IEEE TRANSACTIONS ON
EMERGING TOPICS
IN COMPUTING

Received 31 August 2014; revised 8 December 2014; accepted 14 December 2014. Date of publication 22 December, 2014;
date of current version 10 June, 2015.

Digital Object Identifier 10.1109/TETC.2014.2385594

TSVD as a Statistical Estimator in the Latent Semantic Analysis Paradigm

GIOVANNI PILATO[1] AND GIORGIO VASSALLO[2]

[1]Istituto di Calcolo e Reti ad Alte Prestazioni, Italian National Research Council, Palermo 90128, Italy
[2]Dipartimento di Ingegneria Chimica, Gestionale, Informatica e Meccanica, University of Palermo, Palermo 90128, Italy
CORRESPONDING AUTHOR: G. PILATO (giovanni.pilato@cnr.it)

This work was supported in part by the Italian Ministry of Instruction, University and Research (MIUR) under Project PON01_01687 named "Security and Intelligence System" - PON 2007/2013.

ABSTRACT The aim of this paper is to present a new point of view that makes it possible to give a statistical interpretation of the traditional latent semantic analysis (LSA) paradigm based on the truncated singular value decomposition (TSVD) technique. We show how the TSVD can be interpreted as a statistical estimator derived from the LSA co-occurrence relationship matrix by mapping probability distributions on Riemanian manifolds. Besides, the quality of the estimator model can be expressed by introducing a figure of merit arising from the Solomonoff approach. This figure of merit takes into account both the adherence to the sample data and the simplicity of the model. In our model, the simplicity parameter of the proposed figure of merit depends on the number of the singular values retained after the truncation process, while the TSVD estimator, according to the Hellinger distance, guarantees the minimal distance between the sample probability distribution and the inferred probabilistic model.

INDEX TERMS LSA, data-driven modeling, Hellinger distance.

I. INTRODUCTION

Latent Semantic Analysis (LSA) is a technique based on linear algebra that tries to roughly capture and code the semantics of words and documents [7], [21], [22]. Many researchers have successfully applied this technique for typical Semantic Computing applications, such as natural language understanding, cognitive modeling, speech recognition, smart indexing, anti-spam filters, dialogue systems and other Statistical Natural Language processing problems [2], [3], [18], [20], [29].

Although LSA has been widely employed in statistical NLP, it is a general paradigm that can be applied in principle to any set of elements belonging to a dyadic domain [15], [22], [24]. As reported by Hoffmann, the main theoretical drawback is that in spite of its success in a wide set of applications, the traditional LSA approach still lacks a sound statistical interpretation [14].

The aim of this paper is to describe an attempt to give a theoretical statistical justification of the Latent Semantic Analysis technique based on the Truncated Singular Value Decomposition (TSVD). We use an information geometry approach for interpreting the TSVD decomposition as a stochastic estimator.

Since all data driven models must satisfy some figure of merit, we introduce a quality factor of the model induced from LSA. According to the approach proposed by Solomonoff [26], this factor considers the balancing of two parameters: the *simplicity* of the model and its *adherence* to the training data. The former, which we define as a function of the number of independent parameters required to describe the model, is related to the number of dimensions retained after the truncation process. The latter is measured according to the Hellinger distance [13], [23] between the sample distribution and the inferred probability.

The TSVD, interpreted as an estimator, guarantees the minimal distance between the sample probability distribution and the inferred probabilistic model, for a given number of retained singular values. This analysis makes it possible to give a statistical interpretation of any data driven model built by using the LSA paradigm and, in general, every time data can be expressed as nonnegative matrices [1], [24].

A. LSA AND ITS CONTEXT

The possibility of coding the semantics of words and text segments according to a geometric model plays a key role for the smart management of tasks like digital libraries indexing and information retrieval. Latent Semantic Analysis (LSA) is an automatic procedure for the effective creation of a mathematical model, whose aim is to give a coarse, sub-symbolic, encoding of word semantics [21]. LSA was originally introduced to overcome the shortcomings of traditional keyword lexical matching methods for information retrieval [8] and it is based on the assumption that some words describing or related to the same concept usually appear in the same contexts. The adjective ''latent'' derives from the hypothesis of the existence of what Dumais et al. called an ''*underlying latent semantic structure in word usage data that is partially obscured by the variability of word choice*'' [8], [21].

The traditional LSA procedure starts from a term-document co-occurrence matrix, whose generic element represents the number of times a given word is present in a specific document. Usually, this kind of matrix is sparse. Its rows are associated with words, while its columns are associated with documents or, more generally, to text segments (i.e. paragraphs, sentences, and so on).

LSA tries to code only the information bound to the semantics of words (or natural language text segments). It aims at excluding the accidental, too specific, information, which is strictly related to the particular example used for the model training [8], [21]. To reach this goal, it exploits a dimensionality reduction methodology by decomposing the term-document co-occurrence matrix. The decomposition is calculated by applying a technique closely related to eigenvector decomposition and factor analysis, named Singular Value Decomposition (SVD). Subsequently, only the most important dimensions, associated with the greatest singular values of the co-occurrence matrix, are retained.

The SVD followed by the selection of the most relevant singular values is named Truncated Singular Value Decomposition (TSVD). LSA finds a low-rank approximation of the original term-document matrix. Both the original matrix and its approximation can be seen as vectors whose dimensionality is equal to the number of elements of the matrices (which is the same for both). We can interpret the traditional approximation given by TSVD as the computation of the vector associated with the lower-rank reconstructed matrix that best approximates, according to the Frobenius distance measure, the vector associated with the original matrix. The Truncation step of TSVD attempts to retain the information that is strictly related to word semantics, discarding the distortion given by the accidental use of specific patterns in sample data. The number of retained singular values constitutes the LSA space dimensionality [21].

B. DIFFICULTY OF LSA STATISTICAL INTERPRETATION

The LSA paradigm has been successfully employed in a large range of applications [3], [7], [10], [20], but the explanation of why LSA works remains considerably unclear. An attempt to develop a statistical theory of LSA has been proposed by Tipping and Bishop [27], who introduced a probabilistic interpretation of principal component analysis that is formulated within a maximum-likelihood framework based on a specific form of Gaussian latent variable model. The most significant effort to present a statistical view on LSA has been proposed by Hofmann [14] who introduced the Probabilistic Latent Semantic Analysis (PLSA) methodology. This approach is statistically well founded, however, it is not based on TSVD and it is only analogous to the LSA paradigm in its original formulation. PLSA starts from a statistical model, called *aspect model* [19], based on a set of hidden *aspect variables*, which are used to express the occurrence probability of the word-document pair as a weighted sum of conditional distributions (such a model is known as *mixture model*). However, the performance of PLSA depends on the initialization of the model before training [9] and, as Hoffmann reports, a comparative evaluation of the computational cost between the traditional LSA and the Probabilistic LSA leads to the conclusion that there are some advantages to the first approach [14].

C. THE PROPOSED APPROACH

Our approach starts from the ''philosophical principle'' that all data-driven models should have some *generalization capability* and that they should satisfy two main targets: the adherence to the sample data and the simplicity of the model itself. This consideration is related to the principles of maximum entropy [4]: the model should be the simplest possible, given a distance measure between estimated data and training data; furthermore, the introduction of not essential information into the model should be avoided.

Solomonoff attempted to formalize this problem [26] by introducing a figure of merit having two components: the first one is the shortest description [25] of the inference algorithm, while the second one is a measure of the adherence of the inferred distribution probability to the sample data. This figure of merit should be minimized in order to find an inferred distribution that is a good compromise between the two aforementioned requirements.

We consider a statistical inference problem with a completely data-driven modeling process. The traditional Latent Semantic Analysis based on TSVD is one of the possible methods to infer data-driven models, and information geometry theory can help in presenting a statistical interpretation of this paradigm. We interpret the TSVD decomposition as a statistical estimator by mapping a sample matrix onto a statistical manifold. We establish our interpretation on the minimization of a quality factor that can help to clarify the meaning of the truncation process involved in LSA.

The dimensionality truncation parameter of TSVD acts as a tuning parameter that rules the trade-off between the model simplicity (i.e. the shortest description of the inference algorithm) and the adherence of the model (i.e. the inferred distribution) to the training data. We introduce the definitions

of "*probability amplitude*" and "*probability distribution*" associated with a matrix. Without any assumption of a Gaussian latent variable model, we observe that the Hellinger distance between the final reconstructed matrix and the normalized sample matrix is upper-bounded by the Frobenius distance between the probability amplitude of the normalized sample matrix and its immediate result from TSVD, which is guaranteed to be the minimum by the properties of TSVD. This allows us to consider the TSVD technique as a means for building a statistical estimator.

After a brief review of the traditional, TSVD-based LSA and PLSA paradigms, we will illustrate our information geometric approach that can provide a theoretical foundation of the LSA methodology for statistical inference. Conclusions will then be outlined.

II. LATENT SEMANTICS APPROACHES REVIEW

Although traditional, TSVD-based, Latent Semantic Analysis is usually exploited in Statistical Natural Language Processing, it has been shown that LSA is a general methodology that can be applied, in principle, to any type of count data over a discrete dyadic domain.[1]

The best-known approach that tries to overcome the lack of a statistical interpretation of LSA paradigm is the Probabilistic LSA (PLSA) methodology, proposed by Hoffmann [14]. PLSA is based on a mixture approximation, which is exploited to model the probability of co-occurrence of elements belonging to a dyadic domain. The mixture decomposition has a well-defined probability distribution and the probabilistic meaning of its factors is clearly expressed by the mixture component distributions. It is worthwhile to point out that the PLSA approach is only *analogous* to the traditional LSA paradigm: PLSA does not use the TSVD technique and it is based on a completely different computation procedure.

In the following subsections we will briefly review both these approaches.

A. THE TSVD BASED LSA

Let $D = E \times F$ be a dyadic domain and M, N two positive integers. The starting point of the LSA methodology requires the construction of an $M \times N$ matrix $\mathbf{A}$ whose $(i, j) - th$ entry is the count of the occurrences of the pair (e_i, f_j) in the dyadic domain D (where e_i is the $i - th$ element of the E dimension and f_j is the $j - th$ element of the F dimension). Let K be the rank of $\mathbf{A}$. The following factorization, called *Singular Value Decomposition* (SVD) holds for the matrix $\mathbf{A}$:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{1}$$

where $\mathbf{U}$ is a $M \times K$ orthogonal matrix, $\mathbf{V}$ is a $N \times K$ orthogonal matrix and $\mathbf{\Sigma}$ is a $K \times K$ diagonal matrix, whose diagonal elements $\sigma_1, \sigma_2, \cdots, \sigma_K$ are called *singular values* of $\mathbf{A}$. It can be shown that the singular value decomposition of $\mathbf{A}$ is unique up to the order of the singular values and of the corresponding columns of $\mathbf{U}$ and $\mathbf{V}$, so there is no loss of generality if we suppose that $\sigma_1, \sigma_2, \cdots, \sigma_K$ are ranked in decreasing order.

Chosen an integer $R < K$, let $\mathbf{U}_R$ be the matrix obtained from $\mathbf{U}$ by removing its last $K - R$ columns, $\mathbf{V}_R$ the matrix obtained from $\mathbf{V}$ in the same way and $\mathbf{\Sigma}_R$ the diagonal matrix obtained from Σ by suppressing both its last $K - R$ rows and $K - R$ columns. It can be shown [11] that the matrix:

$$\mathbf{A}_R = \mathbf{U}_R\mathbf{\Sigma}_R\mathbf{V}_R^T \tag{2}$$

is the best rank R approximation to $\mathbf{A}$ according to the Frobenius distance[2]. $\mathbf{A}_R$ is called the *reconstructed matrix*. The process by which $\mathbf{A}_R$ is obtained from $\mathbf{A}$ is called *Truncated Singular Value Decomposition* (TSVD). Further details about the Singular Value Decomposition technique can be found in [12].

B. THE PROBABILISTIC LATENT SEMANTIC ANALYSIS (PLSA)

The most significant attempt to present a statistical view on LSA has been proposed by Hofmann [14] by introducing the Probabilistic Latent Semantic Analysis (PLSA) technique. This approach starts from a statistical model, called *aspect model* [19] that uses a set of R hidden *aspect variables* $z_1, z_2, \cdots, z_R$.

If $D = E \times F$ is a dyadic domain, the aspect variables are used to express the occurrence probability of the pair $(e, f) \in D$ as a weighted sum of conditional distributions (such a model is known as *mixture model*):

$$P(e, f) = \sum_{z=z_1,z_2,\cdots,z_R} P(z)P(e|z)P(f|z) \tag{3}$$

The value of the parameter R is chosen to be much smaller than the cardinality of the dyadic domain dimensions. R, in this case, is the number of relevant features that the mixture model retains from the entire data corpus. By suitably adjusting the values of $P(z)$, this mixture model gives a statistically significant estimation of $P(e, f)$ which only takes into account the chosen relevant features.

Adjustments for $P(z)$ are evaluated by a maximum likelihood iterative algorithm called *Tempered Expectation Maximization* (TEM, [14]). This ensures that the estimated probability distribution is as close as possible to the given sample data according to the Kullback-Leibler distance[3] [14],

[1] A *dyadic domain* D is a set of elements that can be written as $D = E \times F$, where E and F are sets and $\times$ is the ordinary cartesian product between sets. The sets E and F are called *dimensions* of the dyadic domain.

[2] Given two $M \times N$ matrices $\mathbf{A} = [a_{ij}]$ and $\mathbf{B} = [b_{ij}]$, their Frobenius distance is defined by:

$$d_F(\mathbf{A}, \mathbf{B}) = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(b_{ij} - a_{ij})^2}$$

[3] Given two discrete (not necessarily finite) probability distributions $p = \{p_1, p_2, \cdots, p_n\}$ and $q = \{q_1, q_2, \cdots, q_n\}$, their Kullback-Leibler divergence or distance is defined as:

$$KL(p, q) = \sum_i p_i \log \frac{p_i}{q_i}$$

thus ensuring an informative significance of the algorithm.

Traditional TSVD-based LSA lacks such significance. If we recall for convenience the main formula of TSVD decomposition: $\mathbf{A}_R = \mathbf{U}_R\mathbf{\Sigma}_R\mathbf{V}_R^T$, an analogy of PLSA with the traditional TSVD based LSA can be established by considering the following correspondences, as reported in [14]:

$$P(z) \leftrightarrow \Sigma_R \tag{4}$$

$$P(e|z) \leftrightarrow \mathbf{U}_R \tag{5}$$

$$P(f|z) \leftrightarrow \mathbf{V}_R \tag{6}$$

However, despite of its advantages from a theoretical standpoint, PLSA presents some limits. Experiments in the text-processing field have shown that text similarity is better estimated in the LSA low-dimension space because synonyms are mapped to nearby locations and noise is reduced, although handling of polysemy is weak. In contrast, the PLSA model distributes the probability mass of a term over the different latent classes[4] corresponding to different senses of a word, and thus better handles polysemy [14]. Moreover, the likelihood function computed over the training data cannot be used as a predictor of model performance across different models.

III. THE INFORMATION GEOMETRY INTERPRETATION OF LSA

An information geometry approach that can help to give a statistical theoretical explanation of the traditional LSA paradigm based on TSVD decomposition will be illustrated.

The co-occurrence matrix $\mathbf{A}$ of section II-A can be processed and interpreted so that it represents a sample probability distribution. A well-known property of TSVD is that it optimizes the Frobenius norm, i.e. the matrix difference between the matrix $\mathbf{A}$ and the approximated matrix $\mathbf{A}_R$ has the minimum Frobenius norm among the matrices that have the same number of non zero singular values of $\mathbf{A}_R$. However, this kind of norm is not the best choice for calculating distances between probability distributions [6], [28]. A more appropriate measure between these kinds of entities would be the Hellinger distance[5] [13], [23].

It will be demonstrated that a simple preprocessing for all the elements of the sample set, followed by the application of the TSVD procedure, and a subsequent post-processing of the approximated matrix, makes possible the interpretation of the optimization metric used by the Truncated Singular Value Decomposition as the Hellinger distance between the original matrix and one easily computed from the truncated one. This allows us to consider the matrix, calculated starting from the TSVD reconstructed matrix, as an *inferred probabilistic model* of the dyadic domain. Then, the TSVD algorithm, as used in the Latent Semantic Analysis paradigm, acts as an *estimator*, which conveys statistically significant information from the sample to the model.

A. MODELING PROBLEM STATEMENT

Let E and F be two finite non-empty sets, and let $D = E \times F$ be the dyadic domain generated by them. Let $p : D \rightarrow [0, 1]$ be a probability distribution over D, so that $\sum_{x \in D} p(x) = 1$; p can be represented by a $M \times N$ matrix, where M is the number of elements in E and N is the number of elements in F. The (i, j)-th element of the matrix is the probability associated with the element $(x_i, y_j) \in D$. We call $\mathbf{P}$ this matrix. Let us suppose we do not know $\mathbf{P}$, but a statistical sample $q : D \rightarrow [0, 1]$ from $\mathbf{P}$, which is in turn represented by a matrix $\mathbf{Q}$. We want to recast p by using q. This is a statistical inference problem, and the algorithm that allows us to recast p using q is a *statistical estimator*.

A statistical inference problem can be seen as a modeling problem. In fact, we may suppose that a stochastic system exists, whose output is in D, and that this system behaves according to the unknown probability distribution p.

Next, we may suppose that we have observed the behavior of the system by counting the occurrences of each element of D. By dividing by the total number of observations, we obtain a sample $q : D \rightarrow [0, 1]$ from the probability distribution p. In this manner, a statistical inference of p from q can be regarded as an attempt to recast the behavior of a stochastic system from some observations of its output, i.e. to model a stochastic system.

We emphasize that we have no *a priori* knowledge of the system to be modeled: our modeling procedure is *entirely data-driven*.

From now on, we will identify the statistical inference problem with the data-driven modeling process for a stochastic system. A model of such a stochastic system should optimize two requirements:

1) It should reproduce the sample data as closely as possible, with respect to a suitable metric that has some information significance.
   In order to meet this requirement, we could even choose, as an extreme case, to assume q as a model of p without any further processing.
2) The model should have some *generalization capability*. For example, if we take q as a model for p, we obtain an over-fitted model that perfectly adheres to the sample data but that, generally, has poor generalization capability.

This is a common issue when dealing with data-driven modeling processes. Solomonoff tackled it [26], by introducing a figure of merit $\mathfrak{F}$ of a model, given by:

$$\mathfrak{F} = Mod + Dist \tag{7}$$

where the summation is extended over the sample space. The Kullback-Leibler divergence expresses the difference in bits between the amounts of information carried by the probability distribution q and the probability distribution p.

[4] A *latent class* is associated with each hidden variable in PLSA

[5] Given two $M \times N$ matrices $\mathbf{A} = [a_{ij}]$ and $\mathbf{B} = [b_{ij}]$, their Hellinger distance is defined by:

$$d_H(\mathbf{A}, \mathbf{B}) = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(\sqrt{b_{ij}} - \sqrt{a_{ij}})^2}$$

where *Mod* is the shortest description [25] of the inference algorithm, and *Dist* is a measure of the adherence of the inferred distribution probability to the sample data. This figure of merit should be minimized in order to find an inferred distribution that meets the two requirements outlined above. The quantities *Mod* and *Dist* are expressed in *bits*; in fact, in his original derivation [26] Solomonoff takes as *Mod* the shortest Turing-like binary description of the modeling algorithm, and as *Dist* the probability that the modeled system outputs the same sample data as the ''true'' system. Consequently, the figure of merit has the significance of a trade-off parameter between model complexity and the adherence of the model to the sample data.

At present, the result obtained from a TSVD based LSA approach cannot be thought of as a statistical estimator, because:

1) The reconstructed matrix may contain negative entries;
2) The distance that is minimized by TSVD, given the truncation parameter R, is the Frobenius distance, which is not covariant with respect to probability distribution space re-parameterizations;
3) The Frobenius distance is not suitable for performing measures on information conveyed by probability distributions.

It is clear that PLSA, which minimizes a covariant distance (namely the Kullback-Leibler distance), does not present these drawbacks, however the aspect model that it uses is *conceptually different* from TSVD-based LSA.

Here we propose an interpretation of traditional LSA, which allows us to overcome the three issues outlined above. We use the geometric theory of information and, in particular, the Hellinger distance, which is a well founded proximity measure for probability distributions [13], [23].

The number of values required for identifying the probability-inferred model is less than the number of values required to describe the sample matrix. In this context, the truncation parameter R of TSVD is the trade-off parameter between the model complexity (i.e. its shortest description) and its adherence to experimental data. In this manner, the traditional TSVD-based approach to LSA acquires a theoretical significance, meeting the requested trade-off between the shortest description issue and the distance minimization between the sample and the model. Furthermore, it does not lead to an increased computational complexity as PLSA or other analogous techniques do.

B. LSA ALGORITHM FOR STATISTICAL INFERENCE

In this section we will show how a particular processing of the sample data can help in giving a theoretical interpretation of the TSVD technique used in LSA. In order to do this, we define below the concepts of *probability amplitude* and *probability distribution* associated with a matrix.

Let M, N two positive integers and let $\mathbb{R}$ be the set of real numbers. Given a $M \times N$ matrix $\mathbf{B} = [b_{ij}]$ with $b_{ij} \in \mathbb{R}$, $i \in [1, 2, \cdots, M], j \in [1, 2, \cdots, N]$ where at least one of its components $[b_{ij}]$ is positive, we define a set J, composed of all the pairs (i, j) that identify the positive components of $\mathbf{B}$, i.e.:

$$J = \{(i, j) : b_{ij} > 0\} \quad i \in [1, 2, \cdots, M], j \in [1, 2, \cdots, N] \tag{8}$$

Subsequently, we define the *probability amplitude* associated with $\mathbf{B}$, the $M \times N$ matrix $\mathbf{\Psi} = [\psi_{ij}]$ resulting from the mapping $p_a(\cdot)$:

$$\mathbf{\Psi} \equiv p_a(\mathbf{B}) : \mathbb{R}^{M \times N} \rightarrow [0, 1]^{M \times N} \tag{9}$$

whose elements $[\psi_{ij}]$ are computed as:

$$\psi_{ij} = \begin{cases} \dfrac{b_{ij}}{\sqrt{\sum_{(i,j) \in J} b_{ij}^2}} & \text{if} \quad b_{ij} > 0 \\ 0 & \text{if} \quad b_{ij} \leq 0 \end{cases} \tag{10}$$

so that $\forall (i, j)$ it is $\psi_{ij} \geq 0$ and $\sum_{i=1}^{M} \sum_{j=1}^{N} \psi_{ij}^2 = 1$.

We define also the *probability distribution* associated with a matrix $\mathbf{B}$ the $M \times N$ matrix resulting from the mapping $p_d(\cdot)$:

$$\mathbf{B}^{(2)} \equiv p_d(\mathbf{B}) : \mathbb{R}^{M \times N} \rightarrow \mathbb{R}^{M \times N} \tag{11}$$

whose elements are the squares of the elements of $\mathbf{B}$, i.e. $\mathbf{B}^{(2)} = [b_{ij}^2]$.

As written in section 3.1, we are trying to build a model whose complexity is qualitatively given by the figure of merit $\mathfrak{F}$ expressed by formula (7). For clarity, we slightly modify formula (7), without loss of generality, highlighting the fact that it is usually present a parameter γ, which balances the two factors *Mod* and *Dist*:

$$\mathfrak{F}_\gamma := \gamma \cdot Mod + (1 - \gamma) \cdot Dist \quad 0 \leq \gamma \leq 1 \tag{12}$$

According to the Solomonoff paradigm, we are attempting to minimize the figure of merit $\mathfrak{F}_\gamma$. Conceptually, the value of γ that optimizes $\mathfrak{F}_\gamma$ depends on the kind of data and on the specific problem that is tackled. The choice of the optimal value of γ somehow reflects the principle of maximum entropy, which states that ''*in making inferences on the basis of partial information we must use that probability distribution which has maximum entropy subject to whatever is known.*'' [17].

Given two sets of elements $E = \{e_i\}$, $i = 1, 2, \cdots, M$, $F = \{f_j\}$, $j = 1, 2, \cdots, N$, where M is the number of elements in E and N is the number of elements in F and a dyadic domain $D = E \times F$, let $\mathbf{Q}$ be the matrix whose generic element $[q_{ij}]$ is the number of co-occurrences of e_i over f_j, divided by the quantity $\sum_{i=1}^{M} \sum_{j=1}^{N} q_{ij}$. $\mathbf{Q}$ represents a sample distribution q over D. We are trying to find a probability distribution $\hat{p}$, identified by a matrix $\hat{\mathbf{P}}$. A quality factor, given by the figure of merit $\mathfrak{F}_\gamma$, which should be optimized, is associated with $\hat{\mathbf{P}}$. In fact, we are looking for a good balance between the complexity of the model and the distance from data, trying to lower the number of parameters of the model. Obviously, if the model is given only by the sample distribution $\mathbf{Q}$, we obtain $Dist = 0$, i.e. the model has zero

distance from data, however the complexity of the model does not guarantee the best value of $\mathfrak{F}_\gamma$.

Our idea is to map $\mathbf{Q}$ onto a statistical manifold and interpret the TSVD technique as a way to find a statistical model optimizing $\mathfrak{F}_\gamma$. A possible idea would be based on the minimization of the Kullback-Leibler divergence, but this cannot be accomplished by a simple application of the traditional TSVD technique. Such an idea would lead to the well-known PLSA technique [14].

In order to solve the problem, we recall that the Hellinger distance between the representative matrices of the inferred probability distribution $\hat{p}$ and the sample distribution q is:

$$d_H(\hat{\mathbf{P}}, \mathbf{Q}) = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(\sqrt{q_{ij}} - \sqrt{\hat{p}_{ij}})^2} \tag{13}$$

If we call $\mathbf{\Psi}_{\hat{P}}$ and $\mathbf{\Psi}_Q$ the matrices:

$$\mathbf{\Psi}_{\hat{P}} = \left[\sqrt{\hat{p}_{ij}}\right] \tag{14}$$

$$\mathbf{\Psi}_Q = \left[\sqrt{q_{ij}}\right] \tag{15}$$

we observe that we can write:

$$d_H\left(\hat{\mathbf{P}}, \mathbf{Q}\right) = d_F\left(\mathbf{\Psi}_{\hat{P}}, \mathbf{\Psi}_Q\right) \tag{16}$$

where $d_F(\cdot, \cdot)$ is the Frobenius distance.

The matrix $\mathbf{\Psi}_Q$ can be decomposed with the SVD technique:

$$\mathbf{\Psi}_Q = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{17}$$

and its best *rank-R* decomposition $\mathbf{\Xi} = [\xi_{ij}]$ is obtained by applying the TSVD technique, which minimizes the Frobenius distance $d_F(\mathbf{\Xi}, \mathbf{\Psi}_Q)$, given R:

$$\mathbf{\Xi} = \mathbf{U}_R\mathbf{\Sigma}_R\mathbf{V}_R^T \tag{18}$$

The application of the TSVD is only the first step towards the research of our model $\hat{\mathbf{P}}$. In fact, some of the elements ξ_{ij} of $\mathbf{\Xi}$ may be negative; besides, generally, it is $\sum_{i,j}\xi_{ij}^2 \neq 1$. For these reasons the matrix $\mathbf{\Xi}$ cannot be interpreted as a probability amplitude, and a matrix $\mathbf{\Xi}^{(2)}$ resulting from the mapping $p_d(\mathbf{\Xi}) = [\xi_{ij}^2]$ cannot be interpreted as a probability distribution either.

However, even if $\mathbf{\Xi}$ is not a probability distribution, it allows us to identify, without any further addition of external information, the distribution we are looking for.

According to the definitions of probability amplitude and distribution associated with a matrix, starting from $\mathbf{\Xi}$ we compute the *probability amplitude* $\mathbf{\Psi}_{\hat{P}} = p_a(\mathbf{\Xi})$ and its associated *probability distribution* $\hat{\mathbf{P}} = p_d(\mathbf{\Psi}_{\hat{P}})$ for which:

$$d_H\left(\hat{\mathbf{P}}, \mathbf{Q}\right) = d_F\left(\mathbf{\Psi}_{\hat{P}}, \mathbf{\Psi}_Q\right) \leq d_F\left(\mathbf{\Xi}, \mathbf{\Psi}_Q\right) \tag{19}$$

Therefore, if the traditional TSVD technique is applied to the matrix $\mathbf{\Psi}_Q$ rather than to the sample matrix $\mathbf{Q}$, the matrix $\hat{\mathbf{P}}$ is the best approximation, according to the Solomonoff principle, to $\mathbf{Q}$ with respect to the Hellinger distance, for the given value of the truncation parameter R of TSVD.

As a matter of fact, the Hellinger distance between $\hat{\mathbf{P}}$ and $\mathbf{Q}$ is upper-bounded by the Frobenius distance between $\mathbf{\Xi}$ and $\mathbf{\Psi}_Q$, which is guaranteed to be the minimum by the TSVD. In order to evaluate the complexity of our model, intended as the number of parameters that are necessary for describing it, we consider the number of retained singular values of the matrix $\mathbf{\Xi}$, and not the rank of the matrix $\mathbf{\Psi}_{\hat{P}}$. As a matter of fact, the probability amplitude computation step after the TSVD does not increase the number of parameters that are necessary to identify the model. This means that, for evaluating the complexity of the model, we are interested in the rank of the matrix $\mathbf{\Xi}$, which is the truncation parameter R of the TSVD, and for our purpose the rank of the matrix $\mathbf{\Psi}_{\hat{P}}$ (or $\hat{\mathbf{P}}$) is absolutely of no concern. Our goal is to justify the model and not the determination of R, which is a tuning parameter, whose evaluation goes beyond the scope of this paper.

The illustrated approach leads to a model that optimizes the figure of merit arising from the Solomonoff principle. The obtained model $\hat{\mathbf{P}}$, derived from $\mathbf{\Xi}$, is simpler than the model given by $\mathbf{Q}$, since it is possible to express $\hat{\mathbf{P}}$ by using the three matrices $\mathbf{U}_R$, $\mathbf{\Sigma}_R$, and $\mathbf{V}_R$, obtained through the computation of the TSVD. The matrix $\mathbf{\Xi}$ represents a relation between the elements of two, not necessarily different, sets. R can be seen as the number of ''hidden'' independent (orthogonal) ''features'' of the elements of the two sets. R can be also seen as the number of conceptual axes that are considered [29] and the corresponding values of $\mathbf{\Sigma}_R$ are the weights associated with the relevance of the conceptual axes in the reconstruction process of the statistical distribution. The truncation parameter R of the TSVD can assume values ranging from 1 to $rank(\mathbf{Q})$. R plays qualitatively the role of the balancing factor γ in formula (12). Choosing a specific value for R means, operatively, trying to find a model which is the simplest, given the facts that are known, i.e. the available data. This reflects the maximum entropy principle. Let us consider the two extreme cases: $R = rank(\mathbf{Q})$ and $R = 1$. Obviously, if the model is given only by the sample distribution $\mathbf{Q}$ (i.e. $R = rank(\mathbf{Q})$, the model has maximum adherence to sample data. In the same manner, if $R = 1$, the complexity of the model is the lowest possible. The optimal value of R cannot be given a priori, since it depends on both the raw data on which the model is built and on the desired application of the model.

The determination of R is similar to the old problem of separating ''signal'' from ''noise''. From another point of view, it is somehow analogous to the well-known problem of the choice of the number of hidden neurons in a Multi-Layer Perceptron (MLP) neural model [30], for which there is no *a priori* method for setting the number of hidden units without knowing the nature of data [16]. One of the easiest approaches to overcome this point is the use of a validation set, also called ''hold-out'' set [5]. The same approach can be used, as an example, for TSVD-based LSA in classification tasks.

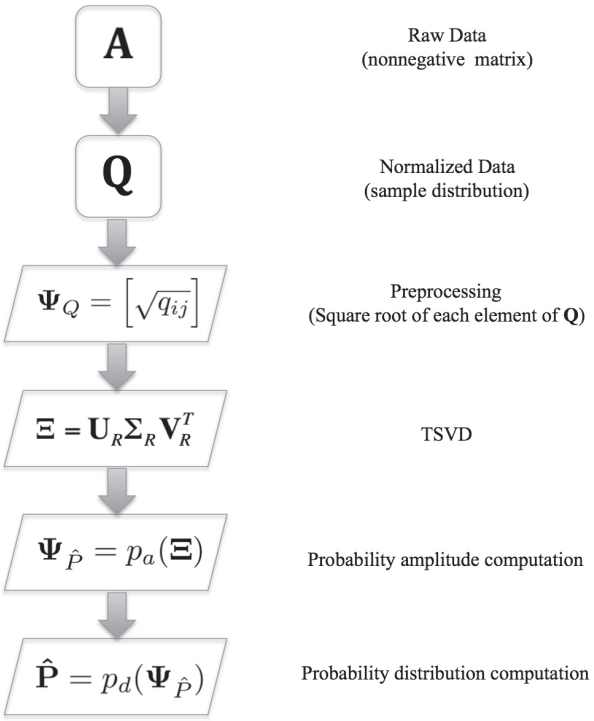

FIGURE 1. The whole procedure that makes it possible to interpret the TSVD-based LSA as a process that leads to the determination of a statistical estimator.

The whole interpretation process, which is summed-up in Figure 1, can explain the theoretical approach to the statistical interpretation of the TSVD-based LSA. We remark that we do not make any assumption of Gaussian latent variable model, as it is done in [27].

C. A GRAPHICAL EXAMPLE OF THE METHODOLOGY

A simplified visual example of the procedure is depicted in Figure 2, where a portion of a unitary circumference C is presented. Let us consider only the positive quadrant. Given a point on C, of co-ordinates $(\hat{x}_1, \hat{x}_2)$, the vector $[\hat{x}_1^2, \hat{x}_2^2]$ can represent a possible probability distribution of a stochastic variable x. Let us have a vector Ψ_Q on the circumference C (see Fig. 2): all the points on the circumference of radius E (i.e. the approximation error), having center on Ψ_Q, represent all the possible vectors that are an approximation of Ψ_Q, according to the Frobenius (Euclidean) metric. Let Ξ be one of these approximations. The transformation of Ξ as $\Psi_{\hat{P}}$ leads to a vector that:

1) is in the compound of the nonnegative quadrant of (x_1, x_2),
2) is projected on the circumference C of radius 1.

The computation of the probability amplitude $\Psi_{\hat{P}} = p_a(\Xi)$ generates an approximation $\Psi_{\hat{P}}$ that is closer (or even equal) to Ψ_Q than Ξ. The difference between $\Psi_{\hat{P}}$ and Ψ_Q can be interpreted as the vector containing the erased information about the not significant variations, from a statistical point of view, in data. It represents the difference between the inferred probability amplitude $\Psi_{\hat{P}}$, and the sample probability amplitude Ψ_Q.

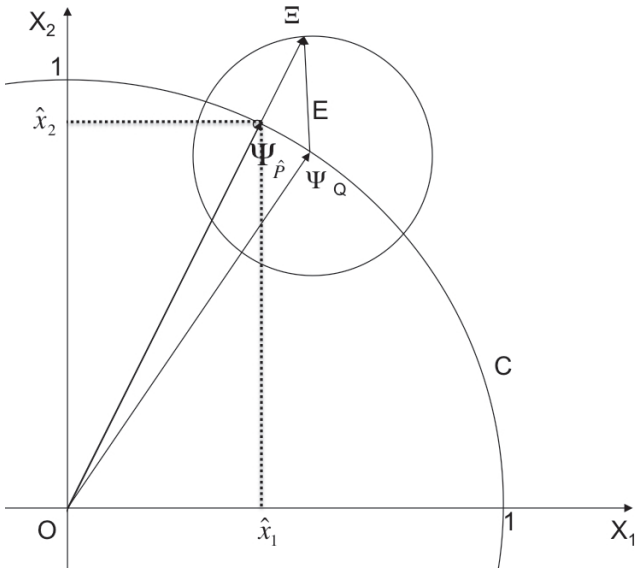

FIGURE 2. A simplified example of the procedure.

IV. CONCLUSIONS

We have illustrated a statistical interpretation of the traditional Latent Semantic Analysis paradigm.

It has been shown that the simple use of the square root for all the entries of the sample set given by the co-occurrence matrix of a dyadic domain and a subsequent mapping as a probability amplitude makes it possible to interpret the TSVD as a statistical estimator.

The idea is to map the matrix that represents a sample distribution over a dyadic domain onto a statistical manifold and to explain the TSVD so that the minimized distance is the Hellinger distance instead of the Frobenius distance. The Hellinger distance is a well-founded proximity measure for probability distributions. The Frobenius distance, computed in the traditional Latent Semantic Analysis approach, determines an upper bound for the Hellinger distance.

We have also introduced a figure-of-merit arising from the Solomonoff approach. This quality measure of the inferred model takes into account both the truncation parameter of the TSVD and the Hellinger distance between the "true" probability distribution of data and the probability approximated by the model inferred with the LSA.

This interpretation allows us to overcome the main drawback of the Latent Semantic Analysis approach, which is the lack of a statistical interpretation of the methodology, and it can be applied to all data driven techniques that exploit the TSVD for the creation of statistical models. It is sufficient to normalize any nonnegative matrix in order to compute Ψ_Q and to obtain a model expressed by the three matrices $\mathbf{U}_R$, $\mathbf{\Sigma}_R$, $\mathbf{V}_R$. The procedure regarding the matrix Ξ, and its subsequent transformation as probability amplitude $\Psi_{\hat{P}}$ are

necessary only for justifying the use of TSVD on the matrix Ψ_Q in order to obtain a statistical estimator.

From the applicative point of view, we have presented a theoretical framework that makes it possible to justify the use of the TSVD for the creation of statistical models every time there are data that can be represented as nonnegative matrices, hence somehow re-interpretable as sample probability distributions.

REFERENCES

[1] F. Agostaro, G. Pilato, G. Vassallo, and S. Gaglio, "A sub-symbolic approach to word modelling for domain specific speech recognition," in *Proc. 7th Int. Workshop Comput. Archit. Mach. Perception (CAMP)*, Jul. 2005, pp. 321–326.

[2] J. R. Bellegarda, "A multispan language modeling framework for large vocabulary speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 456–467, Sep. 1998.

[3] J. R. Bellegarda, "Exploiting both local and global constraints for multi-span statistical language modeling," in *Proc. ICASSP* vol. 2. May 1998, pp. 677–680.

[4] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguistics*, vol. 22, no. 1, pp. 39–71, Mar. 1996.

[5] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). New York, NY, USA: Springer-Verlag, 2006.

[6] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *Int. J. Math. Models Methods Appl. Sci.*, vol. 1, no. 4, pp. 300–307, 2007.

[7] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[8] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman, "Using latent semantic analysis to improve access to textual information," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 1988, pp. 281–285.

[9] A. Farahat and F. Chen, "Improving probabilistic latent semantic analysis with principal component analysis," in *Proc. 11th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, Trento, Italy, 2006, pp. 105–112.

[10] P. W. Foltz and S. T. Dumais, "Personalized information delivery: An analysis of information filtering methods," *Commun. ACM*, vol. 35, no. 12, pp. 51–60, Dec. 1992.

[11] G. H. Golub and C. Reinsch, *Handbook for Matrix Computation II, Linear Algebra*. New York, NY, USA: Springer-Verlag, 1971.

[12] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD, USA: The Johns Hopkins Univ. Press, 1996.

[13] E. Hellinger, "Neue Begründung der theorie quadratischer Formen von unendlichvielen Veränderlichen," *J. Reine Angewandte Math.*, vol. 1909, no. 136, pp. 210–271, 1909.

[14] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. Uncertainty Artif. Intell. (UAI)*, 1999, pp. 289–296.

[15] T. Hofmann, J. Puzicha, and M. I. Jordan, "Learning from dyadic data. in *Advances in Neural Information Processing Systems*, M. S. Kearns, Solla, and D. Cohen, Eds. Cambridge, MA, USA: MIT Press, 1999, pp. 466–472.

[16] D. Hunter, D. Yu, M. S. Pukish, J. Kolbusz, and B. M. Wilamowski, "Selection of proper neural network sizes and architectures—A comparative study," *IEEE Trans. Ind. Informat.*, vol. 8, no. 2, pp. 228–240, May 2012.

[17] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, no. 4, pp. 620–630, May 1957.

[18] J. Koeman and W. Rea. (2014). "How does latent semantic analysis work? A visualisation approach." [Online]. Available: http://arxiv.org/abs/1402.0543

[19] A. Kontostathis and W. M. Pottenger, "Detecting patterns in the LSI term-term matrix," in *Proc. IEEE ICDM Workshop Found. Data Mining Knowl. Discovery (FDM)*, Maebashi, Japan, Dec. 2002, pp. 243–248.

[20] T. K. Landauer and S. T. Dumais, "A solution to PlatoÕs problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge," *Psychol. Rev.*, vol. 104, no. 2, pp. 211–240, 1997.

[21] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse Process.*, vol. 25, nos. 2–3, pp. 259–284, 1998.

[22] T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, Eds., *Handbook of Latent Semantic Analysis*. Evanston, IL, USA: Routledge, 2011.

[23] M. Nikulin, '*Hellinger Distance* (Encyclopedia of Mathematics), M. Hazewinkel, Ed. Kluwer Acad. Publ. Norwell, MA, USA, 2001.

[24] G. Pilato, F. Vella, G. Vassallo, and M. La Cascia, "A conceptual probabilistic model for the induction of image semantics," in *Proc. IEEE 4th Int. Conf. Semantic Comput. (ICSC)*, Sep. 2010, pp. 91–96.

[25] J. Rissanen, (15 Aug, 2006). "Minimum description length principle," *Encyclopedia Statist. Sci.*, DOI: 10.1002/0471667196.ess1641.pub2

[26] R. J. Solomonoff, "The discovery of algorithmic probability," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 73–88, Aug. 1997.

[27] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *J. Roy. Statist. Soc., Ser. B*, vol. 61, no. 3, pp. 611–622, 1999.

[28] M. Tumminello, F. Lillo, and R. N. Mantegna, "Kullback–Leibler distance as a measure of the information filtered from multivariate data," *Phys. Rev. E*, vol. 76, no. 3, p. 031123, 2007.

[29] G. Vassallo, G. Pilato, A. Augello, and S. Gaglio, "Phase coherence in conceptual spaces for conversational agents," in *Semantic Computing*, P. C.-Y. Sheu, H. Yu, C. V. Ramamoorthy, A. K. Joshi, and L. A. Zadeh, Eds. New York, NY, USA: Wiley, IEEE Press, 2010, pp. 357–371.

[30] P. D. Wasserman and T. Schwartz, "Neural networks. II. What are they and why is everybody so interested in them now?" *IEEE Expert*, vol. 3, no. 1, pp. 10–15, Spring 1988.

GIOVANNI PILATO received the Laurea (*cum laude*) degree in electronics engineering and the Ph.D. degree in computer science from the University of Palermo, Palermo, Italy, in 1997 and 2001, respectively. He is currently a Staff Research Scientist with the Istituto di Calcolo e Reti ad Alte Prestazioni, Italian National Research Council, Palermo. He is also a Lecturer with the Department of Computer Science, University of Palermo. His research interests include geometric techniques for knowledge representation, Web data mining, and natural language processing.

GIORGIO VASSALLO received the Laurea degree in physics from the University of Palermo, Palermo, Italy, in 1982, where he is currently a Research Scientist with the Computer Science and Artificial Intelligence Laboratory, Department of Chemical, Management, Computer, and Mechanical Engineering. His research interests include innovative graphic processors, neural networks, geometric techniques for data mining, and natural language processing.