Characterizing Pure High-order Entanglements in Lexical Semantic Spaces via Information Geometry

Yuexian Hou[1] , Dawei Song[2]

[1] School of Computer Sci. & Tech., Tianjin University, China
yxhou@tju.edu.cn
[2] School of Computing, The Robert Gordon University, United Kingdom.
d.song@rgu.ac.uk

Abstract. An emerging topic in Quantuam Interaction is the use of lexical semantic spaces, as Hilbert spaces, to capture the meaning of words. There has been some initial evidence that the phenomenon of quantum entanglement exists in a semantic space and can potentially play a crucial role in determining the embeded semantics. In this paper, we propose to consider pure high-order entanglements that cannot be reduced to the compositional effect of lower-order ones, as an indicator of high-level semantic entities. To characterize the intrinsic order of entanglements and distinguish pure high-order entanglements from lower-order ones, we develop a set of methods in the framework of Information Geometry. Based on the developed methods, we propose an expanded vector space model that involves context-sensitive high-order information and aims at characterizing high-level retrieval contexts. Some initial ideas on applying the proposed methods in query expansion and text classification are also presented.

Keywords: Information geometry, Pure high-order entanglement, Semantic emergence, Extended vector model

1 Introduction

An emerging line of research in Quantum Interaction (QI) is on capturing the meaning of words based on lexical semantic spaces (as Hilbert spaces) [13][14][17]. The intuition is that humans encountering a new concept often derive its meaning via the accumulative experience of contexts in which the concept appears. Therefore, the meaning of a word can be captured by examining its co-occurrence patterns with other words in the language use (e.g., a corpus of texts). A typical semantic space model is the Hyperspace Analogue to Language (HAL) [15]. The semantic space models have demonstrated a cognitive compatibility with human information processing [15][16].

More formally, in this paper, we generalize a semantic space to a Hilbert space induced by a set of words, in which all possible combinations of the words form the basis vectors. For example, given a word set W≡{Napoleon, invasion, Spain}, we have eight basis vectors $|000\rangle,|001\rangle,\ldots,|111\rangle$, where the basis vector $|001\rangle$ stands for the occurrence of 'Napoleon' and the absence of 'invasion' and 'Spain'. A pure state of this semantic space can be written as $a_{000}|000\rangle + \ldots + a_{111}|111\rangle$, where the linear combination coefficients a_{ijk} meet the normalization condition $|a_{000}|^2 + \ldots + |a_{111}|^2 = 1$. In quantum mechanics, the squared norm of a linear combination coefficient is considered as the probability of the corresponding basis event observed. According to this interpretation, it is clear that we can readily recover the marginal probability of any word occurrence from a given pure state of the semantic space. For example, the marginal probability of the occurrence of the first word (Napoleon) can be given by $|a_{100}|^2 + |a_{101}|^2 + |a_{110}|^2 + |a_{111}|^2$. In this sense, a pure state of the semantic space gives a more comprehensive description of the space than the conventional Vector Space Model (VSM or VM for short) [1].

A kind of quantum states of particular importance is the entangled state [2], in which the quantum states of two or more objects are dependent on each other so that one object can no longer be adequately described without a full mention of its counterparts. Technically, several objects are entangled if the state of the compositional system cannot be expressed as the tensor of individual systems' states. Recent study by Bruza el. al. revealed some initial evidence that the phenomenon of entanglement also exist in semantic spaces [12]. Then, the next fundamental research question arise: how to characterize and utilize the entanglements in semantic spaces? This is the very aim of this paper. We are particularly interested in the pure high-order[1] entanglements in semantic spaces, i.e., high-order entanglements that cannot be reduced to the compositional effect of lower-order interactions, which often indicate the emergence of high-level semantic entities.

For illustration, let us consider the example semantic space shown earlier. Given a pure state of this semantic space[2],

$$\begin{aligned}|\psi\rangle = &\sqrt{0.3296}|000\rangle + \sqrt{0.0002}|001\rangle + \sqrt{0.0900}|010\rangle + \sqrt{0.0001}|011\rangle \\ &+ \sqrt{0.3000}|100\rangle + \sqrt{0.0001}|101\rangle + \sqrt{0.2000}|110\rangle + \sqrt{0.0800}|111\rangle\end{aligned}$$

it is easy to check that $|\psi\rangle$ cannot be expressed as the tensor of the pure state of its 1-order subsystems, i.e.,

$$|\psi\rangle \neq (x_0|0\rangle + x_1|1\rangle) \otimes (y_0|0\rangle + y_1|1\rangle) \otimes (z_0|0\rangle + z_1|1\rangle)$$

for arbitrary x_0, x_1, y_0, y_1, z_0 and z_1 meeting $|x_0|^2+|x_1|^2=1$, $|y_0|^2+|y_1|^2=1$ and $|z_0|^2+|z_1|^2=1$. Hence, we conclude that $|\psi\rangle$ is an entangled state.

In this paper, we focus on the pure high-order entanglement, i.e., the high-order entanglement that cannot be expressed as the tensor of any lower-order systems that might be entangled too. For example, it is easy to check that the above $|\psi\rangle$ cannot be expressed as the tensor of state vectors of any two subsystems, e.g.,

$$|\psi\rangle = (u_0|0\rangle + u_1|1\rangle) \otimes (v_{00}|00\rangle + v_{01}|01\rangle + v_{10}|10\rangle + v_{11}|11\rangle),$$

where $|u_0|^2+|u_1|^2=1$, $|v_{00}|^2+|v_{01}|^2+|v_{10}|^2+|v_{11}|^2=1$. In this case, we conclude that $|\psi\rangle$ has a pure 3-order entanglement.

---

1 In this paper, the "high-order entanglement" corresponds to the "multipartite entanglement" in Quantum Mechanics.

2 Note that the coefficients of $|\psi\rangle$ meet the normalization condition, i.e., 0.3296+0.0002+0.0900+0.0001+0.3000+0.0001+0.2000+0.0800=1

The purpose of this paper is to characterize pure high-order entanglements and investigate their semantic implications in semantic spaces. To this end, there are two fundamental issues. One is how to measure entanglements, and the other is how to distinguish pure high-order entanglements from the compositional effect of lower-order entanglements so that we can illuminate the semantic implication of pure high-order entanglements by a computational method.

For the first issue, there are several well-known statistics measuring 2-order entanglement of a pure state, e.g., Von Neumann entropy [2], relative entropy of entanglement [3], robustness of entanglement [3] and squashed entanglement [3]. The measurement of high-order entanglement is more complicated. Some measures on high-order entanglement are derived by a direct generalization or a simple combination of 2-order measures, e.g., relative entropy of entanglement [3], robustness of entanglement [3] and global entanglement [3]. In addition, there are high-order entanglement measures that do not inherently depend on 2-order measures, e.g., Tangle [3] and Schmidt measure [3]. Although they are useful in general, most of the above statistics have some limitations in certain contexts. For example, many of these statistics cannot effectively distinguish pure high-order entanglements from the lower-order ones. Although, based on the above statistics, a rather satisfactory understanding has been achieved in the bipartite case, there is a certain degree of consensus that there is no universal way to define pure high-order entanglement, even in the simplest case of pure states [18, 19]. The existing pure high-order entanglement statistic often has to depend on some strong presupposition, e.g., symmetric Gaussianity [20].

The second issue requires a method which can not only measure pure high-order entanglements but also easily find or construct surrogate states so that we can investigate their semantic implications exclusively. Here, a surrogate state refers to the state that shares the same (k-i)-order entanglements, where $0<i<k$, with the original state but does not have pure k-order entanglement. Hence, by comparing the manifestation of the original state and the surrogate state in a proper context, e.g., information retrieval, we can evaluate the semantic implication of the pure k-order entanglement. In our opinion, the pure k-order entanglements are an important indicator of specific semantic entities.

In this paper, we propose the use of Information Geometry (IG) [4][5] to characterize the pure high-order entanglements. IG provides useful tools and concepts for this purpose, including the orthogonality of coordinate parameters and the Pythagoras relation in the KL-divergence [6][7]. For example, based on parametric orthogonality, we can give a set of statistics and methods for analyzing word occurrence patterns by decomposing the word entanglements into various orders. As a result, pure 2-order, 3-order, and higher-order entanglements are singled out.

It should be emphasized that, owing to the lack of a proper quantum statistic, the proposed IG method in this paper is classical in itself. The usefulness of IG method in a quantum framework roots on the following observation: In a post-measurement configuration, the entanglement degenerates into the statistical dependence between the measurement results. Specifically, it can be shown that several objects are entangled only if the corresponding random variables denoting the measurement results of these objects are statistically dependent on each other (see Subsection 2.1 for details). Since the occurrence and co-occurrence patterns of words can be naturally explained as the measurement results of semantic spaces, we believe that the proposed IG method is sufficient for our purpose.

2 Preliminaries of Information Geometry

Information Geometry (IG) represents probabilistic distributions as parametric coordinate systems, and hence could establish a connection between the properties of statistical distributions and some well-known notions in differential geometry to capture statistical dependencies from a geometric point of view. In this section, we will first discuss the connections between quantum entanglement and Information Geometry (IG), and then give a brief introduction to some relevant concepts and theorems from IG. Note that most theorems presented in Subsections 2.2, 2.3 and 2.4 have been formally proved or implied by the pioneering work in IG, e.g., the work by Rao [8], Jeffreys [9] and Sun-Ichi Amari [4][5]. Here, we restate and interpret them for our purpose in the context of semantic spaces, as their original expressions are heavily dependent on the notions and symbols of differential geometry and are thus not easy to follow for readers without a strong mathematical background.

2.1 On the Connection of Quantum Entanglements and Classical Dependences

Although IG is expressed in a classical framework of probability theory and originally aims at characterizing classical interactions[3], it can be naturally applied in the quantum framework because of the intrinsic connection between quantum entanglements and statistical dependences. For illustration, let $|\psi\rangle$ be a pure state of a two-qubit system A. Then $|\psi\rangle$ can determine a joint distribution on the basis events of A. For instance, if $|\psi\rangle = a_{00}|00\rangle + a_{01}|01\rangle + a_{10}|10\rangle + a_{11}|11\rangle$, then $|\psi\rangle$ determines a joint distribution: $P_{|\psi\rangle} = \left\{p_{00} = |a_{00}|^2, p_{01} = |a_{01}|^2, p_{10} = |a_{10}|^2, p_{11} = |a_{11}|^2\right\}$. Let $X_{|\psi>}$ be the (classical) random variable obeying the joint distribution $P_{|\psi>}$. We call $X_{|\psi>}$ the denotative random variable induced from $|\psi\rangle$, and denote the value of $X_{|\psi>}$ by x_ψ. For example, if $|\psi\rangle = |10\rangle$, then x_ψ=10. The following proposition, which can be generalized to general cases of multi-compositional systems, illuminates the equivalence between entanglements and statistical dependences in the post-measurement configuration.

Proposition 1: Let $|\psi\rangle$ be a pure state of a quantum system A, $\{B,C\}$ be a bipartition of A such that $A = B \otimes C$, and $|u\rangle$ and $|v\rangle$ be the pure states of B and C respectively. Then, $|\psi\rangle = |u\rangle \otimes |v\rangle$ *iff* $\Pr\left(X_{|\psi\rangle} = x_u \circ x_v\right) = \Pr\left(X_{|u\rangle} = x_u\right) \cdot \Pr\left(X_{|v\rangle} = x_v\right)$ where $X_{|\psi\rangle}$, $X_{|u\rangle}$ and $X_{|v\rangle}$ are denotative random variables induced from $|\psi\rangle$, $|u\rangle$

[3] In this paper, we use the term 'interaction' or 'dependence' to be the classical counterpart of the quantum entanglement. The connection between these notions is shown in Proposition 1.

and $|v\rangle$ respectively, and $\circ$ stands for the conjunction of x_u and x_v, e.g., if $x_u = 01, x_v = 10$, then $x_u \circ x_v = 0110$.

Proof: Let $|\psi\rangle = a_{0...0}|0...0\rangle + ... + a_{1...1}|1...1\rangle$ is a state vector of 2^n-dimensional Hilbert space A, $|u\rangle = b_{0...0}|0...0\rangle + ... + b_{1...1}|1...1\rangle$ is a state vector of $A's$ 2^k-dimensional subspace B and $v = c_{0...0}|0...0\rangle + ... + c_{1...1}|1...1\rangle$ is a state vector of $A's$ 2^l-dimensional subspace $A - B$, where $n = k + l$.

If $|\psi\rangle = |u\rangle \otimes |v\rangle$, i.e.,

$$a_{0...0}|0...0\rangle + ... + a_{1...1}|1...1\rangle = \left(b_{0...0}|0...0\rangle + ... + b_{1...1}|1...1\rangle\right) \otimes \left(c_{0...0}|0...0\rangle + ... + c_{1...1}|1...1\rangle\right)$$

it turns out that $a_{x_1...x_n} = b_{x_1...x_k} \cdot c_{x_{k+1}...x_n}$ for any $x_1,...,x_n \in \{0,1\}$, i.e., the probability of a basis event $|x_1,...,x_n\rangle$ is equal to the product of probabilities of corresponding basis events in subsystems. Sufficiency follows directly from this observation.

Assumes that denotative random variables induced by $|\psi\rangle$, $|u\rangle$ and $|v\rangle$ satisfy $\Pr\left(X_{|\psi\rangle} = x_u \circ x_v\right) = \Pr\left(X_{|u\rangle} = x_u\right) \cdot \Pr\left(X_{|v\rangle} = x_v\right)$. Based on the observation that $\Pr\left(X_{|\psi\rangle} = x_u \circ x_v\right) = \left|a_{x_u \circ x_v}\right|^2$, $\Pr\left(X_{|u\rangle} = x_u\right) = \left|b_{x_u}\right|^2$ and $\Pr\left(X_{|v\rangle} = x_v\right) = \left|c_{x_v}\right|^2$, it is easy to check the necessity.

The main tenet of IG is that many important structures in probability theory and statistics can be treated as structures in differential geometry by regarding a space of probabilities as a differential manifold endowed with a Riemannian metric and a family of affine connections [4]. In particular, IG provides a novel method to characterize pure high-order interactions among random variables. According to Proposition 1, IG is relevant to the task of entanglement identification in the post-measurement configuration. Note that most current applications of semantic spaces are essentially in the post-measurement configuration. Hence we can directly investigate the entanglement in semantic spaces using IG.

2.2 Statistical Manifold and Orthogonality

We represent a co-occurrence pattern of words by a random vector with binary components so that the joint distribution of co-occurrence can be exactly expanded by a log-linear model [10]. Let $\mathbf{X} \equiv [X_1, X_2, \ldots, X_n]^T$, $X_i \in \{0,1\}$ be a $n \times 1$ random vector and let $p \equiv p(\mathbf{x})$, $x \equiv [x_1, x_2, \ldots, x_n]^T$, $\mathbf{x}_i \in \{0,1\}$ be its joint probability distribution. Each X_i indicates that the i^{th} word is present ($X_i = 1$) or absent ($X_i = 0$).

Each distribution $p(\mathbf{x})$ is defined by 2^n probabilities:

$$p_{i_1...i_n} \equiv \Pr\{X_1 = i_1, \ldots, X_n = i_n\} > 0, i_k \in \{0,1\}, 1 \le k \le n, \sum_{i_1,\ldots,i_n} p_{i_1...i_n} = 1$$

Hence, the set of all distributions forms a (2^n-1)-dimensional manifold $\mathbf{S}_n$, where the subscript n of $\mathbf{S}$ denotes the number of random variables. Note that we require $p(\mathbf{x})>0$ for all $\mathbf{x}$ since the case of a various support set[4] of $p(\mathbf{x})$ poses rather significant

4 In mathematics, the support of a function is the set of points where the function is not zero, or the closure of that set. Here, the support set refers to the set of terms with nonzero probabilities.

difficulties for analysis. This requirement can be met by any common statistical smoothing method, e.g., Good-Turing estimator [11]. A direct coordinate system of $\mathbf{S}_n$ can be constructed by any 2^n-1 terms among p(**x**). We refer to this coordinate system as *p-coordinates*.

Another coordinate system of $\mathbf{S}_n$ is given by the expectation parameters:

$$\eta_i = E[x_i], i = 1,\ldots,n; \eta_{ij} = E[x_i x_j], i < j; \cdots; \eta_{12\cdots n} = E[x_1 \cdots x_n] \quad \textbf{(1)}$$

which have also 2^n-1 components. This coordinate system is called *η-coordinates*.

On the other hand, p(**x**) can be expanded by

$$\log p(\mathbf{x}) = \sum_i \theta_i x_i + \sum_{i<j} \theta_{ij} x_i x_j + \cdots + \theta_{1\cdots n} x_1 \cdots x_n - \psi \quad \textbf{(2)}$$

where ψ is the normalization term corresponding to ψ≡logp(**0**). It is easy to check that the formula (2) is an exact expansion since all x_is are binary. In addition, if **x**=$[0,\ldots,0]^T$, we have logp(**x**)=logp(**0**). All θ_{ijk}s together have 2^n-1 components and form the so-called *θ-coordinates*.

To characterize pure high-order interactions, we first introduce Riemannian metric tensor which is derived from the Fisher information and orthogonality. We will first give their mathematical definitions in general and then illuminate their meaning in a specific context.

Definition 1 (**Fisher Information** and **Riemannian metric tensor**): Given a probability distributions p(**x**;ξ) parameterized by $\xi \equiv [\xi_1,\ldots,\xi_n]^T \in \Xi$, the **Fisher information** of two coordinate parameters ξ_i and ξ_j is defined by

$$g_{ij}(\xi) \equiv E\left[(\partial/\partial\xi_i) l(\mathbf{x};\xi) \cdot (\partial/\partial\xi_j) l(\mathbf{x};\xi)\right] \quad \textbf{(3)}$$

where l(**x**;ξ)≡logp(**x**;ξ) and E[•] denotes the expectation with respect to p(**x**;ξ). If Fisher information matrix $G(\xi) \equiv (g_{ij}(\xi))$ is nondegenerate for any $\xi \in \Xi$, the parameterized family S≡{p(**x**;ξ)} is a Riemannian manifold, and G(ξ) is a **Riemannian metric tensor**.

Definition 2 (**Orthogonality**)**:** Two coordinate parameters ξ_i and ξ_j are **orthogonal** if the Fisher information of ξ_i and ξ_j vanishes for any $\xi \in \Xi$, i.e.,

$$E\left[(\partial/\partial\xi_i) l(\mathbf{x};\xi) \cdot (\partial/\partial\xi_i) l(\mathbf{x};\xi)\right] = 0 \quad \textbf{(4)}$$

We explain the meaning of Definition 2 by a 3-word example. Using three binary variables X_1, X_2 and X_3 to denote the occurrence of the word w_1, w_2 and w_3 respectively, the joint distribution of X_1, X_2 and X_3 is given by p(**x**)≡p_{ijk}=Pr{x_1=i, x_2=j, x_3=k}>0, i, j, k∈{0,1}, where **x**=$[x_1, x_2, x_3]^T$. It is clear that we need seven free parameters to characterize a distribution because of the constraint $\sum_{i,jk} p_{ijk}=1$. Hence, the p-coordinates (Note that the p-coordinates is not unique), η-coordinates and θ-coordinates of this system can be given by:

$$\mathbf{p} \equiv [p_{001}, p_{010}, p_{011}, p_{100}, p_{101}, p_{110}, p_{111}]^T, \boldsymbol{\eta} \equiv [\eta_1, \eta_2, \eta_3, \eta_{12}, \eta_{13}, \eta_{23}, \eta_{123}]^T, \boldsymbol{\theta} \equiv [\theta_1, \theta_2, \theta_3, \theta_{12}, \theta_{13}, \theta_{23}, \theta_{123}]^T.$$

Given any p-coordinates of a distribution, the computation of η-coordinates is direct, and the θ-coordinates can be obtained by formula (2). For example, it is easy to check that $\theta_1 \equiv \log(p_{100}/p_{000}), \theta_{12} \equiv \log(p_{110}p_{000}/p_{100}p_{010}), \theta_{123} \equiv \log(p_{111}p_{100}p_{010}p_{001}/p_{110}p_{101}p_{011}p_{000})$ etc. The components of η-coordinates, except the unary marginals, can reflect interactions of words. For example, η_{12} measures the co-occurrence between w_1 and w_2 in the sense that the larger η_{12} is, the more frequent the co-occurrence between w_1 and w_2 is.

The effect of an interaction can be evaluated with respect to a likelihood or log-likelihood function. To be specific, given a η-coordinates $\boldsymbol{\eta}$, η_{12} is natural measure of the interaction between w_1 and w_2. An increment $\Delta\eta_{12}$ of η_{12} will result in increments of log-likelihood function at different **x**s. It is convenient to write these increments in the vector form $\Delta\mathbf{l}(\Delta\eta_{12})\equiv[\Delta l_{000}(\Delta\eta_{12}),\ldots,\Delta l_{111}(\Delta\eta_{12})\ ]^T$, where $\Delta l_{ijk}(\Delta\eta_{12})\equiv l([i, j, k]^T,\boldsymbol{\eta}')-l([i, j, k]^T,\boldsymbol{\eta})$, $i, j, k\in\{0,1\}$, and $\boldsymbol{\eta}'$ is the same as $\boldsymbol{\eta}$ except that the parameter η_{12} becomes $\eta_{12}+\Delta\eta_{12}$. A natural intuition is that, if another component ξ of $\boldsymbol{\eta}$ is irrelevant to the interaction between word w_1 and w_2, then the vector $\Delta\mathbf{l}(\Delta\xi)$ should be orthogonal to the vector $\Delta\mathbf{l}(\Delta\eta_{12})$. It is easy to check that the parameter orthogonality given in Definition 2 is only a weighted generalization of the orthogonality between the above incremental vectors of the log-likelihood function, and hence shares the essentially identical meaning with the original one. It turns out that we have an intuitive reason to consider a parameter ξ independent of all 2-order interactions if ξ is orthogonal to all η_{ij}s. More technically, this is summarized in Theorem 1.

Theorem 1: Given a coordinate system $\boldsymbol{\xi}\equiv[\xi_1\ldots\xi_n]^T$, if ξ_i is orthogonal to ξ_j, then the Maximum Likelihood Estimation (MLE) of ξ_i is independent of the value of ξ_j.

Theorem 1 technically confirms our intuition on the independence between parameters. It guarantees a nice property of orthogonal parameters, which remarkably simplifies some common procedures of hypothesis test relevant to our purpose. We will revisit this issue in later.

According to the above discussion, it is natural to require that any measure reflecting pure k-order interactions should be orthogonal to all parameters reflecting lower-order interactions. The requirement cannot be met by η-coordinates or θ-coordinates alone. For example, there might often be the dependence between η_{123} and η_{12}. Hence η_{123} can not reflect the pure 3-order interaction. On the other hand, Information geometry assures that the η-coordinates and θ-coordinates are dually orthogonal coordinates.

Theorem 2: Let the η-coordinates and θ-coordinates of $\mathbf{S}_n$ be $\boldsymbol{\eta}\equiv[\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_n]^T$ and $\boldsymbol{\theta}\equiv[\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_n]^T$ respectively, where $\boldsymbol{\theta}_1\equiv[\theta_1,\ldots,\theta_n]^T$, $\boldsymbol{\theta}_2\equiv[\theta_{12},\theta_{13},\ldots,\theta_{(n-1)n}]^T$ and so on, and let $\boldsymbol{\eta}_{k-}\equiv[\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_k]^T$ and $\boldsymbol{\theta}_{k+}\equiv[\boldsymbol{\theta}_{k+1},\ldots,\boldsymbol{\theta}_n]^T$, then in the k-cut mixed coordinate $\boldsymbol{\zeta}_k\equiv[\boldsymbol{\eta}_{k-},\boldsymbol{\theta}_{k+}]$, any θ parameter is orthogonal to all η parameters, and vice versa.

Hence, we can construct the mixed-coordinates, e.g., $\boldsymbol{\zeta}_2\equiv[\eta_1,\eta_2,\eta_3,\eta_{12},\eta_{13},\eta_{23},\theta_{123}]^T$, such that θ_{123} is orthogonal to all η_i and η_{ij}. It can also be shown that θ_{123} is orthogonal to other common interaction measures, e.g., $cov_{ij}\equiv\eta_{ij}-\eta_i\eta_j$ and the correlation coefficient ρ_{ij}. Furthermore, it is easy to check that, if we generalize the definition of cov and ρ to the high-order case, e.g., $cov_{ijk}\equiv\eta_{ijk}-\eta_i\eta_j\eta_k$, the above claim still holds accordingly. Another important observation is that the independence of $X_1,\ldots,X_k$ implies $\theta_{1\ldots k}=0$[5]. Hence, $\theta_{1\ldots k}$ is a relevant measure of pure k-order interactions. By now, we are able to construct the proper coordinate system aiming at measuring pure high-order interactions. In practice, the measuring procedure of pure high-order interactions can follow two threads: one is directly parametric estimation

[5] We should not require that the converse proposition holds, since $\theta_{1\ldots k}=0$ does not entail the independence of $X_1, X_2,\ldots,X_k$ if there are lower-order dependences among them.

of mixed coordinates; the other is computing the KL-divergence between the original state and the surrogate state using the Pythagoras relation entailed by the dual orthogonality of mixed coordinates.

2.3 Parametric Estimation of Mixed Coordinates

It is natural to investigate the pure k-order interaction in the (k-1)-cut mixed coordinate $\zeta_{k-1} \equiv [\boldsymbol{\eta}_{(k-1)-}, \theta_{1\ldots k}]^T$ of $\mathbf{S}_k$, since the dual orthogonality gives a simple form of the Fisher information metric, and hence simplifies the estimation procedure of $\theta_{1\ldots k}$.

Given $[\boldsymbol{\eta}_{(k-1)-}, \theta_{1\ldots k}]^T$, let us consider a standard procedure of hypothesis test concerning the null hypothesis H_0: $\theta_{1\ldots k} = \theta^{(0)}_{1\ldots k}$ against H_1: $\theta_{1\ldots k} \neq \theta^{(0)}_{1\ldots k}$. Let the log likelihood of models H_0 and H_1 be

$$l_0 = \max_{\boldsymbol{\eta}_{(k-1)-}} \log p\left(\mathbf{x}_1, \ldots, \mathbf{x}_N; \boldsymbol{\eta}_{(k-1)-}, \theta^{(0)}_{1\ldots k}\right), \quad l_1 = \max_{\boldsymbol{\eta}_{(k-1)-}, \theta_{1\ldots k}} \log p\left(\mathbf{x}_1, \ldots, \mathbf{x}_N; \boldsymbol{\eta}_{(k-1)-}, \theta_{1\ldots k}\right)$$

where N is the number of observations.

The likelihood ratio test uses the test statistic $\lambda \equiv 2\log(l_1/l_0)$. It can be shown that $\lambda \sim \chi^2(1)$, where the degree of freedom in Chi-squared distribution is determined by the difference of the free parameter number between l_0 and l_1. Since the distribution of test statistics is known, we can obtain the estimated value of $\theta_{1\ldots k}$. However, the free parameters of l_1 and l_0 are often considerably huge. As a consequence, the computational cost might be prohibitive for the coordinates without dual orthogonality. In the mixed coordinates with dual orthogonality, the likelihood maximization with respect to $\boldsymbol{\eta}_{(k-1)-}$ and $\theta_{1\ldots k}$ can be performed independently, and hence we have

$$l_0 = \log p\left(\mathbf{x}_1, \ldots, \mathbf{x}_N; \hat{\boldsymbol{\eta}}_{(k-1)}, \theta^{(0)}_{1\ldots k}\right), \quad l_1 = \max_{\theta_{1\ldots k}} \log p\left(\mathbf{x}_1, \ldots, \mathbf{x}_N; \hat{\boldsymbol{\eta}}_{(k-1)}, \theta_{1\ldots k}\right)$$

where $\hat{\boldsymbol{\eta}}_{(k-1)}$ can be estimated independently and kept unchangeable for both l_1 and l_0. Hence, the parametric space is remarkably reduced and the likelihood ratio test becomes feasible.

2.4 Kullback-Leibler Divergence and Pythagoras Relation

The properties of dual orthogonal coordinates entail the generalized Pythagoras theorem, which gives a decomposition of the Kullback-Leibler divergence (KL-divergence for short) such that we can examine different contributions in the discrepancy of two probability distributions, or contributions of different ordered interactions of words.

The KL-divergence between two probabilities p($\mathbf{x}$) and q($\mathbf{x}$) is defined by $D[p:q] \equiv \sum_{\mathbf{x}} p(\mathbf{x})\log[p(\mathbf{x})/q(\mathbf{x})]$. Given a distribution $p \in \mathbf{S}_k$, let p_m be the distribution that is the closest to p and without pure k-order interactions, We then have $p_m = \arg\min_{q \in \mathbf{E}_{(k-1)+}(0)} D[p:q]$, where $E_{(i-1)+}(0)$ is the set of all distributions having no k-order interactions, i.e., $\theta_{1\ldots k}=0$. We refer to p_m as the m-projection of p to $E_{(i-1)+}(0)$. Let the mixed coordinates of p be $[\boldsymbol{\eta}_{(k-1)-}, \theta_{1\ldots k}]^T$, then the coordinates of p_m is $[\boldsymbol{\eta}_{(k-1)-}, 0]^T$.

An important result of Information Geometry guarantees that KL-divergence can been approximated subject to the Riemannian metric tensor derived from Fisher information:

$$ds^2 = \sum_{i,j} g_{ij}(\xi) d\xi_i d\xi_j = 2D\left[p(\mathbf{x};\xi) : p(\mathbf{x};\xi + d\xi)\right] \quad \textbf{(5)}$$

This approximation would remarkably simplify the computation of KL-divergence between a distribution p and its m-projection p_m. To explain the Pythagoras relation, we need the following definitions:

Definition 3: A coordinate curve is called an *e-geodesic* if it is given by a linear function $\boldsymbol{\theta}$(t)=t**a**+**b** in the θ-coordinates, where **a** and **b** are constant vector. A coordinate curve is called a *m-geodesic* if it is given by a linear function $\boldsymbol{\eta}$(t)=t**a**+**b** in the η-coordinates, where **a** and **b** are constant vectors.

Theorem 3 (Pythagoras relation): Let p, q and r be three distributions. If the m-geodesic connecting p and q is orthogonal at q to the e-geodesic connecting q and r, then we have D[p:r] = D[p:q] + D[q:r].

Based on Theorem 3, given any p_0 with the coordinate $[\boldsymbol{\eta}'_{(k-1)-},0]^T$, we have $D[p:p_0]= D[p:p_m]+ D[p_m:p_0]$. The first decomposing term of KL-divergence, i.e., $D[p:p_m]$ offers us another relevant statistic to quantitatively evaluate the level of high-order interactions. Note that the $D[p:p_m]$ can be computed by formula (5).

3 Characterizing High-order Entanglements in Semantic Spaces

3.1 On Semantic Implications of Pure High-order Interactions

In this section, we illustrate by two artificial examples the semantic implication of pure high-order entanglements in semantic spaces. Our fundamental idea is: If a set of words as a whole has a significant interaction that cannot be reduced to the compositional effect of lower-order interactions, then this pure high-order interaction implies the emergence of some semantic entity.

Example 1: Given a corpus related to the history of French wars, a word set {w_1=revolution, w_2=Waterloo, w_3=Napoleon} and their occurrence/co-occurrence probabilities:

$$\eta_1 \equiv \frac{\#chunk_1}{\#chunk} = 0.43001, \quad \eta_2 \equiv \frac{\#chunk_2}{\#chunk} = 0.40011, \quad \eta_3 \equiv \frac{\#chunk_3}{\#chunk} = 0.67000$$

$$\eta_{12} \equiv \frac{\#chunk_{12}}{\#chunk} = 0.18001 \quad \eta_{13} \equiv \frac{\#chunk_{13}}{\#chunk} = 0.40000, \quad \eta_{23} \equiv \frac{\#chunk_{23}}{\#chunk} = 0.44000$$

$$\eta_{123} \equiv \frac{\#chunk_{123}}{\#chunk} = 0.18000$$

where η_i is the marginal occurrence probability of w_i's in all chunks (a chunk is a unit fragment of text, e.g., within a window, a paragraph, a section or a document.), η_{ij} is the co-occurrence probability of w_i and w_j, η_{123} is the joint co-occurrence probability of w_1, w_2 and w_3, '#chunk' is the total number of chunks, $\#chunk_i$ is the number of chunks in which w_i occurs, $\#chunk_{ij}$ is the number of chunks in which w_i and w_j co-occur simultaneously and so on.

In example 1, there is a correlation between the occurrences of 'revolution' and 'Napoleon' since the early life of Napoleon is closely related to the France revolution. There is also a correlation between the occurrences of 'Waterloo' and 'Napoleon' since the Waterloo battle ended the myth of Napoleon. Because both 'revolution' and 'Waterloo' are correlated with 'Napoleon', there is also a significant interaction among these three words. We consider the interaction of these three words *significant* if $\eta_{123}>\eta_1\eta_2\eta_3$, e.g., the joint occurrence probability is significantly greater than the product of marginal occurrence probabilities.

It is clear that the set {revolution, Napoleon, Waterloo} cannot be naturally mapped to a realistic event or a specifically semantic entity even if there is an obvious interaction among these three words. One may argue that the whole of these three words is still meaningful since both 'revolution' and 'Waterloo' are related to 'Napoleon', and hence the combination of 'revolution' and 'Waterloo' offers a more complete picture on 'Napoleon'. However, this 3-word correlation is not a *pure* 3-order correlation. Specifically, let us assume that we have already known there were two significant 2-word correlations, i.e., the correlation between 'revolution' and 'Napoleon' and the correlation between 'Waterloo' and 'Napoleon', then it is natural to consider that 'Napoleon' is related to 'revolution' and 'Waterloo' even if we have no any knowledge on the 3-word interaction. It turns out that the extra knowledge on the existence of a 3-word interaction offers nothing new for us. The above insight is confirmed by the observation of $\eta_{123}\approx\eta_{13}\eta_{23}$, which implies that the obvious interaction of w_1, w_2 and w_3 can be explained by a coincidence of two pairwise events. Consequently, in many applications, e.g., query expansion in information retrieval, the 3-order correlation between {revolution, Napoleon, Waterloo} may not bring much added value then the consideration of the individual 2-order correlations, i.e., between 'revolution' and 'Napoleon' and between 'Waterloo' and 'Napoleon'.

Example 2: Given the same corpus, another word set {w_3=Napoleon, w_4= invasion, w_5= Spain} and the corresponding occurrence/co-occurrence probabilities:

$$\eta_3 \equiv \frac{\#chunk_3}{\#chunk} = 0.5801,\quad \eta_4 \equiv \frac{\#chunk_4}{\#chunk} = 0.3701,\quad \eta_5 \equiv \frac{\#chunk_5}{\#chunk} = 0.0804$$

$$\eta_{34} \equiv \frac{\#chunk_{34}}{\#chunk} = 0.2800,\quad \eta_{35} \equiv \frac{\#chunk_{35}}{\#chunk} = 0.0801,\quad \eta_{45} \equiv \frac{\#chunk_{45}}{\#chunk} = 0.0801$$

$$\eta_{345} \equiv \frac{\#chunk_{345}}{\#chunk} = 0.0800$$

The high-order interaction that makes better sense semantically is the pure high-order interaction. In 3-word cases, roughly speaking, a pure 3-order interaction should meet the condition $\eta_{123}>\eta_{12}\eta_{23}$, $\eta_{123}>\eta_{13}\eta_{23}$, $\eta_{123}>\eta_{12}\eta_{13}$, $\eta_{123}>\eta_1\eta_2\eta_3$, $\eta_{123}>\eta_1\eta_{23}$, $\eta_{123}>\eta_2\eta_{13}$ and $\eta_{123}>\eta_3\eta_{12}$, i.e., the joint probability indicating a pure high-order interaction should be greater than any possible compositional effect of lower-order correlations. In Example 2, since Napoleon launched a series of famous invasions, there is a high correlation between 'Napoleon' and 'invasion'. On the other hand, since Spain is not very important during Napoleon's life except for a short period during Spain war, there is only a relatively low correlation between 'Napoleon' and 'Spain'. However, η_{345} is approximately equal to η_{35} and η_{45} since Napoleon's invasion to Spain is the most important event relating Napoleon to Spain. Hence we have $\eta_{345}>\eta_{34}\eta_{35}$. Furthermore, it is easy to check that we have $\eta_{345}>\eta_{34}\eta_{45}$, $\eta_{345}>\eta_{35}\eta_{45}$, $\eta_{345}>\eta_3\eta_4\eta_5$, $\eta_{345}>\eta_3\eta_{45}$ and so on. Therefore, η_{345} is significant greater than any possibly compositional effect of lower-order interactions. Hence, we can conclude that there exists a pure 3-order interaction of w_3, w_4 and w_5, which cannot be explained by a coincidence of lower-order events and implies an emergence of a semantic entity corresponding to the event of Napoleon's invasion to Spain.

It should be noted that we can also define η_i to be consistent with the conventional vector model if the η_i is computed with respect to the chunk of a word. In this case, all the above discussions are essentially similar subject to a minor modification.

The above discussion seems to imply a method identifying pure high-order interaction, i.e., by checking whether η_{345}-$\eta_3\eta_4\eta_5$, η_{345}-$\eta_{35}\eta_{45}$, η_{345}-$\eta_{34}\eta_{45}$, and so on, are greater than zero. However, this naïve method is in general difficult to be applied. For illustration, let us consider the task of identifying k-order pure interactions by an exhaustive search. First, we have to check whether the k-order interaction is significant than any possible bipartition coincidence. Hence we need to compare $\sum_{i=0}^{k} C_k^i = 2^k$ configurations. Second, we have to check whether the k-order interaction is significant than any possible tri-partition coincidence. It turns out that we also have to check all possible l-partitions ($l \le k$). In summary, the number of configurations that we need to check is given by the Bell number B_k. Recall that the exponential generating function for Bell numbers is $\sum_{n=0}^{\infty} \frac{B_n}{n!} z^n = e^{e^z-1}$, it is, in general, prohibitively complex. Furthermore, the difficulty of an exhaustive search strategy also lies in its intrinsic unstableness in practice, especially for small corpus since we can only control the search procedure by a set of ad-hoc thresholds, which is lack of theoretical guarantees. On the other hand, by IG method, the measure of any k-order pure interaction can be given by a closed-form formula. In addition, we can perform some rigorously-established estimation procedure, e.g., the likelihood ratio test introduced in Subsection 2.2, to quantitatively determine how significant our decision is.

3.2 Characterizing Pure High-order Interactions by Information Geometry

Information Geometry offers a promising method to estimate pure high-order interactions. The likelihood ratio test described in Subsection 2.3 can be directly applied to estimate the statistic $\theta_{1\ldots k}$ which measures pure k-order interactions. Moreover, as described in Subsection 2.4, we can measure the level of high-order interactions by decomposing the KL-divergence with respect to a proper m-projection. As a demonstration, we can directly derive θ-parameters from the p-coordinates as shown in the following.

In Example 1, the η-coordinates is given. It is easy to obtain the p-coordinates from η-coordinates by solving a simple linear system. According to p-coordinates, its θ-parameters are θ_{12}=-0.0004, θ_{13}=5.3932, θ_{23}=11.264, θ_{123}=-3.4584. The negative value of θ_{123} indicates that, although η_{123} is large in absolute value, there is no pure 3-order interaction among the corresponding words. Moreover, the interaction level among w_1, w_2 and w_3 is lower than the compositional effect of lower-order interactions. In Example 2, the θ-parameter are θ_{34}=0.8926, θ_{35}=0.5991, θ_{45}=0.6049 and θ_{345}=6.4852.

The positive value of θ_{345} indicates that, although η_{345} is small in the absolute value, there is still a significant pure 3-order interaction among w_3, w_4 and w_5.

3.3 An Extended Vector Model with Pure High-order Interactions

To investigate semantic implications of high-order interactions, we extend the conventional vector model so that it can incorporate high-order interactions. Traditionally, the marginal distribution of words has acted as the language model in IR (Information Retrieval), MT (Machine Translation) and NLP (Natural Language Processing) because a general higher-order model is often computationally expensive even in the 2-order case. However, in many practical applications, it is unnecessary to construct a general high-order model involving all high-order interactions. On the other hand, it is often sufficient to comprise only a small proportion of high-order interactions in a context-sensitive way, e.g., the pure high-order interaction corresponding to some specific subject. This idea is formalized in the following.

Definition 4 (Vector Model): Given a word set $\{w_1,\ldots,w_n\}$ derived from a corpus C, a text's (corpus's) Vector Model (VM) with respect to $\{w_1,\ldots,w_n\}$ is the marginal distribution $[p_1,\ldots,p_n]^T$ of this text (corpus), where p_i is the marginal probability of w_i.

Definition 5 (Extended Vector Model): Given a word set $\{w_1,\ldots,w_n\}$ derived from a corpus, a text's (corpus's) Extended Vector Model (EVM) is composed by the marginal distribution and some statistics measuring the pure high-order interaction, and has the following form: $\left[p_1,\ldots,p_n,\theta_{i_1\ldots i_{k_1}},\theta_{j_1\ldots j_{k_2}},\ldots\right]^T$ or $\left[p_1,\ldots,p_n,D_{i_1\ldots i_{k_1}},D_{j_1\ldots j_{k_2}},\ldots\right]^T$, where $\theta_{i_1\ldots i_{k_1}}$ is the θ parameter subject to the joint distribution of $\left\{w_{i_1},\ldots,w_{i_{k_1}}\right\}$, $D_{i_1\ldots i_{k_1}}$ is the KL-divergence between p and p_m subject to $\left\{w_{i_1},\ldots,w_{i_{k_1}}\right\}$ (see Subsection 2.4), $p_1,\ldots,p_n$ is the marginal probability of $w_1,\ldots,w_n$.

3.4 Practical Applications in Text Classification and Query Expansion

The remaining issue is to determine what θ or D should be included in an EVM. This issue can only be clarified in specific application backgrounds. We give two examples to explain this issue.

In the task of supervised text classification, it is useful to extract a set of words for each class representing the class subject so that the classification model can be designed accordingly. These sets of theme words can be obtained, in principle, by finding out the word set having significantly pure high-order interactions with respect to the joint distribution of the corresponding class. This finding procedure can be efficient by the aid of prior knowledge. For example, if a few initial theme words are given, it is natural to only search possible pure high-order interactions involving some of prior theme words. Even if there is no prior knowledge on class' subjects, the pure high-order interactions relevant to a specific class can be found by checking, e.g., the mutual information between high-order interactions and class labels. Another method evaluating the relevance between pure k-order interactions and class subjects is to compare the class label of the original state and the surrogate state (see Section 1) with vanishing pure k-order interactions. The surrogate states can be obtained by direct searching over the corpus or manual construction. In the latter case, the fictitious class label of a surrogate state is determined by the classification model trained with respect to the EVM involving the pure k-order interactions.

In query expansion tasks, it is desirable to mine the pure high-order interactions involving some of query words so that the marginal language model can be expanded accordingly. We suggest that the pure high-order interaction involving query words would be an indication of relevance of the query theme. The following is a brief algorithmic framework:

1 Collect top ranked initial retrieval results into a set S_I

2 Search word subsets involving some query words and other words, and compute the pure high-order interactions.

3 Construct S_I's EVM by incorporating the pure high-order interactions mined in step 2.

4 Get new search results based on the derived EVM. There can be a number of ways to do that, for example, by using the EVM as a relevance language model to filter or re-rank S_I or to expand the initial query using words with pure high-order interactions with query words; etc.

4 Conclusions and Further Work

Pure high-order entanglements in lexical semantic spaces indicate the emergence of high-level semantic entities. To characterize the intrinsic order of entanglements and distinguish pure high-order entanglements from lower-order ones, we develop a set of methods in the framework of Information Geometry. Based on the developed method, we present an expanded vector space model that involves context-sensitive high-order information and aims at characterizing high-level context. Several examples with specific application backgrounds, e.g., query expansion and text classification, are discussed, and an algorithmic framework incorporating our method in query expansion are proposed. The further work is to carry out practical experiments and develop more efficient algorithms to implement the proposed framework. To this end, some nice properties of pure high-order correlations, e.g., sub-inheritance, can be used to improve the computational efficiency.

5 Acknowledgements

This research is funded in part by the UK's Engineering and Physical Sciences Research Council, grant number EP/F014708/1 and Natural Science Foundation of China (Grant 60603027).

References

1. Salton, Gerard and C. Buckley (1988). Term-weighting approaches in automatic text retrieval, Information Processing & Management, vol. 24(5), pp. 513-523.
2. M. Nielsen and I. Chuang (2000). Quantum Computation and Quantum Information. Cambridge University Press.
3. R. Horodecki, P. Horodecki, M. Horodecki, K. Horodecki: Quantum entanglement, http://arxiv.org/abs/quant-ph/0702225
4. S. Amari et al: Method of Information Geometry. AMS series, Oxford University press (2000)
5. S. Amari (1985). Differential Geometrical Method in Statistics, Springer-Verlag.
6. S. Amari (2001). Information Geometry on Hierarchy of Probability Distributions. IEEE Transactions on Information Theory, 47: 1701-1711.
7. H. Nakahara and S. Amari (2002). Information-Geometric Measure for Neural Spikes. Neural Computation, 14:2269—2316.
8. C. R. Rao (1945). Information and accuracy attainable in the estimation of statistical parameters. Bull. Calcutta. Math. Soc., 37.
9. H. Jeffreys (1946). An invariant form for the prior probability in estimation problems. Proc. Roy. Soc., A, 196.
10. R. Christensen (1997). Log-Linear Models and Logistic Regression, Springer-Verlag.
11. A. Orlitsky, N. Santhanam, and J. Zhang, "Always Good-Turing: Asymptotically optimal probability estimation," S cience, vol. 302.
12. P.D. Bruza, K. Kitto, D. Nelson and C. L. McEvoy (2008). Entangling Words and Meaning. Proceedings of the Second Quantum Interaction Symposium, pp. 118-124, College Publications.
13. P.D. Bruza and J. Woods (2008). Quantum Collapse in Semantic space: Interpreting Natural Language Argumentation. Proceedings of the Second Quantum Interaction Symposium, pp. 141-147, College Publications.
14. P.D. Bruza and R.J. Cole (2005). Quantum Logic of Semantic Space: An Exploratory Investigation of Context Effects in Practical Reasoning. In S. Artemov, H. Barringer, A. S. d'Avila Garcez, L.C. Lamb, J. Woods (eds.) We Will Show Them: Essays in Honour of Dov Gabbay. College Publications.
15. C. Burgess, L. Livesay and K. Lund (1998). Explorations in Context Space: Words, Sentences, Discourse, In Foltz, P.W. (Ed) Quantitative Approaches to Semantic Knowledge Representation. Discourse Processes, 25(2&3), pp. 179-210.
16. T. Landauer and S. Dumais (1997). A Solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211-240.
17. D. Widdows (2004). The Geometry of Meaning. Center for the Study of Language and Information/SRI.
18. M. B. Plenio and S. Virmani, Quant. Inf. Comp. 7, 1 (2007).
19. J. Eisert and D. Gross, in Lectures on Quantum Information, edited by D. Bruß and G. Leuchs (Wiley-VCH,Weinheim, DE, 2007).
20. Gerardo Adesso and Fabrizio Illuminati1, Genuine multipartite entanglement of symmetric Gaussian states: Strong monogamy, unitary localization, scaling behavior, and molecular sharing structure, arXiv:0805.2942v2 [quant-ph] 11 Oct 2008.