# Co-regularized PLSA For Multi-Modal Learning

**Xin Wang**[1], **Ming-Ching Chang**[1], **Yiming Ying**[2], **Siwei Lyu**[1]
[1]Department of Computer Science, [2]Department of Mathematics and Statistics
University at Albany, State University of New York
Albany, NY 12222

## Abstract

Many learning problems in real world applications involve rich datasets comprising multiple information modalities. In this work, we study co-regularized PLSA (coPLSA) as an efficient solution to probabilistic topic analysis of multi-modal data. In coPLSA, similarities between topic compositions of a data entity across different data modalities are measured with divergences between discrete probabilities, which are incorporated as a *co-regularizer* to augment individual PLSA models over each data modality. We derive efficient iterative learning algorithms for coPLSA with symmetric KL, $\ell_2$ and $\ell_1$ divergences as co-regularizers, in each case the essential optimization problem affords simple numerical solutions that entail only matrix arithmetic operations and numerical solution of 1D nonlinear equations. We evaluate the performance of the coPLSA algorithms on text/image cross-modal retrieval tasks, on which they show competitive performance with state-of-the-art methods.

## Introduction

Numerous real-world applications of machine learning involve rich datasets comprising multiple and heterogeneous information modalities. For instance, Wikipedia pages typically include both texts and images, articles recounting emerging news stories often have multiple versions translated into different languages, and datasets of social network usually contain user profiles as well as friendship links. In a multi-modal dataset, each data modality may only reveal partial yet relevant information of the data entities being studied, and only in combination do they yield a complete description. Combining multiple modalities can improve the performance of learning algorithms, and the resulting multi-modal learning methods have found wide ranges of applications, such as image annotation (Srivastava and Salakhutdinov 2012), multi-media retrieval (Pereira et al. 2014; Rasiwasia et al. 2010; Mao et al. 2013), and audio-visual speech classification (Ngiam et al. 2011).

As each data modality may have intrinsically different representations, simply concatenating them cannot lead to a satisfactory solution to multi-modal learning problems. On the other hand, borrowing ideas from probabilistic topic analysis for text documents (Hofmann 1999; Blei, Ng, and Jordan 2003), we can model each data entity in a multi-modal dataset as a probabilistic mixture of "topics" that corresponds to common thematic concepts. Then the same topic is instantiated as multiple probabilistic distributions, each of which is defined over the basic representation of one data modality. Representing data entities using their topic composition thus discounts the difference of basic representations in the data modalities. As such, learning methods based on the topic representations of the multi-modal dataset can leverage compatible and complementary conceptual themes encompassed within each modality, and are often more effective than methods that use features from direct concatenation of all modalities.

In several previous works (Nallapati and Cohen 2008; Nallapati et al. 2008; Liu, Niculescu-Mizil, and Gryc 2009), integrating multi-modal data at topic level is achieved by requesting each data entity to have the same topic compositions across different modalities. This is equivalent to require multiple representations of one data entity over different modalities to *share* their topic compositions. This approach fares well when different modalities contain information largely consistent with each other. But this may not always be the case, for instance, an image of the *Central Park* may be associated with texts in a Wiki document about *New York City*, but its contents may also be related with another document about *Parks*. A less restrictive approach is to model topic compositions of each data entity across different modalities separately, and the learning algorithms encourage them to be *similar* but not necessarily identical. As such, it has the flexibility of capturing non-overlapping topic compositions over different modalities, and can recover more diverse topics across different modalities to summarize the thematic concepts embodied in the dataset.

In this work, we study co-regularized probabilistic latent semantic analysis (coPLSA) as a general method for topic analysis of multi-modal datasets. We describe a general framework of coPLSA where the co-regularizers are divergences between discrete probability distributions that correspond to topic compositions of a data entry across different modalities. Optimizing the objective of PLSA while minimizing such divergences serve to encourage similarity between topic compositions of a data entry across differ-

ent modalities during topic learning. For three widely used divergences (*i.e.*, symmetric KL, $\ell_2$ and $\ell_1$ divergences), we derive efficient algorithms for learning topics and topic compositions on multi-modal datasets, all based on simple matrix operations and numerical solution of 1D nonlinear equations. Unlike previous works (Jiang et al. 2012a; 2012b), our algorithms follow the correct optimization of the objective functions, and afford theoretical guarantee of convergence. The coPLSA algorithms are applied to cross-modal retrieval tasks on benchmark text/image datasets, and compared favorably with the current state-of-the-art methods.

## Related Works

Many existing works on multi-modal learning are based on seeking latent representations where the difference among multiple modalities are minimized. Such latent representations can be directly obtained in the form of nonlinear manifolds with kernelized canonical correlation analysis (Vinokourov, Shawe-taylor, and Cristianini 2002), cross-modal factor analysis (Pereira et al. 2014), manifold alignment approach (Mao et al. 2013), joint dimension reduction (Mahadevan et al. 2011) or deep network models (Ngiam et al. 2011; Srivastava and Salakhutdinov 2012). Another popular approach to find such joint latent representations is through the use of co-regularized nonnegative matrix factorization (NMF)(Jialu Liu and Han 2013; He et al. 2014). Though each of these methods has its merits, the learned joint latent representations usually do not afford explicit probabilistic interpretations.

The probabilistic topic analysis of multi-modal data was first formulated in (Cohn and Hofmann 2001), in which the citations in a document is considered as another modality to the documents in a corpus, and the shared topics of two modalities are the weighted combination of the topic that learned form the individual modalities. Subsequently, many methods, *e.g.*, (Nallapati et al. 2008; Rosen-Zvi et al. 2004; Liu, Niculescu-Mizil, and Gryc 2009) were developed to jointly model document topics and other auxiliary information provided with the corpus. However, these models all assume that each data entity and its associated auxiliary data share the same topic composition across different modality. As pointed out in the previous section, this assumption may be too restrictive when applied to real-world multi-modal datasets that contain non-text data types such as images.

A Bayesian treatment of multi-modal topics that incorporates similarities between associated topic compositions of different data modalities leads to a Markov random field augmented probabilistic topic model that has been studied recently in (Virtanen et al. 2012). Though achieving good performances in multi-modal learning tasks, the Bayesian MRF model suffers from increased complexity in the learning and inference algorithms that have to be implemented with Monte-Carlo methods. Therefore, it is useful to extend simpler topic analysis methods such as probabilistic latent semantic analysis (PLSA) (Hofmann 1999) to multi-modal learning, 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 (Jiang et al. 2012a; 2012b)[1]. The co-regularizer used in (Jiang et al. 2012a) is based on the mutual similarities of data in the topic space, and that of (Jiang et al. 2012b) is the $\ell_2$ 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 (we pointed out in Section 4). As such, the algorithms in these previous works do not afford guarantees to converge and usually lead to inferior performance.

In comparison with the previous works, the main contributions of this work can be summarized as follows: (i) We describe the general method of coPLSA based on divergence between discrete probability distributions as co-regularizer; (ii) for $\ell_2$ divergence as co-regularizer, our coPLSA algorithm is more efficient and guarantees convergence (see Section 4 for detail); and (iii) we describe new coPLSA algorithms using symmetric KL and $\ell_1$ divergences as co-regularizers and demonstrate that they are more effective than that based on $\ell_2$ divergence.

## Review of PLSA Algorithm

We first introduce notations and definitions to be used hereafter. A $d$-dimensional vector $\mathbf{v}$ is *stochastic* if $v_i \geqslant 0$ and $\sum_{i=1}^{d} v_i = 1$, and corresponds to a categorical probability distribution over $d$ outcomes. A $d \times n$ nonnegative matrix $V$ is *stochastic* if its column vectors are stochastic.

For two $d$-dimensional stochastic vectors $\mathbf{v}$ and $\mathbf{w}$, we define their Kulback-Leibler (KL), $\ell_2$ and $\ell_1$ divergence, in sequence, as: $\mathcal{D}_{\text{KL}}(\mathbf{v},\mathbf{w}) = \sum_{i=1}^{d} v_i \log \frac{v_i}{w_i}$, $\mathcal{D}_{\ell_2}(\mathbf{v},\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{d}(v_i - w_i)^2$, $\mathcal{D}_{\ell_1}(\mathbf{v},\mathbf{w}) = \sum_{i=1}^{d}|v_i - w_i|$, and their symmetric KL divergence is defined as $\mathcal{D}_{\text{sKL}}(\mathbf{v},\mathbf{w}) = \mathcal{D}_{\text{KL}}(\mathbf{v},\mathbf{w}) + \mathcal{D}_{\text{KL}}(\mathbf{w},\mathbf{v})$. Accordingly, we define divergence between two stochastic matrices $V$ and $W$ as the sum of the divergence between their corresponding columns, as

$$\mathcal{D}_*(W,V) = \textstyle\sum_j \mathcal{D}_*(W_{\cdot,j}, V_{\cdot,j}), \tag{1}$$

where $\mathcal{D}_*$ can be replaced with $\mathcal{D}_{\text{KL}}, \mathcal{D}_{\text{sKL}}, \mathcal{D}_{\ell_2}$ or $\mathcal{D}_{\ell_1}$. For stochastic vectors/matrices, these divergences are nonnegative 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 of a dataset, where each data entity (a "document") is represented as the normalized frequencies over some basic features ("words" in a "vocabulary"). Probabilistic latent semantic analysis (PLSA) (Hofmann 1999) is based on a simple probabilistic generative model of the dataset (Blei, Ng, and Jordan 2003): each word

[1]Because of the close relation between PLSA and NMF algorithms (Gaussier and Goutte 2005; Ding, Li, and Peng 2008), coPLSA can also be regarded as a co-regularized NMF problem. However, most existing co-regularized NMF methods use $\ell_2$ divergence for both main objective and co-regularizer, and do not consider the normalization constraint.

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, for $n$ 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 $V$ of dimension $d \times n$ to represent conditional probabilities, as $V_{ij} \equiv \texttt{Prob}(\text{word} = \mathbf{w}_i | \text{doc} = \mathbf{d}_j)$. Assuming the documents are associated with $m$ topics, $(\mathbf{t}_1, \cdots, \mathbf{t}_m)$, we use stochastic matrices $W$ of dimension $d \times m$ and $H$ of dimension $m \times n$ to represent conditional probabilities, as $W_{ik} \equiv \texttt{Prob}(\text{word} = \mathbf{w}_i | \text{topic} = \mathbf{t}_k)$ and $H_{kj} \equiv \texttt{Prob}(\text{topic} = \mathbf{t}_k | \text{doc} = \mathbf{d}_j)$, respectively. According to the document generation model, documents and words are conditionally independent from each other. As such, these probabilities satisfy $\texttt{Prob}(\text{word} = \mathbf{w}_i | \text{doc} = \mathbf{d}_j) = \sum_k \texttt{Prob}(\text{word} = \mathbf{w}_i | \text{topic} = \mathbf{t}_k) \texttt{Prob}(\text{topic} = \mathbf{t}_k | \text{doc} = \mathbf{d}_j)$. With the matrix notations, this is equivalent to $V = WH$. Given a dataset represented in matrix $V$, PLSA attempts to find its decomposition into $W$ and $H$, formulated as an optimization problem: $\min_{W,H} \mathcal{D}_{KL}(V, WH)$, with the constraint that both $W$ and $H$ are stochastic matrices. After dropping irrelevant constant terms, minimizing the KL divergence is equivalent to maximizing

$$\mathcal{J}(W, H) = \textstyle\sum_{ij} V_{ij} \log (WH)_{ij}. \tag{2}$$

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

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

$$\begin{aligned} \mathcal{F}(W, \hat{W}) &= \sum_{ijk} \frac{V_{ij} \hat{W}_{ik} H_{kj}}{\left(\hat{W}H\right)_{ij}} \log \left( \frac{W_{ik}}{\hat{W}_{ik}} \left(\hat{W}H\right)_{ij} \right) \\ &= \sum_{ik} M_{ik} \log W_{ik} + \text{const.} \end{aligned} \tag{3}$$

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

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

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

$$\max_W \ \textstyle\sum_{ik} M_{ik} \log W_{ik}, \text{ s.t. } W_{ij} \geqslant 0 \ \& \ \sum_i W_{ij} = 1. \tag{4}$$

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

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

$$\begin{aligned} \mathcal{G}(H, \hat{H}) &= \sum_{ijk} \frac{V_{ij} W_{ik} \hat{H}_{kj}}{\left(W\hat{H}\right)_{ij}} \log \left( \frac{H_{kj}}{\hat{H}_{kj}} \left(W\hat{H}\right)_{ij} \right) \\ &= \sum_{kj} Q_{kj} \log H_{kj} + \text{const}, \end{aligned} \tag{5}$$

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

$$\max_H \ \textstyle\sum_{kj} Q_{ik} \log H_{kj}, \text{ s.t. } H_{kj} \geqslant 0 \ \& \ \sum_k H_{kj} = 1, \tag{6}$$

of which the solution is given by $H_{kj} = \frac{Q_{kj}}{\sum_{k'} Q_{k'j}}$ (proof given in Supplementary Materials).

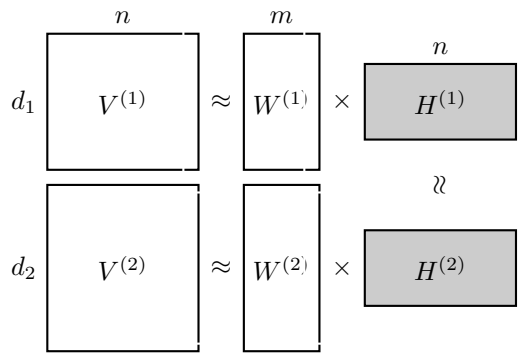
## The coPLSA Algorithm

In coPLSA, our goal is to perform *joint* PLSA of a multimodal dataset across different modalities, based on the assumption that different data modalities admit similar underlying semantic structure of the data. Formally, given two modalities of the dataset[2] represented with stochastic matrices $V^{(1)}$ and $V^{(2)}$ of size $d_1 \times n$ and $d_2 \times n$, we seek factorization $V^{(l)} \approx W^{(l)} H^{(l)}$, with stochastic matrices $W^{(l)}$ of size $d_\ell \times m$ and matrix $H^{(l)}$ of size $m \times n$ representing the $m$ modality-specific topic matrices and the topic compositions of the dataset, respectively. In coPLSA, association of different modalities to their common data entry is achieved by coupling the factorizations $V^{(l)} \approx W^{(l)} H^{(l)}$, *i.e.*, besides individual PLSA objectives to each modality, the algorithm also introduce co-regularizers to minimize the difference of $H$ matrices, corresponding to the topic compositions of each modality (illustrated in the left panel of Fig.1). Specifically, coPLSA is formulated as a constrained optimization problem as

$$\min_{W^{(\ell)}, H^{(\ell)}} \sum_{\ell=1,2} \mathcal{D}_{KL}(V^{(l)}, W^{(l)} H^{(l)}) + \lambda \mathcal{D}_*(H^{(1)}, H^{(2)}), \tag{7}$$

with the constraint that $W^{(l)}$ and $H^{(l)}$ are stochastic matrices. Parameter $\lambda > 0$ balances the contribution of the

[2]The algorithm described subsequently can be easily extended to more than two data modalities.




Initialize $W^{(1)}, W^{(2)}, H^{(1)}, H^{(2)}$;
**While** not converge
  **While** not converge
    update $W^{(l)}$ using Eq.(4);
  **While** not converge
    **If** $\mathcal{D}_* = \mathcal{D}_{\text{sKL}}$, update $H^{(l)}$ with (11) and (14);
    **If** $\mathcal{D}_* = \mathcal{D}_{\ell_2}$, update $H^{(l)}$ with (12) and (14);
    **If** $\mathcal{D}_* = \mathcal{D}_{\ell_1}$, update $H^{(l)}$ with (13) and (14);

Figure 1: **Left**: Illustration of coPLSA as two stochastic matrix factorization problems co-regularized through similarities on factors. **Right**: Pseudo-code of the coPLSA algorithm in this work.

PLSA objectives of each modality and the co-regularization term. In the following, $\mathcal{D}_*$ will be replaced with the symmetric KL, $\ell_2$ or $\ell_1$ divergences[3]. Dropping irrelevant constant terms, the objective function of (7) can be further simplified to

$$\max_{W^{(\ell)},H^{(\ell)}} \sum_{\ell=1,2} \mathcal{J}(W^{(l)}, H^{(l)}) - \lambda \mathcal{D}_*(H^{(1)}, H^{(2)}), \quad (8)$$

with the same constraints on the factors.

As in the case of PLSA, in the learning step of coPLSA, the objective function in (8) is optimized with a block-coordinate descent scheme, alternating between steps that optimizing each of $W^{(1)}$, $W^{(2)}$, $H^{(1)}$ or $H^{(2)}$ while fixing the other factors. In the following, we describe the steps of these sub-problems.

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

**Optimizing** $H^{(l)}$: The optimization of $H^{(l)}$ is different because of the co-regularizer. For simplicity, we use $H^{(\backslash l)}$ to denote the other $H$ factor other than $H^{(l)}$. The optimization of $H^{(l)}$ with fixed $W^{(l)}$ and $H^{(\backslash l)}$, after removing irrelevant constant terms, becomes

$$\max_{H^{(l)}} \ \mathcal{J}(W^{(l)}, H^{(l)}) - \lambda \mathcal{D}_*(H^{(l)}, H^{(\backslash l)}),$$
$$\text{s.t. } H^{(l)}_{kj} \geqslant 0 \ \& \ \sum_k H^{(l)}_{kj} = 1.$$

Using the auxiliary function $\mathcal{G}$ defined in Eq.(5), we can also obtain a tight lower-bound of the above objective function, as: $\mathcal{G}(H^{(l)}, \hat{H}) - \lambda \mathcal{D}_*(H^{(l)}, H^{(\backslash l)}) \leqslant \mathcal{J}(W^{(l)}, H^{(l)}) - \lambda \mathcal{D}_*(H^{(l)}, H^{(\backslash l)})$ with equality when $\hat{H} = H^{(l)}$, which follows from the property of $\mathcal{G}$. Note that in this lower-bound, the second term $\lambda \mathcal{D}_*(H^{(l)}, H^{(\backslash l)})$ does not depend on the auxiliary variable $\hat{H}$.

Then, a similar EM algorithm can be developed to optimize $H^{(l)}$ iteratively, which improves the lower-bound in each iteration: starting with an initial values $H^{(l)} = H^{(l,0)}$, we iteratively solve for

$$H^{(l,t+1)} \leftarrow \operatorname*{argmax}_{H^{(l)}} \mathcal{G}(H^{(l)}, H^{(l,t)}) - \lambda \mathcal{D}_*(H^{(l)}, H^{(\backslash l)}),$$
$$\text{s.t. } H^{(l)}_{kj} \geqslant 0 \ \& \ \sum_k H^{(l)}_{kj} = 1. \quad (9)$$

In each iteration, it is guaranteed that the objective function will not be decreased, as

$$\mathcal{J}(W^{(l)}, H^{(l,t)}) - \lambda \mathcal{D}_*(H^{(l,t)}, H^{(\backslash l)}) = \mathcal{G}(H^{(l,t)}, H^{(l,t)}) - \lambda \mathcal{D}_*(H^{(l,t)}, H^{(\backslash l)}) \leqslant \mathcal{G}(H^{(l,t+1)}, H^{(l,t)}) - \lambda \mathcal{D}_*(H^{(l,t+1)}, H^{(\backslash l)}) \leqslant \mathcal{J}(W^{(l)}, H^{(l,t+1)}) - \lambda \mathcal{D}_*(H^{(l,t+1)}, H^{(\backslash l)}).$$

As such, the sequence $\left(H^{(l,0)}, H^{(l,1)}, \cdots\right)$ monotonically increases the objective function $\mathcal{J}(W^{(l)}, H^{(l)}) - \lambda \mathcal{D}_*(H^{(l)}, H^{(\backslash l)})$ until reaching a local minimum.

Solving the optimization problem in (9) is key to the optimization. For the $\ell_2$ co-regularizer, a method is given in (Jiang et al. 2012b), where one first updates $H^{(l)}_{kj}$ using the PLSA EM step, Eq.(6), followed by another update of $H^{(l)}_{kj}$ to decrease $\lambda \mathcal{D}_{\ell_2}(H^{(l,t)}, H^{(\backslash l)})$. Unfortunately, the two update steps may undo the effect of each other as they are performed independently. As such, there is no clear guarantee that the overall algorithm will converge or converge to the optimal solution.

In this work, we provide efficient algorithms for symmetric KL, $\ell_2$ and $\ell_1$ divergences with convergence guarantees. Using the equivalent matrix definition of function $\mathcal{G}$ in Eq.(5) with $Q$ obtained from $V^{(l)}$, $W^{(l)}$ and $\hat{H}$, the essential optimization problem we need to solve is

$$\max_{H^{(l)}} \ \sum_{kj} Q_{kj} \log H^{(l)}_{kj} - \lambda \mathcal{D}_*(H^{(l)}, H^{(\backslash l)}),$$
$$\text{s.t. } H^{(l)}_{kj} \geqslant 0 \ \& \ \sum_k H^{(l)}_{kj} = 1. \quad (10)$$

With respect to three types of co-regularizer, namely, symmetric KL, $\ell_2$ and $\ell_1$ divergences, the optimal solution to (10) are given as non-linear functions of a scalar variable $\eta_j$ that corresponds to the Lagrangian multiplier of the normalizing constraint, $\sum_k H^{(l)}_{kj} = 1$ and is shared by all elements in one column of $H^{(l)}$. Specifically, these solutions are given

[3]It is also possible to incorporate other regularization terms on the factors $W^{(l)}$ and $H^{(l)}$ to express other preference on the factors such as sparsity. Furthermore, we can us similar methods to enforce consistencies in parts of factor $W^{(l)}$. However, for simplicity, in the current work we do not consider these types of regularizers.

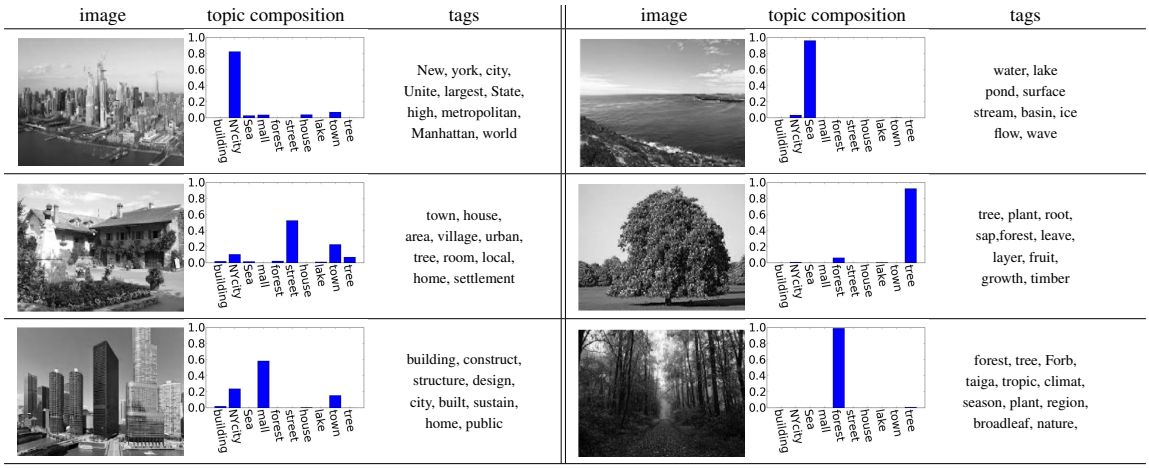



Figure 2: *Illustrating images with their annotation and topic proportion.*

in the following equations (proofs given in the Supplementary Materials):

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

$$H_{kj}^{(l)}(\eta_j) = \frac{Q_{kj}+\lambda H_{kj}^{(\backslash l)}}{\lambda \mathcal{W}_0\left(\frac{Q_{kj}+\lambda H_{kj}^{(\backslash l)}}{\lambda H_{kj}^{(\backslash l)}} \exp\left(1+\frac{\eta_j}{\lambda}\right)\right)}, \tag{11}$$

  where $\mathcal{W}_0(\cdot)$ is the principal branch of the Lambert $\mathcal{W}$ function (Corless et al. 1996) that is defined implicitly as $z = W(z)e^{W(z)}$ for $z > 0$ [4].

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

$$H_{kj}^{(l)}(\eta_j) = \frac{1}{2}\sqrt{\left(H_{kj}^{(\backslash l)} - \frac{\eta_j}{\lambda}\right)^2 + \frac{4Q_{kj}}{\lambda}} + \frac{1}{2}\left(H_{kj}^{(\backslash l)} - \frac{\eta_j}{\lambda}\right). \tag{12}$$

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

$$H_{kj}^{(l)}(\eta_j) = \begin{cases} \frac{Q_{kj}}{\eta_j+\lambda}, & -\lambda < \eta_j < \frac{Q_{kj}}{H_{kj}^{(\backslash l)}} - \lambda, \\ H_{kj}^{(\backslash l)}, & \frac{Q_{kj}}{H_{kj}^{(\backslash l)}} - \lambda \leqslant \eta_j \leqslant \frac{Q_{kj}}{H_{kj}^{(\backslash l)}} + \lambda, \\ \frac{Q_{kj}}{\eta_j-\lambda}, & \frac{Q_{kj}}{H_{kj}^{(\backslash l)}} + \lambda < \eta_j. \end{cases} \tag{13}$$

  Compared with the two other divergence types, the update steps for $\ell_1$ divergence in Eq.(13) corresponds to a piecewise function. The computation only involves arithmetic operations and thresholding, and is substantially simpler and more efficient. The use of $\ell_1$ co-regularizer has another important property that the resulting $H^{(l)}$ can have identical components as $H^{(\backslash l)}$, while for the $\ell_2$ and symmetric KL co-regularizers, this is not usually the case.

We use $H_{kj}^{(l)}(\eta_j)$ in Eqs.(11,12,13) to emphasize the fact that they are functions of the scalar parameter $\eta_j$. To determine the value of $\eta_j$, which in turn leads to the optimal solution to $H^{(l)}$, we can solve the following 1D nonlinear equation corresponding to the normalization constraint in (10),

$$\textstyle\sum_k H_{kj}^{(l)}(\eta_j) = 1. \tag{14}$$

For each type of co-regularizers, we use the corresponding $H_{kj}^{(l)}(\eta_j)$ in Eq.(11,12,13). For each column index $j$, Eq.(14) is solved numerically, *e.g.*, with Newton-Raphson when $H_{kj}^{(l)}(\eta_j)$ 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 coPLSA problem with an iterative algorithm that alternates between the optimization of individual $W$ and $H$ factors while fixing the others. The optimization of $W$ factor is performed with another iterative EM algorithm based on individual optimization steps given in (4), and the optimization of $H$ factor is achieved by iterating steps that first solve Eq.(14) and then determine the factors with Eq.(11), (12) or (13). In practice, all iterative algorithms converges within 5-10 steps. The right panel of Fig.1 provides the pseudo-code of the overall algorithm.

## Application to Cross-Modal Retrieval

One important multi-modal learning task is cross-modal Text/Image retrieval. With the proliferation of online multi-modal data (Wikipedia, Youtube, etc), queries are frequently made in one modality (*e.g.*, image or videos) with input from another modality (*e.g.*, texts). In this section, we apply the coPLSA algorithms to cross-modal retrieval tasks involving documents containing text and images. Specifically, we consider two tasks: text retrieval from an image query (*i2t*), and image retrieval from a query with a text document (*t2i*). With the coPLSA model, the difference in basic data representations of text and images are encapsulated into the corresponding topics. And with their topic compositions, texts and images are projected into a compatible semantic space, which can be used to establish links between images and text documents and facilitates cross-modal retrieval.

We use two benchmark text/image datasets in our experiments: *TVGraz* (Khan, Saffari, and Bischof 2009) and *Wikipedia* (Rasiwasia et al. 2010), in which images are associated with long text documents and the texts and images do not necessarily have direct correspondence as in the case

[4] The Lambert $\mathcal{W}$ function can be numerically evaluated and is provided in popular numerical tools such as MATLAB (function `lambertw`) or SciPy (function `scipy.special.lambertw`). It has been used in algorithms that enforce entropic priors (Brand 1999). It also appears in a variant of PLSA to encourage sparsity over the obtained $W$ or $H$ factors (Shashanka, Raj, and Smaragdis 2007).

| Methods | Topic space similarity | | | | Semantic space similarity | | | |
|---|---|---|---|---|---|---|---|---|
| | TVGraz | | Wikipedia | | TVGraz | | Wikipedia | |
| | i2t | t2i | i2t | t2i | i2t | t2i | i2t | t2i |
| SCM (Pereira et al. 2014) | 0.460 | 0.450 | 0.267 | 0.219 | 0.664 | 0.649 | 0.362 | 0.273 |
| Link PLSA (Cohn and Hofmann 2001) | 0.349 | 0.349 | 0.247 | 0.247 | 0.803 | 0.803 | 0.605 | 0.605 |
| $\ell_1$ coPLSA | 0.359 | 0.365 | 0.317 | 0.307 | 0.723 | 0.726 | 0.667 | 0.658 |
| $\ell_2$ coPLSA | 0.450 | 0.445 | 0.360 | 0.358 | 0.846 | 0.845 | 0.706 | 0.701 |
| sKL coPLSA | **0.481** | **0.481** | **0.413** | **0.413** | **0.850** | **0.850** | **0.726** | **0.724** |

Table 1: Performances of multi-modal learning methods measured by the mean average precision on two text/image datasets.

of image tagging tasks (Hwang and Grauman 2012; Gong et al. 2014). The *TVGraz* dataset contains 2058 text/image pairs of 10 semantic categories with an average document length of 289 words, and is split into a training set of 1558 text/image pairs and a testing set of 500 text/image pairs. The *Wikipedia* dataset contains by 2866 text/image pairs of 30 semantic categories, and is split into training and testing sets with 2173 and 693 text/image 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 coPLSA learning algorithm on the training sets. We extract 50 topics from the *TVGraz* dataset and 100 topics from the *Wikipedia* dataset, and choose the balance parameter $\lambda$ in coPLSA algorithms with cross-validation on a subset of the training data. When performing retrieval tasks on the testing set, we first recover the topic composition of the queried image or text using the PLSA algorithm (only the optimization of $H$ matrix) using the learned modality-specific topics. The similarities between topic compositions of the queried image and texts in testing set (in task *i2t*) or queried text and images in testing set (in task *t2i*) 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 (Pereira et al. 2014). 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 1 summarizes the overall performances of coPLSA algorithms with symmetric KL, $\ell_2$ and $\ell_1$ co-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 implemented as described in (Cohn and Hofmann 2001). Furthermore, all results were compared to a baseline established by the method of *semantic correlation matching* (SCM) (Pereira et al. 2014), which represents the state-of-the-art performance in text/image cross-modal retrieval tasks. Results in Table 1 suggests that for the two cross-modal retrieval tasks, coPLSA 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 text/images) obtained with coPLSA, 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 coPLSA algorithms, suggesting the diminished representational difference between the two modalities in the topic space found by coPLSA. Furthermore, all three variants of the coPLSA algorithms achieves better performance and efficient computation, but symmetric KL co-regularizer leads to the best overall performance. Last, combining with more semantic abstraction, as concluded in (Pereira et al. 2014), can also significantly improve the retrieval performance.

In Fig.2 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 coPLSA 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 coPLSA span wide semantic ranges, and can shed some light on their effects in improving the precisions of semantic matching with the queried text document.

## Conclusions

We study co-regularized PLSA (coPLSA) for topic analysis of multi-modal data and derive efficient iterative learning algorithms for coPLSA with three types of divergences as co-regularizers, in each case the essential optimization problem affords simple solutions that entail only matrix arithmetic operations and numerical solution of 1D nonlinear equations. We evaluate the performance of the proposed coPLSA algorithms on cross-modal retrieval tasks involving text/image documents and show competitive performance with state-of-the-art methods. In future work, we can extend the coPLSA algorithms to more than two data modalities, and to datasets in which each data entity may not associate with all modalities.

## Acknowledgement

This work is supported by US National Science Foundation Research Grant (CCF-1319800) and National Science Foundation Early Faculty Career Development (CAREER) Award (IIS-0953373).

## References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. volume 3, 993–1022. Journal of Machine Learning Research.

Brand, M. 1999. Pattern discovery via entropy minimization. In *In Uncertainty 99: AISTATS 99*.

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

Corless, R. M.; Gonnet, G. H.; Hare, D. E. G.; Jeffrey, D. J.; and Knuth, D. E. 1996. On the lambert w function. In *Advances in Computational Mathematics*, 329–359.

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

Gaussier, E., and Goutte, C. 2005. Relation between PLSA and NMF and implications. In *Special Interest Group on Information Retrieval*, 601–602.

Gong, Y.; Ke, Q.; Isard, M.; and Lazebnik, S. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision* 106(2):210–233.

He, X.; Kan, M.-Y.; Xie, P.; and Chen, X. 2014. Comment-based multi-view clustering of web 2.0 items. In *International Conference on World Wide Web*.

Hofmann, T. 1999. Probabilistic latent semantic analysis. In *Association for Uncertainty in Artificial Intelligence*, 289–296.

Hwang, S. J., and Grauman, K. 2012. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *International Journal of Computer Vision* 134–153.

Jialu Liu, Chi Wang, J. G., and Han, J. 2013. Multi-view clustering via joint nonnegative matrix factorization. In *SIAM Data Mining Symposium*.

Jiang, Y.; Liu, J.; Li, Z.; Li, P.; and Lu, H. 2012a. Co-regularized PLSA for multi-view clustering. In *Asian Conference on Computer Vision*.

Jiang, Y.; Liu, J.; Li, Z.; and Lu, H. 2012b. Collaborative PLSA for multi-view clustering. In *International Conference on Pattern Recognition*.

Khan, I.; Saffari, A.; and Bischof, H. 2009. Tvgraz: Multi-modal learning of object categories by combining textual and visual features. In *AAPR Workshop*, 213–224.

Liu, Y.; Niculescu-Mizil, A.; and Gryc, W. 2009. Topic-link lda: Joint models of topic and author community. In *ICML*, ACM International Conference Proceeding Series, 84. ACM.

Mahadevan, V.; Wah Wong, C.; Costa Pereira, J.; T. Liu, T.; Vasconcelos, N.; and K. Saul, L. 2011. Maximum Covariance Unfolding: Manifold Learning for Bimodal Data. In *Advances in Neural Information Processing Systems*, 918–926.

Mao, X.; Lin, B.; Cai; He, X.; and Pei, J. 2013. Parallel field alignment for cross media retrieval. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, 897–906.

Nallapati, R., and Cohen, W. 2008. Link-PLSA-LDA: A new unsupervised model for topics and influence in blogs. In *Association for the Advancement of Artificial Intelligence*, 84 – 92.

Nallapati, R.; Ahmed, A.; Xing, E. P.; and Cohen, W. W. 2008. Joint latent topic models for text and citations. In *Knowledge Discovery and Data Mining*, 542–550. ACM.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *International Conference on Machine Learning (ICML)*.

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

Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G.; Levy, R.; and Vasconcelos, N. 2010. A New Approach to Cross-Modal Multimedia Retrieval. In *ACM International Conference on Multimedia*, 251–260.

Rosen-Zvi, M.; Griffiths, T. L.; Steyvers, M.; and Smyth, P. 2004. The author-topic model for authors and documents. *Uncertainty in Artificial Intelligence*.

Shashanka, M.; Raj, B.; and Smaragdis, P. 2007. Sparse overcomplete latent variable decomposition of counts data. In Platt, J.; Koller, D.; Singer, Y.; and Roweis, S., eds., *Advances in Neural Information Processing Systems 20*. 1313–1320.

Srivastava, N., and Salakhutdinov, R. 2012. Multimodal learning with deep Boltzmann machines. In *Advances in Neural Information Processing Systems 25*, 2231–2239.

Vinokourov, A.; Shawe-taylor, J.; and Cristianini, N. 2002. Inferring a semantic representation of text via cross-language correlation analysis. In *Advances in Neural Information Processing Systems*.

Virtanen, S.; Jia, Y.; Klami, A.; and Darrell, T. 2012. Factorized multi-modal topic model. *Uncertainty in Artificial Intelligence*.