x_i|y_j)$, $i = 1, 2, \ldots$, on the left of "log" above, represents a sampling distribution (note that a sample or sub-sample is also conditional) to test the hypothesis y_j or model θ_j. If $y_j = f(Z|Z\epsilon C_j)$, then $P(X|y_j) = P(X|Z\epsilon C_j) = P(X|C_j)$.

Although Akaike [16] revealed the relationship between likelihood and the KL divergence [15]. This relationship has attracted more attention in recent decades (see Cover and Thomas's text book ([17], Chap. 11.7)). In the following, we try to show that the relationship between the semantic information measure and likelihood is clearer.

Assume that the size of a sample used to test y_j is N_j, and the sample points come from independent and identically distributed random variables. Among N_j points, the number of x_i is N_{ij}. When N_j is infinite, $P(X|y_j) = N_{ij}/N_j$. Hence there is the following equation:

$$\log\prod_i\left[\frac{P(x_i|\theta_j)}{P(x_i)}\right]^{N_{ji}}=N_j\sum_i P(x_i|y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=N_jI(X;\theta_j) \tag{10}$$

After averaging the above likelihood for different y_j, $j = 1, 2, \ldots, n$, we have

$$\begin{aligned}\frac{1}{N}\sum_j\log\prod_i\left[\frac{P(x_i|\theta_j)}{P(x_i)}\right]^{N_{ji}}&=\sum_j P(y_j)\sum_i P(x_i|y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}\\ =\sum_i P(x_i)\sum_j P(y_j|x_i)\log\frac{T(\theta_j|x_i)}{T(\theta_j)}&=I(X;\Theta)=H(X)-H(X|\Theta)\end{aligned} \tag{11}$$

where $N = N_1 + N_2 + \ldots + N_n$, $H(X)$ is the Shannon entropy of X, Θ is one of a group of models (θ_1, θ_2, ..., θ_n), $H(X|\Theta)$ is the generalized posterior entropy of X, and $I(X; \Theta)$ is the semantic mutual information. This equation shows that the ML criterion is equivalent to the maximum semantic mutual information criterion or the minimum generalized posterior entropy criterion. It is easy to find that when $P(X|\theta_j) = P(X|y_j)$ (for all j), the semantic mutual information $I(X; \Theta)$ will be equal to the Shannon mutual information $I(X; Y)$; the latter is the special case of the former.

3.2 The Optimization of Predictive Models or Semantic Channels

About how we get membership functions, we accept the statistical explanation of random sets [18]. However, to understand the evolution of membership functions, we may explain that membership functions evolves when they match the transition probability function, or say, semantic channels evolve when they match Shannon channels.

Optimizing a predictive model Θ is equivalent to optimizing a semantic Channel $T(\Theta|X)$. For given y_j, optimizing θ_j is equivalent to optimizing $T(\theta_j|X)$ by

$$T*(\theta_j|X) = \underset{T(\theta_j|X)}{\arg\max} I(X;\theta_j) \tag{12}$$

$I(X; \theta_j)$ can be written as the difference of two KL divergences:

$$I(X;\theta_j) = \sum_i P(x_i|y_j)\log\frac{P(x_i|y_j)}{P(x_i)} - \sum_i P(x_i|y_j)\log\frac{P(x_i|y_j)}{P(x_i|\theta_j)} \tag{13}$$

Because the KL divergence is greater than or equal to 0, when

$$P(X|\theta_j) = P(X|y_j) \tag{14}$$

$I(X; \theta_j)$ reaches its maximum and is equal to the KL information $I(X; y_j)$. Let the two sides be divided by $P(X)$; then

$$\frac{T(\theta_j|X)}{T(\theta_j)} = \frac{P(y_j|X)}{P(y_j)} \text{ and } T(\theta_j|X) \propto P(y_j|X) \tag{15}$$

Set the maximum of $T(\theta_j|X)$ to 1. Then we obtain

$$T*(\theta_j|X) = P(y_j|X)/P(y_j|x_j*) \tag{16}$$

where x_j* is the x_i that makes $P(y_j|x_j*)$ be the maximum of $P(y_j|X)$. Generally, it is not easy to get $P(y_j|X)$. Yet, for given $P(X|y_j)$ and $P(X)$, it is easier to get $T(\theta_j|X)$ than to get $P(y_j|X)$ since from Eq. (16), we can obtain

$$T*\left(\theta_j|X\right)=\left[P\left(X|y_j\right)/P(X)\right]/\left[P\left(x_j*|y_j\right)/P\left(x_j*\right)\right] \tag{17}$$

The Eq. (12) fits parameter estimations with smaller samples, and Eqs. (16) and (17) fit non-parameter estimations with larger samples.

Similar to the Maximum-A-Posteriori (MAP) estimation, the above Maximum Semantic Information (MSI) estimation also uses the prior. The difference is that the MAP uses the prior of Y or Θ, whereas the MSI uses the prior of X. The MSI is more compatible with Bayesian prediction.

4 The CM Algorithm for Tests, Estimations, and Mixture Models

4.1 The Semantic Channel of a Medical Test

According to Eq. (11), we can obtain a new iterative algorithm, the CM algorithm, to achieve MMI and ML for uncertain Shannon channels.

For medical tests (see Fig. 4), $A = \{x_0, x_1\}$ where x_0 means no-infected person and x_1 means infected person, and $B = \{y_0, y_1\}$ where y_0 means test-negative and y_1 means test-positive.

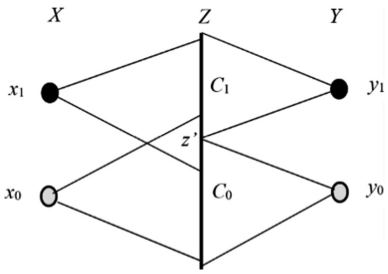

Fig. 4. A 2 × 2 Shannon nosy channel for tests. The channel changes with partition point z'.

In medical tests, the conditional probability in which the test-positive for an infected testee is called sensitivity, and the conditional probability in which the test-negative for an uninfected testee is called specificity [19]. The sensitivity and specificity form a Shannon channel as shown in Table 1.

Table 1. The sensitivity and specificity form a Shannon's channel $P(Y|X)$

Y	Infected x_1	Uninfected x_0
Test-positive y_1	$P(y_1\|x_1)$ = sensitivity	$P(y_1\|x_0)$ = 1−specificity
Test-negative y_0	$P(y_0\|x_1)$ = 1-sensitivity	$P(y_0\|x_0)$ = specificity

Assume that the no-confidence level of y_1 and y_0 are b_1' and b_0' respectively. Table 2 shows the semantic channel for a medical test.

Table 2. Two degrees of disbelief forms a semantic channel $T(\Theta|X)$

Y	Infected x_1	Uninfected x_0
Test-positive y_1	$T(\Theta_1\|x_1) = 1$	$T(\Theta_1\|x_0) = b_1$'
Test-negative y_0	$T(\Theta_0\|x_1) = b_0$'	$T(\Theta_0\|x_0) = 1$

According to Eq. (16), two optimized no-confidence levels are

$$b_1'* = P(y_1|x_0)/P(y_1|x_1); b_1'* = P(y_0|x_1)/P(y_0|x_0) \tag{18}$$

If we use popular likelihood method, when the source $P(X)$ is changed, the old likelihood function $P(X|\theta_j)$ will be improper. However, the above semantic channel is still proper for Bayesian prediction (see Eq. (4)) as well as the Shannon channel in Table 1. Therefore, the SIM can improve the popular likelihood method for variable sources.

4.2 Matching and Iterating

Matching I (Right-step): The semantic channel matches the Shannon channel. We keep the Shannon channel $P(Y|X)$ constant, and optimize the semantic channel $T(\Theta|X)$ (on the right of the log in Eq. (11)) so that $P(X|\theta_j)$ is equal or close to $P(X|y_j)$, or $T(\theta_j|X)$ is proportional or proximately proportional to $P(y_j|X)$ for all j,and hence $I(X; \Theta)$ reaches or approaches its maximum $I(X;Y)$.

Matching II (Left-step): The Shannon channel matches the semantic channel. While keeping the semantic channel $T(\Theta|X)$ constant, we change the Shannon channel $P(Y|X)$ (on the left of the log in Eq. (11)) to maximize $I(X; \Theta)$.

Iterating: The two channels mutually match in turn and iterate. The iterative convergence can be proved pictorially [20].

4.3 The Iterative Process for a Test

For the test as shown in Fig. 4, optimizing the Shannon channel is equivalent to optimizing the dividing point z'. When $Z > z$', we choose y_1; otherwise, we choose y_0.

As an example of the test, $Z \in C = \{1, 2, \ldots, 100\}$ and $P(Z|X)$ is a Gaussian distribution function:

$$P(Z|x_1) = K_1 \exp\left[-(Z - c_1)^2/\left(2d_1^2\right)\right], P(Z|x_0) = K_0 \exp[-(Z - c_0)^2/\left(2d_0^2\right)]$$

where K_1 and K_0 are normalizing constants. From $P(X)$ and $P(Z|X)$, we can obtain $P(X|Z)$. After setting the starting z', say $z' = 50$, as the input of the iteration, we perform the iteration as follows.

Right-Step: Calculate the following items in turn.

(1) Four transition probabilities $P(y_j|x_i)$ $(i,j = 0,1)$ for the Shannon channel:
(2) The b_1'* and b_0'* according to Eq. (18);
(3) $T(\theta_1) = P(x_1) + b_1$'*$P(x_0)$ and $T(\theta_0) = P(x_0) + b_0$'*$P(x_1)$;
(4) $I(X; \theta_1|Z)$ and $I(X; \theta_0|Z)$ for given Z (displaying as two curves):

$$I(X;\theta_j|z_k) = \sum_i P(x_i|z_k)I(x_i;\theta_j), \quad k = 1,2,\ldots,100; j = 0,1 \tag{19}$$

Left-Step: Compare two information function curves $I(X; \theta_1|Z)$ and $I(X; \theta_0|Z)$ over Z to find their cross point. Use z under or over this point as new z'. If the new z' is the same as the last z' then let $z^* = z'$ (z^*: optimal z') and quit the iteration; otherwise go to the Right-step. We may also use the following formula as classification function for new Shannon channel:

$$P(y_j|Z) = \lim_{s\to\infty} \frac{P(y_j)[\exp(I(X;\theta_j|Z))]^s}{\sum_{j'} P(y_{j'})[\exp(I(X;\theta_{j'}|Z))]^s}, \quad j = 1,2,\ldots,n \tag{20}$$

4.4 Two Iterative Examples for Tests and Estimations

Iterative Example 1 (for a 2 × 2 Shannon Channel)

Input Data: $P(x_0) = 0.8$; $c_0 = 30$, $c_1 = 70$; $d_0 = 15$, $d_1 = 10$. The start point $z' = 50$.
Iterative Process: After the first Left-step, we get $z' = 53$; after the second Matching II, we get $z' = 54$; after the third Left-step, we get $z^* = 54$.

Iterative Example 2 (for a 3 × 3 Shannon channel)

This example is to examine a simplified estimation. The semantic channel is little complicated. The principle is the same as that for the test. A pair of good start points and a pair of bad start points are used to examine the reliability and speed of the iteration.

Input Data: $P(x_0) = 0.5$, $P(x_1) = 0.35$, and $P(x_2) = 0.15$; $c_0 = 20$, $c_1 = 50$, and $c_2 = 80$; $d_0 = 15$, $d_1 = 10$, and $d_2 = 10$.

Iterative Results:

(1) With the good start points: $z_1' = 50$ and $z_2' = 60$, the number of iterations is 4; $z_1^* = 35$ and $z_2^* = 66$.

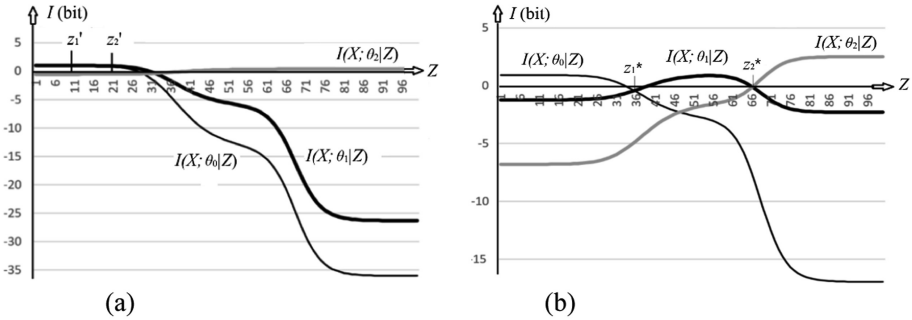

Fig. 5. The iteration with bad start points shows that the convergence is reliable. At the beginning (a), three information curves have small positive areas. At the end (b), three information curves have large positive areas so that $I(X; \Theta)$ reaches its maximum.

(2) With the bad start points: z_1' = 9 and z_2' = 20, the number of iterations is 11; z_1* = 35 and z_2* = 66 also. Figure 5 shows the information curves over Z before and after the iterative process.

4.5 Explaining the Evolution of Semantic Meaning

We may apply the CM algorithm to general predictions, such as weather forecasts. The difference is that the truth functions of predictions may be various. Then we can explain semantic evolution. A Shannon channel indicates a language usage, whereas a semantic channel indicates the comprehension of the audience. The Right-step is to let the comprehension match the usage, and the Left-step is to let the usage (including the observations and discoveries) match the comprehension. The mutual matching and iterating of two channels means that linguistic usage and comprehension mutually match and promote. Natural languages should have been evolving in this way.

4.6 The CM Algorithm for Mixture Models

A popular iterative algorithm for mixture models is the EM algorithm [7]. We can also use the CM algorithm to solve maximum likelihood mixture models or minimum relative entropy mixture models [21]. The convergence proof of the CM algorithm, without using Jensen's inequality, is clearer than that of the EM algorithm.

Table 3 shows an example [21]. Two Gussian distribution components with a group of real parameters produce the mixed distribution $P(X)$. Some guessed parameters (including $P(Y)$) are used to produce the mixed distribution $Q(X)$. The less the relative entropy or KL divergence $H(Q||P)$ is, the better the model is. For $H(Q||P) < 0.001$ bit, the number of iterations is 5.

In this example, Shannon mutual information with real parameters is less than that with start parameters. This example is a challenge to all authors who prove the standard EM algorithm convergent. For this example, maximizing likelihoods Q (in [7, 22]) or $Q + H(y)$ (in [23]) cannot be successful because Q or $Q + H(y)$ with true parameters

Table 3. Real and guessed model parameters and iterative results

	Real parameters in $P^*(X\|Y)$ & $P^*(Y)$			Starting parameters; $H(Q\|\|P)$ = 0.680 bit			Parameters after 5 right-steps; $H(Q\|\|P)$ = 0.00092 bit		
	c	d	$P^*(Y)$	c	d	$P(Y)$	c	d	$P(Y)$
y_1	35	8	0.1	30	8	0.5	38	9.3	0.134
y_2	65	12	0.9	70	8	0.5	65.8	11.5	0.866

may be less than Q or $Q + H(y)$ with starting parameters. About how the CM algorithm solves this problem, see [21] for details.

5 Further Studies

5.1 Optimizing Membership Functions Under the Frame of Factor Space Theory

The factor space theory proposed by Wang [18] is a proper frame for knowledge representation and reasoning. Under this frame, many researchers have made meaningful results [24]. In these studies, the background distribution of objective facts in the factor space is not probability distribution, and the membership function is also not related to the probability distribution of facts. Now, we can use the probability distribution of facts in the factor space as the background distribution, by which we can set up the mutually matching relationship between the membership function and the probability distribution.

We use fuzzy color classification as example. The factor value of a color is a three primary color vector (r, g, b). The factor space R-G-B is a cubic with side length 1. The color vectors (0, 0, 0), (1, 0, 0), (1, 1, 0), (0, 1, 0), (0, 1, 1), (0, 0, 1), (1, 0, 1), and (1, 1, 1) represent typical black, red, yellow, green, cyan, blue, magenta, and white colors respectively. Assume the prior probability distribution of all possible colors in the R-G-B space is $P(\mathbf{X})$. For given y_1 = "red color", the distribution of $\mathbf{X}$ is $P(\mathbf{X}|y_1)$. If the sample is big enough, we can have continuous $P(\mathbf{X}|\ y_1)$ and $P(\mathbf{X})$. Using Eq. (17), we can obtain the optimized truth function $T^*(\theta_1|\mathbf{X})$ of y_1 = "$\mathbf{X}$ is red" or the membership function of fuzzy set θ_1. If the sample is not big enough, we may use Eq. (12) to obtain $T^*(\theta_1|\mathbf{X})$ with some parameters.

If we classify people into fuzzy sets {childhood}, {juvenile}, {youth}, {adult}, {middle-ager}… or classify weathers into {no rain}, {small rain}, {moderate rain}, {moderate to heavy rain}, {heavy rain}, … we may use similar method to obtain the membership function of each fuzzy set. This method does not require that these subsets form a partition or a fuzzy partition of A. That means we may allow fuzzy sets {adult} and {middle-ager}, or {moderate rain} and {moderate to heavy rain} (one may imply another), in A at the same time.

For a GPS device, main factors are the distances between the device and three satellites. Using this Semantic Information Method (SIM), we may eliminate the systematical deviation from other factors. Assume the Shannon channel of a GPS device is:

$$P(y_j|X) = K\exp[-|X - x_j - \Delta x|^2/(2d^2)],\ j = 1, 2, \ldots, n \tag{21}$$

where x_j denotes pointed position, y_j = "$X = x_j$", K is a constant, $\mathbf{\Delta}$e is systematical deviation, and d denotes the precision. According Eq. (16), the corresponding semantic channel is

$$T(\theta_k|X) = \exp[-|X - x_k|^2/(2d^2)],\ k = 1, 2, \ldots, n \tag{22}$$

where $x_k = x_j + \Delta x$. About the applications of the SIM with the factor space theory, we need further studies.

5.2 Optimizing Predictions with Information Value as Criterion

The incremental entropy proposed by Lu [25] is a little different from that introduced by Cover and Thomas [17] (Chap. 16). The information for Lu's incremental entropy is semantic information. The incremental entropy is

$$U(\mathbf{X}|\theta_j) = \sum_{i=1}^{W} P(\mathbf{x}_i|\theta_j)\log R_i = \sum_{i=1}^{W} P(\mathbf{x}_i)\log \sum_{k=0}^{N} q_k R_{ik} \tag{23}$$

where x_i is a price vector of a portfolio; there are W possible price vectors. The $\boldsymbol{\theta}_j$ is the model parameters of prediction y_j. The number of securities in the portfolio is N. R_{ik} means the input-output ratio of the k-th security when $\mathbf{X} = \mathbf{x}_i$, and hence R_i is the input-output ratio of the portfolio as $\mathbf{X} = \mathbf{x}_i$. The q_k is investment ratio in the k-th security and q_0 means the ratio of safe asset or cash. U is the doubling rate of the portfolio.

If without the prediction y_j, $U = U(\mathbf{X})$, then there is the increment of U or information value brought by the information:

$$V(\mathbf{X};\theta_j) = \sum_{i}^{W} P(\mathbf{x}_i|\theta_j)\log[R_i(\mathbf{q}(y_j)*)/R_i(\mathbf{q}*)] \tag{24}$$

where $\mathbf{q}*$ is the optimal vector of investment ratios without the prediction, and $\mathbf{q}(y_j)*$ is the optimal vector based on the prediction.

In some cases, the information value criterion should be better than the information criterion. We need further studies for predictions with the information value criterion.

6 Conclusion

This paper restates Lu's semantic information method to clarify that his semantic information measure is defined with average log normalized likelihood, discusses the semantic channel and its optimization and evolution, and reveals that by letting the semantic channel and Shannon channel mutually match and iterate, we can achieve the maximum Shannon mutual information and maximum average log-likelihood for tests, estimations, and mixture models. Several iterative examples show that the CM algorithm has high speed, clear convergence reasons [20, 21], and wide potential applications[1].

The paper also concludes that the tight combination of Shannon information theory with likelihood method and fuzzy sets theory is necessary for hypothesis testing; with Lu's semantic information method, the combination is feasible.

Acknowledgement. The author thanks Professor Peizhuang Wang for his long term supports. Without his recent encouragement, the author wouldn't have continued researching to find the channels' matching algorithm.

References

1. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–429 (1948). 623–656
2. Popper, K.: Conjectures and Refutations. Routledge, London/New York (1963/2005)
3. Fisher, R.A.: On the mathematical foundations of theoretical statistics. Philo. Trans. Roy. Soc. **222**, 309–368 (1922)
4. Davidson, D.: Truth and meaning. Synthese **17**, 304–323 (1967)
5. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**(3), 338–353 (1965)
6. Kok, M., Dahlin, J., Schon, B., Wills, T.B.: A Newton-based maximum likelihood estimation in nonlinear state space models. IFAC-PapersOnLine **48**, 398–403 (2015)
7. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc., Ser. B **39**, 1–38 (1977)
8. Barron, A., Roos, T., Watanabe, K.: Bayesian properties of normalized maximum likelihood and its fast computation. In: IEEE IT Symposium on Information Theory, pp. 1667–1671 (2014)
9. Lu, C.: B-fuzzy set algebra and a generalized cross-information equation. Fuzzy Syst. Math. (in Chin.) **5**(1), 76–80 (1991)
10. Lu, C.: A Generalized Information Theory (in Chinese). China Science and Technology University Press, Hefei (1993)
11. Lu, C.: Meanings of generalized entropy and generalized mutual information for coding. J. China Inst. Commun. (in Chin.) **15**(6), 37–44 (1994)
12. Lu, C.: A generalization of Shannon's information theory. Int. J. Gen. Syst. **28**(6), 453–490 (1999)

[1] More examples and the excel files for demonstrating the iterative processes can be found at http://survivor99.com/lcg/CM.html.

13. Dubois, D., Prade, H.: Fuzzy sets and probability: misunderstandings, bridges and gaps. In: Second IEEE International Conference on Fuzzy Systems, 28 March, 1 April (1993)
14. Bar-Hillel, Y., Carnap, R.: An outline of a theory of semantic information. Technical report No.247, Research Lab. of Electronics, MIT (1952)
15. Kullback, S., Leibler, R.: On information and sufficiency. Ann. Math. Stat. **22**, 79–86 (1952)
16. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **19**, 716–723 (1974)
17. Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. Wiely, New York (2006)
18. Wang, P.W.: Fuzzy Sets and Random Sets Shadow (in Chinese). Beijing Normal University Press, Beijing (1985)
19. Thornbury, J.R., Fryback, D.G., Edwards, W.: Likelihood ratios as a measure of the diagnostic usefulness of excretory urogram information. Radiology **114**(3), 561–565 (1975)
20. Lu, C.: The Semantic Information Method for Maximum Mutual Information and Maximum Likelihood of Tests, Estimations, and Mixture Models. https://arxiv.org/abs/1706.07918, 24 June 2017
21. Lu, C.: Channels' matching algorithm for mixture models. In: Proceedings of International Conference on Intelligence Science, Shanghai, pp. 25–28, October 2017
22. Wu, C.F.J.: On the convergence properties of the EM algorithm. Ann. Stat. **11**(1), 95–103 (1983)
23. Neal, R., Hinton, G.: A view of the EM algorithm that justifies incremental, sparse, and other variants. In: Jordan, M.I. (ed) Learning in Graphical Models, pp. 355–368. MIT Press, Cambridge (1990)
24. Wang, P.Z.: Factor space and data science. J. Liaoning Tech. Univ. **34**(2), 273–280 (2015)
25. Lu, C.: Entropy Theory of Portfolio and Information Value (in Chinese). Science and Technology University Press, Hefei (1997)

AQ2

Author Query Form

Book ID : **466506_1_En**

Chapter No : **19**

Please ensure you fill out your response to the queries raised below and return this form along with your corrections.

Dear Author,
During the process of typesetting your chapter, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query Refs.	Details Required	Author's Response
AQ1	Please note that the Equations are sequentially renumbered from Eq. (2.1) on. Kindly check.	
AQ2	Please check and confirm the edit made in the page range for Ref. [22].	

MARKED PROOF

Please correct and return this set

Please use the proof correction marks shown below for all alterations and corrections. If you wish to return your proof by fax you should ensure that all amendments are written clearly in dark ink and are made well within the page margins.

Instruction to printer	*Textual mark*	*Marginal mark*
Leave unchanged	··· under matter to remain	New matter followed by
Insert in text the matter indicated in the margin		or (x2)
Delete	through single character, rule or underline or through all characters to be deleted	or (x2)
Substitute character or substitute part of one or more word(s)	through letter or through characters	new character or new characters
Change to italics	under matter to be changed	
Change to capitals	under matter to be changed	
Change to small capitals	under matter to be changed	
Change to bold type	under matter to be changed	
Change to bold italic	under matter to be changed	
Change to lower case	Encircle matter to be changed	
Change italic to upright type	(As above)	
Change bold to non-bold type	(As above)	
Insert 'superior' character	through character or where required	or under character e.g. or
Insert 'inferior' character	(As above)	over character e.g.
Insert full stop	(As above)	
Insert comma	(As above)	
Insert single quotation marks	(As above)	or and/or or
Insert double quotation marks	(As above)	or and/or or
Insert hyphen	(As above)	
Start new paragraph		
No new paragraph		
Transpose		
Close up	linking characters	
Insert or substitute space between characters or words	through character or where required	
Reduce space between characters or words	between characters or words affected	