Efficient Algorithms for Graph Regularized PLSA for Probabilistic Topic Modeling

Xin Wang[a,b], Ming-Ching Chang[b], Siwei Lyu[b,*]

[a] *CuraCloud Corporation, Seattle, WA, 98104, USA*
[b] *Department of Computer Science, University at Albany, SUNY, NY, 12222, USA*

Abstract

Probabilistic Latent Semantic Analysis (PLSA) is a popular data analysis method with the objective to discover the underlying semantic structure of input data. In this work, we describe a method for probabilistic topic analysis in image and text based on a new representation of graph-regularized PLSA (GPLSA). In GPLSA, data entities are mapped to an undirected graph, where similarities between topic compositions on the graph are measured by the divergence between discrete probabilities. Such divergence is essentially incorporated as a graph-regularizer that augments the original PLSA algorithm. Furthermore, we extend the GPLSA algorithms to multiple data modalities based on the connections between data entities of each modality. We propose efficient multiplicative iterative algorithms for GPLSA with three popular regularizers, namely ℓ_1, ℓ_2 and symmetric KL divergences. In each case, we derive simple efficient numerical solutions that require only matrix arithmetic operations during the optimization. Experimental results demonstrate the efficacy of GPLSA over state-of-the-art methods.

Keywords: Probabilistic Latent Semantic Analysis, Graph Regularization, Topic Analysis, Clustering

*Corresponding author
Email addresses: xinw@curacloudcorp.com (Xin Wang), mchang2@albany.edu (Ming-Ching Chang), lsw@cs.albany.edu (Siwei Lyu)

1. Introduction

Probabilistic topic modeling, which aims to discover hidden thematic structures in large archival documents and to annotate a large document corpus with thematic information, has been studied extensively in recent years. Typically, a connection is established between the high-dimensional word distribution vectors of the documents and the lower-dimensional topic vectors, where the semantic properties of these words and documents can be expressed in terms of probabilistic topic models. Two basic probabilistic topic models are the *probabilistic latent semantic analysis* (PLSA) [1] and *latent Dirichlet allocation* (LDA) [2]. Topic analysis has wide range of applications including activity detection [3], image analysis [4], pattern recognition [5, 6, 7, 8], natural scene categories [9], video processing [10, 11, 12], information retrieval [13, 14], document analysis [15, 16], and co-authorship network analysis [17, 18, 19], multiple modalities learning, instead of single modality topic analysis, the tasks in topic analysis of multiple modalities attempt to learn dependencies between these modalities [20, 21, 22].

A simple treatment in probabilistic topic modeling is to assume all the topics to be independent from each other. However, in many application domains, the relation between topics are complicated and can be modeled more precisely with a graph. For instance, social networks (such as the Facebook) operate based on a huge set of user profiles and friendship connections; publication archives such as the Digital Bibliography and Library Project (DBLP) contain a vast authorship network. Several recent works [23, 24] have considered the integration of a graph structure with topic modeling in terms of *regularizer* to restrict the relations among the learned topics to be consistent with the graph structure. However, The works of [23, 24] focus on single mortality modelling and the learned topic representations in [24] usually do not afford explicit probabilistic interpretations. Other approaches rely on Bayesian inference performed on the topic network suffers from increased complexity problems for learning and inference [25, 26].

In this work, we study efficient algorithms for the *graph-regularized prob-*

abilistic latent semantic analysis (GPLSA) for probabilistic topic modeling. GPLSA is capable of handling both single- or multiple- modalities of data by introducing a graph structure into the PLSA topic modeling. In our general formulation of GPLSA, the data entries are defined on a graph encoding the semantic relations among the data entries, and the graph regularizers are in the form of the divergences between the discrete probability distributions corresponding to the composition of topics from each data entry. Our formulation enables the entries with the same semantic to smooth out their effects with each other. We propose efficient algorithms for the learning of topics and their compositions using three widely used divergence definitions, namely, ℓ_1, ℓ_2 and symmetric KL divergences. GPLSA optimization is casted as the minimization of the underlying divergences, which encourages similarities between topic compositions of each data entry and its nearest neighbors on the graph. Our algorithm is efficient because the optimization steps consist of only simple matrix operations and the derivation of scalar nonlinear equations. Our GPLSA algorithms also afford theoretical guarantee of convergence, unlike other state-of-the-art works [27, 28]. We apply the GPLSA algorithms in image clustering, cross-modal retrieval and multi-lingual topic analysis applications on public benchmarks for evaluation and comparison. Experimental results show noticeable improvements of GPLSA against other state-of-the-art methods.

Main contributions of this work are summarized in the following:

(1) We propose efficient algorithms for graph-regularized PLSA (GPLSA) as a general framework for single- or multi- modality topic analysis, where the graph regularizer is based on the divergence between discrete probability distributions. Similarities between topics are enforced in a joint latent space constraint by the graph, and topic distributions are enhanced by their nearest neighbors on the graph.

(2) We also study L_1 graph regularizer in this problem, which has not been studied in the previous works. We show improved results for the ℓ_1 regularizer over the baseline. When using ℓ_2 divergence as the regularizer, our GPLSA algorithm is more efficient and with a convergence guarantee than the exist-

ing method based on GNMF [24]. We further describe a new algorithm using symmetric KL divergences as the regularizer, and demonstrate that it is more effective compared to the ℓ_2 divergence.

(3) The proposed algorithms extend naturally probabilistic topic analysis of a single modality to multiple modalities. Our method enables capturing of similarities between documents across modalities, by learning a joint latent space for documents of different modalities. Our topic learning representation leverages the compatible yet complementary conceptual themes among each modality. Thus it is more effective than other methods relying on features derived from direct concatenation of modalities.

An early version of this work focusing on multi-modal learning was published in [29]. This paper improves our previous work in the following aspects. (1) We derive the efficient algorithms for the GPLSA problem in a general framework based on the extended works. We also provide in depth motivations and full technical details. In this regard, our previous work of [29] can be considered as a special case of the general formulation in this paper. (2) We provide simple efficient numerical solutions that require only matrix arithmetic operations for the optimization. We also provide a proof of convergence for the general form of the algorithm. (3) We provide additional diagnostic experiments regarding the clustering performance and multi-lingual topic analysis to demonstrate the effectiveness of our solutions.

The remaining of the paper is organized as follows. After introducing background works in Section 2, we review the PLSA algorithm with mathematical representations using matrix formation in Section 3. This formulation should facilitate the description of the GPLSA algorithm for the clustering task for single modality in section 4.1 and the extension for multi-modality retrieval in section 4.2. Section 5 describes experimental validation by applying the GPLSA algorithms to the image clustering, image/text cross-modal retrieval and multi-lingual topic analysis applications. Section 6 concludes the paper with discussions and future works.

2. Related Works

2.1. Probabilistic Topic Modeling for Single Modal Data

Much of the existing works regarding the integration of a network structure with probabilistic topic modeling have focused on data with single modality (e.g. images,texts, etc), where topic selection preferences are smoothed among nearest neighbors on the graph [24, 30, 31]. For instance, the work in [23] combines topic modeling and social network to analyze the topics in an co-authorship network. On a related front, graph regularizer is exploited to mode the intrinsic structure of data distributions in *graph-regularized non-negative matrix factorization* (GNMF) [24], where encouraging results are obtained in document and image clustering. We will show later that GNMF can be regarded as a special case of the graph-regularized PLSA model, due to the close relationship between the PLSA and NMF [32, 33]. GNMF aims to find a compact representation which uncovers the hidden semantics from the documents and in the meantime represents the intrinsic geometric structure. A semantic representation space is found based on two assumptions that (1) if two data points are connected along a graph edge, they should be sufficiently close to each other, and (2) the representations of these two data points with respect to the new basis are also close to each other. However, such jointly learned latent representations may not have explicit probabilistic interpretations. Other approaches rely on Bayesian inference performed on the topic network [25, 26]. The *Relational Topic Model* (RTM) [25] uses LDA to model the documents and the relationships between them, but suffers from increased complexity problems for learning and inference.

2.2. Probabilistic Topic Modeling for Multiple Modal

Recently several topic analysis techniques for multi-modal data have been proposed [34, 30, 31]. In an early work [34], the interdependencies between published documents take the form of citations which allow instant access to the referenced documents, the citations in the documents are considered as a separate modality in the document corpus. The topics learned form the individual

modalities are weighted combined as the shared topics of the two modalities. More recently, there has been an effort to jointly model the documents topics and other auxiliary information provided within the dateset [17, 18, 19]. For example, the Wikipedia data documents typically include both texts and images, and the work in [20] uses Markov random field of topic models to associate images and texts based on their similarity. However, it assumes that each data entity and its associated auxiliary data share the same topic compositions across modalities in these methods. Such assumption might be too restrictive to be applied on the real-world multi-modal datasets.

Topic analysis methods are also widely applied on multi-lingual text which can be regarded as multi-modal data [26]. For instance, [35] incorporates a bilingual dictionary based on translation bipartite graph into cross-lingual PLSA to extract common topics in cross-languages. Similarity, [36] presents a novel multilingual topic model, they first build a bipartite graph matching over terms in both languages assuming that words have similarity on document level contexts, then the matching topics are learned as the distributions of these matching pairs instead of being distributions over terms. However, both method relay on term pairs in the dictionary and their assumption of matching terms may result the loss of correlated information between the languages. In contrast, in this work, we aim to extract topics from different information sources (images and texts, or texts in different languages) that reflect their intrinsic conceptual similarities. It inherits the advantage of topic models that the learned topics are often intuitive and interpretable.

Recently, there have also been several Bayesian approaches to multi-modal probabilistic topic modeling, a Markov random field (MRF) augmented probabilistic topic model is proposed in [26], which incorporates the similarities between associated topic compositions of different data modalities using MRF. Although good performance in [26] are achieved, the Bayesian MRF method has the problem of increased complexity in both of learning and inference algorithms related with Monte-Carlo methods. Several works also suggested that connecting multiple modalities using a graph structure is crucial for strong per-

formance of the learning algorithms on multi-modality datasets [37, 38, 39, 20], with applications shown in [40, 41, 26].

Therefore, it is useful to extend simpler topic analysis methods such as GPLSA to learning from multi-modal data, whose efficient implementation can be used for rapid analysis of large multi-modal dataset and initializations of more sophisticated Bayesian methods. Two specific methods of extending PLSA to multi-modal learning with co-regularization has been studied in two recent works [27, 28]. The co-regularizer used in [27] is based on the mutual similarities of data in the topic space, and that of [28] is the ℓ_2 divergence between the topic assignments in the latent space. The common drawback of both methods, however, is that the optimization procedure cannot guarantee monotonic improvement of the objective function before a stationary point is reached. As such, the algorithms in these previous works do not afford guarantees to converge and usually lead to inferior performance.

3. Background: Generalized PLSA Algorithm

The proposed Generalized PLSA algorithm is based on the generalized PLSA Algorithm [42]. We first introduce some notations and definitions to be used throughout the paper. A d-dimensional vector $\mathbf{a}$ is stochastic if $a_i \geqslant 0$ and $\sum_{i=1}^{d} a_i = 1$, and corresponds to a categorical probability distribution over d outcomes. A $d \times n$ nonnegative matrix A is stochastic if its column vectors are stochastic.

Consider any two d-dimensional stochastic vectors $\mathbf{a}$ and $\mathbf{u}$, we define their ℓ_1, ℓ_2 and Kulback-Leibler (KL) divergences, as: $\mathcal{D}_{\ell_1}(\mathbf{a}, \mathbf{u}) = \sum_{i=1}^{d} |a_i - u_i|$, $\mathcal{D}_{\ell_2}(\mathbf{a}, \mathbf{u}) = \frac{1}{2} \sum_{i=1}^{d} (a_i - u_i)^2$, $\mathcal{D}_{\mathrm{KL}}(\mathbf{a}, \mathbf{u}) = \sum_{i=1}^{d} a_i \log \frac{a_i}{u_i}$, and their symmetric KL divergence is defined as $\mathcal{D}_{\mathrm{sKL}}(\mathbf{a}, \mathbf{u}) = \mathcal{D}_{\mathrm{KL}}(\mathbf{a}, \mathbf{u}) + \mathcal{D}_{\mathrm{KL}}(\mathbf{u}, \mathbf{a})$. Accordingly, we define the divergence between two stochastic matrices A and U as the sum of the divergences between their corresponding columns, as $\mathcal{D}_*(A, U) = \sum_j \mathcal{D}_*(A_{\cdot,j}, U_{\cdot,j})$, where $\mathcal{D}_*$ can be replaced with $\mathcal{D}_{\ell_1}$, $\mathcal{D}_{\ell_2}$, $\mathcal{D}_{\mathrm{KL}}$ or $\mathcal{D}_{\mathrm{sKL}}$. For stochastic vectors/matrices, these divergences are non-negative and equal to

zero if and only if the two vectors/matrices are identical.

Making analogy to a collection of text documents, we use a "bag-of-word" representation [1] of a dataset, where each data entity (a "document") is represented as the normalized frequencies over some basic features ("words" in a "vocabulary"). PLSA is performed based on a simple probabilistic generative model of the dataset [2]: each word in a document is a sample from a mixture model; each component of the mixture model is a categorical distributions over the vocabulary (a "topic"); the mixing weights of the mixture model correspond to a probability distribution over the topics, and provides the topic composition of the data entity.

Specifically, given n documents $(\mathbf{d}_1, \cdots, \mathbf{d}_n)$ over a vocabulary of size d, $(\mathbf{w}_1, \cdots, \mathbf{w}_d)$, we use stochastic matrix P of dimension $d \times n$ to represent conditional probabilities, as $P_{ij} \equiv \texttt{Prob}(\text{word} = \mathbf{w}_i | \text{doc} = \mathbf{d}_j)$. Assuming the documents are associated with m topics, $(\mathbf{t}_1, \cdots, \mathbf{t}_m)$, we use stochastic matrices A of dimension $d \times m$ and U of dimension $m \times n$ to represent conditional probabilities, as $A_{ik} \equiv \texttt{Prob}(\text{word} = \mathbf{w}_i | \text{topic} = \mathbf{t}_k)$ and $U_{kj} \equiv \texttt{Prob}(\text{topic} = \mathbf{t}_k | \text{doc} = \mathbf{d}_j)$, respectively. According to the document generation model, documents and words are conditionally independent from each other. As such, these probabilities satisfy:

$$\begin{aligned} \texttt{Prob}(\text{word} = \mathbf{w}_i | \text{doc} = \mathbf{d}_j) = \sum_k & \texttt{Prob}(\text{word} = \mathbf{w}_i | \text{topic} = \mathbf{t}_k) \\ & \texttt{Prob}(\text{topic} = \mathbf{t}_k | \text{doc} = \mathbf{d}_j) \end{aligned} \tag{1}$$

With the matrix notations, the equation (1) is equivalent to $P = AU$. Formally, given a dataset represented in stochastic matrix P of size $d \times n$, PLSA attempts to find its decomposition into A and U, where A and U are stochastic matrix of size $d \times m$ and $m \times n$ respectively, formulated as an optimization problem: $\min_{A,U} \mathcal{D}_{KL}(P, AU)$, with the constraint that both A and U are stochastic matrices. After dropping irrelevant constant terms, minimizing the KL divergence is equivalent to maximizing

$$\mathcal{J}(A, U) = \textstyle\sum_{ij} P_{ij} \log (AU)_{ij}. \tag{2}$$

This optimization problem can be solved with block coordinate ascent by iteratively optimizing A or U while fixing the other until converging to a local optimum. The individual optimization step for A and U is solved with the EM algorithm[43, 44, 42]. To facilitate subsequent discussions, we briefly review the EM algorithm using the matrix notations introduced early in this section.

Optimizing A: Introducing a different stochastic matrix $\hat{A}$, we first define an auxiliary function

$$\mathcal{F}(A, \hat{A}) = \sum_{ijk} \frac{P_{ij}\hat{A}_{ik}U_{kj}}{\left(\hat{A}U\right)_{ij}} \log\left(\frac{A_{ik}}{\hat{A}_{ik}}\left(\hat{A}U\right)_{ij}\right) = \sum_{ik} M_{ik} \log A_{ik} + \text{const.} \quad (3)$$

In the last step, terms irrelevant to A are collected into a constant. Nonnegative matrix $M = \hat{A} \otimes \left[(P \oslash (\hat{A}U))U^T\right]$ is formed with element-wise matrix multiplication $\otimes$ and division $\oslash$. An application of the Jensen's inequality shows that $\mathcal{F}(A, \hat{A}) \leqslant \mathcal{J}(A, U)$ with equality holds when $A = \hat{A}$, *i.e.*, $\mathcal{F}(A, \hat{A})$ is a tight lower-bound of $\mathcal{J}(A, U)$. Derivation of Eq.(3) and proof of $\mathcal{F}(A, \hat{A})$ being a tight lower-bound of $\mathcal{J}(A, U)$ are provided in the Appendix A.

The EM algorithm optimizing A uses the above lower-bound to improve the objective function in an iterative manner: Starting with an initial values $A = A^{(0)}$, we iteratively solve for $A^{(t+1)} \leftarrow \text{argmax}_A \mathcal{F}(A, A^{(t)})$ with the constraint A being stochastic. As we have $\mathcal{J}(A^{(t)}, U) = \mathcal{F}(A^{(t)}, A^{(t)}) \leqslant \mathcal{F}(A^{(t+1)}, A^{(t)}) \leqslant \mathcal{J}(A^{(t+1)}, U)$, the sequence $(A^{(0)}, A^{(1)}, \cdots)$ monotonically increases $\mathcal{J}(A, U)$ until reaching a local maximum.

During each iteration step of the EM algorithm, we solve for $\text{argmax}_A \mathcal{F}(A, A^{(t)})$, which using Eq.(3) reduces to

$$\max_A \quad \sum_{ik} M_{ik} \log A_{ik}, \text{ s.t. } \; A_{ij} \geqslant 0 \;\; \& \;\; \sum_i A_{ij} = 1. \quad (4)$$

The solution to this problem is given by $A_{ik} = \frac{M_{ik}}{\sum_{i'} M_{i'k}}$ (proof given in the Appendix A), in which the normalization step and the non-negativity of M assures A to be a stochastic matrix.

Optimizing U: The EM algorithm optimizing U with fixed A proceeds similarly. First using an auxiliary stochastic matrix $\hat{U}$ we define function

$$\mathcal{G}(U, \hat{U}) = \sum_{ijk} \frac{P_{ij}A_{ik}\hat{U}_{kj}}{\left(A\hat{U}\right)_{ij}} \log\left(\frac{U_{kj}}{\hat{U}_{kj}}\left(A\hat{U}\right)_{ij}\right) = \sum_{kj} Q_{kj} \log U_{kj} + \text{const}, \quad (5)$$

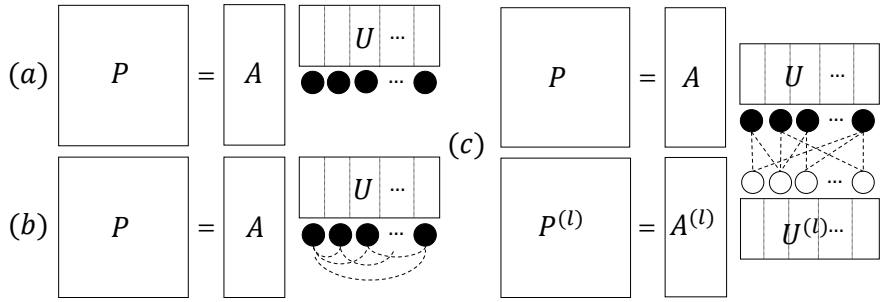

Figure 1: (a) PSLA: vertices (columns of U) has no constrains, (b) GPLSA: vertices with graph regularizer, (c) Multi-Modal GPLSA: vertices among multi-modal data with graph regularizer.

with matrix $Q = \hat{U} \otimes \left[A^T(P \oslash (A\hat{U}))\right]$. With a similar argument, we can show that $\mathcal{G}(U, \hat{U})$ is also a tight lower-bound of $\mathcal{J}(A, U)$ (proof given in the Appendix A), on the basis of which the EM algorithm is obtained. Specifically, each step of the EM algorithm solves

$$\max_U \quad \sum_{kj} Q_{ik} \log U_{kj}, \text{ s.t. } U_{kj} \geqslant 0 \ \ \& \ \ \sum_k U_{kj} = 1, \tag{6}$$

of which the solution is given by $U_{kj} = \frac{Q_{kj}}{\sum_{k'} Q_{k'j}}$. More details of the algorithms can be found in [42]. For completeness, we include a proof of this result in Appendix A.

4. Algorithm for Graph-regularized PLSA

We start with a general setting of the single model GPLSA problem, then extend the formulation to multi-modality problem.

4.1. General Formulation

In this section, we introduce our GPLSA algorithms. More specifically, following the examples in Fig. 1 (a) and (b). Consider a graph with n vertices, where each vertex corresponds to a data point represented by each column of the stochastic matrix U, we use u_i ($i \in [1, n]$) to denote the i-th column of U. For each data point u_i, we add edges between u_i and the other data points to

formulate the relation matrix R among the vertices[1]. The graph regularizers are the divergences between discrete probability distributions given by the corresponding composition of topics from a data entry, which enable effects from entities of the same semantics to be smoothed by each other. Specifically, we formulate GPLSA algorithms as a graph regularizer constrained optimization problem as,

$$\min_{W,H} \sum_{kj} \mathcal{D}_{KL}(P, AU) + \lambda \mathcal{D}_*(U_{kj}, U_{kl}) R_{lj}, \tag{7}$$

which satisfy the constraint that A and U are stochastic matrices. Where the regularization parameter $\lambda > 0$ controls the contribution of the GPLSA objectives of the standard PLSA term and the co-regularization term. In the following, we replace $\mathcal{D}_*$ with the ℓ_1, ℓ_2 and symmetric KL divergences. After dropping irrelevant constant terms of the objective function in Eq.(7), we can simplify the objective as

$$\max_{A,U} \sum_{kj} \mathcal{J}(A, U) - \lambda \mathcal{D}_*(U_{kj}, U_{kl}) R_{lj}, \tag{8}$$

with the same constraints on the factors A and U, with $\mathcal{J}(A, U)$ described in Eq.(2).

Similar to the case of PLSA, in the learning step of GPLSA, the objective function in Eq.(8) is optimized with a block-coordinate descent scheme, by alternating steps between the optimization of A and U while fixing the other factors. We first discuss the EM steps of these sub-problems.

Optimizing A: The optimization of A is the same as the optimization of A in standard PLSA. As such, the solution can be obtained via solving the optimization problem given in Eq.(4).

Optimizing U: The step optimizing U is different because of the graph-regularizer. The optimization of U_{kj} with fixed A and the other columns in U, i.e. U_{kl}, where $l \in n$ and $l \neq j$, after removing irrelevant constant terms,

[1] In this work, we follow the default settings of GNMF using the 0-1 weighting scheme for the relation matrix R, where $R_{jl} = R_{lj}$, $l, j \in n$.

becomes

$$\max_{U} \sum_{kj} \mathcal{J}(A,U) - \lambda \mathcal{D}_*(U_{kj},U_{kl})R_{lj}, \\ \text{s.t. } U_{kj} \geqslant 0 \ \ \& \ \ \sum_k U_{kj} = 1. \tag{9}$$

A tight lower-bound of the above objective function can be derived using the auxiliary function $\mathcal{G}$ defined in Eq.(5), as: $\mathcal{G}(U,\hat{U}) - \lambda \mathcal{D}_*(U_{kj},U_{kl})R_{lj} \leqslant \mathcal{J}(A,U) - \lambda \mathcal{D}_*(U_{kj},U_{kl})R_{lj}$ with equality when $\hat{U}_{kj} = U_{kj}$, which follows from the property of $\mathcal{G}$. We note that the second term $\lambda \mathcal{D}_*(U_{kj},U_{kl})R_{lj}$ does not depend on the auxiliary variable $\hat{U}_{kj}$ in this lower-bound.

Then, we give a similar EM algorithm that optimizes U_{kj} iteratively, which improves the lower-bound in each iteration. Starting with the initial $U_{kj} = U^0_{kj}$, we iteratively solve for

$$U^{(t+1)}_{kj} \leftarrow \operatorname*{argmax}_{U_{kj}} \mathcal{G}(U_{kj},U^{(t)}_{kj}) - \lambda \mathcal{D}_*(U_{kj},U_{kl})R_{lj}, \\ \text{s.t. } U_{kj} \geqslant 0 \ \ \& \ \ \sum_k U_{kj} = 1. \tag{10}$$

We provide efficient algorithms for symmetric KL, ℓ_2 and ℓ_1 divergences with convergence guarantees. Finally, the essential objective function we need to optimize is

$$\max_{U_{kj}} \sum_{kj} Q_{kj} \log U_{kj} - \lambda \mathcal{D}_*(U_{kj},U_{kl})R_{lj}, \\ \text{s.t. } U_{kj} \geqslant 0 \ \ \& \ \ \sum_k U_{kj} = 1. \tag{11}$$

With respect to three types of graph-regularizer, namely, symmetric KL, ℓ_2 and ℓ_1 divergences, the optimal solution to Eq.(11) can be expressed as non-linear functions of a scalar variable η_j, which is the Lagrangian multiplier of the normalizing constraint, $\sum_k U_{kj} = 1$.

Then, we have the solution to each type of graph-regularizer in the following equations (proof given in the Appendix B):

- For $\mathcal{D}_* = \mathcal{D}_{\text{sKL}}$,

$$U_{kj}(\eta_j) = \frac{Q_{kj} + \lambda U_{kl} R_{lj}}{\lambda C \mathcal{W}\Big(\frac{Q_{kj}+\lambda U_{kl}R_{lj}}{\lambda C} \exp(\frac{\eta_j}{\lambda C} - \frac{S_{kj}}{C} + 1)\Big)}, \tag{12}$$

where $\mathcal{W}_0(\cdot)$ is defined implicitly as $x = W(x)e^{W(x)}$ for $x > 0$, which is the principal branch of the Lambert $\mathcal{W}$ function [45][2]. R_{lj} represents the relation weight between document l and j, $\sum_l R_{lj}$ denotes the weight summation between document j and all other documents l, $l \in \{1, n\}$ and $l \neq j$. $(\log U)_{kl}R_{lj}$ and $U_{kl}R_{lj}$ are the constant terms with respect to the current variable, i.e., the j th column of matrix U, we denote $C = \sum_l R_{lj}$, $S_{kj} = (\log U)_{kl}R_{lj}$ for simplicity.

- For $\mathcal{D}_* = \mathcal{D}_{\ell_2}$,

$$U_{kj}(\eta_j) = \frac{\lambda U_{kl}R_{lj} - \eta_j + \sqrt{(\eta_j - \lambda U_{kl}R_{lj})^2 + 4\lambda(\sum_l R_{lj})Q_{kj}}}{2\lambda(\sum_l R_{lj})}, \tag{13}$$

- For $\mathcal{D}_* = \mathcal{D}_{\ell_1}$,

$$U_{kj}(\eta_j) = \begin{cases} \frac{Q_{kj}}{\eta_j + \lambda \sum_l R_{lj}} & -\lambda \sum_l R_{lj} < \eta_j < \frac{Q_{kj}}{U_{kl}} - \lambda \sum_l R_{lj}, \\ U_{kl} & \frac{Q_{kj}}{U_{kl}} - \lambda \sum_l R_{lj} \leq \eta_j \leq \frac{Q_{kj}}{U_{kl}} + \lambda \sum_l R_{lj}, \\ \frac{Q_{kj}}{\eta_j - \lambda \sum_l R_{lj}} & \frac{Q_{kj}}{U_{kl}} + \lambda \sum_l R_{lj} < \eta_j. \end{cases} \tag{14}$$

Compared with the two other divergence types, the update steps for ℓ_1 divergence in Eq.(14) corresponds to a piecewise function. The computation only involves arithmetic operations and thresholding, which is substantially more simple and efficient. Another important property of using ℓ_1 regularizer is that the resulting U_{kj} can have identical components as U_{kl}. This is usually not the case for the ℓ_2 and symmetric KL regularizers.

We use $U_{kj}(\eta_j)$ in Eqs.(12,13,14) to emphasize the fact that they are functions of the scalar parameter η_j. To determine the value of η_j, which in turn leads to the optimal solution to U, we can solve the following 1D nonlinear equation corresponding to the normalization constraint in Eq.(11),

$$\sum_k U_{kj}(\eta_j) = 1. \tag{15}$$

For each type of graph regularizers, we use the corresponding $U_{kj}(\eta_j)$ in Eqs.(12,13,14). For each column index j, Eq.(15) can be solved numerically,

[2]The Lambert $\mathcal{W}$ function has been independently introduced in number of different applications, such as encouraging sparsity over the obtained A or U factors in a variant of PLSA [46], it has also been used in algorithms that enforce entropic priors [47]. It can be numerically evaluated and is provided in popular numerical tools such as MATLAB (function `lambertw`) or SciPy (function `scipy.special.lambertw`).

e.g., with Newton-Raphson when $U_{kj}(\eta_j)$ is differentiable (*e.g.*, $\mathcal{D}_* = \mathcal{D}_{\ell_2}$ or $\mathcal{D}_{\text{sKL}}$) or bi-section when otherwise (*e.g.*, $\mathcal{D}_* = \mathcal{D}_{\ell_1}$).

In summary, we solve the GPLSA problems with an iterative algorithm that alternates between the optimization of individual A and U factors while fixing the others. The optimization of A factor is performed with another iterative EM algorithm based on individual optimization steps given in Eq.(4). The optimization of U factor is achieved by iterating steps that first solve Eq.(15) and then determine the factors with Eq.(12), Eq.(13) or Eq.(14). In practice, all iterative algorithms converges within 5-10 steps, which corresponds to about 20-30 seconds using Matlab implementation running on a desktop with Intel Core(TM) i7 CPU and 8GB RAM.

4.2. Extend GPLSA to Multi-Modal Topic Analysis

Our GPLSA algorithms can be extended to the case with more than one modalities. The overall architecture is shown in Fig. 1 (c). Formally, given a data set with L modalities, we represent it as stochastic matrices $P^{(l)}$ ($l \in 1, ..., L$) of size $d_l \times n$, and aims to find factorization $P^{(l)} \approx A^{(l)}U^{(l)}$, with stochastic matrices $A^{(l)}$ of size $d_{(l)} \times m$ and a matrix $U^{(l)}$ of size $m \times n$ representing the m modality-specific topic matrices and the topic compositions of the dataset, respectively. In GPLSA, association of different modalities to their common data entry is achieved by coupling the factorizations $P^{(l)} \approx A^{(l)}U^{(l)}$, *i.e.*, besides individual GPLSA objectives to each modality. The relation matrix R of the graph regularizers is constructed among $U^{(l)}$ based on the assumption that different data modalities admit similar underlying semantic structure, thus the algorithms aim to minimize the difference of U matrices from different modalities corresponding to each respective topic compositions.

Specifically, GPLSA algorithms are formulated as a constrained optimization problem as

$$\min_{A^{(l)},U^{(l)}} \sum_{l=1,...,L} \mathcal{D}_{KL}(P^{(l)}, A^{(l)}U^{(l)}) + \lambda \mathcal{D}_*(U^{(l)}, U^{(\setminus l)}), \tag{16}$$

which satisfy the constraint that $A^{(l)}$ and $U^{(l)}$ are stochastic matrices, for simplicity, we refer to $U^{(\backslash l)}$ as the other U factor other than $U^{(l)}$. Parameter $\lambda > 0$ balances the contribution of the GPLSA objectives of each modality and the graph-regularization term. Then we replace $\mathcal{D}_*$ with the symmetric KL, ℓ_2 or ℓ_1 divergences[3]. Dropping irrelevant constant terms, the objective function of Eq.(16) can be further simplified to

$$\max_{A^{(l)},U^{(l)}} \sum_{\ell=1,\ldots,L} \mathcal{J}(A^{(l)}, U^{(l)}) - \lambda \mathcal{D}_*(U^{(l)}, U^{(\backslash l)}), \tag{17}$$

with the same constraints on the factors.

Similar to PLSA, in the learning step of GPLSA, the objective function in Eq.(17) is optimized with a block-coordinate descent scheme, by alternating between steps optimizing each of $A^{(l)}$, $U^{(l)}$, while fixing the other factors. In the following, we describe the steps of these sub-problems.

Optimizing $A^{(l)}$: The step optimizing each $A^{(l)}$ is the same as the optimization of A in PLSA. As such, the optimal solution can be obtained via solving a sequence of optimization problem given in Eq.(4).

Optimizing $U^{(l)}$: The optimization of $U^{(l)}$ with fixed $A^{(l)}$ and $U^{(\backslash l)}$, after dropping irrelevant constant terms, we have

$$\begin{aligned} \max_{U^{(l)}} \quad & \sum_{kj} Q_{kj} \log U^{(l)}_{kj} - \lambda \mathcal{D}_*(U^{(l)}, U^{(\backslash l)}), \\ & \text{s.t.} \;\; U^{(l)}_{kj} \geqslant 0 \;\; \& \;\; \sum_k U^{(l)}_{kj} = 1. \end{aligned} \tag{18}$$

Similar to GPLSA on single modality, with respect to three types of graph-regularizer for difference modality (namely, symmetric KL, ℓ_2 and ℓ_1 divergences), the optimal solution to Eq.(18) can be expressed as non-linear functions of a scalar variable η_j, which is referred to as a Lagrangian multiplier of the normalizing constraint, $\sum_k U_{kj} = 1$. Specifically, these solutions are given in the following equations, with proofs given in the Appendix C:

- For $\mathcal{D}_* = \mathcal{D}_{\text{sKL}}$,

$$U^{(l)}_{kj}(\eta_j) = \frac{Q_{kj} + \lambda U^{(\backslash l)}_{kj}}{\lambda \mathcal{W}_0\left(\frac{Q_{kj}+\lambda U^{(\backslash l)}_{kj}}{\lambda U^{(\backslash l)}_{kj}} \exp\left(1+\frac{\eta_j}{\lambda}\right)\right)}, \tag{19}$$

[3] One could have other regularization terms on the factors $A^{(l)}$ and $U^{(l)}$ for other preference on the factors such as sparsity. Furthermore, it is also possible to use similar methods to enforce consistencies in parts of factor $A^{(l)}$. However, for simplicity, in the current work we do not consider these types of regularizers.

- For $\mathcal{D}_* = \mathcal{D}_{\ell_2}$,

$$U_{kj}^{(l)}(\eta_j) = \frac{1}{2}\sqrt{\left(U_{kj}^{(\backslash l)} - \frac{\eta_j}{\lambda}\right)^2 + \frac{4Q_{kj}}{\lambda}} + \frac{1}{2}\left(U_{kj}^{(\backslash l)} - \frac{\eta_j}{\lambda}\right), \tag{20}$$

- For $\mathcal{D}_* = \mathcal{D}_{\ell_1}$,

$$U_{kj}^{(l)}(\eta_j) = \begin{cases} \frac{Q_{kj}}{\eta_j + \lambda}, & -\lambda < \eta_j < \frac{Q_{kj}}{U_{kj}^{(\backslash l)}} - \lambda, \\ U_{kj}^{(\backslash l)}, & \frac{Q_{kj}}{U_{kj}^{(\backslash l)}} - \lambda \leqslant \eta_j \leqslant \frac{Q_{kj}}{U_{kj}^{(\backslash l)}} + \lambda, \\ \frac{Q_{kj}}{\eta_j - \lambda}, & \frac{Q_{kj}}{U_{kj}^{(\backslash l)}} + \lambda < \eta_j. \end{cases} \tag{21}$$

5. Experiments

In this section, we first perform some simple experiment to evaluate the GPLSA algorithms on image clustering to justify how well they perform on topic analysis from data of a single modality. We then apply them to cross-modality retrieval from documents containing both images and texts. Finally, we also apply them on multi-lingual topic analysis problems. An overview of the proposed framework for experiments is shown in Fig.2.

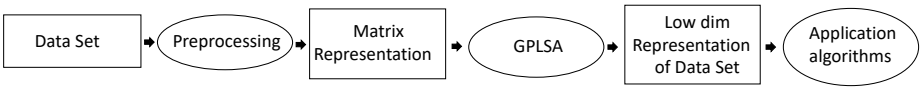

Figure 2: *An overview of the proposed framework.*

5.1. Image Clustering

We use two data sets in this experiment, the first one is the CMU Pose, Illumination, and Expression (PIE) face database [48], which contains of 32 × 32 gray scale facial images of 68 persons. There are 42 facial images taken for each person under different poses, illumination and expression conditions. The second one is the Columbia Image Library (COIL-20) data set [24], which contains 32 × 32 gray scale images of 20 classes (objects). There are 72 images taken for each object with different view angles.

	Accuracy (%)				Normalized Mutual Information (%)			
K	Baseline	Regularizer	GNMF [24]	GPLSA	Baseline	Regularizer	GNMF[24]	GPLSA
30	26.83	L1	—	**38.81**	48.38	L1	—	**57.25**
		L2	82.86	**84.44**		L2	90.49	**91.19**
		KL	61.51	**76.27**		KL	78.11	**88.12**
50	25.48	L1	—	**39.43**	50.93	L1	—	**62.89**
		L2	76.14	**80.48**		L2	89.26	**90.22**
		KL	62.57	**70.14**		KL	81.20	**86.28**
68	23.99	L1	—	**35.75**	53.77	L1	—	**63.56**
		L2	75.56	**77.66**		L2	88.07	**89.14**
		KL	63.55	**76.23**		KL	83.20	**89.16**

Table 1: *Comparison of results regarding clustering performance of GPLSA* vs. *GNMF [24] on the PIE face dataset in terms of accuracy and normalized mutual information.*

The clustering results are evaluated by two metrics, the accuracy and the normalized mutual information metric (NMI) [49]. The accuracy is defined as:

$$AC = \frac{\sum_{i=1}^{n} \delta(r_i, s_i)}{n}, \tag{22}$$

where r_i is the estimated cluster label, s_i is the ground truth label, and n is the number of examples. $\delta(r_i, s_i)$ is the delta function that equals 1 if $r_i = s_i$ and equals 0 otherwise. The NMI are defined as follow:

$$NMI(C, \bar{C}) = \frac{MI(C, \bar{C})}{max(H(C), H(\bar{C}))}, \tag{23}$$

where $H(C)$ and $H(\bar{C})$ are the entropies of $H(C)$ and $H(\bar{C})$, the mutual information (MI) is defined as:

$$MI(C, \bar{C}) = \sum_{c_i \in C, \bar{c}_j \in \bar{C}} p(c_i, \bar{c}_j) \log_2 \frac{p(c_i, \bar{c}_j)}{p(c_i)p(\bar{c}_j)}, \tag{24}$$

where C is the set of ground truth clusters, $\bar{C}$ is the clustering result from the testing algorithm. $p(c_i)$ and $p(\bar{c}_j)$ are the probabilities of a randomly selected image, which belongs to clusters c_i and $\bar{c}_j$, respectively. $p(c_i, \bar{c}_j)$ is the joint probability that this randomly selected image belongs to the clusters c_i and $\bar{c}_j$ at the same time. In case two sets of clusters are identical, NMI is 1. In case two sets are independent, NMI is 0.

We compare GPLSA with (1) a baseline algorithm with K-means clustering in the original image space and (2) the GNMF method [24], which has been shown to achieve superior performance against classic clustering algorithms (i.e. K-means, SVD, NCut, standard NMF)[4]. After applying GPLSA and GNMF, a low dimensional representation with the size of topics numbers for each image is obtained. The clustering of images is then performed upon this low dimensional representation using standard K-means for a final evaluation. We follow the default settings of GNMF using the 0-1 weighting scheme and use 5 nearest neighbors for the relation matrix R; and we set the number of topics equals to the number of clusters, the optimized balance parameter λ in GPLSA algorithms is chosen by cross-validation on a subset of the training data. We evaluate the performance with different number of clusters (K in Table 2), where the evaluated clusters (classes) are randomly chosen from 20 classes (objects).

Table 1 and 2 shows the results of GPLSA algorithms with ℓ_1, ℓ_2 and symmetric KL regularizers on the PIE and COIL-20 dataset, respectively. The results using the ℓ_1 regularizer consistently outperform the baseline. We note that the use of ℓ_1 regularizer in GNMF is not available, thus the comparison is omitted. GPLSA outperforms GNFM in both cases of symmetric KL and ℓ_2 regularizers. Finally, the symmetric KL graph-regularizer achieves the best overall performance. This is expectable as we solve the objective function with the log term using the Lambert $\mathcal{W}$ function, and in comparison GNMF relies on a rough approximate to solve a nonlinear equation.

5.2. Cross-Modal Image/Text Retrieval

For multi-modality data, the difference in original data representations of images and text are encoded into the corresponding topics with the GPLSA model. And with their topic compositions, images and texts are projected into a compatible semantic space, this new representation can then be used

[4]The GNMF code and dataset are released and available at http://www.cad.zju.edu.cn/home/dengcai/Data/GNMF.html

	Accuracy (%)				Normalized Mutual Information (%)			
K	Baseline	Regularizer	GNMF [24]	GPLSA	Baseline	Regularizer	GNMF[24]	GPLSA
10	43.89	L1	—	**60.00**	54.14	L1	—	**65.95**
		L2	79.58	**85.00**		L2	87.92	**88.62**
		KL	84.31	**87.36**		KL	88.44	**90.18**
15	65.37	L1	—	**67.87**	71.08	L1	—	**76.54**
		L2	84.35	**85.83**		L2	85.31	**86.22**
		KL	84.17	**86.39**		KL	87.60	**89.18**
20	60.49	L1	—	**73.68**	73.86	L1	—	**81.69**
		L2	72.22	**80.97**		L2	87.60	**88.36**
		KL	73.68	**81.81**		KL	85.18	**90.14**

Table 2: *Comparison of results regarding clustering performance of GPLSA* vs. *GNMF [24] on the COIL20 dataset in terms of accuracy and normalized mutual information.*

to do establish connections between images and text documents and facilitates cross-modality retrieval. Specifically, our evaluation focused on the task of text retrieval from an image query (*i-2-t*), and image retrieval from a query with a text document (*t-2-i*). Two standard benchmark image/text datasets were used in our experiments: *Wikipedia* [38] and *TVGraz* [50]. The *Wikipedia* dataset consists of 2866 image/text pairs of 30 semantic categories, includes a standard training and testing sets split with 2173 and 693 image/text pairs. The *TVGraz* dataset consists of 2058 image/text pairs of 10 semantic categories with an average document length of 289 words, and is split into training and testing sets with 1558 and 500 image/text pairs. Images and texts in both datasets are converted to bag-of-word representation, where for images we used 1024 visual keywords as a result of clustering the SIFT features from all training images, and 6203 unique text words were selected after stemming and removal of the common stop words.

In the learning phase, we determine modality-specific topics using the GPLSA learning algorithm (Section 4.2) on the training sets. We extract 100 topics from the *Wikipedia* dataset and 50 topics from the *TVGraz* dataset, and the optimized balance parameter λ in GPLSA algorithms is chosen by cross-validation on a subset of the training data. When performing retrieval tasks on the testing

Methods	Topic space similarity				Semantic space similarity			
	TVGraz		Wikipedia		TVGraz		Wikipedia	
	i-2-t	t-2-i	i-2-t	t-2-i	i-2-t	t-2-i	i-2-t	t-2-i
SCM [37]	0.460	0.450	0.267	0.219	0.664	0.649	0.362	0.273
Link PLSA [34]	0.349	0.349	0.247	0.247	0.803	0.803	0.605	0.605
ℓ_1 GPLSA	0.359	0.365	0.317	0.307	0.723	0.726	0.667	0.658
ℓ_2 GPLSA	0.450	0.445	0.360	0.358	0.846	0.845	0.706	0.701
sKL GPLSA	**0.481**	**0.481**	**0.413**	**0.413**	**0.850**	**0.850**	**0.726**	**0.724**

Table 3: *Performance comparison of GPLSA with 3 regularizers and two other multi-modal learning algorithms in terms of mean average precision (mAP) on two standard public image/text benchmark datasets. See texts for details.*

set, we first recover the topic composition of the queried image or text using the PLSA algorithm (only the optimization of U matrix) using the learned modality-specific topics. The similarities between topic compositions of the queried image and texts in testing set (in task *i-2-t*) or queried text and images in testing set (in task *t-2-i*) are then evaluated and ranked. Two similarity measures are used in our experiments, the centered normalized correlation between the topic compositions and the centered normalized correlations between topic compositions transformed by a multi-class logistic regression function learned during training, which maps topic compositions to the semantic categories pre-defined for each dataset. As such, the former evaluates correlations of topic compositions directly, while the latter can be regarded as the correlation in a more semantically meaningful space induced from the topic compositions [37]. We use the mean average precision (MAP) scores over all testing data as performance metric. The average precision score for each query is computed as the mean precision value for the top 10 relevant retrievals. Here, we determine a relevant retrieval occurs if the retrieved text/image is from the same semantic category as the image/text used in query.

Table 3 shows the results of GPLSA algorithms with ℓ_1, ℓ_2 and symmetric KL regularizers on the two datasets. For comparison, we also include retrieval performance based on a link-PLSA model that requires the topic compositions of associated text and image to be identical. The link-PLSA algorithm can be

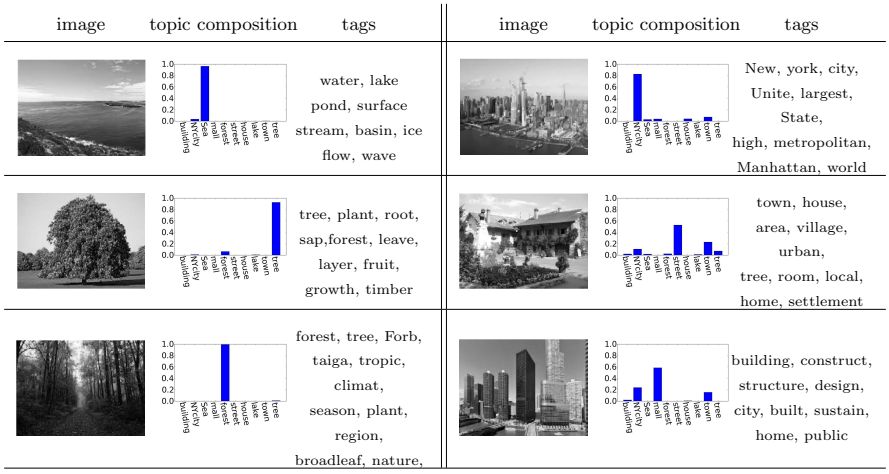

Figure 3: Example images with generated topic compositions and tags obtained with sKL-GPLSA from the Wikipedia dataset.

implemented as described in [34]. Furthermore, all results were compared to a baseline established by the method of *semantic correlation matching* (SCM) [37], which represents the state-of-the-art performance in text/image cross-modal retrieval tasks. Results in Table 3 suggests that for the two cross-modal retrieval tasks, GPLSA algorithms in general achieve better performance than the link PLSA algorithm, and also outperform the SCM method that is based on kernel canonical correlation analysis. This may be attributed to, on the one hand, the more semantic relevance of the representation (as probability mixture of thematic topics of the images/text) obtained with GPLSA, and on the other hand, its less restrict assumption that allows for mis-match of topic compositions of associated text and images. This is further corroborated by observing that the MAP scores for *i2t* and *t2i* tasks are similar with GPLSA algorithms, suggesting the diminished representational difference between the two modalities in the topic space found by GPLSA. Furthermore, all three variants of the GPLSA algorithms achieves better performance and efficient computation, but symmetric KL graph-regularizer leads to the best overall performance. Last, combining with more semantic abstraction, as concluded in [37], can also significantly improve the retrieval performance.

Topic 1 - (Turkey nuclear)		Topic 2 - (Europe peace)	
German	English	German	English
turkei (turkey)	Turkey	israel (israel)	peace
frag (question)	nuclear	europa (Europe)	Israel
prasident (president)	question	palastinens (palestinians)	Palestinian
moglich (possible)	President	union (union)	European
kommission (commission)	people	staat (State)	negoti
antwort (answer)	Commission	region (region)	territory
nuklear (nuclear)	concern	ost (east)	Union
fall (event)	time	friedensprozess (peace process)	East
zeit (time)	Council	abkomm (Agreem)	agreement
bedenk (Bedenk)	possibl	gebiet (area)	State
Topic 3 - (fishing environment)		Topic 4 - (vehicle)	
German	English	German	English
schiff (ship)	fish	herstell (producible)	car
fischerei (fishing)	disaster	fahrzeug (vehicle)	manufacture
katastroph (disaster)	fisher	kost (costly)	cost
erika (heather)	affect	autos (cars)	recycle
fisch (fish)	pollute	verantwort (responsible)	vehicle
europa (Europe)	ship	standpunkt (viewpoint)	environment
schad (defective)	control	gemeinsam (common)	industry
kontroll (control)	damage	automobilindustri (automotive-industrial)	common
umwelt (environment)	sea	recycling (recycling)	consume
fischereisektor (fishing sector)	environment	verbrauch (consumption)	response

Table 4: Top words of leaned topics on the German-English corpus. See text for details.

In Fig.3 we further show several test images from the *Wikipedia* dataset with their corresponding topic compositions over a subset of topics obtained with the symmetric KL GPLSA algorithm (the names of each topic is manually assigned based on the top words from each topic to facilitate understanding), together with text tags that are generated by sampling from the topic mixtures associated with each image. The visualized topic compositions and the generated text tags of these images obtained with GPLSA span wide semantic ranges, and can shed some light on their effects in improving the precisions of semantic matching with the queried text document.

5.3. Multi-Lingual Topic Analysis

In multi-lingual topic analysis, our purpose it to produce a topic-level summarization of documents using GPLSA in different languages. From such a summarization, one can quickly grasp the basic subjects concerning a document

English document: The UN Economic Commission for Europe has also examined the objectives for reductions in the same sources of emissions as in the proposal for a directive now under discussion, and, as a result of these talks, the so-called Gothenburg Protocol was signed. There is a clear difference between this proposal and the Commission's. Mr President, the industry committee, after much discussion and serious consideration, is overwhelmingly opposed to the Commission's proposed ceilings, and this is across all groups and nationalities. I understand that the groups, the whole Parliament, are split on this issue, between supporters of the Committee on the Environment, Public Health and Consumer Policy and supporters of the industry committee's line.
Summarization in German: europa, kommission, union, wirtschaft, parlament, bericht, land, frag, wichtig, prasident, moglich, polit, ziel, mitgliedstaat, sozial
German document: In Kreisen der UNO-Wirtschaftskommission fr Europa sind ebenfalls Ziele fr die Verringerung derselben Emissionsquellen untersucht worden wie in dem jetzt debattierten Vorschlag fr eine Richtlinie, und im Ergebnis dieser Verhandlungen wurde das sogenannte Gteborger Protokoll unterzeichnet. Zwischen diesem Vorschlag und dem Vorschlag der Kommission gibt es einen deutlichen Unterschied. Herr Prsident, der Industrieausschu lehnt nach ausfhrlicher Diskussion und ernsthafter berlegung die von der Kommission vorgeschlagenen Hchstgrenzen mit der berwltigenden Mehrheit aller Fraktionen und Nationalitten ab. Soweit mir bekannt ist, sind die Fraktionen, ist das gesamte Parlament in zwei Lager gespalten, von denen das eine den Ausschu fr Umweltfragen, Volksgesundheit und Verbraucherpolitik untersttzt und das andere den Industrieausschu.
Summarization in English: European, commission, policy, President, Union, service, region, import, report, social, Member, develop, State, country, economic

Table 5: *German-English document summarization results. See text for details.*

written in unknown language by another familiar language. As such analysis bypasses the need of machine translation, it can be used in occasions where fast analysis of a vast number of documents in foreign languages is required.

We used a data set that is a subset of the multilingual corpora in the European Parliament Proceedings Parallel Corpus (EPPPC) [51], which contains newswire articles written in different European languages with aligned sentences. In our experiments, we selected 1100 documents of German $\Leftrightarrow$ English pairs. Considering different languages as different information sources, topics are learned from a training set containing 1000 German-English documents. Performance of the topic learning algorithms are evaluated on a test set of the remaining 100 documents. After stemming and removing a standard list of the

stop words using NLTK[5], the resulting English dictionary contains 10115 words, while the German dictionary contains 19532 words.

We then applied the GPLSA algorithm on this corpus to extract 100 common topics in the two languages. Table 4 shows the top words of two languages from 4 examples of the learned topics (the German words are shown with their English translations in parentheses). As these result shows, common topics capture the correspondence of German and English words without precisely aligning them. More importantly, these words are grouped together under topics of the same concept.

As a simple application, we generate English summarization for German documents that were not covered by the training set. Similar to the previous experiments on image-text documents, we take the simple method of first recovering the topic assignments of the German documents using the learned German topic matrix, then combine it with the English topic matrix to generate corresponding keywords in English. Two examples from the results are shown in Table 5. For readers who do not know German, the extracted keywords in English can provide informative guidelines about the topics of the document (in this case, both are about 'European politics').

6. Conclusion

We have presented efficient algorithms for graph regularized PLSA (GPLSA) to probabilistic topic analysis of both single- and multiple- modality data representation. In GPLSA, topic compositions of a data entity are mapped to a graph and the similarities between topic compositions on the graph are measured with divergences between discrete probabilities. We propose efficient multiplicative iterative algorithms for GPLSA with ℓ_1, ℓ_2 and symmetric KL divergences as regularizers. The optimization problem for each case affords simple numerical solutions that require only matrix arithmetic operations and 1D nonlinear

[5] http://www.nltk.org/

equations. Experimental results in various real-data sets show that the proposed algorithms enhances the performance of the sate-of-the art frameworks. Unfortunately, We have not proven if the algorithm converges to a limit point in the interior of the feasible region, that this point is indeed a stationary point [52], we will investigate more on this direction in the future work.

There are several directions that the current work can be further improved. First, the correlation matrix controls the smoothness of topics in GPLSA model. Thus, learning a suitable correlation matrix is critical to GPLSA algorithms. Secondly, we are working on adapting the GPLSA algorithms to datasets with more sophisticated structures over topics, such as allowing a hierarchical structure of the topics with higher layers capturing more abstract semantic notions. At last, we are also interested in incorporating other type of constraints such as sparseness in multi-modal topic analysis. We will also seek more applications of GPLSA algorithms, for instance to video analysis or multi-modal social data analysis.

Acknowledgment. This material is based upon work supported by the National Science Foundation under Grant No. (IIS-1537257).

Appendix A.

Proof. **Proof of $\mathcal{J}(A,U)$ is tight lower-bounded by $\mathcal{F}(A,\hat{A})$.** First, note that $\left(\hat{A}U\right)_{ij} = \sum_k \hat{A}_{ik}U_{kj}$, or $\sum_k \frac{\hat{A}_{ik}U_{kj}}{\left(\hat{A}U\right)_{ij}} = 1$. Next, consider the concavity of the logarithm function, we first rearrange terms in the definition of $\mathcal{F}(A,\hat{A})$ and then apply Jensen's inequality to the inner term to get

$$\begin{aligned}
&\sum_{ij} V_{ij} \left\{ \sum_k \frac{\hat{W}_{ik}H_{kj}}{\left(\hat{W}H\right)_{ij}} \log\left(\frac{W_{ik}H_{kj}}{\hat{W}_{ik}H_{kj}} \left(\hat{W}H\right)_{ij} \right) \right\} \\
&\leqslant \sum_{ij} V_{ij} \log\left(\sum_k W_{ik}H_{kj} \right) = \sum_{ij} V_{ij} \log(WH)_{ij}.
\end{aligned}$$

As the last term is $\mathcal{J}(A,U)$, this proves the inequalities $\mathcal{F}(A,\hat{A}) \leqslant \mathcal{J}(A,U)$. Furthermore, equality trivially holds if we have $A = \hat{A}$.

Next, we show the other part of Eq.(3), $\mathcal{F}(A,\hat{A}) = \sum_{ik} M_{ik} \log A_{ik} + \text{const.}$ We start with the definition of $\mathcal{F}(A,\hat{A})$, as

$$\sum_{ijk} \frac{P_{ij}\hat{A}_{ik}U_{kj}}{\left(\hat{A}U\right)_{ij}} \log\left(\frac{A_{ik}}{\hat{A}_{ik}}\left(\hat{A}U\right)_{ij}\right) \sum_{ik}\hat{A}_{ik}\sum_{j}\left\{\frac{P_{ij}}{\left(\hat{A}U\right)_{ij}}(U^T)_{jk}\right\} \log A_{ik} + \text{const.}$$

where $\sum_j \left\{\frac{P_{ij}}{\left(\hat{A}U\right)_{ij}}(U^T)_{jk}\right\} = \left[(P \oslash (\hat{A}U))U^T\right]_{ik}$, it then follows that the term in front of $\log A_{ik}$ is the element of matrix $M = \hat{A} \otimes \left[(P \oslash (\hat{A}U))U^T\right]$.

Optimal solution to Eq.(4). We first introduce Lagrangian multiplier for each equality constraint η_k in the optimization problem, and form the Lagrangian as:

$$\sum_{ik} M_{ik} \log A_{ik} - \sum_k \eta_k \left(\sum_i A_{ik} - 1\right).$$

Taking derivative of the Lagrangian with regards to each A_{ik} and solving the equation when setting the result to zero yield $A_{ik} = \frac{M_{ik}}{\eta_k}$. Further considering the constraint $\sum_{i'} A_{i'k} = 1$, we have $\eta_k = \sum_{i'} M_{i'k}$, thus proves the result.

Proof of $\mathcal{J}(A,U)$ is tight lower-bounded by $\mathcal{G}(U,\hat{U})$. As in the case of showing $\mathcal{F}(A,\hat{A}) \leqslant \mathcal{J}(A,U)$, we first use the fact that $\left(A\hat{U}\right)_{ij} = \sum_k A_{ik}\hat{U}_{kj}$, or $\sum_k \frac{A_{ik}\hat{U}_{kj}}{\left(A\hat{U}\right)_{ij}} = 1$. We then rearrange terms of $\mathcal{G}(U,\hat{U})$ and then apply Jensen's inequality to obtain

$$\begin{aligned}
&\sum_{ij} V_{ij}\left\{\sum_k \frac{W_{ik}\hat{H}_{kj}}{\left(W\hat{H}\right)_{ij}} \log\left(\frac{W_{ik}H_{kj}}{W_{ik}\hat{H}_{kj}}\left(W\hat{H}\right)_{ij}\right)\right\} \\
&\leqslant \sum_{ij} V_{ij} \log\left(\sum_k W_{ik}H_{kj}\right) = \sum_{ij} V_{ij}\log(WH)_{ij}.
\end{aligned}$$

As the last term is $\mathcal{J}(A,U)$, this proves that $\mathcal{G}(U,\hat{U}) \leqslant \mathcal{J}(A,U)$. Furthermore, equality is trivially held when we have $U = \hat{U}$.

Next, we show the other part of Eq.(5), $\mathcal{G}(U,\hat{U}) = \sum_{kj} Q_{kj} \log U_{kj} + \text{const.}$

We start with the definition of $\mathcal{G}(U, \hat{U})$, as

$$\sum_{ijk} \frac{V_{ij} W_{ik} \hat{H}_{kj}}{\left(W\hat{H}\right)_{ij}} \log \left(\frac{H_{kj}}{\hat{H}_{kj}} \left(W\hat{H}\right)_{ij} \right)$$
$$= \sum_{kj} \hat{H}_{kj} \sum_{i} \left\{ (W^T)_{ki} \frac{V_{ij}}{\left(W\hat{H}\right)_{ij}} \right\} \log H_{kj} + \text{const.}$$

Where $\sum_i \left\{ (A^T)_{ki} \frac{P_{ij}}{\left(A\hat{U}\right)_{ij}} \right\} = \left[A^T (P \oslash (A\hat{U})) \right]_{kj}$, it then follows that the term in front of $\log U_{kj}$ is the element of matrix that can be written as $Q = \hat{U} \otimes \left[A^T (P \oslash (A\hat{U})) \right]$.

Optimal solution to Eq.(6). We first introduce Lagrangian multiplier for each equality constraint η_j in the optimization problem, and form the Lagrangian as:

$$\textstyle\sum_{kj} Q_{kj} \log U_{kj} - \sum_j \eta_j \left(\sum_k U_{kj} - 1 \right).$$

Taking derivative of the Lagrangian with regards to each U_{kj} and solving the equation when setting the result to zero yield $U_{kj} = \frac{Q_{kj}}{\eta_j}$. Further considering the constraint $\sum_{k'j} U_{k'j} = 1$, we have $\eta_j = \sum_{k'} Q_{k'j}$, thus proves the result.

Appendix B.

Proof. **Proof of Eq.(12)**. We first simplify the objective function of Eq.(11) by dropping irrelevant constant terms to obtain the new objective function as

$$\sum_{kj} Q_{kj} \log U_{kj} - \lambda \sum_{j=1}^{n} \sum_{l=1}^{n} \sum_{k=1}^{r} \left(U_{kj} \log \frac{U_{kj}}{U_{kl}} + U_{kl} \log \frac{U_{kl}}{U_{kj}} \right) R_{lj}.$$

Next, we introduce Lagrangian multiplier η_j to each equality constraint on the column of U_{kj} in Eq.(11) and form the Lagrangian. Ignoring the nonnegative constraint (it will be shown to be satisfied in our setup later in this section), the first KKT condition of the problem with regards to U_{kj} is that the derivative of the Lagrangian with regards to U_{kj} vanishes, as

$$\frac{Q_{kj}}{U_{kj}} - \eta_j - \lambda((\sum_l R_{lj}) \log U_{kj} + (\sum_l R_{lj}) - (\log U)_{kl} R_{lj} - \frac{U_{kl} R_{lj}}{U_{kj}}) = 0.$$

We define $C = \sum_l R_{lj}$, $S_{kj} = (\log H)_{kl} R_{lj}$, $T = \frac{Q_{kj}+\lambda S_{kj}}{\lambda C}$ and $B = \frac{\eta_j}{\lambda C} - \frac{S_{kj}}{C} + 1$, and rearrange term followed by exponentiation of both sides of the above equation to obtain

$$\frac{T}{U_{kj}} = \log U_{kj} + B \Rightarrow T e^B = \exp\big(\log\big(U_{kj} e^B\big)\big) \log\big(U_{kj} e^B\big).$$

Using the definition of the principal branch of the Lambert $\mathcal{W}$-function, this further reduces to

$$\begin{aligned} \log H_{kj} e^B =& \mathcal{W}_0\left(A e^B\right) \Rightarrow H_{kj} e^B = e^{\mathcal{W}_0\left(A e^B\right)} \Rightarrow \\ & H_{kj} e^B = \frac{A e^B}{\mathcal{W}_0\left(A e^B\right)} \Rightarrow H_{kj} = \frac{A}{\mathcal{W}_0\left(A e^B\right)}, \end{aligned}$$

where in the last step we use the property of $\mathcal{W}$ function that $e^{\mathcal{W}_0(z)} = \frac{z}{\mathcal{W}_0(z)}$.

Now replacing T and B with their definitions yields Eq.(12).

Proof of Eq.(13). With the ℓ_2 regularizer, the objective function of Eq.(11) becomes

$$\sum_{kj} Q_{kj} \log U_{kj} - \lambda \sum_{j=1}^{n} \sum_{l=1}^{n} \sum_{k=1}^{r} \frac{1}{2} \left(U_{kj} - U_{kl}\right)^2 R_{lj}. \tag{B.1}$$

Similar to the symmetric KL divergence case, we form Lagrangian with Lagrangian multiplier η_j to each equality constraint of Eq.(11), and take its derivative with regards to U_{kj} and set it to zero,

$$\frac{Q_{kj}}{U_{kj}} - \eta_j - \lambda \left[(\sum_l R_{lj}) U_{kj} - U_{kl} R_{lj} \right] = 0, \tag{B.2}$$

rearranging terms leads to a quadratic equation, as

$$\lambda (\sum_l R_{lj}) U_{kj}^2 + \left(\eta_j - \lambda (U_{kl} R_{lj})\right) U_{kj} - Q_{kj} = 0, \tag{B.3}$$

where the positive root of this equation is given by Eq.(13), and the other root is negative.

Proof of Eq.(14). For the ℓ_1 divergence regularizer, the objective function of Eq.(11) becomes

$$\sum_{kj} Q_{kj} \log U_{kj} - \lambda \sum_{j=1}^{n} \sum_{l=1}^{n} \sum_{k=1}^{r} \left(|U_{kj} - U_{kl}|\right) R_{lj}. \tag{B.4}$$

First we form the Lagrangian by introducing a multiplier η_j for each column of U_{kj}. Since the resulting Lagrangian is separable for each column of matrix U_{kj}, we focus on terms that are relevant to one element U_{kj}, which is

$$\sum_k Q_{kj} \log U_{kj} - \lambda \sum_{j=1}^{n} \sum_{l=1}^{n} \sum_{k=1}^{r} \left(|U_{kj} - U_{kl}| \right) R_{lj} - \eta_j \left(\sum_k U_{kj} - 1 \right).$$

After rearranging terms to remove the absolute value and dropping constants, we have

$$\sum_k Q_{kj} \log U_{kj} - \lambda \sum_{U_{kj} < U_{kl}} (U_{kj} - U_{kl}) - \lambda \sum_{U_{kl} < U_{kj}} (U_{kl} - U_{kj}) - \eta_j \sum_k U_{kj}.$$

Next, we take derivative with regards to this function, which will be in two cases. For $U_{kj} > U_{kl}$, setting the derivative with regards to U_{kj} becomes $\frac{Q_{kj}}{U_{kj}} - \lambda \sum_l R_{lj} - \eta_j = 0 \Rightarrow U_{kj} = \frac{Q_{kj}}{\eta_j + \lambda \sum_l R_{lj}}$. This holds when $U_{kj} = \frac{Q_{kj}}{\eta_j + \lambda \sum_l R_{lj}} > U_{kl} \Rightarrow \frac{Q_{kj}}{U_{kl}} - \lambda > \eta_j \sum_l R_{lj}$. To make U_{kj} nonnegative, it is easy to see that $\eta_j > -\lambda \sum_l R_{lj}$. For $U_{kj} < U_{kl}$, setting the derivative with regards to U_{kj} becomes $\frac{Q_{kj}}{U_{kj}} + \lambda \sum_l R_{lj} - \eta_j = 0 \Rightarrow U_{kj} = \frac{Q_{kj}}{\eta_j - \lambda \sum_l R_{lj}}$. This holds when $U_{kj} = \frac{Q_{kj}}{\eta_j - \lambda \sum_l R_{lj}} < U_{kl} \Rightarrow \frac{Q_{kj}}{U_{kl}} + \lambda \sum_l R_{lj} < \eta_j$. When these two conditions are not satisfied, the optimal solution is given by setting $U_{kj} = U_{kl}$. Combining all these results yields Eq.(14). □

Appendix C.

Proof of Eq.(19). We first simplify the objective function of Eq.(18) by dropping irrelevant constant terms to obtain the new objective function as

$$\sum_{kj} Q_{kj} \log U_{kj}^{(l)} - \lambda \sum_{kj} \left(U_{kj}^{(l)} \log \frac{U_{kj}^{(l)}}{U_{jk}^{(\backslash l)}} - U_{jk}^{(\backslash l)} \log U_{kj}^{(l)} \right)$$

Next, we introduce Lagrangian multiplier η_j to each equality constraint on the column of $U^{(l)}$ in Eq.(18) and form the Lagrangian. Ignoring the nonnegative constraint (it will be shown to be satisfied automatically later), the first KKT condition of the problem with regards to $U_{kj}^{(l)}$ is that the derivative of the Lagrangian with regards to $U_{kj}^{(l)}$ vanishes, as

$$\frac{Q_{kj} + \lambda U_{kj}^{(\backslash l)}}{U_{kj}^{(l)}} + \lambda \log \frac{U_{kj}^{(\backslash l)}}{U_{kj}^{(l)}} - \lambda - \eta_j = 0.$$

We define $T = \frac{Q_{kj}}{\lambda} + U_{kj}^{(\backslash l)}$ and $B = 1 - \log U_{kj}^{(\backslash l)} + \frac{\eta_j}{\lambda}$, and rearrange term followed by exponentiation of both sides of the above equation to obtain

$$\frac{T}{U_{kj}^{(l)}} = \log U_{kj}^{(l)} + B \Rightarrow Te^B = \exp\left(\log\left(U_{kj}^{(l)} e^B\right)\right) \log\left(U_{kj}^{(l)} e^B\right)$$

Using the definition of the principal branch of the Lambert $\mathcal{W}$-function, this further reduces to

$$\log H_{kj}^{(l)} e^B = \mathcal{W}_0\left(Ae^B\right) \Rightarrow H_{kj}^{(l)} e^B = e^{\mathcal{W}_0\left(Ae^B\right)} \Rightarrow$$
$$H_{kj}^{(l)} e^B = \frac{Ae^B}{\mathcal{W}_0\left(Ae^B\right)} \Rightarrow H_{kj}^{(l)} = \frac{A}{\mathcal{W}_0\left(Ae^B\right)},$$

where in the last step we use the property of $\mathcal{W}$ function that $e^{\mathcal{W}_0(z)} = \frac{z}{\mathcal{W}_0(z)}$. Now replacing T and B with their definitions yields Eq.(19).

Proof of Eq.(20). With the ℓ_2 co-regularizer, the objective function of Eq.(18) becomes

$$\sum_{kj} Q_{kj} \log U_{kj}^{(l)} - \lambda \sum_{kj} \frac{1}{2} \left(U_{kj}^{(l)} - U_{kj}^{(\backslash l)}\right)^2$$

Similar to the symmetric KL divergence case, we form Lagrangian with Lagrangian multiplier η_j to each equality constraint of Eq.(18), and take its derivative with regards to $U_{kj}^{(l)}$ and set it to zero,

$$\frac{Q_{kj}}{U_{kj}^{(l)}} - \lambda(U_{kj}^{(l)} - U_{kj}^{(\backslash l)}) - \eta_j = 0.$$

Rearranging terms, this leads to a quadratic equation

$$\lambda \left(U_{kj}^{(l)}\right)^2 - \left(\lambda U_{kj}^{(\backslash l)} - \eta_j\right) U_{kj}^{(l)} - Q_{kj} = 0,$$

the positive root of which is given by Eq.(20) (the other root is negative).

Proof of Eq.(21). For the ℓ_1 divergence co-regularizer, the objective function of Eq.(18) becomes

$$\sum_{kj} Q_{kj} \log U_{kj}^{(l)} - \lambda \sum_{kj} \left|U_{kj}^{(l)} - U_{kj}^{(\backslash l)}\right|.$$

First we form the Lagrangian by introducing a multiplier η_j for each column of $U^{(l)}$. Since the resulting Lagrangian is separable for each column of matrix $U^{(l)}$, we focus on terms that are relevant to one element $U_{kj}^{(l)}$, which is

$$\sum_k Q_{kj} \log U_{kj}^{(l)} - \lambda \sum_k \left|U_{kj}^{(l)} - U_{kj}^{(\backslash l)}\right| - \eta_j \left(\sum_k U_{kj}^{(l)} - 1\right).$$

After rearranging terms to remove the absolute value and dropping constants, we have

$$\sum_k Q_{kj} \log H_{kj}^{(l)} - \lambda \sum_{H_{kj}^{(l)} < H_{kj}^{(\backslash l)}} \left(H_{kj}^{(l)} - H_{kj}^{(\backslash l)} \right)$$
$$- \lambda \sum_{H_{kj}^{(l)} < H_{kj}^{(\backslash l)}} \left(H_{kj}^{(\backslash l)} - H_{kj}^{(l)} \right) - \eta_j \sum_k H_{kj}^{(l)}.$$

Next, we can take derivative with regards to this function, which will be in two cases. For $U_{kj}^{(l)} > U_{kj}^{(\backslash l)}$, setting the derivative with regards to $U_{kj}^{(l)}$ becomes $\frac{Q_{kj}}{U_{kj}^{(l)}} - \lambda - \eta_j = 0 \Rightarrow U_{kj}^{(l)} = \frac{Q_{kj}}{\eta_j + \lambda}$. This holds when $U_{kj}^{(l)} = \frac{Q_{kj}}{\eta_j + \lambda} > U_{kj}^{(\backslash l)} \Rightarrow \frac{Q_{kj}}{U_{kj}^{(\backslash l)}} - \lambda > \eta_j$. To make $U_{kj}^{(l)}$ nonnegative, it is easy to see that $\eta_j > -\lambda$. For $U_{kj}^{(l)} < U_{kj}^{(\backslash l)}$, setting the derivative with regards to $U_{kj}^{(l)}$ becomes $\frac{Q_{kj}}{U_{kj}^{(l)}} + \lambda - \eta_j = 0 \Rightarrow U_{kj}^{(l)} = \frac{Q_{kj}}{\eta_j - \lambda}$. This holds when $U_{kj}^{(l)} = \frac{Q_{kj}}{\eta_j - \lambda} < U_{kj}^{(\backslash l)} \Rightarrow \frac{Q_{kj}}{U_{kj}^{(\backslash l)}} + \lambda < \eta_j$. When these two conditions are not satisfied, the optimal solution is given by setting $U_{kj}^{(l)} = U_{kj}^{(\backslash l)}$. Combining all these results yields Eq.(21). □

References

[1] T. Hofmann, Probabilistic latent semantic analysis, in: UAI, 1999, pp. 289–296.

[2] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet allocation, in: J. Mach. Learn. Res., Vol. 3, JMLR, 2003, pp. 993–1022.

[3] X. Wang, X. Ma, E. Grimson, Unsupervised activity perception by hierarchical bayesian models, in: CVPR, 2007.

[4] S. Nikolopoulos, S. Zafeiriou, I. Patras, I. Kompatsiaris, High order PLSA for indexing tagged images, in: Signal Processing, 2013.

[5] T. Hospedales, S. Gong, T. Xiang, A Markov clustering topic model for mining behaviour in video, in: IJCV, 2012.

[6] J. Sivic, B. C. Russell, A. Zisserman, I. Ecole, N. Supérieure, Unsupervised discovery of visual object class hierarchies, in: In CVPR, 2008.

[7] D. Küttel, M. D. Breitenstein, L. J. V. Gool, V. Ferrari, What s going on? discovering spatio-temporal dependencies in dynamic scenes, in: CVPR, 2010.

[8] J. Li, S. Gong, T. Xiang, Global behaviour inference using probabilistic latent semantic analysis, in: BMVC, 2008, pp. 1–10.

[9] L. Fei-fei, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: CVPR, 2005, pp. 524–531.

[10] J. C. Niebles, H. Wang, L. Fei-fei, Unsupervised learning of human action categories using spatial-temporal words, in: BMVC, 2006.

[11] J. Varadarajan, J.-M. Odobez, Topic models for scene analysis and abnormality detection, in: 9th International Workshop in Visual Surveillance, 2009.

[12] X. Wang, K. T. Ma, G.-W. Ng, W. E. L. Grimson, Trajectory analysis and semantic region modeling using a nonparametric bayesian model, in: CVPR, 2008.

[13] X. Wei, W. B. Croft, LDA-based document models for ad-hoc retrieval, in: SIGIR, 2006.

[14] R. Fernandez-Beltran, F. Pla, Latent topics-based relevance feedback for video retrieval, in: Pattern Recognition, Vol. 51, 2016, pp. 72–84.

[15] M. Kandemir, T. Keke, R. Yeniterzi, Supervising topic models with Gaussian processes, in: Pattern Recognition, Vol. 77, 2018, pp. 226–236.

[16] D. Tu, L. Chen, M. Lv, H. Shi, G. Chen, Hierarchical online NMF for detecting and tracking topic hierarchies in a text stream, in: Pattern Recognition, Vol. 76, 2018, pp. 203–214.

[17] R. Nallapati, A. Ahmed, E. P. Xing, W. W. Cohen, Joint latent topic models for text and citations, in: KDD, 2008, pp. 542–550.

[18] M. Rosen-Zvi, T. L. Griffiths, M. Steyvers, P. Smyth, The author-topic model for authors and documents, in: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, 2004.

[19] Y. Liu, A. Niculescu-Mizil, W. Gryc, Topic-link LDA: Joint models of topic and author community, in: ICML, 2009.

[20] Y. Jia, M. Salzmann, T. Darrell, Learning cross-modality similarity for multinomial data, in: ICCV, 2011.

[21] L. Song, J. Liu, M. Luo, B. Qian, K. Yang, Sparse Relational Topical Coding on multi-modal data, in: Pattern Recognition, Vol. 72, 2017, pp. 368–380.

[22] X. Gao, T. Mu, J. Y. Goulermas, M. Wang, Topic driven multimodal similarity learning with multi-view voted convolutional features, in: Pattern Recognition, Vol. 75, 2018, pp. 223–234.

[23] Q. Mei, D. Cai, D. Zhang, C. Zhai, Topic modeling with network regularization, in: WWW, 2008, pp. 101–110.

[24] D. Cai, X. He, J. Han, T. S. Huang, Graph regularized non-negative matrix factorization for data representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (8) (2011) 1548–1560.

[25] J. Chang, D. Blei, Relational topic models for document networks, in: Artificial Intelligence and Statistics, 2009.

[26] S. Virtanen, Y. Jia, A. Klami, T. Darrell, Factorized multi-modal topic model, in: UAI, 2012.

[27] Y. Jiang, J. Liu, Z. Li, P. Li, H. Lu, Co-regularized PLSA for multi-view clustering, in: ACCV, 2012.

[28] Y. Jiang, J. Liu, Z. Li, H. Lu, Collaborative PLSA for multi-view clustering, in: International Conference on Pattern Recognition, 2012.

[29] X. Wang, M.-C. Chang, Y. Ying, S. Lyu, Co-regularized PLSA for multi-modal learning, in: AAAI, 2016.

[30] J. G. Jialu Liu, Chi Wang, J. Han, Multi-view clustering via joint nonnegative matrix factorization, in: SIAM Data Mining Symposium, 2013.

[31] X. He, M.-Y. Kan, P. Xie, X. Chen, Comment-based multi-view clustering of web 2.0 items, in: International Conference on World Wide Web, 2014.

[32] E. Gaussier, C. Goutte, Relation between PLSA and NMF and implications, in: SIGIR, 2005, pp. 601–602.

[33] C. Ding, T. Li, W. Peng, On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing factorization, in: Computational Statistics and Data Analysis, Vol. 52, 2008, pp. 3913–3927.

[34] D. Cohn, T. Hofmann, The missing link - a probabilistic model of document content and hypertext connectivity, in: Advances in Neural Information Processing Systems, 2001.

[35] D. Zhang, Q. Mei, C. Zhai, Cross-lingual latent topic extraction., in: ACL, 2010, pp. 1128–1137.

[36] J. Boyd-Graber, D. M. Blei, Multilingual topic models for unaligned text, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI '09, 2009, pp. 75–82.

[37] J. C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, N. Vasconcelos, On the role of correlation and abstraction in cross-modal multimedia retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (2014) 521–535.

[38] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: ACM International Conference on Multimedia, 2010, pp. 251–260.

[39] X. Mao, B. Lin, Cai, X. He, J. Pei, Parallel field alignment for cross media retrieval, in: Proceedings of the 21st ACM International Conference on Multimedia, 2013, pp. 897–906.

[40] N. Srivastava, R. Salakhutdinov, Multimodal learning with deep Boltzmann machines, in: Advances in Neural Information Processing Systems 25, 2012, pp. 2231–2239.

[41] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng, Multimodal deep learning, in: International Conference on Machine Learning, 2011.

[42] S. Lyu, X. Wang, On algorithms of sparse multi-factor nonnegative matrix factorization, in: Advances in Neural Information Processing Systems (NIPS), Lake Tahoe, NV, 2013.

[43] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, Series B (1) (1977) 1–38.

[44] R. Neal, G. E. Hinton, A view of the EM algorithm that justifies incremental, sparse, and other variants, in: Learning in Graphical Models, 1998, pp. 355–368.

[45] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, D. E. Knuth, On the Lambert W function, in: Advances in Computational Mathematics, 1996, pp. 329–359.

[46] M. Shashanka, B. Raj, P. Smaragdis, Sparse overcomplete latent variable decomposition of counts data, in: J. Platt, D. Koller, Y. Singer, S. Roweis (Eds.), Advances in Neural Information Processing Systems 20, 2007, pp. 1313–1320.

[47] M. Brand, Pattern discovery via entropy minimization., in: AISTATS, 1999.

[48] T. Sim, S. Baker, M. Bsat, The cmu pose, illumination, and expression (PIE) database, in: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, FGR 2002.

[49] W. Xu, X. Liu, Y. Gong, Document clustering based on non-negative matrix factorization, in: SIGIR, 2003.

[50] I. Khan, A. Saffari, H. Bischof, TVGraz: Multi-modal learning of object categories by combining textual and visual features, in: AAPR Workshop, 2009, pp. 213–224.

[51] European parliament proceedings parallel corpus (EPPPC), http://www.statmt.org/europarl/.

[52] M. W.Berry, MurrayBrowne, A. N.Langville, V. PaulPauca, R. J.Plemmons, Algorithms and applications for approximate nonnegative matrix factorization, in: Computational Statistics and Data Analysis, Vol. 52, 2006, pp. 155–173.

Author Biographies

Dr. Xin Wang is currently a Senior Machine Learning Scientist at the CuraCloud Corporation. He received his Ph.D. degree in Computer Science from the University at Albany, State University of New York in 2015. His research interests are in artificial intelligence, machine learning and computer vision.

Dr. Ming-Ching Chang is an Assistant Professor at the Department of Electrical and Computer Engineering, College of Engineering and Applied Sciences (CEAS), University at Albany - SUNY. He was a lead computer scientist at GE Global Research Center from 2008 to 2016. He received his Ph.D. degree from Brown University in 2008. He has published over 50 refereed journal and conference papers. His research projects are funded by GE Global Research, DARPA, NIJ, VA, and University at Albany, SUNY. He is the recipient of the

IEEE Advanced Video and Signal-based Surveillance (AVSS) 2011 Best Paper Award - Runner-Up, the IEEE Workshop on the Applications of Computer Vision (WACV) 2012 Best Student Paper Award, the GE Belief - Stay Lean and Go Fast Management Award in 2015, and the IEEE Smart World NVIDIA AI City Challenge 2017 Honorary Mention Award. Dr. Chang serves as the Publication and Associate Chair of the IEEE Advanced Video and Signal-based Surveillance (AVSS) 2017, co-chair of the International Workshop on Traffic and Street Surveillance for Safety and Security (IWT4S) 2017, and the Application Track co-chair of the NVIDIA AICity Challenge in 2017.

Siwei Lyu is currently a tenured Associate Professor at the Department of Computer Science, the College of Engineering and Applied Sciences, the University at Albany, State University of New York. Dr. Lyu received his Ph.D. degree in Computer Science from Dartmouth College in 2005, and his M.S. degree in Computer Science in 2000 and B.S. degree in Information Science in 1997, both from Peking University, China. Dr. Lyus research interests include digital image forensics, computer vision, computational neuroscience and machine learning. Dr. Lyu has published over 80 refereed journal and conference papers. He is the recipient of the IEEE Signal Processing Society Best Paper Award in 2011 and the National Science Foundation (US) CAREER Award in 2010.