Reconsideration of Potential Problems of Applying EMIM for Text Analysis

D. Cai
School of Computing and Engineering
University of Huddersfield, HD1 3DH, UK
Email: d.cai@hud.ac.uk

***Abstract*—It seems that the term dependence methods developed using the expected mutual information measure (EMIM) have not achieved their potential in many areas of science, involving statistical text analysis or document processing. This study examines the reasons for the failure and highlights potential problems of applications. Several interesting questions are arisen, including, does a term provide any information if it occurs in all the sample documents? how the mutual information of two terms, under their status values, makes contribution to EMIM? are two terms highly dependent for their co-occurrence if they receive a high positive EMIM value? what may imply for dependence of two term pairs when they receive the same EMIM value? how can properly verify two terms to be high dependent for their co-occurrence? how can properly apply EMIM? does the size of the sample set matter? This study attempts to answer these questions in order to clarify confusions caused by the problems and/or suggest solutions to the problems. Some interesting examples are provided to clarify our viewpoints.**

***Index Terms*—text analysis; term dependence; term co-occurrence; the expected mutual information measure (EMIM).**

I. Introduction

The *expected mutual information measure* (EMIM) quantifies how much knowing one of two variables reduces our uncertainty about the other. The effectiveness of measuring the *mutual information of terms* (MIT) is an active research subject in many areas of science. This subject has been motivated by the concern: to developed a variety of techniques in order to assign a 'dependence' ('relatedness', 'proximity', 'association') value to each term pair, and then to make some decision based on those values. Many studies have used EMIM for a variety of tasks in, for instance, feature selection [1]–[4], document classification [5], face image clustering [6], noise and redundancy reduction [7], multi-modality image registration [8], information retrieval [9]–[13].

Despite the attractiveness of EMIM, however, it seems that the term dependence methods developed using EMIM have not achieved their potential. There may be two main issues for this. First, it is practically difficult to estimate the probability distributions required in EMIM. Second, different estimations conclude to different properties of EMIM and it is theoretically challenging to apply EMIM without clearly understanding the properties. This study focuses on the second issue.

There exist potential problems in applying EMIM. This study examines the reasons for the failure by analysing the properties, particularly when considering the binary probability estimation, denoted by $P_{\equiv}(\delta_i)$ and $P_{\equiv}(\delta_i, \delta_j)$, widely used in many areas of science. We highlight eight problems through respective eight questions below: For two arbitrary distinct terms t_i and t_j (where $I(\delta_i;\delta_j)$ is EMIM and $emim(\delta_i;\delta_j)$ is a simplified form, which will be given in the next section),

- Q1: does t_i provide any information on t_j if it occurs in all the sample documents?
- Q2: what is a fact given from the relation between $I(\delta_i;\delta_j)$ and $emim(\delta_i;\delta_j)$?
- Q3: how the mutual information of t_i and t_j, under their status values, makes contribution to $I(\delta_i;\delta_j)$?
- Q4: are t_i and t_j highly dependent for their co-occurrence if (t_i, t_j) receives a high positive value of $I(\delta_i;\delta_j)$?
- Q5: what may imply for dependence of two term pairs (t_i, t_j) and (t'_i, t'_j) when $I(\delta_i;\delta_j) = I(\delta'_i;\delta'_j)$?
- Q6: how can properly verify t_i and t_j to be high dependent for their co-occurrence?
- Q7: how can properly apply $emim(\delta_i;\delta_j)$?
- Q8: does the size of the sample set matter?

This study attempts to answer the above questions in order to clarify confusions caused by the problems and/or suggest solutions to the problems. As it will be seen from this study, for instance, the occurrence of a term in all samples (which may be regarded as a good term in some applications) does not provide any information about the occurrence of other terms in the samples; two terms receiving a high positive EMIM value may not be necessarily high dependent for their co-occurrence; two term pairs receiving the same EMIM value may be dependent of each other in different implications; an inequality has to be verified, in order to properly apply EMIM or $emim$, to ensure two terms are high dependent for their co-occurrence. Some interesting examples are provided to clarify our viewpoints, and each question Qk is answered through a corresponding remark Remark-k ($k = 1, 2, ..., 8$).

The remainder of the paper is organized as follows. Section 2 gives notation, the expressions of EMIM and $emim$. Section 3 considers the properties of EMIM and answers Q1 and Q2. Section 4 analyses the properties of four MIT measures, derived from EMIM, and answers Q3–Q7. Section 5 explains the sensitivity to the size of the sample set and answers Q8. Conclusions are drawn in Section 6 and detailed proofs of all the theorems given in this study are presented in Appendix.

II. Background

This section gives notation, expressions of EMIM and its simplified form.

Let D be a *collection* of documents, $\Xi \subseteq D$ a sample set of documents interested, and V be a *vocabulary* of terms used to index individual documents in D. Denote $V_d \subseteq V$ as the set of terms occurring in document d, and $V_\Xi \subseteq V$ as the set of terms occurring in at least one of sample documents in Ξ.

In order to clarify our idea presented in this study, let us first give term state value distributions. A term is usually thought of having its state values $present$ or $absent$ in a document or a set of documents. For an arbitrary term $t \in V$, it will be convenient to introduce a variable δ taking values from set $\Omega = \{1, 0\}$, where $\delta = 1$ expresses that t is present and $\delta = 0$ expresses that t is absent. Denote $t^\delta = t$, $\bar{t}$ when $\delta = 1, 0$, respectively. We call $\Omega = \{1, 0\}$ a *state value space*, and each element in Ω a *state value*, of the term t. Thus, for a given term $t \in V_d$, its state distribution, denoted by $P_d(\delta) = P(t^\delta|d)$, is over Ω. Similar discussions can be given to $P_\Xi(\delta) = P(t^\delta|\Xi)$ over Ω for $t \in V_\Xi$, and to $P_\Xi(\delta_i, \delta_j) = P(t_i^{\delta_i}, t_j^{\delta_j}|\Xi)$ over $\Omega \times \Omega = \{(1,1), (1,0), (0,1), (0,0)\}$ for $(t_i, t_j) \in V_\Xi \times V_\Xi$ (where $i \neq j$).

There exists dependence between two terms if the state value of one of them provides mutual information about the probability of the state value of another. The study [14] also showed that there is a relationship between the frequencies (or probabilities) and the mutual information of terms. Therefore, term t taking some state value δ should be looked upon as complex because another state value of t, and state values of many other terms, may be dependent on this state value [?].

To enable to analyse and understand the properties of EMIM and its a simplified form, let us further denote $n_\Xi(t)$ as the number of samples in Ξ in which t occurs, and $n_\Xi(t_i, t_j)$ as the number of samples in Ξ in which t_i and t_j co-occur (where $i \neq j$). Then, under the binary assumption, using the statistics of the sample frequencies concerning the set Ξ, we can introduce the following two theorems, which are essential for estimating probability distributions required in EMIM.

Theorem 2.1 For an arbitrary term $t \in V$, the state value distribution, denoted by $P_\Xi(\delta)$, given by

$$\begin{aligned} P_\Xi(\delta = 1) &= P_\Xi(t) = \frac{n_\Xi(t)}{|\Xi|} \\ P_\Xi(\delta = 0) &= P_\Xi(\bar{t}) = 1 - \frac{n_\Xi(t)}{|\Xi|} \end{aligned} \tag{1}$$

is a probability distribution over Ω. For two arbitrary distinct terms $t_i, t_j \in V$, the state value distribution, denoted by $P_\Xi(\delta_i, \delta_j)$, given by

$$\begin{aligned} P_\Xi(\delta_i = 1, \delta_j = 1) &= P_\Xi(t_i, t_j) = \frac{n_\Xi(t_i, t_j)}{|\Xi|} \\ P_\Xi(\delta_i = 1, \delta_j = 1) &= P_\Xi(t_i, t_j) = \frac{n_\Xi(t_i, t_j)}{|\Xi|} \\ P_\Xi(\delta_i = 0, \delta_j = 1) &= P_\Xi(\bar{t}_i, t_j) = \frac{n_\Xi(t_j) - n_\Xi(t_i, t_j)}{|\Xi|} \\ P_\Xi(\delta_i = 0, \delta_j = 0) &= P_\Xi(\bar{t}_i, \bar{t}_j) \\ &= \frac{|\Xi| - n_\Xi(t_i) - n_\Xi(t_j) + n_\Xi(t_i, t_j)}{|\Xi|} \end{aligned} \tag{2}$$

is a probability distribution over $\Omega \times \Omega$. And $P_\Xi(\delta_i)$ and $P_\Xi(\delta_j)$ are the marginal distributions of $P_\Xi(\delta_i, \delta_j)$.

Theorem 2.2 For two arbitrary distinct terms $t_i, t_j \in V$, suppose $P_\Xi(\delta)$ and $P_\Xi(\delta_i, \delta_j)$ are given in Eq.(1) and Eq.(2), respectively. Then $P_\Xi(\delta_i, \delta_j)$ is absolutely continuous with respect to product $P_\Xi(\delta_i)P_\Xi(\delta_j)$ for $\delta_i, \delta_j = 1, 0$.

With Theorems 2.1 and 2.2, we can now substitute Eq.(1) and Eq.(2) into EMIM:

$$I_\Xi(\delta_i; \delta_j) = \sum_{\delta_i, \delta_j = 0,1} P_\Xi(\delta_i, \delta_j) \ln \frac{P_\Xi(\delta_i, \delta_j)}{P_\Xi(\delta_i)P_\Xi(\delta_j)} \tag{3}$$

where $\ln$ is the natural logarithm, which measures the amount of information that δ_j provides about δ_i, and vice versa.

In order to give a simplified form of EMIM, denoted by $emim_\Xi(\delta_i; \delta_j)$, let us adopt the notation given in [15]:

$$\begin{aligned} n_{1\cdot} &= n_\Xi(t_i) \\ n_{\cdot 1} &= n_\Xi(t_j) \\ n_{11} &= n_\Xi(t_i, t_j) \\ n_{10} &= n_\Xi(t_i) - n_\Xi(t_i, t_j) \\ n_{01} &= n_\Xi(t_j) - n_\Xi(t_i, t_j) \\ n_{0\cdot} &= |\Xi| - n_\Xi(t_i) \\ n_{\cdot 0} &= |\Xi| - n_\Xi(t_j) \\ n_{00} &= |\Xi| - n_\Xi(t_i) - n_\Xi(t_j) + n_\Xi(t_i, t_j) \end{aligned} \tag{4}$$

Then we can write

$$\begin{aligned} emim_\Xi(\delta_i; \delta_j) = n_{11} \ln \frac{n_{11}}{n_{1\cdot} n_{\cdot 1}} + n_{10} \ln \frac{n_{10}}{n_{1\cdot} n_{\cdot 0}} + \\ n_{01} \ln \frac{n_{01}}{n_{0\cdot} n_{\cdot 1}} + n_{00} \ln \frac{n_{00}}{n_{0\cdot} n_{\cdot 0}} \end{aligned} \tag{5}$$

which is well-known to many researchers, in particular, to information retrieval (IR) researchers. It was initially introduced by van Rijsbergen in his earlier book and papers [15], [16].

We will give the relation between EMIM and $emim$ and provide an example to illustrate the computation involved in EMIM and $emim$ in next section. In what follows, we will always assume, when mentioning two arbitrary terms $t_i, t_j \in V$, that they are distinct terms (i.e., $i \neq j$).

III. Properties of EMIM

In order to enable us to gain an insight into $I_\Xi(\delta_i; \delta_j)$ and $emim_\Xi(\delta_i; \delta_j)$, this section introduces three theorems. These give interesting properties of EMIM and $emim$, and then give answers to questions Q1 and Q2.

Theorem 3.1 For two arbitrary terms $t_i, t_j \in V$, suppose $P_\Xi(\delta)$ and $P_\Xi(\delta_i, \delta_j)$ are given in Eq.(1) and Eq.(2), respectively. Then $I_\Xi(\delta_i; \delta_j) = 0$ if $n_\Xi(t_i) = |\Xi|$ or $n_\Xi(t_j) = |\Xi|$.

Remark-1: Theorem 3.1 tells us, when EMIM is used with the estimation given Eq.(1) and Eq.(2), that the occurrence of t_i in all samples does not provide any information about the occurrence of t_j in the samples. Thus, t_i and t_j are statistically independent of one another with respect to Ξ. Consequently, in order to capture the dependence information of terms, we should always avoid many terms having $n_\Xi(t) = |\Xi|$ and take the sample set Ξ with a relatively larger size satisfying, for instance,

$$|\Xi| \geq \alpha + \beta \times \max\{n_\Xi(t) \mid t \in V_\Xi\}$$

where $\alpha, \beta \geq 1$ are integers. $\Diamond$

Theorem 3.2 For two arbitrary terms $t_i, t_j \in V$, suppose $I_{\Xi}(\delta_i;\delta_j)$ and $emim_{\Xi}(\delta_i;\delta_j)$ are given in Eq.(3) and Eq.(5), respectively. Then

$$I_{\Xi}(\delta_i;\delta_j) = \frac{1}{n} \times emim_{\Xi}(\delta_i;\delta_j) + \ln(n) \tag{6}$$

where $n = |\Xi|$.

Remark-2: Theorem 3.2 gives the relation between $I_{\Xi}(\delta_i;\delta_j)$ and $emim_{\Xi}(\delta_i;\delta_j)$. Many applications use $emim_{\Xi}(\delta_i;\delta_j)$, rather than $I_{\Xi}(\delta_i;\delta_j)$, as a scale factor $\frac{1}{n}$ and a constant $\ln(n)$ are independent of all term pairs $(t_i, t_j) \in V \times V$, and thus they are eliminated for simplifying computation. It is clear that an essential difference between Eq.(3) and Eq.(5) is: the former is normalized by n but the latter is not. An important fact given by the above relation to notice is: $I_{\Xi}(\delta_i;\delta_j) \geq 0$ cannot infer $emim_{\Xi}(\delta_i;\delta_j) \geq 0$. Theorem 3.3 below is interesting. $\Diamond$

Theorem 3.3 For two arbitrary terms $t_i, t_j \in V$, suppose $emim_{\Xi}(\delta_i;\delta_j)$ is given in Eq.(5). Then $emim_{\Xi}(\delta_i;\delta_j) \leq 0$.

Example 3.1 Suppose $\Xi = \{d_1, d_2, d_3\} \subseteq D$ is a sample set, $V_{d_1} = \{t_1, t_2, t_4, t_5, t_6, t_8\}$, $V_{d_2} = \{t_1, t_3, t_4, t_5, t_6, t_7\}$ and $V_{d_3} = \{t_2, t_4, t_6\}$. From $n_{\Xi}(t_1, t_2) = 1$, $n_{\Xi}(t_1) = 2$ and $n_{\Xi}(t_2) = 2$, we have

$$\begin{aligned}
I_{\Xi}(\delta_1;\delta_2) &= \frac{1}{3}\ln\frac{\frac{1}{3}}{\frac{2}{3}\frac{2}{3}} \\
&\quad + \frac{2-1}{3}\ln\frac{\frac{2-1}{3}}{\frac{2}{3}(1-\frac{2}{3})} \\
&\quad + \frac{2-1}{3}\ln\frac{\frac{2-1}{3}}{(1-\frac{2}{3})\frac{2}{3}} \\
&\quad + \frac{3-2-2+1}{3}\ln\frac{\frac{3-2-2+1}{3}}{(1-\frac{2}{3})(1-\frac{2}{3})} \\
&= \frac{1}{3}\ln\frac{3}{4} + \frac{1}{3}\ln\frac{3}{2} + \frac{1}{3}\ln\frac{3}{2} + 0\ln 0 \\
&\approx -0.0959 + 0.1352 + 0.1352 - 0.0000 \\
&= 0.1745 \\
emim_{\Xi}(\delta_1;\delta_2) &= 1 \times \ln\frac{1}{2\times 2} \\
&\quad + (2-1)\ln\frac{2-1}{2\times(3-2)} \\
&\quad + (2-1)\ln\frac{2-1}{(3-2)\times 2} \\
&\quad + (3-2-2+1)\ln\frac{3-2-2+1}{(3-2)\times(3-2)} \\
&= \ln\frac{1}{4} + \ln\frac{1}{2} + \ln\frac{1}{2} + 0\ln\frac{0}{1} \\
&\approx -1.3863 - 0.6931 - 0.6931 - 0.0000 \\
&= -2.7725.
\end{aligned}$$

Also, with the expression given in Eq.(6), we can see

$$\begin{aligned}
&\frac{1}{3} \times emim_{\Xi}(\delta_1;\delta_2) + \ln(3) \\
&\approx \frac{1}{3} \times (-2.7725) + 1.0986 \\
&\approx 0.1745 = I_{\Xi}(\delta_1;\delta_2)
\end{aligned}$$

which verifies the relation between $I_{\Xi}(\delta_1;\delta_2)$ and $emim_{\Xi}(\delta_1;\delta_2)$ for terms t_1 and t_2. $\triangle$

IV. Properties of MIT Measures

This section gives four measures of mutual information of terms (MIT), and then clarifies our viewpoints, which are used for answering questions Q3–Q7. The answers are essential for guiding practical applications.

Following the studies in [17] [18], we express EMIM given in Eq.(3) with the sum of four items,

$$\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j}) = P_{\Xi}(\delta_i, \delta_j) \ln \frac{P_{\Xi}(\delta_i, \delta_j)}{P_{\Xi}(\delta_i) P_{\Xi}(\delta_j)} \tag{7}$$

where $\delta_i, \delta_j = 0, 1$, each of which can be regarded as '**m**utual **i**nformation of **t**erms, t_i and t_j, in support of dependence rejecting independence under state value (δ_i, δ_j). Thus, we can regard it as a general form of a MIT measure, computing the extent of the contributions made by t_i and t_j under the corresponding state values to $I_{\Xi}(\delta_i;\delta_j)$. The four MIT measures and example below enable a simple answer to the third question.

Example 4.1 Substituting the probability distributions given in Eq.(1) and Eq.(2) into the MIT measure in Eq.(7), we can write four concrete MIT measures for $\delta_i, \delta_j = 0, 1$. For instance, taking $\delta_i = 1$ and $\delta_j = 1$, we can write the first item of $I_{\Xi}(\delta_i;\delta_j)$:

$$\begin{aligned}
\mathbf{mit}_{\Xi}(t_i, t_j) &= \mathbf{mit}_{\Xi}(t_i^{\delta_i=1}, t_j^{\delta_j=1}) \\
&= P_{\Xi}(t_i, t_j) \ln \frac{P_{\Xi}(t_i, t_j)}{P_{\Xi}(t_i) P(t_j)} \\
&= \frac{n_{\Xi}(t_i, t_j)}{|\Xi|} \ln\Big(\frac{\frac{n_{\Xi}(t_i,t_j)}{|\Xi|}}{\frac{n_{\Xi}(t_i)}{|\Xi|}\frac{n_{\Xi}(t_j)}{|\Xi|}}\Big)
\end{aligned} \tag{8}$$

which is the MIT measure of terms t_i and t_j for their occurrence in Ξ. Also, if taking $\delta_i = 1$ but $\delta_j = 0$, then we have the second item of $I_{\Xi}(\delta_i;\delta_j)$:

$$\begin{aligned}
\mathbf{mit}_{\Xi}(t_i, \bar{t}_j) &= \mathbf{mit}_{\Xi}(t_i^{\delta_i=1}, t_j^{\delta_j=0}) \\
&= P_{\Xi}(t_i, \bar{t}_j) \ln \frac{P_{\Xi}(t_i, \bar{t}_j)}{P_{\Xi}(t_i) P(\bar{t}_j)} \\
&= \frac{n_{\Xi}(t_i) - n_{\Xi}(t_i, t_j)}{|\Xi|} \ln\Big(\frac{\frac{n_{\Xi}(t_i)-n_{\Xi}(t_i,t_j)}{|\Xi|}}{\frac{n_{\Xi}(t_i)}{|\Xi|}(1-\frac{n_{\Xi}(t_j)}{|\Xi|})}\Big)
\end{aligned}$$

which is the MIT measure of term t_i occurring but term t_j not occurring in Ξ. $\triangle$

Remark-3: The expressions Eq.(3) and Eq.(7) tell us, in order to measure the term mutual information, we have to consider the mutual information under the individual state values. That is, we need to measure the extent of the contribution made by the respective four state value pairs (δ_i, δ_j) using the corresponding measure $\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j})$, where $\delta_i, \delta_j = 0, 1$, to the expected mutual information. $\Diamond$

Generally, each MIT measure, $\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j})$, can be positive or negative (which can be seen in Example 3.1). The following theorem, which considers the relation between $\frac{n_{\Xi}(t_i,t_j)}{|\Xi|}$ and $\frac{n_{\Xi}(t_i)}{|\Xi|}\frac{n_{\Xi}(t_j)}{|\Xi|}$, is interesting.

Theorem 4.1 For two arbitrary terms $t_i, t_j \in V$, the four measures, $\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j})$, where $\delta_i, \delta_j = 0, 1$, given in Eq.(7) have the following property.

(1) if $\frac{n_{\Xi}(t_i,t_j)}{|\Xi|} = \frac{n_{\Xi}(t_i)}{|\Xi|}\frac{n_{\Xi}(t_j)}{|\Xi|}$ then

$$\mathbf{mit}_{\Xi}(t_i,t_j)=0, \quad \mathbf{mit}_{\Xi}(t_i,\bar{t}_j)=0,$$
$$\mathbf{mit}_{\Xi}(\bar{t}_i,t_j)=0, \quad \mathbf{mit}_{\Xi}(\bar{t}_i,\bar{t}_j)=0.$$

(2) if $\frac{n_{\Xi}(t_i,t_j)}{|\Xi|} > \frac{n_{\Xi}(t_i)}{|\Xi|}\frac{n_{\Xi}(t_j)}{|\Xi|}$ then

$$\mathbf{mit}_{\Xi}(t_i,t_j)>0, \quad \mathbf{mit}_{\Xi}(t_i,\bar{t}_j)\le 0,$$
$$\mathbf{mit}_{\Xi}(\bar{t}_i,t_j)\le 0, \quad \mathbf{mit}_{\Xi}(\bar{t}_i,\bar{t}_j)>0.$$

(3) if $\frac{n_{\Xi}(t_i,t_j)}{|\Xi|} < \frac{n_{\Xi}(t_i)}{|\Xi|}\frac{n_{\Xi}(t_j)}{|\Xi|}$ then

$$\mathbf{mit}_{\Xi}(t_i,t_j)<0, \quad \mathbf{mit}_{\Xi}(t_i,\bar{t}_j)\ge 0,$$
$$\mathbf{mit}_{\Xi}(\bar{t}_i,t_j)\ge 0, \quad \mathbf{mit}_{\Xi}(\bar{t}_i,\bar{t}_j)<0.$$

Remark-4: By the property given in Theorem 4.1, it can be easily seen, when $\frac{n_{\Xi}(t_i,t_j)}{|\Xi|} < \frac{n_{\Xi}(t_i)}{|\Xi|}\frac{n_{\Xi}(t_j)}{|\Xi|}$, that the positive value $I_{\Xi}(\delta_i;\delta_j)$ is dominated by the positive quantities $\mathbf{mit}_{\Xi}(t_i,\bar{t}_j)$ and/or $\mathbf{mit}_{\Xi}(\bar{t}_i,t_j)$. Thus, the higher value the $I_{\Xi}(\delta_i;\delta_j)$ has, the larger quantities the $\mathbf{mit}_{\Xi}(t_i,\bar{t}_j)$ and/or $\mathbf{mit}_{\Xi}(\bar{t}_i,t_j)$ provide, and the more they indicate that t_i and t_j are highly dependent under state values $(1,0)$ and $(0,1)$, and that they should not co-occur in samples in Ξ. Consequently, a high positive value of $I_{\Xi}(\delta_i;\delta_j)$ may not indicate that t_i and t_j are highly dependent for their occurrence, namely, that the occurrence (absence) of term t_i accompanies the absence (occurrence) of term t_j. ♢

The answer to the fourth question is now apparent. We can clarify our viewpoint by an example below, which can also help to answer the fifth and sixth questions.

Example 4.2 Suppose $\Xi=\{d_1,d_2,d_3\}$, $V_{d_1}=\{t_1,t_2,t_3,t_4,t_5\}$, $V_{d_2}=\{t_1,t_4,t_5,t_7\}$ and $V_{d_3}=\{t_4,t_7,t_8\}$. Then, it has $|\Xi|=3$, $n_{\Xi}(t_1)=2$, $n_{\Xi}(t_2)=1$, $n_{\Xi}(t_1,t_2)=1$, and

$$I_{\Xi}(\delta_1;\delta_2)=\frac{1}{3}\ln\frac{\frac{1}{3}}{\frac{2}{3}\cdot\frac{1}{3}}+\frac{1}{3}\ln\frac{\frac{1}{3}}{\frac{2}{3}\cdot\frac{2}{3}}+\frac{0}{3}\ln\frac{\frac{0}{3}}{\frac{1}{3}\cdot\frac{1}{3}}+\frac{1}{3}\ln\frac{\frac{1}{3}}{\frac{1}{3}\cdot\frac{2}{3}}$$
$$\approx 0.1352-0.0959-0.0000+0.1352=0.1745.$$

In this case, the value $I_{\Xi}(\delta_1;\delta_2)$ is dominated by both the quantities $\mathbf{mit}_{\Xi}(t_1,\bar{t}_2)$ and $\mathbf{mit}_{\Xi}(\bar{t}_1,t_2)$, and t_1 and t_2 are highly dependent for their co-occurrence in set Ξ. Also, from $n_{\Xi}(t_5)=2$, $n_{\Xi}(t_7)=2$ and $n_{\Xi}(t_5,t_7)=1$, it has

$$I_{\Xi}(\delta_5;\delta_7)=\frac{1}{3}\ln\frac{\frac{1}{3}}{\frac{2}{3}\cdot\frac{2}{3}}+\frac{1}{3}\ln\frac{\frac{1}{3}}{\frac{2}{3}\cdot\frac{1}{3}}+\frac{1}{3}\ln\frac{\frac{1}{3}}{\frac{1}{3}\cdot\frac{2}{3}}+\frac{0}{3}\ln\frac{\frac{0}{3}}{\frac{1}{3}\cdot\frac{1}{3}}$$
$$\approx -0.0959+0.1352+0.1352-0.0000=0.1745.$$

In this case, the value $I_{\Xi}(\delta_5;\delta_7)$ is dominated by both the quantities $\mathbf{mit}_{\Xi}(t_5,\bar{t}_7)$ and $\mathbf{mit}_{\Xi}(\bar{t}_5,t_7)$, and t_5 and t_7 are highly dependent for their not-co-occurrence in set Ξ. △

Remark-5: It can be seen, from Example 4.2, that two term pairs (t_1,t_2) and (t_5,t_7) receive the same value, $I_{\Xi}(\delta_1;\delta_2)=I_{\Xi}(\delta_5;\delta_7)$. However, the implications of the dependence information under the individual state values are entirely different: terms t_1 and t_2 provide the information highly supporting for either their co-occurrence or none of them occurrence (i.e., co-not-occurrence); whereas terms t_5 and t_7 provide the information highly supporting for one of them occurrence but another not occurrence (i.e., not-co-occurrence). ♢

Remark-6: In a practical application, we normally concentrate on the statistics of co-occurrence of terms. That is, the dependence under which we are really interested is state value $(\delta_i,\delta_j)=(1,1)$. In this case, what we need is:

- to use the measure $\mathbf{mit}_{\Xi}(t_i,t_j)$ given in Eq.(8), and for every $(t_i,t_j)\in V\times V$, to verify an inequality,

$$\frac{n_{\Xi}(t_i,t_j)}{|\Xi|} > \frac{n_{\Xi}(t_i)}{|\Xi|}\times\frac{n_{\Xi}(t_j)}{|\Xi|} \tag{9}$$

- to select those term pairs (t_i,t_j) satisfying the above inequality as they guarantee both $\mathbf{mit}_{\Xi}(t_i,t_j)>0$ (i.e., co-occurrence) and $\mathbf{mit}_{\Xi}(\bar{t}_1,\bar{t}_2)>0$ (i.e., co-not-occurrence).

Then, we remove the term pairs not carrying the information supporting not-co-occurrence. ♢

Example 4.3 (Example 4.2 continued). Consider terms t_1 and t_2, we have

$$\frac{3}{9}=\frac{1}{3}=\frac{n_{\Xi}(t_1,t_2)}{|\Xi|} > \frac{n_{\Xi}(t_1)}{|\Xi|}\frac{n_{\Xi}(t_2)}{|\Xi|}=\frac{2}{3}\frac{1}{3}=\frac{2}{9}$$

From which we know that $\mathbf{mit}_{\Xi}(t_1,t_2)>0$, $\mathbf{mit}_{\Xi}(t_1,\bar{t}_2)<0$, $\mathbf{mit}_{\Xi}(\bar{t}_1,t_2)<0$, $\mathbf{mit}_{\Xi}(\bar{t}_1,\bar{t}_2)>0$, and that t_1 and t_2 are statistically dependent for their co-occurrence in Ξ, Also, if we consider terms t_5 and t_7, then $n_{\Xi}(t_5)=2$, $n_{\Xi}(t_7)=2$, $n_{\Xi}(t_5,t_7)=1$, and

$$\frac{3}{9}=\frac{1}{3}=\frac{n_{\Xi}(t_5,t_7)}{|\Xi|} < \frac{n_{\Xi}(t_5)}{|\Xi|}\frac{n_{\Xi}(t_7)}{|\Xi|}=\frac{2}{3}\frac{2}{3}=\frac{4}{9}$$

From which we know that $\mathbf{mit}_{\Xi}(t_5,t_7)<0$, $\mathbf{mit}_{\Xi}(t_5,\bar{t}_7)>0$, $\mathbf{mit}_{\Xi}(\bar{t}_5,t_7)>0$, $\mathbf{mit}_{\Xi}(\bar{t}_5,\bar{t}_7)<0$, and that t_5 and t_7 are highly dependent for their not co-occurrence in Ξ). △

The following two Corollaries give properties of the MIT measures, that is, of the individual items of $I_{\Xi}(\delta_i;\delta_j)$ and $emim_{\Xi}(\delta_i;\delta_j)$. Their proofs are given in the proofs of Theorem 3.2 and Theorem 3.3, respectively.

Corollary 4.1 For two arbitrary terms $t_i,t_j\in V_{\Xi}$, if $n_{\Xi}(t_i)=|\Xi|$ or $n_{\Xi}(t_j)=|\Xi|$, then $\mathbf{mit}_{\Xi}(t_i^{\delta_i},t_j^{\delta_j})=0$ for $\delta_i,\delta_j=0,1$.

Corollary 4.2 For two arbitrary terms $t_i,t_j\in V_{\Xi}$, the individual items of $emim_{\Xi}(\delta_i;\delta_j)$ are always non-positive.

Remark-7: In order to apply $emim(\delta_i;\delta_j)$ properly, let us compare the first item of $I_{\Xi}(\delta_i;\delta_j)$ given in Eq.(8) and the first item of $emim_{\Xi}(\delta_i;\delta_j)$ given in Eq.(5). Note that we have

$$\frac{n_{\Xi}(t_i,t_j)}{|\Xi|}=\frac{n_{11}}{n} \quad \text{and} \quad \frac{n_{\Xi}(t_i)}{|\Xi|}\frac{n_{\Xi}(t_j)}{|\Xi|}=\frac{n_{1\cdot}}{n}\frac{n_{\cdot 1}}{n}$$

Thus, from the expressions in the respective ln functions of the two first items:

- from the relation between $\frac{n_{\Xi}(t_i,t_j)}{|\Xi|}$ and $\frac{n_{\Xi}(t_i)}{|\Xi|}\frac{n_{\Xi}(t_j)}{|\Xi|}$ given in Theorem 4.1, we can infer all the signs of $\mathbf{mit}_{\Xi}(t_i^{\delta_i},t_j^{\delta_j})$ for $\delta_i,\delta_j=1,0$, and then determine whether term pair (t_i,t_j) is statistically dependent under the individual state values.
- however, the inference and determination cannot be made from the relation between n_{11} and $n_{1\cdot}n_{\cdot 1}$; in fact, by Corollary 4.2, we know that the individual items of $emim_{\Xi}(\delta_i,\delta_j)$ are always non-positive.

Therefore, to solve the problem arisen by Q7, with Remark-6, we need to verify Eq.(9) or, equivalently, to verify a simpler inequality,

$$n_{11}=n_{\Xi}(t_i,t_j) > \frac{1}{|\Xi|}n_{\Xi}(t_i)n_{\Xi}(t_j)=\frac{1}{n}n_{1\cdot}n_{\cdot 1} \tag{10}$$

which is a straightforward way to the solution. ♢

TABLE I
THE DEPENDENCE VALUES AGAINST SIZES OF Ξ

$\lvert\Xi\rvert$	$\mathbf{mit}_{\Xi}(t_1,t_4)$	$\mathbf{mit}_{\Xi}(t_1,\bar{t}_4)$	$\mathbf{mit}_{\Xi}(\bar{t}_1,t_4)$	$\mathbf{mit}_{\Xi}(\bar{t}_1,\bar{t}_4)$	$I_{\Xi}(\delta_1,\delta_4)$
3	0.0000	0.0000	0.0000	0.0000	0.0000
4	0.1438	0.0000	-0.1014	0.1733	0.2157
5	0.2043	0.0000	-0.1176	0.2043	0.2910
6	0.2310	0.0000	-0.1155	0.2027	0.3182
7	0.2421	0.0000	-0.1089	0.1923	0.3255
8	0.2452	0.0000	-0.1014	0.1798	0.3236
9	0.2441	0.0000	-0.0941	0.1675	0.3175
10	0.2408	0.0000	-0.0875	0.1562	0.3095
15	0.2146	0.0000	-0.0637	0.1145	0.2654
20	0.1897	0.0000	-0.0497	0.0896	0.2296
30	0.1535	0.0000	-0.0343	0.0621	0.1813
50	0.1125	0.0000	-0.0212	0.0384	0.1297
100	0.0701	0.0000	-0.0108	0.0196	0.0789
1000	0.0116	0.0000	-0.0011	0.0020	0.0125
10000	0.0016	0.0000	-0.0001	0.0002	0.0017

$n_{\Xi}(t_1) = 2$, $n_{\Xi}(t_4) = 3$, $n_{\Xi}(t_1,t_4) = 2$

V. SIZE OF SAMPLE SET

The binary estimation methods derive their importance from the fact that their simplicity of computation easily enables us to have an insight into the term dependence. However, the methods may be sensitive to the size of the sample set. This sections explains the sensitivity, using the probability estimation given in Eq.(1) and Eq.(2) as an example, and gives an answer to the last question Q8 through a simple example.

Example 5.1 (Example 4.2 continued) Suppose we have a sample set $\Xi \subseteq D = \{d_1, d_2, ..., d_{10000}\}$. Consider two terms t_1 and t_4 with fixed numbers $n_{\Xi}(t_1,t_4) = 2$, $n_{\Xi}(t_1) = 2$ and $n_{\Xi}(t_4) = 3$. Then, when $|\Xi| = 3$, by Theorem 3.1,

$$\begin{aligned} I_{\Xi}(\delta_1;\delta_4) &= \sum_{\delta_1,\delta_4=1,0} \mathbf{mit}_{\Xi}(t_1^{\delta_1}, t_4^{\delta_4}) \\ &= 0.0000 - 0.0000 - 0.0000 + 0.0000 = 0.0000. \end{aligned}$$

Next, taking $|\Xi| = 10$, then

$$\begin{aligned} I_{\Xi}(\delta_1;\delta_4) &= \frac{2}{10}\ln\frac{\frac{2}{10}}{\frac{2}{10}\frac{3}{10}} \\ &\quad + \frac{2-2}{10}\ln\frac{\frac{2-2}{10}}{\frac{2}{10}\left(1-\frac{3}{10}\right)} \\ &\quad + \frac{3-2}{10}\ln\frac{\frac{3-2}{10}}{\left(1-\frac{2}{10}\right)\frac{3}{10}} \\ &\quad + \frac{10-2-3+2}{10}\ln\frac{\frac{10-2-3+2}{10}}{\left(1-\frac{2}{10}\right)\left(1-\frac{3}{10}\right)} \\ &= \frac{2}{10}\ln\frac{10}{3} + 0\ln 0 + \frac{1}{10}\ln\frac{10}{24} + \frac{7}{10}\ln\frac{10}{8} \\ &\approx 0.2408 - 0.0000 - 0.0875 + 0.1562 = 0.3095. \end{aligned}$$

There are more dependence values of t_1 and t_4 against the increasing sizes of Ξ in Table I, in which, the numbers underlined are the maximum (in absolute values) for the corresponding EMIM and MIT measures. As it can been seen from Table I, the values vary as changing of $|\Xi|$ and the variation tells us about the behaviour of the individual measures. $\triangle$

The five different measures give us useful information; each indicates a different aspect about the dependence of terms and so should be interpreted in an appropriate way. Let us now carefully examine Table I to look at what insight it can give regarding $|\Xi|$ for terms t_1 and t_4.

- When $|\Xi| = 3$, it has $n_{\Xi}(t_4) = |\Xi|$, namely, t_4 occurs in all samples in Ξ. In this case, the occurrence of t_4 does not provide any information about the occurrence of t_1 in samples. Thus, t_1 and t_4 is statistically independent of each other, and $\mathbf{mit}_{\Xi}(t_1^{\delta_1}, t_4^{\delta_4}) = 0$ for $\delta_1, \delta_4 = 1, 0$, so $I_{\Xi}(\delta_1;\delta_4) = 0$.
- As increasing of $|\Xi|$, the individual dependence values in each of the columns are increasing (in absolute values) till to the maximum. This is because if t_1 or t_4 occur in several (not many) samples, and also co-occur in some of these, then the values indicate that t_1 and t_4 are dependent to some extent.
- For larger and larger $|\Xi|$, t_1 and t_4 co-occur in less and less samples in Ξ (compared with $|\Xi|$) and they receive lower and lower dependence values. The values drop greatly when $|\Xi| = 100$ and almost are equal to zero when $|\Xi| = 10000 = |D|$.

Generally, when the numbers $n_{\Xi}(t_i,t_j)$, $n_{\Xi}(t_i)$ and $n_{\Xi}(t_j)$ are fixed, we have $\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j}) \to 0$ (for $\delta_i, \delta_j = 1, 0$) and hence $I_{\Xi}(\delta_i;\delta_j) \to 0$, when $|\Xi| \to \infty$. The mathematical reason for this is simple. As it can be seen from the probability estimation given in Eq.(2) and the MIT measures given Eq.(7),

- except the last one, the individual probabilities $P_{\Xi}(\delta_i, \delta_j)$ approach 0, so the corresponding measures $\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j})$ approach $0 \times \ln(\alpha|\Xi|) = 0$ (where α is a constant), as $|\Xi| \to \infty$.
- the last probability $P_{\Xi}(\delta_i = 0, \delta_j = 0)$ approaches 1, so the measure $\mathbf{mit}_{\Xi}(t_i^{\delta_i=0}, t_j^{\delta_j=0})$ approaches $1 \times \ln 1 = 0$, as $|\Xi| \to \infty$.

Remark-8: It worth mentioning that the binary estimation method given in Eq.(1) and Eq.(2) rely on statistics $n_{\Xi}(t)$, $n_{\Xi}(t_i,t_j)$ and $|\Xi|$; it is thus sensitive to the sample size. A large sample size might overwhelm useful statistical information carried by those important terms having smaller statistics (or, concentrating in a few documents), thereby weaken and dilute the potential capability of EMIM and the MIT measures. $\Diamond$

The sample size is an important feature of any empirical study, and generally a larger sample size leads to increased precision when estimating unknown (probability distribution) parameters. According to study given in [19], an appropriate sample size for a qualitative research depends on a number of factors, including: the quality of the data, the scope of the study, the nature of the topic, the amount of useful information obtained from the participants (samples), the qualitative method, experimental design and settings, and so on. It seems not clear at present how to determine an appropriate sample size against a set of term pairs in practical applications. It would be helpful to consider appropriateness of the sample size prior to determining some probability estimation method for applying EMIM in a specific application.

CONCLUSION

This study examined the reasons for the failure of applying EMIM and highlighted some potential problems of applications. We attempted to clarify confusions caused by the problems and/or suggest solutions to the problems by analysing a various of properties of $I_{\Xi}(\delta_i;\delta_j)$ and $emim_{\Xi}(\delta_i;\delta_j)$. The key points of this study were emphasised and formally discussed through a series of remarks, some of them are listed as follows.

- The occurrence of term t in all samples does not provide any information about the occurrence of other terms in the samples; in order to effectively capture the dependence information of terms, we should always avoid many terms having $n_{\Xi}(t) = |\Xi|$.
- It can be seen, from the relation given in Eq.(6), that $I_{\Xi}(\delta_i;\delta_j) \geq 0$ cannot infer $emim_{\Xi}(\delta_i;\delta_j) \geq 0$; in fact, we have $emim_{\Xi}(\delta_i;\delta_j) \leq 0$ for two arbitrary terms $t_i, t_j \in V$.
- Two term pairs, (t_i, t_j) and (t'_i, t'_j), receiving the same EMIM value, $I_{\Xi}(\delta_i;\delta_j) = I_{\Xi}(\delta_{i'};\delta_{j'})$, may be dependent of each other in entirely different implications under the individual state values.
- A high positive value of $I_{\Xi}(\delta_i;\delta_j)$ may not be necessary to indicate that t_i and t_j are highly dependent for their occurrence; we should always verify the inequality given in Eq.(9) to ensure $\mathbf{mit}_{\Xi}(t_i, t_j) > 0$, and that terms are high dependent for their co-occurrence.
- In order to apply $emim(\delta_i;\delta_j)$ properly, we should always verify the inequality given in Eq.(10).
- The binary estimation method given in Eq.(1) and Eq.(2) is sensitive to the sample size; a large sample size might overwhelm useful statistical information carried by those terms concentrating in a small number of documents.

It is essential for this study to point out that different probability estimations may conclude to different properties of EMIM and the MIT measures, and therefore it is theoretically challenging to apply EMIM without clearly understanding the properties. A widely used binary estimation method is considered in this study as a good example to reveal practical application problems and to clarify our viewpoints. A more general discussion on this subject can be found in our another study [18]. Due to its generality, this study can be regarded as a useful tool for many areas of science, involving statistical text analysis and document processing.

REFERENCES

[1] A. Akadi, A. Abdeljalil El Ouardighi, and D. Aboutajdine, "A powerful feature selection approach based on mutual information," *International Journal of Computer Science and Network Security*, vol. 8, no. 4, pp. 116–121, 2008.

[2] M. Ait Kerroum, A. Hammouch, and D. Aboutajdine, "Textural feature selection by joint mutual information based on Gaussian mixture model for multispectral image classification," *Pattern Recognition Letters*, vol. 31, no. 10, pp. 1168–1174, 2010.

[3] H.-W. Liu, J.-G. Sun, L. Liu, and H.-J. Zhang, "Feature selection with dynamic mutual information," *Pattern Recognition*, vol. 42, pp. 1330–1339, 2009.

[4] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[5] G. Wang, F. Lochovsky, and Q. Yang, "Feature selection with conditional mutual information maximin in text categorization," in *Proceedings of the 10th International Conference on Information and Knowledge Management*, 2004, pp. 342–349.

[6] N. Vretos, V. Solachidis, and I. Pitas, "A mutual information based face clustering algorithm for movies," in *Proceedings of the 2006 IEEE International Conference on Multimedia and Expo (ICME'06)*, 2006, pp. 1013–1016.

[7] X. Zhang, K. Liu, Z. Liu, B. Duval, J. Richer, X. Zhao, J. Hao, and L. Chen, "Narromi: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference," *Bioinformatics*, vol. 29, no. 1, p. 106113, 2013.

[8] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, 1997.

[9] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Journal of the American Society for Information Science*, vol. 16, no. 1, pp. 22–29, 1990.

[10] H. Fang and C. X. Zhai, "Semantic term matching in axiomatic approaches to information retrieval," in *Proceedings of the 29th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 115–122.

[11] S. Gauch, J. Wang, and S. M. Rachakonda, "A corpus analysis approach for automatic query expansion and its extension to multiple databases," *ACM Transactions on Information Systems*, vol. 17, no. 3, pp. 250–269, 1999.

[12] M. Kim and K. Choi, "A comparison of collocation-based similarity measures in query expansion," *Information Processing & Management*, vol. 35, no. 1, pp. 19–30, 1999.

[13] R. Mandala, T. Tokunaga, and H. Tanaka, "Query expansion using heterogeneous thesauri," *Information Processing & Management*, vol. 36, no. 3, pp. 361–378, 2000.

[14] R. M. Losee, Jr., "Term dependence: A basis for Luhn and Zipf models," *Journal of the American Society for Information Science and Technology*, vol. 52, no. 12, pp. 1019–1025, 2001.

[15] C. J. van Rijsbergen, *Information Retrieval*, 2nd ed. London: Butterworths, 1979.

[16] ——, "A theoretical basis for the use of co-occurrence data in information retrieval," *Journal of Documentation*, vol. 33, no. 2, pp. 106–119, 1977.

[17] D. Cai and T. McCluskey, "A simple method for computing term mutual information," *Journal of Computing*, vol. 4, no. 6, pp. 1–6, 2012.

[18] ——, "A general framework of generating estimation functions for computing the mutual information of terms," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 11, pp. 198–207, 2013.

[19] J. Morse, "Determining sample size," *Qualitative Health Research*, vol. 10, no. 1, p. 35, 2000.

APPENDIX

Theorem 2.1 Suppose $P_{\Xi}(\delta)$ and $P_{\Xi}(\delta_i, \delta_j)$ are given in Eq.(1) and Eq.(2), respectively. Then $P_{\Xi}(\delta)$ and $P_{\Xi}(\delta_i, \delta_j)$ are

probability distributions on Ω and $\Omega \times \Omega$, respectively; $P_{\Xi}(\delta_i)$ and $P_{\Xi}(\delta_j)$ are the marginal distributions of $P_{\Xi}(\delta_i, \delta_j)$.

Proof. For arbitrary terms $t, t_i, t_j \in V$ (where $i \neq j$), using the statistics of the document frequencies concerning the set Ξ, it is easy to estimate the probability distributions.

First, notice that the (total) number of documents in the sample set is $|\Xi|$. Thus, the probability that t occurs in some sample is $\frac{n_{\Xi}(t)}{|\Xi|}$ as the number of samples in which t occurs is $n_{\Xi}(t)$, and thus the probability that t does not occur is $1 - \frac{n_{\Xi}(t)}{|\Xi|}$. Therefore, we can write a probability distribution, $P_{\Xi}(\delta)$, over Ω as expressed by Eq.(2).

Second, with the size of the sample set, the probability that t_i and t_j co-occur is $\frac{n_{\Xi}(t_i,t_j)}{|\Xi|}$ as the number of samples in which t_i and t_j co-occur is $n_{\Xi}(t_i, t_j)$; the probability that t_i occurs but t_j does not occur is $\frac{n_{\Xi}(t_i)-n_{\Xi}(t_i,t_j)}{|\Xi|}$ as the number of samples in which t_i occurs but t_j does not occur is $n_{\Xi}(t_i) - n_{\Xi}(t_i, t_j)$; similarly, the probability that t_i does not occur but t_j occurs is $\frac{n_{\Xi}(t_j)-n_{\Xi}(t_i,t_j)}{|\Xi|}$; the probability that neither of t_i nor t_j occur is $\frac{n_{\Xi}(\bar{t}_i,\bar{t}_j)}{|\Xi|}$, where $n_{\Xi}(\bar{t}_i, \bar{t}_j) = |\Xi| - n_{\Xi}(t_i) - n_{\Xi}(t_j) + n_{\Xi}(t_i, t_j)$ is the number of samples in which none of t_i and t_j occur. Therefore, we can write a probability distribution, $P_{\Xi}(\delta_i, \delta_j)$, over $\Omega \times \Omega$ as expressed by Eq.(3).

Finally, it is easy to see: $P_{\Xi}(\delta_i = 1) = \sum_{\delta_j=1,0} P_{\Xi}(\delta_i = 1, \delta_j) = \frac{n_{\Xi}(t_i)}{|\Xi|}$ and $P_{\Xi}(\delta_i = 0) = \sum_{\delta_j=1,0} P_{\Xi}(\delta_i = 0, \delta_j) = 1 - \frac{n_{\Xi}(t_i)}{|\Xi|}$. Hence, $P_{\Xi}(\delta_i)$ is the marginal distributions of $P_{\Xi}(\delta_i, \delta_j)$. A similar discussion may be given for $P_{\Xi}(\delta_j)$. □

An alternative way to derive $P_{\Xi}(\delta_i, \delta_j)$ is to use a conditional probability formula. The conditional probability of observing t_j occurs, given that t_i occurred, is $P_{\Xi}(\delta_j = 1|\delta_i = 1) = \frac{n_{\Xi}(t_i,t_j)}{n_{\Xi}(t_i)}$, since before the observation there were $n_{\Xi}(t_i)$ documents in Ξ, in which t_i occurred. The conditional probability of observing t_j does not occur, given that t_i occurred, is $P_{\Xi}(\delta_j = 0|\delta_i = 1) = 1 - \frac{n_{\Xi}(t_i,t_j)}{n_{\Xi}(t_i)}$, and similarly, we have $P_{\Xi}(\delta_i = 0|\delta_j = 1) = 1 - \frac{n_{\Xi}(t_i,t_j)}{n_{\Xi}(t_j)}$. Then, we can immediately write the expressions:

$$
\begin{aligned}
P_{\Xi}(\delta_i = 1, \delta_j = 1) &= P_{\Xi}(\delta_i = 1) P_{\Xi}(\delta_j = 1|\delta_i = 1) \\
&= \frac{n_{\Xi}(t_i)}{|\Xi|} \frac{n_{\Xi}(t_i, t_j)}{n_{\Xi}(t_i)} \\
&= \frac{n_{\Xi}(t_i, t_j)}{|\Xi|} \\
P_{\Xi}(\delta_i = 1, \delta_j = 0) &= P_{\Xi}(\delta_i = 1) P_{\Xi}(\delta_j = 0|\delta_i = 1) \\
&= \frac{n_{\Xi}(t_i)}{|\Xi|} \Big[1 - \frac{n_{\Xi}(t_i, t_j)}{n_{\Xi}(t_i)}\Big] \\
&= \frac{n_{\Xi}(t_i)}{|\Xi|} - \frac{n_{\Xi}(t_i, t_j)}{|\Xi|} \\
P_{\Xi}(\delta_i = 0, \delta_j = 1) &= P_{\Xi}(\delta_j = 1) P_{\Xi}(\delta_i = 0|\delta_j = 1) \\
&= \frac{n_{\Xi}(t_j)}{|\Xi|} \Big[1 - \frac{n_{\Xi}(t_i, t_j)}{n_{\Xi}(t_j)}\Big] \\
&= \frac{n_{\Xi}(t_j)}{|\Xi|} - \frac{n_{\Xi}(t_i, t_j)}{|\Xi|} \\
P_{\Xi}(\delta_i = 0, \delta_j = 0) &= 1 - P_{\Xi}(\delta_i = 1, \delta_j = 1) \\
&\quad - P_{\Xi}(\delta_i = 1, \delta_j = 0) \\
&\quad - P_{\Xi}(\delta_i = 0, \delta_j = 1) \\
&= 1 - \frac{n_{\Xi}(t_i)}{|\Xi|} - \frac{n_{\Xi}(t_i)}{|\Xi|} + \frac{n_{\Xi}(t_i, t_j)}{|\Xi|}.
\end{aligned}
$$

The results are in agreement with one given in Eq.(2).

It worth mentioning that the reason why we give the detailed proofs of Theorem 2.1 is to interpret mathematical meaning of the estimation of the probability distributions. The proof may be greatly simplified by directly using the nature of the expressions given in Eq.(1) and Eq.(2), that is,

$$P_{\Xi}(\delta) \geq 0 \quad \text{and} \quad P_{\Xi}(\delta_i, \delta_j) \geq 0$$

for $\delta, \delta_i, \delta_j = 1, 0$, and

$$\sum_{\delta=1,0} P_{\Xi}(\delta) = 1 \quad \text{and} \quad \sum_{\delta_i,\delta_j=1,0} P_{\Xi}(\delta_i, \delta_j) = 1$$

Therefore, $P_{\Xi}(\delta)$ and $P_{\Xi}(\delta_i, \delta_j)$ are probability distributions.

Theorem 2.2 Suppose $P_{\Xi}(\delta)$ and $P_{\Xi}(\delta_i, \delta_j)$ are given in Eq.(1) and Eq.(2), respectively. Then $P_{\Xi}(\delta_i, \delta_j)$ is absolutely continuous with respect to product $P_{\Xi}(\delta_i)P_{\Xi}(\delta_j)$, denoted by $P_{\Xi}(\delta_i, \delta_j) \ll P_{\Xi}(\delta_i)P_{\Xi}(\delta_j)$, for $\delta_i, \delta_j = 1, 0$.

Proof. For two arbitrary terms $t_i, t_j \in V$, according to whether $P_{\Xi}(t_i) = 1$ and/or $P_{\Xi}(t_j) = 1$, there are four cases to be considered, that is,

(C1) $0 < P_{\Xi}(t_i) < 1$ and $0 < P_{\Xi}(t_j) < 1$,
(C2) $P_{\Xi}(t_i) = 1$ but $0 < P_{\Xi}(t_j) < 1$,
(C3) $0 < P_{\Xi}(t_i) < 1$ but $P_{\Xi}(t_j) = 1$,
(C4) $P_{\Xi}(t_i) = 1$ and $P_{\Xi}(t_j) = 1$.

We first prove (C1) and then prove (C2). Similar proofs can be given to (C3) and (C4).

In order to prove (C1), let us further consider four cases:

(a) $t_i, t_j \in V_{\Xi}$;
(b) $t_i \in V_{\Xi}$ but $t_j \notin V_{\Xi}$;
(c) $t_i \notin V_{\Xi}$ but $t_j \in V_{\Xi}$;
(d) $t_i, t_j \notin V_{\Xi}$.

Notice that, for (a), $P_{\Xi}(\delta_i, \delta_j) \ll P_{\Xi}(\delta_i)P_{\Xi}(\delta_j)$ as $0 < P_{\Xi}(\delta_i), P_{\Xi}(\delta_j) < 1$ for $\delta_i, \delta_j = 0, 1$ by Eq.(1). We now prove (b), and similar proofs can be given for (c) and (d). The proof is to verify four distinct state values, respectively.

On one hand, when $t_i \in V_{\Xi}$ but $t_j \notin V_{\Xi}$, it has $0 < P_{\Xi}(t_i) < 1$, $P_{\Xi}(t_j) = 0$, and $P_{\Xi}(t_i, t_j) = 0$ by Eq.(1). Thus, by Eq.(3),

$$
\begin{aligned}
&P_{\Xi}(\delta_i = 1, \delta_j = 1) = 0 \\
&P_{\Xi}(\delta_i = 1, \delta_j = 0) = P_{\Xi}(t_i) > 0 \\
&P_{\Xi}(\delta_i = 0, \delta_j = 1) = 0 \\
&P_{\Xi}(\delta_i = 0, \delta_j = 0) = 1 - P_{\Xi}(t_i) > 0
\end{aligned}
$$

On the other hand, by Eq.(2), we have $0 < P_{\Xi}(\delta_i) < 1$ for $\delta_i = 1, 0$ when $t_i \in V_{\Xi}$; $P_{\Xi}(\delta_j = 1) = 0$ and $P_{\Xi}(\delta_j = 0) = 1$ when $t_j \notin V_{\Xi}$. Thus,

$$
\begin{aligned}
&P_{\Xi}(\delta_i = 1)P_{\Xi}(\delta_j = 1) = 0 \\
&P_{\Xi}(\delta_i = 1)P_{\Xi}(\delta_j = 0) = P_{\Xi}(\delta_i = 1) > 0 \\
&P_{\Xi}(\delta_i = 0)P_{\Xi}(\delta_j = 1) = 0 \\
&P_{\Xi}(\delta_i = 0)P_{\Xi}(\delta_j = 0) = P_{\Xi}(\delta_i = 0) > 0
\end{aligned}
$$

Therefore, $P_{\Xi}(\delta_i,\delta_j) \ll P_{\Xi}(\delta_i)P_{\Xi}(\delta_j)$ for $\delta_i,\delta_j = 1,0$.
In order to prove (C2), let us suppose we are given $t_i, t_j \in V_{\Xi}$ satisfying $n_{\Xi}(t_i) = |\Xi|$ and $n_{\Xi}(t_j) < |\Xi|$ (namely t_i occurs in all samples in Ξ, but t_j does not). In this case, it has $P_{\Xi}(t_i) = 1$ and $0 < P_{\Xi}(t_j) < 1$, and $n_{\Xi}(t_j) = n_{\Xi}(t_i,t_j)$. Thus,

(a) $P_{\Xi}(\delta_i = 1)\cdot P_{\Xi}(\delta_j = 1) > 0$ since $P_{\Xi}(\delta_i = 1) = 1$, and $0 < P_{\Xi}(\delta_j = 1) < 1$. Thus, $P_{\Xi}(\delta_i = 1, \delta_j = 1) \ll P_{\Xi}(\delta_i = 1)\cdot P_{\Xi}(\delta_j = 1)$ for $(\delta_i,\delta_j) = (1,1)$.

(b) $P_{\Xi}(\delta_i = 1)\cdot P_{\Xi}(\delta_j = 0) > 0$ since $P_{\Xi}(\delta_i = 1) = 1$ and $0 < P_{\Xi}(\delta_j = 0) < 1$. Thus, $P_{\Xi}(\delta_i = 1, \delta_j = 0) \ll P_{\Xi}(\delta_i = 1)\cdot P_{\Xi}(\delta_j = 0)$ for $(\delta_i,\delta_j) = (1,0)$.

(c) $P_{\Xi}(\delta_i = 0)\cdot P_{\Xi}(\delta_j = 1) = 0$ since $P_{\Xi}(\delta_i = 0) = 0$ and $0 < P_{\Xi}(\delta_j = 1) < 1$. Also, $P_{\Xi}(\delta_i = 0, \delta_j = 1) = \frac{1}{|\Xi|}\big[n_{\cdot 1} - n_{11}\big] = 0$. Thus, $P_{\Xi}(\delta_i = 0, \delta_j = 1) \ll P_{\Xi}(\delta_i = 0)\cdot P_{\Xi}(\delta_j = 1)$ for $(\delta_i,\delta_j) = (0,1)$.

(d) $P_{\Xi}(\delta_i = 0)\cdot P_{\Xi}(\delta_j = 0) = 0$ since $P_{\Xi}(\delta_i = 0) = 0$ and $0 < P_{\Xi}(\delta_j = 0) < 1$. Also, $P_{\Xi}(\delta_i = 0, \delta_j = 0) = \frac{1}{|\Xi|}\big[|\Xi| - n_{1\cdot} - n_{\cdot 1} + n_{11}\big] = \frac{1}{|\Xi|}\big[(|\Xi| - n_{1\cdot}) - (n_{\cdot 1} - n_{11})\big] = 0$. Thus, $P_{\Xi}(\delta_i = 0, \delta_j = 0) \ll P_{\Xi}(\delta_i = 0)\cdot P_{\Xi}(\delta_j = 0)$ for $(\delta_i,\delta_j) = (0,0)$.

Therefore, $P_{\Xi}(\delta_i,\delta_j) \ll P_{\Xi}(\delta_i)\cdot P_{\Xi}(\delta_j)$ for $\delta_i,\delta_j = 1,0$. □

Theorem 3.1 Suppose $P_{\Xi}(\delta)$ and $P_{\Xi}(\delta_i,\delta_j)$ are given in Eq.(1) and Eq.(2), respectively. Then $I_{\Xi}(\delta_i;\delta_j) = 0$ if $n_{\Xi}(t_i) = |\Xi|$ or $n_{\Xi}(t_j) = |\Xi|$.

Proof. We prove that each item of $I_{\Xi}(\delta_i;\delta_j)$ is zero for $n_{\Xi}(t_i) = |\Xi|$. A similar proof can be given to $n_{\Xi}(t_j) = |\Xi|$. Notice that, we have $n_{\Xi}(t_j) = n_{\Xi}(t_i,t_j)$, Thus,

1) for $(\delta_i,\delta_j) = (1,1)$, with $n_{11} = n_{\Xi}(t_i,t_j) = n_{\Xi}(t_j)$, it has

$$\frac{n_{11}}{|\Xi|}\ln\Big(\frac{n_{11}}{|\Xi|}\Big/\frac{n_{\Xi}(t_i)}{|\Xi|}\frac{n_{\Xi}(t_j)}{|\Xi|}\Big) = \frac{n_{11}}{|\Xi|}\ln\frac{n_{11}}{1\times n_{\Xi}(t_j)} = \frac{n_{\Xi}(t_i,t_j)}{|\Xi|}\ln 1 = 0$$

2) for $(\delta_i,\delta_j) = (1,0)$, with $n_{10} = n_{\Xi}(t_i) - n_{\Xi}(t_i,t_j) = |\Xi| - n_{\Xi}(t_j)$, it has

$$\frac{n_{10}}{|\Xi|}\ln\Big(\frac{n_{10}}{|\Xi|}\Big/\frac{n_{\Xi}(t_i)}{|\Xi|}\big(1-\frac{n_{\Xi}(t_j)}{|\Xi|}\big)\Big) = \frac{n_{10}}{|\Xi|}\ln\frac{n_{10}}{1\times\big(|\Xi| - n_{\Xi}(t_j)\big)} = \frac{n_{10}}{|\Xi|}\ln 1 = 0$$

3) for $(\delta_i,\delta_j) = (0,1)$, with $n_{01} = n_{\Xi}(t_j) - n_{\Xi}(t_i,t_j) = n_{\Xi}(t_j) - n_{\Xi}(t_j) = 0$, is has

$$\frac{n_{01}}{|\Xi|}\ln\Big(\frac{n_{01}}{|\Xi|}\Big/\big(1-\frac{n_{\Xi}(t_i)}{|\Xi|}\big)\frac{n_{\Xi}(t_j)}{|\Xi|}\Big) = \frac{0}{|\Xi|}\ln\frac{0}{0\times n_{\Xi}(t_j)} = 0\ln\frac{0}{0} = 0$$

4) for $(\delta_i,\delta_j) = (0,0)$, with $n_{00} = |\Xi| - n_{\Xi}(t_i) - n_{\Xi}(t_j) + n_{\Xi}(t_i,t_j) = |\Xi| - |\Xi| - n_{\Xi}(t_j) + n_{\Xi}(t_j) = 0$, it has

$$\frac{n_{00}}{|\Xi|}\ln\Big(\frac{n_{00}}{|\Xi|}\Big/\big(1-\frac{n_{\Xi}(t_i)}{|\Xi|}\big)\big(1-\frac{n_{\Xi}(t_j)}{|\Xi|}\big)\Big) = \frac{0}{|\Xi|}\ln\frac{0}{0\times\big(|\Xi| - n_{\Xi}(t_j)\big)} = 0\ln\frac{0}{0} = 0$$

The proof is completed. □

Theorem 3.2 Suppose $I_{\Xi}(\delta_i,\delta_j)$ and $emim_{\Xi}(\delta_i,\delta_j)$ are given in Eq.(3) and Eq.(5), respectively. Then

$$I_{\Xi}(\delta_i,\delta_j) = \frac{1}{n}\times emim_{\Xi}(\delta_i,\delta_j) + \ln(n)$$

where $n = |\Xi|$.

Proof. With the above notation n_{11}, $n_{1\cdot}$ and $n_{\cdot 1}$ given in Eq.(5), we can write an alternative, but fully equivalent, expression:

$$\begin{aligned}
I_{\Xi}(\delta_i;\delta_j) &= \frac{n_{11}}{n}\ln\Big(\frac{n_{11}}{n_{1\cdot}n_{\cdot 1}}n\Big) + \frac{n_{1\cdot}-n_{11}}{n}\ln\Big(\frac{n_{1\cdot}-n_{11}}{n_{1\cdot}(n-n_{\cdot 1})}n\Big) + \frac{n_{\cdot 1}-n_{11}}{n}\ln\Big(\frac{n_{\cdot 1}-n_{11}}{(n-n_{1\cdot})n_{11}}n\Big) + \frac{n-n_{1\cdot}-n_{\cdot 1}+n_{11}}{n}\times \ln\Big(\frac{n-n_{1\cdot}-n_{\cdot 1}+n_{11}}{(n-n_{1\cdot})(n-n_{\cdot 1})}n\Big)\\
&= \Big[\frac{n_{11}}{n}\ln\frac{n_{11}}{n_{1\cdot}n_{\cdot 1}} + \frac{n_{1\cdot}-n_{11}}{n}\ln\frac{n_{1\cdot}-n_{11}}{n_{1\cdot}(n-n_{\cdot 1})} + \frac{n_{\cdot 1}-n_{11}}{n}\ln\frac{n_{\cdot 1}-n_{11}}{(n-n_{\cdot 1})n_{\cdot 1}} + \frac{n-n_{1\cdot}-n_{\cdot 1}+n_{11}}{n}\times \ln\frac{n-n_{1\cdot}-n_{\cdot 1}+n_{11}}{(n-n_{1\cdot})(n-n_{\cdot 1})}\Big]\\
&\quad + \Big[\frac{n_{11}}{n} + \frac{n_{1\cdot}-n_{11}}{n} + \frac{n_{\cdot 1}-n_{11}}{n} + \frac{n-n_{1\cdot}-n_{\cdot 1}+n_{11}}{n}\Big]\times\ln(n)\\
&= \Big[n_{11}\ln\frac{n_{11}}{n_{1\cdot}n_{\cdot 1}} + (n_{1\cdot}-n_{11})\ln\frac{n_{1\cdot}-n_{11}}{n_{\Xi}(t_i)(n-n_{\cdot 1})} + (n_{\cdot 1}-n_{11})\ln\frac{n_{\cdot 1}-n_{11}}{(n-n_{1\cdot})n_{\cdot 1}} + (n-n_{1\cdot}-n_{\cdot 1}+n_{11})\times \ln\frac{n-n_{1\cdot}-n_{\cdot 1}+n_{11}}{(n-n_{1\cdot})(n-n_{\cdot 1})}\Big]\times\frac{1}{n} + \ln(n)\\
&= emim_{\Xi}(\delta_i;\delta_j)\times\frac{1}{n} + \ln(n)
\end{aligned}$$

The proof is completed. □

Theorem 3.3 Suppose $emim_{\Xi}(\delta_i,\delta_j)$ is given expression Eq.(5). Then $emim_{\Xi}(\delta_i,\delta_j) \le 0$.

Proof. We prove each item of $emim_{\Xi}(\delta_i;\delta_j)$ non-positive. The proof is simple with an inequality $\frac{a}{a_1\cdot a_2} \le 1$ if $a \le a_1$ and $a \le a_2$.

1) we have $\frac{n_{11}}{n_{1\cdot}n_{\cdot 1}} \le 1$ since,

$$n_{11} = n_{\Xi}(t_i,t_j) \le n_{\Xi}(t_i) = n_{1\cdot}$$
$$n_{11} = n_{\Xi}(t_i,t_j) \le n_{\Xi}(t_j) = n_{\cdot 1}$$

2) we have $\frac{n_{10}}{n_{1\cdot}\, n_{\cdot 0}} \leq 1$ since,

$$\begin{aligned} n_{10} &= n_{\Xi}(t_i) - n_{\Xi}(t_i, t_j) \leq n_{\Xi}(t_i) = n_{1\cdot} \\ n_{10} &= n_{\Xi}(t_i) - n_{\Xi}(t_i, t_j) \leq |\Xi| - n_{\Xi}(t_i, t_j) \\ &\leq |\Xi| - n_{\Xi}(t_j) = n_{\cdot 0} \end{aligned}$$

3) the proof is similar to 2).
4) we have $\frac{n_{00}}{n_{0\cdot}\, n_{\cdot 0}} \leq 1$ since,

$$\begin{aligned} n_{00} &= |\Xi| - n_{\Xi}(t_i) - n_{\Xi}(t_j) + n_{\Xi}(t_i, t_j) \\ &= |\Xi| - n_{\Xi}(t_i) - \big[n_{\Xi}(t_j) - n_{\Xi}(t_i, t_j)\big] \\ &\leq |\Xi| - n_{\Xi}(t_i) = n_{0\cdot} \\ n_{00} &= |\Xi| - n_{\Xi}(t_j) - \big[n_{\Xi}(t_i) - n_{\Xi}(t_i, t_j)\big] \\ &\leq |\Xi| - n_{\Xi}(t_j) = n_{\cdot 0} \end{aligned}$$

The proof is completed. □

Note that the fact that the individual items of $emim_{\Xi}(\delta_i, \delta_j)$ are non-positive can also be seen directly by the relations:

$$\begin{aligned} n_{1\cdot} &= n_{11} + n_{10}, \quad n_{0\cdot} = n_{01} + n_{00}, \\ n_{\cdot 1} &= n_{11} + n_{01}, \quad n_{\cdot 0} = n_{10} + n_{00}. \end{aligned}$$

Theorem 4.1 Suppose the four measures, $\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j})$, where $\delta_i \delta_j = 0, 1$, are given in Eq.(7). Then we have the following property.
(1) If $\frac{n_{\Xi}(t_i,t_j)}{|\Xi|} = \frac{n_{\Xi}(t_i)}{|\Xi|}\frac{n_{\Xi}(t_j)}{|\Xi|}$ then

$$\begin{aligned} &\mathbf{mit}_{\Xi}(t_i, t_j) = 0, \quad \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) = 0, \\ &\mathbf{mit}_{\Xi}(\bar{t}_i, t_j) = 0, \quad \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) = 0. \end{aligned}$$

(2) If $\frac{n_{\Xi}(t_i,t_j)}{|\Xi|} > \frac{n_{\Xi}(t_i)}{|\Xi|}\frac{n_{\Xi}(t_j)}{|\Xi|}$ then

$$\begin{aligned} &\mathbf{mit}_{\Xi}(t_i, t_j) > 0, \quad \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) \leq 0, \\ &\mathbf{mit}_{\Xi}(\bar{t}_i, t_j) \leq 0, \quad \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) > 0. \end{aligned}$$

(3) If $\frac{n_{\Xi}(t_i,t_j)}{|\Xi|} < \frac{n_{\Xi}(t_i)}{|\Xi|}\frac{n_{\Xi}(t_j)}{|\Xi|}$ then

$$\begin{aligned} &\mathbf{mit}_{\Xi}(t_i, t_j) < 0, \quad \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) \geq 0, \\ &\mathbf{mit}_{\Xi}(\bar{t}_i, t_j) \geq 0, \quad \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) < 0. \end{aligned}$$

Proof. The proof of (1) is obvious. We only prove (2) here. A similar proof can be given to (3).

Now, substituting $P_{\Xi}(\delta)$ and $P_{\Xi}(\delta_i, \delta_j)$ in Eq.(1) and Eq.(2) into $\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j})$ in Eq.(7), we can rewrite the four MIT measures as follows (also see Example 4.1):

$$\begin{aligned} \mathbf{mit}_{\Xi}(t_i, t_j) &= P_{\Xi}(t_i, t_j) \ln \frac{P_{\Xi}(t_i, t_j)}{P_{\Xi}(t_i) P_{\Xi}(t_j)} \\ \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) &= \big(P_{\Xi}(t_i) - P_{\Xi}(t_i, t_j)\big) \ln \frac{P_{\Xi}(t_i) - P_{\Xi}(t_i, t_j)}{P_{\Xi}(t_i)\big(1 - P_{\Xi}(t_j)\big)} \\ \mathbf{mit}_{\Xi}(\bar{t}_i, t_j) &= \big(P_{\Xi}(t_j) - P_{\Xi}(t_i, t_j)\big) \ln \frac{P_{\Xi}(t_j) - P_{\Xi}(t_i, t_j)}{\big(1 - P_{\Xi}(t_i)\big) P_{\Xi}(t_j)} \\ \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) &= \big(1 - P_{\Xi}(t_i) - P_{\Xi}(t_j) + P_{\Xi}(t_i, t_j)\big) \times \\ &\qquad \ln \frac{1 - P_{\Xi}(t_i) - P_{\Xi}(t_j) + P_{\Xi}(t_i, t_j)}{\big(1 - P_{\Xi}(t_i)\big)\big(1 - P_{\Xi}(t_j)\big)} \end{aligned}$$

Thus, on one hand, from $\frac{n_{\Xi}(t_i,t_j)}{|\Xi|} > \frac{n_{\Xi}(t_i)}{|\Xi|}\frac{n_{\Xi}(t_j)}{|\Xi|}$, we have

$$\begin{aligned} P_{\Xi}(t_i, t_j) &> P_{\Xi}(t_i) P_{\Xi}(t_j) \\ P_{\Xi}(t_i) - P_{\Xi}(t_i, t_j) &< P_{\Xi}(t_i) - P_{\Xi}(t_i) P_{\Xi}(t_j) \\ &= P_{\Xi}(t_i)\big(1 - P_{\Xi}(t_j)\big) \\ P_{\Xi}(t_j) - P_{\Xi}(t_i, t_j) &< P_{\Xi}(t_j) - P_{\Xi}(t_i) P_{\Xi}(t_j) \\ &= P_{\Xi}(t_j)\big(1 - P_{\Xi}(t_i)\big) \\ 1 - P_{\Xi}(t_i) - P_{\Xi}(t_j) + P_{\Xi}(t_i, t_j) &> 1 - P_{\Xi}(t_i) - P_{\Xi}(t_j) + P_{\Xi}(t_i) P_{\Xi}(t_j) \\ &= \big(1 - P_{\Xi}(t_i)\big)\big(1 - P_{\Xi}(t_j)\big) \end{aligned}$$

which are equivalent respectively to

$$\begin{aligned} &\frac{P_{\Xi}(t_i, t_j)}{P_{\Xi}(t_i) P_{\Xi}(t_j)} > 1 \\ &\frac{P_{\Xi}(t_i) - P_{\Xi}(t_i, t_j)}{P_{\Xi}(t_i)\big(1 - P_{\Xi}(t_j)\big)} < 1 \\ &\frac{P_{\Xi}(t_j) - P_{\Xi}(t_i, t_j)}{P_{\Xi}(t_j)\big(1 - P_{\Xi}(t_i)\big)} < 1 \\ &\frac{1 - P_{\Xi}(t_i) - P_{\Xi}(t_j) + P_{\Xi}(t_i, t_j)}{\big(1 - P_{\Xi}(t_i)\big)\big(1 - P_{\Xi}(t_j)\big)} > 1 \end{aligned}$$

then we obtain

$$\begin{aligned} &\ln \frac{P_{\Xi}(t_i, t_j)}{P_{\Xi}(t_i) P_{\Xi}(t_j)} > 0 \\ &\ln \frac{P_{\Xi}(t_i) - P_{\Xi}(t_i, t_j)}{P_{\Xi}(t_i)\big(1 - P_{\Xi}(t_j)\big)} < 0 \\ &\ln \frac{P_{\Xi}(t_j) - P_{\Xi}(t_i, t_j)}{P_{\Xi}(t_j)\big(1 - P_{\Xi}(t_i)\big)} < 0 \\ &\ln \frac{1 - P_{\Xi}(t_i) - P_{\Xi}(t_j) + P_{\Xi}(t_i, t_j)}{\big(1 - P_{\Xi}(t_i)\big)\big(1 - P_{\Xi}(t_j)\big)} > 0 \end{aligned}$$

On the other hand, for $t, t_i, t_j \in V_{\Xi}$, from

$$\begin{aligned} &0 < P_{\Xi}(t) = \frac{n_{\Xi}(t)}{|\Xi|} \leq 1 \\ &0 \leq 1 - P_{\Xi}(t) < 1 \\ &P_{\Xi}(t_i) = \frac{n_{\Xi}(t_i)}{|\Xi|} \geq \frac{n_{\Xi}(t_i, t_j)}{|\Xi|} = P_{\Xi}(t_i, t_j) \\ &P_{\Xi}(t_j) = \frac{n_{\Xi}(t_j)}{|\Xi|} \geq \frac{n_{\Xi}(t_i, t_j)}{|\Xi|} = P_{\Xi}(t_i, t_j) \end{aligned}$$

we obtain

$$\begin{aligned} P_{\Xi}(t_i, t_j) &> P_{\Xi}(t_i) P_{\Xi}(t_j) > 0 \\ P_{\Xi}(t_i) - P_{\Xi}(t_i, t_j) &\geq 0 \\ P_{\Xi}(t_j) - P_{\Xi}(t_i, t_j) &\geq 0 \\ 1 - P_{\Xi}(t_i) - P_{\Xi}(t_j) + P_{\Xi}(t_i, t_j) &> \big(1 - P_{\Xi}(t_i)\big)\big(1 - P_{\Xi}(t_j)\big) \geq 0 \end{aligned}$$

Hence, from the above four rewritten MIT measures, we can see that the four inequalities in (2) hold. □